Phoneme representation and classification in primary auditory cortex

Nima Mesgarani, Stephen V. David, Jonathan B. Fritz, and Shihab A. Shamma[a)]
Electrical and Computer Engineering & Institute for Systems Research, University of Maryland, College Park, Maryland 20742

(Received 5 June 2007; revised 15 October 2007; accepted 30 October 2007; corrected 20 March 2008)

A controversial issue in neurolinguistics is whether basic neural auditory representations found in many animals can account for human perception of speech. This question was addressed by examining how a population of neurons in the primary auditory cortex (A1) of the naïve awake ferret encodes phonemes and whether this representation could account for the human ability to discriminate them. When neural responses were characterized and ordered by spectral tuning and dynamics, perceptually significant features including formant patterns in vowels and place and manner of articulation in consonants, were readily visualized by activity in distinct neural subpopulations. Furthermore, these responses faithfully encoded the similarity between the acoustic features of these phonemes. A simple classifier trained on the neural representation was able to simulate human phoneme confusion when tested with novel exemplars. These results suggest that A1 responses are sufficiently rich to encode and discriminate phoneme classes and that humans and animals may build upon the same general acoustic representations to learn boundaries for categorical and robust sound classification. © *2008 Acoustical Society of America.*

[DOI: 10.1121/1.2816572]

PACS number(s): 43.64.Sj, 43.71.Qr [WPS] Pages: 899–909

I. INTRODUCTION

Humans reliably identify many phonemes and discriminate them categorically, despite considerable natural variability across speakers and distortions in noisy and reverberant environments that limit the performance of even the best speech recognition algorithms.[1,2] Trained animals have also been shown to discriminate phoneme pairs categorically and to generalize to novel situations.[3–10] The neurophysiological basis of these perceptual abilities in humans and animals remains uncertain. However, there is experimental evidence for cortical encoding of phonetic acoustic features regarded as critical for distinguishing classes of consonant-vowel (CV) syllables, such as voice-onset time.[11–14] Key questions include the nature and location of the neural representations of different phonemes and, more specifically, whether the neural responses of the primary auditory cortex (A1) are sufficiently rich to support the phonetic discriminations observed in humans and animals.

The general issue of the neural representation of complex patterns is common to all neuroscience and has been investigated in many sensory modalities. In the visual system, recent studies have shown that responses of approximately 100 cells in the inferior temporal cortex are sufficient to account for the robust identification and categorization of several object categories.[15] In the auditory system, a recent study has shown that neurometric functions derived from single unit recordings in the ferret primary auditory cortex closely parallel human psychometric functions for complex sound discrimination.[16] An important aspect of our approach in the present study is the inclusion of temporal features of the response in the analysis. This is crucial because phonemes are *spectro-temporal* patterns, and hence analyzing their neural representation at a single cell or ensemble level requires consideration of the interactions between the stimuli and the intrinsic dynamics of individual neurons.

In the present study, we recorded responses of A1 neurons to a large number of American English phonemes in a variety of phonemic contexts and derived from many speakers. Our results demonstrate that (I) time-varying responses from a relatively small population of primary auditory cortical neurons (<100) can account for distinctive aspects of phoneme identification observed in humans,[17] and that (II) well known acoustic features of phonemes are indeed explicitly encoded in the population responses in A1.

The analysis of the categorical representation of phonemes across a neuronal population presented in this paper remains largely model-independent in that only relatively raw response measures (e.g., peri-stimulus time histograms, PSTHs) are used in the computations and illustrations. The one key departure from this rule is necessitated by the desire to organize the display of the population responses according to their best frequency, spectral scale, and temporal dynamics. These response properties are quantified using the measured spectro-temporal receptive field (STRF) model of the neurons.[18,19]

[a)] Author to whom correspondence should be addressed. Electronic mail: sas@isr.umd.edu

II. EXPERIMENTAL PROCEDURES

The protocol for all surgical and experimental procedures was approved by the Institutional Animal Care and Use Committee (IACUC) at the University of Maryland and consistent with NIH Guidelines.

A. Surgery

Four young adult, female ferrets were used in the neurophysiological recordings reported here. To secure stability of the recordings, a stainless steel head post was surgically implanted on the skull. During implant surgery, the ferrets were anesthetized with Nembutal (40 mg/kg) and Halothane (1–2%). Using sterile procedures, the skull was exposed and a headpost was mounted using dental cement, leaving clear access to primary auditory cortex in both hemispheres. Antibiotics and analgesics were administered as needed.

B. Neurophysiological recording

Experiments were conducted with awake head-restrained ferrets. The animals were habituated to this setup over a period of several weeks, and usually remained relaxed and relatively motionless throughout recording sessions that may last 2–4 h. Recordings were conducted in a double-walled acoustic chamber. Small craniotomies (1–2 mm in diameter) were made over the primary auditory cortex before recording sessions. Physiological recordings were made using tungsten microelectrodes (4–8 MΩ). Electrical signals were amplified and stored using an integrated data acquisition system (Alpha Omega). Spike sorting of the raw neural traces was done offline using a custom principal component analysis (PCA) clustering algorithm. Our requirements for single unit isolation of stable waveforms included (1) that the waveform and spike rate remained stable throughout the recording, and (2) that the inter-spike interval for each neuron was distributed exponentially with a minimum latency of 2 ms.

C. Speech stimuli and data analysis

Stimuli were phonetically transcribed continuous speech from the TIMIT database.[20] Thirty different sentences (3 s, 16 kHz sampling) spoken by different speakers (15 male and 15 female) were used to sample a variety of speakers and contexts. A large stimulus set was used, that extended the original set from 30 to 90 sentences, and also increased speaker diversity to 45 male and 45 female speakers. In all recordings, each sentence was presented five times.

D. Mean phoneme representation

The TIMIT phonetic transcriptions were used to align the responses of each neuron to all the instances of a given phoneme and then averaged to compute the peri-stimulus time histogram (PSTH) response to that phoneme, as illustrated in Fig. 1(A) (10 ms time bins). We did not attempt to compensate for the relatively short latency of neural responses in the ferret, since this was roughly constant and consistent for all A1 neurons (15–20 ms). We also computed the auditory spectrogram of each phoneme using the following procedure: Let $S(t,f)$ be the auditory spectrogram of the speech stimulus computed using a model of cochlear frequency analysis,[21] and let $r(t)$ be the corresponding neural response. For phoneme k, which occurs at times t_{k_1}, $t_{k_2},\ldots,t_{k_n}$, the average spectrogram is

$$\hat{S}_k(t,f) = \frac{1}{n}\sum_{i=1}^{n} S(t_{k_i}+t,f) \tag{1}$$

and the average neural response is

$$\hat{r}_k(t) = \frac{1}{n}\sum_{i=1}^{n} r(t_{k_i}+t). \tag{2}$$

The total number of occurrences of each phoneme, n, ranged from 7 (e.g. /g/) to 72 (e.g., /ɨ/) in the chosen sentences.

