Different Timescales for the Neural Coding of Consonant and Vowel Sounds

Claudia A. Perez, Crystal T. Engineer, Vikram Jakkamsetti, Ryan S. Carraway, Matthew S. Perry and Michael P. Kilgard

Cognition and Neuroscience Program, School of Behavioral and Brain Sciences, University of Texas at Dallas, 800 West Campbell Road, Richardson, TX 75080, USA

Address correspondence to Michael P. Kilgard, Cognition and Neuroscience Program, School of Behavioral and Brain Sciences, University of Texas at Dallas, GR41 800 West Campbell Road Richardson, TX 75080, USA. Email: kilgard@utdallas.edu.

Psychophysical, clinical, and imaging evidence suggests that consonant and vowel sounds have distinct neural representations. This study tests the hypothesis that consonant and vowel sounds are represented on different timescales within the same population of neurons by comparing behavioral discrimination with neural discrimination based on activity recorded in rat inferior colliculus and primary auditory cortex. Performance on 9 vowel discrimination tasks was highly correlated with neural discrimination based on spike count and was not correlated when spike timing was preserved. In contrast, performance on 11 consonant discrimination tasks was highly correlated with neural discrimination when spike timing was preserved and not when spike timing was eliminated. These results suggest that in the early stages of auditory processing, spike count encodes vowel sounds and spike timing encodes consonant sounds. These distinct coding strategies likely contribute to the robust nature of speech sound representations and may help explain some aspects of developmental and acquired speech processing disorders.

Keywords: multiplexed temporal coding, processing streams, rate code, spatiotemporal activity pattern

Introduction

A diverse set of observations suggest that consonants and vowels are processed differently by the central auditory system. Compared with vowels, consonant perception 1) matures later (Polka and Werker 1994), 2) is more categorical (Fry et al. 1962; Pisoni 1973), 3) is less sensitive to spectral degradation (Shannon et al. 1995; Xu et al. 2005), 4) is more sensitive to temporal degradation (Shannon et al. 1995; Kasturi et al. 2002; Xu et al. 2005), and 5) is more useful for parsing the speech stream into words (Bonatti et al. 2005; Toro et al. 2008). Clinical studies, imaging, and stimulation experiments also suggest that consonants and vowels are differentially processed. Brain damage can impair consonant perception while sparing vowel perception and vice versa (Caramazza et al. 2000). Electrical stimulation of the temporal cortex that impaired consonant discrimination did not impair vowel discrimination (Boatman et al. 1994, 1995, 1997). Human brain imaging studies also suggest that consonant and vowel sounds are processed differently (Fiez et al. 1995; Seifritz et al. 2002; Poeppel 2003; Carreiras and Price 2008). Unfortunately, none of these studies provides sufficient resolution to document how the neural representations of consonant and vowel sounds differ. In this study, we recorded action potentials from inferior colliculus (IC) and primary auditory cortex (A1) neurons of rats to test the hypothesis that consonant and vowel sounds are represented by neural activity patterns occurring on different timescales (Poeppel 2003).

Previous neurophysiology studies in animals revealed that both the number of spikes generated by speech sounds and the relative timing of these spikes contain significant and complementary information about important speech features. From auditory nerve to A1, the spectral shape of vowel and fricative sounds can be identified in plots of evoked spike count as a function of characteristic frequency (Sachs and Young 1979; Delgutte and Kiang 1984a, 1984b; Ohl and Scheich 1997; Versnel and Shamma 1998). Information based on spike timing can be used to identify the formants (peaks of spectral energy) of steady state vowels (Young and Sachs 1979; Palmer 1990). Relative spike timing can also be used to identify the onset spectrum, formant transitions, and voice onset time of many consonants (Steinschneider et al. 1982, 1995, 1999, 2005; Miller and Sachs 1983; Carney and Geisler 1986; Deng and Geisler 1987; Engineer et al. 2008). Behavioral discrimination of consonant sounds is highly correlated with neural discrimination using spike timing information collected in A1 (Engineer et al. 2008). However, no study has directly compared potential coding strategies for consonants and vowels with behavior.

In this study, we trained 19 rats to discriminate consonant or vowel sounds and compared behavioral and neural discrimination using activity recorded from IC and A1. It is reasonable to expect that sounds which evoke similar neural activity patterns will be more difficult to discriminate than sounds which evoke more distinct patterns. We tested whether the neural code for speech sounds was best described as the average number of action potentials generated by a population of neurons or as the average spatiotemporal pattern across a population or as both depending on the particular stimulus discrimination. Our results suggest that the brain represents consonant and vowel sounds on distinctly different timescales. By comparing responses in IC and A1, we confirmed that speech sound processing, like sensory processing generally, occurs as the gradual transformation from acoustic information to behavioral category (Chechik et al. 2006; Hernández et al. 2010; Tsunada et al. 2011).

