UNIVERSITY OF CALIFORNIA SAN DIEGO

The Unreasonable Effectiveness of Machine Learning in Neuroscience: Understanding High-dimensional Neural Representations with Realistic Synthetic Stimuli

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Neurosciences with a Specialization in Computational Neurosciences

by

Marvin Thielk

Committee in charge:

Professor Tim Gentner, Chair
Professor Saket Navlakha
Professor Terrence Sejnowski
Professor John Serences
Professor Tatyana Sharpee

2019

The dissertation of Marvin Thielk is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

DEDICATION

To Alex, for reminding me not to sacrifice too much of my physical, mental, and emotional well-being as I tried to find out how deep the rabbit hole goes.

And to my family, especially my parents, who taught me to focus on three things: Do good. Work hard. And never stop learning and improving.

“Always take time to sharpen your axe.”

EPIGRAPH

If I have seen further it is by standing on the shoulders of giants.

—Isaac Newton

There is only one thing which is more unreasonable than the unreasonable effectiveness of mathematics in physics, and this is the unreasonable ineffectiveness of mathematics in biology.

—Israel Gelfand

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Thanks to the universe, without which, this probably wouldn't exist.

Thanks to Alex, for telling me if what I'm generating is real or not.

Thanks to my parents, Mike and Minh, for making me the man I am today.

Thanks to my brothers, Aaron and Alex, for reminding me that the people who care don't matter and the people who matter don't care. RL is much better in co-op.

Thanks to my PI, Tim, for years of both scientific and non-scientific guidance, and for sometimes knowing what I needed more than I did, like when he came into lab at 11:30 PM on Friday night to help me without me having to ask. I am the scientist I am today because of him.

Thanks to my lab. Dan, Justin, and Krista, for being my role models and creating the environment that made me join the lab. Brad, for being the opposite of me in a lot of ways, but the same as me in all the best ways. Tim, for noticing I say wavenet at every presentation. Michael, for doing all the things I've never even tried. Zeke, for being the postdoc I never knew I needed. My undergrads, Karen, Kevin, Mingcheng, and Darvesh, for helping me and reminding me how much I enjoy mentorship. Nasim, Sasen, Kai, Srihita, Anna, Daril, and Sean, for being a constant source of encouragement and reminding me of all the progress I've made during grad school.

Thanks to my cohort, especially Aki, Andrea, Cailey, Claire, Ethan, Geoff, Geoff, Kathleen, Kyle, Landon, Laura, Matt, Melissa, Sequoyah, Stephen, Tom, and Wilmer, for making UCSD NGP the best and getting me into the right amount of trouble and knowing when to "just let him go." I'll never forget running from helicopter spotlight with you guys.

Thanks to my past-life friends, Allic, Michael, Yuliy, Hernan, Will, Bryan, and Cam, for bursting my bubble and knowing I'm an introvert pretending to be an extrovert.

Thanks to Erin and Linh, for solving all the problems I caused.

Finally, thanks to my co-PI, Tanya, for providing theoretical spice to my work, as well as the rest of my committee, Terry, John, and Saket, for their guidance and providing reference to my work.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Thielk, Marvin; Sainburg, Timothy; Sharpee, Tatyana; Gentner, Timothy. The dissertation author was the primary investigator and author of this manuscript

VITA

2012 B. A. with majors in Applied Mathematics *with honors* and Computer Science *with honors*, University of California, Berkeley

2019 Ph. D. in Neurosciences with a Specialization in Computational Neurosciences, University of California, San Diego

ABSTRACT OF THE DISSERTATION

The Unreasonable Effectiveness of Machine Learning in Neuroscience: Understanding High-dimensional Neural Representations with Realistic Synthetic Stimuli

by

Marvin Thielk

Doctor of Philosophy in Neurosciences with a Specialization in Computational Neurosciences

University of California San Diego, 2019

Professor Tim Gentner, Chair

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. We record simultaneous firing from populations of 10's of neurons in a secondary auditory cortical region of anesthetized starlings. By estimating how fast population level neural representation change with respect to the stimuli, we produce a measure along a path in stimuli space that is shared between birds and descriptive of the psychophysically determined parameters in other birds. Consistent with this, we predict the behavioral psychometric function along one dimension by fitting the behavior for other dimensions to the population level neural activity. Thus, knowing how the animal responds in one sub-region of the parametrized space informs responses in other sub-regions. 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. This thesis also explores other applications of machine learning to solve neuroscience problems, in particular, the curse of dimensionality and exploring predictive coding and surprise. A model explicitly designed to predict future states allows the compression of high-dimensional time-varying signals into a lower-dimensional representation encoding exclusively predictive and predictable information and has many practical applications.

Chapter 1

Introduction

1.1 What is categorical perception?

Categorical Perception (CP) is a phenomenon that occurs when categories held by the observer, either learned or inherent, influence the observers perception. This means that the perception of changes in the stimuli does not depend solely on the physical changes in the stimuli. Small changes in stimuli that lie near or across category boundaries are very noticeable while more substantial changes in other regions may not be noticeable at all. This means that our perceptual systems transform relatively linear sensory signals into relatively nonlinear internal representations.

1.2 What is required for Categorical Perception research?

The fundamental difficulty with Categorical Perception research is that it generally occurs in the perception of a high dimensional space where naturally occurring ecologically relevant stimuli exist only on non-uniformly distributed sub-manifold of that space. In terms of auditory Categorical Perception in human speech, the high dimensional space is the space of all possible

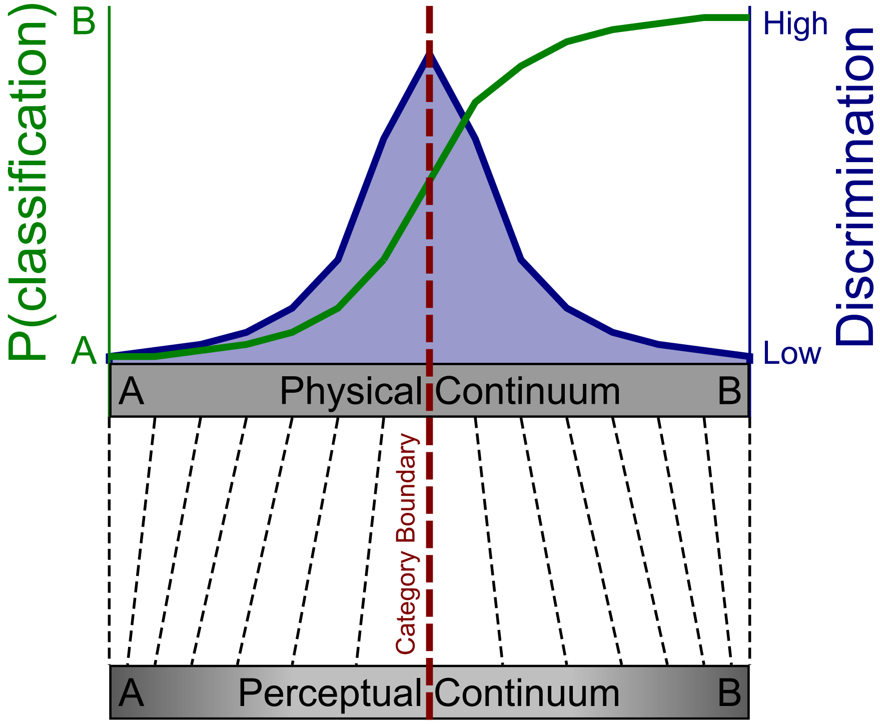


Figure 1.1: *A schematic of idealized categorical perception. Top green*: The psychometric function or the probability of classifying a stimulus as A or B as you move along a physical continuum between A and B. The category boundary occurs at the vertical red dotted line and there is some amount of uncertainty near this boundary. The boundary is defined as the point of perceived equality or the stimuli location on the physical continuum where a participant is equally likely to classify the stimuli as A or B. *Top Blue*: is the discrimination curve describing how well a participant can discriminate between nearby stimuli. Categorical Perception has been described as a peak in the discrimination curve near the category boundary. The lower axis indicates how the perceptual space is stretched compared to the physical continuum causing stimuli that are within category limits to be perceived as closer than stimuli that lie across category boundaries but are an equal distance apart.

sounds that could possibly be heard, most of which would sound like random noise. The naturally occurring ecologically relevant stimuli are all the possible spoken sounds or phonemes that are useful for human speech production and comprehension, which exist in a much smaller lower-dimensional subset of all possible sounds. Even within this subset of sounds, not all possible phonemes are heard with equal probability, and this distributional information may begin to shape our perceptual systems even before birth.

Because of this, the majority of auditory Categorical Perception has focused on human speech sounds, predominantly English, leveraging our intuitions and years of linguistic modeling to identify the relevant dimensions where Categorical Perception occur. Categorical Perception work concerning animal vocalizations must first identify stimuli that are categorically perceived, a process that was very labor intensive [Nelson and Marler, 1989, Prather et al., 2009, Lachlan and Nowicki, 2015] before the development of more modern techniques [Sainburg et al., sion]. The variation within the songs of European starlings also makes this task easier because clustering song elements is more straightforward with more distinct clusters.

Furthermore, the identification of category boundaries is more difficult in non-human research than with human subjects. Since we can't provide instructions, we are forced to use multiple shaping or pre-training stages of operant behavioral conditioning. Also, if we permit a fully adaptive double staircase procedure, our subjects learn to exploit this and force the boundary all the way to one side or another. To solve this, we use a new staircase technique.

Last, but certainly not least, Categorical Perception requires the generation of a synthetic continuum between the categories or category exemplars. In human speech, the existence of speech synthesizers has made this task easy, at least in English. In Chinese for example, the technical difficulty in creating a Chinese-based synthetic continuum suitable for a Categorical Perception study has been proposed as a reason for the large number of papers focused on Categorical Perception in lexical tones in the Chinese language [Zhang, 2013]. Categorical Perception research on animal vocalization has been limited to exploiting natural variation in recorded calls [Nelson and Marler, 1989, Prather et al., 2009, Lachlan and Nowicki, 2015] which invariably limits the sampling and control over the stimuli. It has also been shown that the quality of the speech synthesizer is correlated with how much Categorical Perception is observed[Van Hessen and Schouten, 1999].

Thus the approach we present streamlines the following steps for Categorical Perception research in non-human vocalizations:

1. Identification of relevant dimensions where Categorical Perception occurs

2. Identification of categorical boundaries

3. Generation of a synthetic continuum between the categories/ category exemplars.

1.3 History of Categorical Perception

There is a long history of categorical perception research. The invention of the formant speech synthesizer known as the Pattern Playback at Haskins Laboratories [Cooper et al., 1951] made the discovery of Categorical Perception possible. The Pattern Playback was a machine that converted pictures of the acoustic patterns of speech in the form of a spectrogram into sound using light.

Categorical Perception was initially described as the absolute definition of completely quantal perception where subjects would be unable to discriminate differences to stimuli within a category, only noticing differences in stimuli that lies across category boundaries [Liberman et al., 1957, Studdert-Kennedy et al., 1970]. Although only ever described as an idealized case, this definition is far too restrictive and has been disproven many times; however, it has caused much criticism of the work done in the Categorical Perception field with critics calling for “The end of categorical perception as we know it” [Schouten et al., 2003] etc.

Later work has instead demonstrated that with-category differences are discriminable

[Pisoni and Tash, 1974, Carney et al., 1977, Massaro and Cohen, 1983] and meaningful [Miller, 1997, McMurray et al., 2002, McMurray et al., 2008]. Thus it seems there is just an increase in sensitivity to differences in stimuli that lie across category boundaries.

Categorical Perception first explored in 1957 concerning the discrimination of speech sounds within and across phenome boundaries [Liberman et al., 1957]. In addition to dubbing the term "Categorical Perception," Liberman suggested that Categorical Perception was unique to human speech, and furthermore, was what made human speech special. He also originally proposed the motor theory of speech perception [Liberman et al., 1967].

Formerly thought to be peculiar to speech and color perception, Categorical Perception turns out to be far more general and may be a result of the way our neural circuits are wired to identify hierarchical patterns.

Categorical Perception also provides an intermediate explanation of equivalence classes which in turn form the basis of high-level symbolic thought and auditory Categorical Perception also forms the basis for human language.

Auditory Categorical Perception was previously thought to be strictly human trait possibly depending on special processing or that there might be some motor theory basis that could inform perception based on a motor production model. It is now thought to be a ubiquitous feature of all perceptual systems.

It has been shown in chinchilla [Kuhl and Miller, 1975] and later in macaques [Kuhl and Padden, 1983], that animal models also have enhanced discriminability at human phonetic boundaries without a phylogenetic history of phenomic knowledge, either acoustic or articulatory. These studies consist of playing human speech sounds to animals and behaviorally measuring if they place the same boundaries we do. This has been used to argue that human languages may have adapted to use phenome category boundaries located in regions with intrinsically higher discriminability [Stevens, 1981, Goldstone and Hendrickson, 2010]

There also exists a host of research on Categorical Perception in non-human vocal communication sounds. These studies either use natural variation in the vocalizations artificially generated hand drawn spectrogram morphs. These studies are only possible after considerable background work establishes the relevant categories, similar to the techniques originally used in humans to identify Categorical Perception.

Categorical boundary location preference has also been explored, mostly in the context of human speech. Early explanations included the motor theory of speech perception, which suggested that the explanation for Categorical Perception lay in the anatomy of speech production, specifically, physical differences during articulation [Liberman et al., 1967]. There have also been arguments made that a biological predisposition influenced the development of human languages pushing phenome boundaries to locations of increased sensitivity [Stevens, 1981, Goldstone and Hendrickson, 2010, Halle and Stevens, 1979]

1.4 Learned Categorical Perception

However, we also know that the boundary location is experience dependent. Native language exposure can shift boundary locations. Also, a sound difference that crosses a boundary of a language is more discriminable to speakers of that language than to speakers of a language that doesnt have a boundary in that location.

1.5 Our Contribution

We use machine learning techniques to approximate unsupervised statistical learning of the distribution of the set of natural motifs while compressing it into a lower dimensional representation. This allows us to generate a continuum of synthetic motifs that lie between exemplars.

These types of machine learning methods allow for the creation of synthetic continua between exemplars when trained upon a corpus of that stimuli type. This allows for the exploration of natural ecologically relevant animal communication signals as we have done here. It also would allow the exploration of categorical perception in non-English languages. For example, technical difficulty in creating the Chinese-based synthetic continuum suitable for a Categorical Perception study has been proposed as a reason for the large number of papers focused on Categorical Perception in lexical tones in the Chinese language [Zhang, 2013].

Furthermore, stimulus naturalness has been shown to be a factor determining the degree of categorical perception[Van Hessen and Schouten, 1999], so as these machine learning techniques improve, we may be able to better explore the Categorical Perception phenomena.

Our method allows for the creation or discovery of boundaries of Categorical Perception in an unsupervised way in animal vocalizations and are thus more ecologically relevant and activate auditory regions that are specifically tuned to conspecific vocalizations.

Chapter 2

Categorical Perception

Synthetic vocalizations reveal a shared perceptual space predicted by relationships in neural population activity shared across birds and recording sites

2.1 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. We record simultaneous firing from populations of 10's of neurons in a secondary auditory cortical region of anesthetized starlings. By estimating how fast population level neural representation change with respect to the stimuli, we produce a measure along a path in stimuli space that is shared between birds and

descriptive of the psychophysically determined parameters in other birds. Consistent with this, we predict the behavioral psychometric function along one dimension by fitting the behavior for other dimensions to the population level neural activity. Thus, knowing how the animal responds in one sub-region of the parametrized space informs responses in other sub-regions. 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.

2.2 Introduction

Categorical Perception (CP) is the phenomenon where categories held by the observer, either learned or inherent, influence the observer's perception. This means that the perception of changes in the stimuli does not depend solely on the physical changes in the stimuli. Small changes in stimuli that move near or across category boundaries are very noticeable while more substantial changes in other regions may not be noticeable at all. This means that our perceptual system transforms relatively linear sensory signals into relatively nonlinear interval representations.

Experimentally we observe Categorical Perception as demonstrated in figure (2.1A). For a defined physical continuum that moves between two (or more) categories, A and B, for example, we can experimentally ask a participant to classify the stimuli as belonging to category A or category B. This provides the sigmoid-shaped psychometric curve shown in green in figure (2.1A) where the probability of classifying a stimulus as category B is high when the stimulus is near the B category and low when the stimulus is near the A category. Another Categorical Perception hallmark is a peak in the discrimination function near the boundary of two categories. This is tested by experimentally asking a participant to distinguish nearby stimuli on the physical continuum, and is shown in blue in figure (2.1A). Thus, the physical continuum is perceived as being stretched near the category boundary and compressed within category limits.

The conventional view of the auditory object processing pathway (Ventral Auditory Pathway) is very similar to the visual object processing pathway (Ventral Visual Pathway). The ventral auditory pathway consists of the hierarchical processing of categorical information where neurons become increasingly sensitive to more complex stimuli and abstract information between the beginning stages and the latter stages. It is therefore theorized that as you move along the ventral auditory pathway, there is a progression of category information processing, where simpler receptive fields are combined to form more and more complex receptive fields encoding more and more complex features of the stimuli until you reach the auditory equivalent of a "grandmother cell." Perhaps not quite as extreme as proposing the existence of the auditory equivalent of a "grandmother cell," there are regions such as the superior temporal gyrus and sulcus which is categorically activated by speech sounds, relative to other sounds [Binder et al., 2000, Leaver and Rauschecker, 2010] and there are distinct regions of superior temporal gyrus and sulcus that are selectively activated by musical instruments sounds [Leaver and Rauschecker, 2010] and even tool sounds and birdsong [Doehrmann et al., 2008].

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 [Gentner and 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 and caudal mesopallium [Gentner and Margoliash, 2003, Gentner, 2004, Thompson and Gentner, 2010a, Jeanne et al., 2011a].

