Neuroscience 258 (2014) 292–306

DETECTION AND IDENTIFICATION OF SPEECH SOUNDS USING CORTICAL ACTIVITY PATTERNS

T. M. CENTANNI, * A. M. SLOAN, A. C. REED, C. T. ENGINEER, R. L. RENNAKER II AND M. P. KILGARD

University of Texas at Dallas, United States

Abstract—We have developed a classifier capable of locating and identifying speech sounds using activity from rat auditory cortex with an accuracy equivalent to behavioral performance and without the need to specify the onset time of the speech sounds. This classifier can identify speech sounds from a large speech set within 40 ms of stimulus presentation. To compare the temporal limits of the classifier to behavior, we developed a novel task that requires rats to identify individual consonant sounds from a stream of distracter consonants. The classifier successfully predicted the ability of rats to accurately identify speech sounds for syllable presentation rates up to 10 syllables per second (up to 17.9 ± 1.5 bits/s), which is comparable to human performance. Our results demonstrate that the spatiotemporal patterns generated in primary auditory cortex can be used to quickly and accurately identify consonant sounds from a continuous speech stream without prior knowledge of the stimulus onset times. Improved understanding of the neural mechanisms that support robust speech processing in difficult listening conditions could improve the identification and treatment of a variety of speech-processing disorders.

Key words: classifier, rat, auditory cortex, coding, temporal patterns.

INTRODUCTION

Speech sounds evoke unique spatiotemporal patterns in the auditory cortex of many species (Kuhl and Miller, 1975; Eggermont, 1995; Engineer et al., 2008). Primary auditory cortex (A1) neurons respond to most consonants, which evoke short, transient bursts of neural activity, but respond with different spatiotemporal patterns for different sounds (Engineer et al., 2008). For example, the consonant /d/ evokes activity first in neurons tuned to high frequencies, followed by neurons tuned to lower frequencies. The sound /b/ causes the opposite pattern such that low-frequency neurons fire approximately 20 ms before the high-frequency neurons (Engineer et al., 2008; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012b). These patterns of activity can be used to identify the evoking auditory stimulus in both human (Steinschneider et al., 2005; Chang et al., 2010; Pasley et al., 2012) and animal auditory cortex (Engineer et al., 2008; Mesgarani et al., 2008; Huetz et al., 2009; Bizley et al., 2010; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012a; Centanni et al., 2013a).

Rats are a good model of human speech sound discrimination as these rodents have neural and behavioral speech discrimination thresholds that are similar to humans. Rats can discriminate isolated human speech sounds with high levels of accuracy (Engineer et al., 2008; Perez et al., 2012; Centanni et al., 2013a). Rats and humans have similar thresholds for discriminating spectrally-degraded speech sounds, down to as few as four bands of spectral information (Ranasinghe et al., 2012b). Rats and humans are both able to discriminate speech sounds when presented at 0-dB signal to noise ratio (Shetake et al., 2011).

In both rats and humans, sounds that evoke different patterns of neural activity are more easily discriminated behaviorally than sounds that evoke similar patterns of activity (Engineer et al., 2008; Shetake et al., 2011; Ranasinghe et al., 2012b). Speech sounds presented in background noise evoke neural response patterns with longer latency and lower firing rate than speech presented in quiet and the extent of these differences is correlated with behavioral performance (Martin and Stapells, 2005; Shetake et al., 2011). Neural activity patterns in anesthetized rats also predict behavioral discrimination ability of temporally degraded speech sounds (Ranasinghe et al., 2012b).

The relationship between neural activity and associated behavior is often analyzed using minimum distance classifiers, but classifiers used in previous studies typically differ from behavioral processes in one key aspect: the classifiers were provided with the stimulus onset time, which greatly simplifies the problem of speech classification (Engineer et al., 2008; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012a; Centanni et al., 2013a,b). During natural listening, stimulus onset times occur at irregular intervals. One possible correction allows a classifier to look through an entire recording sweep, rather than only considering activity immediately following stimulus onset. The classifier then guesses the location and

*Corresponding author. Address: University of Texas at Dallas, 800W Campbell Road, GR 41, Richardson, TX 75080, United States. Tel: +1-972-883-2376; fax: +1-972-883-2491.
E-mail address: tmcentanni@gmail.com (T. M. Centanni).

Abbreviations: A1, primary auditory cortex; CF, characteristic frequency; CVC, consonant–vowel–consonant; FM, frequency modulated; IR, infrared; ISI, inter-stimulus-interval; NM, normalized metric; SPL, sound pressure level; sps, syllables per second.

0306-4522/13 $36.00
http://dx.doi.org/10.1016/j.neuroscience.2013.11.030

identity of the sound post hoc by picking the location most similar to a template (Shetake et al., 2011). While this method is highly accurate and predicts behavioral ability without the need to provide the onset time, the method could not be implemented in real time and assumes that a stimulus was present. We expected that large numbers of recording sites would be able to accurately identify a sound's onset, since the onset response in A1 to sound is well known (Anderson et al., 2006; Engineer et al., 2008; Dong et al., 2011; Centanni et al., 2013b). We hypothesized that with many recording sites, A1 activity can also be used for identification of the sound with a very brief delay consistent with behavioral performance in humans and animals.

EXPERIMENTAL PROCEDURES

Speech stimuli

For this study, we used the same stimuli as several previous studies in our lab (Engineer et al., 2008; Floody et al., 2010; Porter et al., 2011; Shetake et al., 2011; Ranasinghe et al., 2012b). We used nine English consonant–vowel–consonant (CVC) speech sounds differing only by the initial consonant: (/bad/, /dad/, /gad/, /kad/, /pad/, /sad/, /tad/, /wad/, and /zad/), which were recorded in a double-walled, soundproof booth spoken by a female native-English speaker. The spectral envelope was shifted up in frequency by a factor of two while preserving all spectral information using the STRAIGHT vocoder (Kawahara, 1997) to better accommodate the rat hearing range. The intensity of each sound was calibrated with respect to its length, such that the loudest 100 ms was presented at 60-dB sound pressure level (SPL) and 5 ms on and off ramps were added to prevent any artifacts.

Surgical procedure – Anesthetized recordings

Multi-unit recordings were acquired from the A1 of anesthetized, experimentally-naïve female Sprague–Dawley rats (Charles River, Wilmington, MA, USA). Recording procedures are described in detail elsewhere (Engineer et al., 2008; Shetake et al., 2011; Ranasinghe et al., 2012b). In brief, animals were anesthetized with pentobarbital (50 mg/kg) and were given supplemental dilute pentobarbital (8 mg/ml) as needed to maintain areflexia, along with a 1:1 mixture of dextrose (5%) and standard Ringer's lactate to prevent dehydration. A tracheotomy was performed to ensure ease of breathing throughout the experiment and filtered air was provided through an air tube fixed at the open end of the tracheotomy. Craniotomy and durotomy were performed, exposing right A1. Four Parylene-coated tungsten microelectrodes (1–2 MΩ) were simultaneously lowered to layer (4/5) of right A1 (~600 μm). Electrode penetrations were marked using blood vessels as landmarks.

Brief (25 ms) tones were presented at 90 randomly interleaved frequencies (1–47 kHz) at 16 intensities (0–75 dB SPL) to determine the characteristic frequency (CF) of each site. A set of four stimuli were created using Adobe Audition for comparison to our behavioral task (described below). Each stimulus consisted of a train of six individual speech sounds such that across all four sequences, all 24 possible sound pairs were presented once (/bad bad gad sad tad dad/, /tad tad sad gad bad dad/, /gad gad tad bad sad dad/, /sad sad bad tad gad dad/). The temporal envelope of the stimuli was compressed so that when presented with a 0-s inter-stimulus-interval (ISI), sounds were presented at 2, 4, 5, 6.7, 10 and 20 syllables per second (sps). All speech stimuli were randomly interleaved, and presented at 20 repeats per recording site. All sounds were presented approximately 10 cm from the left ear of the rat. Stimulus generation, data acquisition and spike sorting were performed with Tucker-Davis hardware (RP2.1 and RX5) and software (Brainware).

Surgical procedure – Awake recordings

Rats were anesthetized and implanted with a chronic array of 16 polyimide-insulated 50-μm diameter tungsten microwires. The implantation surgery and microwire arrays have been previously reported in detail (Rennaker et al., 2005). Briefly, subjects were anesthetized with an intramuscular injection of a mixture of ketamine, xylazine and acepromazine (50, 20, 5 mg/kg, respectively). Atropine and dexamethazone were administered subcutaneously prior to and following surgery. A midline incision was made, exposing the top of the skull, and a section of the right temporalis muscle was removed to access A1. Six bone screws were fixed to the dorsal surface of the skull (two in each parietal bone and one in each frontal bone) to provide structural support for the head cap. The two middle screws had attached leads to serve as a reference wire and a grounding wire. Craniotomy and durotomy were performed to expose the cortex in the region of A1. The microwire array was then inserted to a depth of 550–600 μm (layer IV/V) in A1 using a custom-built mechanical inserter (Rennaker et al., 2005). The area was sealed with a silicone elastomer (Kwik-Cast, World Precision Instruments Inc., Sarasota, FL, USA) and the head cap was built with a connector secured with acrylic. Finally, the skin around the implant was sutured in the front and the back of the head cap. Subjects were given prophylactic minocycline in water *ad libitum* for 2 days prior to and 5 days following surgery to lessen immune responses (Rennaker et al., 2005), and were also given Rimadyl tablets for 3 days after surgery to minimize discomfort. Topical antibiotic was applied to the incision to prevent infection. After a minimum of 5 days of recovery, neural activity was collected in a single 2.5-h session and saved using a custom MATLAB program. The session included an abridged tuning curve (to assess each site's best frequency) and the same set of speech sequence stimuli presented to the anesthetized animals. All passive sound sets were created and run through custom MATLAB programming.