E. Measurement of neuronal tuning properties

We characterized each neuron by its spectro-temporal receptive field (STRF), estimated by normalized reverse correlation of the neuron's response to the auditory spectrogram of the speech stimulus.[18] Although methods such as normalized reverse correlation can produce unbiased STRF estimates in theory, practical implementation requires some form of regularization to prevent overfitting to noise along the low-variance dimensions. This in effect imposes a smoothness constraint on the STRF. The regression parameters were adjusted using a jackknife validation set to maximize the correlation between actual and predicted responses.[22] Figure 1(B) illustrates the STRF of one such neuron. We measured several tuning properties from each STRF: Best frequency (BF) was defined as the largest positive peak value of the STRF along its frequency dimension. The STRF scale and rate were estimated from the two-dimensional (2D) modulation transfer function (MTF) (Fig. 1(B)). The MTF is the 2D Fourier transform of the STRF that is then collapsed along its temporal or spectral dimensions (known also as the *rate* and *scale*) to obtain the purely *spectral* (*sMTF*) or *temporal* (*tMTF*) modulation transfer functions (Fig. 1(B)). The *best scale* (related to the inverse bandwidth) of an STRF is defined as the centroid of the sMTF (in "cycles/octave"), whereas "speed" or *best rate* of the STRF is defined as the centroid of the tMTF (in Hz), as illustrated in Fig. 1(B). To display the neural *population responses* for each phoneme, we generated two-dimensional "topographic" plots in which each row contained the average PSTH response of one neuron, sorted according to neural BF, scale or rate. The distribution of these three tuning properties in our sample was fairly broad, covering most BFs, best scales, and best rates (see Appendix). However, because the parameters were not distributed exactly uniformly, we interpolated the vertical axis of the smoothed PSTH (2D disk filter: 60 ms * 6 neurons) to have uniform spacing and then smoothed the PSTH display with the same 2D filter. We characterized each phoneme according to the *locus* of maximal response within the neural population along the BF, scale and rate dimensions. For example, to find the locus along the BF dimension, we determined the position of the maximum PSTH responses over time for neurons ordered along the BF axis. The same

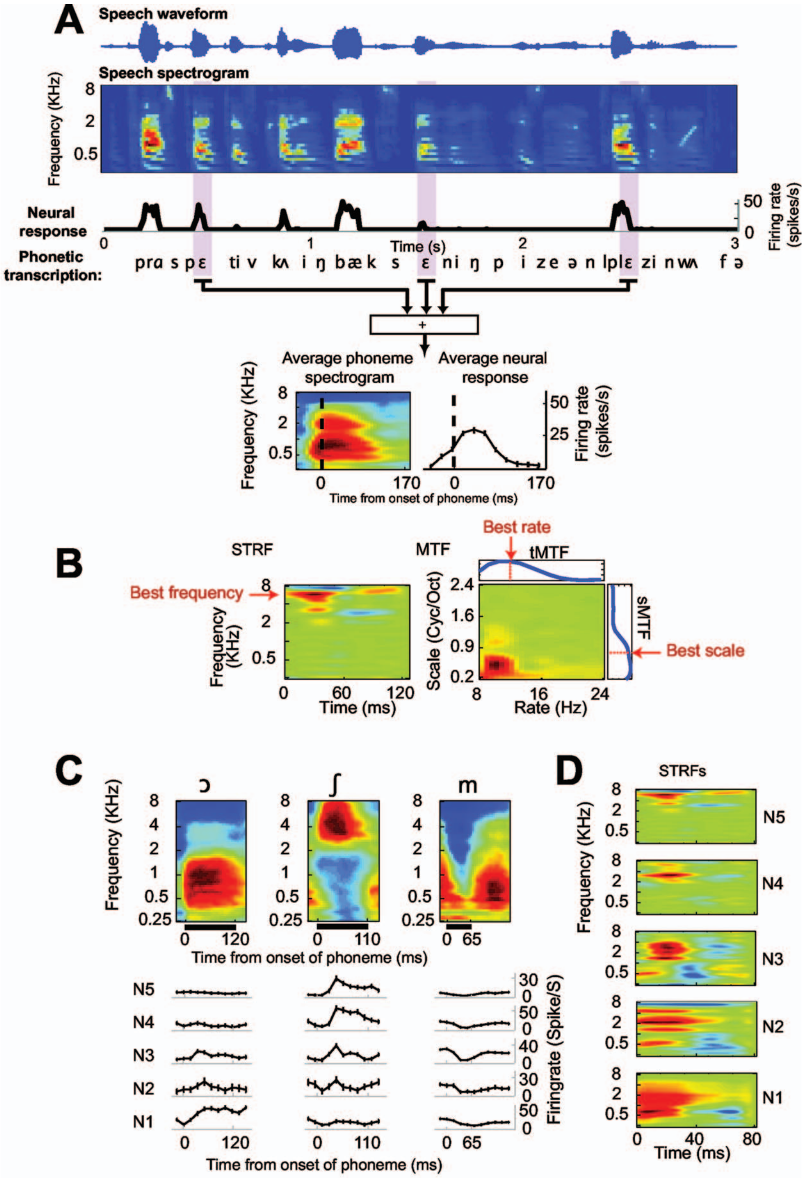

FIG. 1. Neuronal responses to phonemes in continuous speech. (A) The spectrograms of all /ɛ/ vowel exemplars were extracted and averaged to obtain one grand average auditory spectrogram (bottom left). In this and following average spectrogram plots, red areas indicate regions of higher than average energy and blue regions indicate weaker than average energy. The corresponding PSTH response to /ɛ/ was computed by averaging neural spike rates over the same time windows (bottom right). (B) The spectro-temporal receptive field (STRF) of a neuron as measured by normalized reverse correlation. Red areas indicate stimulus frequencies and time lags correlated with an increased response, and blue areas indicate stimulus features correlated with a decreased response. The neuron's BF was defined to be the excitatory peak of the STRF (red arrow). The modulation transfer function (MTF) is computed by taking the absolute value of the 2D Fourier transform of the STRF. We then collapse along the temporal or spectral dimensions (known also as the *rate* and *scale*) to obtain the purely *spectral (sMTF)* or *temporal (tMTF)* modulation transfer functions. The *best scale* (proportional to the inverse of bandwidth) of a STRF was defined as the centroid of the sMTF (in "cycles/octave"), whereas "speed" or *best rate* of the STRF is defined as the centroid of the tMTF (in Hz). The choice of *centroid* for best-scale parameter results in a compressed range but it does not affect the ordering of neurons along this dimension. (C) Average auditory spectra of three phonemes (/ɔ/, /ʃ/, /m/). Below each spectrogram is the PSTH response of five example neurons (labeled N1–N5). (D) The STRFs of these neurons indicate a diversity of spectro-temporal tuning properties.

procedure was repeated for PSTHs ordered along the scale and rate axes to obtain the three coordinates of the locus.

F. Phoneme classification and confusions

To examine the separation or overlap among the representations of different phonemes, we trained linear binary classifiers to discriminate each phoneme from all the others based on the neuronal population response. Formally, the neurons project the phoneme acoustic signals into a high dimensional space (i.e., the total number of neurons × the number of samples in each PSTH $=90\times 22$). Because of the different selectivity of each neuron, different phonemes fall in specific subregions of this space.

A linear Support Vector Machine (SVM[35]) was trained to find the optimal hyperplanes for each phoneme, such that the hyperplane has the maximum distance (or "margin") to the closest data points (or "support vectors") in the two classes it separates. Using linear hyperplanes is intuitively appealing because the classifier's output is a weighted sum of the neural responses that can be interpreted easily. The output of each classifier is a scalar value indicating the distance of the data point to the hyperplane. Novel sounds are identified by choosing the classifier that produces the maximum distance to the boundary. We should emphasize that the order of the neural responses is not important in any way for classification.

G. Statistical analysis

The significance of correlations between the pattern of phoneme confusion predicted by the neural classifier and confusion observed for human perception[17] was ascertained by a randomized t test. Random correlations were computed between neural and perceptual confusion matrices after randomly shuffling phoneme identity (20 000 shuffles). The significance of the correlation between the actual confusion matrices was taken as the probability that such a correlation could be produced by the randomly shuffled matrices.

H. Measuring the acoustic distance among phonemes

The average auditory spectrogram of each phoneme was computed as described above.[21] The acoustic similarity between any pair of phonemes was then defined as the Euclidean distance between their average spectrograms.