Materials and Methods

Speech Stimuli

Twenty-eight English consonant-vowel-consonant words were recorded in a double-walled soundproof booth. Twenty of the sounds ended in "ad" (/æd/ as in "sad") and were identical to the sounds used in our earlier study (Engineer et al. 2008). The initial consonants of these sounds differ in voicing, place of articulation, or manner of articulation. The remaining 8 words began with either "s" or "d" and contained the vowels (/æ/, /i/, and /u/ as in "said," "sad," "seed," and "sood," Fig. 1). To confirm that vowel discrimination was not based on coarticulation during the preceding "s" or "d" sounds (Soli 1981), we also tested vowel discrimination on a set of 5 sounds in which the "s" was replaced with a 10 ms burst of white noise (60 dB sound pressure level [SPL], 1–32
is \( X \) Spectrograms of the 5 vowel sounds with 2 initial consonants. The initial recordings were shifted one octave higher to match the rat hearing range using the STRAIGHT vocoder (Kawahara 1997).

### Operant Training Procedure and Analysis

Nineteen rats were trained using an operant go/no-go procedure to discriminate words differing in their initial consonant sound or in their vowel sound. Each rat was trained for 2 sessions a day (1 h each), 5 days per week. Rats first underwent a shaping period during which they were taught to press the lever. Each time the rat was in close proximity to the lever, the rat heard the target stimulus ("dad" or "sad") and received a food pellet. Eventually, the rat began to press the lever without assistance. After each lever press, the rat heard the target sound and received a pellet. The shaping period lasted until the rat was able to reach the criteria of obtaining at least 100 pellets per session for 2 consecutive sessions. This stage lasted on average 3.5 days. Following the shaping period, rats began a detection task where they learned to press the lever each time the target sound was presented. Silent periods were randomly interleaved with the target sounds during each training session. Sounds were initially presented every 10 s, and the rat was given an 8 s window to press the lever. The sound interval was gradually decreased to 6 s, and the lever press window was decreased to 3 s. Once rats reached the performance criteria of a \( d' \geq 1.5 \) for 10 sessions, they advanced to the discrimination task. The quantity \( d' \) is a measure of discriminability of 2 sets of samples based on signal detection theory (Green and Swets 1966).

During discrimination training, rats learned to discriminate the target ("dad" or "sad") from the distractor sounds, which differed in initial consonant or vowel. Training took place in a soundproof double-walled training booth that included a house light, video camera for monitoring, speaker, and a cage that included a lever, lever light, and pellet receptacle. Trials began every 6 s, and silent catch trials were randomly interleaved 20–33% of the time. Rats were only rewarded for lever presses to the target stimulus. Pressing the lever at any other time resulted in a timeout during which the house light was extinguished and the training program paused for a period of 6 s. Rats were food deprived to motivate behavior but were fed on days off to maintain between 80% and 90% ad lib body weight.

Each discrimination task lasted for 20 training sessions over 2 weeks. Eight rats were trained to discriminate between vowel sounds. During the first task, 4 rats were required to press to "sad" and not "said," "sad," "seed," "sood." The other 4 rats were required to press to "dad" and not "dead," "dud," "deed," "dood." After 2 weeks of training, the tasks were switched and the rats were trained to discriminate the vowels with the initial consonant switched ("d" or "s"). After 2 additional weeks of training, both groups of rats were required to press to either "sad" or "dad" and reject any of the 8 other sounds (Fig. 1). For 2 sessions during this final training stage, the stimuli were replaced with noise burst-vowel-consonant syllables (described above).

Eleven rats were trained to discriminate between consonant sounds. Six rats performed each of 4 different consonant discrimination tasks for 2 weeks each ("dad" vs. "sad," "dad" vs. "rad," "dad" vs. "lad," and "dad" vs. "bad" and "gad"), and 5 rats performed each of 4 different discrimination tasks for 2 weeks each ("dad" vs. "dad" with a lower pitch, "mad" vs. "ndad," "chad" vs. "lad," and "had" vs. "lad," "sad," and "had"). These 11 consonant trained rats were the same rats as in our earlier study (Engineer et al. 2008).

