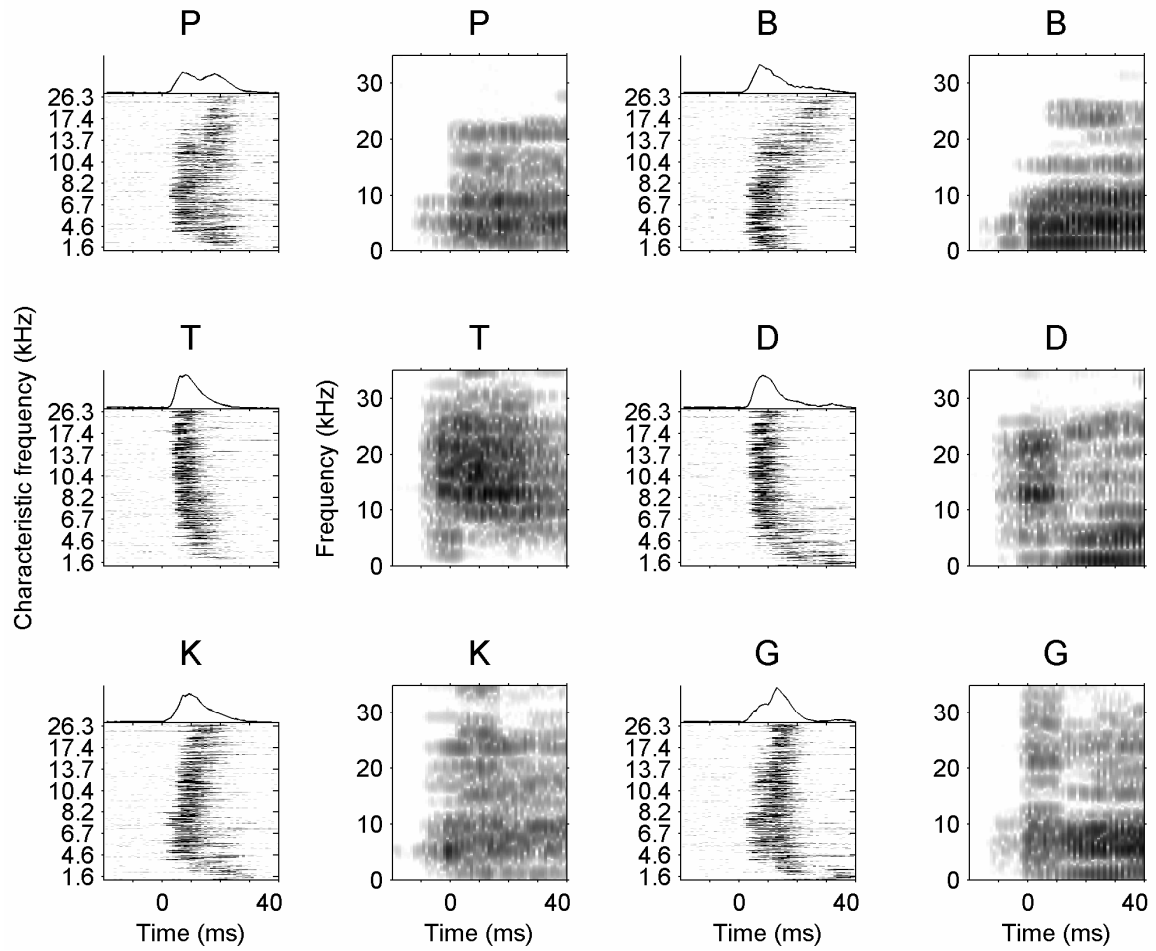
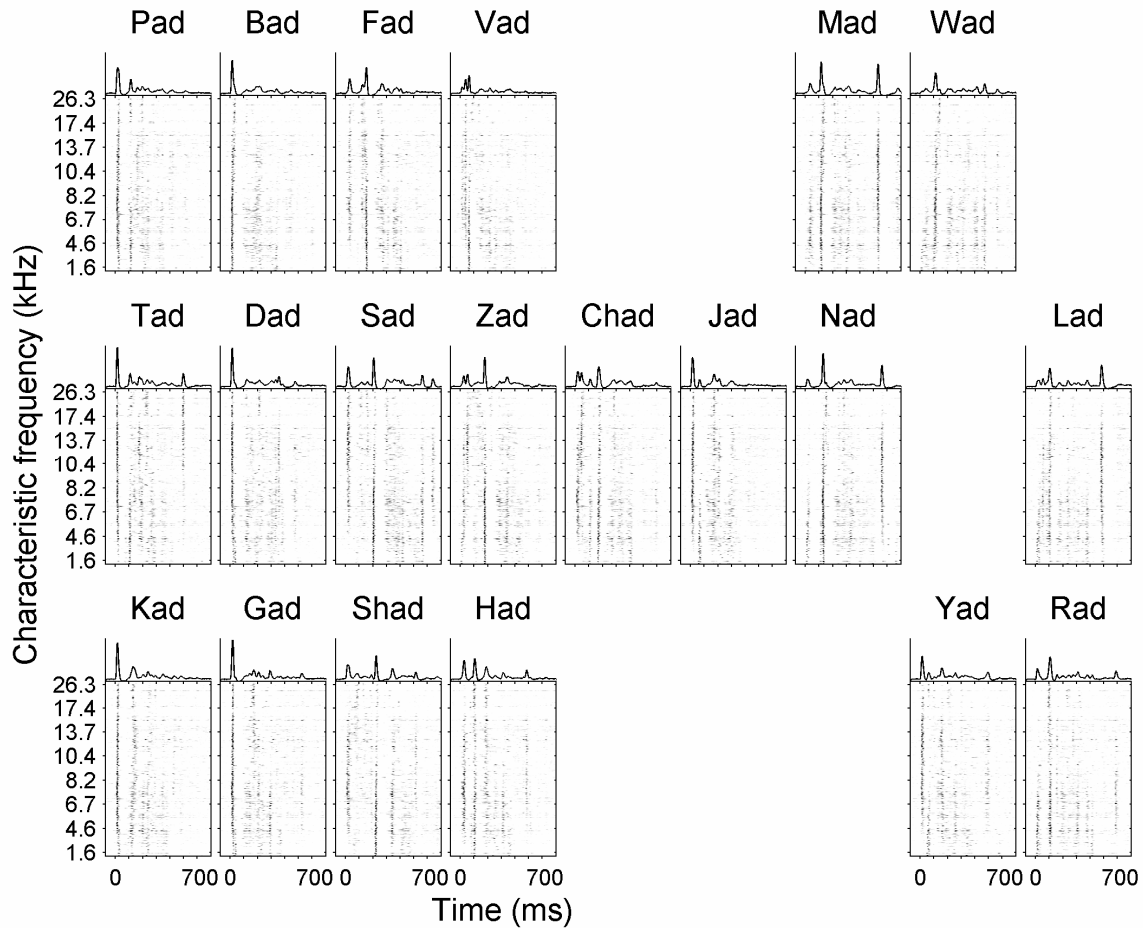


