REPRESENTATION OF PHONEMES IN PRIMARY AUDITORY CORTEX: HOW THE BRAIN ANALYZES SPEECH

Nima Mesgarani, Stephen David, Shihab Shamma

Electrical and Computer Engineering Department
University of Maryland, College Park, MD 20742

ABSTRACT

Many transformations inspired by the auditory system have improved the performance of Automatic Speech Recognition (ASR) systems. However, humans perform substantially better than today's ASR systems, suggesting that ASR systems can further benefit from understanding how the brain represents speech. To learn about the cortical representation of speech, we measured the neural responses in the primary auditory cortex to sentences from the TIMIT database. Here we examine how individual phonemes activate different subsets of auditory neurons, reflecting the diversity of neural tuning properties. We find that neurons with different spectro-temporal tuning provide an explicit multidimensional representation of articulatory features independent of speaker and context. This representation that matches the human perception could provide a framework for ASR in adverse conditions.

Index Terms— Auditory System, speech processing, speech recognition

1. INTRODUCTION

The brain is a working example of a robust speech recognition system. Humans perceive speech in noisy and reverberant environments despite large variability across different speakers and contexts. Experience with automatic speech recognition has amply demonstrated that the performance of ASR systems is considerably enhanced through use of early and central auditory-inspired transformations (PLP [1], MFCC, RASTA [2] and spectro-temporal features [3]). Learning more about how the brain analyzes continuous speech will enable us to design systems that operate robustly in natural conditions.

Numerous important acoustic features of speech, such as voice onset time and articulatory features, have been found to be robustly represented in auditory cortex [4] and sub-cortical nuclei [5]. These experimental findings suggest that the basic machinery for the robust encoding of speech already exists in mammalian auditory systems independent of learning. However, previous studies of the auditory system have focused on simple synthetic sounds, and very little is known about the representation of more complex natural sounds especially continuous speech. To learn more about how the brain analyzes speech, we measured the representation of continuous speech in the primary auditory cortex neurons and examined how individual phonemes modulate activity across the population of auditory neurons. The selectivity of neurons is characterized by their spectro-temporal receptive field (STRF), which is the linear approximation of the transfer function between sound spectrogram input and neural output. We demonstrate how neurons with different frequency tuning, temporal modulation tuning (rate) and spectral shape tuning (scale) generate a multidimensional representation in which different phoneme categories form unique patterns that may provide a neural basis for low-level phoneme discrimination.

2. EXPERIMENTAL SETUP

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 [6].

Speech Stimuli: Stimuli were phonetically transcribed continuous speech from TIMIT database [7]. 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.

Neurons spectro-temporal receptive field: One of the most successful models of the auditory cortical neurons is the linear spectro-temporal receptive field (STRF), $h(t, f)$ that maps the spectrogram

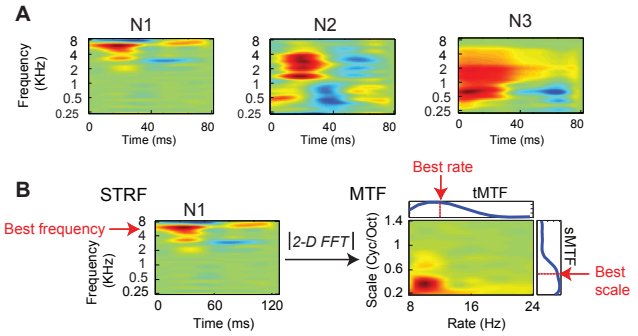

Fig. 1. A) Spectro-temporal receptive fields of three cortical neurons as measured by normalized reverse correlation. Red/blue areas indicate stimulus frequencies and time lags correlated with an increased/decreased response. B) The neurons best frequency is defined to be the excitatory peak of the STRF (red arrow). The modulation transfer function (MTF) is computed by taking the absolute value of the 2-D Fourier transform of the STRF. The best scale and the best rate of the neuron is defined as the centroid of the collapsed MTF along rate and scale axes (red arrows).