Starlings rely on object-level representations of motifs [Meliza et al., 2010, Gentner and Hulse, 1998, Comins and Gentner, 2013, Comins and Gentner, 2014a], and the neural correlates of motif perception are localized in the regions of the auditory forebrain *that are evolutionarily homologous to the mammalian auditory cortex[Wang et al., 2010].* The songbird auditory system follows the vertebrate plan[Webster, 1992]. Field L2a is the primary telencephalic target of the auditory thalamus[Karten, 1968], and is the input layer for a columnar circuit spanning L1, L3, and caudal mesopallium, anatomically[Wang and Karten, 2010, Jarvis et al., 2005], genetically[Dugas-Ford et al., 2012], and functionally[Calabrese and Woolley, 2015] analogous to the mammalian auditory cortical micro-circuit. Field L sub-regions also project to the caudomedial nidopallium, a region that, along with the caudaolateral mesopallium, shares reciprocal connections with the caudaomedial mesopallium. Neurons in caudomedial nidopallium have been shown to have complex composite receptive fields[Kozlov and Gentner, 2016] that are reminiscent of the multidimensional receptive fields found in cat A1[Atencio et al., 2008]. Multidimensional receptive field characteristics can be reproduced by a Deep Neural Network[Kozlov and Gentner, 2016] trained to represent starling songs. Neurons throughout the avian cortex are selective for species-specific (conspecific) vocalizations[Bonke et al., 1979, Leppelsack and Vogt, 1976, Muller and Leppelsack, 1985], with selectivity increasing from Field L2, to L1 and L3[Theunissen et al., 2004, Theunissen and Doupe, 1998, Theunissen et al., 2000], and again in caudomedial nidopallium and caudal mesopallium[Calabrese and Woolley, 2015, Muller and Leppelsack, 1985, Grace et al., 2003, Bonke et al., 1979, Leppelsack and Vogt, 1976, Gentner and Margoliash, 2003, Gentner et al., 2004, Thompson and Gentner, 2010b, Jeanne et al., 2011b]. This is consistent with a functional hierarchy tuned to conspecific song[Hsu et al., 2004, Woolley et al., 2005], refined by experience[Gentner and Margoliash, 2003, Sockman et al., 2002, Sockman et al., 2005, Phan et al., 2006, Thompson and Gentner, 2010b, Jeanne et al., 2011b], with selectivity for acoustic features at multiple timescales increasing in a feed-forward manner[Rose et al., 1988, Kaas and Hackett, 2000, Binder et al., 2000, Van Essen et al., 1992]. Because of its compact organization, close relation to well-studied sensorimotor/vocal production structures, and the rich behavioral history of birdsong, the system is ideal for understanding the processing of natural acoustic communication signals. Starlings respond to natural complex stimuli patterns such as the presence or absence of a center-embedded recursive structure[Gentner et al., 2006, Comins and Gentner, 2014a, Comins and Gentner, 2014b], a characteristic previously thought unique to human language (at least by Chomsky), allowing access to signals that would be invasive in humans, and cost prohibitive in non-human primates.

2.2.1 The difficulties of using European Starlings

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 and extensively characterize both the information that these regions encode, and how this information is encoded. 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 wave stimuli that can drive more primary sensory regions.

Furthermore, starlings are not among the most popular model organisms, so many of the off the shelf tools available for other model organisms haven't been established for the Starlings. This includes the genetic and viral tools as well as off the shelf hardware and tools.

Lastly, increased mobility (3-dimensional freedom and frequent backflips) makes recordings while behaving more difficult, even though there have been examples of it[Knudsen and Gentner, 2013, Bluvas and Gentner, 2013].

2.2.2 What is required for Categorical Perception research?

The fundamental difficulty with Categorical Perception research is that it generally occurs in the perception of a high dimensional space where naturally occurring ecologically relevant stimuli exist only on non-uniformly distributed sub-manifold of that space. In terms of auditory Categorical Perception in human speech, the high dimensional space is the space of all possible sounds that could possibly be heard, most of which would sound like random noise. The naturally occurring ecologically relevant stimuli are all the possible spoken sounds or phonemes that are useful for human speech production and comprehension, which exist in a much smaller lower-dimensional subset of all possible sounds. Even within this subset of sounds, not all possible phonemes are heard with equal probability, and this distributional information may begin to shape our perceptual systems even before birth.

Because of this, the majority of auditory Categorical Perception has focused on human speech sounds, predominantly English, leveraging our intuitions and years of linguistic modeling to identify the relevant dimensions where Categorical Perception occur. Categorical Perception work concerning animal vocalizations must first identify stimuli that are categorically perceived, a process that was very labor intensive[Nelson and Marler, 1989, Prather et al., 2009, Lachlan and Nowicki, 2015] before the development of more modern techniques[Sainburg et al., sion]. The variation within the songs of European starlings also makes this task easier because clustering song elements is more straightforward with more distinct clusters.

Furthermore, the identification of category boundaries is more difficult in non-human research than with human subjects. Since we can't provide instructions, we are forced to use multiple shaping or pre-training stages of operant behavioral conditioning. Also, if we permit a fully adaptive double staircase procedure, our subjects learn to exploit this and force the boundary all the way to one side or another. To solve this, we use a new staircase technique.

Last, but certainly not least, Categorical Perception requires the generation of a synthetic continuum between the categories or category exemplars. In human speech, the existence of speech synthesizers has made this task easy, at least in English. In Chinese for example, the technical difficulty in creating a Chinese-based synthetic continuum suitable for a Categorical Perception study has been proposed as a reason for the large number of papers focused on Categorical Perception in lexical tones in the Chinese language[Zhang, 2013]. Categorical Perception research on animal vocalization has been limited to exploiting natural variation in recorded calls[Nelson and Marler, 1989, Prather et al., 2009, Lachlan and Nowicki, 2015] which invariably limits the sampling and control over the stimuli. It has also been shown that the quality of the speech synthesizer is correlated with how much Categorical Perception is observed[Van Hessen and Schouten, 1999].

Thus the approach we present streamlines the following steps for Categorical Perception research in non-human vocalizations:

1. Identification of relevant dimensions where Categorical Perception occurs

2. Identification of categorical boundaries

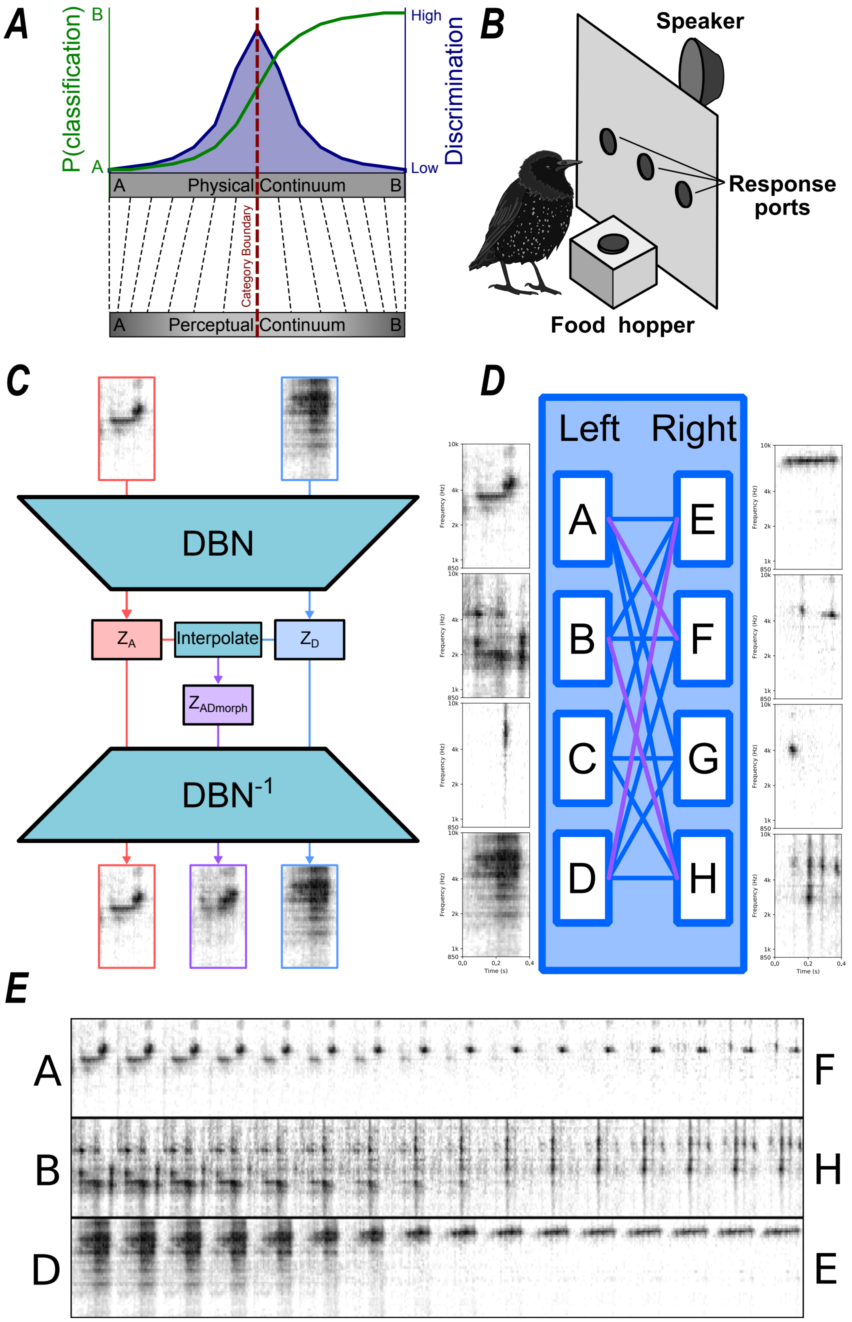
3. Generation of a synthetic continuum between the categories/ category exemplars.

2.3 Results

2.3.1 Generating smoothly varying morphs

Using a large corpus of recorded starling song, a generative machine learning model is trained to non-linearly autoencode 400 mS motifs (song segments) of starling song in a low (64) dimensional latent space. The generative model attempts to model the distribution of song segments the model is trained on, and as a result, it tries to completely fill the latent space with projected representations due to an information bottleneck. The consequence of this is that for any point in the latent space, the spectrogram reconstruction of that point could lie within the actual distribution of starling song. Thus, the model can be used to interpolate between any two

Figure 2.1: *Task outline* (A) A schematic of idealized categorical perception. *Top green*: The psychometric function or the probability of classifying a stimuli as A or B as you move along a physical continuum between A and B. The category boundary occurs at the vertical red dotted line and there is some amount of uncertainty near this boundary. The boundary is defined as the point of perceived equality or the stimuli location on the physical continuum where a participant is equally likely to classify the stimuli as A or B. *Top Blue*: is the discrimination curve describing how well a participant can discriminate between nearby stimuli. Categorical Perception has been described as a peak in the discrimination curve near the category boundary. The lower axis indicates how the perceptual space is stretched compared to the physical continuum causing stimuli that are within category limits to be perceived as closer than stimuli that lie across category boundaries, but are an equal distance apart. (B) A diagram of the operant behavioral apparatus which consists of response ports with IR beak detection sensors, a food hopper to provide access to a food reward, and a speaker hidden behind the panel to present auditory stimuli. (C) A diagram of generating an interpolating morph using a Deep Belief Network autoencoder. A spectrogram representation of two 400 ms song motif, A and D, are fed into a compressive network to create a latent representation, $Z_A$and Z_D. We then linearly interpolate between $Z_A$and Z_D to create $Z_{ADmorph}$ which we can use to reconstruct a spectrogram motif that lies between A and D. Three example morph interpolations are shown below in (E) (D) Diagram of the behavioral task. Spectrograms of the initial 8 randomly chosen 400 ms long motifs, labeled A-H, and their reward associated responses. Once the performance on these 8 reached a sufficient stable level, interpolated morph motifs, indicated by the 16 connecting lines were probed using a ratcheting double staircase procedure to allow the birds to determine their own behavioral boundaries.The 3 example morph dimensions displayed to the left are highlighted in blue. (E) Three example interpolated morph dimensions generated using the Deep Belief Network. 16 (of 128 used) example motifs for each morph dimension. Spectrogram representation with frequency on y-axis and time on the x-axis. Each motif is 400 ms long.


A
P(classification)
B
A
High
Low
Discrimination
Physical Continuum
A
B
Category Boundary
Perceptual Continuum
A
B
B
Speaker
Response
ports
Food hopper
C
DBN
Z_A
Interpolate
Z_D
$Z_{ADmorph}$
DBN^{-1}
D
Left
Right
A
B
C
D
E
F
G
H
Frequency (Hz)
10k
4k
2k
1k
850
0,0
0,2
0,4
Time (s)
E
A
F
B
H
D
E

arbitrarily chosen starling motifs projected into the latent space to produce a smoothly varying continuum of morphed motifs that shift from one target motif to the other as outlined in 2.1(C). Instead of sounding like a simple linear crossfade between the two motifs, the network tries to produce a motif that could be in the distribution of recorded motifs, resulting in more realistic sounding motifs. Three examples of these smooth morphs are shown in 2.1(E)

2.3.2 Behavioral Measurement of Perceptual Space

Using an operant behavioral apparatus diagrammed in 2.1(B), 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 as summarized in figure 2.1(D). After training to a stable performance criterion, the interpolated morph motifs generated by the Deep Belief Network 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 us to iteratively (and independently) estimate the categorical boundary along each of the 16 morph dimensions for each bird.

2.3.3 Psychometric curves are conserved across subjects

This provides independent binary behavioral responses along each of the 16 possible dimensions interpolating from each of the 4 left associated motifs to each of the right associated motifs.

We fit a ***psychometric curve*** of the form $P(R) = A + \frac{K-A}{1+e^{-B(x-M)}}$ using maximum log likelihood for each of the 16 morph dimensions. A and K determine the maximum accuracy achieved near each endpoint. M determines the boundary location or the point of subjective equality between each of the endpoints. B determines boundary sensitivity or slope at the boundary.

Figure 2.2: *Psychometric curves are conserved among different birds* (A) Construction of a single psychometric curve: The x-axis represents presentations of stimuli that vary between motif A on the left and motif E on the right. The middle tick mark is the location of a stimuli that lies exactly between motif A and motif E (according to the Deep Belief Network). Thus the 3 tick marks indicate the location of stimuli that is (left) 75% A, 25% E, (middle) 50% A, 50% E, and (right) 25% A, 75% E. The y-axis is the probability of a right response (that has been operantly conditioned to be associated with motif E). The black dots represent the binary decision of the birds (responses are jittered to demonstrate relative density). All the responses ordered by morph position are binned into 16 equally sized bins and the 95% confidence intervals of the mean estimates are plotted as vertical blue lines. The maximum-likelihood 4 parameter logistic fit of the probability of the bird responding right as a function of morph position between motif A and motif E, which we will call the psychometric curve, is plotted in blue. (B) Psychometric curve variation: The set of all 16 behaviorally determined psychometric curves for a single bird. Color indicates each of the 16 different morph dimensions. (C-D) Exemplar psychometric curves that are conserved between subjects: 2 of the 16 morph dimensions, (C) motif B to motif F and (D) motif A to motif H, are plotted for 4 different birds trained on the same categorization task. The full 16 dimensions are plotted in supplemental figure 2.10. (E) To measure how the B parameter (determines boundary sensitivity or slope at the boundary) of the psychometric curves is grouped we take the one-sided KolmogorovSmirnov metric between the distribution of pairwise distances between the B parameter of all psychometric curves, to the distribution of pairwise distances between the B parameter of psychometric curves from the same morph dimension (blue), and to the distribution of pairwise distances between the B parameter of psychometric curves from the same bird (red). The vertical blue line represents the KolmogorovSmirnov statistic between the distribution of pairwise distances between psychometric sensitivities that share the same morph dimension to that of all measured psychometric curves. The blue shaded cumulative survival curve is the distribution of expected KolmogorovSmirnov statistic values if we randomly split the psychometric curves into groups that share the same size as the morph dimension groups, or the null distribution as estimated by $2^{17} = 131,072$ shuffles. Thus, the interception of the vertical blue line with the blue shaded cumulative survival curve provides an estimate of the likelihood that the psychometric sensitivities are as close to the psychometric sensitivities of other birds on the same morph dimension by chance, $p = 3E-4$. We use an 8-way bonferroni correction to account for the 2 groupings and 4 parameters we test. * indicates $p < 0.00625$, **, $p < 0.00125$, and ***, $p < 0.000125$. The vertical red line is the KolmogorovSmirnov statistic between the distribution of pairwise distances between psychometric sensitivities from the same bird to the distribution of pairwise distances between all psychometric sensitivities. Its corresponding shuffled null distribution is plotted as the red shaded region providing an estimate of $p = 0.044$. Together, this indicates that psychometric sensitivities are closer than would be expected by chance when grouped by morph dimension, but not when grouped by bird. (continued on following page)

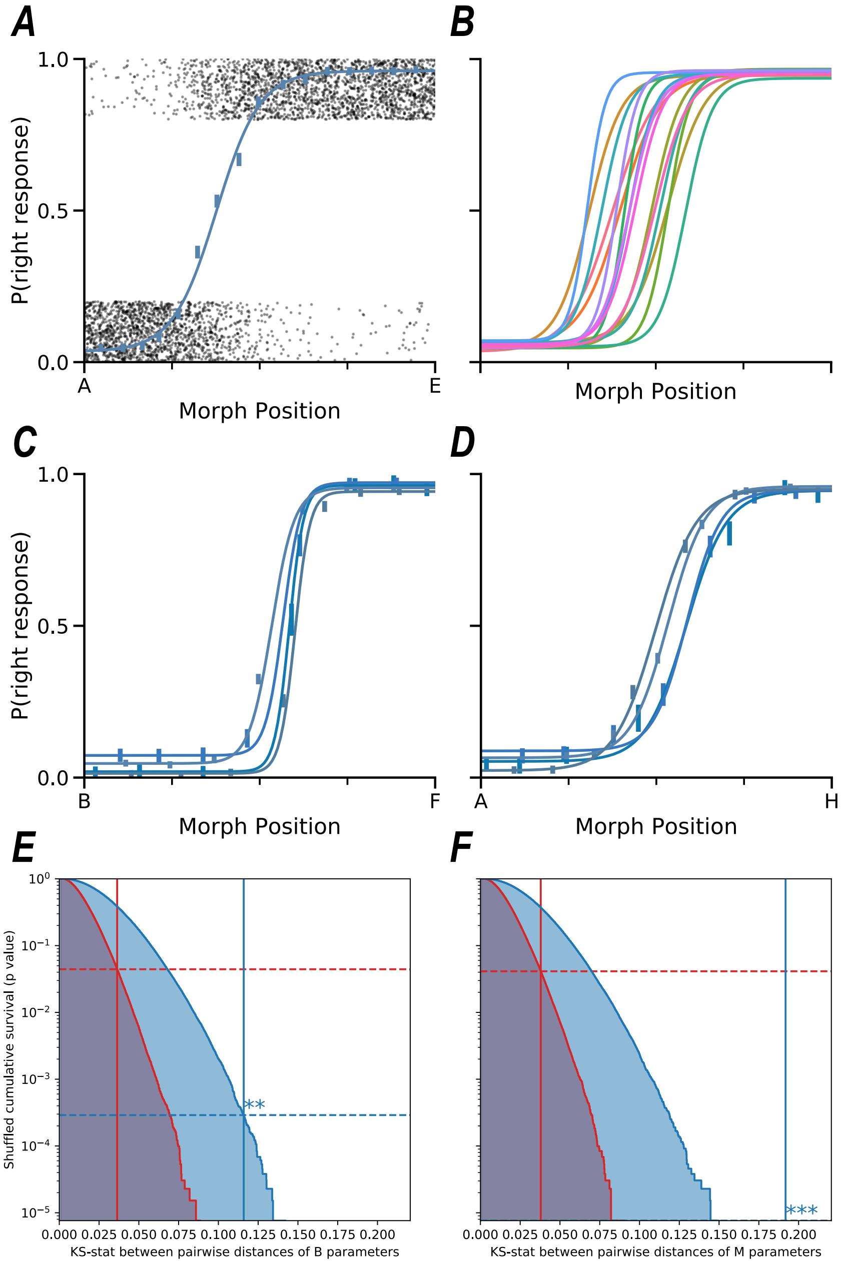


Figure 2.2: *Psychometric curves are conserved among different birds* Continued from previous page. (F) The corresponding figure for the M parameter (boundary location). This shows that the likelihood that the psychometric boundaries are as close to the psychometric boundaries of other birds on the same morph dimension by chance (blue), $p < 8E-6$ because never once in all the 2^{17} shuffles, was the KS statistic as great as the value measured. When distances between psychometric boundaries within a single bird are considered (red), $p = 0.041$. This indicates that psychometric boundaries are also closer than would be expected by chance when grouped by morph dimension, but not when grouped by bird. Parameters A and K, the maximum accuracy achieved near each endpoint, are plotted in supplemental figure 2.11 and show the opposite trend.