### Recording Procedure

Multiunit and single unit responses from the right inferior colliculus (IC, \( n = 187 \) recording sites) or primary auditory cortex (A1, \( n = 445 \) recording sites) of 23 experimentally naïve female Sprague-Dawley rats were obtained in a soundproof recording booth. During acute recordings, rats were anesthetized with pentobarbital (50 mg/kg) and received supplemental dilute pentobarbital (8 mg/mL) every half hour to 1 h as needed to maintain areflexia. Heart rate and body temperature were monitored throughout the experiment. For most cortical recordings, 4 Parylene-coated tungsten microelectrodes (1-2 MΩ, FHC Inc., Bowdoin, ME) were simultaneously lowered to 600 μm below the surface of the right primary auditory cortex (layer 4/5). Electrode penetrations were marked using blood vessels as landmarks.
Consonant responses from these +45 A1 recordings sites were also used in our earlier study (Engineer et al. 2008).

For collicular recordings, 1 or 2 Parylene-coated tungsten microelectrodes (1–2 MΩ, FHC Inc.) were introduced through a hole located in the dorsal ventral axis to a depth of 1100 and 5690 μm. Blood vessels were used as landmarks as the electrode placement was changed in the caudal-rostral direction. Based on latency analysis and frequency topography, the majority of IC recordings are likely to be from the central nucleus (Palombi and Caspary 1996).

At each recording site, 25 ms tones were played at 81 frequencies (1-32 kHz) at 16 intensities (0-75 dB) to determine the characteristic frequency of each site. All tones were separated by 560 ms and randomly interleaved. Sounds were presented approximately 10 cm from the left ear of the rat. Twenty-nine 60 dB speech stimuli were randomly interleaved and presented every 2000 ms. Each sound was repeated 20 times. Stimulus generation, data acquisition, and spike sorting were performed with Tucker-Davis hardware (RP2.1 and RX5) and software (Brainware). Protocols and recording procedures were approved by the University of Texas at Dallas IACUC.

**Data Analysis**

The distinctness of the spatial activity patterns generated by each pair of sounds (i.e., “dad” and “dud”) was quantified as the average difference in the number of spikes evoked by each sound across the population of neurons recorded. As in our earlier study, the analysis window for consonant sounds was 40 ms long beginning at sound onset (Engineer et al. 2008). The analysis window for vowel sounds was 300 ms long and began immediately at vowel onset.

A nearest-neighbor classifier was used to quantify neural discrimination accuracy based on single trial activity patterns (Foffani and Moxon 2004; Schnupp et al. 2006; Engineer et al. 2008; Malone et al. 2010). The classifier binned activity using 1–400 ms intervals and then compared the response of each single trial with the average activity pattern (poststimulus time histogram, PSTH) evoked by each of the speech stimuli presented. PSTH templates were constructed from 19 presentations since the trial currently considered was not included in the average activity pattern to prevent artifact. This model assumes that the brain region reading out the information in the spike trials has previously heard each of the sounds 19 times and attempts to identify which of the possible choices was most likely to have generated the trial under consideration. Euclidean distance was used to determine how similar each response was to the average activity evoked by each of the sounds. The classifier guesses that the single trial pattern was generated by the sound whose average pattern it most closely resembles (i.e., minimum Euclidean distance). The response onset to each sound was defined as time point at which neural activity exceeded the spontaneous firing rate by 3 standard deviations. Error bars indicate standard error of the mean. Pearson’s correlation coefficient was used to examine the relationship between neural and behavioral discrimination on the 11 consonant tasks and 9 vowel tasks.

**Results**

Rats were trained on a go/no-go task that required them to discriminate between English monosyllabic words with different vowels. The rats were rewarded for pressing a lever in response to the target word (“dad” or “sad”) and punished with a brief timeout for pressing the lever in response to any of the 4 distracter words (“dead,” “dud,” “deed,” and “doed”) or “said,” “said,” “seed,” and “sood” (Fig. 1). Performance was well above chance after only 3 days of discrimination training ($P < 0.005$, Fig. 2a). By the end of training, rats were correct on 85% of trials (Fig. 2c). The pattern of false alarms suggests that rats discriminate between vowels by identifying differences in the vowel power spectrums (Fig. 3a,b). Vowels with similar spectral peaks (i.e., “dad” vs. “dead”; Peterson and Barney 1952) were difficult for rats to discriminate, and vowels with distinct spectral peaks (i.e., dad vs. deed) were easy to discriminate (Fig. 3c,d). For example, rats were able to correctly discriminate “dad” from “dead” on 58 ± 2% of trials but were able to correctly discriminate “dad” from “doed” on 78 ± 2% of trials ($P < 0.02$). Discrimination of vowel pairs was highly correlated with the difference between the sounds in the feature space created by the peak of the first and second formants ($R^2 = 0.77$, $P < 0.002$, Fig. 3a,b) and with the Euclidean distance between the complete power spectrum of each vowel used ($R^2 = 0.79$, $P < 0.001$). These results suggest that rats, like humans, discriminate between vowel sounds by comparing formant peaks (Delattre et al. 1952; Peterson and Barney 1952; Fry et al. 1962).

