UNIVERSITY OF CALIFORNIA SAN DIEGO

Temporal organization in vocal communication: sequential structure, perceptual integration, and neural foundations

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

in

Experimental Psychology with a Specialization in Anthropogeny

by

Tim Sainburg

Committee in charge:

Professor Timothy Q. Gentner, Chair
Professor Vikash Gilja
Professor Cory Miller
Professor Eran Mukamel
Professor Terrence Sejnowski
Professor Ed Vul

2021

The Dissertation of Tim Sainburg is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

TABLE OF CONTENTS

Dissertation Approval Page iii

Table of Contents iv

List of Figures ix

List of Tables xiv

Acknowledgements xv

Vita xviii

Abstract of the Dissertation xix

Chapter 1 Towards a computational neuroethology of vocal communication 1
1.1 Introduction 1
1.2 Signal processing and denoising 3
1.3 Signal representation 7
1.4 Identifying, segmenting, and labeling vocalizations 10
1.5 Extracting relational structure and clustering 15
1.6 Inferring temporal and sequential structure 22
1.7 Synthesizing vocalizations 28
1.8 Mapping vocal communication to perception, behavior, and physiology 33
1.9 Discussion 39
1.10 Acknowledgments 42

Chapter 2 Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires 43
2.1 Introduction 44
2.1.1 Latent models of acoustic communication 45
2.2 Results 48
2.2.1 Dimensionality reduction 48
2.2.2 Choosing features to represent vocalizations 51
2.2.3 Discrete latent projections of animal vocalizations 54
2.2.4 Temporally continuous latent trajectories 75
2.3 Discussion 85
2.3.1 Limitations 89
2.3.2 Future work 92
2.4 Methods 95
2.4.1 Datasets 95
2.4.2 Reducing noise in audio 95
2.4.3 Segmentation 97
2.4.4 Spectrogramming 99

2.4.5 Projections . . . 99
2.4.6 Clusterability . . . 101
2.4.7 Clustering vocalizations . . . 103
2.4.8 Comparing algorithmic and hand-transcriptions . . . 104
2.4.9 Hidden Markov Models (HMMs) . . . 105
2.4.10 Data Availability . . . 105
2.4.11 Code Availability . . . 105
2.4.12 Ethics statement . . . 105
2.5 Supporting information . . . 106
2.6 Acknowledgments . . . 113

Chapter 3 Parametric UMAP . . . 114
3.1 Introduction . . . 114
3.2 Parametric and non-parametric UMAP . . . 116
3.2.1 Graph Construction . . . 116
3.2.2 Graph Embedding . . . 118
3.2.3 Attraction and repulsion . . . 119
3.2.4 Parametric UMAP . . . 120
3.3 Related Work . . . 121
3.4 UMAP as a regularization . . . 122
3.4.1 Autoencoding with UMAP . . . 123
3.4.2 Semi-supervised learning . . . 123
3.4.3 Preserving global structure . . . 124
3.5 Experiments . . . 125
3.5.1 Embeddings . . . 126
3.5.2 Training and embedding speed . . . 129
3.5.3 Capturing additional global structure in data . . . 131
3.5.4 Autoencoding with UMAP . . . 135
3.5.5 Semi-supervised learning . . . 137
3.5.6 Comparisons with indirect parametric embeddings . . . 140
3.6 Discussion . . . 144
3.7 Acknowledgments . . . 145

Chapter 4 Parallels in the sequential organization of birdsong and human speech . . . 146
4.1 Introduction . . . 147
4.2 Results . . . 148
4.2.1 Modeling . . . 148
4.2.2 Speech . . . 152
4.2.3 Birdsong . . . 155
4.3 Discussion . . . 158
4.4 Methods . . . 161
4.4.1 Birdsong data sets. . . . 161
4.4.2 Speech corpora. . . . 162
4.4.3 Corpus annotation for European starlings. . . . 163

4.4.4 Song bouts. 166
4.4.5 Mutual information estimation. 166
4.4.6 Mutual information decay fitting. 167
4.4.7 Model selection. 168
4.4.8 Curvature of decay fits. 169
4.4.9 Sequence analyses. 169
4.4.10 Computational models. 170
4.5 appendix 172
4.5.1 AICc 172
4.6 Acknowledgments 194

Chapter 5 Long-range sequential dependencies precede complex syntactic production in language acquisition 195
5.1 Introduction 196
5.2 Methods 199
5.2.1 Datasets 199
5.2.2 Mutual information 200
5.2.3 Fitting mutual information decay 202
5.2.4 Controls 203
5.3 Results 203
5.4 Discussion 207
5.4.1 Data Availability 209
5.4.2 Author contributions 209
5.5 Acknowledgments 216

Chapter 6 Prediction and probabilistic integration underlie learned context-dependent categorical vocal sequence perception and sensory physiology 217
6.1 Introduction 218
6.2 Results 220
6.2.1 Paradigm 220
6.2.2 Context dependent shift in perceptual decision making 222
6.2.3 Context dependent perceptual shift increases with uncertainty 223
6.2.4 Reaction time represent likelihood and prior probability 224
6.2.5 Physiology paradigm 228
6.2.6 Quantifying response similarity and estimating a neurometric 229
6.2.7 Neurometric slope reflects psychometric uncertainty 229
6.2.8 Within subject perceptual variability is reflected in neural response 231
6.2.9 Context modulates neural response 231
6.2.10 Expectation suppresses spike rate in predicted stimuli 233
6.2.11 Predictive response modulation is consistent with a shift in the likelihood of Bayesian model 234
6.3 Discussion 237
6.3.1 How is predictive information actively maintained and integrated? 237
6.3.2 How do populations of neurons represent predictive information? 238

6.3.3 Is there a distinction between categorical perception and perceptual decision making? . . . 239
6.3.4 How do prediction, attention, and integration differ? . . . 239
6.3.5 Are more natural stimulus spaces better poised for probing the complexities of vocal communication? . . . 240
6.3.6 Final note . . . 240
6.4 Methods . . . 241
6.4.1 Summary . . . 241
6.4.2 Subjects . . . 241
6.4.3 Ethical note . . . 241
6.4.4 Datasets . . . 241
6.4.5 Stimulus generation . . . 243
6.4.6 Training dataset . . . 243
6.4.7 Neural network . . . 244
6.4.8 Sampling and synthesis . . . 244
6.4.9 Behavioral shaping . . . 245
6.4.10 Behavioral training paradigm . . . 245
6.4.11 Training parameters . . . 246
6.4.12 Cue stimulus . . . 246
6.4.13 Psychometric fit . . . 247
6.4.14 Bayesian integration hypothesis . . . 247
6.4.15 Bayesian fit . . . 248
6.4.16 Response time . . . 249
6.4.17 Chronic electrophysiology . . . 250
6.4.18 Behavioral neural acquisition interfacing with PiOperant . . . 250
6.4.19 Microdrives and head caps . . . 250
6.4.20 Electrode implant procedure . . . 251
6.4.21 Recordings and behavior blocks . . . 251
6.4.22 Chronic behavior blocks . . . 252
6.4.23 Chronic passive playback blocks . . . 252
6.4.24 Spikesorting and merging over long-term chronic recordings . . . 252
6.4.25 Stimulus alignment . . . 253
6.4.26 Acute recording sessions . . . 253
6.4.27 Localizing units . . . 253
6.4.28 Clustering unit spike shapes . . . 254
6.4.29 Neural feature representation and response similarity . . . 255
6.4.30 Estimating a neurometric function from the similarity matrix . . . 257
6.4.31 Categoricality metric . . . 257
6.4.32 Subsetting categorical units . . . 258
6.4.33 Comparing spike rate across units, cues, and morphs . . . 258
6.4.34 Morph class and cue interactions by subject, brain region, unit type, and morph . . . 261
6.4.35 Differences in spike rate as a function of time . . . 262
6.4.36 Within cue response similarity . . . 262

6.4.37 Data and code availability 267
6.4.38 Acknowledgments 267

Bibliography 268

LIST OF FIGURES

Figure 1.1. Stationary and non-stationary spectral gating noise reduction. 3

Figure 1.2. Examples of several different spectral representations of a five-second red deer (*Cervus elaphus*) vocalization. 8

Figure 1.3. Levels of organization and architectural design features for identifying, segmenting, and labeling vocalizations. 11

Figure 1.4. Local and global embeddings. (A) The steps outlined in Section 1.5 exhibit the differences between the relationships preserved in local and global embeddings. 16

Figure 1.5. Capturing long and short-range sequential organization with different models. 23

Figure 1.6. Steps involved in synthesizing vocalization from a Variational Autoencoder (VAE) trained on spectrograms. 28

Figure 1.7. An outline of mappings between perceptual, acoustic, and physiological signals. 33

Figure 2.1. Graph-based dimensionality reduction. 47

Figure 2.2. Comparison between dimensionality reduction and manifold learning algorithms. 50

Figure 2.3. Comparison between dimensionality reduction on spectrograms versus computed features of syllables. 52

Figure 2.4. Individual identity is captured in projections for some datasets. 56

Figure 2.5. Comparing species with latent projections. 58

Figure 2.6. Comparing notes of swamp sparrow song across different geographic populations. 60

Figure 2.7. Latent projections of consonants. 61

Figure 2.8. UMAP projections of vocal repertoires across diverse species. 62

Figure 2.9. HDBSCAN density-based clustering. 65

Figure 2.10. Clustered UMAP projections of Cassin's vireo syllable spectrograms. 67

Figure 2.11. Comparing latent and known features in swamp sparrow song. 69

Figure 2.12. Latent visualizations of Bengalese finch song sequences. 71

Figure 2.13. Latent comparisons of hand- and algorithmically-clustered Bengalese finch song. 73

Figure 2.14. Comparison of Hidden Markov Model performance using different hidden states. 76

Figure 2.15. Continuous UMAP projections of Bengalese finch song from a single bout produced by one individual. 77

Figure 2.16. Starling bouts projected into continuous UMAP space. 79

Figure 2.17. USV patterns revealed through latent projections of a single mouse vocal sequence. 82

Figure 2.18. Speech trajectories showing coarticulation in minimal pairs. 84

Figure 2.19. Segmentation algorithm (A) The dynamic threshold segmentation algorithm. 97

Figure 2.20. Continuous projections from vocalizations. (A) A spectrogram of each vocalization is computed. 100

Figure 2.21. UMAP projections Cassin's vireo syllables with syllable features overlaid generated from the BioSound python package. 107

Figure 2.22. Example vocal elements from each of the species used in this paper. 108

Figure 2.23. Latent projections of vowels. 109

Figure 2.24. Comparing latent and known features in swamp sparrow song. 110

Figure 2.25. Comparison of Hidden Markov Model performance using different latent states. 111

Figure 2.26. Silhouette score of UMAP projections with different levels background noise added to spectrogram. 111

Figure 3.2. Varients of UMAP used in this paper. 123

Figure 3.3. An example of semi-supervised learning with UMAP on the moons dataset. 124

Figure 3.4. Comparison of projections from multiple datasets using UMAP, UMAP in Tensorflow, ... 125

Figure 3.5. Embedding metrics for 2D projections. 127

Figure 3.6. Training times comparison between UMAP and Parametric UMAP. 129

Figure 3.7. Comparison of embedding speeds using parametric UMAP and other embedding algorithms on a held-out testing dataset. 130

Figure 3.8. Global loss applied to Parametric UMAP embeddings with different weights. 132

Figure 3.9. Comparison of pairwise global and local relationship preservation across embeddings for MNIST and Macosko [262] 134

Figure 3.10. Reconstruction accuracy measured as mean squared error (MSE). 135

Figure 3.11. Reconstruction and interpolation. 136

Figure 3.12. Baseline classifier with an additional UMAP loss with different numbers of labeled training examples. 138

Figure 3.13. Comparison of baseline classifier and augmentation 139

Figure 3.14. Non-parametric UMAP projections of activations in the last layer of a trained classifier for MNIST, FMNIST, and CIFAR10. 141

Figure 3.15. Cross entropy UMAP loss 143

Figure 4.1. Latent and graphical representations of songbird vocalizations. 150

Figure 4.2. MI decay of sequences generated by three classes of models. 151

Figure 4.3. Mutual information decay in human speech. 153

Figure 4.4. Mutual information decay in birdsong. 156

Figure 4.5. Utterance length in phones for English and Japanese 175

Figure 4.6. MI decay between phones in shuffled speech for different languages 176

Figure 4.7. MI decay between words, syllables, mora, and parts-of-speech plotted as a function of sequential distances between each of these elements in three languages. 177

Figure 4.9. Mutual information decay between syllables in the songs of four songbird species 179

Figure 4.10. MI decay in the four largest data sets from individual songbirds in each species. 180

Figure 4.11. Relative decay model fits. 182

Figure 4.12. The decay in MI between syllables in the 18 individual songbirds with the longest available recordings in all data sets. Vireo and thrasher decay. 183

Figure 4.13. The intersyllable interval time in seconds for each songbird species. 185

Figure 4.14. The goodness of fit of the composite decay model for each language as a function of the MI analysis length. 186

Figure 4.15. The goodness of fit of the composite decay model for each songbird species as a function of the MI analysis length. 187

Figure 4.16. Decay of MI between syllables in the birdsong data sets after removing sequentially repeated syllables. 188

Figure 4.17. Decay in MI between song and speech signal components arbitrarily parsed at multiple timescales. 189

Figure 5.1. Comparison of long-range statistical dependencies between sequences with and without deep latent relationships. 197

Figure 5.2. Mutual Information decay over words and phonemes during development. 204

Figure 5.3. MI decay between phones under different shuffling conditions. 205

Figure 5.4. Distribution of sequence lengths for each dataset. 210

Figure 5.5. MI decay between phones under different shuffling conditions. 211

Figure 5.6. MI decay between words under different shuffling conditions. 212

Figure 5.7. MI decay with repeated elements removed across each dataset. 213

Figure 5.8. MI decay and best fit model of five largest transcripts for each age group across PhonBank. 214

Figure 5.9. MI decay and best fit model of five largest transcripts for each age group across CHILDES. 215

Figure 6.1. Overview of behavior and hypothesis. 221

Figure 6.2. Overview of behavioral results. 225

Figure 6.3. Overview of physiological paradigm and data set. 227

Figure 6.4. Neurometric functions of single units reflect psychometric functions of perceptual behavior. 230

Figure 6.5. Predictive syllables modulate response to morph syllable. 232

Figure 6.6. Modulation of response similarity as a function of predictive cue probability. 236

Figure 6.7. Sample fit of response time decay for a single morph (AE) for a single bird (B1174). 249

Figure 6.8. Spike widths and rates for each unit type. 254

Figure 6.9. Comparison of unit categoricality using cosine similarity, ... 256

Figure 6.10. Method for computing a neurometric function from a similarity matrix. . . 257

Figure 6.11. Categorical and non-categorical units, sorted by categoricality 258

Figure 6.12. Spike rate modulation by cue. 260

Figure 6.13. Interaction between cue probability and morph stimulus class on spike rate. 261

Figure 6.14. Spike rate differences between minus within cue categories over time. 262

Figure 6.15. Spike vector cosine similarity and shift in similarity as a function of class probability. 264

Figure 6.16. Morph, brain region, subject, and unit type similarity. 265

Figure 6.17. The relationship between the probability of the stimulus class and the shift in similarity from baseline. 266

LIST OF TABLES

Table 2.1. Cluster similarity to hand labels for two Bengalese finch and one Cassin's vireo dataset. 66

Table 2.2. Overview of the species and datasets used in this paper. 106

Table 2.3. BioSound features used in feature statistics analysis. 112

Table 4.1. Birdsong dataset statistics. 191

Table 4.2. Language dataset statistics. 192

Table 4.3. Language corpus model fit results at 100 phones of distance. 193

Table 4.4. Birdsong dataset model fit results at 100 syllables of distance. 194

Table 6.1. Behavioral datasets 242

Table 6.2. Neural datasets 243

Table 6.3. Variational autoencoder architecture outline 244

ACKNOWLEDGEMENTS

I would first like to thank my advisor, Tim Gentner for everything he has done for me in the past six years. I'm deeply grateful for both the guidance and freedom he gave me, without which, this dissertation would not have been possible.

Many thanks to my labmates, whose diversity in thoughts, ideas, and knowledge were essential to pursuing questions I would have otherwise been unequipped to ask. Marvin for his guidance and willingness to teach at the drop of a hat. Zeke for helping me learn chronic ephys. Anna for helping me to understand the linguist's perspective. Michael, who has been alongside me from boot camp to defense. Trevor for helping me finish strong and keeping those chronic rigs running. Brad for helping me design and build the chronic rigs. Srihita, Pablo, Daril, Kai, Lauren, Jeffrey, and everyone else for their support.