A psychometric curve is fit to these behavioral responses and figure 2.2(A) demonstrates how well a four-parameter psychometric curve fits the average response along a single example dimension. We measure 16 of these curves for each bird.

In all cases, birds show very clear categorical perception as evidenced by the steepness each psychometric function regardless of bird 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. 2.2(B)). This variability across dimensions is presumably a result of the non-linear nature of the Deep Belief Network 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.

Despite the significant variability within a single bird across multiple morph dimensions, we observed a remarkable degree of consensus between birds. This included the 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 2.2(C-D) gives an example of the typical agreement between birds, where two of the 16 morph dimensions are plotted for 4 different birds.

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 position between the left associated motif to the right associated motif.

A and *K*, the left and right scaling parameters are statistically closer than would be expected when compared across all morphs within a single bird indicating that they are bird specific parameters as demonstrated by supplementary figure 2.11. This makes sense because they correspond to the bird's absolute performance on the left and right endpoints. Figure 2.2(E-F) also show that M, the category boundary and B, the sensitivity are conserved within the same morph dimension across different birds, but not within a single bird. This indicates the existence of a shared perceptual space of these synthetic natural-like sounds in these wild-caught birds.

2.3.4 Different training context sometimes results in reliable shifts in the psychometric curves

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. The category assignments for the three different cohorts are shown in figure 2.3(A). 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 the same location as shown in figure 2.3(B) while in other interpolating morph dimensions each cohort has a separate boundary (but consistent within that cohort), as shown in figures 2.3(C-D). 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.

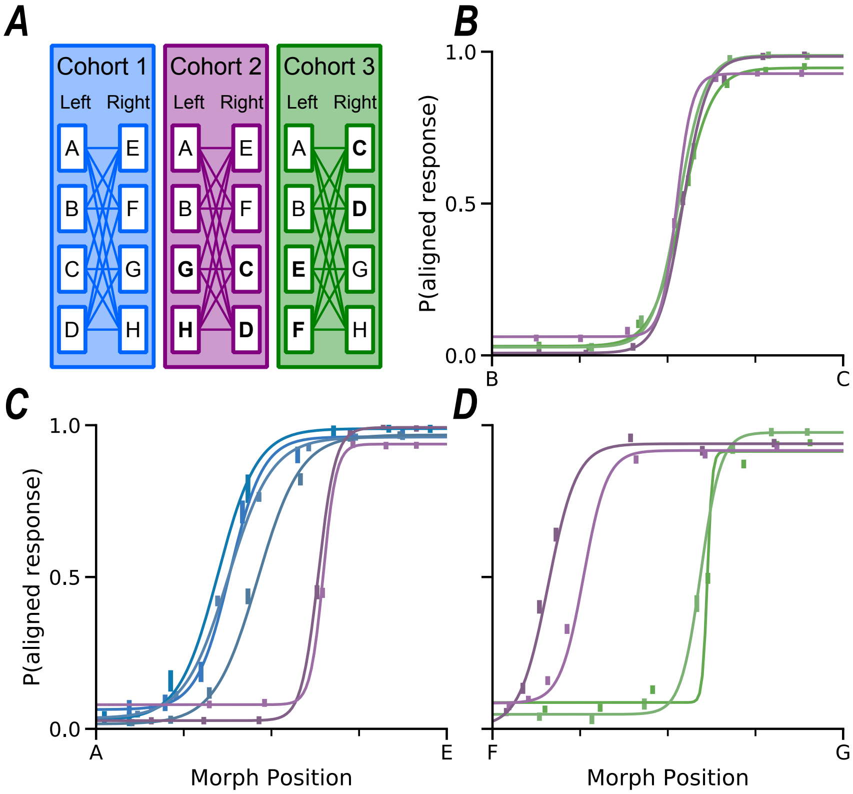


Figure 2.3: *Different training context sometimes results in reliable shifts in the psychometric curves* (A) Diagram of alternative training contexts. Task structure remains the same in each cohort, but some motif category assignments change. Category assignment differences from Cohort 1 are bolded in Cohort 2 and 3. This allows a subset of the morph dimensions to be compared across training on different category permutations. Colors correspond to colors used in adjacent plots. (B-D) Exemplar psychometric curves from subjects trained on different category permutations. (B) On most dimensions the boundary is conserved as if the training was the same as shown in the psychometric curves from these 4 birds, 2 from each of cohort 2 (purple) and cohort 3 (green) along the morph dimension between motif B and motif C. (C-D) However, on some of the dimensions, such as A to E (left) and F to G (right), the behaviorally measured boundaries depend on the initial categorical assignment. When there is a difference, the boundaries are still conserved within cohorts.

2.3.5 Electrophysiological Recordings

To explore the neural underpinnings of categorical perception, we record from secondary auditory regions of anesthetized starlings and present the generated motifs. We record from a total of 2019 neurons in 47 recorded population sites in 8 different starlings, 4 having been trained on the task, 4 naive, never having heard the stimuli before. We use a 32 channel silicon electrode and stereotaxically target caudaomedial mesopallium and caudaolateral mesopallium as described in the methods. We sort the data using MountainSort[Chung et al., 2017].

2.3.6 Single neurons have reliable temporally precise responses

We find the units are extremely time-locked and stimuli specific, as demonstrated by figure 2.4. Because of this we choose to include time in our neural representation by convolving the spike train with a Gaussian($\sigma = 10\,\text{ms}$). This provides us with a continuous estimate of firing rate through time for each trial for every unit. If we look at the average of this representation as we move along a morph dimension, we find that there is a large amount of variation in how the units respond. Some units respond very smoothly to changes along a morph dimension, whereas others are quite categorical.

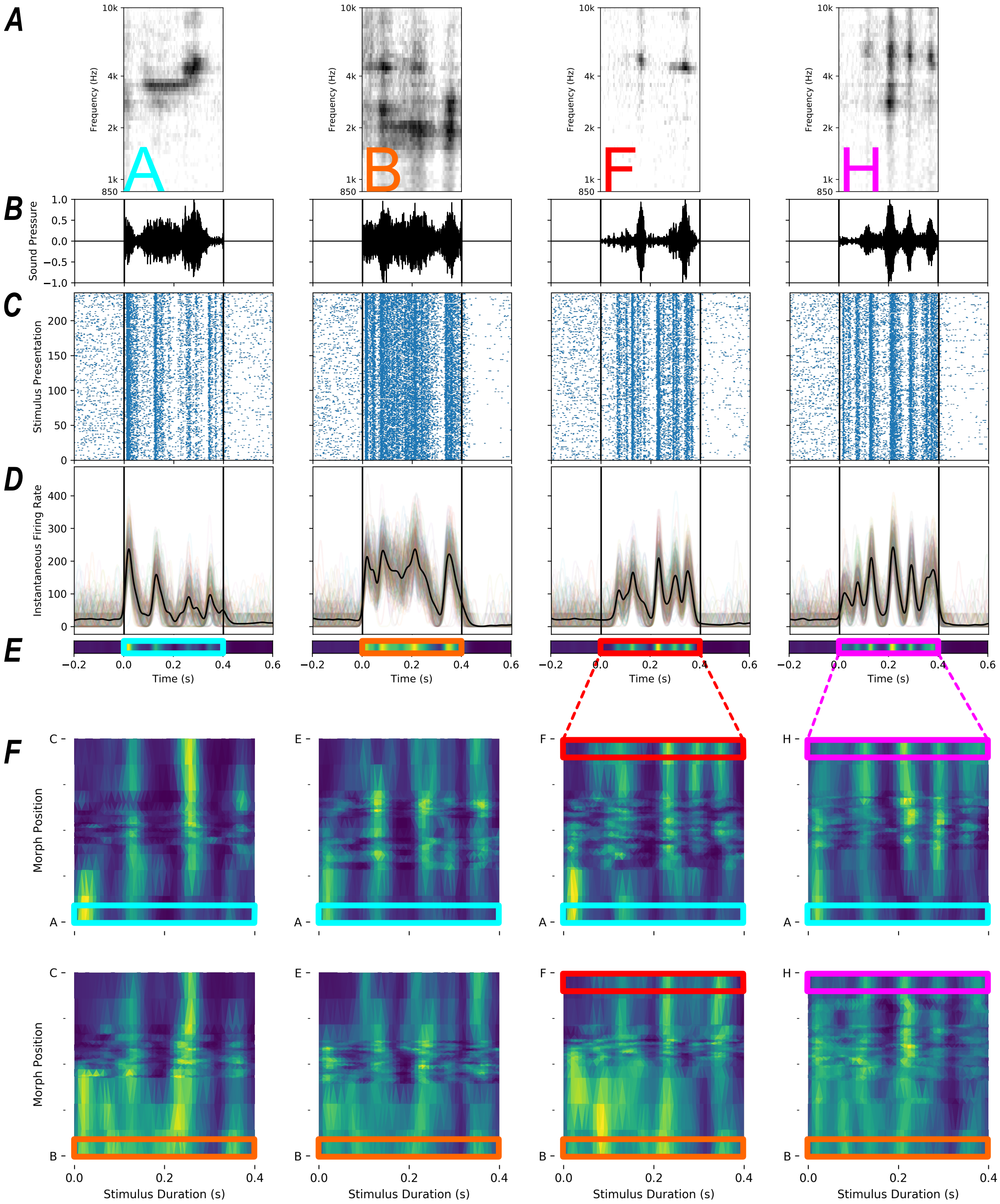
We restrict our analysis to units that can correctly identify the identity of our eight template motifs (averaged in a pairwise fashion so that 50% is chance) more than 60% of the time. This leaves 705 remaining ”behaviorally relevant” units, as seen in figure 2.5.

2.3.7 The Neural Stretching Curve

We are interested in the brain’s ability to discriminate changes along our morph dimension. We use distance in our neural representation to approximate the brain’s ability to discriminate.

We pose the following regression problem by defining the Neural Stretching curve: Given a stimuli representation space S containing $\{s_i\}$ that lie along a 1D path s and a neural

Figure 2.4: *Single neurons have reliable temporally precise responses* (A) Spectrogram representations of example stimuli presented to the anesthetized starlings of four example motifs presented, motifs A, B, F, and H. (B) The equivalent audio waveform pressure representations. Black lines mark the start and end of the auditory stimuli. (C) The raster plots of a single neuron for 240 presentations of each of these stimuli. Vertical marks are plotted at each time point of the occurrence of a spike in the 400 ms stimuli. The y-axis denotes the stimuli presentation. Black lines mark the start and end of the auditory stimuli. (D) Average Gaussian convolved spike train representation in black. We convolve each individual trial with a Gaussian with $\sigma = 10ms$ to get an instantaneous estimate of spike firing rate. These are plotted faintly to demonstrate trial to trial reliability and variance of this representation. The average of these 240 faint lines is plotted in black for each of these stimuli. (E) Heatmap representation of the above representation to be used below. Lighter represents higher estimated instantaneous firing rate. These plots are all normalized to the max firing rate. The colored outlines indicate where the data is repeated below. (F) Variation of the average Gaussian convolved single neurons representation across morph dimensions. Smoothly interpolated using triangulation for surface estimation. Morph position is plotted on the Y axis as the motif presented goes from motif A (highlighted in blue as a band along the bottom) to C (along the top) for example in the top left subplot. Time (during the stimuli presentation) is plotted on the X-axis. The representations of motif F (red) and motif H (Magenta) are diagrammed to demonstrate how the representation is using the average response of this neuron as the top row in the right two subplots in the top row of this representation. Other colors outline other areas that include the averaged representations from figure E. Not all morph positions were presented the same number of times so confidence (not represented) varies along morph position axis.


A
B
C
D
E
F
10k
4k
2k
1k
850
Frequency (Hz)
A
B
F
H
Sound Pressure
1.0
0.5
0.0
−0.5
−1.0
Stimulus Presentation
200
150
100
50
0
Instantaneous Firing Rate
400
300
200
100
0
−0.2
0.0
0.2
0.4
0.6
Time (s)
Morph Position
C
A
E
F
H
B
0.0
0.2
0.4
Stimulus Duration (s)

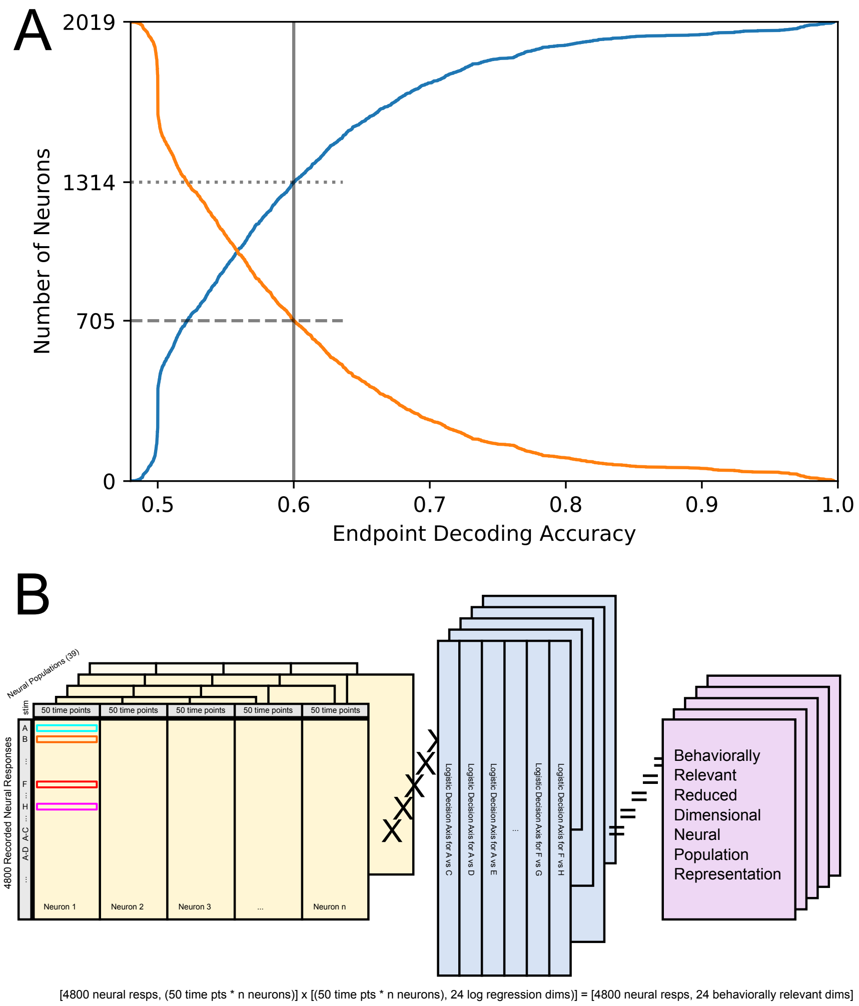
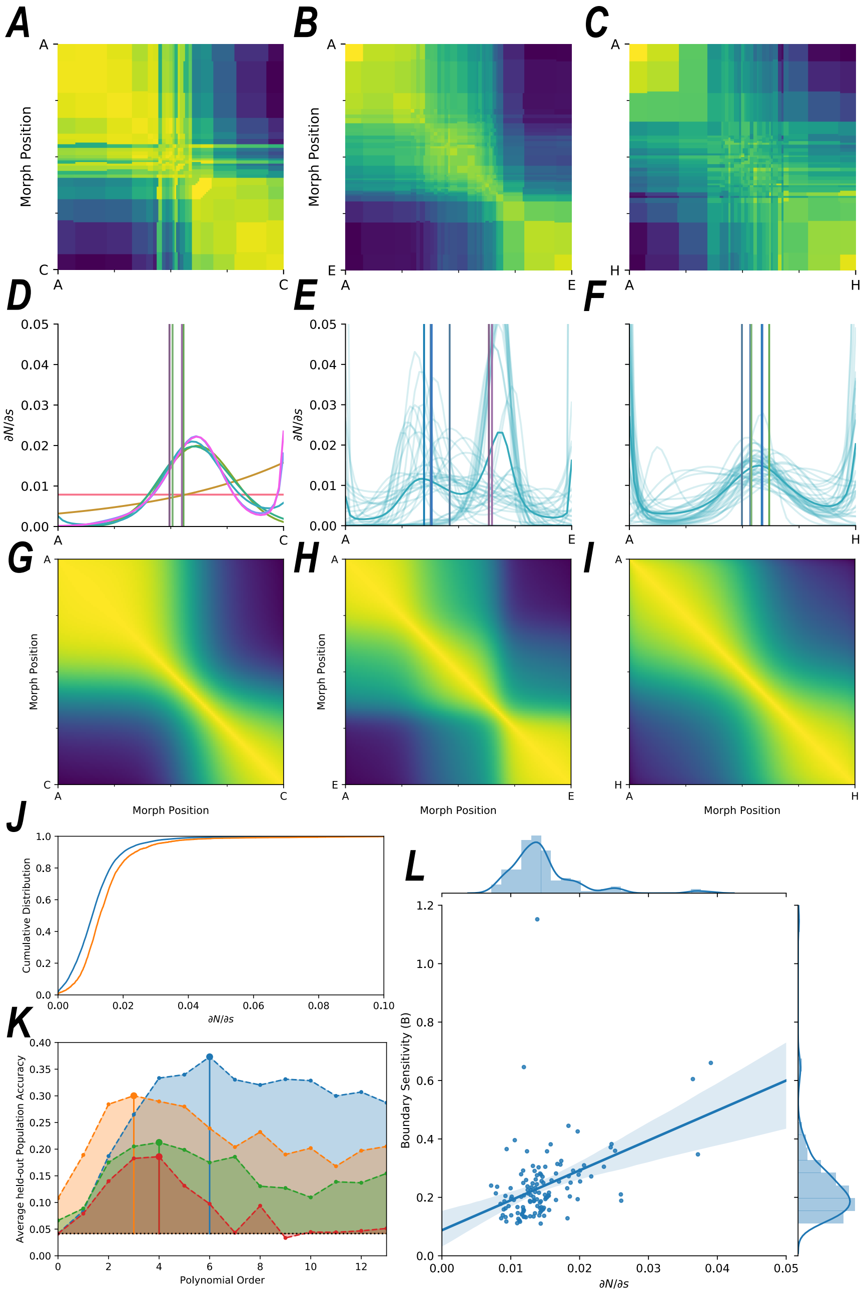


Figure 2.5: *Behaviorally relevant neural representation* (A) The cumulative (blue) and survival (orange) curves for the distribution of units average accuracy on decoding pairs of the endpoint template motifs. The 0.6 accuracy cutoff was chosen mainly for computational tractability reasons. Subsequent analysis only includes activity of the 705 behaviorally relevant neurons with an average accuracy above 0.6. The 1314 neurons with average endpoint decoding accuracy below 0.6 were ignored by subsequent analysis. (B) Diagram of the construction of the behaviorally relevant reduced dimensional neural population representation. We take the recorded neural representations from a population of behaviorally relevant neurons, sampled at 50 time points and concatenated to form a neural representation. We take the logistic decision axis for a logistic regression trained on the endpoints of the 24 morph dimensions to form the columns of a projection matrix which projects the neural population representation into 24 behaviorally relevant dimensions. The colored boxes correspond to the data which would be averaged to form the representations in the previous figure 2.4. This behaviorally relevant neural representations are then used as the input to predict the Neural Stretching curves (figure 2.6) and and to predict the behavioral responses (figure 2.7).