We also tested whether rats could quickly generalize the task to a new set of stimuli with a different initial consonant to confirm that the rats were discriminating the speech sounds based on the vowel and not some other acoustic cue. Rats that had been trained for 10 days on words with the same initial consonant (either “d” or “s”) correctly generalized their performance when presented with a completely new set of stimuli with a different initial consonant (“s” or “d,” Fig. 2b). The average performance on the first day was well above...
chance (60 ± 2% correct performance, \(P < 0.005\)). Performance was above chance even during the first 4 trials of testing with each of the new stimuli (60 ± 4%, \(P < 0.05\)), which confirms that generalization of vowel discrimination occurred rapidly. The fact that rats can generalize their performance to target sounds with somewhat different formant structure (Fig. 3a) suggests the possibility that preceding consonants may alter the perception of vowel sounds, as in human listeners (Holt et al. 2000). However, additional studies with synthetic vowel stimuli would be needed to test this possibility.

During the final stage of training, rats were able to respond to the target vowel when presented in either of 2 contexts ("d" or "s") and withhold responding to the 8 other distracters (85 ± 4% correct trials on the last day, Fig. 2c). To confirm that rats were using spectral differences during the vowel portion of the speech sounds instead of cues present only in the consonant, we tested vowel discrimination for 2 sessions (marked by open circles in Fig. 2c) using a set of sounds derived from the stimuli beginning with "s," but where the "s" sound was replaced with a 30 ms white noise burst that was identical for each sound. Performance was well above chance (64 ± 3%, \(P = 0.016\)) and well correlated with formant differences even when the initial consonant was replaced by a noise burst (\(R^2 = 0.65, P < 0.008\)). These results demonstrate that rats are able to accurately discriminate English vowel sounds and suggest that rats are a suitable model for studying the neural mechanisms of speech sound discrimination (Engineer et al. 2008).

We recorded multiunit neural activity from 187 IC sites and 445 A1 sites to evaluate which of several possible neural codes best explain the behavioral discrimination of consonant and vowel sounds. A total of 29 speech sounds were presented to each site, including 20 consonants and 5 vowels. As observed in earlier studies using awake and anesthetized preparations (Watanabe 1978; Steinschneider et al. 1982; Chen et al. 1996; Sinex and Chen 2000; Cunningham et al. 2002; Engineer et al. 2008), speech sounds evoked a brief phasic response as well as a sustained response (Fig. 4 for IC responses and Fig. 5 for A1 responses). As expected, IC neurons responded to speech sounds earlier than A1 neurons (3.1 ± 1.2 ms earlier, \(P < 0.05\)). The sustained spike rate was more pronounced in IC but was also observed in A1 neurons (IC: 62 ± 4 Hz, Fig. 4; A1: 26 ± 0.9 Hz, Fig. 5). The pattern of neural firing was distinct for each of the vowels tested. For example, even though "dad" and "dood" evoked approximately the same number of spikes in IC (18 ± 1.3 and 19 ± 1.2 Hz) and A1 (8.5 ± 0.3 and 8.3 ± 0.2 Hz), the spatial pattern differed such that the average absolute value of the difference in the response to "dad" and "dood" across all IC and A1 sites was 28.4 ± 2.7 and 6.7 ± 0.4 Hz, respectively. The absolute value of the difference in the firing rate between

---

**Figure 3.** Lever press rates for each of the 5 vowel sounds tested. (a) Peak frequency for the first and second formant of each sound tested is plotted in Hertz. Note that original speech sounds were shifted one octave higher to match the rat hearing range using the STRAIGHT vocoder (Kawahara 1997). (b) The relationship between the Euclidean distance in F1-F2 space in octaves is well correlated with behavioral discrimination. (c) and (d) Rats pressed the lever in response to the target vowel/æ/ (as in "sad" and "dad") significantly more than for any of the other vowel sounds. The solid line indicates the frequency that rats pressed the lever during silent catch trials. The dashed lines and the error bars represent standard error of the mean. Behavioral responses were from all 10 days of training.
each pair of vowel sounds was well correlated with the distance between the vowels in the feature space formed by the first and second formant peaks (IC: $R^2 = 0.88$, $P < 0.01$; A1: $R^2 = 0.69$, $P < 0.01$, Fig. 6). The finding that different vowel sounds generate distinct spatial activity patterns in the central auditory system is consistent with earlier studies (Sachs and Young 1979; Delgutte and Kiang 1984a; Ohl and Scheich 1997; Versnel and Shamma 1998) and suggests that differences in the average neural firing rate across a population of neurons could be used to discriminate between vowel sounds.