Neural analysis and classifier

We designed a classifier that does not require precise information about the sound's onset time by modifying

a well-established classifier (Foffani and Moxon, 2004; Schnupp et al., 2006; Engineer et al., 2008; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012b; Centanni et al., 2013a,b). The classifier searched neural activity (from a randomly selected subgroup of sites) for the onset of a sound by locating a pattern of activity observed in the average response to many repetitions. We trained the classifier to recognize patterns of activity evoked by several auditory stimuli (Fig. 1A), by providing the mean activity across 19 repeats of each stimulus (Fig. 1D–G, right panels). The classifier then analyzed the neural activity generated by a single presentation (Fig. 1B) to determine whether one of the trained sounds occurred. The classifier calculated a unique decision threshold for each possible consonant sound in order to allow the classifier to determine which sound most likely caused the activity. A classifier decision was registered within 40 ms of stimulus onset because this was the duration of the stored template for each onset pattern. To calculate the thresholds, the classifier compared the similarity between each average pattern of activity (template) to the response of each single repeat (20 repeats $\times$ 20 speech sounds). To reduce false alarms caused by spontaneous activity, data were smoothed across similarly tuned recording sites using a Gaussian filter (Giraud et al., 2000; Langers et al., 2003, 2007) (Fig. 1C). The classifier evaluated a variety of half-widths for this spatial filter; from 1% of the total number of sites all the way to 50% of sites (Fig. 2A). The most effective filter had a half-width including 15% of the total number of sites, and was used for all results reported here.

Euclidean distance was used to measure the similarity between the single trial and each template and was calculated by taking the square root of the sum of the squared differences between two patterns of neural activity. The threshold value for each sound was set to ensure the maximum number of correct responses while minimizing the number of false alarms (Fig. 2B). The threshold was calculated using the equation:

$$th = \max(|ED_m \sim -ED_u \sim|)$$

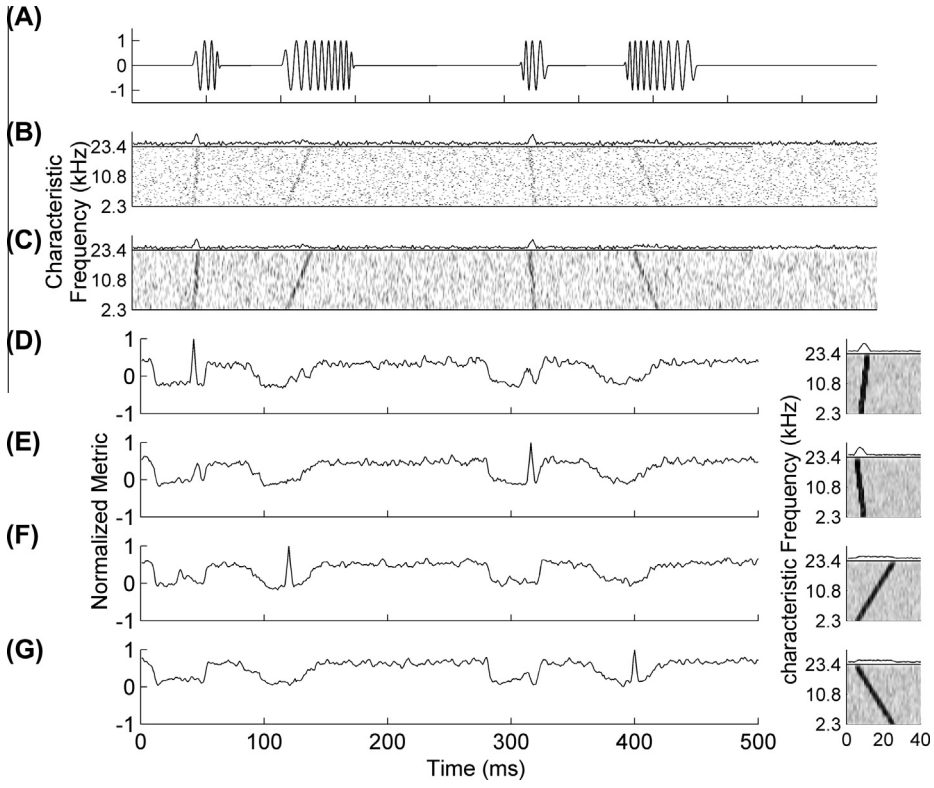

Fig. 1. Schematic of the classifier using simulated neural data. (A) Example waveform of four frequency-modulated (FM) sweeps; one fast sweep from low to high frequency, one slow sweep from low to high frequency, one fast sweep from high to low frequency, and one slow sweep from high to low frequency. (B) Simulated single-trial neural response from 200 sites, organized by characteristic frequency. This single trial is shown without any smoothing or other manipulation and shows that the evoked activity patterns are difficult to distinguish from spontaneous firing. (C) The same activity from panel B after a Gaussian filter was applied to the spectral dimension. This filter had a half-width of 15% of the total number of sites and effectively highlighted the evoked patterns while minimizing the influence of spontaneous action potentials. (D–G) Examples of the classifier locating and identifying each of the four FM sweeps. The template is shown to the right of each panel and is created from the average neural activity across 19 sweeps. The normalized metric (NM) shown is calculated by comparing the Euclidean distance between the single-trial response and the template, and then normalizing by the threshold values. NM values at 1 indicate a guess by the classifier. In a case of two templates making a guess, the classifier would use the larger of the two raw values as a tie breaker (e.g. the value that is most similar).

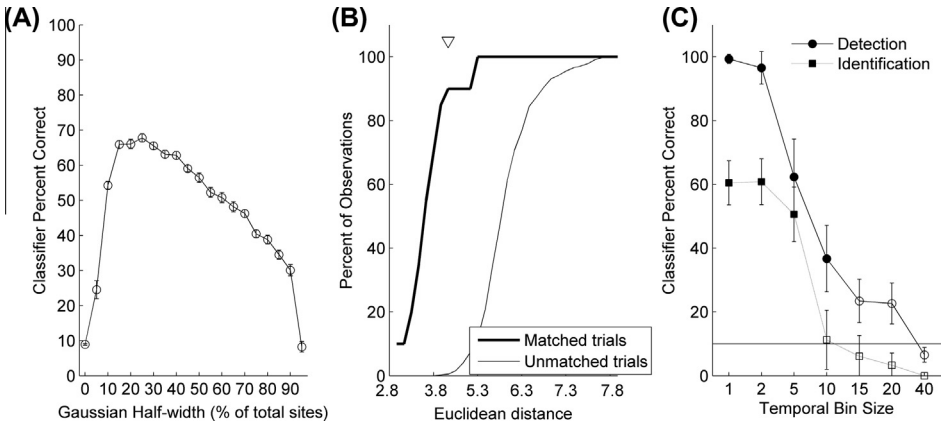

Fig. 2. (A) Average classifier performance using 150 neural recording sites after different amounts of Gaussian smoothing. Data were smoothed using a Gaussian filter with varying half-widths from 0% to 90% of the total number of sites. The classifier then used the resulting datasets to attempt to locate and identify each of nine consonant sounds. Though half-width between 10% and 20% were highly accurate, a half-width of 15% of the total number of sites was optimal. (B) Decision thresholds were calculated by comparing single-trial neural responses to the average evoked response to each consonant sound. For example, to create the decision threshold for the sound /sad/, the average response to this sound (over 19 repeats) was compared to all single-trial responses to every sound. The similarity of the single trials to the template was calculated using Euclidean distance. We then plotted the distributions of Euclidean distance values generated when the single trials were evoked by the template sound (e.g. Matched trials: when template and single trial were both evoked by /sad/) vs. the Euclidean distance values when the template did not match the single trial (e.g. Unmatched trials: when the template was evoked by /bad/ while the template was evoked by /sad/). The decision threshold was then set at the point at which the distributions were most different, as marked by a triangle in the bottom half of the figure. This maximized the sensitivity index so that the most correct answers were preserved while excluding the maximum number of false alarms. (C) Mean detection (circle markers) and identification (square markers) performance of the classifier using different temporal bin sizes. Error bars represent standard error of the mean. Filled circles represent values significantly above chance performance (10%). As expected from previous studies, our classifier performs significantly better when spike timing information is preserved (e.g. temporal bins smaller than 10 ms). The classifier is still able to correctly signal that a sound occurred using bins between 10 and 20 ms, but begins to false alarm to silence when spike timing information is completely removed (e.g. 40 ms bins). This result suggests that the number of action potentials can be used to locate the onset of a sound, but precise spike timing information is required for consonant identification.

where th is the threshold being calculated, $ED_m\sim$ is the discretized distribution of Euclidean distance values calculated between the template and the single-trial responses evoked by the template sound (matched comparisons: e.g. the average response to /dad/ compared to a single-trial response to /dad/) and $ED_u\sim$ is the discretized distribution of Euclidean distance values calculated between the template and the single-trial responses evoked by a different sound (unmatched comparisons: e.g. the average response to /dad/ compared to a single-trial response to /bad/, /gad/, /pad/, /kad/, /tad/, /sad/, /zad/, or /wad/).

There is significant variability in the difference between templates because some sounds trigger a larger neural response than others (Fig. 3A). To compensate for the variability in neural response to each sound, the classifier normalized the data to center all comparisons on a single scale. This was accomplished by calculating a normalized metric (NM) of Euclidean distance values for each single trial so that the values centered on 0 and templates similar to the single trial returned positive values while templates less similar to the single trial returned negative values (Fig. 3B). This was done using the equation:

$$\mathrm{NM}_c = -\left(\frac{(Ed_c - ED_{sp}) * th_{\min}}{th_c}\right)$$

where c is the window currently being analyzed, ED_c is the Euclidean distance between that point and the template, ED_{sp} is the Euclidean distance between the template and spontaneous activity, th_c is the threshold for the template and $th_{\min}$ is the minimum threshold across all nine templates.

The classifier searched each single-trial recording sweep and identified when a pattern of activity occurred (when a threshold was crossed) and which stimulus caused that pattern of activity. If more than one template crossed the threshold within a single bin, the classifier chose the template with the highest NM value; e.g. the template that was closest to the single trial being analyzed. To count as a correct guess, the classifier must recognize the sound within 40 ms of the stimulus onset, which is considerably shorter than the 500-ms hit window the rats were given to respond behaviorally. The longer behavioral hit window allowed for processing time and motor movement of the animal, while the information needed for the classifier to guess was contained in the first 40 ms (Miller and Nicely, 1955; Kuhl and Miller, 1975; Engineer et al., 2008; Porter et al., 2011; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012b). The classifier was run thirty times, with a different randomly-sampled neural population for each run. For comparison to behavioral performance, the average percent correct for each run was calculated and plotted with the average last day behavioral performance of rats trained on speech sound discrimination tasks. The strength of the correlation was measured using the Pearson correlation coefficient.