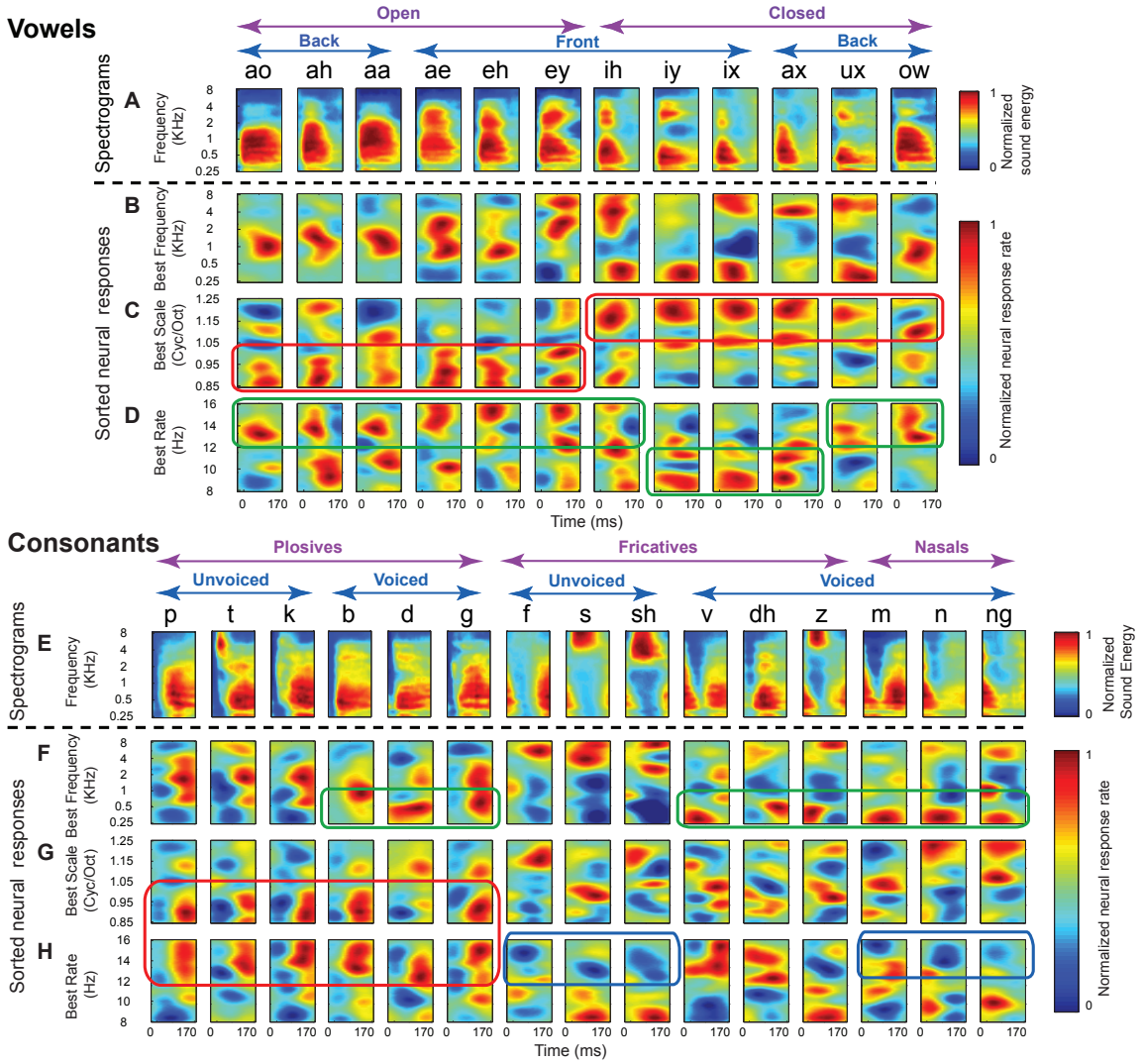

Fig. 2. Average population responses to vowels and consonants. (A) The average spectrogram of 12 vowels. The neural population response to these vowels is displayed in the heat maps in rows B-D. Each horizontal line of the heat map shows the timecourse of the average response of a single neuron to the corresponding vowel. Red regions indicate strong responses and blue regions indicate weak responses. The population responses are displayed in three ways: sorted by their best frequency (B), best scale (C) and best rate (D). (E) Average spectrogram of 15 consonants. Rows F-H show the average neural population responses sorted by neurons best frequency, scale and rate.

of the sound, $s(t, f)$, to the neural output, $r(t)$:

$$r(t) = \sum_{f=1}^{F} \int h(\tau, f) s(t - \tau, f) d\tau \qquad (1)$$

In the STRF model, the output (neural response) is the sum of F LTI filters each operating on one frequency of the sound spectrogram. The STRF of the neurons can be estimated using normalized reverse correlation techniques from any sound-response pairs [8]. Some example STRFs obtained from the speech data are shown in figure 1A. These examples show how neurons differ in their frequency tuning and in their temporal and spectral tuning shape, characterized by best frequency, best rate and best scale [6]. These parameters are measured from the STRF and its 2-D Fourier transform as shown in figure 1B.

3. AVERAGE PHONEME REPRESENTATIONS

Population responses to phoneme classes: To appreciate the unique response patterns evoked by different phonemes, and in particular, to highlight the acoustic features enhanced in the neural representation, it is best to view the ordered activity of the entire population. 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 [6]. Here we consider the ordered representation of phoneme responses along three different dimensions: best frequency, best scale, and best rate (figure 1B) [9]. We used the speech-based STRFs to estimate these parameters for each neuron.