Although individual formants were not clearly visible in the spatial activity patterns of A1 or IC, the global characteristics of the spatial activity patterns (rate–place code) evoked by each vowel sound were clearly related to the stimulus acoustics. The difference plots in Figure 4 show that broad regions of the IC frequency map respond similarly to different vowel sounds. For example, the words “seed” and “deed” evoked more activity among high frequency IC sites (>9 kHz) and less activity among low frequency sites (<6 kHz), which was expected from their power spectrums (Fig. 7).

Figure 4. Response of the entire population of IC neurons to each of the vowel sounds tested. Neurograms are constructed from the average PSTH of 187 IC recording sites ordered by the characteristic frequency of each recording site. The average PSTH for all the sites recorded is shown above each neurogram. The peak firing rate for the population PSTH was 550 Hz. To illustrate how the spatial activity patterns differed across each vowel, subplots to the right of each neurogram show the difference between the response to each sound and the mean response to all 5 sounds. A white line is provided at zero to make it clearer which responses are above or below zero. Compared with the average response, the words “seed” and “deed” evoked more activity among high frequency sites (>9 kHz) and less activity among low frequency sites (<6 kHz), which was expected from their power spectrums (Fig. 7).
The average correlation between the responses to the 10 speech sounds was 0.41 ± 0.01. The average correlation was significantly less for A1 sites (0.29 ± 0.01, \( P < 10^{-10} \)). Only 10% of nearby A1 pairs exhibited a correlation coefficient above 0.8 compared with 24% of IC sites. The observation that A1 responses to vowels are less determined by characteristic frequency compared with IC responses is consistent with earlier reports that neurons in A1 are sensitive to interactions among many different acoustic features which leads to a reduction in the information redundancy compared with IC (Steinschneider et al. 1990; DeWeese et al. 2005; Chechik et al. 2006; Mesgarani et al. 2008).

Since some vowel sounds are more difficult for rats to distinguish than others (Fig. 3), we expected that the difference in the neural responses would be relatively small for vowels that were difficult to distinguish and larger for vowels that were easier to distinguish. For example, the average absolute value of the difference in the response to "dad" and "dead" across all A1 and IC sites was 5.3 ± 0.3 and 15.4 ± 1.3 Hz, respectively, which is significantly less than the absolute value of the difference between "dad" and "deed" \((P < 0.05)\). Consistent with this prediction, the absolute value of the difference in the firing rate between each pair of vowel sounds was correlated \((R^2 = 0.58, P < 0.02)\) with the behavioral discrimination ability (Fig. 6). These results support the hypothesis that differences in the average neural firing rate are used to discriminate between vowel sounds.

It is important to confirm that any potential neural code can duplicate the accuracy of sensory discrimination. Analysis using a nearest-neighbor classifier makes it possible to document neural discrimination based on single trial data and allows the direct correlation between neural and behavioral discrimination in units of percent correct (Engineer et al. 2008). The classifier compares the PSTH evoked by each stimulus presentation with the average PSTH evoked by each stimulus and selects the most similar (see Materials and Methods). The

Figure 5. Response of the entire population of A1 neurons recorded to each of the vowel sounds tested. The conventions are the same as Figure 4, except that the height of the PSTH scale bar represents a firing rate of 250 Hz.
number of bins used (and thus the temporal precision) is a free parameter. In the simplest form, the classifier compares the total number of spikes recorded on each trial with the average number of spikes in response to 2 speech sounds and classifies each trial as the sound that evoked the closest number of spikes on average (Fig. 8b). Vowel discrimination using this simple method was remarkably similar to behavioral discrimination (Fig. 9a). For example, using this method, the number of spikes evoked at a single IC site on a single trial is sufficient to discriminate between "dad" and "deed" 78 ± 1% of the time.

Figure 6. The distinctness of the spatial patterns evoked in IC (a) and A1 (b) by each vowel pair was correlated with the distance between the vowels in the feature space formed by the first and second formant peaks. The distinctness of the spatial patterns evoked in IC (c) and A1 (d) was also correlated with behavior discrimination of the vowel pairs. In this figure, neural distinctness was quantified as the absolute value of the difference in the average firing rate recorded in response to each pair of vowel sounds. A 300 ms long analysis window beginning at vowel onset was used to quantify the vowel response.