To consider the amount of information present in the neural response to each speech sound, we calculated

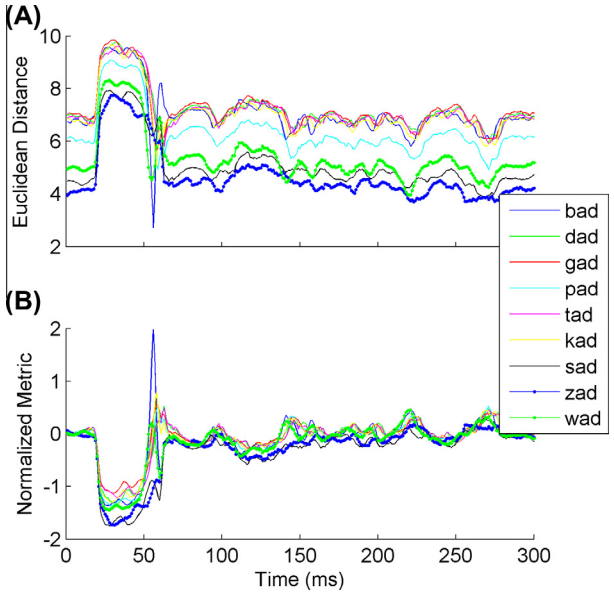

Fig. 3. Example of normalized metric calculation using speech sounds. (A) An example of the Euclidean distance values calculated between a single trial and each of the templates. The Euclidean distance between each template and the spontaneous firing at the beginning of the sweep were highly variable and each comparison was therefore on a different scale, making comparisons difficult. (B) To allow for a more accurate comparison across templates, we normalized the Euclidean distance values using the comparison between the template and spontaneous firing as well as the threshold value for that template (see Experimental procedures for detailed equation). The comparison values were then centered on 0 and values indicating similarity were positive and values indicating difference were negative.

the number of bits present in the neural activity pattern evoked by each stimulus. Bits were calculated using the following equation:

$$\text{bits} = \left(pxy * \log_2\left(\frac{pxy}{(px * py)}\right)\right) * pr$$

where *pxy* is the percent of correct guesses, *px* is the number of target sounds, *py* is the number of guesses for this sound, and *pr* is the number of sounds presented in one second (Brillouin, 2013). Chance performance for classifier tasks was 10% since each sound was approximately 400-ms long and the classifier was only correct if it guessed within 40 ms of the sound's onset.

Simulation of correlated data

Neural recordings were acquired in groups of four simultaneously-recorded electrode locations. This arrangement caused our neural recordings to be uncorrelated with each other. To evaluate whether this would bias our classifier performance, we simulated the amount of correlated firing activity that would be expected if all sites were recorded at the same time. To pseudo-correlate the data, we evaluated the average percentage of sites that fired at each time point during a recording sweep using 1-ms bins. We then compared these values to those in the general population. At any time point where the proportion of firing sites in the full dataset was less than the proportion of firing sites across simultaneously recorded sites, single action potentials were iteratively added at that time point to randomly chosen sites. Action potentials were added until the proportion of sites firing across the entire dataset matched the proportion of simultaneously recorded sites firing at each millisecond time point.

Behavioral paradigm

Sprague–Dawley albino rats were trained using either an established lever press paradigm for the isolated speech task (Engineer et al., 2008; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012b) or an operant nose poke paradigm (for the speech sequence task), developed by Dr. Robert Rennaker (Sloan et al., 2009). Each rat trained for two 1-h sessions per day, 5 days per week. For the isolated speech task, the behavioral training procedures and data reported here were the same as was reported in Engineer et al. (2008). In brief, six rats were trained to press a lever when a target speech sound was presented and to withhold pressing when a distracter sound was presented (/d/ vs. /s/, /d/ vs. /t/, and /d/ vs. /b/ and /g/). Rats were trained for 2 weeks on the tasks in the order given and performance was assessed after training on each task to obtain the values reported in Engineer et al. (2008) and the current study.

For the speech sequence task, all animals were first trained to associate the infrared (IR) nose poke with the sugar pellet reward. Each time the rat broke the IR beam, the target speech sound (/dad/) was played and a 45-mg sugar pellet was dispensed. After each animal earned over 100 pellets, each rat was moved onto a series of training stages, during which d' was used as a criterion for advancing to the next stage (Green and Swets, 1966). During the first training stage, rats were trained to wait patiently in the nose poke and withdraw their nose from the nose poke after hearing the target. This stage lasted until the animal performed with a d' greater than 1.5 for 10 sessions. For these first two stages, the animal had a response window of 800 ms to withdraw their nose in response to the target.

Rats were then introduced to the four possible distracters by presenting a string of repeats of a single distracter prior to the presentation of the target. The ISI was 1 s and the response window was also reduced to 650 ms during these stages. Since the task involved random patterns of distracters, we trained the animals on a fixed pattern of distracters to introduce the concept of multiple distracters per trial. For each trial in this stage, two or three of the four distractors were randomly selected and alternated. In the final two training stages, a sequence for each trial was randomly generated using all four possible distracters and presented to ensure that the rat could not memorize the pattern or time their responses. In addition, the ISI was reduced to 0 s and the response window was reduced to 500 ms. Once rats performed with a $d' > 1.5$ for at least two sessions, they were introduced to each presentation rate. During this period of training, rats were presented with blocks of 20 trials each. Each trial contained a random hold time (the time before the onset of the target sound) between 2 and 7 s, with the sounds prior to the target consisting of

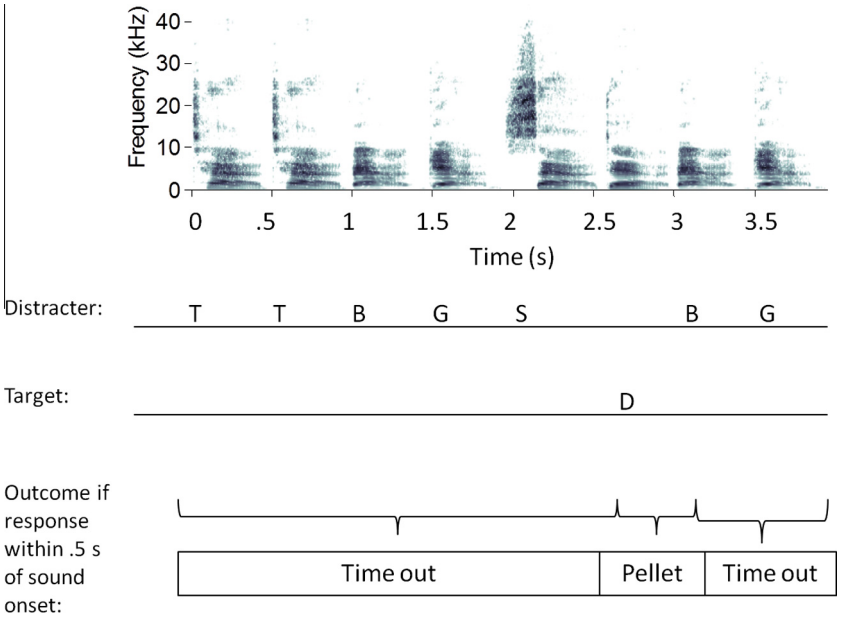

Fig. 4. Schematic of the behavioral task. Speech sounds are presented in random order beginning when a rat breaks the infrared (IR) beam. Target sound (/dad/) was presented in a single random location anywhere from the third sound of the sequence until the end of the 2–7 s trial. From the onset of the target sound, rats had 500 ms to respond by withdrawing from the IR beam. If the target sound was less than 500 ms long, additional distracters were added afterward to avoid the use of silence as a cue. Correct responses to the target were rewarded with a 45-mg sugar pellet. Incorrect responses to distracter sounds or missed responses to the target were punished by a 5-s timeout in which the booth lights were extinguished and the IR beam was disabled.

randomly selected distracters (Fig. 4). The presentation rate of each block was either 2 sps or one of the additional presentation rates. These blocks were presented in random order. 20% of trials were catch trials in which no target was presented to ensure the rats were listening for the target and not attempting to time the target location (Fig. 4).

Animals were tested for a minimum of 10 sessions during which all six presentation rates were randomly interleaved. The animals were individually housed with free access to water and were food deprived to no less than 85% body weight while experimenting. When not experimenting, they were given free access to water and food and housed on a reverse 12:12 light/dark schedule. The behavioral paradigm was written and executed via a custom-designed MATLAB (The Mathworks Inc., Natick, MA, USA) program and run through a PC computer with an external real time digital-to-analog processor (RP2.1; Tucker-Davis Technologies, Alachua, FL, USA), which monitored the IR nose poke and controlled the stimuli presentation and lights. Each of the five sounds was analyzed for response rate (number of responses/number of presentations * 100). Target responses are referred to as hits and the summed response to all four distracters is referred to as false alarm rate. Overall performance is reported in terms of hits-false alarms per presentation rate. All protocols and recording procedures were approved by the University of Texas at Dallas Institutional Animal Care and Use Committee (Protocol Number: 99-06). All surgeries were performed under either pentobarbital or ketamine anesthesia and all efforts were made to minimize suffering.