III. RESULTS

A. Diversity of single-unit responses to phonemes

Physiological responses were recorded from 90 single units in A1 of four ferrets (*Mustela putorius*) during the monaural presentation of continuous speech stimuli (see Fig. 1(A)). The recorded neurons were broadly distributed in their spectral tuning and dynamic response properties as shown by population range of their best frequency (BF), best scale, and best rate (documented in the scatter plots in Fig. 6 in the Appendix). These neural tuning properties are based on measurements of the spectro-temporal receptive fields of the neurons (STRFs) as depicted in Fig. 1(B) and described in detail earlier in Sec. II. Figure 1(C) illustrates the PSTH responses of five single units (N1–N5) to three different phonemes (vowel /ɔ/, fricative /ʃ/ and nasal /m/) whose average auditory spectra are depicted in Fig. 1(C). The spectro-temporal receptive fields (STRFs) of the five selected neurons are shown in Fig. 1(D).

Each phoneme activates these five neurons differentially, depending on the match between the neuron's STRF and the spectro-temporal structure of the stimulus. For instance, the vowel /ɔ/ drives N1 very effectively because of the low BF of the neuron (~700 Hz). By contrast, the fricative /ʃ/ maximally activates N4 and N5, which have the highest BFs (~3 and ~7 kHz, respectively). Finally, the response pattern of the nasal /m/ is unique in that it causes a depression of responses in N2 and N3, reflecting the energy dip midway through the phoneme over all frequencies, but especially in the middle frequencies (~0.5–4 kHz).[23,24] In this manner, each phoneme evokes a unique response pattern across the population of A1 cells that differs from the evoked responses elicited by other phonemes.

B. Population responses to phoneme classes

To appreciate the unique response patterns evoked by different phonemes and, in particular, in order to highlight the acoustic features enhanced in the neural representation, it is best to view the ordered activity of the entire population simultaneously. This ordering depends entirely on the neuronal tuning properties to be emphasized. For instance, inspired by the tonotopic organization of the auditory pathway, the most common way to organize neural PSTHs has been by frequency according to the BF of the units.[25,26] However, unlike the receptive fields of fibers in the auditory nerve, A1 neurons exhibit systematic variations of tuning along multiple feature axes, including bandwidth, asymmetry, and temporal dynamics.[14,27,28]

Here we consider the ordered representation of phoneme responses along BF and two other dimensions derived from the STRF: best scale and best rate (see Sec. II and Fig. 1(B)). Best scale is *inversely proportional* to bandwidth and indicates how wide a range of sound frequencies are integrated into the neural response. Best rate indicates the dynamic agility of a neuron's responses and hence reflects the temporal modulation of the stimulus spectrum that best drives the neuron. The coordinates of each cell along these three dimensions can be estimated using a variety of techniques and stimuli. The most common techniques include tuning curves or iso-response functions measured from tones[28] and STRFs measured from ripples.[29] We use the speech-based STRFs to estimate these parameters for each cell.[18]

1. Encoding of vowels

Population responses to 12 American-English vowels are summarized in Fig. 2. Panels in the top row (Fig. 2(A)-I) display the average auditory spectrogram of each vowel computed from all of its samples encountered in the speech database (see Sec. II for details). The vowels are organized according to their articulatory configurations along the Open/Closed and Front/Back axes,[23] as illustrated at the top of Fig. 2: /o/, /ɔ/, /ɑ/, /ʌ/, /æ/, /ɛ/, /e/, /ə/, /i/, /ɪ/, /ɨ/, /ʉ/. The three middle vowels (/ɛ/, /e/, /ə/) are tightly clustered near the midpoint of the Front/Back and Open/Closed axes, and are difficult to order accurately along this one-dimensional representation of the vowels.

The averaged spectra (top row) reveal that Mid/Back vowels (/o/, /ɔ/, /ɑ/, and /ʌ/) have relatively concentrated activity at low to medium frequencies (~0.4–2 KHz), whereas Front vowels sometimes have two peaks spaced over a larger frequency range (~0.3 and ~4 KHz). This is consistent with the known distribution of the three formants (F1, F2, and F3) in these vowels,[23] namely, that they have F1 and F2 that are closely spaced, creating compact single broad peak spectra at intermediate frequencies (reminiscent of the center-of-gravity hypothesis of Chistovich and Lublinskaya[30]). As the vowels become more "Front"ed, the single peak broadens and splits (/æ/ to /ə/). Continuing this trend, Front/Closed vowels (/i/, /ɪ/, /ɨ/, /ʉ/) exhibit relatively narrow and well separated formant peaks with F1 at low and F2 at high frequencies.

These averaged phoneme spectra are broadly reflected in the response distributions ordered along the BF axis; neurons with BFs matching regions of high energy in a phoneme spectrum tend to give strong responses to that phoneme (Fig. 2(A)-II). However, notable differences of unknown significance exist such as the relative weakness of the low BF peaks in /e/ and /ə/, and of the high BF peak in /i/). More striking, however, are the response distributions along the best scale axis, which roughly indicates the *inverse* of the vowels' spectral bandwidths (Fig. 2(A)-III). Here, consistent with the bandwidths of the spectral peaks discussed earlier, Central/Open vowels tend to evoke maximal responses in broadly tuned cells commensurate with their broad spectra (low scales <1 Cyc/Oct) while Closed vowels evoke maximal responses in narrowly tuned cells (scales >1 Cyc/Oct), as indicated by the blue and red boxes in Fig. 2(A)-III, respectively.[31] Response distributions in the best rate panels (Fig. 2(A)-IV) reveal a trend in the dynamics of the vowels as one moves along the Front/Back axis. Specifically, Front