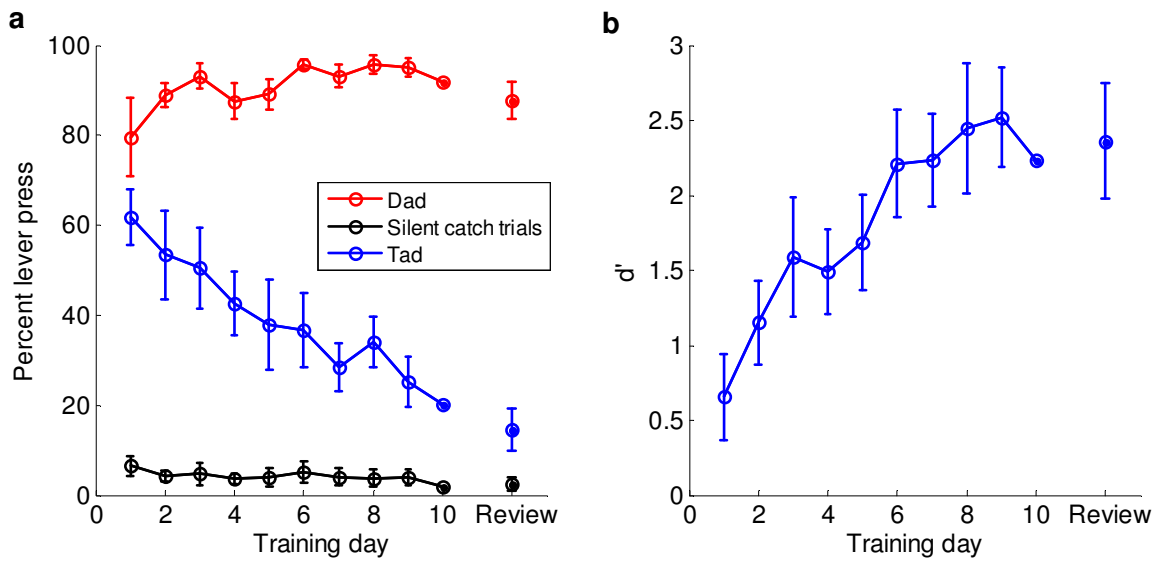
Supplementary Figure 1 Spectrograms of each speech sound grouped by manner and place of articulation. Words with unvoiced initial consonants are underlined. All conventions are identical to **Figure 1**, except that the x axis has been limited to 40 ms to facilitate comparison with **Figure 2**.