RESULTS

Neural activity patterns predict stimulus identity

Our classifier was tested using previously published neural activity evoked by nine different consonant sounds (Engineer et al., 2008) (Fig. 5). The first test of the classifier used 2-ms temporal bins over an 80-ms sliding window (which created an analysis window of 40 units), which is similar to the temporal parameters used in previous studies (Engineer et al., 2008). Overall, this classifier performed at chance levels (10% chance vs. 10.7 ± 0.6% correct; unpaired *t*-test, $p = 0.86$; Fig. 2A). We hypothesized that the poor performance of the classifier was due to the influence of un-correlated, spontaneous activity across channels. Since our recordings were acquired in groups of four sites at a time, some sites fired in the absence of the population response, causing many false alarms for the classifier. To attain a more reliable estimate of the neural activity in each frequency band, we used a Gaussian filter to smooth activity along the tonotopic spatial dimension. This method ensured low variability, e.g. noise in the neural signal (Bear et al., 1987; Sengpiel and Kind, 2002; Poirazi et al., 2003a,b; Hao et al., 2009) (Fig. 5B). Although a range of Gaussian filter half-widths generated high accuracy from the classifier (10–30% of the total number of sites), a half-width of 15% of the total number of sites was optimal (Fig. 2A).

After Gaussian smoothing, the classifier was able to perform the identification task with high levels of accuracy (58.3 ± 5.5%; Fig. 6A). As expected from previous studies, this classifier relied on spike timing information (2-ms temporal bins) and was not

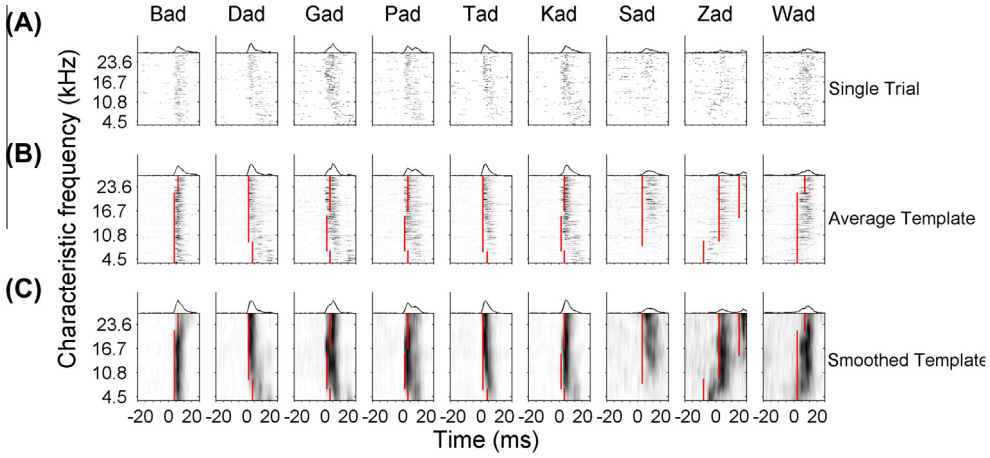

Fig. 5. A Gaussian filter was necessary for highlighting evoked activity. (A) Single-trial neural activity patterns in A1 without any smoothing. The first 40 ms of average evoked activity from each site is organized by characteristic frequency. Each consonant evoked a unique pattern of activity such that each group of neurons fire at a different latency depending on the characteristic frequency of the group. (B) Average activity over 20 trials plotted without smoothing. Red lines mark the onset response of each frequency group. (C) The same neural activity plotted in panel B after a Gaussian filter has been applied to the spectral dimension. We used a filter with a half-width of 15% of the total number of sites. This ensured that spontaneous activity is not as influential on the classifier as evoked activity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

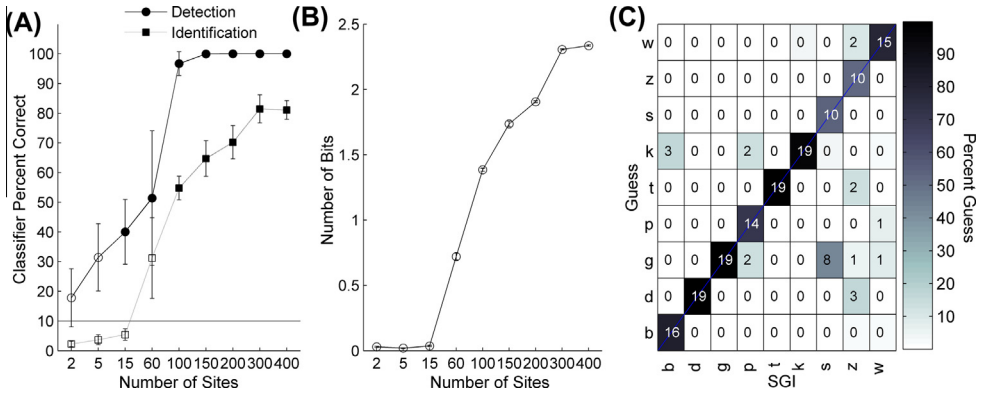

Fig. 6. A Euclidean distance classifier could locate and identify nine consonant speech sounds with high levels of accuracy. (A) The classifier was able to locate the onset of a speech stimulus with high levels of accuracy (circle markers), but required a larger number of sites to accurately identify the speech sound (square markers). Error bars represent standard error of the mean. Filled markers represent values significantly above chance performance (10%). This is likely due to the limited frequency range included in small groups of sites. Previous classifiers provided the stimulus onset time and were able to achieve high levels of accuracy using single sites of neural activity. (B) Number of bits encoded in various subgroups of sites. 60 sites were able to locate the sound onset, but could not identify the sound, as this number of sites contained less than 0.8 ± 0.03 bits of information. Larger groups of sites contained up to 2 bits of information (1.6 ± 0.01 with 400 sites) and were better able to perform the task. (C) Confusion matrix of classifier performance on nine English consonant sounds. The classifier performed the task with high levels of accuracy at every sound presented. The number of classifier guesses (out of 20 trials) is listed in each square of the matrix and the shading represents overall percent correct.

significantly different from chance performance (10% is chance level) when temporal bins of 10 ms were used (11.3 ± 9.3% correct using 10 ms bins; *t*-test vs. chance level, $p = 0.89$; Fig. 2C). When spike timing information was removed (by summing the number of action potentials in the 40-ms window) the classifier fell below chance level at both detection and identification tasks (Fig. 2C). As shown previously, when spike timing was no longer preserved, spontaneous activity could not be distinguished from evoked activity, and the classifier lost accuracy (Engineer et al., 2008; Huetz et al., 2009; Panzeri and Diamond, 2010; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012a).

As expected, spatiotemporal patterns of evoked neural activity are identifiable when neurons with a variety of CFs are recorded (Creutzfeldt et al., 1980; Wang et al., 1995; Bizley et al., 2010). For example, if the stimulus onset time was unknown and only one recording site was available for analysis, the classifier did not perform significantly above chance. Small numbers of sites (as few as 15) were able to detect the location of a stimulus onset significantly above chance (10% was chance performance; 60 site detection at 51.4 ± 22.6% correct, one-tailed *t*-test vs. chance performance, $p = 0.04$; Fig. 6A), but performed at chance when asked to identify the sound (60-site

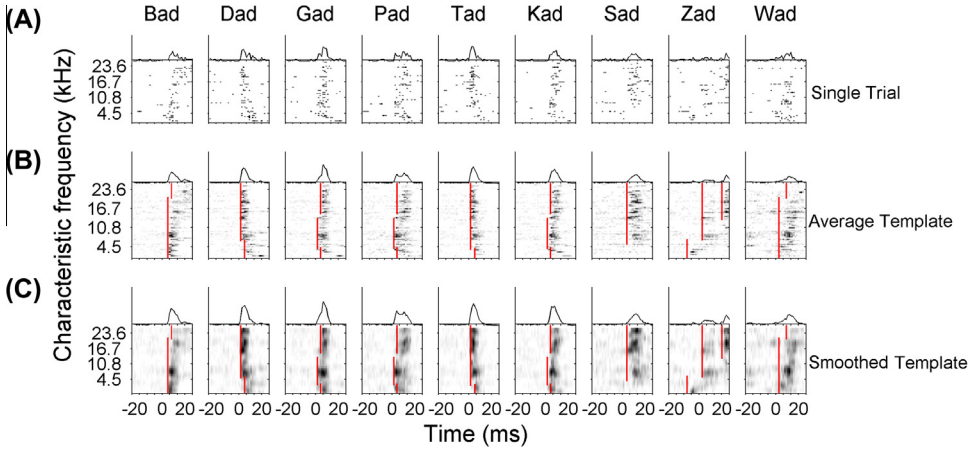

Fig. 7. Large numbers of sites are needed to encompass the complete frequency range. (A) Single-trial neural activity evoked by each of nine consonant sounds, organized by characteristic frequency and shown without any smoothing. As compared to the responses by 200 sites (see Fig. 2), these responses are less distinct even without averaging or smoothing. (B) Average activity over 20 trials plotted without smoothing. Red lines mark the onset response of each frequency group. (C) The same neural activity plotted in panel B after a Gaussian filter has been applied to the spectral dimension. We used a filter with a half-width of 15% of the total number of sites. Although this smoothing does highlight evoked activity over spontaneous firing, responses from 60 sites are not sufficient to produce clearly distinguishable patterns, especially as compared to the responses from 200 sites (see Fig. 2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

discrimination performance at 31.2 ± 13.6% correct, one-tailed *t*-test vs. chance performance, $p = 0.07$; Fig. 6A). This level of performance using 60 sites is likely due to the reduction in frequency information represented by this number of recording sites (Fig. 7) and corresponds with the amount of bits encoded with this number of sites (0.8 ± 0.03 bits of information) compared to almost 2.5 bits of information in a group of 400 sites (2.3 ± 0.01; Fig. 6B) (Brillouin, 2013).

This more comprehensive frequency range represented in larger site groups is needed for consonant identification (Fig. 6). For example, in response to the sound /dad/, sites with a CF above ~7 kHz responded to the consonant /d/ first, while lower frequency sites fired only to the onset of the vowel (Fig. 5). To the sound /tad/, the same pattern occurred, but the latency of the low-frequency sites was later than in response to /tad/. If only high-frequency sites were sampled, these two sounds would be indistinguishable (Fig. 5). When the classifier was given sites with a small band of CFs (below 6 kHz), average performance was 24.4 ± 9.5% (*t*-test vs. 10% chance performance; $p = 0.13$). Using large numbers of sites, this entire frequency range was represented and the classifier was able to perform the task well above chance level (Fig. 6A, C). These results validate that our new classifier is able perform the task using neural activity without specific knowledge of the stimulus onset time. In addition, our results show that a Euclidean distance classifier can perform with high levels of accuracy without being forced to guess.