A special thanks to the Anthropogeny program, especially Pascal Gagneux, for his tireless effort and unparalleled dedication to his students.

I am grateful to the advisers who gave me the opportunities that led to this Ph.D. My undergraduate advisor Dan Weiss who gave me an excellent first research experience which ultimately led to this Ph.D. in animal communication nearly a decade later. My post-bacc advisor Cat Hobaiter and the Budongo Research Field Station who gave me the incredible opportunity to study in the field how our closest living relatives live and communicate. My post-bacc advisors John Sustersic and Brad Wyble, who gave me the opportunity to explore topics in machine perception and learning that continue to shape my ideas and interests today.

A big thanks to the UCSD Psychology department and staff, especially Danny, Todd, Silas, and Thanh in IT for helping me keep everything running.

Thank you to my dissertation committee, for their encouragement and insights over the past years.

I would also like to acknowledge my animal research subjects for their role in this Ph.D.

Thanks also to all of the friends that I've made here in San Diego, who have provided me with support and made California home. Jarrett, Rocio, Drew, Meredith, Taylor, Tommy,

Brendan, Ross, Maddie, Bryan, Tim, my cycling friends, surfing friends, basketball friends, volleyball friends, and Ph.D. cohort.

Many thanks to my parents for instilling into me from a young age an interest in science and the support needed to pursue it.

Finally, I thank my wife, Elisabeth, who I met on day 1 of my Ph.D. and married around day 2200. Her support has been invaluable.

The research in this thesis was supported by an NSF GRFP (2017216247) an Institute for Neural Computation Training Fellowship (5T32MH020002-20), a William Orr Dingwall dissertation fellowship, a CARTA fellowship, and an Anette Merle-Smith fellowship.

Chapter 1, in full, is a reprint of a manuscript under review. Sainburg, Tim, Gentner, Timothy Q (2021) The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reprint of the material as it appears in PLOS Computational Biology, 2020, Sainburg, Tim, Thielk, Marvin, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Neural Computation, 2021, Sainburg, Tim, McInnes, Leland, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Nature Communications, 2019, Sainburg, Tim, Theilman, Brad, Thielk, Marvin, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is a reprint of a manuscript under review. Sainburg, Tim, Mai, Anna, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 6, in full, is a reprint of a manuscript in preparation. Sainburg, Tim, McPherson, Trevor S, Arneodo, Ezequiel M., Rudraraju, Srihita, Turvey, Michael, Thielman, Brad, Marcos, Pablo Tostado, Thielk, Marvin, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

VITA

2014 Bachelor of Arts, Psychology minor in Biology, The Pennsylvania State University

2021 Doctor of Philosophy, Experimental Psychology with a specialization in Anthropogeny, University of California San Diego

ABSTRACT OF THE DISSERTATION

Temporal organization in vocal communication: sequential structure, perceptual integration, and neural foundations

by

Tim Sainburg

Doctor of Philosophy in Experimental Psychology with a Specialization in Anthropogeny

University of California San Diego, 2021

Professor Timothy Q. Gentner, Chair

Our interactions with the world unfold over time. Whether it's speaking, where one word follows the next, or walking, where each step follows another, the organization of our behaviors in time tends to follow a predictable pattern. Those patterns are dictated by a multitude of underlying factors, influenced both by endogenous physiological factors like the rhythmic nature of our gait as well as by exogenous factors, like the social dynamics underlying turn-taking while speaking. Despite decades of research studying the temporal organization of behavior, dating back to the work of influential biologists like Tinbergen, Lashley, and Dawkins, little is known about the physiological substrates that underlie either the production of the sequential organization of most

aspects of behavior. Despite widespread acknowledgment that physiological motor programs and many non-linguistic behaviors are hierarchical, for example, few physiological investigations into the dynamics of behavior extend beyond low-order (Markovian) transition statistics.

In this thesis, I build onto the emerging field of computational neuroethology to further our understanding of what structure underlies the sequential organization of behavior, what physiological mechanisms might be involved in producing, perceiving, and representing sequential behavioral organization, and how sequential behavioral organization might have emerged developmentally and evolutionarily. Throughout the thesis, I draw primarily upon birdsong and human speech, developing methods to analyze the acoustic and temporal structure in vocal signals and then behaviorally and physiologically probing the underpinnings of sequential organization in the songbird. This work advances the field of computational neuroethology in several ways. I uncover novel acoustic structure in vocal signals separating avian and mammalian vocalizations along a spectrum of vocal stereotypy. I observe that both human speech and birdsong are characterized by a combination of long and short-range temporal patterning. I find that the long-range temporal patterning characterizing human speech, believed to be underlied by hierarchical linguistic organization, is present at the earliest developmental stages of human speech, well before complex syntax is produced. I find that the perceptual integration of birdsong syllable sequences can be well explained by Bayesian models of probabilistic perceptual decision-making. Finally, I find that sensory neural representations of syllable sequences are modulated by sequential context and that this modulation reflects the animals underlying perceptual behavior. In the following paragraphs, I give a brief overview of the methods and major results of the chapters comprising this thesis.

In Chapter 1 I give an introduction to the emerging field of vocal computational neuroethology. This introduction contextualizes the following chapters in a review of current work. I emphasize current tools, challenges, and future directions in vocal neuroethology. I start with a discussion of low-level bioacoustics challenges and build up to a discussion of behavioral organization and physiology. I first discuss challenges in signal processing such as dealing

with noise and signals and representing vocal signals as time-frequency representations. I then discuss machine learning approaches used to identify, segment, and label vocalizations. Next, I discuss how to extract relational structure between vocalizations, and cluster latent projections of vocalizations. I then give an overview of methods for capturing temporal relationships in vocal sequences, outlining traditional Markovian descriptions of vocal structure, and new tools for capturing long-range structure, enabled by large datasets. I then move on to machine learning tools that can be used to systematically control and synthesize vocal signals from learned vocal spaces. Finally, I discuss how these techniques are being utilized in several active areas of neuroethology research.

In Chapter 2 I develop a set of methods to visualize and quantify relational structure in vocalizations, which enable the analyses and experiments performed in the following chapters. I use graph-based dimensionality reduction to uncover local structure in vocal communication signals and apply that technique to 19 datasets consisting of vocalizations from 29 species, including songbirds, primates, cetaceans, rodents, and bats. I observe that these methods uncover novel structure in animal vocal signals, including vocal dialects, acoustic units, behaviorally relevant signal information, and sub-syllabic structure.

In Chapter 3, I extend the methods from Chapter 2 by introducing Parametric UMAP, a graph-based dimensionality reduction algorithm that parametrically learns the relationship between data (here vocal signals) and latent embeddings. Parametric UMAP enables the methods from Chapter 2 to be applied in real-time closed-looped settings over larger datasets due to the learned parametric embeddings. I show that this algorithm has applications in semi-supervised settings, and provides additional control over the trade-off between capturing global and local structure in embeddings.

In Chapter 4 I explore the long and short-range temporal patterning of vocal sequences in birdsong and human speech. I use an information-theoretic framework to analyze statistical dependencies as a function of the distance between elements in vocal sequences. I find that both birdsong and human speech exhibit two forms of structure: short-range relationships captured by

Markovian dynamics over short-timescales, and long-range relationships that follow a power-law occurring over longer timescales. In language, the observed short-range organization conforms to phonological processes, which are well-described by finite-state dynamics, while long-range organization suggests more complex dynamics such as underlying hierarchical organization. Previous analyses of birdsong have only identified short-range Markovian dynamics, making our observation of long-range dynamics in birdsong novel.

In Chapter 5 I extend our experiment from chapter 4 over human speech to language acquisition. By analyzing corpora of speech throughout language development, we can observe the time course of the emergence of long and short-range relationships over development. Surprisingly, I find that long-range statistical dependencies are present in children's speech as early as 6-12 months, well before complex syntactic structure is present. I discuss these results alongside emerging evidence from computational ethology that long-range relationships are also common to non-linguistic behavioral signals from animals as diverse as zebrafish, drosophila, and whales. Although previous analyses of long-range relationships have suggested that long-range relationships are the product of hierarchical linguistic structure such as syntax and discourse structure, our observations in developmental speech and non-linguistic behaviors suggest that other mechanisms may also be at play.

Finally, in Chapter 6 I probe how sequential dependencies in vocal sequences are integrated behaviorally and physiologically. I developed a behavioral task in which European starlings are trained to classify morphs of syllables of starling song synthesized from an interpolation between two points in the latent space of a neural network (a Variational Autoencoder). These morph syllables are preceded with a separate syllable (a cue syllable), which holds predictive information about the category of the following morph syllable. I find that classification of the morph syllable is contextually modulated by the predictive probability of the cue syllable, which can be well explained by a model of Bayesian integration. With the same behavioral paradigm, I then record chronic electrophysiology data from auditory nuclei while birds performed this context-dependent categorical perceptual decision-making task. I find that

neural activity patterns reflect several aspects of our model of perceptual behavior, including the uncertainty in decision making, and prediction-related perceptual modulation.

Chapter 1

Towards a computational neuroethology of vocal communication

Abstract

Recently developed methods in computational neuroethology have enabled increasingly detailed and comprehensive quantification of animal movements and behavioral kinematics. Vocal communication behavior is well poised for application of similar large-scale quantification methods in the service of physiological and ethological studies. This review describes emerging techniques that can be applied to acoustic and vocal communication signals with the goal of enabling study beyond a small number of model species. We review a range of modern computational methods for bioacoustics, signal processing, brain-behavior mapping, and physiological data analysis. Along with a discussion of recent advances and techniques, we include challenges and broader goals in establishing a framework for the computational neuroethology of vocal communication.

1.1 Introduction

Over the past several years emerging methods have enabled biologists to capture and quantify ethological data in ways that yield new insights into the structure and organization of behavior. These methods capitalize on two advances: the ability to record and annotate very-large

behavioral datasets, and the use of new computational tools to reveal structure within and between these datasets. The ethological and neuro-ethological study of animal communication has a long history, and its future stands to benefit greatly from these new methods. Here, we discuss this emerging set of tools available to the animal communication researcher. We contextualize these computational methods within the emerging field of computational ethology more broadly and discuss how these tools can be applied in behavior and neurophysiology.

Many of the challenges that exist in the computational neuroethology of vocal behavior are neither new nor unique and parallel those in other areas of human and animal behavior. For example, the algorithmic discovery of vocal units and sequential organization in animal communication parallels the zero-speech challenge in language acquisition: given limited sensory information, can we build a system that discovers subwords, words, and sequential and syntactic organization present in speech [446]. In animal communication the challenge is similar: can we infer vocal segment boundaries, categories, and temporal organization from the physical and temporal characteristics of the signal. The computational neuroethology of vocal communication also parallels the emerging field of motion sequencing and the mapping behavioral kinematics, where new technologies allowing researchers to map postures and behavioral kinematics have facilitated new understandings of behavioral dynamics across scales [28, 81, 46, 69, 9, 342]. It is the goal of computational neuroethology to not only develop an understanding of the organization of behaviors, but also the neural and cognitive mechanisms that facilitate behavior. This review synthesizes work from several fields including bioacoustics, systems neuroscience, and computational neuroethology to discuss emerging methodologies and frameworks which span these fields and are available to vocal communication researchers.

The review begins with considerations in bioacoustics and signal processing and then shifts to a consideration of acoustic structure, sequential organization, and eventually to mapping the acoustic and sequential structure of vocal communication to neurophysiology correlates of behavior and perception. Throughout our review of current approaches, we relay ongoing challenges, discuss future directions, and attempt to give practical advice on vocal analyses.

1.2 Signal processing and denoising

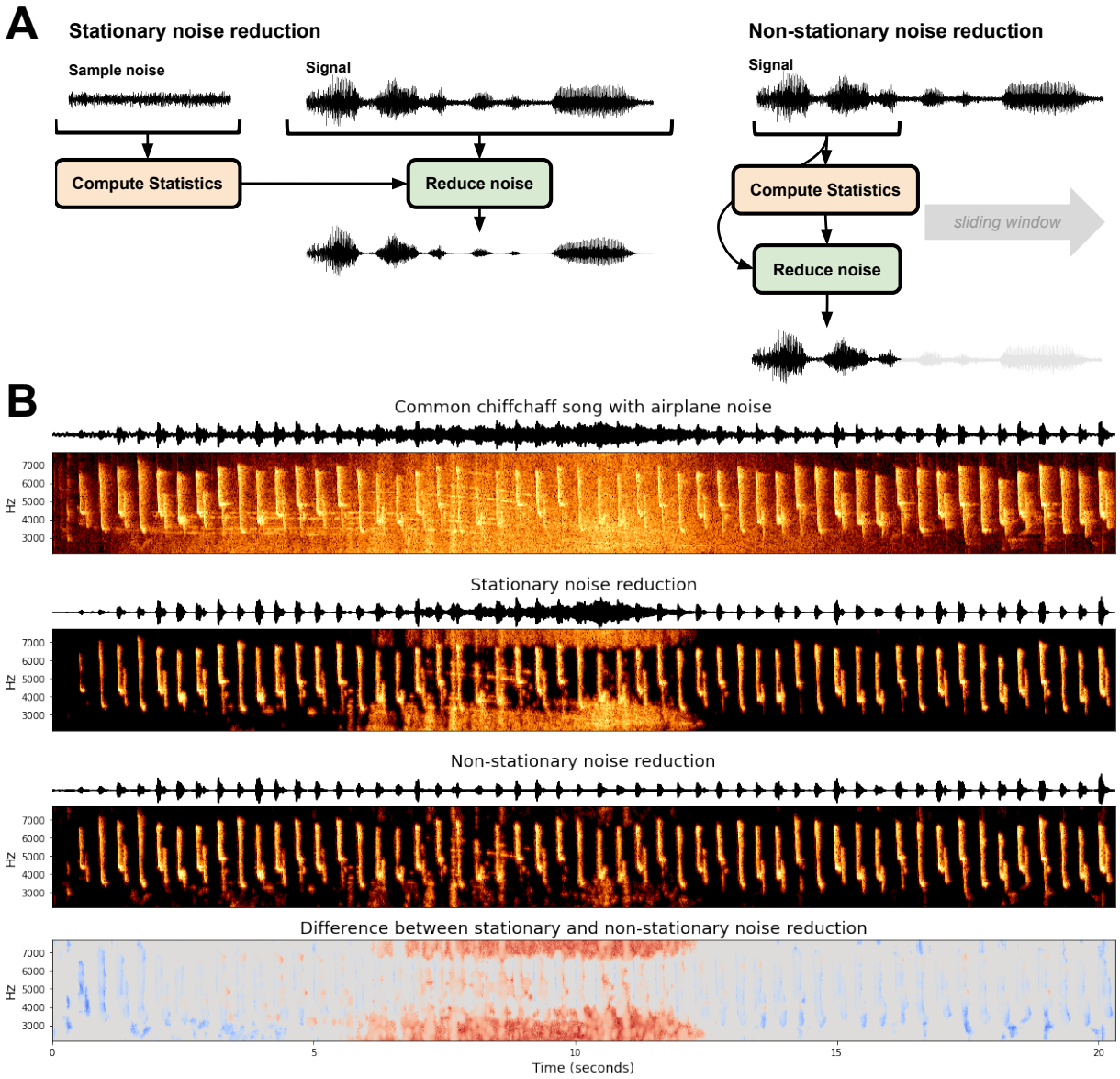


Figure 1.1. Stationary and non-stationary spectral gating noise reduction. (A) An overview of each algorithm. Stationary noise reduction typically takes in an explicit noise signal to calculate statistics and performs noise reduction over the entire signal uniformly. Non-stationary noise reduction dynamically estimates and reduces noise concurrently. Stationary and non-stationary spectral gating noise reduction using the noisereduce Python package [378] applied to a Common chiffchaff (*Phylloscopus collybita*) song [415] with an airplane noise in the background. The bottom frame depicts the difference between the two algorithms.

Recorded sounds typically contain a mixture of both relevant and irrelevant components.

Computational ethology often relies on modeling structure in data without making assumptions about the relevant features. Thus it is often important to remove irrelevant features (i.e. background noise) prior to analysis. Ones operationalization of noise can vary based upon the end goal of the analysis. A simple example is band-pass filtering: because vocalizations typically occur in a confined frequency range, it is reasonable to consider signal outside of that range noise and filter it away. When a recording contains vocalizations from two animals, a songbird with song in a high-frequency range, and heterospecific calls in a low-frequency range, if the subject of interest is the songbird, a simple high-pass filter can be applied to attenuate the non-target calls. When frequency ranges overlap between signal and noise, however, the problem of noise reduction becomes more difficult.

