Combining Biological and Artificial Approaches to Understand Perceptual Spaces for Categorizing Natural Acoustic Signals

Marvin Thielk (mthielk@ucsd.edu)
Neuroscience, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Tim Sainburg (tsainbur@ucsd.edu)
Psychology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Tatyana Sharpee (sharpee@salk.edu)
Salk Institute, 10010 N Torrey Pines Rd, La Jolla, CA 92037

Timothy Gentner (tgentner@ucsd.edu)
Neuroscience, Psychology, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Abstract

Parametrizing complex natural stimuli is a difficult and long-standing challenge. We used a generative deep convergent network to represent and parametrize a large corpus of song from European starlings, a songbird species, into a compressed low-dimensional space. We applied psychophysical methods to probe categorical perception of natural starling song syllables, which reveal a shared categorical perceptual space. Some categorical boundaries are sensitive to the category assignment of training syllables, indicating that the consensus is context dependent and that underlying dimensions of the space are not independent. Consistent with this, we predict the behavioral psychometric function along one dimension by fitting the behavior for other dimensions to artificial neural network activations. Similar predictions are obtained by fitting spike timings of in-vivo neuronal populations, recorded simultaneously from 10's of neurons in a secondary auditory cortical region. Thus, knowing how the animal responds in one sub-region of the parametrized space informs responses in other sub-regions of both the artificial and in-vivo spaces. Our results implicate the importance of experience in shaping shared perceptual boundaries among complex communication signals, and suggest the categorical representation of natural signals in secondary sensory cortices is distributed much more densely than predicted by traditional hierarchical object recognition models.

Keywords: categorical perception; generative DNN; birdsong

Introduction

Understanding how secondary auditory regions encode communication signals and how this representation gives rise to psychophysical measures such as categorical perception is a longstanding goal of sensory neuroscience. European starlings (*Sturnus vulgaris*) are an excellent established model organism to study auditory processing and categorical perception. Like human speech, starling song is composed of learned, spectrally complex, temporally-patterned acoustic objects (called ***motifs***), that are produced in long, well-organized temporal sequences (T. Q. Gentner & Margoliash, 2003), and that function in a wide range of natural behaviors. As with other complex natural signals, our understanding of how birdsongs are represented in higher cortical regions, both benefits from and is hindered by the complex spectro-temporal character of these sounds. Multiple physiological studies have used conspecific vocalizations, and reveal a strong selectivity for songs that emerges across the auditory forebrain and strengthens from field L to caudomedial nidopallium (NCM) and caudal mesopallium (CM) (T. Q. Gentner & Margoliash, 2003; T. Gentner, 2004; Thompson & Gentner, 2010; Jeanne, Thompson, Sharpee, & Gentner, 2011). However, the lack of parametric control over the complex acoustic features composing birdsongs (and other communication signals in other species) had rendered it difficult to more rigorously extensively characterize the information that these regions encode (and how). Ideally, we would like to parametrically control the complex natural stimuli to which high-order sensory regions are tuned, with the same precision and control that past studies have manipulated more simple stimuli like white noise and simple sine stimuli that can drive more primary sensory regions.

Here we present *a novel method to parametrize the natural auditory stimulus space*. We then apply classical behavioral psychophysics to describe how starlings categorize motifs that vary systematically within the parameterized space of their natural song. Our method has the potential to reveal novel insights into the representations of complex stimuli in secondary perceptual regions. Currently, our understanding of these secondary perceptual regions is unsatisfactory and often, what is known is a result of happening upon the precise stimuli that drives a specific neuron such as in the case of the infamous Jennifer Aniston neuron (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Our methods represent a way to traverse complex natural stimuli spaces in a focused, quantitative, flexible, yet rigorous manner.

Results

Unsupervised Parametrization Using a large corpus of recorded starling song, a compressive Deep Belief Network

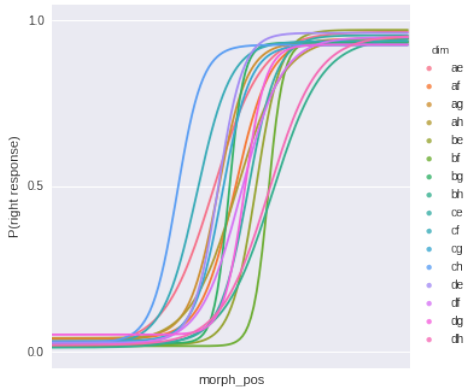

Figure 1: The set of 16 psychometric curves for a single bird. Each psychometric curve is fit using the maximum likelihood estimate of the probability of the bird choosing to respond right as the stimulus is morphed from a left-associated motif to a right-associated motif along the x axis. Note the large amount of variability in where this bird places the categorical boundaries on different motif dimensions and the different slopes (sensitivity) associated with each dimension.