Our new classifier performed significantly above chance at identifying speech sounds without prior knowledge of the stimulus onset time, but may not have performed the task as well as rats could behaviorally. We used behavioral data published in our previous report (Engineer et al., 2008) for comparison with this new classifier. Six rats were trained to press a lever when a target speech sound was presented and to withhold pressing when a distracter sound was presented (/d/ vs. /s/, /d/ vs. /t/, /d/ vs. /b/, and /d/ vs. /g/). Using groups of 150 recording sites, we ran the new classifier on these same, two-alternative forced-choice tasks. Our classifier performed with accuracy levels that were not significantly different from rats' behavioral performance (average classifier performance was 81.8 ± 13.0% vs. 88.3 ± 2.4% correct by the rats; unpaired *t*-test, $p = 0.59$). This result suggests that our new classifier performance was comparable to the rats' behavioral performance and may be applicable to a range of speech stimuli and new behavior tasks.

The classifier is able to identify speech sounds in sequences

Identifying the onset of a speech sound using neural activity is relatively easy when speech sounds are isolated (Fig. 6A). To test whether the classifier was limited to sounds presented in isolation, we tested the classifier's performance on neural responses to speech sounds imbedded in sequences. When the classifier crossed the threshold for an evoked response, it chooses the identity of the sound as the template with the highest value, even if multiple templates are similar enough to trigger a response (Fig. 8). The classifier was able to guess the location of the target sound (within 40 ms of the onset of the sound) with an accuracy of 65.5 ± 11.1% using random groups of 150 sites. Our new classifier is able to identify speech sounds without prior knowledge of the stimulus onset time, and does not rely on silent context.

The clear speech sequences we used were a bit unnatural and slower than conversational speech. We compressed the speech sounds so that our sequences could be presented at a variety of speeds that not only closely resemble conversational speed, but also test the

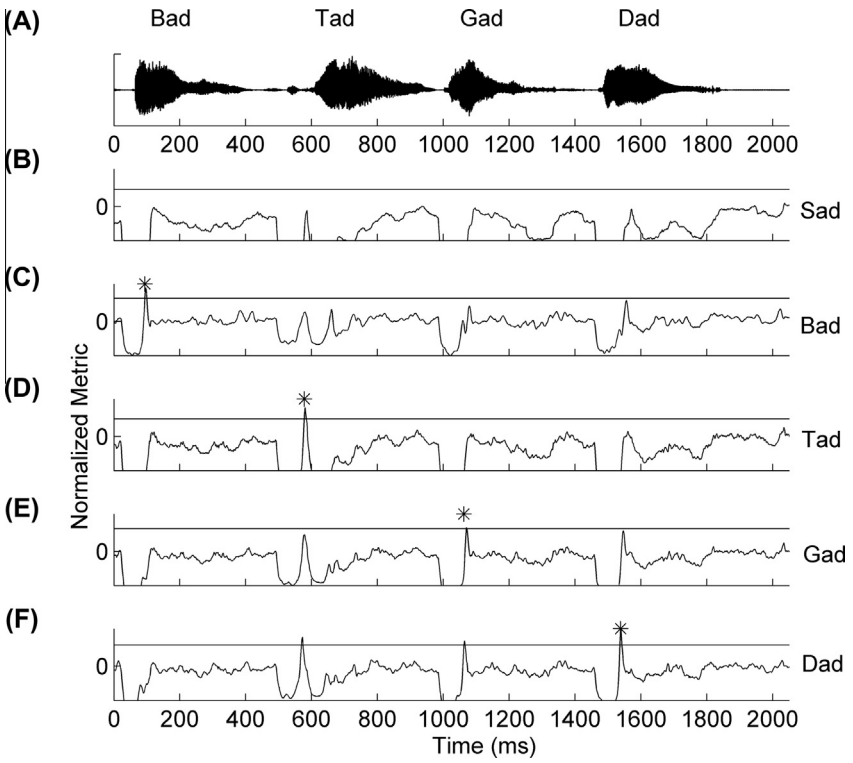

Fig. 8. Example classifier run on a four speech sound sequence. A single trial example of the classifier's performance on a speech sequence is plotted by template. The classifier analyzed a single-trial neural response to the sequence 'bad tad gad dad' by comparing the response to each of five templates. (A) Waveforms of the four speech sounds presented during anesthetized and awake mapping. (B–F) Examples of the comparison between each of five templates and the single-trial response to /bad tad gad dad/. The classifier detects that a sound has occurred when the NM reaches a value of 1, and the identity of the sound is the template with the highest NM value at that time point. Guesses are marked in the figure by asterisks.

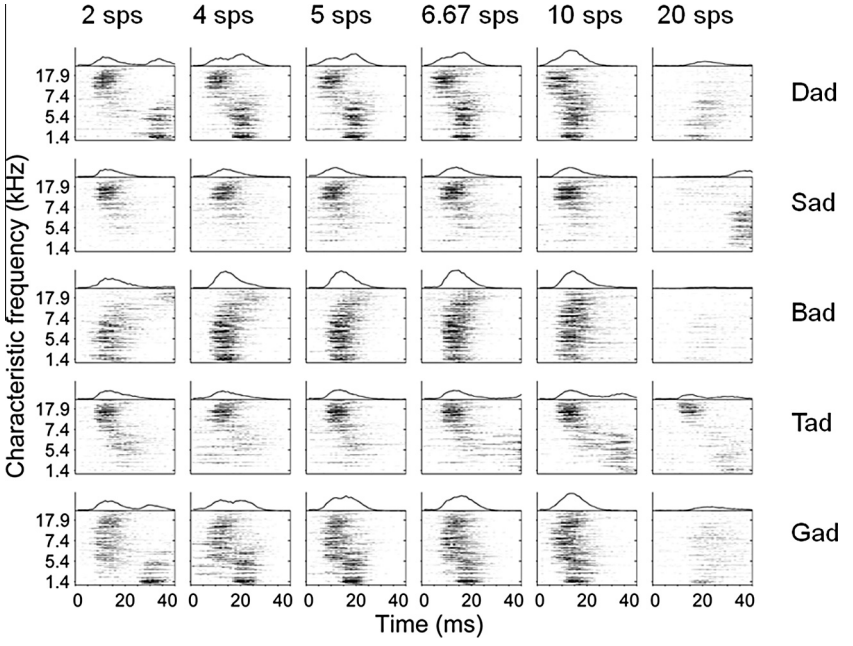

Fig. 9. Cortical speech-evoked activity patterns were robust up to 10 sps. Neural responses were averaged for each site and plotted organized by characteristic frequency. Each consonant speech sound (by row) evoked a unique pattern of activity at 2 sps (first column). The response of these patterns was robust through the 10-sps presentation rate. At 20 sps, responses were visibly weaker and were less distinct that at the previous presentation rates. This drastic change in neural responses may be the reason that both behavior and classifier performance fall at this speed.

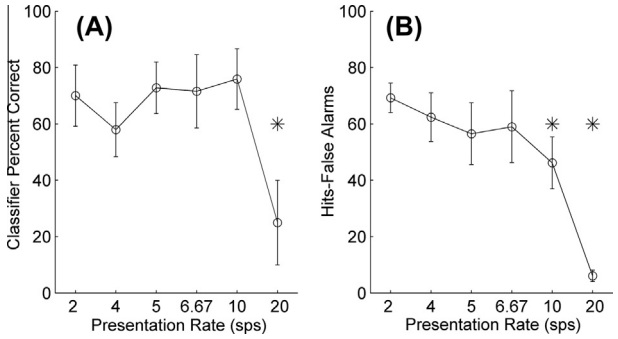

Fig. 10. Average performance of rats and the classifier on the speech sequence task. (A) Average classifier performance at each of the six presentation rates. Performance was calculated by counting the number of correct responses per sequence over 20 repeats of each sequence. This process was repeated 30 times with random groups of sites and average performance across the 30 runs is plotted. The classifier generated the expected performance curve for the behavioral task in rats. (B) Average behavioral performance by rats was measured by hits-false alarms for each of six presentation rates tested. Performance was plotted across a minimum of 10 sessions per rat of the testing stage in which all presentation rates were randomly interleaved in blocks of 20 trials per block (see Experimental procedures). Performance was robust until 10 and 20 sps (compared to performance at 2 sps; $^{*}p < 0.01$). The task was almost impossible for rats when sounds were presented at 20 sps ($^{**}p < 0.001$ as compared to 2 sps). Behavioral ability of rats was not significantly different from classifier performance (unpaired *t*-tests, $p = 0.65$, $p = 0.78$, $p = 0.35$, $p = 0.58$, $p = 0.16$, and $p = 0.14$ at each presentation rate, respectively).

temporal limits of the classifier. Neural activity patterns were strong and distinguishable at rates up to 10 sps (Fig. 9), and performance of the classifier was similarly robust up until 10 sps, and then performed significantly worse at 20 sps than at 2 sps (Fig. 10A). The significant reduction in neural firing strength at 20 sps as well as the impaired performance of the classifier at this speech mimics the temporal thresholds seen in human participants on rapid speech discrimination tasks (Ahissar et al., 2001; Poldrack et al., 2001; Ghitza and Greenberg, 2009). This result suggests that as long as neural response patterns are unique and are distinguishable from spontaneous firing, A1 activity can be used to locate and identify speech sounds in a sequence.

Since our classifier was able to accurately mimic behavioral ability on a two-alternative forced-choice task, we hypothesized that our real time classifier could predict rats' ability to identify a target speech sound in a stream of speech distracters. Rats were trained to initiate trials by engaging an IR nose poke, and to withdraw from the nose poke upon presentation of the target sound /dad/ (within a 500-ms hit window) and to withhold responding during preceding random sequences of four distracter sounds (Fig. 4; /bad/, /gad/, /sad/, and /tad/). This task required a longer learning period than previous studies of speech sound discrimination. Our rats required 38.2 ± 1.7 days to reach performance of $d' \geqslant 1.5$ compared to 17.4 ± 2.3 days for isolated speech tasks (Engineer et al., 2008). Behavioral discrimination accuracy gradually decreased as the presentation rate was increased.