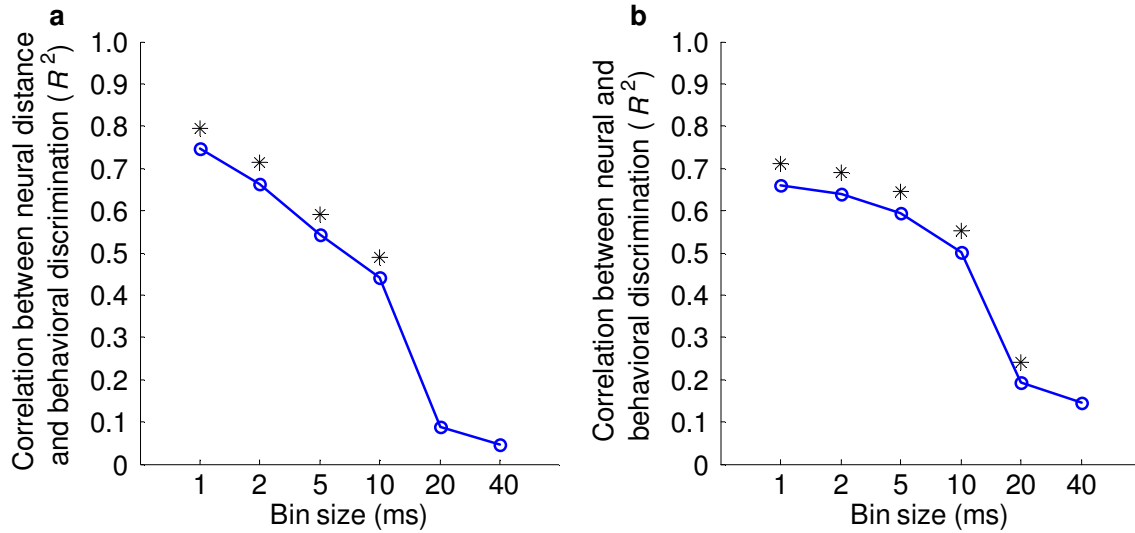
Supplementary Figure 2 Onset neurograms and spectrograms of each stop consonant are shown to facilitate visual comparison. All conventions are identical to **Figure 2** and **Supplementary Figure 1**.



Supplementary Figure 3 Neurograms and population PSTH's of rat A1 neurons to twenty English consonants. Post-stimulus time histograms (PSTH) were ordered by the characteristic frequency (kHz) of each recording site (*y* axis). Multi-unit data was collected from each of 445 multi-unit recording sites in eleven experimentally naïve adult rats. Conventions are identical to **Figure 2**, except for the time interval included on the *x* axis (–75 to 700 ms). Late response peaks were due to the onset of the vowel /a/ and the terminal /d/ (as in 'sad').