Figure 2.6: *The Neural Stretching curve captures how neural representation changes as we move along a morph dimension* (A) The cross correlation matrix for a single recording on the morph dimension from motif A to motif C. Average cosine distance in the task relevant neural representation subspace between pairs of morph positions are initially calculated. These values are then upsampled to the full 128 pixel grid using nearest neighbor interpolation. (B-C) Average cross correlation matrix for all recordings (using the final stimuli set) for morph dimensions A to E and A to H. (D) Using a polynomial of order 0 through 6 to estimate the Neural Stretching curve for a single recording and morph dimension, A to C (same as above). Vertical lines represent the behaviorally determined psychometric boundaries for all birds tested across this morph dimension. The horizontal pink line is the 0th order fit, the exponential gold line is the 1st order fit. (E-F) 4th order mean Neural Stretching curves across all recordings on morph dimension AE (E) and AH (F). Individual 4th order Neural Stretching curves for each recorded population are faintly plotted to show variance. Vertical lines represent the behaviorally determined psychometric boundaries for all birds tested across this morph dimension. (G) Cross correlation matrix reconstructed from the 4th order Neural Stretching curve for the same recording used in (A) and (D). (H-I) Average reconstructed cross correlation matrix from all 4th order Neural Stretching curves on AE (H) and AH (I) (J) Cumulative distributions of the value of the 4th order Neural Stretching curve at psychometric boundary locations of the same dimension vs different dimensions as a control. A one-sided KS test provides a P value of $p = 1.47E - 121$ (K) Average accuracy of predicting the morph dimension of a Neural Stretching curve from a held out recorded neural population as we increase the order of the polynomial. The green and orange curves use the optimized polynomial coefficients to predict the morph dimension label. The red and blue curves use the normalized Neural Stretching curve evaluated at 50 evenly spaced points along the morph dimension. The red and green curves use a multinomial logistic regression as the classification model. The orange and blue curves use XGBoost as the classification model. The vertical lines highlight the maximum average held-out prediction accuracy achieved by each method. Chance performance $(1/24)$ plotted as a horizontal dotted line. (L) Correlation between the boundary sensitivity parameter (B) of the psychometric curve and the value of the 4th order Neural Stretching curve for the given boundary location. Plotted above and to the right are the projected histograms and associated kernel density estimate of the distribution. Plotted as a line is the regression between these two variables with 1000 fold bootstrapped 95% confidence interval plotted as a shaded region around the regression.


A
B
C
Morph Position
A
C
E
H
D
E
F
∂N/∂s
0.00
0.01
0.02
0.03
0.04
0.05
G
H
I
Morph Position
J
Cumulative Distribution
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.02
0.04
0.06
0.08
0.10
∂N/∂s
K
Average held-out Population Accuracy
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0
2
4
6
8
10
12
Polynomial Order
L
Boundary Sensitivity (B)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.00
0.01
0.02
0.03
0.04
0.05
∂N/∂s

representation space N for a given recorded neural population where we have samples of noisy projection, through a brain, of examples of s_i into N, which we will call $\mathcal{B}(s_i) : s_i \rightarrow N$. We would like to define the following curve as $T(s_i) = \frac{\partial N}{\partial s}(s_i) : s_i \in s$, such that for a pair of presented stimuli, s_i and s_j, and a distance, $y = |\mathcal{B}(s_i) - \mathcal{B}(s_j)|$ in neural representation space between these to presentations, $y = \int_{s_i}^{s_j} T(s)ds + \varepsilon$. This assumes that the neural representation of these stimuli smoothly vary (up to some noise) as I move along the morph dimension. We then perform the regression to minimize the Mean Squared Error of ε for all pairs of stimuli presented on a single morph dimension.

The Neural Stretching curve measures how fast the neural representation changes as we move along a path in stimuli space.

If we consider how the Neural Stretching curve is related to the cross-correlation matrices plotted in figure 2.6(A-C), it is a function defined along the diagonal and the value of each i, j is the distance y fit by the integral of the function from (i,i) to (j,j). Figure 2.6(G-I) show how much of the structure in the cross-correlation matrices in (A-C) are described by the Neural Stretching curves in (D-F)

We would like the Neural Stretching curve $T(s_i)$ to have the following properties: 1. Strictly positive. 2. Smooth over the range D. To keep the function positive I have used $T(s_i) = e^{f(x)}$ and to keep it smooth I have used $f(x)$ as a polynomial of varying orders. Several other parameterizations have been explored, but they provided qualitatively similar results and took significantly longer to fit.

The polynomial parameterization and s_i sampling make the Neural Stretching curve highly susceptible to Runge's phenomenon near the ends of our morph dimension, especially for higher order polynomials. Therefore we do not interpret any spikes near the endpoints of our morph dimension as being meaningful.

Ignoring 0th and 1st order polynomial fits, higher order fits seem to all robustly fit the same curve shape as seen in figure 2.6(D) with a peak near the behaviorally determined psychometric boundaries from other birds. This means that the neural representation was changing the fastest near the location on the morph dimension that other birds choose to place their boundary. This is less surprising when we see figure 2.6(E-F) and see that the shape of the 4th order Neural Stretching curve is relatively similar for every single neural population recorded. Thus the Neural Stretching curve is a somewhat recording invariant measure of changes in the neural representation. To statistically test this, we try to predict the dimension a Neural Stretching curve describes without seeing any other data from a recording in figure 2.6(K). We use two different prediction models and two different representations of the Neural Stretching curve, and for each test, our validation accuracy is significantly greater than the chance value of $1/24$. Each method-representation pair achieved max performance on held out populations using a polynomial order near a 4th order polynomial, which further reinforces our decision to use a 4th order polynomial fit.

Lastly, if we consider the morph dimension from motif A to motif E (column BEH) where we observe a shift in the psychometric boundary between cohort 1 and 2, we see that many of the individual population Neural Stretching curves have two peaks, and there is undoubtedly two peaks in the average Neural Stretching curve which correspond to the two locations of the psychometric boundaries.

If we compare the distribution of the value of the normalized Neural Stretching curve at the psychometric boundary location for psychometric curves on the same morph dimension to the values for different morph dimension in figure 2.6(J), we see a definite shift in the distribution. A 1-sided KolmogorovSmirnov test confirms this with a p-value of $p = 1.47E - 121$ indicating that it is highly improbable that morph dimensions boundaries do not occur at locations on the morph dimension with higher Neural Stretching values than would be expected by chance.

Lastly, we use the natural variation in boundary location to test if higher Neural Stretching values are correlated with increased boundary sensitivity (B) in the measured psychometric curves. The relationship in the average Neural Stretching value of the psychometric boundary location to the boundary sensitivity (B) parameter has a Pearson's R correlation value of 0.42 which means we have a $p = 6.5E - 7$ that there is no correlation between these two variables.

2.3.8 Predicting psychometric curves from neural population activity

Using a hold-one-dimension-out cross-validation strategy, 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. For computational tractability, we used the behaviorally relevant reduced dimensional neural population representation described in figure 2.5(B). The mean squared error for each held out dimension are combined to provide 16 performances for each neural recording (39) for each set of behaviorally determined psychometric fits (8). We create a null distribution by shuffling the labels of the morph dimensions 2048 times and measuring the KS distance to the distribution of 16 dimension performances by 8 behavioral bird's psychometric functions. Overall we find the fits are significantly better than would we expected by chance. Furthermore, if we compare the distribution of p values for recordings from untrained naive birds against the recordings of the birds that had been behaviorally trained we find that there's no difference between how well the neural representation of naive birds or trained birds can predict the location of the boundary or the sensitivity to changes near the boundary. That is not to say that there is no difference between the population representation in trained or untrained birds, or that more information about boundary location isn't encoded elsewhere in the brain, but that the information that allows generalization of categorical boundary parameters to other morph dimensions is equally present in both trained and untrained birds.

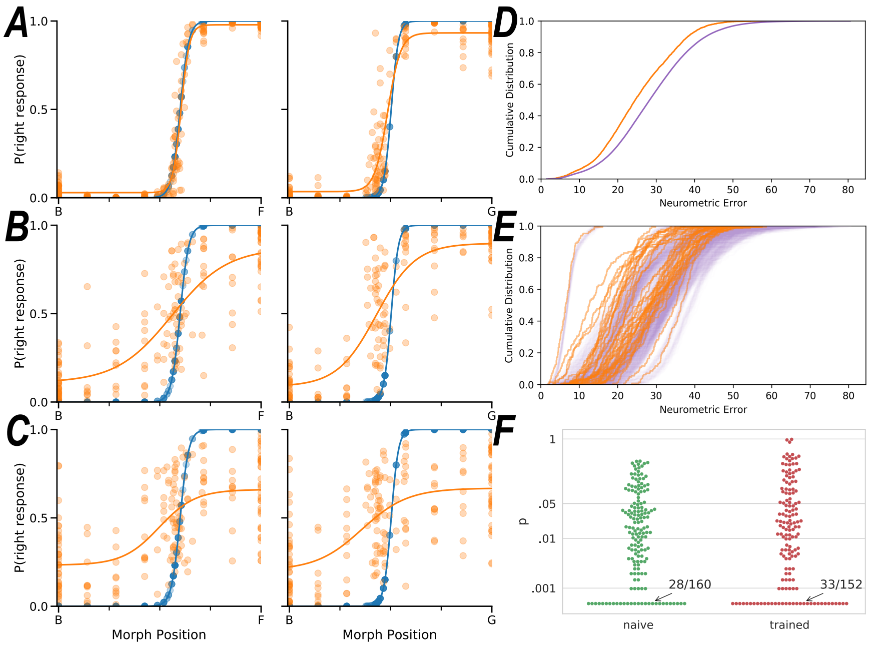


Figure 2.7: *Predicting psychometric curves from neural population activity* (A-C) Predictions of behavioral psychometric curves for two morph dimensions, left: B to F, right:B to G, for the same behavioral bird, using three different recorded neural populations of differing quality. (A) shows an excellent fit, (B) shows a good fit, and (C) shows a fair fit for these two morph dimensions. Plotted in blue is the behavioral target, and plotted in orange are the held out predictions from a logistic regression trained on the other 15/16 morph dimensions. The orange dots are the predicted probability of each stimuli presentation and the orange line is a 4 parameter logistic curve fit using Mean Squared Error to the probabilities. Some populations (A) do a lot better job at predicting the psychometric curves than others (C). (D) Cumulative Distribution of the Mean Squared Error for neural predictions of the psychometric functions in orange and the cumulative distribution of the Mean Squared Error for neural predictions of shuffled psychometric functions in purple. This is the complete cumulative distribution for the Mean Squared Error for each of the 16 dimensions, for predictions against the psychometric curves of each behavioral bird (8), for all recorded neural populations (39). (E) Same as (D) except plotted individually for each neural population in orange, and individually for each of 64 shuffles in purple. (F) Shuffling each population against the behavioral curves 2048 times and comparing the KolmogorovSmirnov metric between the distributions allows us to estimate the p values for each neural population. Overall we find most population fits to perform better than would be expected by chance. Furthermore, if we split the recorded populations depending on whether they were from birds trained on this task or naive (never having heard the stimuli before) we find no difference in the distribution of p values. That is to say that a KolmogorovSmirnov test between the distribution of p values obtained for trained birds against the distribution of p values obtained for naive birds fails to reject the null hypothesis with a p value of $p = 0.565$.

2.4 Discussion

To the best of our knowledge, this is the first single-unit level study where a vast (and arbitrarily chosen) selection of a species own vocal communication signals are used to understand categorization. Generalization in either response domain (behavioral or neural) was measured, and the only study where cross-domain (from neural to behavioral) generalization is shown.

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 observed shift in psychometric boundaries between cohorts was difficult to statistically explore because they only occurred in three of the 24 of the morph dimensions measured between the three cohorts. These shifts may indicate that depending on initial training conditions; birds may end up using different stimuli features to perform the classification or that discrimination along certain features may be altered by learning.

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 [Sainburg et al., sion].

The field of machine learning is rapidly evolving, and there are several 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 caudal mesopallium are sensitive to these small physical differences in physical stimuli that are imperceptible to human ears.

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.

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., 2011a, Kozlov and Gentner, 2016] that opposes the strongest version of sparse hierarchical models of perception, where neurons with simpler receptive fields converge onto neurons with 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*.

Using A DBN as a generative model of birdsong is another limitation of this work due to the amount of recent progress in machine learning. In particular, we recommend using [Sainburg et al., 2018] for these techniques as it can produce much more realistic sounding birdsong and is explicitly constructed such that the stimuli space is convex and interpolations between exemplars are realistic. There also exist a variety of other methods[Kawahara and Matsui, 2003, Prenger et al., 2019] that could be used to interpolate the songs.

Another possible limitation of these results is the anesthetized recording prep that was used. These measurements might not necessarily reflect the actual awake representations and especially attentional modulation, but certainly measure the functional tendencies of the network representation [Bluvas and Gentner, 2013, Knudsen and Gentner, 2013]. Additionally, this doesn't change our interpretation, which is based on the fact that we can predict categorical boundary parameters and not precisely what representation we are using to make the prediction.

While regions of the auditory forebrain are homologous to mammalian auditory cortex [Wang et al., 2010], there are significant structural differences. Thus, while solutions to the categorical perception of complex auditory stimuli used by birds aren't necessarily the same as those in our own, they do provide another example of a neural system capable of solving problems (recognizing complex hierarchical features and recursive grammars) previously thought to be strictly within human capacity. Thus, while it might not allow exploration of the neural circuitry we use, this line of work can perhaps indicate what kinds of circuit patterns and architectures could be useful in a system that needs this capacity.

2.5 Methods

2.5.1 Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Marvin Thielk (Marvin.Thielk@gmail.com).

2.5.2 Experimental Model and Subject Details

European Starlings

European starlings that have been wild-caught in southern California have served as the experimental subjects. All subjects have full adult plumage when acquired, indicating they were at least one year old. Otherwise, I do not control for age or sex of the subjects. Subjects are housed in a large, mixed-sex, conspecific aviary with *ad libitum* access to food and water from the time of capture until being moved into the testing chamber. The photoperiod in the aviary and testing chambers followed the seasonal variation in local sunrise and sunset times.

2.5.3 Method Details

Deep Belief Network latent space morphs

Using a large corpus of recorded starling song, a compressive Deep Belief Network [Hinton and Salakhutdinov, 2006] is trained to autoencode 400 mS sections of starling song. The Analysis & Resynthesis Sound Spectrograph (ARSS) package was used to generate the spectrograms and perform the spectrogram inversion. We used 400 pixels per second, with 44 frequency bands logrithmically spaced between 850 Hz to 10 kHz. We created a dataset of 3,000,000 samples of (possibly overlapping) 400 mS long spectrogram slices of size 44x160 = 7040 dimensions. We found that contrast normalizing the spectrogram using a logistic on the outputs of the ARSS generated spectrograms. We tested several layer size configurations and settled on a network with 5 encoding fully-connected hidden layers of sizes 1024, 512, 256, 128, and 64, which is then decoded back into 128, 256, 512, 1024, and finally back into the 7040-dimensional spectrogram space. The network is trained in a greedy layerwise fashion where the first layer is trained, then fixed, then the next layer is trained with the previous layers fixed, and so on until all layers are trained. Once the network has been trained, we can project a 400 mS motif, A, into the latent space, which we'll refer to as $Z_A = DBN(A)$, and then reconstruct the original spectrogram with the decoder network, $DBN^{-1}(DBN(A))$. The contrastive divergence learning algorithm used by the network tries to minimize the reconstruction error, $|A - DBN^{-1}(DBN(A))|$. We can also linearly interpolate between latent space representations of two motifs, A and D, to generate a smoothly varying path in spectrogram space that morphs between A and D. $\text{morph}(A, D, \tau) = DBN^{-1}(\tau DBN(A) + (1-\tau)DBN(D))|\tau \in [0,1]$. For the experimental morphs I used $\tau \in \{i/127 \mid i \in [0,127] \cap \mathbb{N}\}$. This process is diagrammed in 2.1(C)

Operant panels

Our testing boxes are acoustically isolated chambers containing a panel, shown in 2.1(B), with three response ports. I use PyOperant, an open sourced python package we have developed in the Gentner Lab, to reward the birds by raising a hopper of food to reinforce correct responses. Acoustic stimuli are delivered through a small full-range audio speaker mounted behind the panel and out of the subject's view. Water is unrestricted.

Shaping

I have coded an approximately unsupervised automated procedure as part of Pyoperant for teaching the birds how to interact with the panel and respond to stimuli to receive the food rewards. This autoshaping routine[Brown and Jenkins, 1968] consists of multiple stages that help the bird become accustomed to receiving a food reward from the hopper and interacting with the panel, and automatically transitions the bird into the task. It takes the subjects an average of 3-5 days to complete the whole autoshaping procedure.