(DBN) (Hinton & Salakhutdinov, 2006) is trained to autoencode 400 mS motifs (song segments) of starling song. The DBN accomplishes a non-linear dimensionality reduction of the starling motif corpus to a low dimensional latent space (64 dimensions), as well as providing a generative model from any point in this latent space to a starling song-like sound. The trained DBN allows one to interpolate between any two arbitrarily chosen starling motifs projected into the latent space, to create a smoothly varying continuum of morphed motifs that shift from one target motif to the other, without sounding like a simple linear crossfade between them.

Behavioral Training Starlings are trained on a two alternative choice task where four arbitrarily chosen motifs (labeled A, B, C, and D) are associated with a left response and another four (E, F, G, and H), are associated with a right response. After training to a stable performance criterion, the interpolated morph motifs generated by the DBN are used to probe the bird's perception as each left-associated motif is transformed into each right-associated motif. We employed a ratcheting double staircase that *allowed each subject to iteratively (and independently) estimate the categorical boundary along each of the 16 morph dimensions*. A ***psychometric curve*** (four parameter logistic function) is fit to the behavioral responses along the entire morph dimension between a left-associated motif and a right-associated motif.

In all cases, birds show very clear categorical perception as evidenced by the steepness each psychometric function regardless of subject or motif dimension. Comparing the psychometric curves within a single bird across all 16 motif-to-motif dimensions reveals a large amount of variation in the point of subjective equality (the category boundary) and in how sensitive the bird is to stimulus changes across the boundary (Fig. 1). This variability across dimensions is presumably a result of the non-linear nature of the DBN song compression and morphing. Thus, we can conclude that *the features the DBN uses to represent the motifs are perceptually relevant to the starlings*, and that the starlings are differentially sensitive to variation along these different feature dimensions.

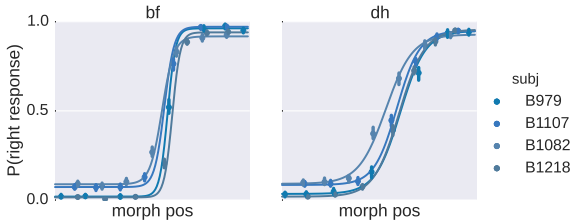

Figure 2: Psychometric curves for four starlings, over several months of training, are highly conserved between individuals, suggesting a shared perceptual space. The y-axis shows the probability of a right response to a stimulus morphed continuously between, for example on the left, motif B (reinforced as left) and motif F (reinforced as right). The x-axis is the morph point between the left associated motif to the right associated motif.

Despite the significant variability within a single bird across multiple morph dimensions, we observed a remarkable degree of consensus between birds. This included strong agreement in where each bird placed the category boundary on a given dimension, and in the sensitivity of all birds to changes along a given stimulus dimension (slope of psychometric function). Figure 2 gives an example of the typical agreement between subjects, where two of the 16 morph dimensions are plotted for 4 different birds. The left and right scaling parameters are more conserved across all morphs within a single bird indicating that they are bird specific parameters. This makes sense because they correspond to the bird's absolute performance on the left and right endpoints. The category boundary and the sensitivity are conserved within a given morph dimension across different subjects indicating a shared perceptual space of these synthetic natural-like sounds in these wild caught birds.

The shared perceptual space for motif categorization may result from either common training and experience, idiosyncrasies of the compressive network transformation, or some combination of the two. To test for this, we permuted the initial motif category assignments for a subset of birds. Instead of associating motifs A, B, C and D with left responses and motifs E, F, G, and H with right responses, a new cohort of birds learned, for example, to associate motifs A, B, E, and F with left responses and motifs C, D, G, and H with right responses. Thus, a subset of the 16 interpolating morph dimensions between left and right associated motifs is shared with the original cohort's interpolating morph dimensions. Comparing these shared boundaries demonstrates that on some of the interpolated morph dimensions, both cohorts of birds place the boundaries in same location as shown in the left of figure 3 while in other interpolating morph dimensions each