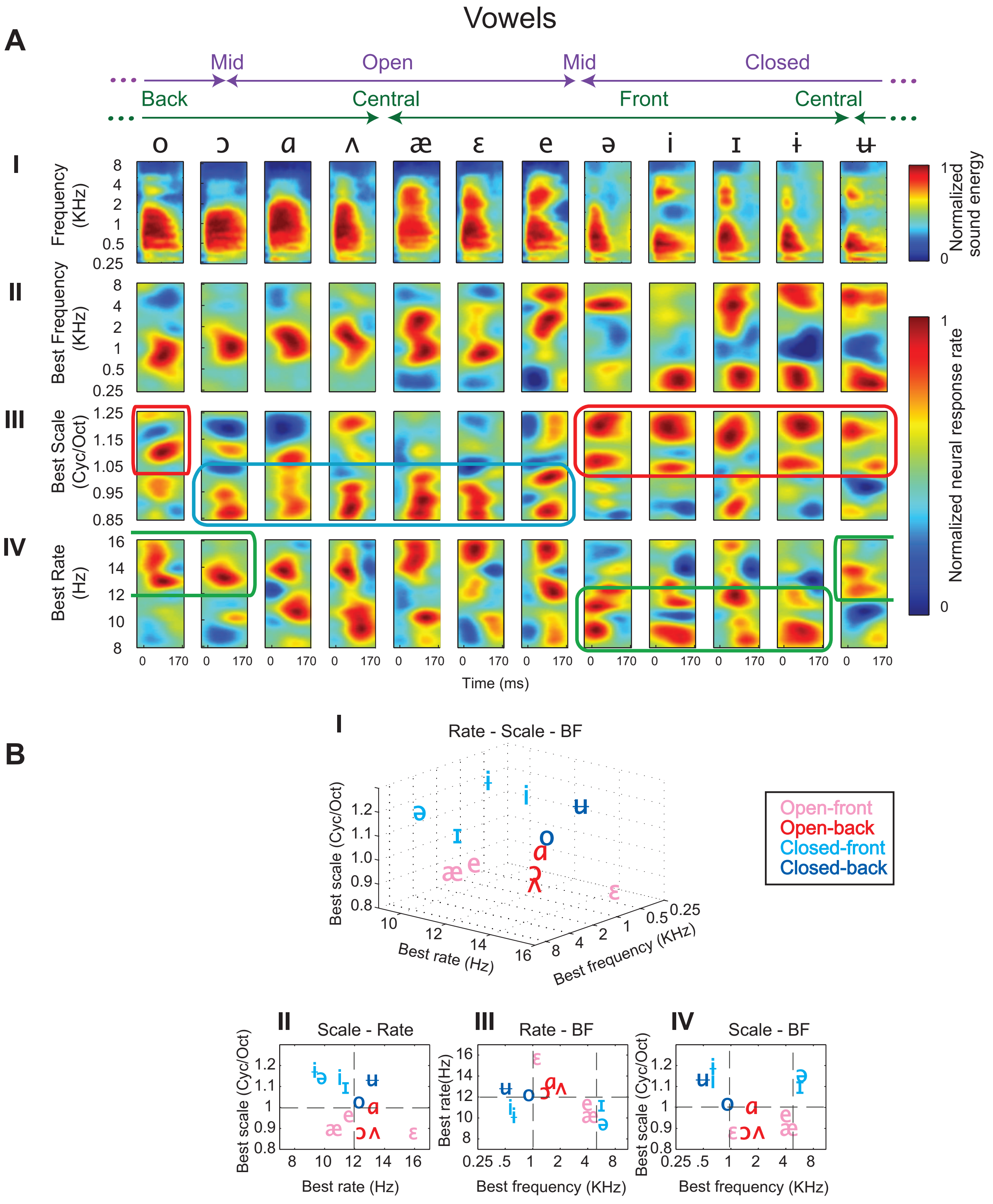

FIG. 2. Population response to vowels. (A) I. Average auditory spectrogram of 12 vowels organized approximately according to their open-closed and front-back articulatory features. The arrows at top indicate the *degree* of these features, with arrow "tips" representing minima (mid or central) and midpoints representing maxima. For example /ʌ/ is maximally open, but is neutral (central) on the front/back axis. Note also that the axes are presumed to loop around the page from right to left (dashed ends joining) creating a circular representation (II, III, IV): Average PSTH responses of 90 neurons to each vowel. Within each heat map, each row indicates the average response of a single neuron to the corresponding phoneme. Red regions indicate strong responses, and blue regions indicate weak responses. The average PSTH responses are sorted by neurons' best frequency (II), best scale (III) and best rate (IV) to emphasize the role of that parameter in the encoding of each vowel. (Details of the analysis and generation of these plots are given in Sec. II). (B) I. Each vowel is plotted at the centroid frequency, rate and scale of its average neuronal population response. The centroid values are calculated from the average PSTH responses sorted by the corresponding parameter (2A). "Open" vowels are shown in red, "Closed" vowels in blue, "Front" vowels with *light* font and "Back" vowels with *dark*. To visualize the contribution of each tuning property to vowel discrimination, the location of each vowel is also shown collapsed in 2–D plots of (II) scale-rate, (III) rate-BF and (IV) scale-BF. All other details of the analysis and generation of these plots are given in Sec. II (Experimental Procedures).

vowels (/ə/, /i/, /ɪ/, /ɨ/) evoke relatively stronger responses in the slower cells (with best rates $<\sim 12$ Hz), as compared to the more Back vowels (/ʉ/, /o/, /ɔ/) as highlighted by the green boxes in Fig. 2(A)-IV. The remaining more Central vowels (/ɑ/, /ʌ/, /æ/, /ɛ/, /e/) exhibit all dynamics. This response pattern may reflect the longer durations required to complete the articulatory excursions toward or away from Closed vowels towards the front of the vocal tract.

Figure 2(B) provides a compact summary of the population response to vowels. Each vowel is placed at the *locus* of maximum response in the neural population along the BF, best scale, and best rate axes. To highlight more clearly which of the three features best segregates them, the 3D display is projected onto each of the three marginal planes (Figs. 2(B)-II and 2(B)-IV). It is readily evident in these displays that the Open and Closed vowels separate along the scale axis above and below 1 Cyc/Oct (horizontal dashed lines in Figs. 2(B)-II and 2(B)-IV). They are also distinguished by BF, with the Open vowels clustering in the range 1.0–4.5 KHz (vertical dashed lines in Fig. 2(B)-III). Finally, the best rate axis segregates the Front/Back vowels (as discussed earlier), with Central and Back vowels located at high rates (>12 Hz), and Front vowels below it. It remains to be confirmed, however, whether these locations, which reflect the vowels' overall spectro-temporal similarity, can explain the perceptual confusion among them[32].

2. Encoding of consonants

Population responses to 15 consonants are shown in Fig. 3 in the same format already described for vowels. Three properties are commonly used to organize and classify consonants: place of articulation, manner of articulation, and voicing.[23,24,33] Here we examined how these three properties are encoded in the responses of the neuron population.

The distributions of the responses to the consonants sorted along the BF axis (Fig. 3(A)-II) approximate the features of their averaged spectra (Fig. 3(A)-I), which in turn are known to be closely related to place of articulation cues. For instance, the difference between the more forward places of constriction for /s/ compared to /ʃ/ is mirrored by the downward shift of the highpass spectral edge. Similarly the high-frequency noise burst at the onset of the forwardly constricted /t/ contrasts with the lower-frequency distribution of the other plosives (/p/ and /k/). However, there are also some notable differences in detail between the two sets of plots. There is generally a slight delay of about 20 ms in the neural responses relative to the spectrograms (presumably due to the latency of cortical responses). In addition, however, there are substantial differences between the responses and spectrograms in certain phonemes. For example, high BF responses to /f/ in Fig. 3(A)-II are strong despite their relative weakness in the spectrograms. Similarly, the low BF responses to /v/ are not consistent with the spectrogram. In other consonants, there are differences in the "timing" of certain frequency regions such as the rapid onset of high frequencies in the spectrogram of /t/ relative to its more delayed response, or in the continuity of the spectral regions in /ʃ/, /d/ and /ŋ/. The origin of all these differences is unclear and may reflect the nonlinearity of neural responses and/or our limited sampling of the neural population (90 neurons).

Response distributions along the best scale and best rate axes (Figs. 3(A)-III and 3(A)-IV) capture well the essential *manner of articulation* cues that supply the information necessary to discriminate plosives, fricatives, and nasals in continuous speech. For example, the broad distinction between "plosives" and "continuants" (e.g. /p/, /t/, /k/, /b/, /d/, /g/ versus /s/, /ʃ/, /z/, /n/, /m/, /ŋ/) is evident in the distribution of responses along the scale and rate axes (Figs. 3(A)-III and 3(A)-IV). Thus, plosives with their sudden and spectrally broad onsets display relatively strong activation in broadly tuned (low scales <1.1 cyc/oct) and fast (rates >12 Hz) cells (regions outlined in red in Figs. 3(A)-III and 3(A)-IV) compared to the more suppressed responses to longer duration unvoiced fricatives and nasals (outlined in blue in Fig. 3(A)-IV). Note also the brief suppressed response preceding the onset of all plosives due to the (silent) voice-onset-time (VOT) in all panels within the red box (Figs. 3(A)-III and 3(A)-IV).