Behavior

Subjects learn to classify two sets of song motifs (4motifs/set) using a well-developed two-alternate choice procedure[Gentner and Margoliash, 2003]. Briefly, subjects initiate each trial by pecking at the center response port to trigger the presentation of the stimuli (chosen at random on each trial). Immediately after playback, the subject must respond to the left or right port to obtain food. For half the motifs, the subject must peck left; for the other half, it must peck right. Correct responses are reinforced with brief access to food. Incorrect responses are punished by turning off the houselight for a few seconds and no access to food. High response rates can be achieved, and stimulus-independent response biases can be ameliorated by manipulating the reinforcement schedules or introducing remedial trials according to established procedures. Subjects can initiate trials from sunrise to sunset, and food intake and weight is monitored throughout training to ensure well-being. The birds are trained with a variable ratio of 4, which means need to get 1-7 (average of 4) in a row correct to be rewarded.

Double staircase

Once the subject meets our performance and response rate criteria, I begin the ratcheting double staircase procedure to estimate the perceptual boundary between pairs of motifs. The procedure works by estimating a window encompassing the boundary and iteratively reducing the range from both sides (independently) based on the subject's performance. The staircase procedure begins by randomly choosing one of the 16 possible motifs pairs, and then selects a motif morph that is outside the window (90%) or just inside the window (10%). For an easy trial, the stimuli is one of the motifs mixed only slightly with the other motif. For the probe trial, the stimuli is a morph just within the window the procedure believes the perceptual boundary to be in. If the subject gets a probe trial correct, the window edge advances to the location of the probe trial and further probe trials along this axis become more difficult. The subjects are rewarded by a variable reinforcement ratio (they must respond correctly up to a variable threshold before being rewarded) so that the birds are forced to perform on each trial but are not necessarily rewarded. This also allows for more trials per day.

Electrophysiology recording

We record from caudal mesopallium, a secondary auditory region, in lightly anesthetized (7-8mL of 20% by volume urethane per kg) starlings. We target caudal mesopallium using established stereotaxic coordinates of 2500 um rostrally, 500 um laterally, and 2000-2500 um deep at an angle of 45° from the plane between a beak bite bar and ear pins [Knudsen and Gentner, 2013].

We record extracellularly using a 32 channel silicon Neuronexus probe (most recordings were performed using an electrode with the edge configuration, although some early recordings may have used other configurations). The recordings used the data-acquisition hardware and Spike2 software from Cambridge Electronic Design (CED).

We performed post-hoc spike-sorting using mountainsort [Chung et al., 2017], which provided a set of spike clusters from putative neurons. We convolved the spike trains with a Gaussian ($\sigma = 10\,\mathrm{ms}$) and sampled the function at 50 points during the stimuli presentation ($400ms$) to provide a 50-dimensional representation of the neuron's response to the stimuli that preserved spike timing information. We then exclude neurons that did not contain stimulus-specific information by performing pairwise logistic regression on the subset of stimuli corresponding to the training motifs or 8 template endpoints of our morph dimensions, validating performance on held out data. We exclude units with average performance lower than 0.6.

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 in the latent space and half were equally spaced in perceptual space to ensure adequate sampling near the relevant behavioral boundaries. Since the perceptual space is behaviorally determined from each bird independently, we randomly sampled the birds to get a fixed set of stimuli that could be used for all recordings.

2.5.4 Quantification and Statistical Analysis

Psychometric curve fitting

I model the behavior as a Bernoulli process with the probability of responding left or right determined by a four-parameter generalized logistic fit dependent on the morph position. I fit the parameters of the logistic using maximum log likelihood.

$P(r) = A + \frac{K-A}{1+e^{-B(x-M)}}$

Psychometric curve conservation

To statistically measure how well psychometric parameters are conserved within morph dimension or within a bird, we consider the pairwise distances between all the fit points of subjective equality (M) across all dimensions and subjects. We then measure the KolmogorovSmirnov distance between this null distribution and the subset of pairwise distances that only between the M fit of psychometric curves of the same dimension, and separately the subset of pairwise distances of the same bird. Since our observations are pairwise distances and therefore not independent, we bootstrap a null distribution by shuffling 2^{17} times. The bootstrapped p-values do not differ qualitatively from the KS test derived p-values. We perform this same analysis on each of the 4 parameters of the psychometric fits.

Task relevant LDA-like dimensionality reduction

We concatenate the 50-dimensional single unit representations into a population representation that varies in size from recording to recording. To compress all representations into the same size lower dimensional space, we perform a linear discriminant analysis-like (LDA-like) dimensionality reduction on the data. Instead of the dimensions described in the LDA algorithm, we project the data onto the dimensions determined by logistic regression performed on the endpoints of the 24 possible comparisons ever behaviorally presented to any cohort of birds. This provided a 24-dimensional representation of the population response that contains the spike timing information of individual units that maximally discriminates the 8 endpoint or template stimuli.

Hold-one-dimension-out neurometric fits

We examine the amount of information about the relationship between the behaviorally determined psychometric parameters that is stored in the neural representation by fitting a logistic regression and predicting on a held out dimension. We use the task-relevant 24-dimensional LDA-like representation of the population response as the input to a logistic regression predicting the probability of left vs. right using the behaviorally determined psychometric curves scaled such that $A = 0$ and $K = 1$. We use 16 fold cross validation structured such that all exemplars of a given dimension are held out at once. This provides us with an estimate of how much the other psychometric curves provide information about the shape of the held out dimension. The sum of the mean squared error between the logistic regression's predicted probabilities and the scaled psychometric curves for each recorded stimuli on the held out morph dimension, which provides a distribution of 16 errors for the representation.

We compare this distribution to null distribution of errors obtained by shuffling the morph dimension labels on the psychometric curves using the KolmogorovSmirnov distances. We shuffle each starling's (8) psychometric curves $N = 2048$ times to get a cumulative null distribution of errors, for each recorded neural population (39). Then we measure the one-sided KS distance from the total cumulative distribution to each shuffle's cumulative distribution to estimate the distribution of KS values, which subsequently allows us to estimate the p-values of each fit.

2.5.5 Data and Software Availability

All analysis scripts and data are available from https://github.com/MarvinT/morphs

2.5.6 Key Resources Table

ARSS

MountainSort

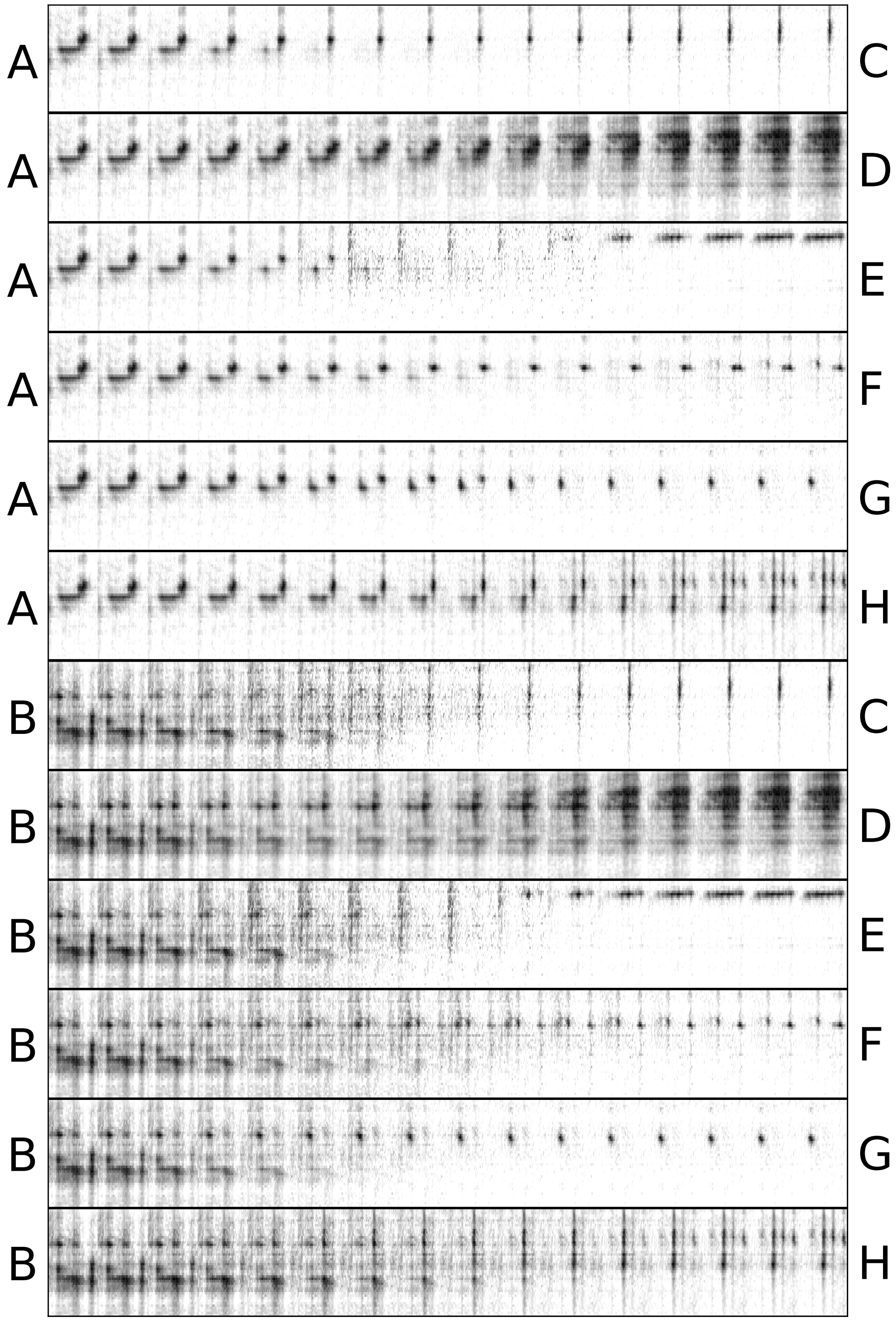
Morphs

Pyoperant

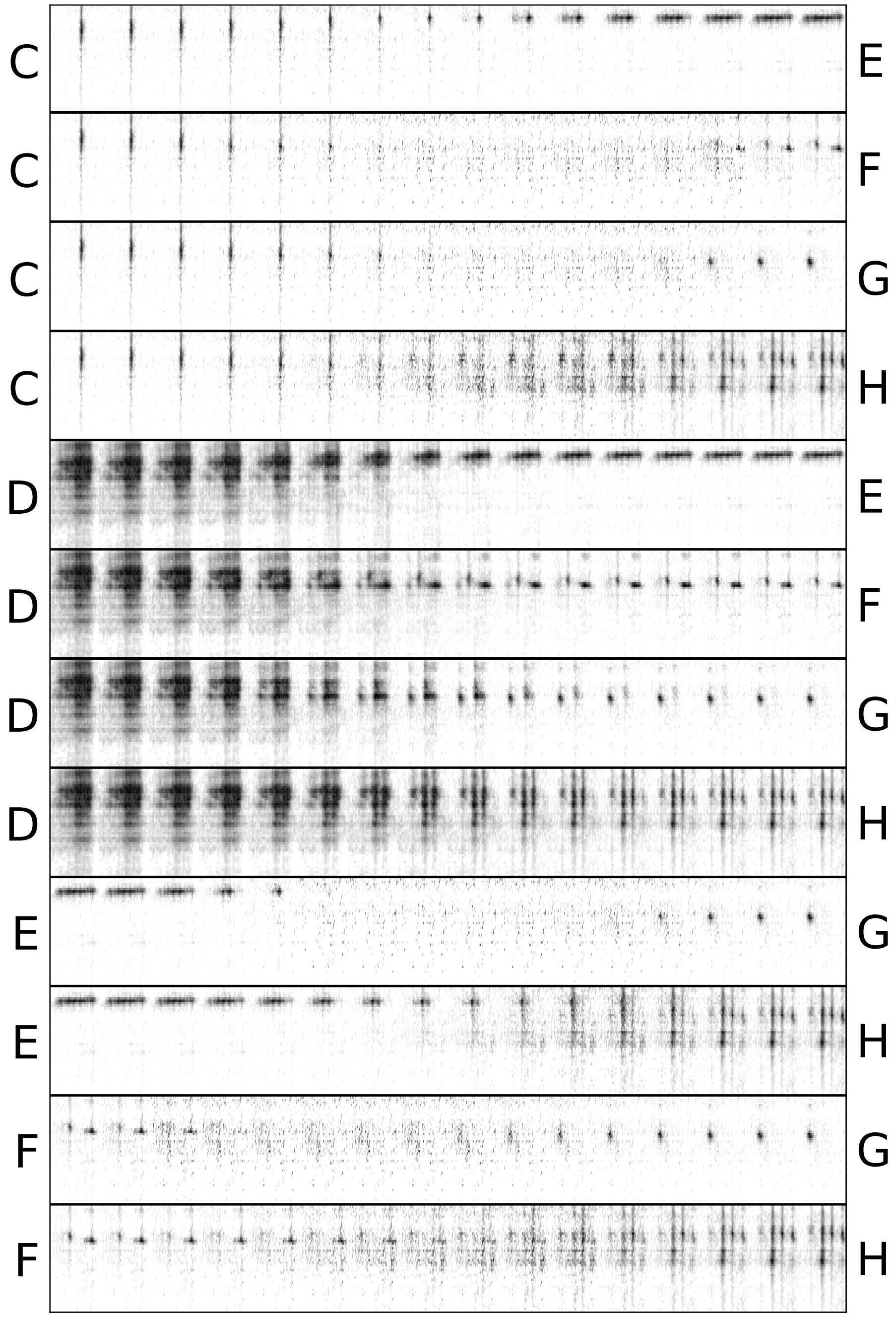
Spike2

2.6 In Preparation Acknowledgement

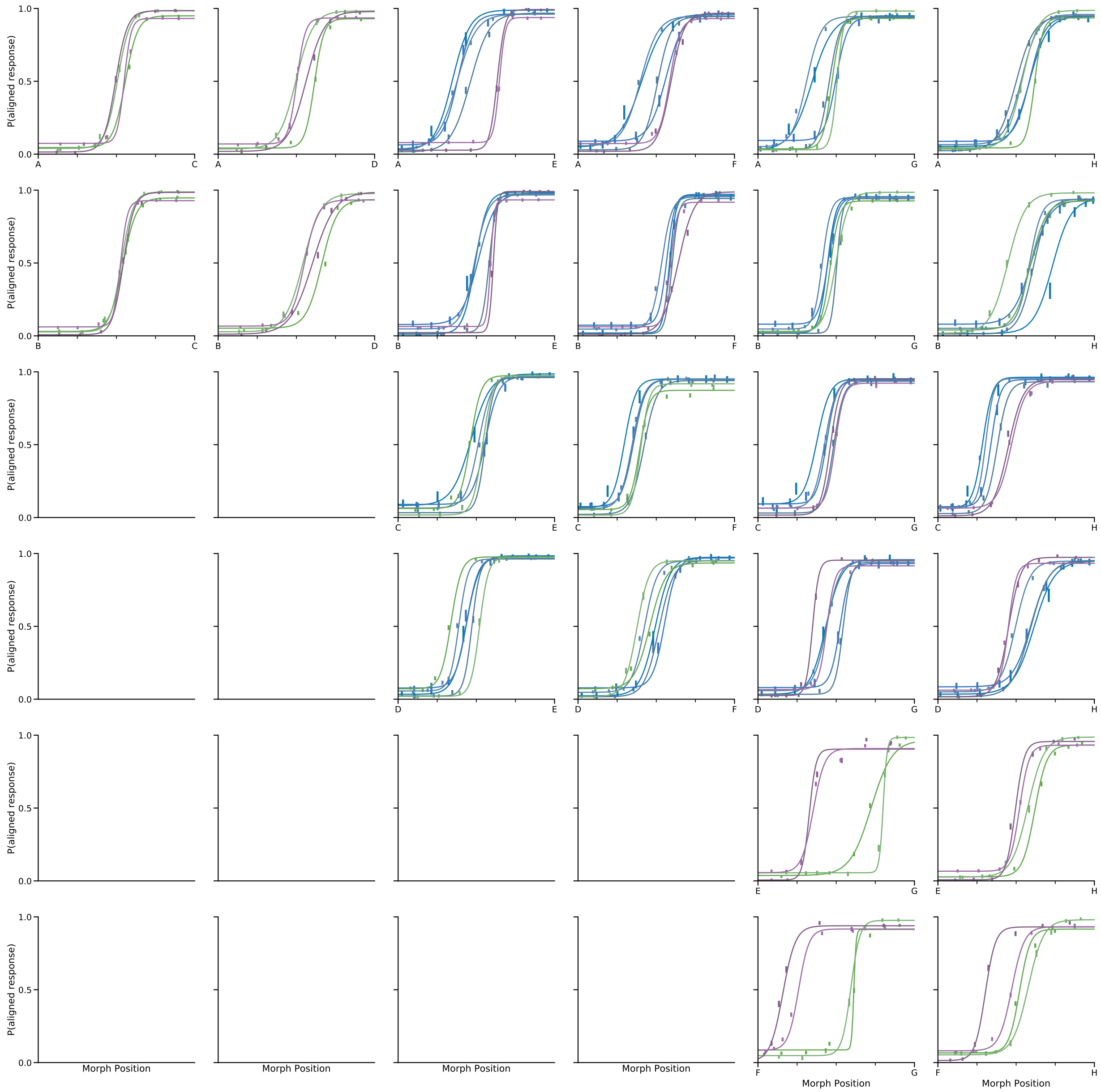
Chapter 2, in part, is currently being prepared for submission for publication of the material. Thielk, Marvin; Sainburg, Timothy; Sharpee, Tatyana; Gentner, Timothy. The dissertation author was the primary investigator and author of this manuscript.