Performance remained well above chance (0%) up to 10 sps (2 sps: 69.2 ± 5.2%, 4 sps: 62.4 ± 8.7%, 5 sps: 56.5 ± 10.9%, 6.67 sps: 59.0 ± 12.7%, 10 sps: 46.1 ± 9.2%), though performance at this rate was significantly worse than performance at 2 sps (46.1 ± 9.2% vs. 69.2 ± 5.2%, 10 sps vs. 2 sps respectively; paired *t*-test; $p = 0.007$). Poor performance at 20 sps (6.1 ± 2.0% correct) was consistent with performance in humans at the same rate (Ahissar et al., 2001; Poldrack et al., 2001; Ghitza and Greenberg, 2009) (Fig. 10B). At this speed, not only did hit rate decrease (paired *t*-test of 18.8 ± 7.1% vs. 47.7 ± 3.6% at 2 sps; $p < 0.01$), but the number of early responses (aborts) significantly increased (paired *t*-tests of 35.1 ± 5.7% vs. 16.3 ± 3.8% misses at 2 sps; $p < 0.01$ and paired *t*-tests of 33.6 ± 3.8% vs. 17.9 ± 2.3% aborts at 2 sps; $p = 0.01$; Fig. 11). At presentation rates faster than 2 sps, false alarm rates did not differ between distracters (two-way analysis of variance; $F(5,3) = 2.11$; $p = 0.07$), which suggests that compression does not drastically alter perception of distracter sounds. Overall, classifier performance was not significantly different from rat behavioral performance (unpaired *t*-tests at each presentation rate; 2 sps; $p = 0.65$, 4 sps; $p = 0.78$, 5 sps; $p = 0.35$, 6.67 sps; $p = 0.58$, 10 sps; $p = 0.16$, and 20 sps; $p = 0.14$). These results show that rats are able to accurately identify speech sounds imbedded in a rapid stream and our classifier was able to predict this performance function.

Spatial smoothing compensates for un-correlated neural activity

Our data were recorded in groups of four channels at a time. It was possible that the de-correlation caused by grouping channels that were not recorded simultaneously would negatively impact the classifier's ability to predict behavioral performance. We calculated the correlation between action potential patterns across different pairs of electrodes for each recording sweep. The average correlation coefficient between pairs of simultaneously recorded channels ($r = 0.15$) was significantly different from the correlation between pairs of channels in different penetrations ($r = 0.05$; $p < 0.01$). To test whether this lower correlation across channels affected the classifier, we added action potentials to the dataset of 196 sites to mimic the correlation observed when sites were recorded simultaneously (see Experimental procedures).

After this process, the average correlation coefficient across pairs of sites was no longer significantly different from the correlation between pairs of simultaneously recorded sites ($r = 0.13$ after adjustment; *t*-test vs. pairs of simultaneously recorded sites, $p = 0.53$). The re-correlated neural data did not require as much spatial smoothing as the un-altered population to achieve the same accuracy (unpaired *t*-test comparing classifier performance using re-correlated data with the un-altered population; $p = 0.11$). Using a Gaussian filter with a half-width of 2% of the total number of sites, the re-

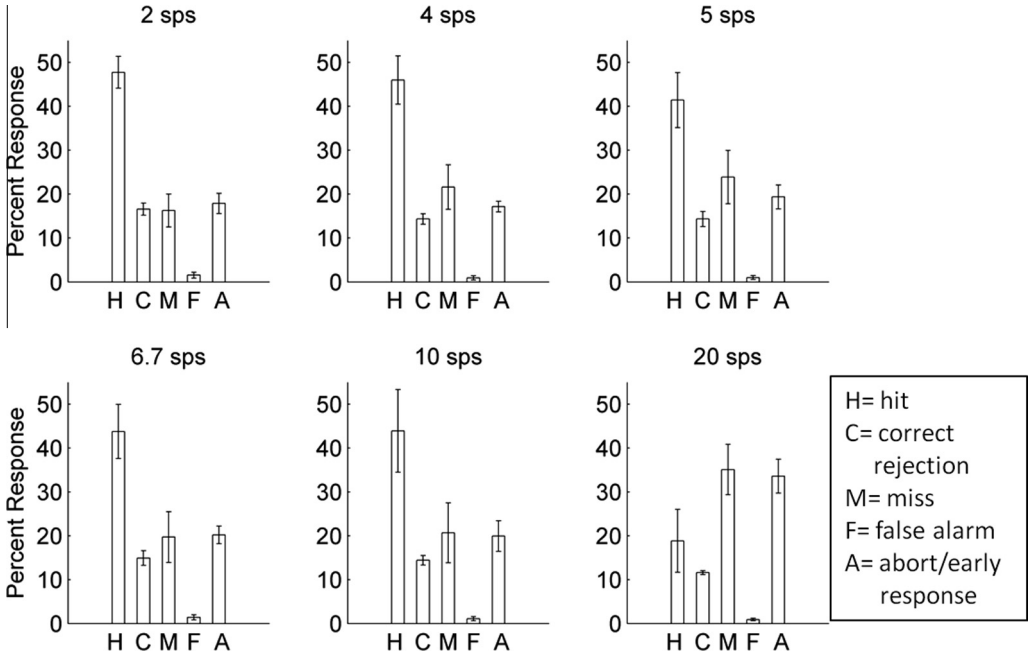

Fig. 11. Behavioral performance was robust at speeds slower than 20 sps. Performance breakdown at each of the six speeds we tested (H = hits, C = correct rejections, M = misses, F = false alarms, A = aborts/early responses). At speeds up to and including 10 sps, the majority of responses were to the target sound, with low rates of misses, false alarms, and aborts (or responses before the target was presented). At the fastest speed (20 sps), hit rate significantly decreased ($p < 0.01$) and both miss and abort rates significantly increased ($p < 0.01$ and $p = 0.01$ respectively. This pattern of response suggests that rats are still able to distinguish speech sounds until the presentation rate exceeds 10 sps. This is the same speed threshold commonly observed in human participants.

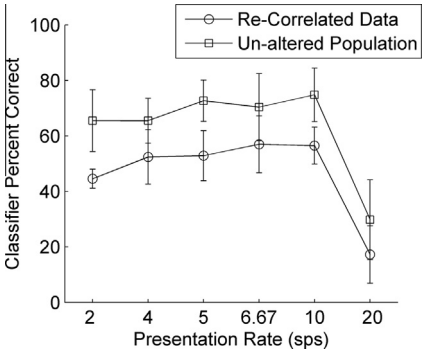

Fig. 12. A simulation of correlated neural data is able to predict rat behavior with less smoothing than the un-altered population. To evaluate the effect of different recording sessions on the performance of the classifier, we altered our dataset to mimic the correlated firing across recording sites acquired simultaneously (see Experimental procedures). The re-correlated data were able to predict rat behavior on the sequence task with less spatial smoothing than was required in the un-altered population (Gaussian filter with a half-width of 2% of the total number of sites; $R^2 = 0.67$, $p = 0.04$). The performance of the re-correlated data is not significantly different than the performance of the un-altered population (unpaired t-test, $p = 0.11$).

correlated data were highly accurate at locating and identifying the target sound /dad/ in a sequence and was significantly correlated with behavioral ability of rats on this task ($R^2 = 0.67$, $p = 0.04$; Fig. 12). The re-correlated data were not significantly different from the un-altered population in five of the six presentation rates (unpaired t-tests between un-altered data and re-correlated data; 2 sps; $p = 0.04$, 4 sps; $p = 0.23$, 5 sps; $p = 0.24$, 6.67 sps; $p = 0.49$, 10 sps; $p = 0.35$, and 20 sps; $p = 0.38$). This result suggests that the technique of smoothing on the spatial dimension may serve as an accurate method of compensation for the de-correlation that occurs when recording sites are not acquired simultaneously.

The classifier is as accurate using awake neural data

It was possible that our classifier would perform differently using awake recordings, due to differences in spontaneous activity or attention effects (Steinmetz et al., 2000; Treue, 2001). A different group of rats were implanted with a chronic array of 16 micro-electrodes. After recovery, we presented four speech sound sequences during a single passive recording session and were able to obtain a total of 123 reliable recording sites. Awake recordings had a higher spontaneous firing rate than anesthetized recordings (64.2 ± 1.8 Hz compared to 23.4 ± 1 Hz in the anesthetized preparation, unpaired t-test; $p < 0.001$; Fig. 13A) but this did not change the effectiveness of the classifier. After spatial smoothing (half-width of 15% of the total number of sites; Fig. 13B), the classifier performed at an average of 36.0 ± 6.6% using random groups of 100 sites (since we did not have enough sites to run in groups of 150). This accuracy mimics what the anesthetized classifier was able to accomplish with groups of 100 sites (Fig. 6A). The result that awake neural activity can perform the neural discrimination task with comparable accuracy to anesthetized recordings is

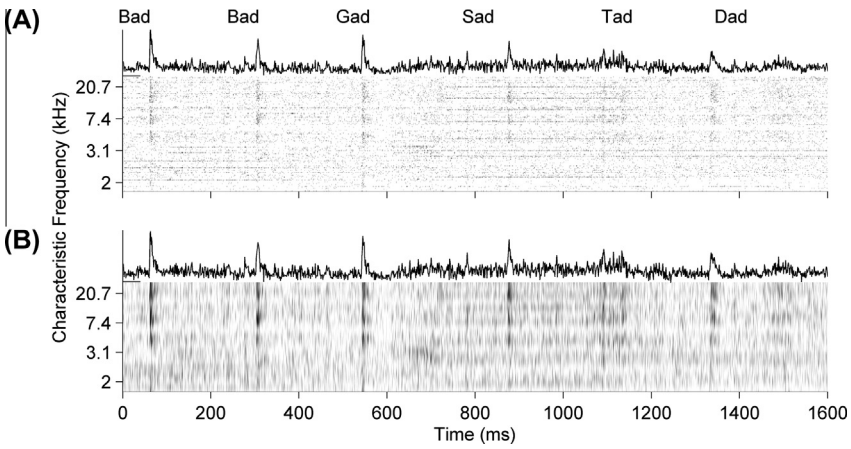

Fig. 13. The classifier can use awake neural data to locate and identify speech sounds in sequences. (A) Raw neural recordings from 123 sites in passively-listening awake rats. Awake data had significantly higher spontaneous firing rates compared to anesthetized (64.2 ± 1.8 Hz compared to 23.4 ± 1 Hz in the anesthetized preparation, unpaired *t*-test; $p < 0.001$). (B) After spatial smoothing with the same Gaussian filter used with anesthetized recordings, evoked activity was averaged and the classifier was able to locate and identify each speech sound in the sequence.

similar to what we saw using our earlier classifier (Engineer et al., 2008). This result suggests that our classifier may be able to predict performance in real time using neural recordings acquired from awake and behaving animals.