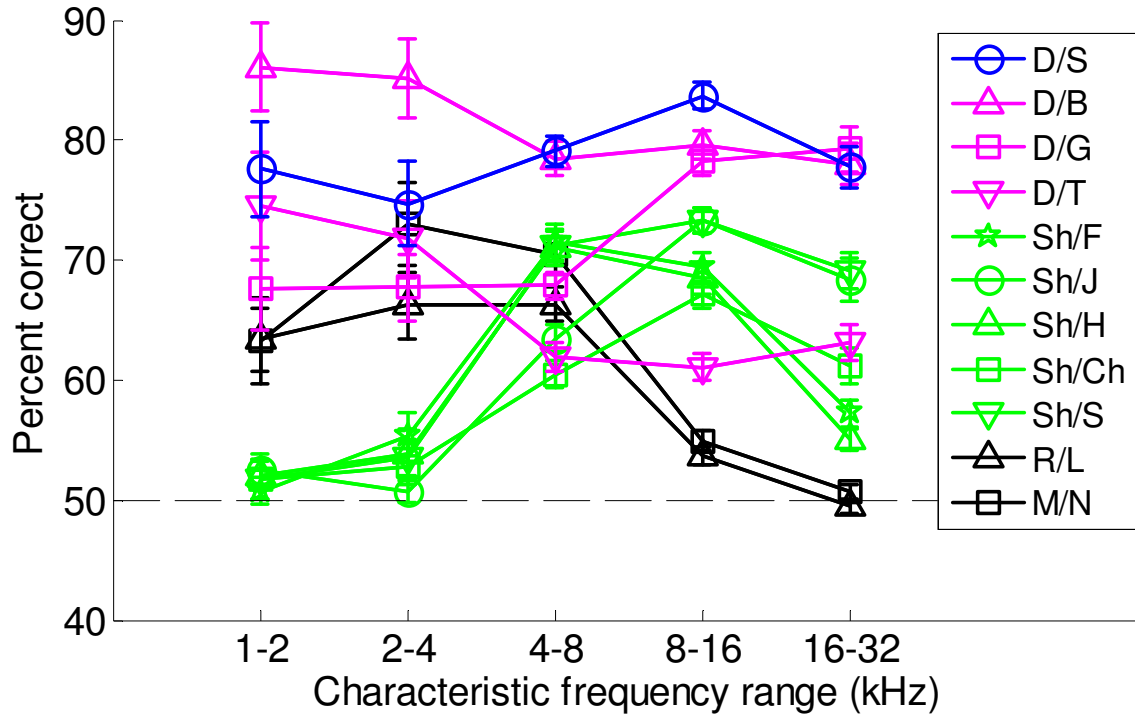


Supplementary Figure 4 Average learning curve illustrates asymptotic performance by the seventh and eighth days of /d/ versus /t/ discrimination ($n = 6$ rats). In **(a)** the percent of trials in which rats pressed the lever is given. In **(b)**, the discrimination between /d/ and /t/ is shown using d' . Review indicates performance during review sessions conducted after several weeks of training on other discriminations. For this task, we only had data from four rats on day 9 and one rat on day ten (see methods).