Noise reduction

Determining what constitutes noise in recordings is non-trivial and impacts what type of noise reduction algorithm can and should be used. In a systematic review of noise reduction methods in bio-acoustics, Xie et al., ([466]) outline six classes of noise reduction algorithms used for bio-acoustics: (1) Optimal FIR filter (e.g. [204]), (2) spectral subtraction (e.g. [37, 202, 384]), (3) minimum-mean square error short-time spectral amplitude estimator (MMSE-STSA; e.g. [114, 45, 5]) (4) wavelet based denoising (e.g. [362, 353]) (5) image processing based noise reduction, and (6) deep learning based noised reduction. These noise reduction algorithms can be broadly divided into two categories: stationary and non-stationary noise reduction (Fig 1.1A). Stationary noise reduction acts on noise that is stationary in intensity and spectral shape over time, such as the constant hum of electronics. Non-stationary noise reduction targets background noise that is nonstationary and can fluctuate in time, like the on-and-off presence of a plane flying overhead (Fig 1.1B). Stationary noise reduction algorithms operationalize noise as stationary signals, for example, the constant hum from a nearby electronic device in a laboratory setting, or insect noise in a field setting.

One approach to stationary noise reduction is spectral gating, a spectral-subtraction

algorithm (e.g. [384, 202]). The general notion is that for each frequency component of the signal, any time-frequency component below a threshold is discarded as noise. Spectral gating computes the mean and standard deviation of each frequency channel of a Short-Time Fourier Transform (STFT) of a signal (e.g. a spectrogram) and optionally a noise clip. A threshold, or gate, for each frequency component is then set at some level above the mean (e.g. three standard deviations). This threshold determines whether a time-frequency component in the spectrogram is considered signal or noise. The spectrogram is then masked based upon this threshold and inverted (with an inverse STFT) back into the time domain.

Non-stationary noise reduction

While stationary noise reduction algorithms can operationalize noise as any stationary acoustic signal, non-stationary algorithms vary in how they determine what is signal and what is noise. Non-stationary noise can be more challenging to remove because it can be difficult to algorithmically define the difference between signal and noise. Because the hum of a computer in the background of a lab-recording is stationary, it can be defined as noise and can be readily removed. A bird hopping around its cage can produce time-varying sounds in the same frequency range as song, making it especially pernicious.

One approach for determining the boundary between signal and non-stationary noise is to determine the timescale on which the signal acts and treat anything outside of that timescale as noise. For example, zebra finch motifs are generally between 0.5-1.5 seconds long repeated one to four times [51]. Any acoustic event that is outside of that time range could be considered noise. Spectral gating can be extended to non-stationary noise reduction by computing a variable gate based upon the current estimate of background noise. In the Python package noisereduce [378], this background estimate is computed using a time-smoothed spectrogram (using a forward and backward IIR filter) on a timescale parameterized by the expected signal length, an approach motivated by the Per-Channel Energy Normalization algorithm (outlined in Section 1.3). An example of this is given in Figure 1.1, where stationary and non-stationary spectral gating noise

reduction is applied to birdsong with an airplane noise occurring in the background of the middle of the recording. Because the airplane noise is non-stationary, The stationary approach fails in two ways relative to the non-stationary approach: the airplane noise is not fully successfully gated at its peak in the middle of the recording (shown as red in the bottom panel) and weaker syllables of song are treated as noise and reduced in the beginning and end of the clip (shown in blue in the bottom panel). Advantages of non-stationary noise reduction are not unique to acoustic noise: when we know the timescale of a signal we can use the same non-stationary principles to remove noise occurring at different timescales. For example in the continuous recording of neural data, action potentials occur within the range of one millisecond. Events occurring over tens or hundreds of milliseconds can therefore be treated as noise.

Reducing noise with deep learning

A promising future avenue for noise reduction is in explicitly training machine learning algorithms to mask or remove noise, as is done in speech enhancement and segregation [452]. At present, however, deep learning based noise reduction has not been utilized directly in bio-acoustics [466]. Xie et al ., ([466]) attribute this to a lack of utility when using denoising in some applications of deep learning-based bio-acoustics detection [217]. The utility of noise reduction exists beyond classification tasks, however. For example, computing spectral features and acoustic similarity between vocalizations can be susceptible to background noise. Recent work by Stowell et al., ([415]) shows that manipulating datasets by superimposing background environment noise on vocal datasets can reduce confounds and improve identification across recording conditions. Similar approaches could be used to remove noise. For example, spectral gating could be extended with neural networks by training a neural network to learn a mask to gate away background noise and recover the lower-noise spectrogram, as has been done in speech enhancement applications [452, 240].

It is also important to consider what information is being removed by pre-processing techniques such as denoising. Pre-processing methods throw away potentially valuable information

that will influence downstream analyses. De-noising vocal data without careful consideration can remove lower amplitude syllables of birdsong or infrequent vocalizations outside of the expected frequency range.

1.3 Signal representation

An important consideration in any analysis pipeline is how to represent the data that goes in. Animal vocalizations are typically recorded using one or more microphones at a sampling rate that can capture the full spectral range of the vocalization. Performing analyses directly upon recorded waveforms is not always optimal for capturing informative structure in vocal data, however. Waveforms are high-dimensional representations of audio that can make it difficult for learning algorithms to capture time-frequency structure in vocalizations. Spectro-temporal representations can be both lower-dimensional, and more explicitly capture complex time-frequency relationships in vocalizations.

Spectrograms are, at present, the most common form of vocalization representation, both for visualization and as input to learning algorithms, both in bio-acoustics and speech. When representing vocal data with a spectrogram, the parameters used to compute the spectrogram can have an important influence on the performance of the algorithm [208, 109]. The most important parameterization of spectrograms is the trade-off between temporal and frequency resolution when computing a spectrogram, a result of the Heisenburg Uncertainty Principle [139, 293]. For example, three spectrograms are shown in Figure 1.2A-C with different windows used to compute the Short-Time Fourier Transform. The first has an intermediate-sized window with intermediate time and frequency resolution (Figure 1.2A), the second uses a short window with high time-resolution and low frequency-resolution (Figure 1.2B), and the third uses a long window with high frequency-resolution but low time-resolution (Figure 1.2C).

A number of approaches exist to improve time and frequency resolution. Time-frequency reassigned spectrograms attempt to improve time-frequency resolution using additional informa-

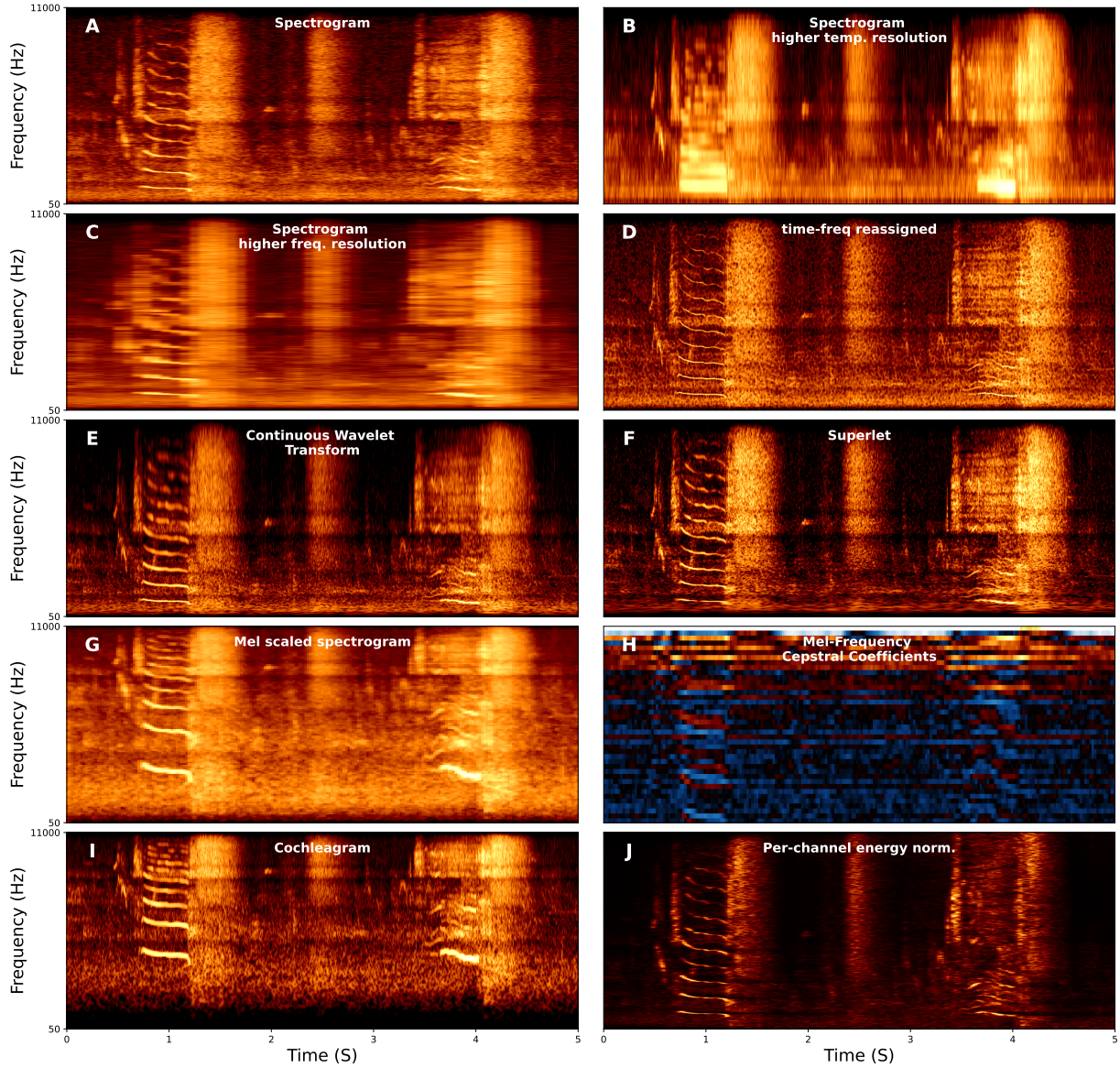


Figure 1.2. Examples of several different spectral representations of a five-second red deer (*Cervus elaphus*) vocalization. For each axis, the *x*-axis corresponds to time, and the *y*-axis corresponds to frequency. The y-axis corresponds to frequency and is linearly spaced in A-F,J between 50-11000 Hz and log-spaced in the same range for G,H and I. (F) Continuous Wavelet Transform using the Morlet (i.e. Gabor) wavelet.

tion from the phase spectrum (Figure 1.2D) [139, 465, 133]. Wavelet transforms (Figure 1.2E) have more recently been used in representing animal vocalizations [353, 178, 354, 265], and allows multi-scaled emphasis on time versus frequency, for example emphasizing frequency resolution at lower frequencies and time-resolution at higher frequencies, intuitively because an uncertainty of 50Hz is more relevant at 500Hz than at 5000Hz. Most recently, the superlet (Figure 1.2F) enables time-frequency super-resolution by geometrically combining sets of wavelets with increasing constrained bandwidths [293].

There are also several variants of spectrograms and time-frequency representations that differentially emphasize time-frequency information. For example, log-scaling spectrograms in frequency emphasizes lower frequency ranges over higher frequency ranges, which parallels both the cochlea and perception [108]. Mel-scaling (Figure 1.2G), is a form of log-scaling fit to fit human perception [414], though the specific scaling range relative to human perception are imperfect [157]. Mel-Frequency Cepstral Coefficients (MFCCs; Figure 1.2H) additionally compute the Discrete Cosine Transform on the Mel-spectrogram, and were, until recently, commonly used for speech recognition because they are generally robust to noise and emphasize the frequency range of speech (Figure 1.2H) [309]. Another model, directly relevant to physiology, is the Cochleagram [47, 116, 359]. Cochleagrams mimic the cochlea by using a filter bank associated with points on the basilar membrane to mimic an impulse response 1.2I).

A new approach that has shown much promise in bio-acoustics is Per-Channel Energy Normalization (PCEN; Figure 1.2J; [453, 257]). Lostanen et al., ([257]) identify three advantages of PCEN: (1) temporal integration, (2) adaptive gain control, and (3) dynamic gain compression. Temporal integration estimates the background noise at each frequency band. Adaptive gain control then adapts the gain of the spectral representation. Finally, dynamic range compression adaptively shifts the range of quiet and loud components of the signal. Adaptive gain control is ubiquitous to mammalian auditory processing and is also often used in cochleagrams [359]. PCEN has been shown to aid in enhancing animal vocalizations relative to background noise across distances from the microphone [256] and reduce biases in bio-acoustics background

settings such as dawn versus dusk [258].

Descriptive basis-features features can also be used to represent vocalizations for downstream analyses. One challenge with using basis-features for vocal analysis is in determining what basis-features are relevant [424, 109]. Very few species have been rigorously examined to determine what acoustic features distinguish vocal units [109, 200]. Swamp sparrow notes, for example, are relatively simple vocalizations and can be well-described using just the length of the note, the peak frequency at the start of the note, and the peak frequency at the end of the note [70]. One approach to determining what features are relevant in a vocal signal is to train classifiers to predict behaviorally-relevant information, such as individual identity, age, or the activity the animal is engaged on a full set to basis features, and retain those features which are highly informative [109, 110].

1.4 Identifying, segmenting, and labeling vocalizations

Vocalization data can be recorded in a number of different settings, ranging from single individuals in well-controlled and acoustically isolated lab settings to multi-individual and multi-species recordings taken next to a busy highway. When vocalizations are produced by isolated, single individuals, segmenting out vocalizations can often be performed simply by thresholding the vocal envelope and assuming any detected noise events that match the statistics of the vocalizing animal (e.g. frequency and length of vocalization) are vocalizations [424]. More complex environments and species with more complex vocal structure require more complex solutions [352].

Experimental paradigms in neuroethology differ from bio-acoustics in that environmental sounds can usually be controlled, but are still faced with the challenge of often being made in colony settings with multiple vocalizing individuals or individuals who make non-vocal sounds such as interaction with a living space. Regardless of context, recent advances in machine learning algorithms for passive monitoring of acoustic environments allow for real-time labeling

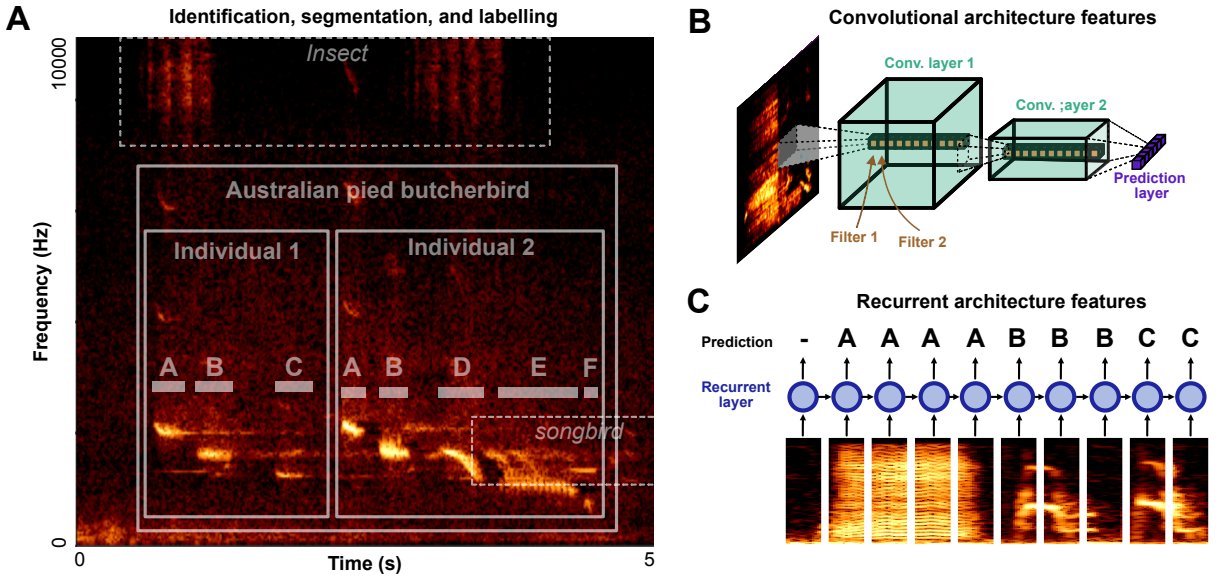


Figure 1.3. Levels of organization and architectural design features for identifying, segmenting, and labeling vocalizations. (A) An example clip of Australian pied butcherbird song [183] is shown containing two male butcherbirds, alongside background noise containing insect noise and another bird's song. A classification task exists at several levels: identifying the target species, differentiating individuals, detecting note boundaries, and classifying notes. (B) A cartoon diagram of a convolutional neural network architecture applied to a spectrogram. Convolutional filters are applied in time-frequency space. Deeper layers have larger spectrotemporal receptive fields and learn more complex filters. (C) A cartoon diagram of a recurrent neural network applied to a song spectrogram. Spectral slices are input to recurrent layers in the network (depicted as a circle) which are recurrent in time, allowing information to be integrated over time.

of species and individuals in noisy environments.