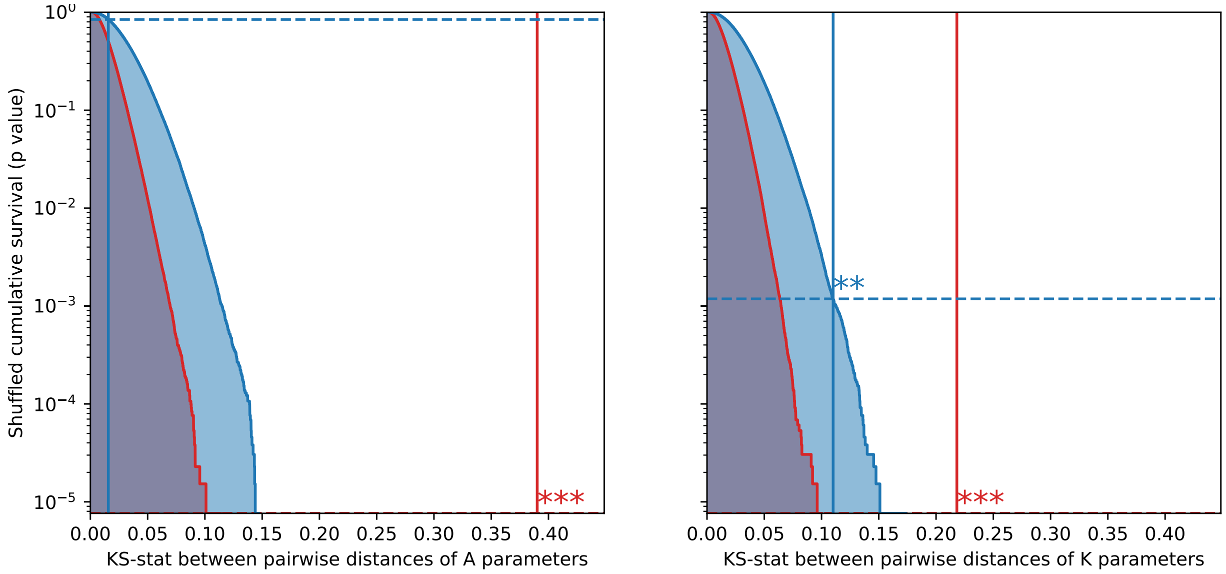
Supplementary Figure 2.8: *Generated morph dimensions 1/2* 12 example interpolated morph dimensions generated using the Deep Belief Network. 16 (of 128 used) example motifs for each morph dimension. Spectrogram representation with frequency on y-axis and time on the x-axis. Each motif is 400 ms long.



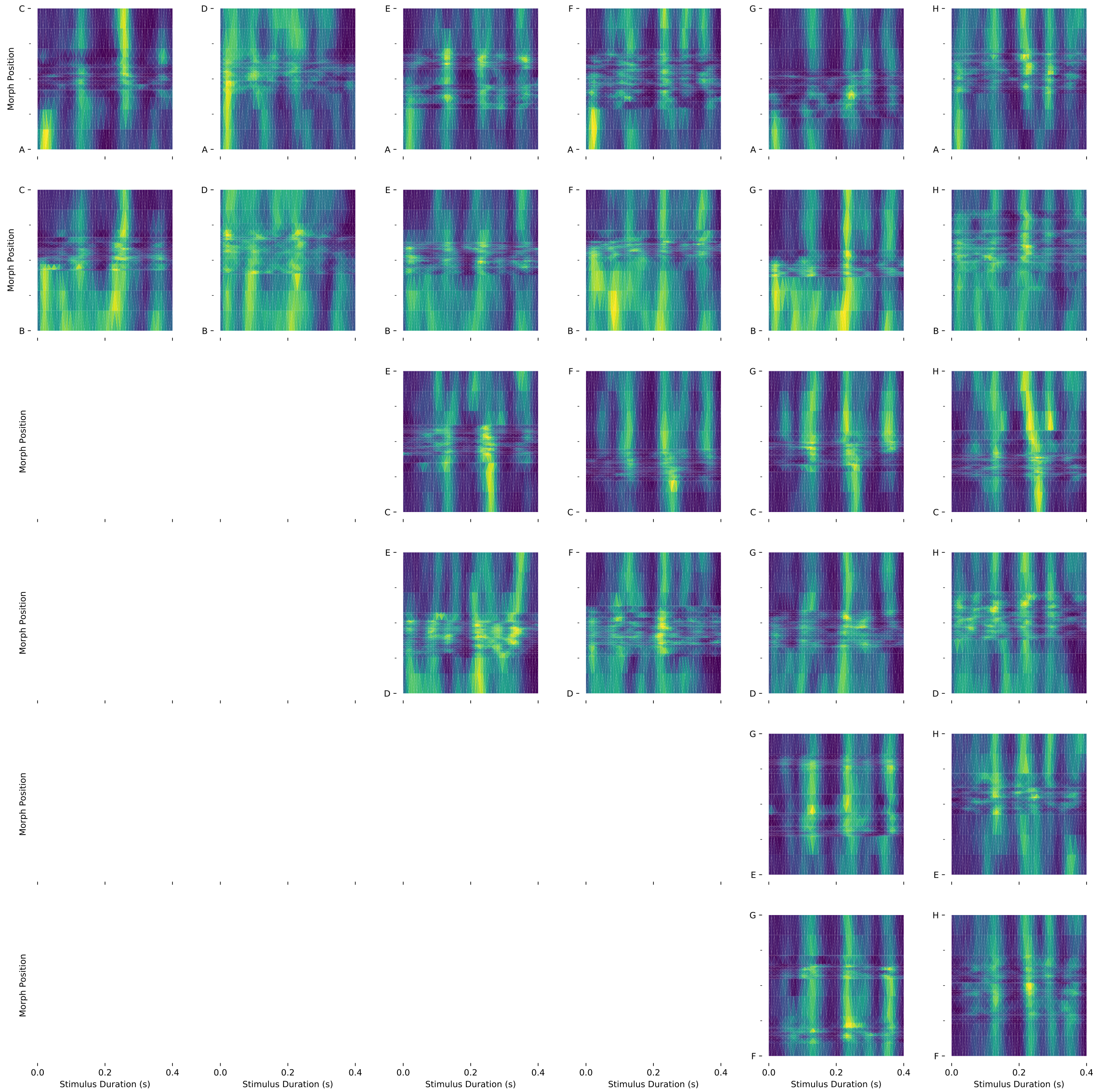
Supplementary Figure 2.9: *Generated morph dimensions 2/2* 12 example interpolated morph dimensions generated using the Deep Belief Network. 16 (of 128 used) example motifs for each morph dimension. Spectrogram representation with frequency on y-axis and time on the x-axis. Each motif is 400 ms long.



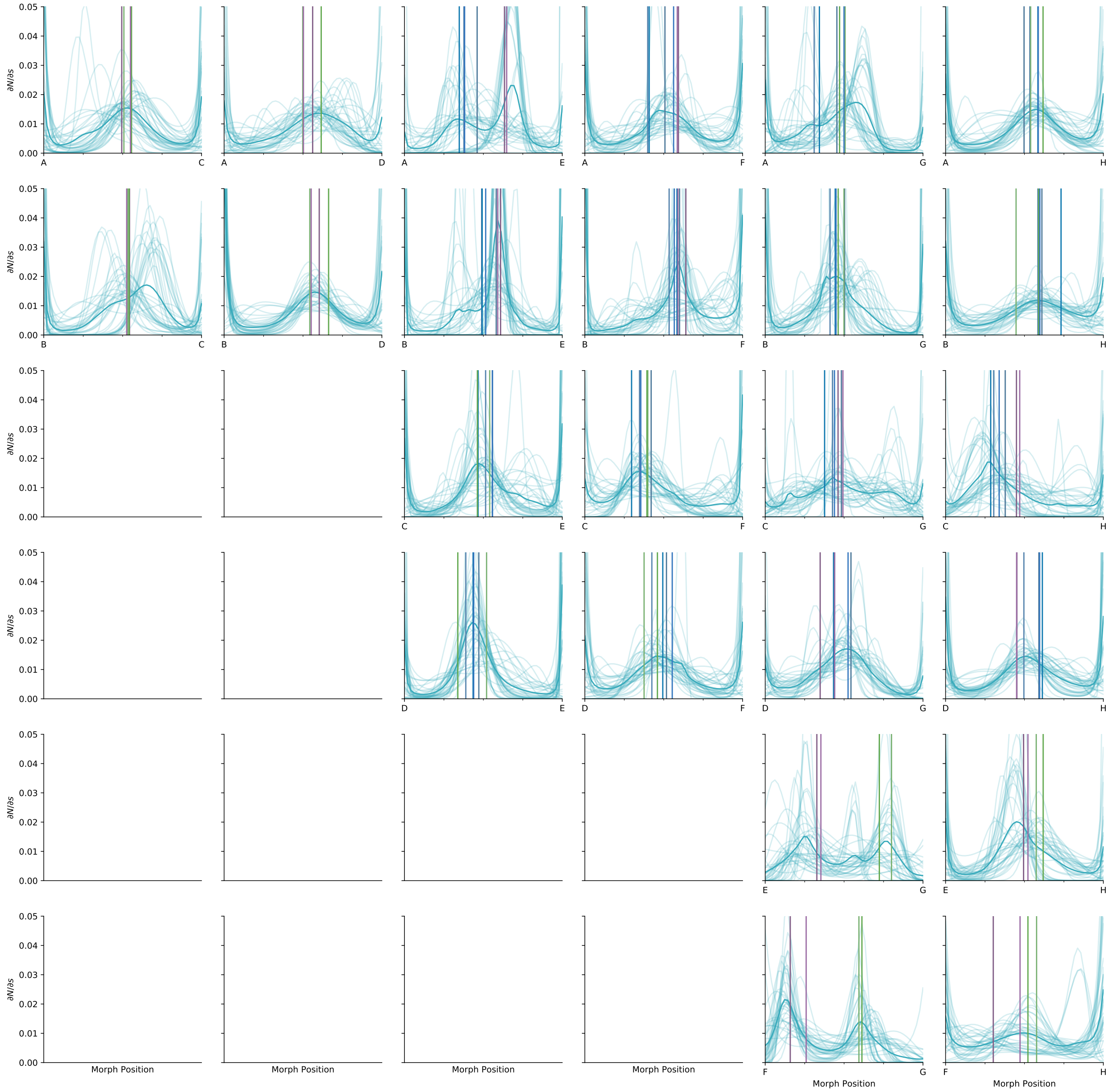
Supplementary Figure 2.10: *All psychometric curves* Psychometric curves for all behavioral subjects on all 24 morph dimensions. Blue indicates the bird was part of cohort 1, purple, cohort 2, and green, cohort 3.



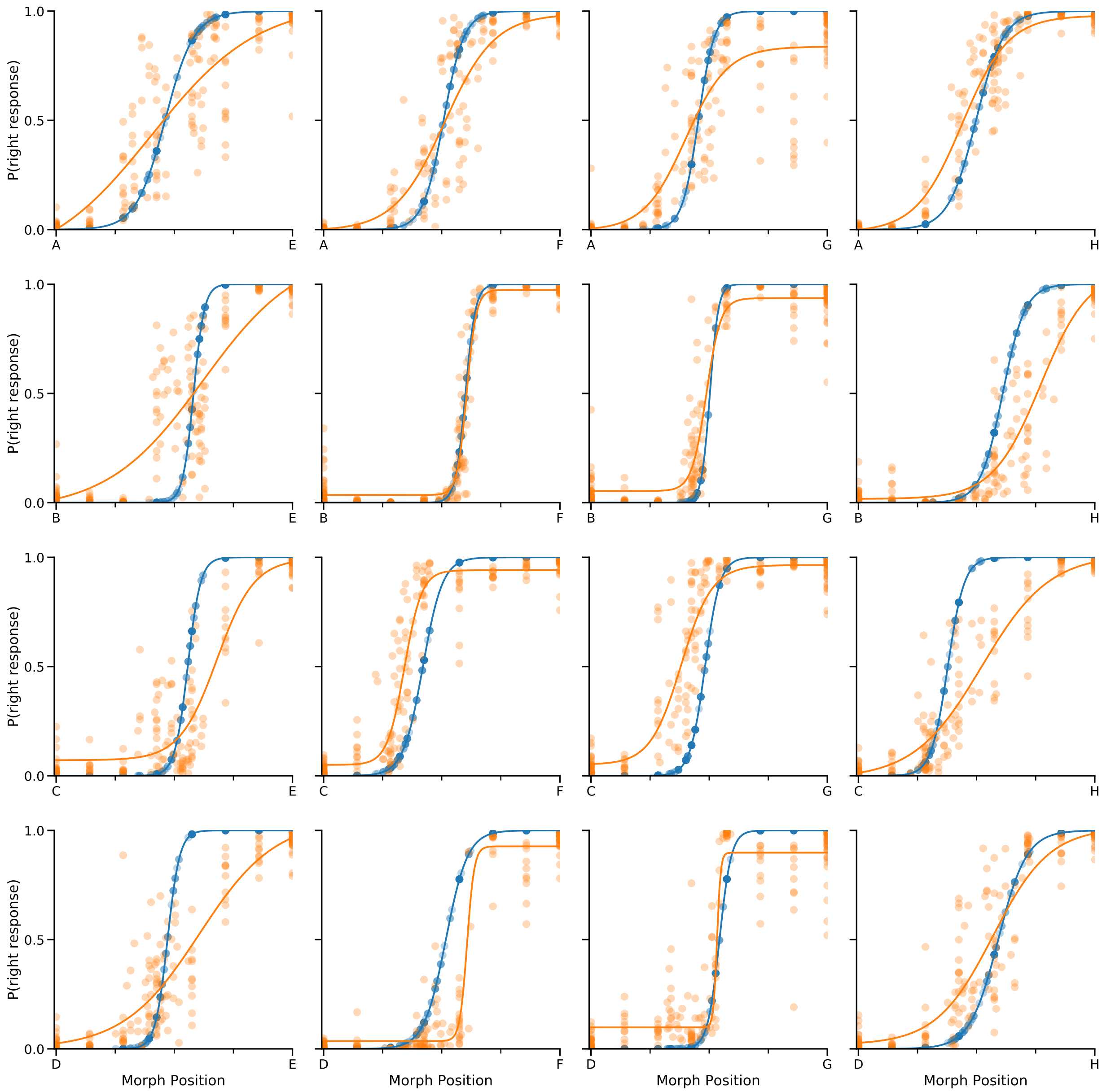
Supplementary Figure 2.11: *Parameters A and K are conserved within subject* The corresponding figure to 2.2(E-F) for the scaling parameters A (left), and K (right).



Supplementary Figure 2.12: *Morph activity for 24 morph dimensions for a single neuron* Variation of the average Gaussian convolved single neurons representation across morph dimensions. Smoothly interpolated using triangulation for surface estimation. All 24 dimensions of the neuron plotted in figure 2.4 (F)



Supplementary Figure 2.13: 4th order mean Thielk curves across all 24 morph dimensions from all recorded neural populations. Corresponding to figure 2.6 (E-F).



Supplementary Figure 2.14: Predictions of behavioral psychometric curves for all morph dimensions for a single behavioral bird in cohort 1. Same neural population and behavioral bird as figure 2.7 (A).

Chapter 3

Predictive Coding

Predictive Coding (PC) is the theory that the “purpose” of the perceptual system is not to encode the maximum amount of information about the current state of the world, but rather to provide a prediction of (the posterior distribution of) the future state of the world.

Indeed, this provides a parsimonious explanation for much of the brain’s function. If you consider a video of two people talking in a living room with an old TV displaying static - most of the Shannon information in the video will be in the TV’s display. While the speck of static may be a result of a particle left by the big bang hitting the antenna, our brain doesn’t devote the majority of its resources to keep track of the white noise patterns. i.e., the brain’s sensory processing system isn’t designed to maximize the mutual information between the sensory signals it receives. Instead, if you consider the temporal information, the brain tries to encode the predictive information within its sensory streams. Predictive Coding has been hailed by some as “The unifying theory of the brain.”

Predictive Coding provided the inspiration and fundamental theory behind temporal difference learning, a class of model-free reinforcement learning methods which learn by bootstrapping from the current prediction of the state of the world. Temporal Difference Learning produced TD-Gammon, which was beating humans even before the more famous Deep Blue eventually beat Kasparov. Temporal Difference Learning is also one of the inspirations for many of the current state-of-the-art reinforcement learning applications such as AlphaGo by DeepMind and Five by OpenAI.

If we were prescribed to the theory of Predictive Coding we would expect that somewhere the brain has to encode 1) the prediction of what is to come, and 2) the error between the prediction and what came to pass. Indeed, some of the most famous work in neuroscience has demonstrated these values in predictions of reward in dopamine signaling in the VTA. However, the theory of Predictive Coding would expect that these values should exist throughout the perceptual system as well for all aspects of the state of the world. An excellent place to look for this would be where the brain processes complex time-varying signals, namely, in auditory processing regions.

A direct corollary of Predictive Coding is the notion of surprise which is a measure of how likely a stimuli is given its context. A simple formulation might be $\text{Surprise} = -\log P(S|C)$ where S is the stimuli and C is the context. Several normalizations or transformations have been applied to this formulation; however, this describes the fundamental idea. Some version of this formulation has been used to show that responses from secondary sensory regions are better described by models of surprise than by models using strictly the physical stimuli in both vision [Itti and Baldi, 2005] and audition [Gill et al., 2008].

This formulation requires a model of the world that provides the predicted probability. Not too long ago, “a model-free approach seems out of the question in the context of video processing, as it would take unreasonably too many data samples and hence unreasonably too long to accumulate sufficient data and allow accurate model-free estimation of the underlying probability density function (PDF) of the data” [Itti and Baldi, 2005]. However, this is no longer true with modern machine learning techniques and methods for handling significantly more data than was possible in 2005. Therefore I have spent some effort building some models that predict future posterior probability density function of spectrograms.

3.1 Models

3.1.1 Deep mixture of beta distributions

My first attempt was to create a neural network that, given a window of a spectrogram, predict the probability distribution of each frequency band in the next time bin of the spectrogram. The network was optimized by maximizing the log likelihood of the model. I modeled the probability distribution as a mixture of 3 beta distributions. The network was two fully connected layers with an output for a weight, w, and the two beta distribution parameters, α and β, for each of the three beta distributions, for each of the frequency bins. This network was very ill-behaved, and I spent considerable time implementing as many numerical analysis tricks to control and regularize the networks' exploding values. Most were relatively simple and included but was not limited to gradient clipping, different beta distribution approximations, the log-sum-exp trick and even an added cost function tied to the standard deviation of the beta distributions to try to prevent them from becoming delta functions.

One interesting method I used was adversarial training [Goodfellow et al., 2014, Lakshminarayanan et al., 2017]. Given an input x with target y, and loss $l(\theta,x,y)$(e.g. $-\log p_\theta(y|x)$) [Goodfellow et al., 2014] proposes the fast gradient sign method which generates an adversarial example as $x' = x + \varepsilon\,\mathrm{sign}(\nabla_x l(\theta,x,y))$ and using x' to augment the dataset by treating (x',y) as an additional training example. Intuitively, this can be interpreted as a smoothing method by smoothing the likelihood around the target of an ε-neighborhood around the training examples. This helped a small amount, but my networks still eventually got corrupted by NaNs. In my case where I'm predicting the likelihood of a full spectrogram slice, its possible adversarial training on my labels, y might have also helped the stability. In the end, however, I began to strip away as many of the complexities of the network as I could to create a minimal viable network.

3.1.2 Deep Gaussian prediction

This minimum network was to simply predict each frequency band as a univariate Gaussian with a mean and a standard deviation, and let the network try and deal with any covariances. This network trained easily enough but didn't do nearly as good of a job at predicting as the deep mixture of beta distributions networks that were stopped before being corrupted by NaNs. However, even with this poor performance, I did notice peaks in the likelihood that corresponded to motif boundaries.