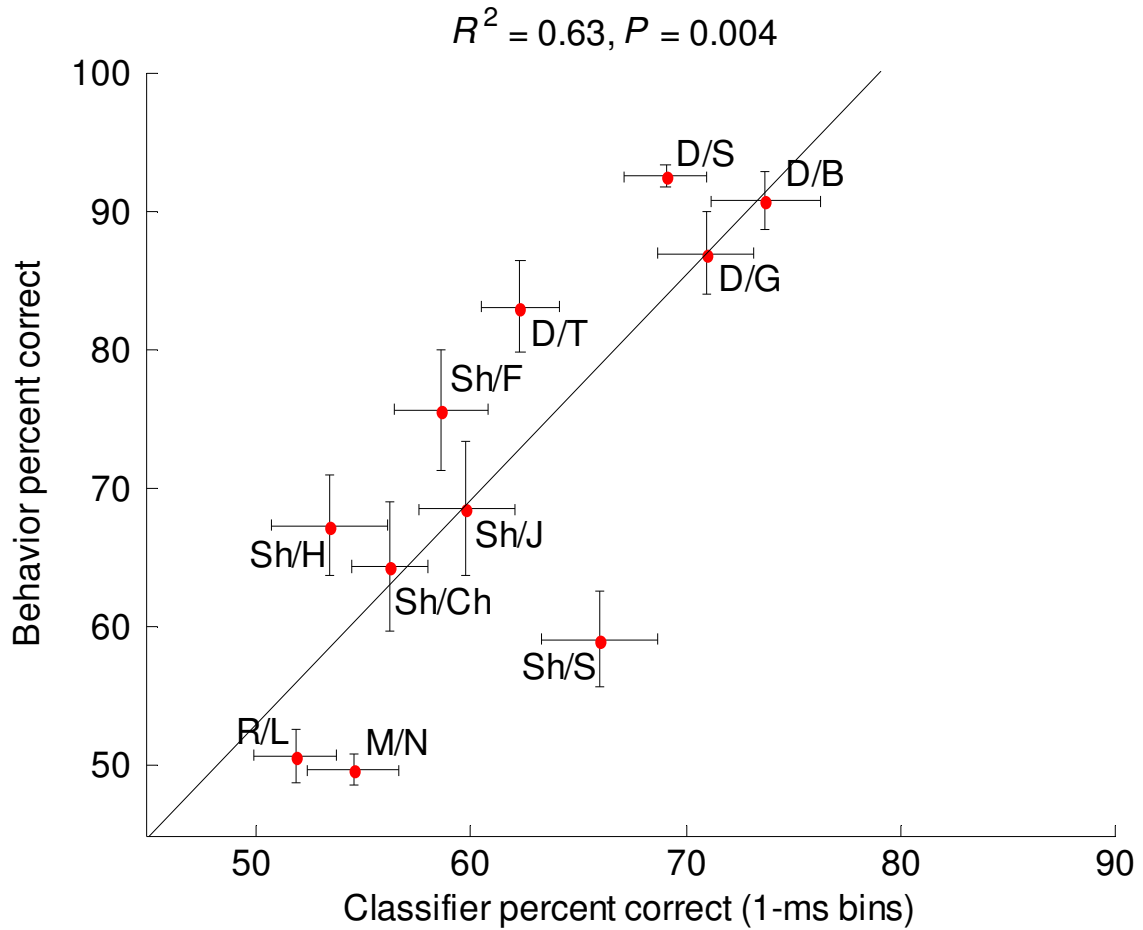


Supplementary Figure 5 Percent of variance explained decreases as bin size increases. **(a)**

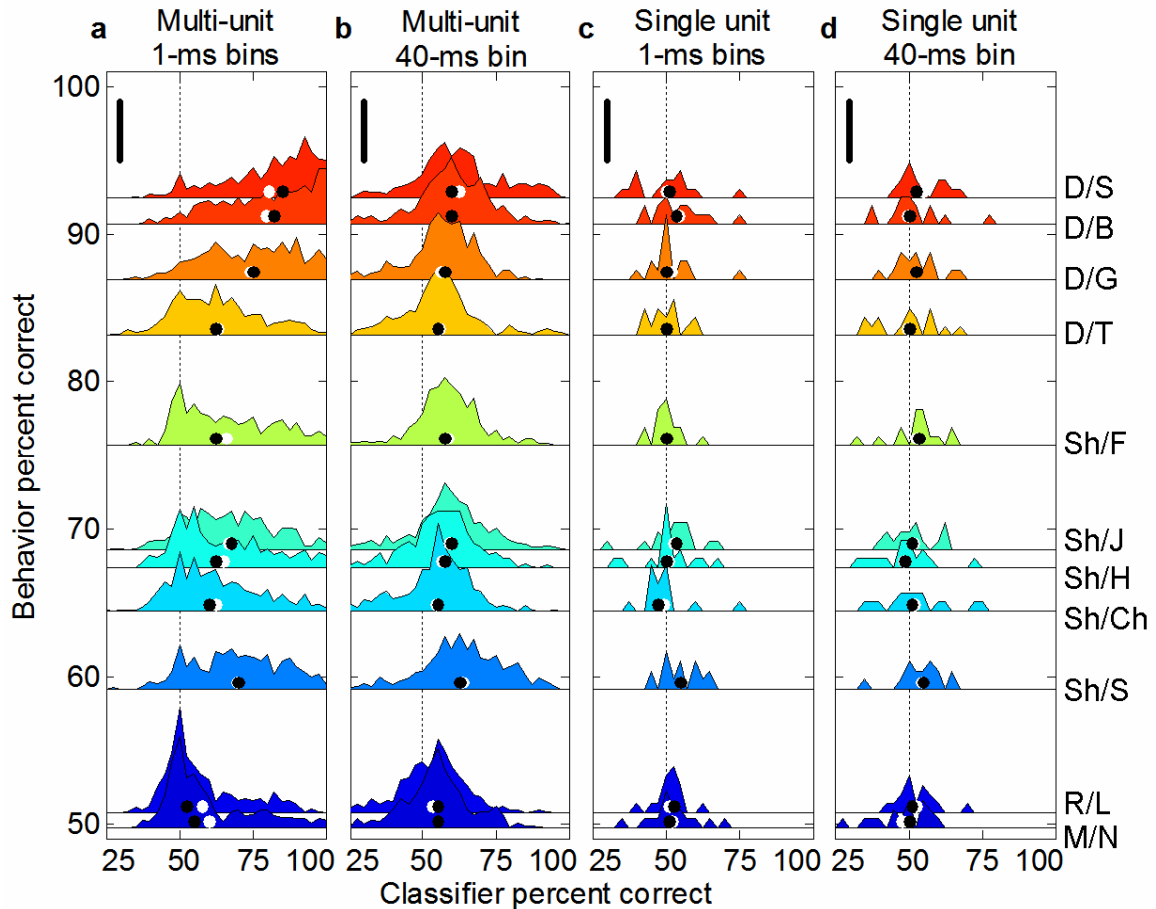
Percent of variance across the 11 consonant discrimination tasks that is explained using Euclidean distance calculated from the 40 ms onset response. **(b)** Percent of variance explained using classifier percent correct. The observation that neural discrimination in the auditory system depends upon temporal information is consistent with earlier studies⁹⁻¹¹. Asterisks indicate correlations that are statistically significant (Pearson's correlation coefficient, $P < 0.05$).



