# Phoneme representation in primary auditory cortex

Shihab Shamma, Nima Mesgarani, Stephen David, and Jonathan Fritz

*Institute for Systems Research, Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA*

We examined the responses of neurons in primary auditory cortex (A1) to phonetically labeled speech stimuli. Sentences were taken from the TIMIT database and chosen to represent a diversity of male and female speakers. We presented these stimuli to awake ferrets while recording the activity of isolated A1 neurons. For analysis, we segmented the continuous speech samples into sequences of phonemes, which represent the smallest significant units of speech. We characterized the response properties of each neuron as the peristimulus time histogram (PSTH) response to each phoneme. Across a population of A1 neurons, we observed distinct patterns of phoneme selectivity that may provide a neural basis or low-level phoneme discrimination. We investigated how features of speech are encoded in A1 using a method for reconstructing the speech stimulus from the neural population responses. Stimuli were reconstructed using a linear spectro-temporal model to map the response to the stimulus spectrogram. We compared the accuracy of reconstruction across phonemes. One important factor involved in stimulus reconstruction is the presence of correlations in complex natural stimulus such as speech. Prior knowledge of regularities in the stimulus can benefit reconstruction in noise and when spectro-temporal coverage is limited. We studied the influence of prior knowledge of stimulus correlations, noise and spectro-temporal coverage on reconstruction accuracy in neural data and in simulation.

## INTRODUCTION

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. 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 (Walker *et al.*, 2006). 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

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.). ISBN: 87-990013-1-4. Print: Centertryk A/S.

small population of primary auditory cortical neurons (< 100) can account for distinctive aspects of phoneme identification observed in humans (Miller and Nicely 1955), and that (II) long-established acoustic features of phonemes are indeed explicitly encoded in the population responses in A1.

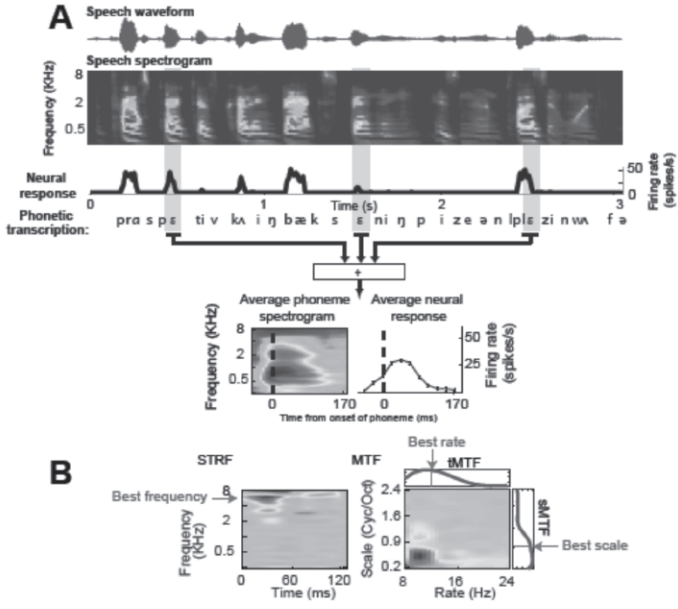


**Fig. 1**: Neuronal responses to phonemes in continuous speech (A) The spectrograms of all /ɛ/ vowel exemplars are extracted and averaged to obtain one grand average auditory spectrogram (bottom left). In this and following average spectrogram plots, red areas indicate regions of higher than average energy and blue regions indicate weaker than average energy. The corresponding PSTH response to /ɛ/ is 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 is 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 2-D 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 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).

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 (Theunissen *et al.*, 2001, Klein *et al.*, 2001). The key questions we address here concern 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.

## METHODS

Physiology: Spiking activity was recorded from isolated single neurons in primary auditory cortex of awake, passive ferrets. The experimental preparation and electrophysiological methods are described in detail in Klein *et al.*, (2006). Speech stimuli were phonetically transcribed continuous speech from TIMIT database (Seneft and Zue, 1988). The samples were chosen to represent a diversity of male and female speakers. Thirty different sentences spoken by 15 male and 15 female different speakers were used to represent a variety of speakers and contexts while keeping the time of neural data acquisition suitably short. Figure 1 illustrates the spectrogram of one such sentence, and the way responses for a given phoneme (/eh/) are averaged. Neurons spectro-temporal receptive field (STRF): The STRF is a linear model of the manner in which auditory cortical neurons respond to complex sounds. It is estimated using normalized reverse correlation techniques from any sound-response pairs (Theunissen *et al.*, 2001). Figure 1B 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 2-D modulation transfer function (MTF) (Fig.1B). The MTF is the 2-D 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. 1B). 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. 1B.