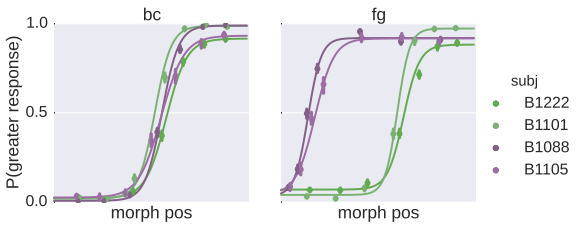

Figure 3: Psychometric curves from subjects trained on different category permutations. B1222 and B1101 (green) were trained to classify motifs ABEF as left and motifs CDGH as right. B1088 and B1105 (purple) were trained on ABGH as left, EFCD as right. This allows a subset of the morph dimensions to be compared across training on different category permutations. On most of these dimensions the boundary is conserved as if the training was the same, however, on some of the dimensions, the boundary depends on the initial categorical assignment. When it does shift, it shifts to the same location for all birds.

cohort has a separate boundary (but consistent within that cohort), as shown in the right of figure 3. The boundaries that are preserved across motif category permutation indicate that these dimensions are independent from the other dimensions, however, the boundaries that are shifted as a result of the motif category permutation indicate an interaction between the interpolated morph dimensions. This is unexpected, especially if one considers that in the latent space of the DBN there is minimal collinearity and no discernable structure between any of the 8 motifs used.

Another way to gauge the independence of the 16 interpolation dimensions is to use the representation learned by the DBN to predict the parameters of the behavioral psychometric function. Because we used the latent space to create linear, independent morphs, there isn't enough information on the relationships between the morph dimensions in the latent space alone to predict behavior. Using the activations of the entire DBN, however, can accurately predict (in a hold-one-out paradigm) the psychometric boundary and slope. Because this is a completely determined system with no noise we can show that the mean square error between the curve predicted by the logistic regression and the behaviorally determined psychometric curve is less than would be expected if the psychometric curves are randomly shuffled. Thus, the stimulus dimensions used for categorization are not independent of each other. Knowing how the subjects treat one subregion of the latent space reveals information about how other portions of the space are perceived.

Electrophysiological Recordings To explore how neural representations of secondary auditory regions, such as ***CM***, varied along these morph dimensions we recorded extracellularly from many single- and multi-neuron sites in CM simultaneously, in lightly anesthetized starlings. CM was targeted using established stereotaxic coordinates and a 32-channel silicon electrode was lowered into CM until neurons

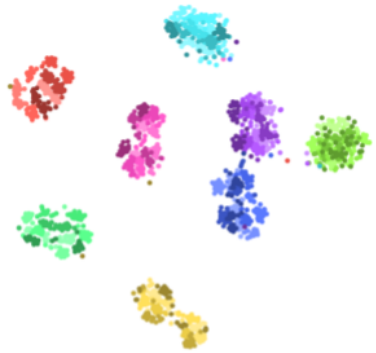

Figure 4: tSNE embedding of the neural response of a population of putative units in CM of many repetitions of the eight template motifs used in our initial operant conditioning paradigm. Each of the 8 colors indicate a different motif. The separation between each cluster indicates the distinct categorically separate representation of this neural population to each of the motifs. Different intensity of color indicates nearly identical spectrograms, but different random initializations of the spectrogram inversion.

responsive to conspecific songs were observed. We then presented a range of morph and training stimuli, selected so that half of the presentations were equally spaced along each of the interpolated morph dimensions and half were equally spaced in perceptual space to ensure adequate sampling near the relevant behavioral boundaries.

For each recorded neuron (sorted post-hoc) in the population we convolved its spike train with a Gaussian to yield an n-wise time varying estimate of firing rates on each trial that preserved spike timing information, where n is the number of simultaneously recorded neurons. A logistic regression is able to very accurately predict the identity of the 8 training motifs from most individual neurons in the recorded population response, and provided an exclusion criterion to remove units that did not contain much stimulus specific information (relevant to our training motifs).

Visualizing the neural population representation (concatenating all the individual unit representations) of the presentations of the 8 original template motifs using tSNE reveals a high degree of separation between clusters of different stimuli as seen in figure 4. Similar patterns are observed in naive and trained animals, regardless of whether they've heard the stimuli before the neural recording.

Using the hold-one-dimension-out prediction strategy, similar to that for the DBN activations, we predicted the behavioral psychometric functions from the neural population representations. The behavioral response is fit given the neural representation of stimuli presented from 15 of the 16 interpolated morph dimensions and then the response is predicted on the neural responses of the remaining dimension. A maximum likelihood four parameter logistic function is fit on the predicted responses of the trained logistic regression. The performance of this method varied between recording sites and different birds. We are still working on characterizing these differences

more in greater detail.