Finally, the third cue of voicing is associated with the harmonic structure of voiced spectra near the low to mid-frequency range (0.2–1 kHz), and to a lesser extent the weak energy at low BFs near the fundamental of the voicing. Only this latter cue seems to distinguish consistently the voiced (/b/, /d/, /g/, /v/, /ð/, /z/, /m/, /n/, /ŋ/) from unvoiced (/p/, /t/, /k/, /f/, /s/, /ʃ/) consonants in our data as indicated by the green outlined region of Fig. 3(A)-II. However, such a strong low BF response as an indicator of "voicing" is missing in many of the vowel responses discussed earlier (e.g., the Open/Back vowels in Fig. 3(A)-II). Instead, its presence seems to correlate with the low F1 of the Closed vowels there. Therefore, our data suggest that the low-frequency voicing is reliably represented only in consonant responses, and perhaps in vowels where the F1 is low enough to amplify it.[34] However, there may well be a different and separate representation of voicing in the auditory cortex, for example, in terms of the pitch it evokes, or the harmonicity of its spectral components.[35]

Figure 3(B) illustrates the locus of the population response to each consonant in a plot of best frequency, best rate and best scale similar to that used with vowels earlier. The lower panels of Fig. 3(B) are projections of the three-dimensional (3D) plot onto its three marginal planes. Members of the three groups of consonants—plosives (red), fricatives (blue), and nasals (green)—are loosely grouped together in this parameter space. For instance, plosives tend to drive broadly tuned (scale <0.9 Cyc/Oct) and fast (rates >12 Hz) cells (Figs. 3(B)-II). Rate is also a distinguishing feature between plosives on the one hand, and nasals and (most) fricatives on the other (above and below 12 Hz, respectively). Similarly, phoneme groups roughly segregate along the BF axis, with unvoiced fricatives occupying the highest frequencies (>4 kHz), unvoiced plosives falling between 2 and 4 kHz, and other voiced phonemes falling below 2 kHz (Figs. 3(A)-III and 3(A)-IV). As with vowels, this plot of the neural loci of consonants reveals the

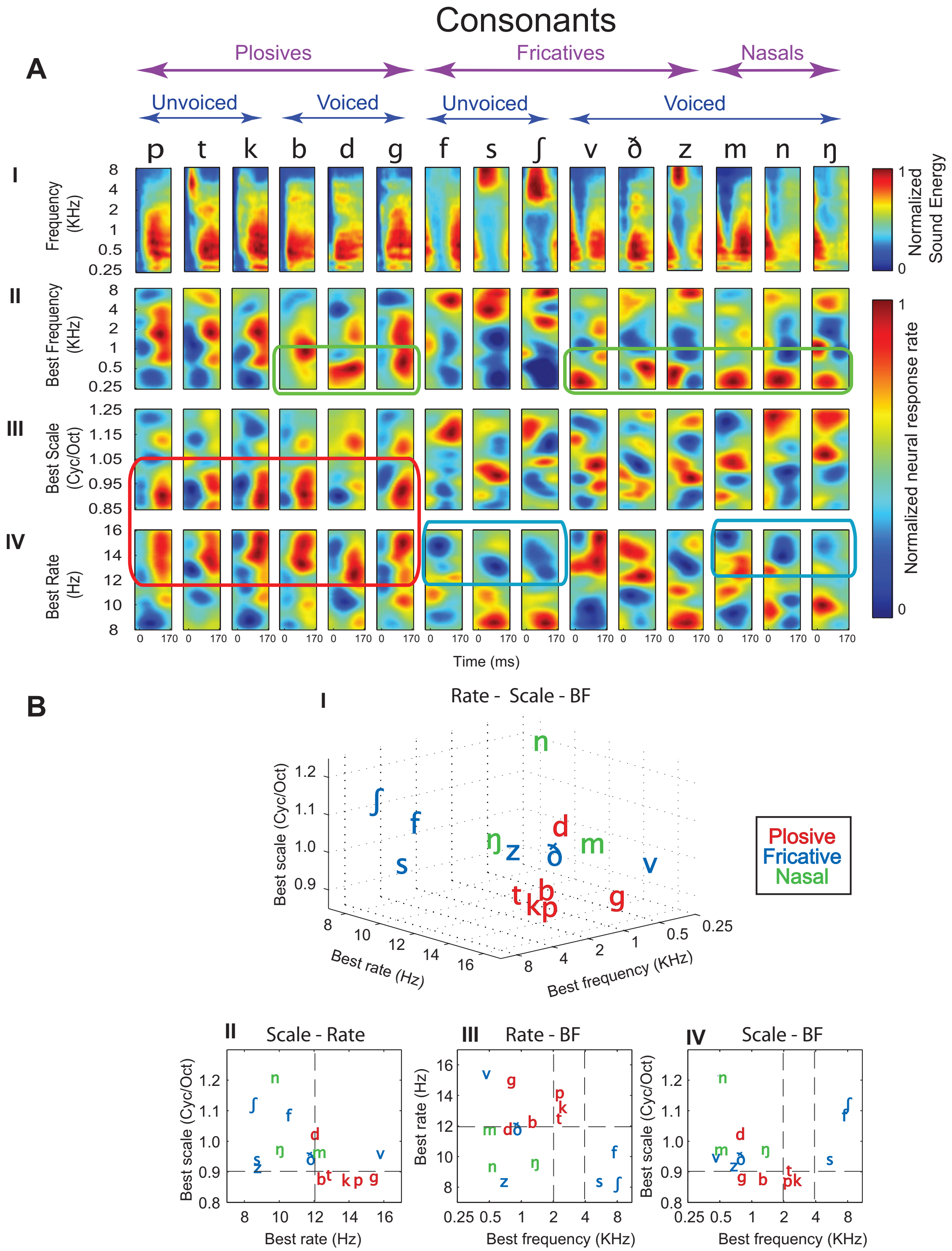

FIG. 3. Population response to consonants. (A) I. Average spectrogram of 15 consonants phonemes grouped as six plosives, six fricatives and three nasals. Each of the plosive and fricative-groups contains three voiced and three unvoiced phonemes (see arrows at top). (II, III, IV) Average PSTH responses of the neural population to each consonant, plotted as in Fig. 2(A). The average PSTH responses are sorted by neurons' best frequency (III), best scale (II) and best rate (IV) to emphasize the role of that parameter in the encoding of consonants. (All other details of the analysis and generation of these plots are given in Sec. II). (B) Each consonant is placed at the centroid frequency, rate and scale of its neuronal population response, measured from the corresponding PSTH responses (A). Plosive phonemes are plotted in red, fricatives in blue and nasals in green. The locus of each consonant is also shown collapsed in 2D plots of (II) scale-rate, (III) rate-BF and (IV) scale-BF. All other details of the analysis and generation of these plots are given in Sec. II (Experimental Procedures).

relative distances among them and perhaps explains the pattern of perceptual confusion observed between them, as we shall elaborate next.

C. Phoneme confusions

Average phoneme responses give useful insights into the mean representation of each phoneme, but they fail to indicate how well the neural population can discriminate pho-

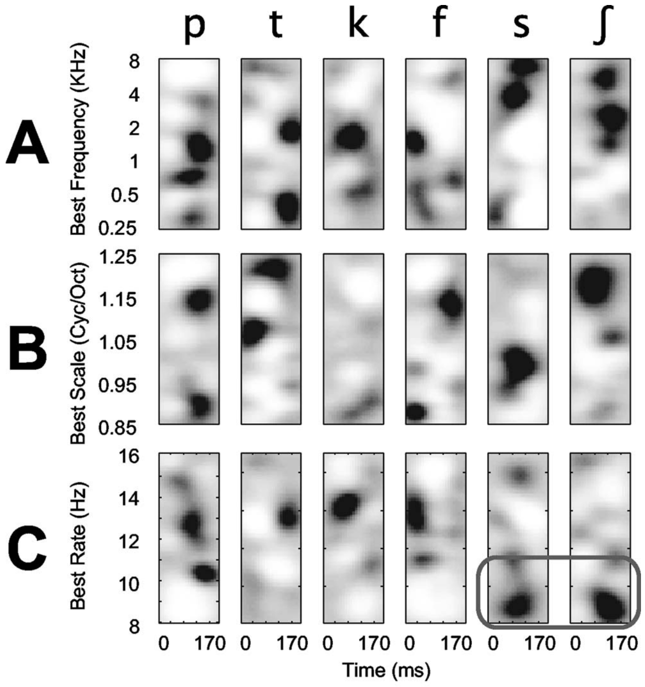

FIG. 4. Phoneme classification based on the population response. Classification masks for three unvoiced plosives (/p/, /t/, /k/) and three unvoiced fricatives (/f/, /s/, /ʃ/) sorted by neurons' best frequency (A), best scale (B) and best rate (C). Gray scale indicates the importance of the presence (*black* regions) or absence (*white* regions) of neural response for the classification of that phoneme. The output of each phoneme classifier is a scalar, computed as the sum of the population PSTH multiplied by the mask. Thus the order of the mask/PSTH is irrelevant to the output of the classifier.