## RESULTS

### Average phoneme representations

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. In primary auditory cortex, unlike in the auditory nerve, receptive fields (tuning curves or STRFs) exhibit systematic variations along a myriad of feature axes including best frequency (BF), bandwidth, asymmetry, and temporal modulations (Simon *et al.*, 2007). Here we consider the ordered representation of phoneme responses along three different dimensions: *best frequency, best scale, and best rate* (Figure 1B). We used the speech-based STRFs to estimate these parameters for each neuron.

## Encoding of vowels

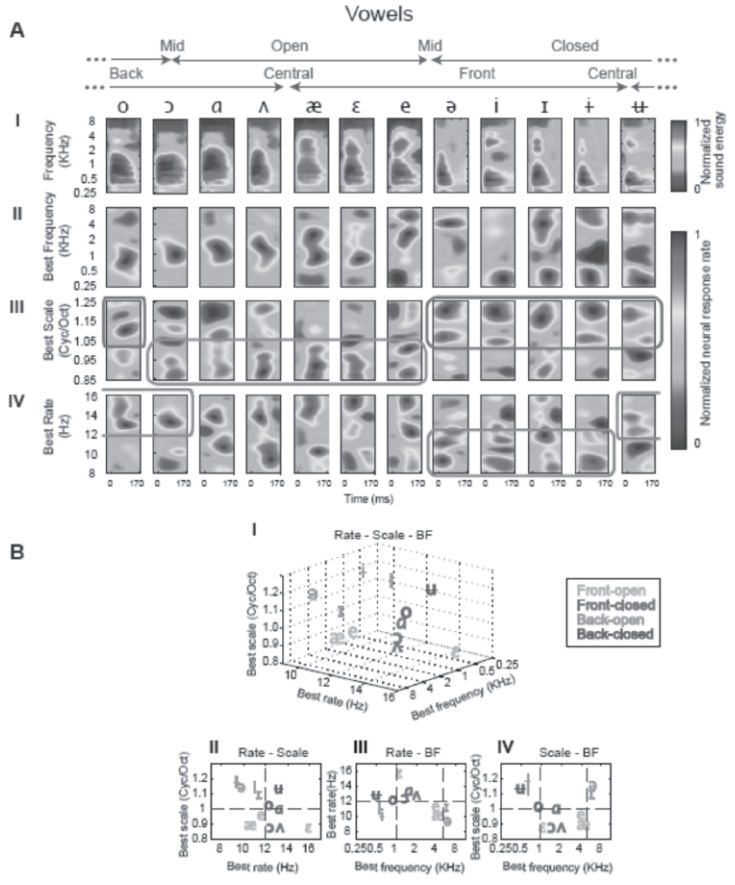


**Fig. 2**: Population response to vowels. (A) I. Average auditory spectrogram of 12 vowels organized approximately according to their open-closed and front-back articulatory features. (II, III, IV): Average PSTH responses of 90 neurons to each vowel. Within each heat map, each row indicates the average 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. (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 hollow font and Back vowels with solid font. 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) rate-scale, (III) rate-frequency and (IV) scale-frequency.

Population responses to 12 American-English vowels are summarized in Fig. 2. Panels in the top row (Fig. 2A-I) display the average auditory spectrogram of each vowel

computed from all of its samples encountered in the speech database. The vowels are organized according to their articulatory configurations along the Open/Closed and Front/Back axes (Ladefoged, 2006), 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 1-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 (Ladefoged, 2006), 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, 1979). 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. 2A-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. 2A-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. 2A-III, respectively. Response distributions in the best rate panels (Fig. 2A-IV) reveal a trend in the dynamics of the vowels as one moves along the Front/Back axis. Specifically, Front vowels (/ə/, /i/, /ɪ/, /ɨ/) evoke relatively stronger responses in the slower cells (with best rates <~ 12 Hz), as compared to the more Back vowels (/ʉ/, /o/, /ɔ/) as highlighted by the green boxes in Fig. 2A-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 2B 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 3-D display is projected onto each of the three marginal planes (Figs. 2B-II and 2B-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. 2B-II and 2B-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. 2B-III). Finally, the best rate axis segregates the Front/Back vowels (as discussed earlier), with Central and Back vowels locates at high rates (> 12 Hz), and Front vowels below it.