Automatic vocalization annotation can be broken down into three related tasks: identification, segmentation, labeling. Identifying refers to what animal is vocalizing and at what times and frequency channels. Segmentation refers to the segmentation of vocalizations into their constituent units, labeling then refers to grouping units into discrete element categories. A spectrogram outlining all three tasks is given in Figure 1.3A. Two individuals in the target species, Australian pied butcherbird *Cracticus nigrogularis* are vocalizing over top of background noise from another, unidentified, species of songbird, as well as an unidentified species of insect. Each bird's song can be divided into segmental units (notes) which can be further categorized into discrete element categories ('A', 'B', 'C', ...). In such a dataset, labeling challenges occur over multiple levels: identifying the species, identifying the individual, segmenting vocal units, and labeling vocal units into discrete categories. Some algorithms perform only one of these steps at a time, while others perform all three.

Detecting species and individuals

To detect species in continuous bio-acoustic data, several open-source tools and datasets have recently been made available for passive acoustic monitoring. A summary of many of these software and their features are given in Priyadarshani et al., ([352] Table 4). Over the past few years machine learning competitions challenging researchers to produce species recognition algorithms have motivated an increasing number of open-source approaches to bioacoustic sound recognition (e.g. [238, 312, 148, 415]). The same tools can be applied to differentiating between individuals in the same recording environment (e.g. [2, 286]). Most recent approaches rely on deep neural networks to detect vocalizations in noisy environments (e.g. [74, 415]). Current neural networks generally rely on some combination of convolutional filters in the temporal-frequency space of spectrograms (Convolutional Neural Networks or CNNs, Fig 1.3B) and temporal-recurrence (Recurrent Neural Networks, or RNNs, Figure 1.3C). Convolutional filters in the time-frequency space of spectrograms allow neural networks to learn complex

spectro-temporal features used to classify sounds (Figure 1.3B). Temporal recurrence allows neural networks to learn sequential and temporal relationships that unfold over long time delays (Figure 1.3C). In combination, recurrent and convolutional architectures allow complex, non-linear spectrotemporal features that occur over arbitrary timescales to be captured by neural network architectures.

Segmenting and labeling vocal units

Beyond identifying individuals and species, many analyses of vocal communication rely on the temporal segmentation and categorization of vocalizations into discrete units. Unlike identifying species or individuals, where an objective measure exists of what animal produced a vocalization, the segmental units that comprise animal vocalizations are less well-defined. In comparison to human language, where linguistic units are determined based on their functional role, substantially less is known about the function each vocal unit plays in most species' communication, or even what should define the beginning and ending of a vocal unit [200, 292]. Analyses of most animals, therefore, rely on easily discernible physical features of vocalizations. For example in songbirds, songs are typically segmented at different hierarchical levels, though no strict definition of these levels of organization are agreed upon by all researchers. Common units of birdsong are notes, corresponding to abrupt changes in frequency, syllables, defined by periods of silence surrounding continuous vocalizations, motifs, stereotyped repetitive combinations of acoustic elements, and phrases, series of stereotyped or commonly associated syllables. Despite the ubiquity with which these terms are used, most vocal units have not been validated in terms of the species' own perceptual system, and those that do, like the Bengalese finch 'syllable' [292] call into question the commonly assumed role they play in communication. It is therefore ideal, but not always feasible, to validate decisions about vocal units based upon perceptual, physiological, or functional roles those vocal units play in the animal's communication [419, 200]. Still, most analyses of animal communications rely on human perceptual decisions at some level, whether it is to label discrete classes of birdsong phrases, or determine the representational space

upon which an 'unsupervised' learning algorithm will discretize units (discussed in Section 1.5).

When vocal units are defined and vocal classes are chosen, machine learning algorithms can be used to systematize and vastly speed up the classification and segmentation of vocal units. Most commonly, supervised recognition algorithms are used, where the algorithm explicitly learns to algorithmically map acoustic data to the researcher's labeling scheme. Over the past decades, vocalization labeling algorithms have paralleled those used in other acoustic domains, such as speech and music recognition. At present, tools rely on deep neural networks. The field of deep learning has changed rapidly over the past decade, with different architectures of neural networks quickly outperforming the previous architectures [314]. Prior to deep learning, automated birdsong element recognition relied on algorithms such as Hidden Markov Models [212], support vector machines [421], template matching [10], or k-Nearest-Neighbors labeling [321], following alongside contemporary speech recognition algorithms. Like sound event detection, current approaches tend to rely on recurrent and convolutional neural network architectures. TweetyNet [74], for example, uses a recurrent and convolutional architecture to capture complex spectro-temporal patterns over long timescales. Future advances in neural network architectures will likely continue to follow those in speech recognition, for example, using transformer architectures [193] as well as semi-supervised and unsupervised pre-training methods such as wav2vec [393]. One important divergence between speech recognition and animal vocalization classification is the reliance upon data availability, however. An ideal animal vocalization classifier works well on very small amounts of labeled data, requiring less experimenter time, whereas speech recognition systems tend to have an abundance of data available (though speech recognition for low-resource languages may be an area to watch).

A second approach to labeling vocalizations is to actively involve the experimenter in the algorithm via human-in-the-loop labeling (e.g. [464, 203]). Human-in-the-loop algorithms rely on a combination of supervised and unsupervised learning. Supervised learning comprises learning algorithms that are trained with labeled data, such as classification tasks. Unsupervised learning refers to algorithms that do not require supervised labels, such as dimensionality reduc-

tion. Human-in-the-loop algorithms leverage both, by proposing an initial coarse segmentation and/or labeling of the dataset through unsupervised learning, which the human then partially revises (e.g. merging or splitting putative classes of vocalizations) via a graphical user interface (GUI). The revised data is then re-processed by the algorithm and sent back to the user to revise, until the experimenter is content with the resulting labeled dataset. Using a combination of human expertise and machine processing enables quicker labeling of large bio-acoustics data with minimal human effort. A further discussion of unsupervised algorithms is discussed below in Section 1.5.

1.5 Extracting relational structure and clustering

Classifying vocal elements into discrete categories (e.g. 'A', 'B', 'C', ...) is for many analyses a necessary abstraction that enables the analysis of recurring events. At the same time, this symbolic abstraction ignores acoustic relationships both within discrete element categories and between them. For example, in Figure 1.3, are the syllables of birdsong labeled 'A' more similar to the syllables labelled 'B' or the syllables labeled 'C'? Determining the relatedness (or distance) between vocalizations can enable the quantification of how vocalizations change over time [285, 216], how vocal repertoires differ across individuals and species [384, 289], and map and visualize broad structure present in vocal repertoires [384, 149].

Operationalizing relatedness

Given a dataset of vocalizations segmented into discrete units, relatedness is a measure quantifying the similarity of vocalizations relative to one another. The basis for operationalizing relatedness can utilize physical properties of signals, perceptual judgments, or behavioral and physiological responses to the signal. Most commonly, the relationships between vocal elements are computed on either spectrotemporal representations or on the basis of descriptive features of the vocalization, such as frequency modulation, pitch, and vocal envelope, or [384, 149, 289].

How different aspects of the vocalization should weigh into a measure of similarity

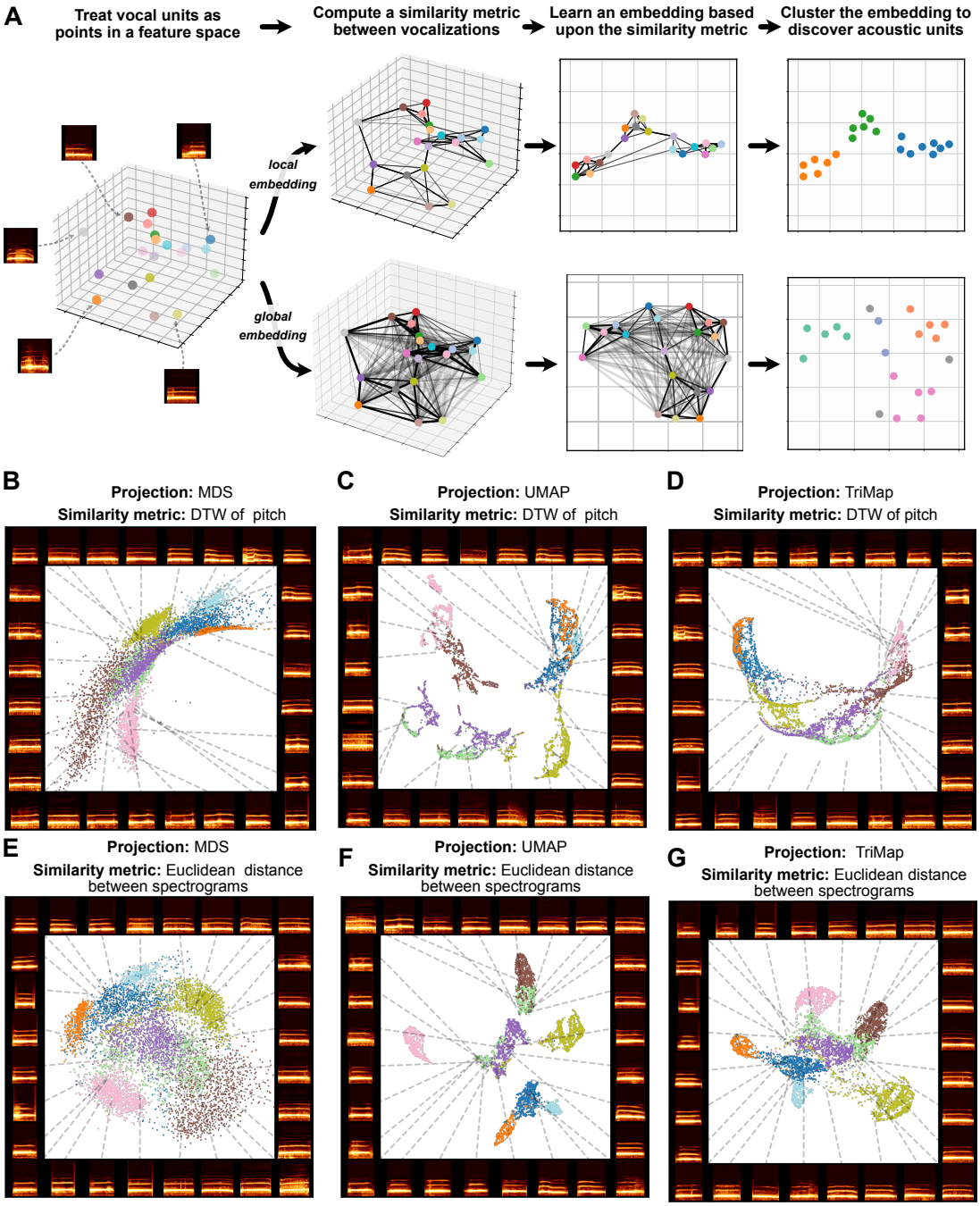


Figure 1.4. Local and global embeddings. (A) The steps outlined in Section 1.5 exhibit the differences between the relationships preserved in local and global embeddings. (B-D) Projections of a dataset of macaque coo calls [130] using two similarity metrics (Dynamic Time Warping over pitch, and Euclidean distance between spectrograms) and three projection algorithms (Multidimensional Scaling, UMAP, and TriMap). Colors represent individual identity.

is non-trivial. No metric for similarity is objectively correct, even when metrics are derived purely from objective physical features. For example, what is the relative importance of a vocalization's duration versus pitch in determining similarity? One ground truth metric for an algorithms judgement of similarity is its relationship with human's perceptual judgment of similarity [424], though there is no guarantee that these measures reflect the animal's own perception and physiology [99]. An ideal measure of similarity could be derived through careful experimentation gleaning the animal's own judgment of similarity [200], but in most cases, this task would be unfeasible and time-consuming. Even when performed carefully, perception varies from animal to animal, based upon experience [231].

In addition, when vocal features are continuous, accounting for differences in duration and temporal alignment requires consideration. Approaches vary from averaging over time [109], pooling using attention mechanisms [302], using dynamic time warping [212], and zero-padding [384]. Similarly, at least some animals rely on spectral shape rather than absolute pitch when recognizing acoustic objects [44]. A recent approach accounting for variability in frequency is dynamic frequency warping [410]. Striking a balance between spectrotemporal tolerance and absolutely discounting spectrotemporal alignment can have substantial impact on the final measure of similarity.

Learning a similarity space

Once a metric for similarity is determined, that distance can be used to infer a structured representation of the relationships in a repertoire as a whole, providing a new representational space with which to quantify vocalizations.

Perhaps the most intuitive and pervading example of a learned embedding space for vocal similarity is multi-dimensional scaling (MDS, e.g. [289, 98, 302]. Multi-dimensional scaling takes a graph of pairwise similarity measures between each vocalization in the dataset and attempts to find an embedding that best preserves the similarity structure of that graph. As the number of vocal elements in a dataset gets larger, however, the number of pairwise

distances between vocal elements increases exponentially. This is computationally an issue because computing 10000 pairwise distances between 100 elements is computationally feasible, but 10,000,000,000 pairwise distances between 100,000 elements is not.

Trying to preserve the pairwise relationships between every element in a dataset can also over-emphasize irrelevant relationships in vocal data. For example, if a bird's vocal repertoire comprises 10 motifs classes all produced with the same frequency, the vast majority of pairwise distance relationships computed (90%) will be between class, while only 10% of pairwise relationships computed will be within class. In many cases, both in animal communication and in dimensionality reduction more broadly, there is utility in weighing relationships between similar vocal elements more highly than relationships between less similar vocalizations. This contrast is defined in the dimensionality reduction literature as the emphasis of local versus global structure [87]. Algorithms that attempt to preserve every pairwise relationship are called global dimensionality reduction algorithms, while algorithms that emphasize capturing relationships only to nearby points in dataspace (more similar vocalizations) are called local dimensionality reduction algorithms. In many vocalization datasets, emphasizing local over global structure better preserves categorical structure such as individual and call identity [385, 302, 149]. A visual demonstration contrasting local and global structure preservation is given in Figure 1.4A. While global embedding algorithms like MDS attempt to preserve every pair-wise relationship, local algorithms preserve only local (e.g. nearest-neighbor) relationships, capturing more within-cluster structure. In Figure 1.4B-G an example is given with macaque coo calls, in which a local structure-preserving algorithm (UMAP, described below) more clearly pulls apart clusters corresponding to individual identity than MDS.

At present, the two dominant local dimensionality reduction algorithms are UMAP and t-SNE. UMAP and t-SNE differ in several important ways beyond the scope of this paper, but their key intuition and the steps underlying the algorithms remain similar: first, compute a (nearest-neighbors) graph of pairwise relationships between nearest neighbors in the original dataset (e.g. using Euclidean distance or an arbitrary similarity metric) then, embed that graph

into an embedding space via gradient descent [380]. UMAP, in particular, has been shown to capture complex structure in vocal repertoires such as differences in vocal dialect, vocal stereotypy, vocal element categories, inter-species similarity, and individual identity, in contrast to classic methods like MDS and PCA [380, 302, 149].

One challenge with graph-based dimensionality reduction algorithms like MDS, UMAP and t-SNE is that they are non-parametric dimensionality reduction algorithms, meaning they do not learn the relationship between input data (e.g. a spectrogram of the vocalization) and their embeddings. Learning a parametric relationship between vocalizations and their embeddings allows a fast mapping between data and embedding, i.e. for applications that necessitate real-time feedback such as brain-machine interfacing.

The most common parametric dimensionality reduction algorithm is PCA, where a linear transform is learned between data and an embedding space. Similarly, neural networks such as autoencoders can be used to learn a set of basis features which can be complex and non-linear [149, 404, 385, 213]. For example, an autoencoder trained on images of faces can learn to linearize the presence of glasses or a beard [380, 386, 358]. Autoencoders trained on animal vocalization data can similarly learn complex non-linear relationships in vocal data. In Section 1.7 we discuss how these complex learned features could be utilized in animal vocalizations to learn acoustic features such animal age, sex, and attractiveness, which can, in principle, be utilized for playback experiments.

A recent extension to UMAP, Parametric UMAP weds the advantages of UMAP with the parametric embedding of neural networks [380]. Parametric UMAP acts by optimizing the UMAP loss function over arbitrary neural networks (e.g. convolutional recurrent networks were used with Cassin's vireo song in [380]) which can be balanced with additional losses such as MDS and autoencoding, to preserve additional global structure in UMAP projections. Parametric neural network-based approaches such as Parametric UMAP can also embed data on a similar timescale as PCA, enabling real-time applications, as opposed to non-parametric methods such as UMAP, t-SNE, and MDS.