nemes, given the natural acoustic variability among samples of the same phoneme during continuous speech. To delineate perceptual boundaries implied by the responses to the phonemes, we trained a linear classifier for each phoneme to separate it from all others, based on the PSTHs of the neural population.[36] To determine the identity of a novel phoneme, the population response was applied to all the classifiers, each computing the likelihood of its designated phoneme. The classifier indicating the maximum likelihood was taken as the identity of the input phoneme. To train and test the classifiers, we divided the speech data into 100 train and test subsets. In each subset, 90% of the data was randomly chosen for training and the remaining 10% was used for testing. The classification accuracy and the confusion matrices reported here are the average results from these 100 subsets.

Once trained, each linear classifier can be viewed as a mask that selects, by multiplication with the population response, the neurons and response latencies that most effectively distinguish the associated phoneme from all others. Figure 4 displays the masks computed for the unvoiced consonants /p/, /t/, /k/, /f/, /s/, /ʃ/. The masks are ordered in the same way as the PSTHs in Fig. 3(A) (i.e., by BF, best scale, and best rate). In the masks, black regions signify neurons and response latencies for which a strong response provides evidence for presence of the phoneme, and white regions signify strong responses that provide evidence against that phoneme. The masks in Fig. 4 differ from the mean neural responses in Fig. 3(A) in that they emphasize the *unique* features of each phoneme. For example, the mean responses

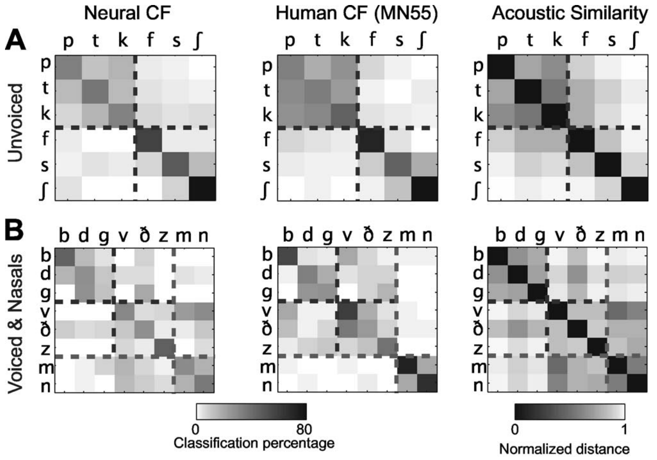

FIG. 5. Neural and human phoneme confusions, and phonemes acoustic similarity. Consonant confusion matrices from neural phoneme classifiers (left panels) and human psychoacoustic studies (Ref. 17) (middle panels). Gray scale indicates the probability of reporting a particular phoneme (column) for an input phoneme (row). (Right panels) The acoustic similarity between phoneme pairs defined as the Euclidian distance between their average auditory spectrograms. (A) Confusion matrices and phonemic distances for unvoiced consonants. Dashed lines separate the plosives /p/, /t/, /k/ from fricatives /f/, /s/, /ʃ/. (B) Confusion matrices and phonemic distances for voiced consonants. Dashed lines separate the plosives /b/, /d/, /g/ from fricatives /v/, /ð/, /z/ and the nasal consonants /m/ and /n/ from the rest.

to /ʃ/ (Fig. 3(A)-II) indicate strong responses in high and medium BF neurons, but in the masks the mid-BF neurons (2 kHz) are given higher weights. This differential weighting reflects the fact that both /ʃ/ and /s/ evoke strong responses from high BF neurons, but only /ʃ/ evokes responses from the mid-BF neurons. Similarly, the /p/, /t/, /k/ masks reflect only the features that distinguish these phonemes from each other. The BF masks (Fig. 4(A)) emphasize the low (750 Hz), high (>2 kHz), and medium (0.3–1.5 kHz) spectral regions for the /p/, /t/, /k/ bursts, respectively. Note also how the rate masks (Fig. 4(C)) distinguish plosives /p/, /t/, /k/ from the long fricatives /s/, /ʃ/ by enhancing the regions outlined in the rectangle, namely the slow rates of the fricatives (<11 Hz) relative to the faster rates of the plosives. It should be noted that the classifier performance does not depend in any way on the *order* of the neural responses, which is solely used for analysis and display purposes.

The extent to which the neural phoneme representations can account for the perception of *individual* phoneme exemplars can be assessed by studying the pattern of pair-wise confusions by the classifier. Figure 5(A) shows the confusion matrix measured from classifications of the neural data. Labels along each row indicate the phoneme presented, and columns report the probability of the phoneme output by the classifier.[17,37] The classifier was trained on two sets of data. In a small set of 20 neurons, we succeeded in measuring responses to 330 s of speech (90 sentences) to be used in the training; these are shown in Fig. 5. In a larger set, training was based on responses from all sampled neurons in which at least 90 s of speech stimuli (30 sentences) were presented; these results are shown in Fig. 7 of the Appendix. In an ideal case in which all phonemes were accurately identifiable, we would expect to see a diagonal confusion matrix. Off-diagonal values represent misidentification. The phonemes are arranged based on voiced-unvoiced and plosive, fricative, nasal consonant categories to facilitate comparison with a previous study of human perception[17,37] (replicated in Fig. 5(B)). The dashed boxes delineate the three major phoneme categories: plosives, fricatives, and nasals. In both neural and perceptual data, phonemes within each category—plosives (/p/, /t/, /k/), fricatives (/f/, /s/, /ʃ/), and nasals (/m/, /n/)—tend to be more confusable within the group than across categories. The correlation coefficient between the complete neural and perceptual matrices is 0.78 ($p = 0.0002$, randomized t test). Ignoring the confusions between voiced and unvoiced consonants improves the similarity to 0.86, with a correlation of 0.95 for only the unvoiced consonants and 0.71 for their voiced counterparts. At least some of the difference between confusion matrices reflects noise due to limited sampling of neural responses, and/or limited data for training the phoneme classifiers. For example, when we computed the same confusion matrix for the entire population of 90 neurons (trained only on 90 s of speech), the correlation between neural and human confusion matrices fell to 0.70 ($p = 0.001$), a change that may reflect the added dimensions and free parameters as new neurons are included in the analysis, while the amount of training data decreases at the same time. (Appendix; Fig. 7).

Alternatively, we explored the sensitivity of the classification in Fig. 5 to the number of neurons included (using the same training material). As expected, the results indicate that percentage of correct classification (averaged across all consonant phonemes) improves as the number of randomly selected neurons is increased (Appendix; Fig. 8). More detailed exploration of this issue should take into account the differential contribution of specific neurons to different phonemes, e.g., high BF neurons to the classification of /s/ and /ʃ/.

Finally, we also explored the extent to which both the neural and human confusion matrices are a reflection of the acoustic similarity (or "distances") among the phonemes at the level of the auditory spectrograms (see Sec. II). Figure 5 illustrates that such a phoneme "similarity matrix" fundamentally resembles the human and neural confusion matrices (with correlation coefficients of 0.66 and 0.93, respectively). In fact, the neural matrix encodes remarkably well the details of the phoneme acoustic similarity, such as the confusions between /v/ and the nasals /m/, /n/, and also between /ð/ and the voiced consonants /b/, /d/, /g/.