Encoding of vowels: Population responses to 12 American-English vowels are summarized in Figure 2. Panels in the top row (2A) display the average auditory spectrogram of each vowel computed from all samples in the database. The vowels are organized according to their articulatory configuration along the Open/Closed and Front/Back axes as illustrated at the top of the figure: /ao, ah, aa, ae, eh, ey, ih, iy, ix, ax, ux, ow/. The averaged spectra (top row) reveal that open vowels have relatively concentrated activity at medium frequencies (0.8- 2 KHz), whereas closed vowels tend to have sometimes two peaks spaced over a larger frequency range (0.5 and 4 KHz). These are consistent with the known distribution of the three formants (F1, F2, and F3) in these vowels [10]. 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 (fiugre 2B) More striking, however, are the response distributions along the best scale axis (figure 2C). Here, open vowels tend to evoke maximal responses in broadly tuned cells commensurate with their broad spectra (low scales at < 1 Cyc/Oct), while closed vowels evoke maximal responses in narrowly tuned cells (scales > 1 Cyc/Oct), as indicated by the red boxes in the figure. There is also a consistent pattern along the best rate axis: front/closed vowels are more likely to excite slower cells with rates $<$ 11-12 Hz compared to all other vowels (as indicated by the green boxes in figure 2D). This pattern may reflect the longer times required to complete the articulatory excursions toward or away from closed vowels at the front of the mouth. Figure 3A 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 separate them, the 3-D display is projected onto each of the three marginal planes (figure 3B-D). It is evident in these displays that the vowels are clearly separated along the scale axis above and below 1 Cyc/Oct (horizontal dashed lines in panels figure 3B,D). They are also distinguished by BF, with the open vowels clustering in the range 1.0 - 4.5 KHz (vertical dashed lines in panel figure 3C.

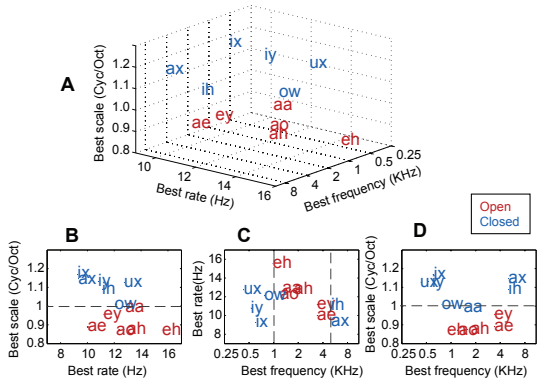

Fig. 3. (a) Each vowel is displayed at the locus of the average frequency, rate and scale of the neurons activated by that vowel. Back vowels are shown in red while front ones are in blue. To reveal the contribution of each dimension to the separation of front-back vowels, we project them onto (b) rate-scale, (c) rate-frequency and (d) scale-frequency planes.

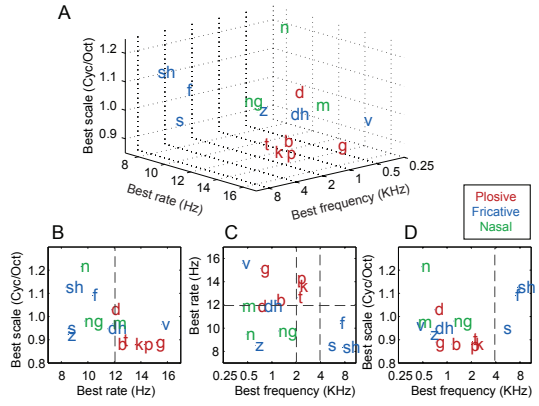

Fig. 4. (a) Each consonant is displayed at the locus of the average frequency, rate and scale of the neurons activated by that consonant. Plosives are shown in red, fricatives in blue and nasals in green. The contribution of each dimension to the separation between categories is evident from the projections onto (b) rate-scale, (c) rate frequency and (d) scale-frequency planes.