3.1.3 Contrastive Predictive Coding

The last method I have explored is Contrastive Predictive Coding [Oord et al., 2018]. Contrastive Predictive Coding is a method that takes high dimensional time-varying data and attempts to do representation learning on it to provide a representation that captures the predictive information of the signal. They do this by encoding the high dimensional signal in a low dimensional latent space, Z, using an encoding network, and then feeding the latent space into a recurrent neural network which outputs a context in another latent space, C. c_t is then used to linearly classify if a sample is actually the next time sample, or drawn from a distribution (I have used other random samples from the spectrogram). The whole model is learned end-to-end, providing an encoding network to a Z space that both contains information that is useful for predictions, as well as information that is predictable. The model also provides the C space, which also contains context from the past.

Figure 3.1 shows an example 8-dimensional Z latent space learned by Contrastive Predictive Coding. It encodes all the predictive and predictable information contained in each spectrogram time slice. Noticeable changes in the spectrogram are reflected in some way in the encoded dimensions. It would be interesting to see if it allocated its weights in a similar distribution as would be predicted by the mel frequency scale.

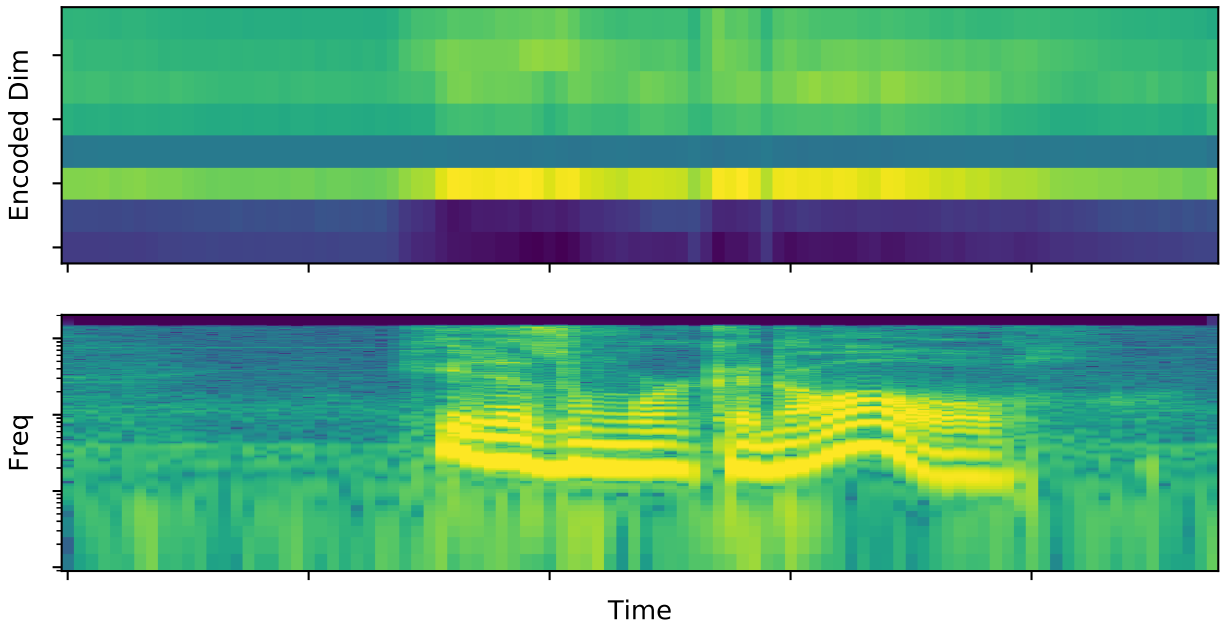


Figure 3.1: *8 dimensional Z space learned by Contrastive Predictive Coding.* The representation learned by the encoder of Contrastive Predictive Coding. Top: The 8-dimensional latent space Z created by pushing each 2048-dimensional spectrogram time slice through the encoder network. Bottom: 2048 Frequency bin spectrogram representation fed into the encoder and the Contrastive Predictive Coding network. The figure has been log scaled in the frequency axis to provide a visualization similar to a mel-scaled spectrogram; however, the 2048 dimensions are not log spaced so the Contrastive Predictive Coding network must learn where to allocate its weights to encode information in the predictive and predictable frequency dimensions. It would be interesting to see if the allocations it learns match that of a mel-scaled spectrogram.

3.1.4 Future Improvements

Contrastive Predictive Coding seems to be working quite well to generate representations, even though it doesn't give us explicit predictions. It would be possible to extract predictions from the Contrastive Predictive Coding network by 1) sampling from the latent space, or 2) explicitly using the linear decision boundary created by the network. There are also several other changes that could be made to the architecture which may improve the performance. A Wavenet [Van Den Oord et al., 2016] style time dilation network architecture might allow for the removal of the LSTM or recurrent neural network and might improve training time. Alternatively, a transformer network [Vaswani et al., 2017] might be more efficient recurrent neural network architecture that is still able to capture the fundamental dynamics. Lastly, further skip-state predictions might improve the learned representations as hinted by [Gregor and Besse, 2018]. There are many possible ways to incorporate skip state predictions into Contrastive Predictive Coding.

3.2 Future Applications

3.2.1 Receptive Field Estimation

These kinds of models have the potential to be used for much more than just posterior distribution prediction. As I alluded to in the introduction, an explicit model of the posterior distribution may allow for improved receptive field models and more accurate spike predictions. This would be done by splitting the stimuli explicitly into what was predicted, and how the actual stimuli differed from this prediction. Since the spiking could be dependent on the surprise at a variety of time lags, the dimensionality of this stimuli space quickly explodes as you consider farther back in time. To restrict the dimensions considered, a linear receptive field model should probably be used first to define the time lag limits and relevant frequency bands.

Alternatively, instead of using an explicit posterior prediction model, the learned Contrastive Predictive Coding encoded representation could be used to make spike predictions. This wouldn't require explicit arbitrary frequency limits which we have currently imposed by estimating the range of frequencies we believe the birds are sensitive to and by using a mel scale based on the human cochlear organization. Furthermore, by varying the dimensionality of the Contrastive Predictive Coding encoded representation, a balance between the encoded stimuli information and the limited experimental statistical power could be formed to combat the curse of dimensionality.

3.2.2 Natural Unsupervised Segmentation of Vocal Objects

Both word segmentation and motif segmentation has proven to be a non-trivial task and has limited efforts towards automatic vocal analysis. In both linguistics and birdsong research, hand segmentation remains the gold-standard technique for separating vocal objects. For human speech, some supervised methods do a reasonable job, mainly due to large amounts of labeled data. For birdsong, this is not available.

Preliminary data seems to indicate the usefulness of the likelihood of each time sample as an unsupervised segmentation signal. Previous methods at song segmentation in our lab mainly consisted of using the sum of the spectral power for each time point. I suspect that an improved method could use the likelihood of each time sample or its derivative. This makes intuitive sense because one would expect that intra-object predictability to be higher than inter-object predictability.

3.2.3 General Encoding of High-dimensional Time-Varying Data

Lastly, these techniques might be useful to extract meaningful information out of general high-dimensional time-varying data, of which we have a lot in Neuroscience. I suspect that the success of techniques like Independent Components Analysis (ICA) [Bell and Sejnowski, 1995], Delay Differential Analysis [Lainscsek and Sejnowski, 2015], and Taken's delay embeddings indicate the kinds of statistical structure that are utilized by Contrastive Predictive Coding are abundant and informative in Neuroscience data. I expect this to be useful for many types of recordings including but not limited to EEG, fMRI, ECoG, LFP, and even recordings of many single units, mainly as an unsupervised dimensionality reduction technique.

Chapter 4

Discussion

I believe that the results in the Categorical Perception chapter would not be possible or significantly reduced if we ignored the temporal component of the neural population activity. Even when single unit activity is available, the way neural activity is usually represented is as an average spike rate during the stimuli playback. This seems incredibly naive and disregards a huge amount of the information that has been recorded, especially when considering the existence of neurons that fire a single time with millisecond precision in response to songs that are seconds to minutes long as shown in figure 4.1. In this light, the absurdity of trying to estimate the discriminability of neural representations of auditory stimuli using estimates of blood-flow that change hundreds of milliseconds after the metabolic activity of hundreds or thousands of neurons change becomes painfully apparent. I suspect that evidence for the neural correlates of Categorical Perception exists in population activity much earlier in the ventral auditory pathway than is measurable using techniques such as fMRI.

It seems likely that in our experiment, the birds did not have to shift their representations, but instead just used preexisting features to classify the stimuli. Similarly, human languages may have adapted to use phenome boundaries located in regions with preexisting features allowing for easier discrimination[Kuhl and Miller, 1975, Stevens, 1981].

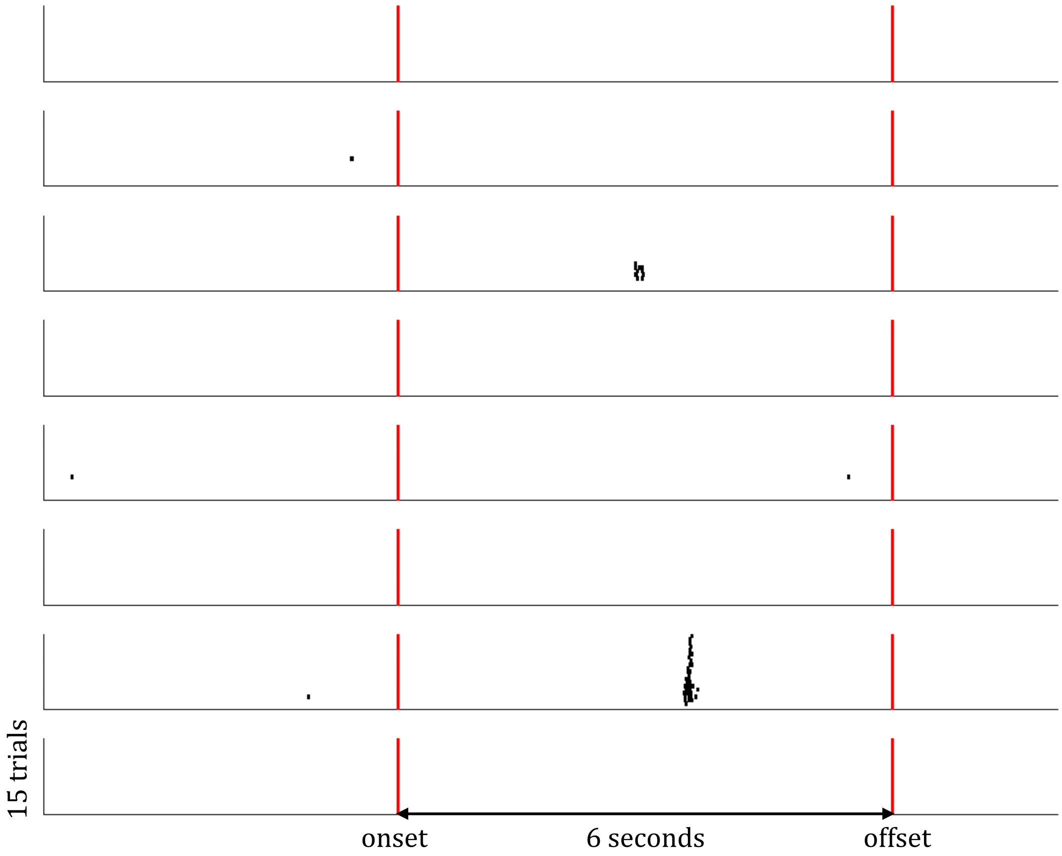


Figure 4.1: *Sparse firing of NCM neuron.* The sparseness of firing by an intracellularly recorded neuron from NCM, a secondary auditory region recorded by Dr. Krista Perks. A raster plot of 8 different 6 second long songs, which were presented to this neuron 15 times each. Each vertical black tick represents a single spike. Stimuli onset and offset are plotted in red.

Overall, I believe these methods represent a new potential direction for research into Categorical Perception and representations in secondary perceptual regions.

A similar method to the Neural Stretching curve analysis could also be applied to the exploration of artificial neural network representations. A separate generative technique could be used to probe category boundaries in a machine learning model trained in a supervised manner on categories. This may be a more effective way to explore how different morph interpolation techniques affect the curves. It has been shown that stimulus naturalness could be a factor in determining the degree of categorical perception[Van Hessen and Schouten, 1999] so it would be interesting to see how stimulus naturalness would affect the Neural Stretching curve. Since there is a remarkable improvement in the state of the art generative models [Sainburg et al., 2018, Van Den Oord et al., 2016, Prenger et al., 2019] far beyond the techniques that the stimulus naturalness effect was originally measured on it would be interesting to see if the Neural Stretching curve shows a similar function of stimulus quality or if the recent advances have hit the ceiling of the effect.

4.1 Analysis Motifs to learn from

This thesis represents my attempts to apply machine learning and develop techniques to handle high dimensional temporal data in neuroscience. As neuroscientific data acquisition techniques have advanced, we now record more data than was previously imaginable. Our attempts to interpret these data in some way parallel the evolutionary forces that drove the brain to process and interpret the many high dimensional sensory signals it receives. In general, there are several analysis motifs that I'd like to incorporate into my repertoire moving forward.

1. Representation matters

2. Dimensionality matters

3. Don't throw out temporal information
4. Don't assume linear scaling
5. Don't assume normal distributions

4.1.1 Representation matters

When I started my Ph.D. in 2012, the AI winter was ending, and machine learning was regaining commercial acceptance and funding was increasing dramatically. There was much talk about the Universal Approximation Theorem of neural networks [Csáji, 2001] which states that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate continuous functions on compact subsets of $\mathbb{R}^n$, under mild assumptions on the activation function. Crucially, however, it does not touch upon the algorithmic learnability of these parameters. There was a consensus that Deep Neural Networks had completely replaced hand-tuned features and that the best approach was feeding in the raw values and letting the neural networks learn the optimal features. While this does work given infinite data, data has proven to be the limiting factor in all real-world situations[Halevy et al., 2009, Sun et al., 2017]. Figure 2.6K is an example of this. The optimized polynomial coefficients completely determine the evaluated Neural Stretching curve; however, the predictive power of even XGBoost differs between them. This is a result of a combination of the kernel effect and how disentangled the data is.

I think a useful metric for the quality of a representation would be the area under the performance curve plotted against the log of the dataset size (scaled by the max performance achieved on the task). This would, of course, be dependent on the task, but it would be a measure of how good a representation is for a given task. The ideal representation (the labels, for example) would take very few samples to reach max performance and would, therefore, have a metric value close to one. This would be especially beneficial for auditory domain questions where many parameters go into the construction of a mel-scaled spectrogram, and I haven't seen too many principled determinations of those parameters.

4.1.2 Dimensionality matters

The curse of dimensionality is real and affects much of the work we do in neuroscience. It refers to the fact that the volume of space increases exponentially as you increase the dimensionality resulting in any available data becoming sparse. Typically downsampling or linear methods such as PCA have been used to reduce the dimensionality, however, with enough data, newer methods such as autoencoder-styled unsupervised techniques can be used as a non-linear analog of the linear techniques. If you have enough data samples to begin to describe non-linear interactions, these techniques can often be a drop-in replacement for the linear techniques. They work by compressing the representation through an information bottleneck, and can often succeed in disentangling the data for an improved representation.

4.1.3 Don't throw out temporal information

This motif was mostly explained in the predictive coding chapter, but in general, there are time components to nearly all of our data and if you ignore time you miss out on a lot of the structure of your data. A movie wouldn't be quite as moving if we collapsed each pixel value across time to form a static image. Even when viewing a static image, our sensory stream consists of a sequence of local perceptions as we move our fovea using saccades which are later integrated into a complete perception. Sequential information is meaningful and important, and we do ourselves a disservice by ignoring it.

4.1.4 Don't assume linear scaling

Very often, the absolute value of a measure is not as meaningful as the order or relationship between elements. Sometimes this can be rectified by explicitly rescaling the data if you have some a priori of how it should be scaled, however, other times it is better to consider the space more as a partially ordered set (poset). This is another reason I like KolmogorovSmirnov test and the KolmogorovSmirnov metric, it only requires a partial ordering of the distribution and is invariant to monotonic transformations like taking the square or log of the values.

In the context machine learning and even stochastic gradient descent, an interesting approach would be to train on pairs (or more) of data points to correctly order them and let softmax deal with it in the end. This could allow for more information to be extracted from some metrics that wouldn't necessarily be available if trained only on the labels. Imagine a training set of memes or videos and you want to predict which will go viral. A model that predicts which of two exemplars got more views would likely be able to learn a lot more about what drives virality than a model that simply outputs a probability of a single exemplar going viral.

Another exciting extension of this idea would be to create sorting layers in machine learning. Instead of explicitly taking the max or min of a pool, the output would be all the values in sorted order. Since it doesn't change any values, you wouldn't even have to worry about differentiating the operation, just matching the assignment. This would be very interesting to form a Hero2vec representation for something like a moba. Each hero could be represented as an N dimensional vector representing N unlabeled attributes. The team composition would be represented as a matrix of team members (usually 5) by N attributes with each attribute in sorted order. This would also make the network permutation invariant which could apply to many applications.

4.1.5 Don’t assume normal distributions

This is simply an extension of don’t assume linear scaling. If you know what the scaling or distribution should be you can perform some transformation; however, there are methods like the KolmogorovSmirnov test that only assume a partial ordering of the distribution.

4.2 Conclusions

The recent explosion of progress in machine learning provides an opportunity to solve many long-standing as well as many newly emerging problems in the field of neuroscience. I expect machine learning solutions to engineering problems in neuroscience will drive much progress in the field for a long time, and I am glad to have participated in its inception/revival.

Bibliography

[Atencio et al., 2008] Atencio, C. A., Sharpee, T. O., and Schreiner, C. E. (2008). Cooperative nonlinearities in auditory cortical neurons. *Neuron*, 58(6):956–966.

[Bell and Sejnowski, 1995] Bell, A. J. and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159.

[Binder et al., 2000] Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Springer, J. A., Kaufman, J. N., and Possing, E. T. (2000). Human temporal lobe activation by speech and nonspeech sounds. *Cereb Cortex*, 10(5):512–28. Binder, J RFrost, J AHammeke, T ABellgowan, P SSpringer, J AKaufman, J NPossing, E TP01 MH51358/MH/NIMH NIH HHS/United StatesR01 NS33576/NS/NINDS NIH HHS/United StatesR01 NS35929/NS/NINDS NIH HHS/United StatesComparative StudyResearch Support, U.S. Gov't, P.H.S.United statesCerebral cortex (New York, N.Y. : 1991)Cereb Cortex. 2000 May;10(5):512-28.