IV. DISCUSSION

Neuronal responses to continuous speech in the primary auditory cortex of the naive ferret reveal an explicit multidimensional representation that is sufficiently rich to support the discrimination of many American English phonemes. This representation is made possible by the wide range of spectro-temporal tuning in A1 to stimulus frequency, scale and rate. The great advantage of such diversity is that there is always a unique subpopulation of neurons that responds well to the distinctive acoustic features of a given phoneme and hence encodes that phoneme in a high-dimensional space.

As an example, consider the perception of the plosive consonant /k/ in a CV syllable, which is identified by a conjunction of several acoustic features: an initial silent voice-onset time (VOT), an onset burst of spectrally broad noise, and the direction of the following formant transitions.[23] Each of these features can be encoded in the cortical responses along different dimensions. Thus, neurons selective for broad spectra respond selectively to the noise burst. Rapid neurons respond well following the VOT, whereas directional neurons selectively encode the vowel formant transitions. In this manner, /k/ is encoded *robustly* by a rich pattern of activation that varies in time across the neural population. This neuronal activation pattern constitutes the phoneme representation in A1 and presumably forms the input to a set of neural "phoneme classifiers" in higher auditory areas. If one acoustic feature is distorted or absent, the pattern along the other dimensions (and hence the percept) remains stable.

We have focused here on describing a few prominent features of the response distributions that correspond to well-known distinctive acoustic features of the consonants considered.[24] There are clearly many other aspects and more details of the responses that reflect intricate articulatory gestures, contextual effects, or speaker-dependent variability that can only be reliably considered with a much larger sample of responses. One example is the distribution of the *directionality index* of the responses in the neighborhood of a consonant,[38] an attribute that would indicate whether the formants are upward or downward sweeping, or if they are converging towards or diverging away from a locus frequency.

Humans confuse the phonemes of their native tongue when placed in unusual or noisy contexts. Typically, phonemes that share some acoustic features tend to be more confusable than those that do not. This was confirmed by the similarity we found between the acoustic distance and the human confusion matrices. Similarly, since A1 responses in our naive ferrets also preserve the relative acoustic distances between the phonemes (as they would presumably for other complex sounds), we are led to the conjecture that human phoneme perception can (in principle) be explained in large measure by basic auditory representations such as the auditory spectrogram and the cortical spectro-temporal analysis common to many mammalian (and also avian) species.[6,9,10,39,40]

The representation of phonemic features across a population of filters tuned to BF, scale and rate suggests a strategy for improved speech recognition systems, and further study may reveal additional strategies for speech processing. However, many questions about the neural representation of phonemes still remain unanswered; for example, how can one extrapolate from such neurophysiological findings to the human perceptual ability to perceive phonemes categorically (also found in monkeys,[11] cats,[8] chinchillas,[3] birds[9] and rats,[41]), and to shift categorical boundaries arbitrarily between phoneme pairs?

While the human ability to discriminate native phonemes is the result of many years of training, naïve ferrets lack such a history. Hence ferret perception of clean phonemes may be comparable to humans perception of noisy phonemes. In both cases, confusion patterns would reflect the acoustic distances between the phonemes. However, if ferrets were trained to actively discriminate phonemes, it is

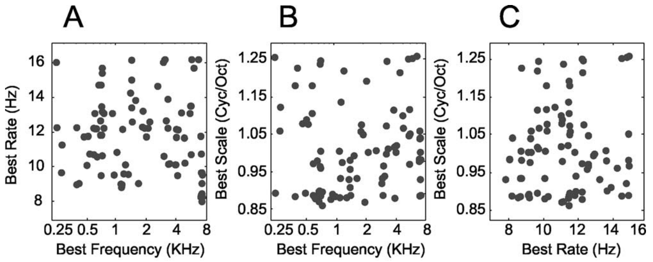

FIG. 6. Joint distribution of neural parameters. Joint distributions of best frequency, best rate (A), best frequency, best scale (B) and best rate, best scale (C) of 90 neurons.

likely that dimensions useful for this specific discrimination would be emphasized, creating the heightened sensitivity necessary to perform the task. This is presumably what happens in humans as they learn the phonemes of a given language, and what the classifier essentially simulates in our analysis when it learns the masks and boundaries that enable robust phoneme discriminations. Therefore, from a neural perspective, one may view the masks as either a subsequent layer of synaptic weights *or* as pattern of behaviorally driven plasticity of A1 receptive fields—the end result of perceptual learning in which neurons adapt their tuning along the dimensions appropriate for the phoneme discrimination task. This same general principle would apply to the discrimination between members of any set of complex sound, using frequency, rate and scale as well as additional cortical response dimensions, such as pitch, spatial location, and loudness.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (Grant Nos. R01DC005779 and F32DC008453), the Air Force Office of Scientific Research and the Southwest Research Institute.

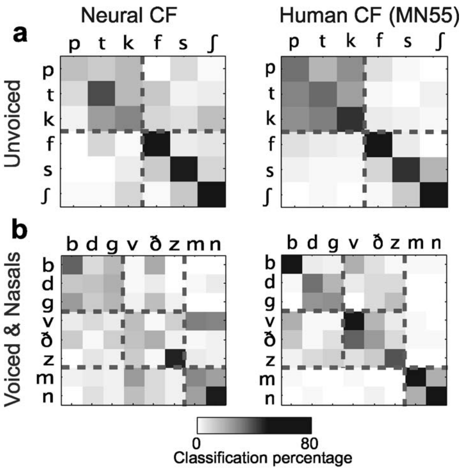

FIG. 7. Phoneme confusions from 90 neurons. (Left column) Consonant confusion matrices from neural phoneme classifiers using entire population of 90 neurons and 90 s of speech. (Right column) human psychoacoustic studies. Gray scale indicates the probability of reporting a particular phoneme (column) for an input phoneme (row). (a) Confusion matrices for unvoiced consonants. Dashed lines separate the plosives /p/, /t/, /k/ from fricatives /f/, /s/, /ʃ/. (b) Confusion matrices for voiced consonants. Dashed lines separate the plosives /b/, /d/, /g/ from fricatives /v/, /ð/, /z/ and the nasal consonants /m/ and /n/ from the rest.

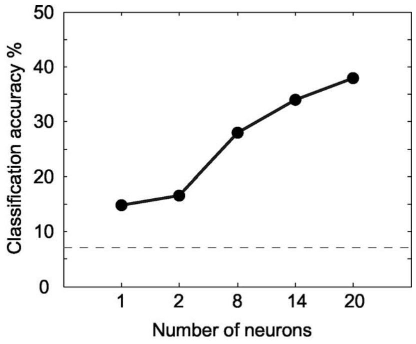

FIG. 8. Dependence of phoneme classification accuracy to the number of neurons. Classification accuracy as a function of the number of neurons used by the classifier. The dashed line indicates chance performance (7% for 14 phonemes) (see Sec. II for details).

APPENDIX

Here we provide additional information regarding: (A) uniformity of the sampling of the neural parameters; (B) the phoneme confusions from an SVM recognizer using a larger number of neurons, but with significantly fewer speech responses on which to train the classifier; (C) an exploration of the recognition accuracy with fewer numbers of neurons.

1. Joint distribution of neural parameters

To ensure that the response patterns in Figs. 2(A) and 3(A) are representative of the neural population in the cortex, we examined the uniformity of the coverage of the parameters of neural STRFs in our sample of 90 neurons. Specifically, the joint distributions of the different neural receptive field parameters (best frequency, best scale, and best rate) are shown in the three panels of Fig. 6, revealing fairly uniform coverage over all frequencies, bandwidths, and different dynamics (see Sec. II for further details).