**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 (Ladefoged, 2006). 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. 3A-II) approximates the features of their averaged spectra (Fig. 3A-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 milliseconds 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. 3A-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 or our limited sampling of the neural population (90 neurons).

Response distributions along the best scale and best rate axes (Figs. 3A-III and 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 (Fig. 3A-III and 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. 3A-III and IV) compared to the more suppressed responses to longer duration unvoiced fricatives and nasals (outlined in blue in Fig. 3A-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. 3A-III and IV).

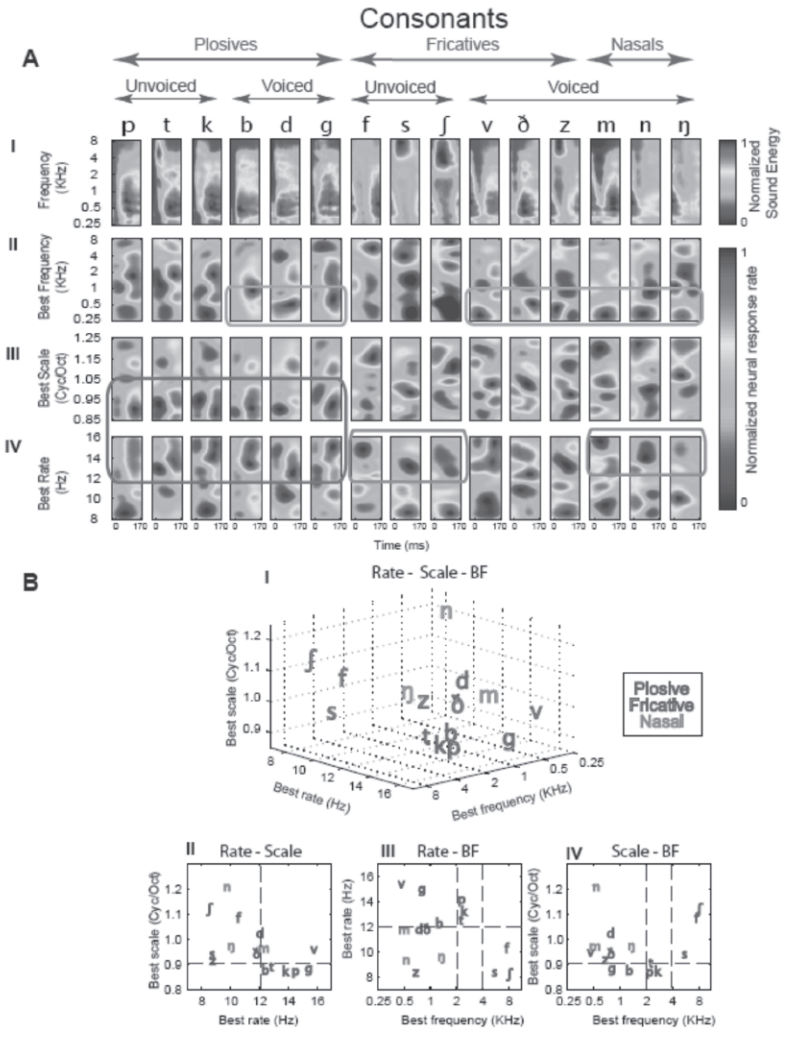


**Fig. 3**: Population response to consonants. (A) I. Average spectrogram of 15 consonants phonemes grouped as 6 plosives, 6 fricatives and 3 nasals. Each of the plosive and fricative groups contains 3 voiced and 3 unvoiced phonemes (see arrows at top). (II, III, IV) Average PSTH responses of the neural population to each consonant, plotted as in Fig. 2A. The average PSTH responses are sorted by neurons' best frequency (II), 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 Section II). (B) Each consonant is placed at the centroid frequency, rate and scale of its neuronal population response, measured from the corresponding PSTH responses (Fig. 3A). Plosive phonemes are plotted in red, fricatives in blue and nasals in green. The locus of each consonant is also shown collapsed in 2-D plots of (II) rate-scale, (III) rate-frequency and (IV) scale-frequency.

Finally, the third cue of voicing is associated with the harmonic structure of voiced spectra near the low to mid-frequency range (0.2 to 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. 3A-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. 3A-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; 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 (Bendor and Wang, 2005).