Another class of neural network based dimensionality reduction algorithms rely on triplet-loss-based similarity preservation. Triplet-based embeddings have been used for birdsong for classification and embedding [363, 302]. Triplet networks learn an embedding space by sampling three types of vocal units: an anchor, a positive sample that is perceptually similar to the anchor point, and a negative sample that is perceptually distant from a vocal unit. The loss then encourages the positive sample to be pulled to the anchor, and the negative sample to be pushed further away. For example, Morfi et al., [302] describe a triplet-loss-based network trained to produce vocal embeddings based upon a metric of perceptual distances. Like graph-based dimensionality reduction algorithms, triplet-loss-based embeddings rely on a pre-defined experimenter-determined notion of distance. Morfi et al. suggest a forthcoming animal-defined metric but in-lieu use a descriptive feature-based metric in the software Luscinia [230] which is correlated with human perceptual judgments of zebra finch song [175].

Finding latent units through clustering

Learned embedding spaces enable the inference of broad structure acoustic structure from the statistics of vocalizations, enabling further downstream discovery of vocal units based upon distributional properties in embedding spaces [384, 199, 200]. Unsupervised clustering of vocal elements lies in contrast with supervised learning, where class labels are determined by experimenters, as in Section 1.4. Sainburg et al., [384] observe that labels obtained by clustering UMAP embeddings of Cassin's vireo and Bengalese finch syllables are more similar to experimenter labels than clustering PCA projections or spectrograms. Further, these latent projections capture additional acoustic and syntactic structure than the ground truth experimenter labels. In addition to acoustic structure, vocal elements can be clustered on the basis of syntactic organization. For example, incorporating transition information through Partially observable Markov Models (POMMs; [187]) and Hidden Markov Models (HMMs; [384, 195]) into a labeling scheme for birdsong better explains sequential structure than hand-labels or clustering without reference to temporal sequencing. An alternative approach is to perform clustering prior

to embedding, directly upon the inferred relational graph [127].

One challenge in unsupervised vocal unit discovery through methods such as UMAP embeddings is their reliance upon pre-defined vocal unit temporal boundaries. Although clustering on latent projections enables an unsupervised extraction of vocal categories from segmental units, latent projections rely on a pre-defined temporal segmentation of acoustic units from the vocal stream. In some species, atomic vocal units can be determined by clearly defined physical features of the signal, like long pauses between syllables, however, even in the case of clearly defined physical features, those units are not necessarily the base units of perception [292]. An open issue in vocal analysis is the unsupervised temporal segmentation of vocalizations into elements when clear physical boundaries are not available. This problem parallels both unsupervised speech discovery (i.e. ZeroSpeech), and the challenge of discovering behavior units in other areas of computational neuroethology (e.g. Motion Sequencing). In speech, phonemes are not clearly defined by physical characteristics, thus approaches for segmentation rely upon a combination of temporal and distributional information alongside imposed priors. Ongoing efforts in unsupervised speech segmentation, syllabic unit discovery, and word discovery can motivate parallel approaches in animal communication. In addition, physiological and kinematic measures such as articulation and breathing rate can aid in determining vocal boundaries. In computational neuroethology, new methods in tracking behavioral kinematics provide similar continuous behavioral datasets to those discussed in this paper (e.g. [463, 462, 29, 343, 273, 269, 103]). For example, MoSeq [463, 462] discovers animal behavioral states using depth camera recordings of animals by fitting the behavioral data to an Autoregressive Hiden Markov Model. They find stereotyped sub-second mouse behavioral states, dubbed syllables, that underlie a syntax of behavior, much like birdsong. Communicative behavior is also not produced solely in the auditory domain. Improving methods for uncovering structure in animal behavior more broadly will facilitate research on the interaction between multi-sensory and multi-modal vocal behavior, like the dances that accompany many bird songs [459].

Data augmentation

Another approach that is largely underutilized in bio-acoustic vocal recognition algorithms is data augmentation, an approach that is currently used in most state-of-the-art machine perception applications. In automatic speech recognition, for example, several current state-of-the-art approaches (e.g. [20, 159]) use SpecAugment [334] in which the classifier learns a policy of various augmentations such as warping and masking frequency channels in time. Lostanlen et al., [258] demonstrate the utility of augmenting bio-acoustics datasets with diverse background acoustics to facilitate better generalization across environments and conditions. Augmentation in settings where little labeled data are available has also proven successful on several semi-supervised learning benchmarks (e.g. [31]). One difficulty with performing data augmentation with bio-acoustics data, however, is the extent to which slight manipulations can affect the perceptual class that vocalizations fall into [302].

1.6 Inferring temporal and sequential structure

Identifying sequential organization typically relies upon the abstraction of vocalizations into discrete sequences of elements, effectively treating vocal data as corpora from which to perform symbolic analyses. Kershenbaum et al., [200] identify six classes of models and analyses for analyzing temporal sequences: Markov chains, Hidden Markov Models, Network-based analyses, Formal grammars analyses, and temporal models. Analyses of temporal organization in animal communication has traditionally been largely influenced by Chomsky's hierarchy of formal grammars, with a focus on trying to understand what class of the Chomsky hierarchy animal's behaviors belong within [303, 369, 186, 162]. For example, Markov models, Hidden Markov Models, and Network models are all finite-state models in the Chomsky hierarchy.

Short-timescale organization and graphical analysis

Broadly, analyses over vocal organization can be broken down into two classes: analyses over short- and long-distance (i.e short- and long-timescale) sequential organization. Short-

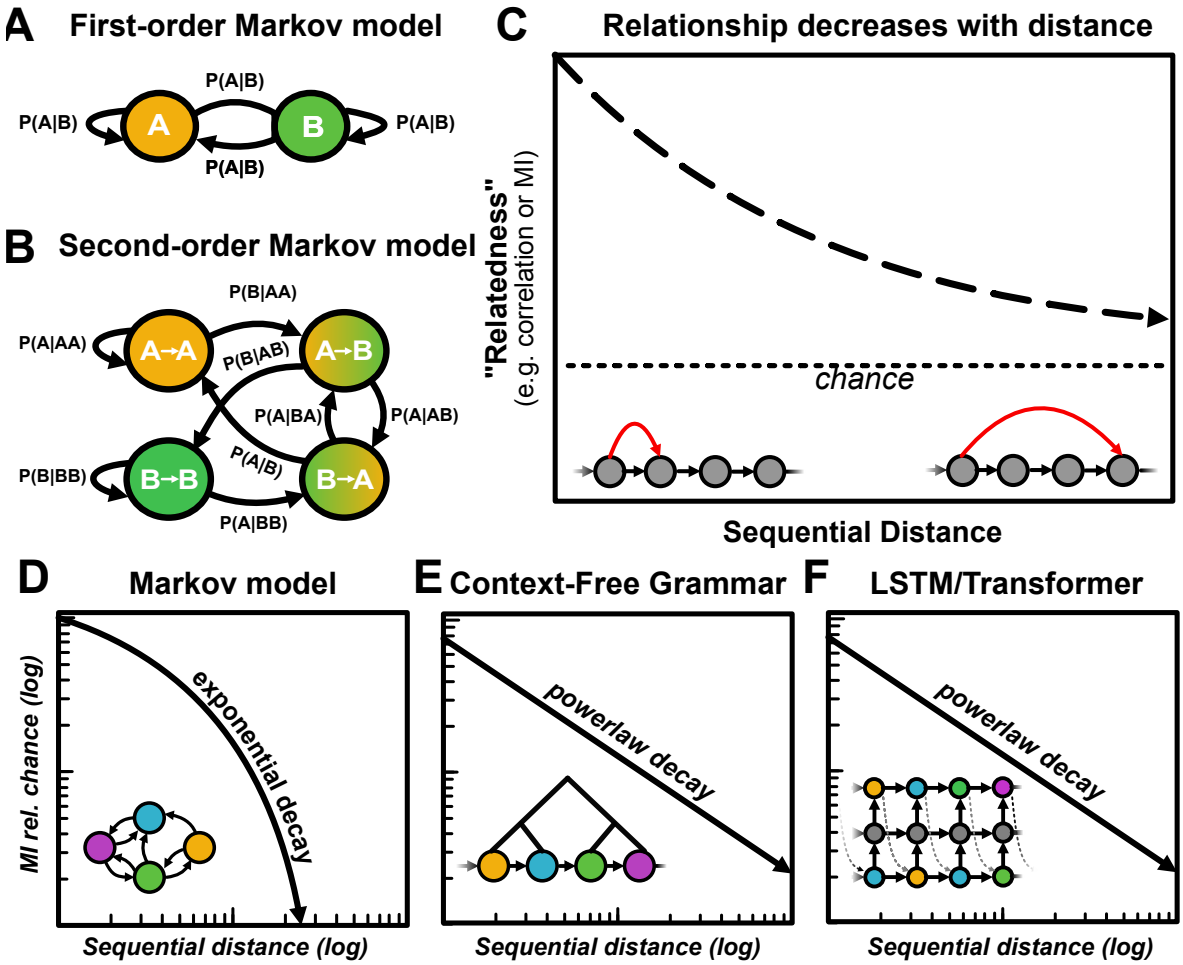


Figure 1.5. Capturing long and short-range sequential organization with different models.(A) An example of a 2-state Markov model, capturing $2^2 = 4$ transitional probabilities between states. (B) An example second-order Markov model, capturing $2^3 = 8$ transition probabilities between states. (C) A visualization of the general principle that as sequential distances increase, the relatedness between elements (measured through mutual information or correlation functions) decays toward chance. (D) Sequences generated by Markov models decay exponentially toward chance. (E) Context-free grammars produce sequences that decay following a power law. (F) Certain neural network models such as LSTM RNNs and Transformer models produce sequences that also decay following a power law.

timescale analyses are concerned with relationships between adjacent, or near adjacent elements in a sequence. Markov models, for example, capture short-timescale dynamics of vocal communication. A typical Markov model of birdsong is simply a transition matrix representing the probability of transitions from each element to each other elements (e.g. $P(B|A)$ Fig 1.5A). As Markov models increase in order, they become increasingly capable of capturing long-distance relationships, though high-order Markov models are rarely used in practice because of the number of parameters and amount of data needed to compute them (Fig 1.5B). Approaches such as Hidden Markov Models [195] and Probabilistic Suffix Trees [266, 75] can compute more succinct high-order Markov relationships, though the amount of data needed to capture these deeply contextualized relationships (e.g. $P(F|A,B,C,D)$) is still a limiting factor in capturing long-range organization with Markov models. Short-range relationships are also often captured graphically, treating any transition probability above zero as an edge in the graph. Graphical representations and metrics for vocal sequencing can explain general sequencing characteristics of vocalizations such as network motifs, communities, and clusters [200, 390, 455, 336, 166].

Mutual information and long-timescale organization

Relationships that extend beyond adjacencies and over longer timescales are called long-range orlong-timescale relationships. For example, how related are two notes within a phrase, two phrases within a bout of song, or two bouts of song sung within a day?

Broadly, elements that are further displaced in a vocalization from one another tend to be less related. When two elements in a sequence are further apart, the relatedness between those two elements tends to be lower. For example, in birdsong, notes within a phrase are more likely to be related than notes separated by multiple phrases. The same is true of most sequential and temporal data: we can better predict what a stock price will look like tomorrow, than in ten years. As we look further and further out into a sequence, the relatedness between elements will decrease alongside our ability to predict the future, until the relatedness approaches chance (Fig 1.5C). We can capture this relatedness over symbolic sequences using information theory. For

example, given a sequence of discrete elements $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow f$, we can estimate the mutual information between pairs of elements at e.g. a distance of 2 elements ($a-c$, $b-d$, $c-e$, and $d-f$) or 3 elements ($a-d$, $b-e$, and $c-f$). As the distance increases between pairs of elements, we expect the relatedness (mutual information) to decay toward chance as a function of sequential distance.

We can estimate the extent to which a signal exhibits long-range relationships by computing how long the mutual information between pairs of elements remains above chance. Such approaches have been used variously across animal vocalization datasets in birds and whales [419, 381]. Similar approaches have also been used to observe long-range structure in animal motion ethology data, such as the long-range structure in *Drosophila* [29] motility.

Inferring structure from sequential relationships

The shape of the decay in relatedness as a function of sequential distance can not only tell us about the timescales that vocal sequences are operating over but can also give indications about the structure underlying sequential organization. For example, sequences generated by Markov processes, such as finite-state grammars decay exponentially [249, 246] (Figure 1.5D). Intuitively, Markov models are memoryless; each state is dictated only by the set of transition probabilities associated with the previous state. As a result, the relatedness between states decays very quickly. When there are deep latent relationships present in the structure underlying the sequence, relatedness between sequentially disparate elements decays more slowly. For example, Probabilistic Context-Free Grammars can produce power-law relationships in mutual information as a function of sequential distance [249] (Fig 1.5E).

Characterizations of statistical relationships over abstracted discrete units enables comparative analyses across species because these measures make no assumptions about units or temporal organization underlying the signal. Characterizing correlations and information decay has an especially rich history in uncovering long-range structure dating back to Claude Shannon's original work [399, 246, 249]. Language corpora such as speech and written text decay in

information following the combination of a power-law over longer distances, and exponential decay over shorter distances, attributed to the finite-state processes underlying phonological organization [381] and the hierarchical organization underlying language at higher levels of organization such as syntax and discourse [381, 249, 6, 7]. At the same time, however, young children's speech contains the same long-range information context before complex syntax is present in speech, indicating possible extra-linguistic mechanisms at play dictating these long-range statistical relationships [379]. Long-range mutual information decay and correlations have also been demonstrated that in animals such as songbirds [381] and humpback whales [419], extending over minute- and hour-long timescales. In particular, birdsong exhibits similar exponential short-range and power-law long-range mutual information decay to human speech, indicating potential parallels in the mechanisms governing how patterns of vocalizations are temporally sequenced. Similar observations in non-vocal behavioral sequences [29, 379] also exhibit these long-range sequential organizations, suggesting similarities in latent dynamics that facilitate long-range statistical relationships.

It is tempting to suggest that these parallels suggest shared underlying structure generating mechanisms, such as universals in the hierarchical organization of behavior (e.g. [84, 237], though we should be wary of making any extended inferences based upon the observation of long-range information decay. For example, we can infer that power-law sequential relationships are produced by non-Markovian mechanisms because the decay is not exponential. However, the set of generative mechanisms that can produce power-law relationships in signals is not understood well enough to attribute the origins of these relationships to, for example, any specific class of formal grammar. Power-law mutual information decay in signals can also be drawn simply from coupling vocal or behavioral 1/f noise found in exogenous environmental signals.

While it is well-acknowledged that many animal vocalizations are organized hierarchically [369, 84], the implications of that hierarchy in terms of underlying cognitive and physiological mechanisms are not well understood. For example, on very short timescales, birdsong motor sequencing is dictated by a hierarchical cascade of motor programs running

originating in the premotor region HVC eventually ending in motor output [100]. Recent physiological evidence shows that these high-level nuclei also contain information about future states displaced from current vocalizations as well [75], though the mechanisms by which those relationships are learned, maintained, and ultimately dictate behavior are not yet clear.

Although we do not have access to the mechanisms underlying the observed long-distance relationships in vocal and non-vocal behavioral sequences, we do know that many vocal and behavioral sequences cannot be well-captured by Markovian models, thus alternative methods for modeling, characterizing, and forming hypotheses about the long-range organization in behavioral sequences are crucial to furthering our understanding of long-range structure in behavioral sequences. One promising approach is the use of deep neural network models such as RNNs and transformer networks [436, 304]. Unlike Markov models, recent neural network models like RNNs and transformer models do capture power-law relationships in sequential data (Fig 1.5E) [400, 249]. In language, transformer networks, in particular, have changed the landscape of natural language processing by capturing deeply contextual and complex implicit relationships in linguistic sequences. In birdsong, the same approaches show promise [304]. For example, Morita et al., [304] train a transformer network on Bengalese finch song and find that it captures long-range dependencies extending well beyond that of a Markov model. Like modeling language sequences, however, neural-network-based approaches suffer from the same issues of being black box and providing little explanatory power over the sequential structure they learn. In addition, the amount of data required to train a model to capture complex sequential dependencies is vast. Although the number of parameters does not increase exponentially with the amount of context the model captures (as in Markov models) state-of-the-art transformer models have billions of parameters requiring training data comprised of billions to trillions of characters [49]. In language, the dataset size needed to train transformer models scales with the number of parameters in the model to prevent overfitting [191]. When dataset sizes are smaller, LSTM RNNs perform better than more state-of-the-art Transformer language models [115], though transformers allow you to explicitly specify the length of temporal context allowed

in the model, making them an attractive model for controlling context when generating vocal sequences [304]. Although non-human animal vocalization repertoires are smaller and syntactic organization is less complex than language, birdsong analyses relying on language models will need to address the same challenges.