Figure 7. Power spectrums of the 5 vowel sounds with 2 initial consonants. The initial recordings were shifted one octave higher using the STRAIGHT vocoder to better match the rat hearing range (Kawahara 1997).
This performance is very similar to the behavioral discrimination of 78 ± 2% correct. The correlation between behavioral and neural discrimination using a single trial from a single IC site was high ($R^2 = 0.59, P = 0.02$) for the 9 vowel tasks tested (4 distracter vowels with 2 different initial consonants plus the pitch task). Although average accuracy was lower compared with IC (57 ± 0.5% vs. 73 ± 2%, $P < 0.05$), neural discrimination based on individual A1 sites was also significantly correlated with vowel discrimination behavior ($R^2 = 0.58, P = 0.02$; Fig. 9c). The difference between A1 and IC neural discrimination accuracy may simply result from the lower firing rates of the multiunit activity in A1 compared with IC (26 ± 0.85 vs. 52 ± 3 Hz, $P < 0.05$). Single unit analysis suggests that the difference in multiunit activity is due at least in part to higher and more reliable firing rates of individual IC neurons compared with individual A1 neurons (6.1 ± 1.4 vs. 1.4 ± 0.4 Hz, $P < 0.05$), as seen in earlier studies (Chechik et al. 2006). These results suggest that vowel sounds are represented in the early central auditory system as the average firing rate over the entire duration of these sounds.

Neural discrimination of consonant sounds was only correlated with behavioral discrimination when the classifier was able to use spike timing information (Engineer et al. 2008). To test the hypothesis that rats might also use spike timing information when discriminating vowel sounds, we compared vowel behavior with classifier performance using spike timing (Fig. 8a). The addition of spike timing information increased discrimination accuracy, such that neural discrimination greatly exceeded behavioral discrimination (Fig. 9b). As a result, there was no correlation between the difficulty of neural and behavioral discrimination of each vowel pair (Fig. 9b,d). This result suggests the possibility that spike timing is used for discrimination of consonant sounds but not for discrimination of the longer vowel sounds. Analysis based on 23 well-isolated IC single units confirmed that vowel discrimination is best correlated with neural discrimination based on spike count, while consonant discrimination was best correlated with neural discrimination that includes spike timing information.

Since our earlier study of consonant discrimination only evaluated neural discrimination by A1 neurons, we repeated the analyses (Fig. 8c,d) using IC responses (Fig. 10) to the set of consonant sounds tested in our earlier study (Engineer et al. 2008). Classifier performance was best correlated ($R^2 = 0.73, P = 0.0008$) with behavioral discrimination when spike timing information was provided (Fig. 11a). The observation that behavior was still weakly correlated with neural discrimination when spike timing was eliminated in IC (Fig. 11b), but not in A1 (Engineer et al. 2008), suggests that the coding strategies used to represent consonant and vowel sounds may become more distinct as the information ascends the central auditory system (Chechik et al. 2006).

As in our earlier report, the basic findings were not dependent on the specifics of the classifier distance metric used or the bin size (Fig. 12). Consonant discrimination was well correlated with classifier performance when spike timing information was binned with 1–20 ms precision. In contrast, vowel discrimination was well correlated with classifier performance when neural activity was binned with bins larger than 100 ms. These observations are consistent with the hypothesis that different stimulus classes may be represented with different levels of temporal precision (Poeppel 2003; Boemio et al. 2005; Panzeri et al. 2010).

**Discussion**

Sensory discrimination requires the ability to detect differences in the patterns of neural activity generated by different stimuli.
Stimuli that evoke different activity patterns are expected to be easier to discriminate than stimuli that evoke similar patterns. There are many possible methods to compare neural activity patterns. This is the first study to evaluate plausible neural coding strategies for consonant and vowel sounds by comparing behavioral discrimination with neural discrimination. We compared behavioral performance on 9 vowel discrimination tasks and 11 consonant discrimination tasks with neural discrimination using activity recorded in inferior colliculus and primary auditory cortex. Vowel discrimination was highly correlated with neural discrimination when spike timing was eliminated and was not correlated when spike timing was preserved. In contrast, performance on 11 consonant discrimination tasks was highly correlated with neural discrimination when spike timing was preserved and was not well correlated when spike timing was eliminated. These results suggest that in the early stages of auditory processing, spike count encodes vowel sounds and spike timing encodes consonant sounds. Our observation that neurons in IC and A1 respond similarly, but not identically, to different speech sounds is consistent with the progressive processing of sensory information across multiple cortical and subcortical stations (Chechik et al. 2006; Hernández et al. 2010; Tsunada et al. 2011).

There has been a robust discussion over many years about the timescale of neural codes (Cariani 1995; Parker and Newsome 1998; Jacobs et al. 2009). Neural responses to speech sounds recorded in human and animal auditory cortex demonstrate that some features of speech sounds (e.g., voice onset time) can be extracted using spike timing (Young and Sachs 1979; Steinschneider et al. 1982, 1990 1995, 1999 2005; Palmer 1990; Eggermont 1995; Liégeois-Chauvel et al. 1999; Wong and Schreiner 2003). Information theoretic analyses of neural activity patterns almost invariably find that neurons contain more information about the sensory world when precise spike timing is included in the analysis (Middlebrooks et al. 1994; Schnupp et al. 2006; Foffani et al. 2009; Jacobs et al. 2009). However, the mere presence of information does not imply that the brain has access to the additional information. Direct comparison of behavioral discrimination and neural discrimination with and without spike timing information can be used to evaluate whether spike timing is likely to be used in behavioral judgments. Such experiments are technically challenging, and surprisingly, few studies have directly evaluated different coding strategies using a sufficiently large set of stimuli to allow for a statistically valid correlation analysis.