DISCUSSION

Calculation of decision thresholds

In our study, we designed a classifier that sweeps neural activity for a pattern of activity evoked by a speech sound and decides which sound caused that activity using predetermined decision thresholds. Our results support the idea that A1 contains information sufficient to perform speech sound identification (Steinschneider et al., 1995; Engineer et al., 2008; Bizley et al., 2010; Shetake et al., 2011; Perez et al., 2012; Ranasinghe et al., 2012b). This information may also be present in other cortical areas, as previous studies showed that removing A1 does not impair the ability of animals to perform speech sound discrimination tasks or to learn new speech sound targets (Floody et al., 2010; Porter et al., 2011). While the information needed to accomplish this task exists in A1, we recently showed that it may also be encoded in other auditory fields by parallel pathways from the thalamus (Centanni et al., 2013b).

In the current study, we did not find any difference in the ability of the classifier to locate the target stimulus in trained animals compared to recordings from naïve animals. This result suggests that training did not enhance the representation of the target sound in A1, though this target-specific effect may be present in other brain regions. For example, when monkeys were asked to identify whether two tactile stimuli were the same or different, primary somatosensory cortex encoded only the current stimulus, while secondary somatosensory cortex was already beginning to compare the two stimuli (Romo and Salinas, 2003). It is likely that higher level brain regions contain integrator neurons that recognize patterns of activity occurring in lower level areas. Neural networks designed to mimic sensory neurons can be trained to integrate basic sensory information into categorical decisions (Buonomano and Merzenich, 1995; Mazurek et al., 2003). Single neurons recorded in premotor cortex of monkeys can also predict the intended motor sequence when a maximum-likelihood decoder analyzes the firing rate (Shanechi et al., 2012). Our classifier does not propose a mechanism for how this threshold is created or where in the brain it is stored, but it is the first to show that a classifier can use A1 activity to predict the location and identity of speech stimuli without being forced to choose between a set of options. As in behavioral tasks, if the decision threshold is not met, the classifier is not required to guess. In addition, if multiple thresholds are met, our classifier is designed to choose the template which is most like the single trial.

Our study does not address the effect of behavioral feedback on the creation or maintenance of this threshold, as our thresholds did not change during testing. It is likely that the brain adapts to real time feedback during testing. If thresholds never changed, the brain would be inept at tasks of generalization. For example, the same word spoken with small changes in pitch, pronunciation and/or context may cause the brain to categorize these as two different words. It is well known that synapses change as a result of real time feedback (Malenka and Nicoll, 1993; Buonomano and Merzenich, 1998; Cohen-Cory, 2002; Malinow and Malenka, 2002), but the question of how the brain monitors these changes and how drastic the adjustments are remains to be answered. A classifier that could adjust its thresholds in relation to real time feedback would provide a more

biologically accurate model and may be able to explain models of learning impairments.

Evaluation of the data set and classifier

The data reported in our study were acquired from many animals and analyzed post hoc. In the anesthetized recordings, four electrodes were recorded simultaneously. In the awake preparation, up to seven electrodes were viable at any given time point. Our result that a simulation of correlated data is able to predict behavioral ability suggests that this classifier would likely perform well if provided over 150 simultaneously recorded sites. We also observed that re-correlated data do not need as much spatial smoothing as de-correlated data. A small amount of integration is likely present in the brain from one neural population to another, so the amount of smoothing still required after re-correlation is biologically plausible (Giraud et al., 2000; Langers et al., 2003, 2007). We suggest that greater amounts of spatial smoothing may therefore compensate for un-correlated data. This hypothesis will require further study using large numbers of simultaneously recorded sites.

The classifier used a fixed window (80 ms) to scan a single trial of neural activity for evoked responses. There is sufficient information present in this window for consonant identification to take place (Miller and Nicely, 1955; Kuhl and Miller, 1975; Engineer et al., 2008). However, it is likely that rats and humans also use information occurring in larger integration windows, especially in difficult hearing environments (Shetake et al., 2011). Our classifier attempted to account for this by analyzing the NM values within 4 ms of the initial guess. This allowed the classifier some flexibility to wait until all similar templates were considered and then make a decision using the strongest signal. This time period of flexibility is biologically plausible as it is well within the minimum amount of time in which the brain can make a decision (Stanford et al., 2010).

We also show in the current study that our classifier fell to chance performance when 10-ms temporal bins were used. This finding is in contrast to recent work showing that this bin size is optimal for single cell discrimination (Wang et al., 1995; Schnupp et al., 2006). This difference may be due to the influence of neighboring neurons in the current study. Our study used multi-unit recordings as the data set for testing the classifier, and the influence of nearby neurons with slightly varying response patterns is likely the cause of the discrepancy between our test of 10-ms bins and other recent work in single units. An additional test of this classifier using many single-unit responses will be critical in determining the effect of multi-unit sites on the efficacy of 10–50-ms temporal bin sizes. In addition to the differences in single unit vs. multi-unit recordings, future work should also investigate the differences in neural activity recorded from difference cortical layers. In the current study, we recorded from layers 4/5 of rat auditory cortex. These are input layers and are a common choice for auditory recording studies (Winer et al., 2005; Christianson et al., 2011; Centanni et al., 2013b). A recent study showed that superficial layers in the rodent often respond with fewer spikes per stimulus and show evidence of larger post-activation suppression (Christianson et al., 2011). Therefore, it is possible that responses in different cortical layers may encode information in a different way than the activity patterns shown here and our classifier should be tested using datasets from different cortical layers.

Our auditory cortex recordings were acquired exclusively in the right hemisphere of rats. There has been considerable discussion in the recent literature about the possible lateralization of rodent auditory cortex and the nature of possible specializations that result from such organization. For example, recent work has shown that there are functional and specific differences in frequency-modulated (FM) discrimination directly related to which hemisphere is lesioned in rodents (Wetzel et al., 2008). Left hemisphere auditory areas have also been shown to be important in pattern discrimination in the cat (Lomber and Malhotra, 2008). It is possible that A1 responses in the left auditory cortex may yield additional insight into whether any functional specialization occurs in a particular hemisphere of the rat auditory cortex. In addition, neural activity patterns from other auditory areas should be tested, as there are differences in the neural encoding of temporal stimuli across cortical regions (Lomber and Malhotra, 2008; Centanni et al., 2013b).

Future applications for the classifier

In the current study, we demonstrate that a classifier can locate and identify speech sound stimuli in real time using single repeats of A1 neural activity. The ability to extract this type of meaningful information from basic auditory processing areas confirms that the relevant speech-coding activity is present at a low sensory level. The functional mechanisms behind many speech-processing disorders are still poorly understood and our classifier may prove to be valuable tool in answering these questions. If, for example, neural activity from A1 is able to accurately encode speech sounds, then a processing deficit is likely to exist in a higher cortical area.

In addition, our classifier may prove to be a useful tool in the early identification of speech-processing disorders. These individuals are often impaired at speech processing in difficult listening conditions. Individuals with dyslexia, for example, often have difficulty processing rapid speech sounds and speech presented in background noise (Helenius et al., 1999; Ziegler et al., 2009; Poelmans et al., 2012). Recent research is elucidating the neural basis for these types of perception impairments (Lehongre et al., 2011; Kovelman et al., 2012; Centanni et al., 2013a; Hornickel and Kraus, 2013), but early detection of these disorders has yet to be optimized. The biologically plausible smoothing parameters used in our study have the potential to improve our current understanding of the mechanisms by which sounds are encoded in A1 and are subsequently passed to higher cortical areas. An understanding of this knowledge may help us better identify when this system is performing inadequately

and may help in the development of early identification and treatment of speech-processing disorders.

CONCLUSION

In the current study, we developed a classifier that can locate the onset and identify consonant speech sounds using population neural data acquired from A1. Our classifier successfully predicted the ability of rats to identify a target CVC speech sound in a continuous stream of distracter CVC sounds at speeds up to 10 sps, which is comparable to human performance. The classifier was just as accurate when using data recorded from awake rats. We also demonstrate that smoothing neural data along the spatial dimension may compensate for the de-correlation that occurs when acquiring neural data in several separate groups. These results demonstrate that the neural activity in A1 can be used to quickly and accurately identify consonant speech sounds with accuracy that mimics performance.

Acknowledgments—We would like to thank N. Lengnick, H. Shepard, N. Moreno, R. Cheung, K. Im, S. Mahioddin, C. Rohloff and A. Malik for their help with behavioral training, as well as A. Reed, D. Gunter, C. Mains, M. Borland, E. Hancik and Z. Abdulali for help in acquiring neural recordings. We would also like to thank K. Ranasinghe for suggestions on earlier versions of this manuscript. This work was supported by the National Institute for Deafness and Communication Disorders at the National Institutes of Health (R01DC010433). The authors declare no competing interests.

REFERENCES

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci 98:13367.

Anderson S, Kilgard M, Sloan A, Rennaker R (2006) Response to broadband repetitive stimuli in auditory cortex of the unanesthetized rat. Hear Res 213:107–117.

Bear Mark F, Cooper Leon N, Ebner Ford F (1987) A physiological basis for a theory of synapse modification. Science 237:42–48.

Bizley JK, Walker KM, King AJ, Schnupp JW (2010) Neural ensemble codes for stimulus periodicity in auditory cortex. J Neurosci 30:5078–5091.