Neural network-based models also provide the ability to capture temporal dependencies that mutual information and correlation functions do not. Mutual-information-function-based and correlation-based analyses compute relationships between vocal elements as a function of sequential distance, ignoring any temporal relationships between disparate elements. This is both a benefit and a shortcoming of correlation methods. Ignoring intermediary temporal relationships enables the characterization of structure at temporal distances without having to additionally model higher-order combinatorial relationships (e.g. $P(F|A)$) vs $P(F|A,B,C,D)$). For the same reason, mutual-information-function-based and correlation-based analyses are coarse descriptions of temporal structure and miss the full temporal dynamics of the signal that neural-network-based models can capture [304].

1.7 Synthesizing vocalizations

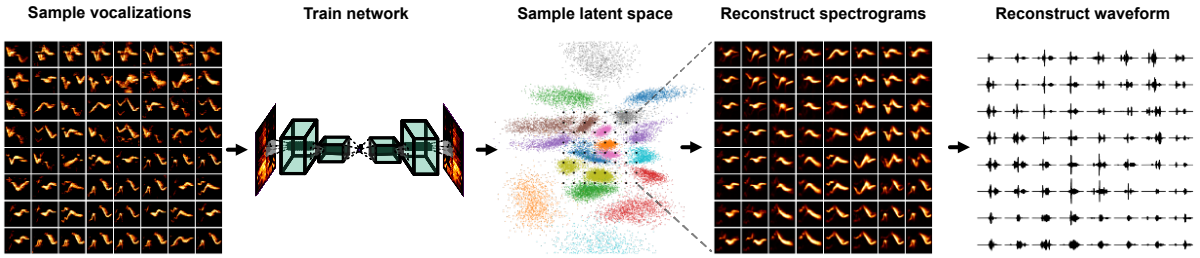


Figure 1.6. Steps involved in synthesizing vocalization from a Variational Autoencoder (VAE) trained on spectrograms.

Although the methods discussed in Section 1.5 allow us to learn representational spaces of animal vocalizations, providing new ways to infer structure in vocal repertoires, analyses on vocal signals alone lack grounding in animal behavior, perception, and physiology. In this section, we give an overview of methods for synthesizing animal vocalizations as a means to

systematically control vocalization stimuli and relate vocal representations to physiology and behavior.

An ideal model for vocal synthesis exhibits several features: (1) it can model the entire vocal repertoire of a species or multiple species, (2) the parameters of the model can be related to physiological properties of the vocalizing species, and (3) the parameters of the model can be explained in terms of understandable features (i.e. it is not a black box algorithm). Throughout this section, we find that current synthesis algorithms have tradeoffs in how they balance aspects of these ideals.

One reason to systematically synthesize animal vocalizations is to probe their perceptual and physiological representations of vocal space, for example, determining how animals categorically perceive the difference between two categories of vocal units [315]. Traditionally, categorical perception in animals has been studied on the basis of human speech sound stimuli [405, 227, 226]. Even with speech, however, the features that can be manipulated are limited. Recently methods in machine learning have furthered our ability to manipulate complex non-linear speech features substantially. These same approaches can often be applied to animal communication [11, 385].

Source-filter models

Source-filter models have their origins in vocoding speech [101], but have been used in numerous animal vocalization synthesis paradigms [92, 63, 134, 14]. Source-filter models decompose vocalizations into the source of the voice and filters [196]. For example, the STRAIGHT algorithm [196, 197] has been used to morph between macaque monkey vocalizations for investigations of monkey and human perception and physiology related to categorization [63, 134]. STRAIGHT breaks down the macaque vocalization into the fundamental frequency (the source) and its harmonics from higher-resonant or formant frequencies (the filter) [63]. It then uses landmarks based upon these estimated parameters from the two sounds being morphed and interpolates between them to generate the morph stimuli. Takafumi et al., [134], for example,

used this method to parametrically vary generated morphs based on source and filter properties to determine the features macaques use to distinguish between conspecifics. Soundgen [11] is a recent open-source GUI-based web app for R that is designed to synthesize nonverbal vocalizations using a source-filter model, including animal vocal signals such as birdsong and primate vocalizations. Related source-filter models have been developed to synthesize birdsong based upon underlying physiological mechanisms [14, 118, 406, 407, 15]. Recently, Arneodo et al., [13] demonstrated that synthetic source-filter models can be coupled with neural recordings accurately reconstruct vocalizations from neural data alone. One drawback of source-filter models is the difficulty with which they can be fitted to the diversity of non-human vocalizations that exist. For example, the source-filter models of birdsong described above can well describe the dynamics of zebra finch song, but not the dual-syringeal dynamics of European starling song. Without reference to explicit hypotheses about underlying production mechanism, HMM based source-filter approaches provide one potential solution to this problem birdsong [38].

Neural network models

An alternative approach to synthesizing animal vocalizations is the use of neural-network-based synthesis algorithms. These neural-network-based algorithms can be used to sample directly from the learned representational spaces described in Section 1.3. A simple example is autoencoder-based synthesis [472, 382]. Autoencoders can be trained on spectral representations of vocal data, and systematically sampled in the learned latent space to produce new vocalizations. Insofar as the neural network or latent projection can learn to represent the entire vocal repertoire, the entire vocal repertoire can be sampled from. In addition to sampling vocalizations from a latent distribution, vocal features can be manipulated in latent space. Well-defined latent spaces and higher-dimensional latent projections can learn to linearize complex non-linear reltionships in data. For example, in pictures of faces, the presence of a glasses, hair color, and the shape of a person's face can all be manipulated as linear features [380, 386, 358]. With more complex features, such as the attractiveness of a call or the age of the vocalizer, a promising avenue for

future research would be to synthesize vocalizations, varying these complex non-linear features for playback studies.

Like most areas of deep learning, substantial progress has been made on the task of audio synthesis in the past few years. Basic methods comprise autoencoders [213, 112, 385], Generative Adversarial Networks (GANd) [385, 431, 330, 96, 111] and autoregressive approaches [281, 327, 351, 189]. One advantage of GAN-based models is that their loss is not defined directly by reconstruction loss, resulting in higher-fidelity syntheses [236]. Typically, approaches for synthesizing vocalizations based on neural networks rely on treating magnitude spectrogram like an image, training a neural network architecture in the same manner as one would an image, and finally inverting the sampled spectrogram into a waveform [384, 472, 330]. When synthesizing vocalizations from neural networks trained on the magnitude spectrogram, the estimation of phase is necessary to invert the spectrogram into a waveform signal for playback. The de-facto algorithm for spectral inversion has been Griffin and Lim [158], though several recent approaches have been shown to improve over the Griffen-Lim algorithm recently [272, 356]. An alternative to Griffen-Lim inversion is to train neural networks to invert spectrograms either directly in the neural network architecture [228], or perform inversion in a second network [272]. Spectrogram-based audio synthesis can also be sidestepped entirely, training the network directly on waveform [327, 281, 112].

Sound texture synthesis

Another approach to sound synthesis is the synthesis of sound texture [275, 387]. For example, McDermott et al., [275] propose an approach that relies on computing a set of statistics over stationary elements of sounds, and synthesizing new sounds based upon the computed statistics. By manipulating or interpolating between sound statistics, they synthesize new sound textures. One application, for example, is to manipulate components of sound textures for stimulus playback to determine what sound texture statistics listeners rely upon for recognition [276].

Generating sequences

A parallel approach to synthesizing vocalizations is to generate vocal sequences from symbolically labeled vocal elements. Synthetic song sequences can be used to understand how animals process and represent temporal and sequential organization. For example, can a songbird differentiate between sounds generated using different underlying models of song syntax? Traditional approaches to song sequence generation rely upon the explicit, hand-crafted, generation of artificial grammars for playback studies. By crafting artificial grammars that differ in underlying structure, such as belonging to different classes of the Chomsky hierarchy [201, 123, 140], playback studies can be used to determine whether animals can learn these grammars. A number of challenges exist with artificial grammar learning studies, however. One such challenge is the difficulty in crafting sequences that can exclusively be learned by inferring the structure that generated them, for example, making it impossible for the animal to learn by brute-force memorizing every sequence [123]. When using artificial grammars, computational and modeling considerations aid in forming hypotheses about how generated grammars can be used. In the context of the neuroethology of vocal communication, these cognitive models can be related to physiological measures [472]. An additional challenge with artificial grammar learning is constructing sequences that are structured in a similar way to natural and behaviorally relevant signals to the animal. For example, artificial grammar studies usually rely on short sequences modeled after human language syntax, rather than the animal's own communication systems. Because the task of generating vocal sequences is performed over symbolic representations of syllables, generating vocal sequences can be performed using the same methods as in text or musical note generation. These approaches can range from generating sequences using Markov models of various orders, to explicitly modeling hierarchical organization in the signal generation algorithm [366].

1.8 Mapping vocal communication to perception, behavior, and physiology

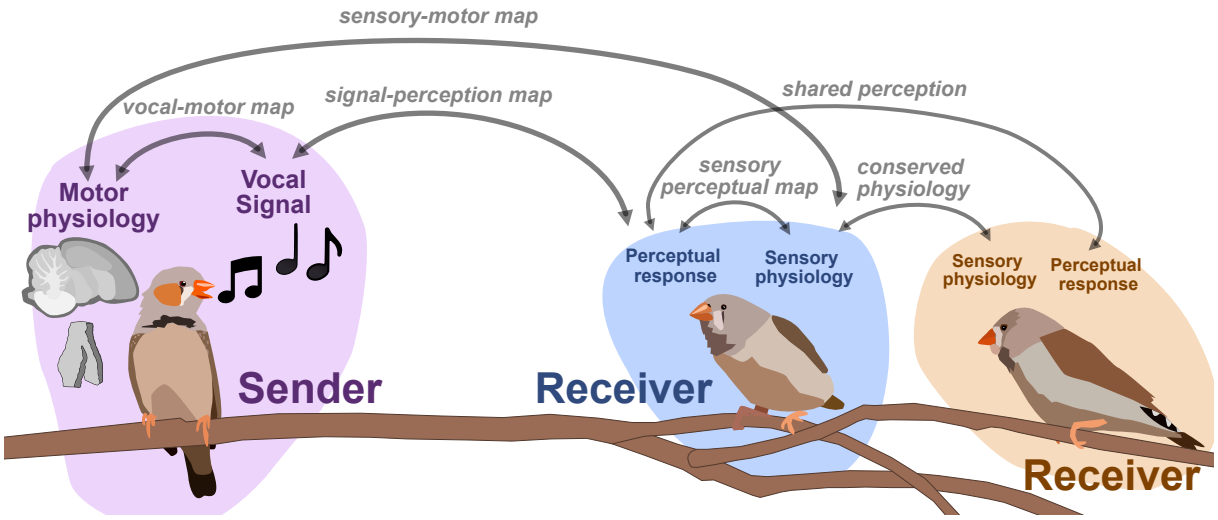


Figure 1.7. An outline of mappings between perceptual, acoustic, and physiological signals. One focus in sensory and motor neuroscience is to learn the relationships between signals, perception, and physiology.

The methods discussed here provide a framework to develop a set of constrained spaces from which to understand and model vocal behavior in relation to perception, production, and physiology. Perceptual or relational vocal spaces, such as UMAP projections of spectrograms, provide a low-dimensional space that can be used to infer structure in vocal repertoires. Likewise, symbolic abstractions of vocal behavior to large corpora provides a categorical representation in which vocal behavior is seen as sequential actions on those category sets. In both cases, the methods provide a constrained behavioral representation for physiological analyses.

Brain-computer interfacing

One of the primary challenges facing the field of brain-computer interfacing is scaling up from simple behavioral spaces like moving a cursor on a screen to complex behaviors [137]. A clear advantage to the approaches discussed in Section 1.5 is that we can learn to bring complex vocal behavioral spaces into a compressive low-dimensional behavior spaces, even without a

prior model of the structure in that space. For example, Arneodo et al., [13] find that directly predicting the acoustic structure of zebra finch song from neural data does not perform as well as predicting the parameters of a low-dimensional biophysical model of song production. In the many species in which we do not have access to a biophysical model of vocal production, learned acoustic spaces may be a viable alternative. In contrast, the current state-of-the-art vocal prostheses for speech bypass biophysical models, directly predicting sentences (i.e. symbolic sequences) with the aid of language models (i.e. a sequence model) [306]. Such methods do not capture important extra-linguistic information such as emotional tone and stress. In future work, a clear pathway forward is to develop BCI models that can both capture symbolic organization aided by sequential models, as well as within-symbol variability in the acoustic signal.

Vocal production

Songbirds as a model for systems neuroscience are perhaps best known for the role they play in our understanding of vocal learning [100]. In addition to songbirds, rodent and non-human primate vocal behavior are becoming increasingly prominant models in the neuroscience of vocal production. In non-human primates, recent evidence has suggested some degree of constrained vocal learning in some species [122]. In rodents, recent focus has been placed upon variability and structure mouse in ultrasonic vocalizations (USVs) [345, 18, 176], singing mice have emerged as a physiological model of turn-taking [325], and the cultural transmission of vocal dialect has been observed in the naked mole rat [22]. In each of these cases, quantification of how vocalizations vary as well as relationships between vocalizations (either within individual, between conspecifics, or from tutor to pupil) is integral to understanding how we learn to navigate vocal space. For example, Kollmorgen et al., [216] use nearest-neighbor graphs and t-SNE projections to quantify and visualize the developmental trajectory of zebra finch song during vocal learning. For each syllable, they compute a nearest-neighbors graph based metric termed the ”neighborhood production time”, which quantifies the developmental time point at which similar (neighboring) syllables were sung. For example, a syllable song on day 45

might have 10 neighbors, sung on days between day 40 and 50, comprising its neighborhood production times. Syllable renditions that are neighbors with predominantly future syllables are deemed anticipations, while syllable repetitions that are neighbors with predominantly past syllables are deemed regressions. They observe that day-by-day, zebra finch songs gradually moves along a constant vocal learning trajectory, but anticipations and regressions differ in how they are consolidated overnight.

The number of neurons we can simultaneously record from physiologically has increased the dimensionality of neural datasets substantially over the past decade, making methods for dimensionality reduction on neural signals such as spike trains increasingly necessary for neural data analysis and opening the door to computational methods that directly link the latent representations of behavioral and neural datasets. Population modeling approaches such as LFADS (Latent Factor Analysis via Dynamical Systems; [333]) reduce large population spiking datasets into low-dimensional trajectories, similar to the approaches discussed here with vocal signals. In the case of LFADS, these embeddings are performed over single trials using a recurrent autoencoder. One promising direction for computational neuroethology is learning the relationship between latent behavioral states and latent physiological states. By developing tools that allow us to learn the relationship between physiological and behavioral representations, we hope to untangle how, for example, movements in behavioral space reflect changes in physiology, and vice-versa. Singh Alvarado et al., [404] developed a joint encoding model in which they used variational autoencoders to learn a joint latent representation of spectrograms of zebra finch song, and corresponding ensemble neural activity of spiny neurons in songbird basal ganglia (average calcium fluorescence of around 60 ROIs, or putative neurons, per bird). In a series of experiments leading up to this joint mapping, Singh Alvarado et al. demonstrated that Area X spiny neurons are involved in the regulation of vocal variability; exhibiting suppressed activity during female-directed song and enhanced activity during practice. Using the joint vocal-neural latent mapping, they were able to uncover the mapping between specific features of song and variants present in neural ensemble activity. In Figure 1.7 we outline several similar maps

between behavior, perception, and neural dynamics. Singh Alvarado et al.'s work exhibit that one such latent map, a vocal-motor mapping between motor physiology and vocal behavior, can uncover complex and detailed relationships that traditional methodology cannot. Similar mappings between the physiology, perception, and behavior of sender-receiver dynamics (e.g. Fig 1.7) are also well poised to benefit from emerging latent approaches.

The physiology of vocal syntax is another area poised to benefit from computational ethology. One example is the role of the songbird premotor nucleus, HVC, in encoding song syntax. Birdsong has a long history of being described sequentially in terms of low-order Markovian transitions between song elements. HVC's role in song syntax, until recently has been described exclusively in terms of these low-order transition statistics [129]. In a recent example, however, Cohen et al., [75] made use of an automated birdsong labeling paradigm and high-order sequence model to observe 'hidden neural states' encoding sequentially displaced (i.e. high-order) transitions in the premotor nucleus HVC of canaries. To identify non-adjacent dependencies in the song, they used a Prediction Suffix Tree [266], which can capture high-order Markovian relationships in the song syntax. Prediction Suffix Trees have previously been used to observe long-range dependencies up to the 7th order in canaries [266]. While birds were singing, Cohen et al., used a miniature microscope to image neurons from HVC, a region involved in the songbird vocal motor circuit. They observed that HVC ROIs were locked to individual song-phrases and transitions, and that this phrase locking is modified by non-adjacent context, displaced by several phrases and seconds. As more recent approaches give access to larger datasets enabling the identification of longer-range dependencies in birdsong, it is currently not clear whether we have yet found an upper bound on the sequential displacement of long-range representations of vocal syntax in physiology.