Fig. 3B 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. 3B are projections of the 3-D plot onto its three marginal planes. Members of the three groups of consonants - plosives (red), fricatives (blue), and nasals (green) - are located roughly close 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. 3B-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 (> 4KHz), unvoiced plosives falling between 2-4 KHz, and other voiced phonemes falling below 2 KHz (Figs. 3B-III and IV). As with vowels, this plot of the neural loci of consonants reveals the relative distances among them and perhaps explains the pattern of perceptual confusion observed among them, as we shall elaborate next.

**Phoneme recognition based on the responses**

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 phonemes, 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 Support Vector Machine (SVM , Vapnik, 1995) for each phoneme to separate it from all others, based on the responses of the neural population. To determine the identity of a novel phoneme, the population response was input 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. Studying the pattern of pair-wise confusions by the classifier can assess the extent to which the neural phoneme representations can account for the perception of individual phoneme exemplars.

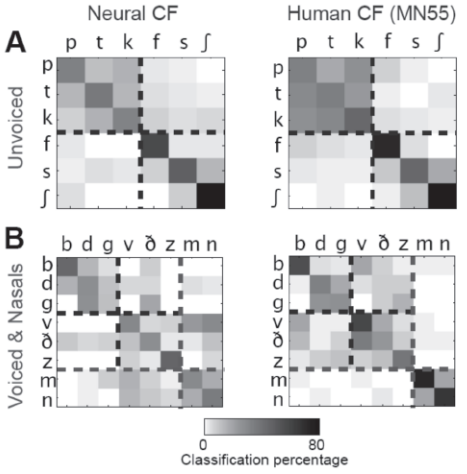


**Fig. 4**: Neural and human phoneme confusions. (left) shows the confusion matrix measured from classifications of the neural data. Labels along each row indicate the phoneme presented, and columns report the probability of the phoneme output by the classifier. The classifier was trained on responses of 20 neurons to 330 seconds of speech (90 sentences). The phonemes are arranged based on voiced-unvoiced (Fig. 6A and B) and plosive, fricative, nasal consonant categories to facilitate comparison with a previous study of human perception (Miller, and Nicely 1955) (replicated in Fig. 6 (right)). The dashed boxes delineate the 3 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).

## RECONSTRUCTIONS

Top-down effects of behavior and attention on cortical responses can best be discerned from the activity of large populations of neurons. This can be done in the auditory cortex by characterizing the stimulus-response relationships of the neural population in terms of their equivalent "forward" and "inverse" models (explained below). During behavior, top-down signals change these models, allowing us in turn to estimate the nature and extent of the effects on the neural representations. From the encoding point of view, the methods that are used to measure the information content of the neural response often fail to specify what aspects of the stimuli are encoded (Theunissen, 1993). One possible way to investigate this question is reverse reconstruction of the stimuli where the best approximation of the input stimuli is estimated from the population response. These reconstructed stimuli can be compared to the original to understand what features are preserved. Here we explain two methods for the reconstruction of the stimuli, the forward and inverse models and investigate the encoding of phonemes using this method.

**Forward model**

Forward model (STRF) (Figure 5, left) is the transformation that explains the mapping of the sound spectrogram to the neural response. Intuitively it explains what a neuron is tuned to and can be used to predict the response of a neuron to a novel stimulus. The forward model of a neuron, $h(t,f)$ transform the sound spectrogram $s(t,f)$ to the neural response $r(t)$:

$$r(t) = \sum_f \sum_t h(\tau, f) s(t - \tau, f) \qquad \text{(Eq. 1)}$$

The function $h$ is estimated by minimizing the mean squared error between actual and predicted response:

$$\min \quad e = \sum_t \left(\hat{r}(t) - r(t)\right)^2 \rightarrow \quad h \; = \; C_{ss}^{-1} C_{sr} \qquad \text{(Eq. 2)}$$

It is not possible to invert this equation to find $s(t,f)$ from $r(t)$ because of the ambiguity of the frequency dimension (response is a function of time and not frequency). However, we can recover the frequency dimension provided we have enough neurons to construct an invertible system of linear equations (full coverage of the frequency space).