[Bluvas and Gentner, 2013] Bluvas, E. C. and Gentner, T. Q. (2013). Attention to natural auditory signals. *Hearing research*, 305:10–18.

[Bonke et al., 1979] Bonke, D., Scheich, H., and Langner, G. (1979). Responsiveness of units in the auditory neostriatum of the guinea fowl (numida meleagris) to species-specific calls and synthetic stimuli. *J Comp Physiol A*, 132(3):1432–1351.

[Brown and Jenkins, 1968] Brown, P. L. and Jenkins, H. M. (1968). Auto-shaping of the pigeon's key-peck. *Journal of the experimental analysis of behavior*, 11(1):1–8.

[Calabrese and Woolley, 2015] Calabrese, A. and Woolley, S. M. (2015). Coding principles of the canonical cortical microcircuit in the avian brain. *Proc Natl Acad Sci U S A*. Calabrese, AnaWoolley, Sarah M NENG2015/02/19 06:00Proc Natl Acad Sci U S A. 2015 Feb 17. pii: 201408545.

[Carney et al., 1977] Carney, A. E., Widin, G. P., and Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in vot. *The Journal of the Acoustical Society of America*, 62(4):961–970.

[Chung et al., 2017] Chung, J. E., Magland, J. F., Barnett, A. H., Tolosa, V. M., Tooker, A. C., Lee, K. Y., Shah, K. G., Felix, S. H., Frank, L. M., and Greengard, L. F. (2017). A fully automated approach to spike sorting. *Neuron*, 95(6):1381–1394.

[Comins and Gentner, 2013] Comins, J. A. and Gentner, T. Q. (2013). Perceptual categories enable pattern generalization in songbirds. *Cognition*, 128(2):113–8. Comins, Jordan AGentner, Timothy QengR01 DC008358/DC/NIDCD NIH HHS/Netherlands2013/05/15 06:00Cognition. 2013 Aug;128(2):113-8. doi: 10.1016/j.cognition.2013.03.014. Epub 2013 May 10.

[Comins and Gentner, 2014a] Comins, J. A. and Gentner, T. Q. (2014a). Auditory temporal pattern learning by songbirds using maximal stimulus diversity and minimal repetition. *Animal cognition*, 17(5):1023–1030.

[Comins and Gentner, 2014b] Comins, J. A. and Gentner, T. Q. (2014b). Temporal pattern processing in songbirds. *Current opinion in neurobiology*, 28:179–187.

[Cooper et al., 1951] Cooper, F. S., Liberman, A. M., and Borst, J. M. (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences of the United States of America*, 37(5):318.

[Csáji, 2001] Csáji, B. C. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24:48.

[Doehrmann et al., 2008] Doehrmann, O., Naumer, M. J., Volz, S., Kaiser, J., and Altmann, C. F. (2008). Probing category selectivity for environmental sounds in the human auditory brain. *Neuropsychologia*, 46(11):2776–2786.

[Dugas-Ford et al., 2012] Dugas-Ford, J., Rowell, J. J., and Ragsdale, C. W. (2012). Cell-type homologies and the origins of the neocortex. *Proc Natl Acad Sci U S A*, 109(42):16974–9. Dugas-Ford, JenniferRowell, Joanna JRagsdale, Clifton WengResearch Support, N.I.H., Extramural2012/10/03 06:00Proc Natl Acad Sci U S A. 2012 Oct 16;109(42):16974-9. doi: 10.1073/pnas.1204773109. Epub 2012 Oct 1.

[Gentner, 2004] 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 et al., 2006] Gentner, T. Q., Fenn, K. M., Margoliash, D., and Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, 440(7088):1204–1207.

[Gentner and Hulse, 1998] Gentner, T. Q. and Hulse, S. H. (1998). Perceptual mechanisms for individual vocal recognition in european starlings, sturnus vulgaris. *Animal behaviour*, 56(3):579–594.

[Gentner et al., 2004] Gentner, T. Q., Hulse, S. H., and Ball, G. F. (2004). Functional differences in forebrain auditory regions during learned vocal recognition in songbirds. *J Comp Physiol A Neuroethol Sens Neural Behav Physiol*, 190(12):1001–10. Gentner, Timothy QHulse, Stewart HBall, Gregory FDC 00389/DC/NIDCD NIH HHS/United StatesR01 35467/PHS HHS/United StatesComparative StudyResearch Support, U.S. Gov't, P.H.S.United StatesJournal of comparative physiology. A, Neuroethology, sensory, neural, and behavioral physiologyJ Comp Physiol A Neuroethol Sens Neural Behav Physiol. 2004 Dec;190(12):1001-10. Epub 2004 Sep 21.

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

[Gill et al., 2008] Gill, P., Woolley, S. M., Fremouw, T., and Theunissen, F. E. (2008). What's that sound? auditory area clm encodes stimulus surprise, not intensity or intensity changes. *J Neurophysiol*, 99(6):2809–20.

[Goldstone and Hendrickson, 2010] Goldstone, R. L. and Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1):69–78.

[Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

[Grace et al., 2003] Grace, J. A., Amin, N., Singh, N. C., and Theunissen, F. E. (2003). Selectivity for conspecific song in the zebra finch auditory forebrain. *J Neurophysiol*, 89(1):472–87. Grace, Julie AAmin, NoopurSingh, Nandini CTheunissen, Frederic EResearch Support, Non-U.S. Gov'tResearch Support, U.S. Gov't, P.H.S.United StatesJournal of neurophysiologyJ Neurophysiol. 2003 Jan;89(1):472-87.

[Gregor and Besse, 2018] Gregor, K. and Besse, F. (2018). Temporal difference variational auto-encoder. *arXiv preprint arXiv:1806.03107*.

[Halevy et al., 2009] Halevy, A., Norvig, P., and Pereira, F. (2009). The unreasonable effectiveness of data.

[Halle and Stevens, 1979] Halle, M. and Stevens, K. N. (1979). Some reflections on the theoretical bases of phonetics. *Frontiers of speech communication research*, pages 335–349.

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

[Hsu et al., 2004] Hsu, A., Woolley, S. M., Fremouw, T. E., and Theunissen, F. E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *J Neurosci*, 24(41):9201–11. Hsu, AnneWoolley, Sarah M NFremouw, Thane ETheunissen, Frederic EResearch Support, U.S. Gov't, P.H.S.United StatesThe Journal of neuroscience : the official journal of the Society for NeuroscienceJ Neurosci. 2004 Oct 13;24(41):9201-11.

[Itti and Baldi, 2005] Itti, L. and Baldi, P. (2005). A principled approach to detecting surprising events in video. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 631–637. IEEE.

[Jarvis et al., 2005] Jarvis, E. D., Güntürkün, O., Bruce, L., Csillag, A., Karten, H., Kuenzel, W., Medina, L., Paxinos, G., Perkel, D. J., Shimizu, T., Striedter, G., Wild, J. M., Ball, G. F., Dugas-Ford, J., Durand, S. E., Hough, G. E., Husband, S., Kubikova, L., Lee, D. W., Mello, C. V., Powers, A., Siang, C., Smulders, T. V., Wada, K., White, S. A., Yamamoto, K., Yu, J., Reiner, A., Butler, A. B., and Consortium, A. B. N. (2005). Avian brains and a new understanding of vertebrate brain evolution. *Nature Reviews Neuroscience*, 6(2):151–9.

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

[Jeanne et al., 2011b] Jeanne, J. M., Thompson, J. V., Sharpee, T. O., and Gentner, T. Q. (2011b). Emergence of learned categorical representations within an auditory forebrain circuit. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 31(7):2595–606. Jeanne, James MThompson, Jason VSharpee, Tatyana OGentner, Timothy QDC008358/DC/NIDCD NIH HHS/MH068904/MH/NIMH NIH HHS/R01 DC008358-02/DC/NIDCD NIH HHS/R01 DC008358-03/DC/NIDCD NIH HHS/R01 DC008358-04/DC/NIDCD NIH HHS/R01 EY019493-03/EY/NEI NIH HHS/R01EY019493/EY/NEI NIH HHS/J Neurosci. 2011 Feb 16;31(7):2595-606.

[Kaas and Hackett, 2000] Kaas, J. H. and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc Natl Acad Sci U S A*, 97(22):11793–9. Kaas, J HHackett, T AUnited statesProceedings of the National Academy of Sciences of the United States of AmericaProc Natl Acad Sci U S A. 2000 Oct 24;97(22):11793-9.

[Karten, 1968] Karten, H. J. (1968). The ascending auditory pathway in the pigeon (columba livia). ii. telencephalic projections of the nucleus ovoidalis thalami. *Brain Res*, 11(1):134–53.

[Kawahara and Matsui, 2003] Kawahara, H. and Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

[Knudsen and Gentner, 2013] Knudsen, D. P. and Gentner, T. Q. (2013). Active recognition enhances the representation of behaviorally relevant information in single auditory forebrain neurons. *Journal of neurophysiology*, 109(7):1690–1703.

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

[Kuhl and Miller, 1975] Kuhl, P. K. and Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science*, 190(4209):69–72.

[Kuhl and Padden, 1983] Kuhl, P. K. and Padden, D. M. (1983). Enhanced discriminability at the phonetic boundaries for the place feature in macaques. *The Journal of the Acoustical Society of America*, 73(3):1003–1010.

[Lachlan and Nowicki, 2015] Lachlan, R. F. and Nowicki, S. (2015). Context-dependent categorical perception in a songbird. *Proceedings of the National Academy of Sciences*, 112(6):1892–1897.

[Lainscsek and Sejnowski, 2015] Lainscsek, C. and Sejnowski, T. J. (2015). Delay differential analysis of time series. *Neural computation*, 27(3):594–614.

[Lakshminarayanan et al., 2017] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413.

[Leaver and Rauschecker, 2010] Leaver, A. M. and Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *Journal of Neuroscience*, 30(22):7604–7612.

[Leppelsack and Vogt, 1976] Leppelsack, H. J. and Vogt, M. (1976). Response to auditory neurons in the forebrainof a song bird to stimulation with species-specific sounds. *Journal of Comparative Physiology*, 107:263–274.

[Liberman et al., 1967] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74(6):431.

[Liberman et al., 1957] Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358.

[Massaro and Cohen, 1983] Massaro, D. W. and Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech communication*, 2(1):15–35.

[McMurray et al., 2008] McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., and Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, 34(6):1609.

[McMurray et al., 2002] McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, 86(2):B33–B42.

[Meliza et al., 2010] Meliza, C. D., Chi, Z., and Margoliash, D. (2010). Representations of conspecific song by starling secondary forebrain auditory neurons: toward a hierarchical framework. *J Neurophysiol*, 103(3):1195–208. Meliza, C DanielChi, ZhiyiMargoliash, DanielDC-007206/DC/NIDCD NIH HHS/United StatesF32 DC-008752/DC/NIDCD NIH HHS/United StatesResearch Support, N.I.H., ExtramuralResearch Support, U.S. Gov't, Non-P.H.S.United StatesJournal of neurophysiologyJ Neurophysiol. 2010 Mar;103(3):1195-208. Epub 2009 Dec 23.

[Miller, 1997] Miller, J. L. (1997). Internal structure of phonetic categories. *Language and cognitive processes*, 12(5-6):865–870.

[Muller and Leppelsack, 1985] Muller, C. M. and Leppelsack, H. J. (1985). Feature extraction and tonotopic organization in the avian auditory forebrain. *Exp Brain Res*, 59(3):587–99.

Muller, C MLeppelsack, H JResearch Support, Non-U.S. Gov'tGermany, westExperimental brain research. Experimentelle Hirnforschung. Experimentation cerebraleExp Brain Res. 1985;59(3):587-99.

[Nelson and Marler, 1989] Nelson, D. A. and Marler, P. (1989). Categorical perception of a natural stimulus continuum: birdsong. *Science*, 244(4907):976–978.

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[Phan et al., 2006] Phan, M. L., Pytte, C. L., and Vicario, D. S. (2006). Early auditory experience generates long-lasting memories that may subserve vocal learning in songbirds. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4):1088–93. Phan, Mimi LPytte, Carolyn LVicario, David SProc Natl Acad Sci U S A. 2006 Jan 24;103(4):1088-93. Epub 2006 Jan 17.

[Pisoni and Tash, 1974] Pisoni, D. B. and Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & psychophysics*, 15(2):285–290.

[Prather et al., 2009] Prather, J. F., Nowicki, S., Anderson, R. C., Peters, S., and Mooney, R. (2009). Neural correlates of categorical perception in learned vocal communication. *Nature neuroscience*, 12(2):221.

[Prenger et al., 2019] Prenger, R., Valle, R., and Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. IEEE.

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

[Rose et al., 1988] Rose, G. J., Kawasaki, M., and Heiligenberg, W. (1988). 'recognition units' at the top of a neuronal hierarchy? prepacemaker neurons in eigenmannia code the sign of frequency differences unambiguously. *J Comp Physiol A*, 162(6):759–72. Rose, G JKawasaki, MHeiligenberg, W1 R01 NS 22244-02/NS/NINDS NIH HHS/United States2 R01-MH 26149-12/MH/NIMH NIH HHS/United StatesNS 22740-02/NS/NINDS NIH HHS/United States-Research Support, U.S. Gov't, Non-P.H.S.Research Support, U.S. Gov't, P.H.S.Germany, westJournal of comparative physiology. A, Sensory, neural, and behavioral physiologyJ Comp Physiol A. 1988 Apr;162(6):759-72.

[Sainburg et al., sion] Sainburg, T., Theilman, B., Thielk, M., and Gentner, T. (2019, in submission). Parallels in the sequential organization of birdsong and human speech.

[Sainburg et al., 2018] Sainburg, T., Thielk, M., Theilman, B., Migliori, B., and Gentner, T. (2018). Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. *arXiv preprint arXiv:1807.06650*.

[Schouten et al., 2003] Schouten, B., Gerrits, E., and Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech communication*, 41(1):71–80.

[Sockman et al., 2002] Sockman, K. W., Gentner, T. Q., and Ball, G. F. (2002). Recent experience modulates forebrain gene-expression in response to mate-choice cues in european starlings. *Proc R Soc Lond B Biol Sci*, 269(1508):2479–85. 0962-8452Journal Article.

[Sockman et al., 2005] Sockman, K. W., Gentner, T. Q., and Ball, G. F. (2005). Complementary neural systems for the experience-dependent integration of mate-choice cues in european starlings. *J Neurobiol*, 62(1):72–81. Sockman, Keith WGentner, Timothy QBall, Gregory Feng41854/PHS HHS/R0135467/PHS HHS/Comparative StudyResearch Support, U.S. Gov't, P.H.S.2004/09/25 05:00J Neurobiol. 2005 Jan;62(1):72-81.

[Stevens, 1981] Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics. In *Advances in psychology*, volume 7, pages 61–74. Elsevier.

[Studdert-Kennedy et al., 1970] Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., and Cooper, F. S. (1970). Motor theory of speech perception: A reply to lane's critical review.

[Sun et al., 2017] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852.

[Theunissen et al., 2004] Theunissen, F. E., Amin, N., Shaevitz, S. S., Woolley, S. M., Fremouw, T., and Hauber, M. E. (2004). Song selectivity in the song system and in the auditory forebrain. *Annals of the New York Academy of Sciences*, 1016:222–45.

[Theunissen and Doupe, 1998] Theunissen, F. E. and Doupe, A. J. (1998). Temporal and spectral sensitivity of complex auditory neurons in the nucleus hvc of male zebra finches. *J Neurosci*, 18(10):3786–802.

[Theunissen et al., 2000] Theunissen, F. E., Sen, K., and Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 20(6):2315–31. Theunissen, F ESen, KDoupe, A JMH 11209/MH/NIMH NIH HHS/MH 59189/MH/NIMH NIH HHS/NS 34385/NS/NINDS NIH HHS/J Neurosci. 2000 Mar 15;20(6):2315-31.

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

[Thompson and Gentner, 2010b] Thompson, J. V. and Gentner, T. Q. (2010b). Song recognition learning and stimulus-specific weakening of neural responses in the avian auditory forebrain. *J Neurophysiol*, 103(4):1785–97.

[Van Den Oord et al., 2016] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. W., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *SSW*, 125.

[Van Essen et al., 1992] Van Essen, D. C., Anderson, C. H., and Felleman, D. J. (1992). Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–23. Van Essen, D CAnderson, C HFelleman, D JResearch Support, Non-U.S. Gov'tResearch Support, U.S. Gov't, Non-P.H.S.Research Support, U.S. Gov't, P.H.S.ReviewUnited statesScience (New York, N.Y.)Science. 1992 Jan 24;255(5043):419-23.

[Van Hessen and Schouten, 1999] Van Hessen, A. J. and Schouten, M. E. H. (1999). Categorical perception as a function of stimulus quality. *Phonetica*, 56(1-2):56–72.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

[Wang et al., 2010] Wang, Y., Brzozowska-Prechtl, A., and Karten, H. J. (2010). Laminar and columnar auditory cortex in avian brain. *Proceedings of the National Academy of Sciences*, 107(28):12676–12681.

[Wang and Karten, 2010] Wang, Y. and Karten, H. J. (2010). Three subdivisions of the auditory midbrain in chicks (gallus gallus) identified by their afferent and commissural projections. *J Comp Neurol*, 518(8):1199–219. Wang, YuanKarten, Harvey JDC-00018/DC/NIDCD NIH HHS/United StatesDC-03829/DC/NIDCD NIH HHS/United StatesDC-04661/DC/NIDCD NIH HHS/United StatesNH 60975-07/PHS HHS/United StatesNS 24560-15/NS/NINDS NIH HHS/United StatesP20 MH060975-07/MH/NIMH NIH HHS/United StatesP30 DC004661-10/DC/NIDCD NIH HHS/United StatesR01 DC003829-04/DC/NIDCD NIH HHS/United StatesR01 NS024560-15/NS/NINDS NIH HHS/United StatesResearch Support, N.I.H., ExtramuralUnited StatesThe Journal of comparative neurologyJ Comp Neurol. 2010 Apr 15;518(8):1199-219.

[Webster, 1992] Webster, D. B., F. R. R. P. A. N., editor (1992). *The Evolutionary Biology of Hearing*, chapter Evolution of the Central Auditory System in Reptiles and Birds, pages 511–544. Springer - Verlag, New York.

[Woolley et al., 2005] Woolley, S. M., Fremouw, T. E., Hsu, A., and Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nat Neurosci*, 8(10):1371–9. Woolley, Sarah M NFremouw, Thane EHsu, AnneTheunissen, Frederic EComparative StudyResearch Support, N.I.H., ExtramuralResearch Support, U.S. Gov't, P.H.S.United StatesNature neuroscienceNat Neurosci. 2005 Oct;8(10):1371-9. Epub 2005 Sep 4.

[Zhang, 2013] Zhang, Y. (2013). Categorical perception. *Unpublished encyclopedia article manuscript*.