Some of the most convincing research on neural correlates of sensory discrimination suggests that sensory information is

![Figure 9.](image-url)

Neural discrimination was well correlated with behavioral discrimination of vowels when the average firing rate over a long analysis window was used but was not significantly correlated with behavior when spike timing was considered. Neural discrimination was based on single trial multunit activity recorded from individual sites in IC (a and b) and A1 (c and d). The nearest-neighbor classifier assigns each sweep to the stimulus that evokes the most similar response on average. In a and c, the classifier used a 300 ms long analysis window beginning at vowel onset and was used to quantify the vowel response. In b and d, the classifier compared the temporal pattern recorded over the same window and binned with 1 ms precision. These results suggest that vowel sounds are represented by the average spatial activity pattern and that spike timing information is not used.
encoded in the mean firing rate averaged over 50–500 ms (Britten et al. 1992; Shadlen and Newsome 1994; Pruett et al. 2001; Romo and Salinas 2003; Liu and Newsome 2005; Wang et al. 2005; Lemus et al. 2009). These studies found no evidence that spike timing information was used for sensory discrimination. However, a recent study of speech sound processing in auditory cortex came to the opposite conclusion (Engineer et al. 2008). In that study, behavioral discrimination of consonant sounds was only correlated with neural discrimination if spike timing information was used in decoding the speech sounds. These apparently contradictory studies seemed at first to complicate rather than resolve the long-standing debate about whether the brain uses spike timing information (Young 2010). Our new observation that vowel discrimination

![Figure 10](image1.png)

**Figure 10.** Response of the entire population of IC neurons recorded to each of the consonant sounds tested. Neurograms are constructed from the average PSTH of 187 IC recording sites ordered by the preferred frequency of each recording site. The average PSTH for all the sites recorded is shown above each neurogram. The height of the scale bar to the left of each average PSTH represents a firing rate of 600 Hz. Only the initial onset response to each consonant sound is shown to allow the relative differences in spike timing to be visible. For example, high frequency neurons respond earlier to “dad” compared with “bad.” The opposite timing occurs for low frequency neurons. These patterns are similar to A1 responses (Engineer et al. 2008).

![Figure 11](image2.png)

**Figure 11.** Neural discrimination was well correlated with behavioral discrimination of consonants when spike timing information was used but was not as well correlated with behavior when the average firing rate over a long analysis window was used. Neural discrimination was based on single trial multiunit activity recorded from individual sites in IC and A1. The nearest-neighbor classifier assigns each sweep to the stimulus that evokes the most similar response on average. In (a), the classifier compared the temporal pattern recorded within 40 ms of consonant onset binned with 1 ms precision. In (b), the classifier used the average firing rate over this same period (i.e., spike timing information was eliminated). These results suggest that spike timing plays an important role in the representation of consonant sounds.
is best accounted for by decoding mean firing rate and ignoring spike timing, while consonant discrimination is best accounted by decoding that includes spike timing supports the hypothesis that the brain can process sensory information within a single structure at multiple timescales (Cariani 1995; Wang et al. 1995, 2005; Parker and Newsome 1998; Victor 2000; Poeppel 2003; Boemio et al. 2005; Mesgarani et al. 2008; Buonomano and Maass 2009; Panzeri et al. 2010; Walker et al. 2011).

Earlier conclusions that the auditory, visual, and somatosensory systems use a simple rate code (and not spike timing) may have resulted from the choice of continuous or periodic stimuli (Britten et al. 1992; Shadlen and Newsome 1994; Preutt et al. 2001; Romo and Salinas 2003; Liu and Newsome 2005; Lemus et al. 2009). Additional behavior and neurophysiology studies are needed to determine whether complex transient stimuli are generally represented using spike timing information (Cariani 1995; Victor 2000). The long-standing debate about the appropriate timescale to analyze neural activity would be largely resolved if it were found that spike timing was required to represent transient stimuli but not steady state stimuli (Cariani 1995; Parker and Newsome 1998; Jacobs et al. 2009; Panzeri et al. 2010). Computational modeling will likely be useful in clarifying the potential biological mechanisms that could be used to represent sensory information on multiple timescales (Buonomano and Maass 2009; Panzeri et al. 2010).