Assuming we have the response of $n$ neurons to the same sound, we construct the following system of linear equations:

$$\begin{aligned} r_1 &= h_1 S \\ &\vdots \qquad \rightarrow \quad R = HS \\ r_n &= h_n S \end{aligned} \qquad \text{(Eq. 3)}$$

Assuming the $H$ matrix has a pseudo inverse, we invert equation 3 to find $S$:

$$S = (H^T H)^{-1} H^T R \qquad \text{(Eq. 4)}$$

**Inverse model**

The inverse model of a neuron (Figure 5, right) is the transformation that maps the neural response back to the sound spectrogram. The inverse model is not as intuitive as STRF because it is the property of a neuron in a population. We can estimate the inverse model, $G(t,f)$ that is defined as follows:

$$S(t, f) = \sum_n \sum_\tau G_n(t - \tau, f) r_n(\tau) \qquad \text{(Eq. 5)}$$

The function $G$ is then estimated by minimizing the mean squared error between actual and reconstructed stimulus which results in normalized reverse correlation:

$$\min \; e = \sum_f \sum_t \left(\hat{S}(t, f) - S(t, f)\right)^2 \rightarrow \; G = C_{rr}^{-1} C_{rs} \qquad \text{(Eq. 6)}$$

Since the statistics of stimulus have been removed in the calculation of H, the forward method does not assume any prior knowledge about the stimulus, however the inverse method takes into account the prior statistics of the stimuli. Furthermore, because

data from many recordings are combined to reconstruct the same stimulus spectrogram, it is reasonable to expect that increasing the number of recorded responses (neurons) results in a faster and cleaner reconstruction of the spectrogram. In summary, by recording simultaneously from many neurons in AI and other auditory cortical fields where STRFs can be reliably measured, one can reconstruct an online approximation of the stimulus S.

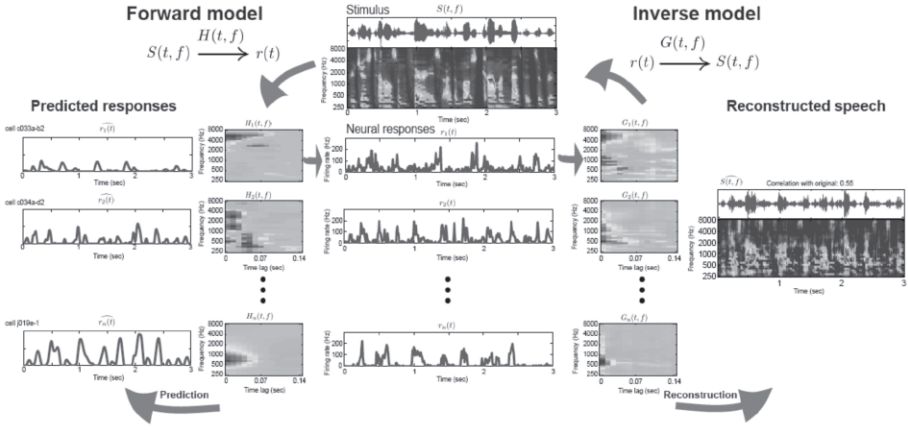


**Fig. 5**: illustration of forward (left) and inverse (right) models of stimuli-response mappings. The forward model of a neuron (H) maps the spectrogram of the sound to the neural response. Having the model, one can predict the response of the neuron to any novel stimuli (left column). The inverse model (G) is a mapping from the neural response back to the sound spectrogram. Using the G functions, we can reconstruct the stimuli spectrogram from the neural population response (right).

**Average phoneme spectrograms from the original and reconstructed stimulus**

Phonemes vary across various spectral and temporal dimensions because of the way they are produced. Some of the parameters affecting the phonemes include the shape of the vocal tract, vibration of vocal cords and manner and place of articulation. To examine the encoding of these features in A1 we estimated the average phoneme spectrograms from the original and reconstructed signals. The average phonemes are shown in Fig 6 for consonants (top) and vowels (bottom). We divided the group of consonants based on their manner of articulation into fricatives, plosives and nasals. Consonants within these categories are different in their place of articulation. For each phoneme, two average spectrograms are shown obtained from original and reconstructed spectrograms. We quantified the similarity between original and reconstructed spectrograms using correlation coefficients as shown in Fig 6e. The correlation coefficients for almost all vowels are high indicating the effective encoding of the formant frequencies by the neurons. Plosives (/p, t, k, b, d, g/) show slightly less correlation (from 0.8 to 0.9), but higher than the fricative group (/s, ʃ, f, z, v/) (0.7 to 0.9). Overall, the high correlation between original and reconstructed phonemes shows a strong encoding of the phonemic features in the primary auditory cortex.