Vocal perception

Similar to vocal production, latent and sequential models are promising avenues for better understanding cognitive and physiological underpinnings of vocal perception. In songbirds,

primates, and rodents, many foundational studies of auditory categorical perception, perceptual decision making, and their underlying physiology rely upon either relatively simple stimuli such as tones or complex stimuli like human speech [467, 374, 226, 227, 426]. Categorization in these stimulus spaces are attractive because they are well-characterized and understood. Across species, however, neural responses are often tied to complex and more behaviorally-relevant acoustic phenomena such as recognizing and discriminating between conspecific vocalizations [254, 21]. When the acoustic features underlying vocal repertoires are simple and known, categorical stimuli can be selected directly based upon those features. For example, Lachlan et al., [233] manipulate a single dimension, the duration of swamp sparrow notes, to determine how notes are categorically perceived in different sequential contexts. In speech, voice onset time (VOT) is a similar single-dimension commonly used for categorical perception paradigms [248]. However, it is rarely the case that categorical perception is driven by a single dimension. Thus building stimuli in more complex feature spaces will be necessary to untangle the relationship between vocal features, perception, and physiology. When biophysical models of vocal structure exist, species relevant stimuli can be generated using biophysical parameters [14]. When the underlying acoustical structure of a vocal repertoire is more complex and biophysical models of vocalizations have not been defined, neural-network synthesized vocalizations are an attractive alternative. For example, as discussed above, birdsong can be synthesized with neural networks for physiological and perceptual playback studies to determine perceptual similarity between syllables or learn categorical boundaries between song-morphs [472, 382, 430]. By systematically controlling the signal space of a vocal repertoire, we can systematically explore how changes in that space relate to changes in physiology.

Algorithmic approaches are similarly well poised to aid in our understanding of how vocal sequences are maintained and represented. Sequence learning research in human and non-human primates is largely dominated by artificial grammar learning (AGL) research, an umbrella category that comprises several different forms of sequence learning ranging from hierarchically nested tree structures to transitional probabilities [89]. Artificial grammar learning

studies aim to determine what structures animals (and humans) are capable of learning, what cognitive mechanisms underlie grammar induction, and what physiological systems underlie those cognitive mechanisms. In the domain of primate sequence learning, neural pathways are generally conserved between humans and non-human primates and involve the ventral regions of cortex [461]. Determining an appropriate stimuli set is requisite for developing an AGL paradigm. Latent representations of vocalizations can aid in choosing stimuli from a well-defined stimulus space. For example, when choosing a stimulus set for an A^nB^n grammar, it is desirable depending on the goal of the task to ensure that the constituent vocalizations comprising A and B belong to equidistant or separate clusters in acoustic or perceptual spaces [472].

While artificial grammar learning has also played a prominent role in birdsong sequence learning [427], the structure underlying an animal's own vocal syntax provides an opportunity to study the neural and cognitive underpinnings of a more ethologically-relevant complex sequential structure. Despite the important role vocal syntax production has played in establishing birdsong as a model in systems neuroscience, a surprising gap exists in our knowledge of the physiological circuits underlying how syntactic information is recognized and sequentially integrated when listening to song. Songbird vocal communication contains often very complex syntax that can be structured over long timescales comprising often tens to hundreds of unique, stereotyped vocal units [71]. Conspecifics pay attention to the structure of that song. Abe and Wantanabe [1] developed a habituation/dishabituation paradigm with Bengalese finches alongside immediate early gene (IEG) expression and lesioning experiments to explore the role of song nuclei on the recognition of grammatical sequences. They found that IEG expression increased when presented with non-conforming/nonpredictive strings in the nuclei LMAN, a basal ganglia output nuclei characterized by recurrent loops that is also involved in vocal learning [39]. Abe and Wantenable then lesioned LMAN and measured song discrimination with their habituation paradigm. They found that discrimination was disturbed in birds where LMAN was lesioned, implicating LMAN in the ability to discriminate syntactic song. How syntactic information is learned, integrated, and maintained in LMAN and associated striatal regions of songbird brain are still open questions.

In contrast to the auditory domain where little is known about syntactic integration, a neural correlate for complex and abstract information integration, NCL, has been well established and characterized in songbird vision with pigeons and corvids [160, 223]. Strong parallels exist between NCL and the primate prefrontal cortex, which is involved in sequence learning. NCL has variously been associated with rule learning [444], numerosity [451, 95], directed forgetting [370, 290, 169], choice behavior [190], working memory [91, 364], sequence learning [170], and reward learning. Anatomically and neurochemically NCL also exhibits strong parallels to the primate prefrontal cortex. NCL is characterized by similar circuitry from auditory and dopaminergic afferents, as well as multi-sensory projections [223, 450]. Surprisingly, however, an auditory equivalent to the visual working-memory region in NCL has not been found, though they have been observed in the multi-sensory audio-visual integration and association [296, 295]. Birdsong is well poised as a signal to be a model of vocal syntax perception, To establish this model, however, it will be imperative to uncover the systems in songbird brain related to working memory and temporal context integration in song. NCL appears to be a likely candidate for processing syntactic vocal signals, though, as yet, this has not been found to be the case.

Although mouse USVs do not appear to contain temporal structure to the same extent as songbirds, mouse USVs are temporally organized [62] and female mice also show preference for more complex syllables and sequences [176], making mouse USVs another potential target for the study of syntactic and sequential integration in vocal perception.

1.9 Discussion

This review covers emerging approaches in the computational neuroethology of vocal communication enabling researchers to engage with large and diverse datasets of vocal signals and to represent them in computationally tractable frameworks.

We started by discussing techniques to process and represent acoustic signals. We then discussed how to parse complex vocal datasets into species, individuals, and discrete vocal

elements. Next, we discussed how relational structure can be extracted from vocal signals, how these signals can be clustered in learned latent spaces, and how these latent spaces capture different aspects of the information contained within the underlying signals. We then discussed how temporal structure can be inferred from vocal units, including emerging work on the non-Markovian dynamics underlying vocal behavior. In the next section, we discussed how vocalizations can be synthesized for use in playback experiments that allow an unprecedented degree of control over non-linear and complex vocal feature spaces. Finally, we discussed how these approaches are being applied to the field of neuroethology and emerging frameworks for understanding vocal signals and their underlying physiology.

The methods discussed here provide a promising avenue for a broader, more diverse, and larger-scale neuroethology of vocal communication, than the research practices that have dominated the past several decades, and hold the promise of expanding both the breadth and depth of our understanding. Instead of focusing on a small number of model species, new computational techniques provide a framework for studying vocal behavior across a wide range of animals.For example, research on vocal learning in songbirds has ignored the majority of species, female birdsong, and most call types [255]. Likewise, because these new computational methodologies can often deal with unstructured data, they enable us to expand beyond simplified, isolated behaviors in controlled environments to more natural or naturalistic behavioral contexts where dynamics involving multi-modal integration and multi-animal social interactions arise. As we capture increasing levels of detail in behavior, our understanding of its sophistication naturally follows. Already, these new computational framework have revealed deep structure in the sequential organization of communication, where large-scale datasets of both symbolic sequences, and latent projections that capture rendition-to-rendition variability, have enabled quantitative analyses of rare (but perhaps meaningful) events, such as long-range syntactic organization. Together, all of these approaches point toward a new framework, in which complex and non-linear behavioral and physiological signals can be represented in compressive and tractable spaces that can capture the complex dynamics and relationships in the increasingly rich

datasets available to researchers.

As with any powerful tool, these techniques require careful consideration when put into practice. Broadly, automation and machine learning in data analysis can be fraught with unexpected complications and confounds that may be hard to spot. For example, automating the labeling of large datasets of birdsong syllables can speed up the task of labeling by days, weeks, or months, but can also leave the experimenter with less intuitive knowledge of the animal's vocal repertoire, resulting in a loss of domain knowledge. As we have noted elsewhere [384], when domain knowledge is available it should be integrated with one computational approach. Another potential pitfall (and a source of much needed research effort) is in understanding the structure of the latent manifolds that are yielded in many of the described methods. In particular, non-linear latent modeling techniques like UMAP or neural networks can capture complex relationships in vocal data, but interpreting these projections requires an understanding of how data are represented within the geometry of the latent space. For example, UMAP captures primarily local structure in datasets that are present in nearest neighbor graphs, meaning that the relative distances of vocal elements have no explicit relation to the data, as is the case in PCA for example.

Attending to the cautions of computational abstraction, the approaches discussed in this manuscript provide a framework from which to quantify vocal signals that promises to yield important new insights into vocal behavior and neurobiology. These approaches enable neuroethologists to project vocalizations onto low dimensional latent manifolds, visualize and quantify the transitional structure and information decay of vocal syntax, and map vocal and neural repertoires into shared neural spaces for functional representation and action. As the richness of datasets grow to capture more of the complexities of behavior and physiology, methods and frameworks for modeling and inferring structure in ethological data are increasingly necessary for hypothesis formulation and testing. The methods and frameworks discussed in this review parallel and supplement those in the broader field of computational neuroethology.

Author Contributions

This manuscript was written by T.S. and T.Q.G.

Funding

This work was supported by a CARTA Felloship to T.S., NIH 5T32MH020002-20 to T.S., and 5R01DC018055-02 to T.G.

1.10 Acknowledgments

Chapter 1, in full, is a reprint of a manuscript under review. Sainburg, Tim, Gentner, Timothy Q (2021) The dissertation author was the primary investigator and author of this paper.

Chapter 2

Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires

Abstract

Animals produce vocalizations that range in complexity from a single repeated call to hundreds of unique vocal elements patterned in sequences unfolding over hours. Characterizing complex vocalizations can require considerable effort and a deep intuition about each species' vocal behavior. Even with a great deal of experience, human characterizations of animal communication can be affected by human perceptual biases. We present a set of computational methods for projecting animal vocalizations into low dimensional latent representational spaces that are directly learned from the spectrograms of vocal signals. We apply these methods to diverse datasets from over 20 species, including humans, bats, songbirds, mice, cetaceans, and nonhuman primates. Latent projections uncover complex features of data in visually intuitive and quantifiable ways, enabling high-powered comparative analyses of vocal acoustics. We introduce methods for analyzing vocalizations as both discrete sequences and as continuous latent variables. Each method can be used to disentangle complex spectro-temporal structure and observe long-timescale organization in communication.

2.1 Introduction

Vocal communication is a common social behavior among many species, in which acoustic signals are transmitted from sender to receiver to convey information such as identity, individual fitness, or the presence of danger. Across diverse fields, a set of shared research questions seeks to uncover the structure and mechanism of vocal communication: What information is carried within signals? How are signals produced and perceived? How does the communicative transmission of information affect fitness and reproductive success? Many methods are available to address these questions quantitatively, most of which are founded on underlying principles of abstraction and characterization of 'units' in the vocal time series [200]. For example, segmentation of birdsong into temporally discrete elements followed by clustering into discrete categories has played a crucial role in understanding syntactic structure in birdsong [200, 32, 381, 194, 266, 71, 165, 220, 141].

The characterization and abstraction of vocal communication signals remains both an art and a science. In a recent survey, Kershenbaum et. al. [200] outline four common steps used in many analyses to abstract and describe vocal sequences: (1) the collection of data, (2) segmentation of vocalizations into units, (3) characterization of sequences, and (4) identification of meaning. A number of heuristics guide these steps, but it is largely up to the experimenter to determine which heuristics to apply and how. This application typically requires expert-level knowledge, which in turn can be difficult and time-consuming to acquire, and often unique to the structure of each species' vocal repertoire. For instance, what constitutes a 'unit' of humpback whale song? Do these units generalize to other species? Should they? When such intuitions are available they should be considered, of course, but they are generally rare in comparison to the wide range of communication signals observed naturally. As a result, communication remains understudied in most of the thousands of vocally communicating species. Even in well-documented model species, characterizations of vocalizations are often influenced by human perceptual and cognitive biases [419, 437, 182, 200]. We explore a class of unsupervised,

computational, machine learning techniques that avoid many of the foregoing limitations, and provide an alternative method to characterize vocal communication signals. Machine learning methods are designed to capture statistical patterns in complex datasets and have flourished in many domains [239, 27, 27, 358, 24, 46, 24]. These techniques are therefore well suited to quantitatively investigate complex statistical structure in vocal repertoires that otherwise rely upon expert intuitions. In this paper, we demonstrate the utility of unsupervised latent models, statistical models that learn latent (compressed) representations of complex data, in describing animal communication.

2.1.1 Latent models of acoustic communication

Dimensionality reduction refers to the compression of high-dimensional data into a smaller number of dimensions, while retaining the structure and variance present in the original high-dimensional data. Each point in the high-dimensional input space can be projected into the lower-dimensional 'latent' feature space, and dimensions of the latent space can be thought of as features of the dataset. Animal vocalizations are good targets for dimensionality reduction. They appear naturally as sound pressure waveforms with rich, multi-dimensional temporal and spectral variations, but can generally be explained by lower-dimensional dynamics [344, 138, 15]. Dimensionality reduction, therefore, offers a way to infer a smaller set of latent dimensions (or features) that can explain much of the variance in high-dimensional vocalizations.

The common practice of developing a set of basis-features on which vocalizations can be quantitatively compared *(also called Predefined Acoustic Features, or PAFs)* is a form of dimensionality reduction and comes standard in most animal vocalization analysis software (e.g. Luscinia [234], Sound Analysis Pro [423, 424], BioSound [110], Avisoft [411], and Raven [64]). Birdsong, for example, is often analyzed on the basis of features such as amplitude envelope, Weiner entropy, spectral continuity, pitch, duration, and frequency modulation [423, 200]. Grouping elements of animal vocalizations (e.g. syllables of birdsong, mouse ultrasonic vocalizations) into abstracted discrete categories is also a form of dimensionality reduction,

where each category is a single orthogonal dimension. In machine learning parlance, the process of determining the relevant features, or dimensions, of a particular dataset, is called *feature engineering*. Engineered features are ideal for many analyses because they are human-interpretable in models that describe the relative contribution of those features as explanatory variables, for example explaining the contribution of the fundamental frequency of a *coo* call in predicting caller identity in macaques [130]. As with other human-centric heuristics, however, feature engineering has two caveats. First, the features selected by humans can be biased by human perceptual systems, which are not necessarily "tuned" for analyzing non-human communication signals [419, 109]. Second, feature engineering typically requires significant domain knowledge, which is time-consuming to acquire and difficult to generalize across species, impairing cross-species comparisons.

An attractive alternative to feature engineering is to project animal vocalizations into low-dimensional feature spaces that are determined directly from the structure of the data. Many methods for data-driven dimensionality reduction are available. PCA, for example, projects data onto a lower-dimensional surface that maximizes the variance of the projected data [102, 200], while multidimensional scaling (MDS) projects data onto a lower-dimensional surface that maximally preserves the pairwise distances between data points. Both PCA and MDS are capable of learning manifolds that are linear or near-linear transformations of the original high-dimensional data space [428].

More recently developed graph-based methods extend dimensionality reduction to infer latent manifolds as non-linear transformations of the original high-dimensional space using ideas from topology [428, 280, 260]. Like their linear predecessors, these non-linear dimensionality reduction algorithms also try to find a low-dimensional manifold that captures variation in the higher-dimensional input data, but the graph-based methods allow the manifold to be continuously deformed, by for example stretching, twisting, and/or shrinking, in high dimensional space. These algorithms work by building a topological representation of the data and then learning a low-dimensional embedding that preserves the structure of the topological representation (Fig

2.1). For example, while MDS learns a low-dimensional embedding that preserves the pairwise distance between points in Euclidean space, ISOMAP [428], one of the original topological non-linear dimensionality reduction algorithms, infers a graphical representation of the data and then performs MDS on the pairwise distances between points within the graph (geodesics) rather than in Euclidean space. These graph-based methods are often preferable to linear methods because they capture more of the local structure of the data, but these benefits do have a cost. Whereas the latent dimensions of PCA, for example, have a ready interpretation in terms of the variance in the data, the ISOMAP dimensions have no specific meaning beyond separability [280]. In addition, in practice, high-level (global) structure in the dataset, like the distances between clusters in low-dimensional embeddings, can be less meaningful in graph-based dimensionality reduction than in PCA or MDS, because current graph-based methods tend to local-notions of distance like nearest neighbors to construct a graphical representation [454].