Supplementary Figure 6 Classifier performance on each task as a function of tone characteristic frequency. High frequency neurons provide little information about nasals and liquids (black). Low frequency neurons provide little information about fricatives and affricates (green). Both high and low frequency neurons provide information about stop consonants (purple) and manner of articulation (blue). Classifier data is identical to that used in **Figure 5c**.



Supplementary Figure 7 Strong correlation between behavioral performance on 11 consonant discrimination tasks and neural discrimination using data from 40 A1 multi-unit recording sites in 6 chronically implanted awake rats. Conventions are identical to **Figure 5c**, except data was recorded without anesthesia. The solid line indicates the best linear fit ($R^2 = 0.63, P = 0.004$ compared to $R^2 = 0.66, P = 0.002$ in anesthetized rats).



Supplementary Figure 8 Neural discrimination using the onset response of individual multi-units with precise temporal information predicts behavioral discrimination. **(a)** The population distribution of the onset response of 445 multi-unit sites using forty 1 ms bins and **(b)** a single 40 ms bin. **(c)** The population distribution of the onset response of 16 single unit sites using forty 1 ms bins and **(d)** a single 40 ms bin. Black circles indicate the median classifier percent correct for each of the 11 consonant discrimination tasks, and white circles indicate the mean classifier percent correct. The dashed line indicates chance performance. The scale bar at the top left of each subplot represents 10% of multi-units sites **(a,b)** and 40% of single unit sites **(c,d)**. Classifier percent correct is binned into 2.5% correct bins. Each task is labeled to the right of the figure.

	Euclidean distance		City block distance		Chebychev distance	
	Mean	Single trial	Mean	Single trial	Mean	Single trial
Temporal onset	$R^2 = 0.75$ $P = 0.0006$	$R^2 = 0.66$ $P = 0.002$	$R^2 = 0.59$ $P = 0.005$	$R^2 = 0.74$ $P = 0.0007$	$R^2 = 0.39$ $P = 0.04$	$R^2 = 0.63$ $P = 0.003$
Rate onset	$R^2 = 0.046$ $P = 0.5$	$R^2 = 0.14$ $P = 0.2$	$R^2 = 0.08$ $P = 0.41$	$R^2 = 0.15$ $P = 0.24$	$R^2 = 0.01$ $P = 0.75$	$R^2 = 0.14$ $P = 0.25$

Supplementary Table 1 Neural activity was only correlated with behavioral performance when temporal precision was preserved. Both mean activity and single trial activity were significantly correlated with behavior when temporal information was preserved using Euclidean distance, City block distance, and Chebychev distance. There was no significant correlation between behavior and neural activity when temporal information was eliminated using mean or single trial activity for each of the three models. Significant correlations are shown in bold.

Cortical activity patterns predict speech discrimination ability

Crystal T. Engineer, Claudia A. Perez, YeTing H. Chen, Ryan S. Carraway, Amanda C. Reed, Jai A. Shetake, Vikram Jakkamsetti, Kevin Q. Chang and Michael P. Kilgard

SUPPLEMENTARY DATA

Neurogram quantification and comparison with earlier physiology studies

Consistent with earlier observations in humans and non-human animals¹⁻⁴, more spikes were evoked during the second peak of the unvoiced stop consonant /p/ than during the second peak of the voiced stop consonant /b/ (1.05 ± 0.08 additional spikes from 70 to 170 ms after consonant onset, two tailed *t*-test, $n = 445$, $P < 0.00001$). An increase in firing rate was also seen in the unvoiced stop consonants /k/ and /t/ compared to the voiced stop consonants /g/ and /d/ (two tailed *t*-test, $n = 445$, $P < 0.00001$, **Supplementary Fig. 3** and **Supplementary Video 2**). The stop consonants generated the strongest onset response (40 ms), the fricatives and affricates had an intermediate response, and the sonorants (nasals, glides, and liquids) generated the weakest response. The six stops had an average peak firing rate of 524 ± 13 Hz, the eight fricatives and affricates had an average peak firing rate of 284 ± 9 Hz, and the six sonorants had an average peak firing rate of 203 ± 7 Hz (ANOVA $F_{19,444} = 284$, $MSE = 43745$, $P < 0.00001$, **Fig. 2**).