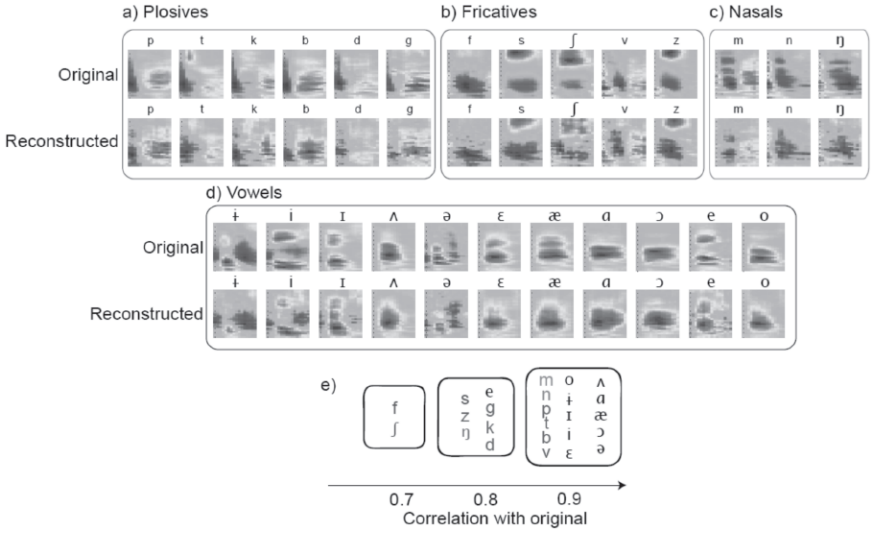


**Fig. 6:** Average phoneme spectrograms from the original and reconstructed signals. The average phonemes are grouped into a) plosives, b) fricatives, c) nasals and d) vowels. For all phoneme groups, the important phonemic features are preserved in the reconstructed signal indicating their proper encoding. The similarity is quantified using correlation coefficients as shown in (e).

## SUMMARY AND CONCLUSION

Responses to speech in primary auditory cortex reveal a multidimensional representation that is sufficiently rich to support the perceptual 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 sub-population of neurons that responds well to the distinctive acoustic features of a given phoneme and hence encodes that phoneme in a high-dimensional space. Three dimensions of neural tuning considered in this study are the best frequency, rate (temporal modulations) and scale (spectral shape). We showed that frequency tuning of neurons provides a representation of the place of articulation and that rate and scale tuning provide a representation of manner of articulation, distinguishing plosives, fricatives and nasals. The explicit representation of phoneme identity across a population of filters tuned to BF, scale and rate suggests a strategy for improved speech recognition systems in noise and other sources of variability.

## ACKNOWLEDGEMENT

Partial funding for this project was obtained from the Air Force Office of Scientific Research, and the National Institutes of Health (NIH) Grants R01DC005779.

## REFERENCES

Walker, K., King, A., Ahmed, B., and Schnupp, J. W. H. (**2006**). "Psychometric and neurometric discrimination of non-conspecific vocalizations," Abstract 430, Mid-Winter Meeting of Association for Research in Otolaryngology, Baltimore.

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

Theunissen, F.E., David, S.V., Singh, N.C., Hsu A., Vinje, W.E., and Gallant, J.L. (**2001**). "Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli," 1: Network. **12**(3): 289-316.

Klein, D.J., Simon, J. Z., Depireux, D. A., and Shamma, S. A. (**2006**). "Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex," J Comput Neurosci., **20**(2): 111-36.

Seneff. S., and Zue, V. (**1988**). "Transcription and alignment of the timit database", J. S. Garofolo," Ed. National Institute of Standards and Technology (NIST), Gaithersburgh, MD.

Simon, J. Z., Depireux, D. A., Klein, D. J., Fritz, J. B., and Shamma, S.A. (**2007**), "Temporal Symmetry in Primary Auditory Cortex: Implications for Cortical Connectivity, Neural Computation," **19**, 583-638.

Ladefoged, P., A. (**2006**). "Course in phonetics. Orlando: Harcourt Brace," 5th ed. Boston: Thomson/Wadsworth.

Chistovich, L. A., and Lublinskaya, V. V. (**1979**). "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.

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

Vapnik, V. N. (**1995**). The Nature of Statistical Learning Theory, Springer.

Theunissen, F. E. (**1993**). "An Investigation of Sensory Coding Principles Using Advanced Statistical Techniques," Thesis, Univ. California, Berkeley.