Brillouin L (2013) Science and information theory. Dover Publications.

Buonomano DV, Merzenich MM (1998) Cortical plasticity: from synapses to maps. Annu Rev Neurosci 21:149–186.

Buonomano DV, Merzenich MM (1995) Temporal information transformed into a spatial code by a neural network with realistic properties. Science:1028–1030 (New York then Washington).

Centanni T, Booker A, Sloan A, Chen F, Maher B, Carraway R, Khodaparast N, Rennaker R, Loturco J, Kilgard M (2013a) Knockdown of the dyslexia-associated gene KIAA0319 impairs temporal responses to speech stimuli in rat primary auditory cortex. Cereb Cortex. http://dx.doi.org/10.1093/Cercor/Bht028.

Centanni TM, Engineer CT, Kilgard MP (2013b) Cortical speech-evoked response patterns in multiple auditory fields are correlated with behavioral discrimination ability. J Neurophysiol 110:177–189.

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432.

Christianson GB, Sahani M, Linden JF (2011) Depth-dependent temporal response properties in core auditory cortex. J Neurosci 31:12837–12848.

Cohen-Cory S (2002) The developing synapse: construction and modulation of synaptic structures and circuits. Science 298:770–776.

Creutzfeldt O, Hellweg F, Schreiner C (1980) Thalamocortical transformation of responses to complex auditory stimuli. Exp Brain Res 39:87–104.

Dong C, Qin L, Liu Y, Zhang X, Sato Y (2011) Neural responses in the primary auditory cortex of freely behaving cats while discriminating fast and slow click-trains. PLoS One 6:E25895.

Eggermont JJ (1995) Representation of a voice onset time continuum in primary auditory cortex of the cat. J Acoust Soc Am 98:911.

Engineer CT, Perez CA, Chen YTH, Carraway RS, Reed AC, Shetake JA, Jakkamsetti V, Chang KQ, Kilgard MP (2008) Cortical activity patterns predict speech discrimination ability. Nat Neurosci 11:603–608.

Floody OR, Ouda L, Porter BA, Kilgard MP (2010) Effects of damage to auditory cortex on the discrimination of speech sounds by rats. Physiol Behav 101:260–268.

Foffani G, Moxon KA (2004) Psth-based classification of sensory stimuli using ensembles of single neurons. J Neurosci Methods 135:107–120.

Ghitza O, Greenberg S (2009) On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. Phonetica 66:113–126.

Giraud AL, Lorenzi C, Ashburner J, Wable J, Johnsrude I, Frackowiak R, Kleinschmidt A (2000) Representation of the temporal envelope of sounds in the human brain. J Neurophysiol 84:1588–1598.

Green DM, Swets JA (1966) Signal detection theory and psychophysics. New York: Wiley.

Hao J, Wang X, Dan Y, Poo M, Zhang X (2009) An arithmetic rule for spatial summation of excitatory and inhibitory inputs in pyramidal neurons. Proc Natl Acad Sci 106:21906–21911.

Helenius P, Uutela K, Hari R (1999) Auditory stream segregation in dyslexic adults. Brain 122:907–913.

Hornickel J, Kraus N (2013) Unstable representation of sound: a biological marker of dyslexia. J Neurosci 33:3500–3504.

Huetz C, Philibert B, Edeline JM (2009) A spike-timing code for discriminating conspecific vocalizations in the thalamocortical system of anesthetized and awake guinea pigs. J Neurosci 29:334–350.

Kawahara H (1997) Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. Acoustics, Speech, and Signal Processing 2:1303–1306.

Kovelman I, Norton ES, Christodoulou JA, Gaab N, Lieberman DA, Triantafyllou C, Wolf M, Whitfield-Gabrieli S, Gabrieli JDE (2012) Brain basis of phonological awareness for spoken language in children and its disruption in dyslexia. Cereb Cortex 22:754–764.

Kuhl PK, Miller JD (1975) Speech perception by the chinchilla: voiced–voiceless distinction in alveolar plosive consonants. Science 190:69–72.

Langers DRM, Backes WH, Dijk P (2003) Spectrotemporal features of the auditory cortex: the activation in response to dynamic ripples. Neuroimage 20:265–275.

Langers DRM, Backes WH, Van Dijk P (2007) Representation of lateralization and tonotopy in primary versus secondary human auditory cortex. Neuroimage 34:264–273.

Lehongre K, Ramus F, Villiermet N, Schwartz D, Giraud AL (2011) Altered low-gamma sampling in auditory cortex accounts for the three main facets of dyslexia. Neuron 72:1080–1090.

Lomber SG, Malhotra S (2008) Double dissociation of 'what' and 'where' processing in auditory cortex. Nat Neurosci 11: 609–616.

Malenka RC, Nicoll RA (1993) NMDA-receptor-dependent synaptic plasticity: multiple forms and mechanisms. Trends Neurosci 16.

Malinow R, Malenka RC (2002) AMPA receptor trafficking and synaptic plasticity. Annu Rev Neurosci 25:103–126.

Martin BA, Stapells DR (2005) Effects of low-pass noise masking on auditory event-related potentials to speech. Ear Hear 26:195.
Mazurek ME, Roitman JD, Ditterich J, Shadlen MN (2003) A role for neural integrators in perceptual decision making. Cereb Cortex 13:1257–1269.
Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123:899.
Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some english consonants. J Acoust Soc Am 27:338.
Panzeri S, Diamond ME (2010) Information carried by population spike times in the whisker sensory cortex can be decoded without knowledge of stimulus time. Front Synaptic Neurosci 2:17.
Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. PLoS Biol 10:E1001251.
Perez CA, Engineer CT, Jakkamsetti V, Carraway RS, Perry MS, Kilgard MP (2012) Different timescales for the neural coding of consonant and vowel sounds. Cereb Cortex 23:670–683.
Poelmans H, Luts H, Vandermosten M, Boets B, Ghesquière P, Wouters J (2012) Auditory steady state cortical responses indicate deviant phonemic-rate processing in adults with dyslexia. Ear Hear 33:134.
Poirazi P, Brannon T, Mel BW (2003a) Arithmetic of subthreshold synaptic summation in a model ca1 pyramidal cell. Neuron 37:977–987.
Poirazi P, Brannon T, Mel BW (2003b) Pyramidal neuron as two-layer neural network. Neuron 37:989–999.
Poldrack RA, Temple E, Protopapas A, Nagarajan S, Tallal P, Merzenich M, Gabrieli JDE (2001) Relations between the neural bases of dynamic auditory processing and phonological processing: evidence from FMRI. J Cogn Neurosci 13:687–697.
Porter BA, Rosenthal TR, Ranasinghe KG, Kilgard MP (2011) Discrimination of brief speech sounds is impaired in rats with auditory cortex lesions. Behav Brain Res 219:68–74.
Ranasinghe KG, Carraway RS, Borland MS, Moreno NA, Hanacik EA, Miller RS, Kilgard MP (2012a) Speech discrimination after early exposure to pulsed-noise or speech. Hear Res 289:1–12.
Ranasinghe KG, Vrana WA, Matney CJ, Kilgard MP (2012b) Neural mechanisms supporting robust discrimination of spectrally and temporally degraded speech. J Assoc Res Otolaryngol:1–16.
Rennaker R, Street S, Ruyle A, Sloan A (2005) A comparison of chronic multi-channel cortical implantation techniques: manual versus mechanical insertion. J Neurosci Methods 142:169–176.
Romo R, Salinas E (2003) Flutter discrimination: neural codes, perception, memory and decision making. Nat Rev Neurosci 4:203–218.
Schnupp JWH, Hall TM, Kokelaar RF, Ahmed B (2006) Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. J Neurosci 26:4785–4795.
Sengpiel F, Kind PC (2002) The role of activity in development of the visual system. Curr Biol 12:R818–R826.
Shanechi MM, Hu RC, Powers M, Wornell GW, Brown EN, Williams ZM (2012) Neural population partitioning and a concurrent brain-machine interface for sequential motor function. Nat Neurosci 15:1715–1722.
Shetake JA, Wolf JT, Cheung RJ, Engineer CT, Ram SK, Kilgard MP (2011) Cortical activity patterns predict robust speech discrimination ability in noise. Eur J Neurosci 34:1823–1838.
Sloan AM, Dodd OT, Rennaker II RL (2009) Frequency discrimination in rats measured with tone-step stimuli and discrete pure tones. Hear Res 251:60–69.
Stanford TR, Shankar S, Massoglia DP, Costello MG, Salinas E (2010) Perceptual decision making in less than 30 milliseconds. Nat Neurosci 13:379–385.
Steinmetz PN, Roy A, Fitzgerald P, Hsiao S, Johnson K, Niebur E (2000) Attention modulates synchronized neuronal firing in primate somatosensory cortex. Nature 404:131–133.
Steinschneider M, Reser D, Schroeder CE, Arezzo JC (1995) Tonotopic organization of responses reflecting stop consonant place of articulation in primary auditory cortex (A1) of the monkey. Brain Res 674:147–152.
Steinschneider M, Volkov IO, Fishman YI, Oya H, Arezzo JC, Howard III MA (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 15:170–186.
Treue S (2001) Neural correlates of attention in primate visual cortex. Trends Neurosci 24:295–300.
Wang X, Merzenich MM, Beitel R, Schreiner CE (1995) Representation of a species-specific vocalization in the primary auditory cortex of the common marmoset: temporal and spectral characteristics. J Neurophysiol 74:2685–2706.
Wetzel W, Ohl FW, Scheich H (2008) Global versus local processing of frequency-modulated tones in gerbils: an animal model of lateralized auditory cortex functions. Proc Natl Acad Sci 105:6753–6758.
Winer JA, Miller LM, Lee CC, Schreiner CE (2005) Auditory thalamocortical transformation: structure and function. Trends Neurosci 28:255–263.
Ziegler JC, Pech-Georgel C, George F, Lorenzi C (2009) Speech-perception-in-noise deficits in dyslexia. Dev Sci 12:732–745.

(Accepted 15 November 2013)
(Available online 26 November 2013)