Discussion

Our results overcome a long-standing impediment to understanding the perception of natural communication signals. We demonstrate a method for parametrizing complex stimuli and generating *smoothly varying morphs between these stimuli*, as well as how to use these morphs to explore the perceptual basis, behaviorally and neurally, of the natural stimulus space. To our knowledge, this marks one of the first naturalistic parametric explorations of non-human auditory communication signals. Our characterization of the perception of this space and its neurological underpinnings, reveals remarkable behavioral consensus between animals for categorical boundaries and a broadly distributed encoding strategy for categorical stimulus information at the neural population level.

The field of machine learning is rapidly evolving and there are number of possible improvements to the processing and methodology, however, this work mainly demonstrates the usefulness of these kinds of techniques for understanding the perception of complex natural communication signals. In addition to changes in network architecture, newer implementations of spectrogram inversion would improve stimulus generation and are currently being tested and developed. In our experiment, however, different initializations of the spectrogram inversion process revealed that neurons in CM are sensitive to these small physical differences in physical stimuli that are imperceptible to human ears figure 4.

The work demonstrates the existence of a shared perceptual space, common across individuals, in which perceived categorical boundaries cluster at consensus locations. This kind of consensus is a pre-requisite to functional communication systems that use discrete signals. Furthermore, the 16 interpolating morph dimensions used in this study are not independent, nor are they a simple linear function of the endpoints, independent of the structure of the network space. If the latter were true, then all (or none) of the boundaries would shift when the endpoints were permuted. Instead, because only some of the dimensions are affected by permutation of the initial categories, not all the dimensions are independent, and the relationships between them are likely complex. Moreover, because there are dimensions that are not changed by the permutation of the initial categories, the decision boundaries learned by birds cannot rely on simple separation of the initial template motifs (as are seen in algorithms such as support vector machines or equivalent). Understanding where these boundaries fall likely requires knowledge of how natural stimuli is distributed in the latent space of the network, and the underlying geometry in which the latent manifold is embedded. Additional work is needed in these areas.

Our ability to predict held out behavioral responses from both the artificial neural network activations as well as the in vivo neural population activities indicate that each of these representations are sufficient, although not uniquely so, to describe the perceptual and behavioral space. One compelling result is that the behavior generalizes across this stimulus space, such that that knowing how perception acts in one sub-region can inform behavioral responses in other sub-regions. This category generalization also deserves further study.

The network representation decoding of the psychophysical parameters indicates the importance of the distribution of natural songs on the perceptual space. While our library of songs provides for an approximation of their natural experience history, it is limited, especially in the context of wild caught animals. A fascinating future direction of this work would be to see if changing the distribution of experienced song before the training would influence category boundaries.

Finally, the fact that categorical behavioral responses can be decoded from a randomly selected set of 10s of neurons contributes to a growing body of work (Jeanne et al., 2011; Kozlov & Gentner, 2016) that opposes the strongest version of sparse hierarchical models of perception, where neurons with simpler receptive fields converge onto neurons with a more complex receptive fields until a complex percept, like Jennifer Aniston emerges (Quiroga et al., 2005). Under this model, decoding a categorical behavioral measure (as we show here) is only possible by matching the right stimuli to the right subset of neurons. Thus, our results imply that *the representation of these secondary auditory regions is much more distributed than would be predicted by a model where increasingly complex features are encoded exclusively by single neurons*.

Acknowledgments

Work supported by NIMH T32 MH020002-16A, UCSD Frontiers of Innovation Scholars Program Fellowship, NIH R56DC016408, and NSF GRF 2017216247.

References

Gentner, T. (2004). Neural systems for individual song recognition in adult birds. *Annals of the New York Academy of Sciences*, *1016*(1), 282–302.

Gentner, T. Q., & Margoliash, D. (2003). Neuronal populations and single cells representing learned auditory objects. *Nature*, *424*(6949), 669–674.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Jeanne, J. M., Thompson, J. V., Sharpee, T. O., & Gentner, T. Q. (2011). Emergence of learned categorical representations within an auditory forebrain circuit. *The Journal of Neuroscience*, *31*(7), 2595–2606.

Kozlov, A. S., & Gentner, T. Q. (2016). Central auditory neurons have composite receptive fields. *Proceedings of the National Academy of Sciences*, 201506903.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102.

Thompson, J. V., & Gentner, T. Q. (2010). Song recognition learning and stimulus-specific weakening of neural responses in the avian auditory forebrain. *Journal of neurophysiology*, *103*(4), 1785–1797.