Consistent with earlier observations in human and monkey auditory cortex, consonants that differ in place of articulation generate different spatial patterns of A1 activity⁵⁻⁷. In the low frequency region of A1 (< 9 kHz), the consonant /d/ had the weakest response of the voiced stop consonants (two tailed *t*-test, $n = 445$, $P < 0.001$, 40 ms onset window), while in the high frequency region of A1 (> 9 kHz), the consonant /b/ had the weakest response (two tailed *t*-test, $n = 445$, $P < 0.00001$). The differential spatial patterns were most apparent when comparing activity produced during 1 ms windows. For example, 8 ms after sound onset, /b/ primarily

activated sites with 2–9 kHz CF's, /g/ activated sites with 6–14 kHz CF's, and /d/ activated sites with 8–25 kHz CF's (**Supplementary Video 1**).

The fact that voicing, manner, and place of articulation alter A1 activity as in earlier studies conducted in humans and non-human animals with or without anesthesia supports the use of the rat model and indicates that frequency shifting speech sounds to match the rat hearing range does not dramatically distort the previously reported pattern of cortical responses^{1–7}.

Behavioral discrimination time course

Average performance during the seventh and eighth days of training was used to estimate the relative difficulty of different consonant discriminations. By the end of day eight, each rat had performed 3544 ± 757 (mean \pm s.d.) trials. All rats reached asymptotic performance by the seventh day of training on each task.

Two additional days of training (i.e. days 9 and 10) were used to confirm asymptotic performance (paired *t*-test, $P > 0.05$). However, for /d/ versus /t/, only one rat was trained for ten days and the others were inadvertently trained for only nine ($n = 4$ rats) or seven ($n = 1$ rat) days. The learning curve for the /d/ versus /t/ task is shown in **Supplementary Figure 4a**. Although the false alarm rate was trending down during the last three days of training, discrimination quantified by d' did not improve because the hit rate was also trending down (**Supplementary Fig. 4b**).

To ensure discrimination ability was not strongly influenced by task order, each rat underwent a review session (after completion of all of the two week consonant discrimination tasks) during which it reviewed each of the tasks it was trained on for two additional days per task. Performance was not significantly different between the seventh and eighth days of training and the two days of review training (ANOVA $F_{9,36} = 0.49$, $MSE = 0.28$, $P = 0.87$).

Onset neurograms only predict behavior when spatial and temporal information is preserved

The correlation between behavioral discrimination and the Euclidean distance between neurogram pairs was highest when the classifier was provided 40 ms of A1 onset activity in 1 ms

bins ($R^2 = 0.75$, $P = 0.0006$, **Fig. 5a**). When spike timing information was removed by using a single 40 ms bin, the distance between neurogram pairs was not significantly correlated with behavior ($R^2 = 0.046$, $P = 0.5$, **Fig. 5b**). When the entire duration (i.e. including the vowel and final /d/) was included, the distances between neurogram pairs (**Supplementary Fig. 3**) were not significantly correlated with behavior whether 1 ms bins were used ($R^2 = 0.07$, $P = 0.4$) or a single 700 ms bin was used ($R^2 = 0.08$, $P = 0.4$). The critical observation is that cortical responses are only correlated with behavior when spike timing is considered.

To test whether spatial information is also required for accurate consonant discrimination, we compared the Euclidean distances between the mean population PSTH's (i.e. the PSTH above each neurogram in **Figure 2**) for each pair of sounds. Although precise temporal information was maintained, behavior was not significantly correlated with the similarity of responses averaged across all 445 A1 sites ($R^2 = 0.23$, $P = 0.1$). Similarly, no correlation was observed when both spatial and temporal information was removed (i.e. comparing the total number of spikes in each neurogram in **Figure 2**, $R^2 = 0.14$, $P = 0.2$). These results indicate that the precise spatiotemporal pattern of A1 activity is best correlated with behavior.

The Euclidean distances between the first 40 ms of the consonant spectrograms (**Supplementary Fig. 1**) were modestly correlated with behavior ($R^2 = 0.40$, $P = 0.038$), and no correlation was observed when the entire spectrogram was used ($R^2 = 0.03$, $P = 0.61$). This result confirms our expectation that behavioral discrimination ability is related to acoustic similarity.