Our observation that consonant and vowel sounds are encoded by neural activity patterns that may be read out at different timescales supports earlier psychophysical, clinical, and imaging evidence that consonant and vowel sounds might be represented differently in the brain (Pisoni 1973; Shannon et al. 1995; Boatman et al. 1977; Caramazza et al. 2000; Poeppel 2003; Bonatti et al. 2005; Carreiras and Price 2008; Wallace and Blumstein 2009). The findings that consonant discrimination is more categorical, more sensitive to temporal degradation, and less sensitive to spectral degradation (compared with vowel discrimination) are consistent with our hypothesis that spike timing plays a key role in the representation of consonant sounds (Pisoni 1973; Shannon et al. 1995; Kasturi et al. 2002; Xu et al. 2005; Engineer et al. 2008). Detailed studies of rat behavioral and neural discrimination of noise-vocoded speech could help explain the differential sensitivity of consonant and vowel sounds to spectral and temporal degradation and may be able to explain why consonant place of articulation information is so much more sensitive to spectral degradation compared with consonant manner of articulation and voicing (Shannon et al. 1995; Xu et al. 2005). The findings that consonant and vowel discrimination develop at different rates and that consonant and vowel discrimination can be independently impaired by lesions or electrical stimulation support the hypothesis that higher cortical fields decode auditory information on different timescales (Boatman et al. 1997; Caramazza et al. 2000; Kudoh et al. 2006; Porter et al. 2011). Regional differences in temporal integration could play an important role in generating sensory processing streams (like the “what” and “where” pathways) and hemispheric lateralization. Physiology and imaging results in animals and humans suggest that the distinct processing streams diverge soon after primary auditory cortex (Shankweiler and Studdert-Kennedy 1967; Schwartz and Tallal 1980; Liégeois-Chauvel et al. 1999; Binder et al. 2000; Rauschecker and Tian 2000; Zatorre et al. 2002; Poeppel 2003; Scott and Johnsrude 2003; Boemio et al. 2005; Obleser et al. 2006; Rauschecker and Scott 2009; Recanzone and Cohen 2010).

Our results are consistent with the asymmetric sampling in time hypothesis (AST, Poeppel 2003). AST postulates that 1) the central auditory system is bilaterally symmetric in terms of temporal integration up to the level of primary auditory cortex, 2) left nonprimary auditory cortex preferentially extracts information using short (20–50 ms) temporal integration windows, and 3) right nonprimary auditory cortex preferentially extracts information using long (150–250 ms) temporal integration windows. Neural recordings from secondary auditory fields and focal inactivation of secondary auditory fields are needed to test whether the AST occurs in nonhuman animals (Lomber and Malhotra 2008). If confirmed, differential temporal integration by different cortical fields is likely to contribute to the natural robustness of sensory processing, including speech processing, to noise and many other forms of distortion.

Future studies are needed to evaluate the potential effect of anesthesia and attention on speech sound responses. Earlier studies have shown qualitatively similar response properties in awake and anesthetized preparations for auditory stations up to the level of A1 (Watanabe 1978; Steinschneider et al. 1982; Chen et al. 1996; Sinex and Chen 2000; Cunningham et al. 2002; Engineer et al. 2008). However, anesthesia has been shown to reduce the maximum rate of click trains that A1 neurons can respond to, so it is likely that anesthesia reduces the A1 response to speech sounds, especially the sustained response to vowel sounds (Anderson et al. 2006; Rennaker et al. 2007).

Although human imaging studies have shown little to no effect of attention on primary auditory cortex responses to speech sounds, single unit studies are needed to better understand how attention might shape the cortical representation of speech (Grady et al. 1997; Hugdahl et al. 2003; Christensen et al. 2008). Single unit studies would also clarify whether speech sound coding strategies differ across cortical layers.
Human listeners can adapt to many forms of speech sound degradation (Miller and Nicely 1955; Strange 1989; Shannon et al. 1995). Our recent demonstration that the A1 coding strategy is modified in noisy environments suggests the possibility that other forms of stimulus degradation also modify the analysis windows used to represent speech sounds (Shetake et al. 2011).

In summary, we would like to thank Kevin Chang, Gabriel Mettlach, Jarod Roland, Roshini Jain, and Dwayne Listhrop for assistance with microelectrode recording technology advances, it may be possible to test our hypothesis that consonant and vowel sounds are represented on different timescales in human auditory cortex. If confirmed, this hypothesis could explain a wide range of psychological findings and may prove useful in understanding a variety of common speech processing disorders.

Funding
National Institute on Deafness and Other Communication Disorders (grant numbers R01DC010433, R15DC006624).

Notes