2. Phoneme confusions from 90 neurons

Phoneme confusions derived from responses of the entire population of 90 neurons, but using only 90 s of speech, are displayed in Fig. 7. The correlation coefficient between the neural and human phoneme confusion (0.70; $p=0.001$) is still reasonable but is significantly less than that of the patterns in Fig. 8 (see method for more details).

3. Dependence of phoneme classification accuracy to the number of neurons

The number of neurons is a crucial variable in determining the accuracy of the phoneme classification as illustrated in the results of Fig. 8. Here the classification accuracy was computed as a function of the number of neurons used in training the classifier. For each condition, 100 random subsets of neurons were taken and the classification accuracy was averaged over all subsets. Note that the accuracy based on the 20 neurons in this plot is still only at 37% (7% is chance performance). Presumably, adding more neurons increases the performance.

[1] R. P. Lippmann, "Speech recognition by machines and humans," Speech Commun. **22**, 1–15 (1997).

[2] S. Greenberg, W. Ainsworth, A. N. Popper, and R. R. Fay, *Speech Processing in the Auditory System* (Springer-Verlag, New York, 2004), Vol. 18.

[3] P. K. Kuhl and J. D. Miller, "Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants," Science **190**, 69–72 (1975).

[4] P. K. Kuhl and D. M. Padden, "Enhanced discriminability at the phonetic boundaries for the place feature in macaques," J. Acoust. Soc. Am. **73**(3), 1003–1010 (1983).

[5] P. K. Kuhl and D. M. Padden, "Enhanced discriminability at the phonetic boundaries for the place feature in macaques," J. Acoust. Soc. Am. **73**(3), 1003–1010 (1983).

[6] K. R. Kluender, A. J. Lotto, L. L. Holt, and S. L. Bloedel, "Role of experience for language-specific functional mappings of vowel sounds," J. Acoust. Soc. Am. **104**(6), 3568–3582 (1998).

[7] F. Pons, "The effects of distributional learning on rats' sensitivity to phonetic information," J. Exp. Psychol. Anim. Behav. Process **32**(1), 97–101 (2006).

[8] R. D. Hienz, C. M. Aleszczyk, and B. J. May, "Vowel discrimination in cats: Acquisition, effects of stimulus level, and performance in noise," J. Acoust. Soc. Am. **99**(6), 3656–3668 (1996).

[9] M. L. Dent, E. F. Brittan-Powell, R. J. Dooling, and A. Pierce, "Perception of synthetic /ba/-/wa/ speech continuum by budgerigars (Melopsittacus undulatus)," J. Acoust. Soc. Am. **102**(3), 1891–1897 (1997).

[10] A. J. Lotto, K. R. Kluender, and L. L. Holt, "Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica)," J. Acoust. Soc. Am. **102**(2 Pt 1), 1134–1140 (1997).

[11] M. Steinschneider, Y. I. Fishman, and J. C. Arezzo, "Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey," J. Acoust. Soc. Am. **114**(1), 307–321 (2003).

[12] M. Steinschneider, D. Reser, C. E. Schroeder, and J. C. Arezzo, "Tonotopic organization of responses reflecting stop consonant place of articulation in primary cortex (A1) of the monkey," Brain Res. **674**, 147–152 (1995).

[13] M. Steinschneider, I. O. Volkov, Y. I. Fishman, H. Oya, J. C. Arezzo, and M. A. Howard, "Intracortical responses in human and monkey auditory cortex support a temporal processing mechanism for encoding of the voice onset time phonetic parameter," Cereb. Cortex **15**, 170–186 (2005).

[14] J. J. Eggermont and C. W. Ponton, "The neurophysiology of auditory perception: From single units to evoked potentials," Audiol. Neuro-Otol. **7**(2), 71–99 (2002).

[15] C. P. Hung, G. K. Kreiman, T. Poggio, and J. J. DiCarlo, "Fast readout of object identity from macaque inferior temporal cortex," Science **310**, 863–866 (2005).

[16] K. Walker, B. Ahmed, and J. W. Schnupp, "Linking cortical spike pattern codes to auditory perception," J. Cogn Neurosci., Oct 5 (Epub) (2007).

[17] G. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352 (1955).

[18] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant, "Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli," Network **12**(3), 289–316 (2001).

[19] D. J. Klein, J. Z. Simon, D. A. Depireux, and S. A. Shamma, "Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex," J. Comput. Neurosci. **20**(2), 111–136 (2006).

[20] S. Seneft and V. Zue, "Transcription and alignment of the timit database," J. S. Garofolo, editor, National Institute of Standards and Technology (NIST), Gaithersburg, MD (1988).

[21] X. Yang, K. Wang, and S. A. Shamma, "Auditory representation of acoustic signals," IEEE Trans. Inf. Theory **38**(2), 824–839 (Special issue on wavelet transforms and multi-resolution signal analysis) (1992).

[22] S. V. David and J. L. Gallant, "Predicting neuronal responses during natural vision," Network **16**(2–3), 239–260 (2005).

[23] P. Ladefoged, *A Course in Phonetics*, 5th ed. (Harcourt Brace, Orlando, 2006).

[24] K. N. Stevens, *Acoustic Phonetics* (MIT Press, Cambridge, MA, 1980).

[25] S. Shamma, "Speech processing in the auditory system. Part I: The representation of speech sounds in the responses of the auditory-nerve," J. Acoust. Soc. Am. **78**(5), 1612–1621 (1985).

[26] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403 (1979).

[27] C. E. Schreiner, H. L. Read, and M. L. Sutter, "Modular organization of frequency integration in primary auditory cortex," Annu. Rev. Neurosci. **23**, 501–529 (2000).

[28] H. L. Read, J. A. Winer, and C. E. Schreiner, "Functional architecture of auditory cortex," Curr. Opin. Neurobiol. **12**(4), 433–440 (2002).

[29] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**, 1220–1234 (2001).

[30] L. A. Chistovich and V. V. Lublinskaya, "The center of gravity effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli," Hear. Res. **1**, 185–195 (1979).

[31] We emphasize that this response pattern is unlikely to be due to a nonuniform sampling of the scale and frequency variables, since no such bias in the joint distribution of the scale frequency is evident in Fig. 6(A). Furthermore, note that high scale neurons can be driven well by spectra with low frequencies as in phoneme /o/. The opposite is true for vowel /e/ where low scale units are driven well by high frequency energy.

[32] W. Klein, R. Plomp, and L. C. Pols, "Vowel spectra, vowel spaces and vowel identification," J. Acoust. Soc. Am. **48**(4), 999–1009 (1970).

[33] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice* (Prentice–Hall, Englewood Cliffs, NJ, 2002).

[34] O. Deshmukh, C. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: Detection of the periodicity and aperiodicity profile of speech," IEEE Trans. Speech Audio Process. **13**(5), 776–786 (2005).

[35] D. Bendor and X. Wang, "The neuronal representation of pitch in primate auditory cortex," Nature (London) **436**, 1161–1165 (2005).

[36] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, New York, 1995).

[37] J. B. Allen, *Articulation and Intelligibility* (Morgan and Claypool, 2005).

[38] D. Depireux, J. Z. Simon, and S. Shamma, "Measuring the dynamics of neural responses in primary auditory cortex," Comments in Theoretical Biology **5**(2), 89–118 (1998).

[39] N. Kowalski, D. Depireux, and S. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex: Prediction of single-unit responses to arbitrary dynamic spectra," J. Neurophysiol. **76**(5), 3524–3534 (1996).

[40] L. M. Miller, M. A. Escabi, H. L. Read, and C. E. Schreiner, "Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex," J. Neurophysiol. **87**, 516–527 (2002).

[41] C. T. Novitski *et al.*, Program 800.18/Poster E45, Neural coding of speech sounds in naïve and trained rat primary auditory cortex, Society for Neuroscience, Atlanta (2006).