Relationship between tone and consonant responses

As expected, tone response properties explain much of the variance in the ability of different recording sites to discriminate each of the eleven pairs evaluated. For example, low frequency neurons were better than high frequency neurons at distinguishing nasals and liquids (**Supplementary Fig. 6** black lines), and the opposite pattern was seen for fricatives and affricates (**Supplementary Fig. 6** green lines). Stops which begin with broadband noise could be

reliably discriminated by neurons from any region of A1. These observations are consistent with the acoustics of the consonants studied (**Supplementary Fig. 1**)⁸. Broader tone frequency tuning, shorter response latency, lower tone thresholds, and greater tone evoked response strength were each positively correlated with neural discrimination ability ($R^2 = 0.15$, $P = 1.1 \times 10^{-17}$; $R^2 = 0.2$, $P = 7.8 \times 10^{-23}$; $R^2 = 0.1$, $P = 3.2 \times 10^{-12}$; $R^2 = 0.38$, $P = 2.1 \times 10^{-47}$). Collectively, these factors account for 47.3% of the variance in average consonant discrimination ability ($P = 0$).

Other analysis models

Although there are an infinite number of potential models that could be used to read out neural activity, we attempted to use common analysis methods that relied on as few assumptions as possible. Euclidean distance was selected to compare average and single trial responses (**Figs. 3** and **5–7**) as it is the most commonly used metric of similarity. Our results are not dependent on the use of Euclidean distance. When Chebychev and city block distances were computed with 1 ms bins, neural discrimination was also well correlated with behavior (**Supplementary Table 1**). Neither method was significantly correlated with behavior when a single bin of activity was used.

SUPPLEMENTARY REFERENCES

1. Steinschneider, M., et al. 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 (2005).
2. Steinschneider, M., Volkov, I.O., Noh, M.D., Garell, P.C. & Howard, M.A., 3rd. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *Journal of neurophysiology* **82**, 2346–2357 (1999).
3. Wong, S.W. & Schreiner, C.E. Representation of CV-sounds in cat primary auditory cortex: intensity dependence. *Speech Communication* **41**, 93–106 (2003).
4. Steinschneider, M., Fishman, Y.I. & Arezzo, J.C. Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. *The Journal of the Acoustical Society of America* **114**, 307–321 (2003).
5. Blumstein, S.E. & Stevens, K.N. Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants. *The Journal of the Acoustical Society of America* **66**, 1001–1017 (1979).
6. Obleser, J., Lahiri, A. & Eulitz, C. Auditory-evoked magnetic field codes place of articulation in timing and topography around 100 milliseconds post syllable onset. *NeuroImage* **20**, 1839–1847 (2003).

7. Steinschneider, M., Reser, D., Schroeder, C.E. & Arezzo, J.C. Tonotopic organization of responses reflecting stop consonant place of articulation in primary auditory cortex (A1) of the monkey. *Brain research* **674**, 147–152 (1995).
8. Stevens, K.N. *Acoustic Phonetics* (MIT Press, Cambridge, MA, 2000).
9. Narayan, R., Grana, G. & Sen, K. Distinct time scales in cortical discrimination of natural sounds in songbirds. *J Neurophysiol* **96**, 252–258 (2006).
10. Schnupp, J.W., Hall, T.M., Kokelaar, R.F. & Ahmed, B. Plasticity of temporal pattern codes for vocalization stimuli in primary auditory cortex. *J Neurosci* **26**, 4785–4795 (2006).
11. Walker, K.M., Ahmed, B. & Schnupp, J.W. Linking cortical spike pattern codes to auditory perception. *J Cogn Neurosci* **20**, 135-152 (2008).

Supplementary Video 1 The video illustrates the spatiotemporal activity patterns evoked in primary auditory cortex by twenty different consonant sounds (**Supplementary Fig. 1**). The color of each polygon indicates the multi-unit activity at each of 63 recording sites in a single rat. Pairs of sounds that evoke similar patterns (i.e. /m/ and /n/) are difficult for rats to discriminate, while pairs that evoke distinct patterns are easy (i.e. /d/ and /b/, **Fig. 4**). The map of characteristic frequency (CF) illustrates the topographic organization of tone frequency tuning. Red indicates instantaneous firing rates above 800 Hz, yellow and orange denote 550–800 Hz, light blue and green denote 100–550 Hz, and dark blue denotes rates below 100 Hz. Activity patterns reflect the average response to twenty repeats. The blue lines under each activity map indicate the average firing rate for all 63 A1 sites. The red lines mark the time at which each spatial activity pattern occurs relative to response onset. The patterns were smoothed using a 3 ms sliding window.

Supplementary Video 2 The video illustrates the spatiotemporal activity patterns evoked in primary auditory cortex by twenty different consonant sounds (**Fig. 1**). The conventions are the same except that the frame rate is faster and the video continues until 250 ms after response onset. Late response peaks were due to the onset of the vowel /a/ (as in 'sad'). The terminal /d/ of each sound was excluded to reduce video size.