Encoding of Consonants Population responses to several consonants are shown in Figure 2E-H 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 [10] . We studied the population responses to determine how these properties are encoded across the population of neurons. The distributions of the responses to the consonants sorted along the BF axis (figure 2F) approximates the features of their averaged spectra (figure 2E), which in turn are known to be closely related to place of articulation cues. For instance, 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/). Response distributions along the best scale and best rate axes (figure 2G-H) 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 "obstruents" and "continuants" (e.g. /p,t,k,b,d,g/ versus /s,sh,z,n,m,ng/) is evident in the distribution of responses along the scale and rate axes (figure 2G-H). Because of their sudden and spectrally broad onsets, obstruents (also called stop consonants) display relatively strong activation in broadly tuned and fast cells (regions outlined in red in figure 2G-H) compared to the more suppressed responses to longer duration unvoiced fricative and nasal continuants (outlined in blue). Finally, the third cue of voicing is primarily associated with the harmonic structure of voiced spectra near the mid-frequency range (0.2 to 1 KHz), and the weak energy at very low BFs near the fundamental of the voicing. Only this latter cue seems to distinguish consistently the voiced (/b,d,g,v,dh,z,m,n,ng/) from unvoiced (/p,t,k,f,s,sh/) consonants in our data as indicated by the green outlined region of figure 2F. Figure 4A 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 Figure 4 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)) fall in distinct clusters. For instance, plosives tend to drive broadly tuned and fast cells. 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. As with vowels, this plot of the neural loci of consonants reveals the relative distances among them and presumably their tendency toward perceptual confusion, as we shall elaborate next.

4. PHONEME RECOGNITION BASED ON NEURAL POPULATION 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 [11]) 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. 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 (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 (figure 5A and B) and plosive, fricative, nasal consonant categories to facilitate comparison with a previous study of human perception [12] (replicated in figure 5 (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, sh/), 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).

5. 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 independent of speaker and context. 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.

6. ACKNOWLEDGEMENTS

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

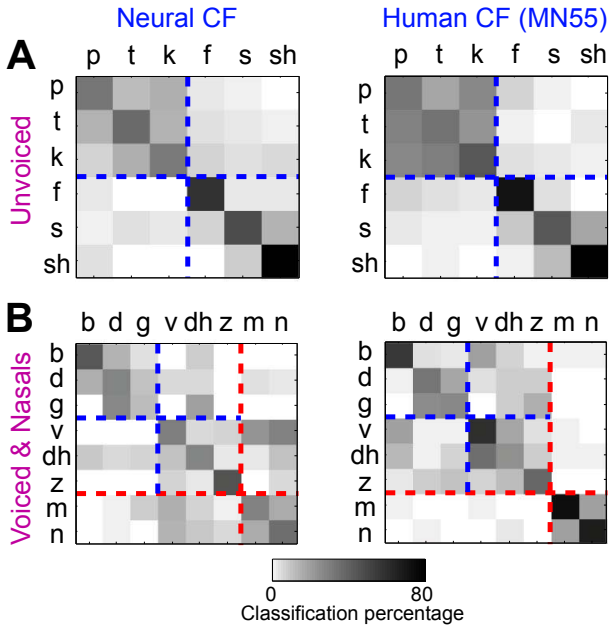

Fig. 5. Confusion matrices obtained from human psychophysics (left column) and neural population responses (right column). The confusion matrices are shown for unvoiced (top row) and voiced and nasal consonants (bottom row). Within each matrix, each row shows the actual phoneme and each column shows the predicted phoneme. Darker regions indicate a greater probability of prediction. The neural confusion shows a similar pattern to human perception, especially for unvoiced consonants.

7. REFERENCES

[1] H. Hermansky, *Perceptual linear predictive (PLP) analysis of speech*, J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, Apr. 1990.

[2] H. Hermansky and N. Morgan, *RASTA processing of speech*, IEEE Trans. on Speech and Audio Proc., vol. 2, no. 4, pp. 578-589, Oct. 1994.

[3] M. Kleinschmidt, and D. Gelbart. *Improving Word Accuracy with Gabor Feature Extraction*, International Conference on Spoken Language Processing, Denver, CO, 2002.

[4] M. Steinschneider, IO. Volkov,M. D. Noh, P. C. Garell, M. A. Howard, *Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex*, J Neurophysiol. 1999 Nov;82(5):2346-57.

[5] K. L. Johnson, T. G. Nicol, N. Kraus, *Brain Stem Response to Speech: A Biological Marker of Auditory Processing*, Review Ear and Hearing. 26(5):424-434, October 2005.

[6] D. J. Klein, J. Z. Simon, D. A. Depireux and S. A. Shamma, *Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex*, Volume 20, Number 2 / April, 2006

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

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

[9] T. Chi, P. Ru, S. A. Shamma, *Multiresolution spectrotemporal analysis of complex sounds*, J Acoust Soc Am. 2005 Aug;118(2):887-906.

[10] P. Ladefoged, *A course in phonetics*. Orlando: Harcourt Brace. 5th ed. Boston: Thomson/Wadsworth 2006.

[11] T. Joachims, *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*, B. Scholkopf, C. Burges and A. Smola (ed.), MIT-Press, 1999.

[12] G. Miller and P. Nicely, *An Analysis of Perceptual Confusions among some English Consonants* J. Acoustical Society America, vol. 27, no. 2, pp. 338-352, 1955.