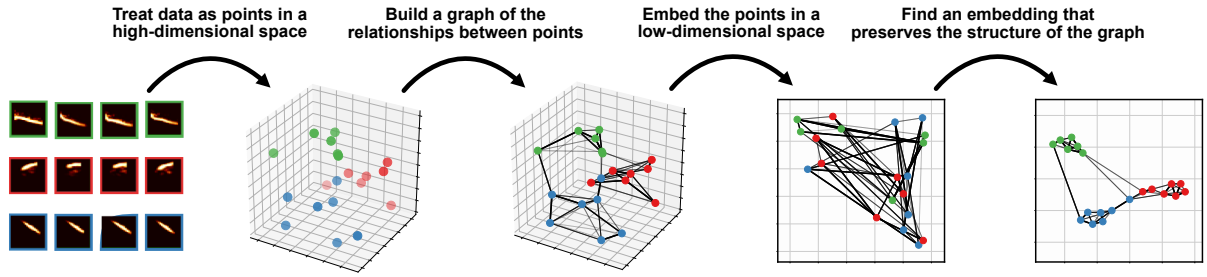


Figure 2.1. Graph-based dimensionality reduction. Current non-linear dimensionality reduction algorithms like TSNE, UMAP, and ISOMAP work by building a graph representing the relationships between high-dimensional data points, projecting those data points into a low-dimensional space, and then finds and embedding that retains the structure of the graph. This figure is for visualization, the spectrograms do not actually correspond to the points in the 3D space.

The utility of non-linear dimensionality reduction techniques are just now coming to fruition in the study of animal communication, for example using t-distributed stochastic neighborhood embedding (t-SNE; [260]) to describe the development of zebra finch song [215], using Uniform Manifold Approximation and Projection (UMAP; [280]) to describe and infer categories in birdsong [150, 381], or using deep neural networks to synthesize naturalistic acoustic stimuli [382, 430]. Developments in non-linear representation learning have helped fuel the most recent

advancements in machine learning, untangling statistical relationships in ways that provide more explanatory power over data than traditional linear techniques [27, 239]. These advances have proven important for understanding data in diverse fields including the life sciences (e.g. [381, 29, 76, 215, 150, 24]), in part due to their utility in rapidly extracting complex features from increasingly large and high-dimensional datasets.

In this paper, we describe a class of nonlinear latent models that learn complex feature-spaces of vocalizations, requiring few *a priori* assumptions about the features that best describe a species' vocalizations. We show that these methods reveal informative, low-dimensional, feature-spaces that enable the formulation and testing of hypotheses about animal communication. We apply our method to diverse datasets consisting of over 20 species (Supporting information), including humans, bats, songbirds, mice, cetaceans, and nonhuman primates. We introduce methods for treating vocalizations both as sequences of temporally discrete elements such as syllables, as is traditional in studying animal communication [200], as well as temporally continuous trajectories, as is becoming increasingly common in representing neural sequences [79]. Using both methods, we show that latent projections produce visually-intuitive and quantifiable representations that capture complex acoustic features. We show comparatively that the spectrotemporal characteristics of vocal units vary from species to species in how distributionally discrete they are and discuss the relative utility of different ways to represent different communicative signals.

2.2 Results

2.2.1 Dimensionality reduction

The current state-of-the-art graph-based manifold learning algorithms are t-SNE [260] and UMAP [280]. Like ISOMAP, t-SNE and UMAP first build a topological (graphical) representation of the data, and then project that graph into a lower-dimensional embedding, preserving as much of the topological structure of the graph as possible. Both embedding

methods are unsupervised, meaning they do not require labeled data. To visually compare the graph-based dimensionality reduction algorithms UMAP and t-SNE to the more classical linear methods PCA and MDS, we projected spectrograms of a dataset of Egyptian fruit bat infant isolation calls from 12 individuals into 2-dimensional PCA, MDS, t-SNE, and UMAP (Fig 2.2). Broadly, we can see that PCA and MDS projections are more diffuse (Fig 2.2A,B), while t-SNE and UMAP capture much more of the local similarity structure across the dataset, tightly packing together calls from the same individuals (Fig 2.2C,D).

Throughout this manuscript, we chose to use UMAP over t-SNE because UMAP has been shown to preserve more global structure, decrease computation time, and effectively produce more meaningful data representations across a number of datasets within the natural sciences (e.g. [24, 381, 150, 280]).

Both t-SNE and UMAP are underlied by functionally similar steps: (1) construct a probabilistically weighted graph and (2) embed the graph in a low-dimensional embedding space (see Fig 2.1). To build a probabilistically weighted graph, UMAP and t-SNE first build a nearest-neighbor graph of the high-dimensional data using some distance metric (e.g. the Euclidean distance between spectrograms). They then compute a probability distribution over the edges of that graph (pairs of nearest neighbors), assigning higher weights to closer pairs, and lower weights to more distant pairs. Embedding that graph in lower-dimensional space is then simply a graph-layout problem. An embedding is first initialized (e.g. using PCA or a spectral embedding of the graph). UMAP and t-SNE then compute the probabilities over the relationships between projections in the embedding space, again where closer pairs of elements are assigned a higher probability and more distant pairs are assigned a lower probability. Using gradient-descent, the embeddings are then optimized to minimize the difference between the probability distribution computed from the nearest-neighbor graph and the probability distribution in the embedding space.

UMAP and t-SNE differ in how these graphs are constructed and how embeddings are optimized. UMAP, in particular, assumes that the high-dimensional space in which the data

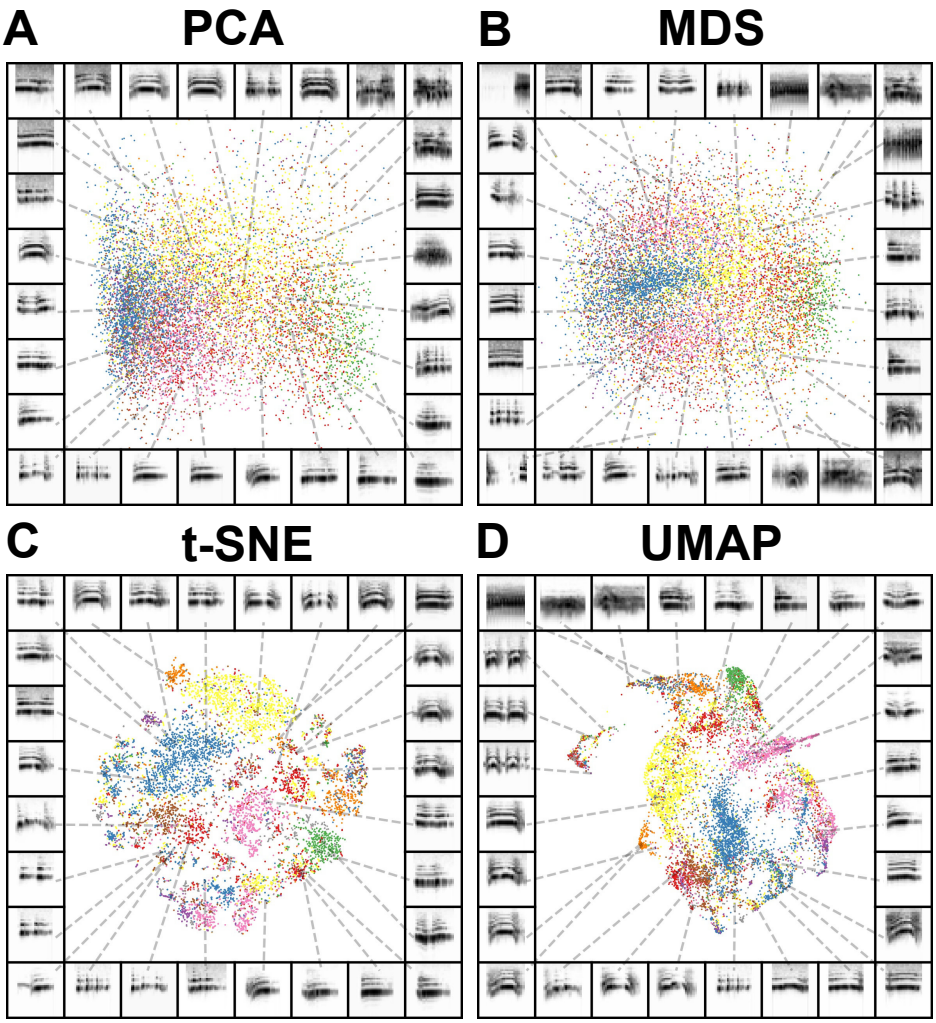


Figure 2.2. Comparison between dimensionality reduction and manifold learning algorithms. Isolation calls from 12 juvenile Egyptian fruit bats, where spectrograms of vocalizations are projected into two dimensions in (A) PCA, (B) MDS, (C) t-SNE, and (D) UMAP. In each panel, each point in the scatterplot corresponds to a single isolation call. The color of each point corresponds to the ID of the caller. The frame of each panel is a spectrogram of an example syllable, pointing to where that syllable lies in the projection.

lives is warped, such that data are uniformly distributed on a non-linear manifold in the original dataspace. UMAP's construction of the graphical representation of the data uses concepts from topology, so that the edges of the graph (the connections between data points) are probabilistically weighted by distance on the uniform manifold. The embeddings are then found by minimizing the cross-entropy between the graph and a probability distribution defined over the relationships

between embeddings. In other words, an embedding is learned that tries to preserve as much of the topological structure of the original graph as possible.

UMAP has several parameters for constructing its graph and embedding it in a low-dimensional space. The four primary UMAP parameters are `n_neighbors` which determines how many neighbors (nearby data points) are used in constructing the nearest-neighbor graph, `min_dist` which determines how spread apart connected embedding are allowed to be, `n_components` which is the dimensionality of the embedding space, and `metric` which defines the distance metric (e.g. Euclidean, cosine) that is used to define distances between points in the high-dimensional dataspace. We use the default parameters for each, except when otherwise noted.

2.2.2 Choosing features to represent vocalizations

Choosing the best features to represent vocal data is difficult without significant domain knowledge. In some species, the features underlying behaviorally-relevant variability in vocalizations are well documented and understood. When such information about a species' vocal repertoire is known, those features can and should be used to make comparisons between vocalizations within species. When analyzing vocalizations across species or within species whose vocal repertoires are less well understood, choosing features to represent vocalizations is more difficult: features that capture only a subset of the true behaviorally relevant variance can bias downstream analyses in unforeseen ways.

Two methods for choosing feature-sets are commonly used by experimenters when the features underlying vocal data are unknown: (1) extract common descriptive statistics of vocalizations, sometimes called Predefined Acoustical Features (PAFs; e.g. mean fundamental frequency, syllable length, spectral entropy) and make comparisons on the basis of PAFs, or (2) make comparisons based upon time-frequency representations of the data (i.e. spectrograms) where the magnitude of each time-frequency component in the spectrogram is treated as an independent feature (or dimension) of the vocalization.

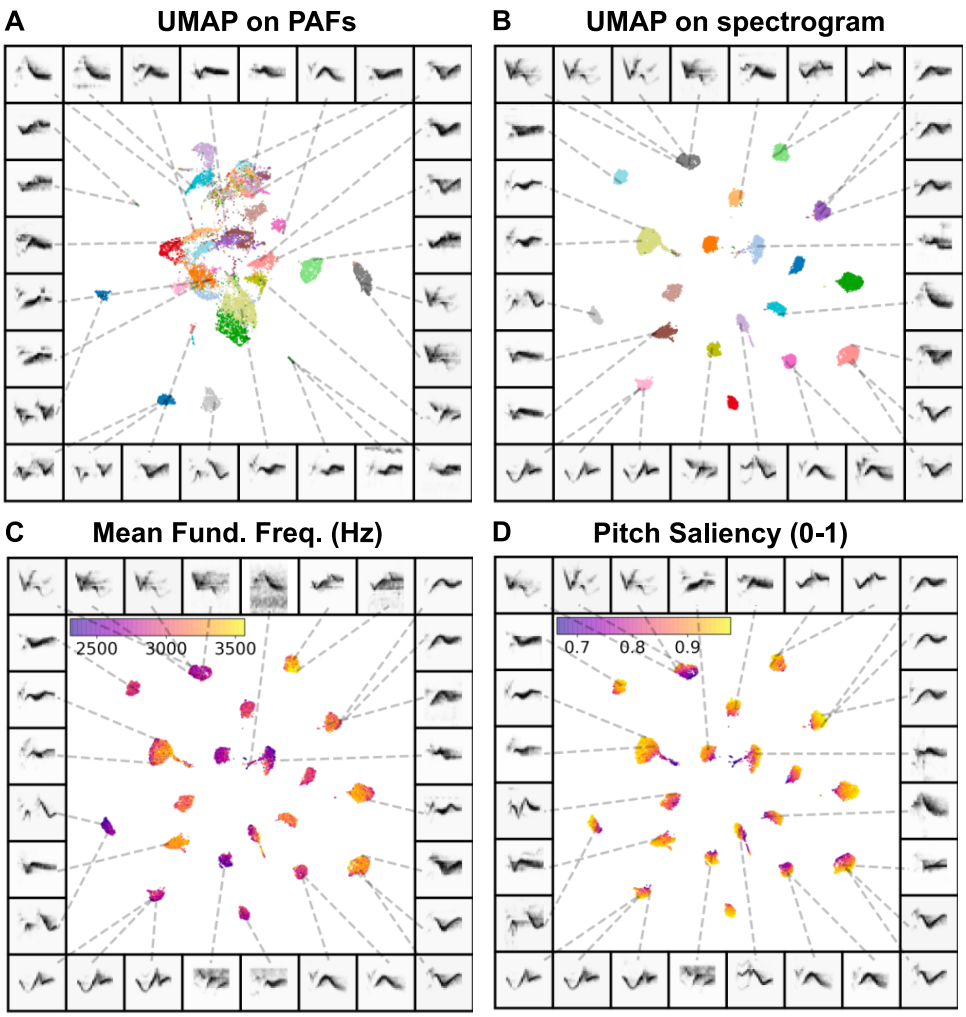


Figure 2.3. Comparison between dimensionality reduction on spectrograms versus computed features of syllables. Each plot shows 20 syllables of Cassin's vireo song. (A) UMAP projections of 18 features (see Supporting information) of syllables generated using BioSound. (B) UMAP applied to spectrograms of syllables. (E) UMAP of spectrograms where color is the syllable's average fundamental frequency (F) The same as (E) where pitch saliency of each syllable, which corresponds to the relative size of the first auto-correlation peak represents color.

To compare and visualize the structure captured by both PAF and spectrogram representations of vocalizations, we used a subset of the 20 most frequent syllable-types from a dataset of Cassin's vireo song recorded in the Sierra Nevada Mountains [165, 19]. We computed both spectrographic representations of syllables as well as a set of 18 temporal, spectral, and fundamental characteristics (Supporting information) over each syllable using the BioSound python package [110]. We then projected both the spectral representation as well as the PAFs into 2D UMAP feature spaces (Fig Comparison between dimensionality reduction on spectrograms versus computed features of syllables A,B). To quantify the difference in how well clustered the different data representations are, we compare the silhouette score (Eq. 2.4; [373]) of each representation. The silhouette score is a measure of how well a dataset is clustered relative to a set of known category labels (e.g. syllable label, species identity). The silhouette score is the mean silhouette coefficient across all of the samples in a dataset, where the silhouette coefficient measures how distant each point is to points in its own category, relative to its distance from the nearest point in another category. It is therefore taken as a measure of how well clustered together elements are that belong to the same category. Silhouette scores range from -1 to 1, with 1 being more clustered.

Overall, the UMAP projections significantly increase the clusterability of syllables in the Cassin's vireo dataset. The UMAP representations of both the PAF and the spectrogram data (Fig 2.3A,B) are more clustered than either PAFs or spectrograms alone. The silhouette score of PAFs (0.054) is significantly lower than that for the UMAP projections of PAFs (0.092; H(2) = 632; p ¡ 10^{-10}; Fig 2.3A), and the silhouette score of spectrograms (0.252) is significantly lower than that of the UMAP projections of spectrograms (0.772; H(2) = 37868; p ¡ 10^{-10}; Fig 2.3B). In addition, comparing between features, the UMAP projections of spectrograms yields more clearly discriminable clusters than UMAP projections of the PAFs (H(2) = 37868; p ¡ 10^{-10}). All the silhouette scores are significantly better than chance (for each, H(2) ¿ 500; p ¡ 10^{-10}; see methods). Thus, for this dataset, UMAP projections yield highly clusterable representations of the data points, and UMAP projections of spectrograms are more clustered than UMAP projections of

PAFs. One should not infer from this, however, that spectrographic representations necessarily capture more structure than PAFs in all cases. For zebra finch vocalizations, PAFs provide more information about vocalization types than spectrograms [109], and in other datasets, smaller basis sets of acoustic features can account for nearly all the dynamics of a vocal element (e.g. [70]). Even when spectrographic representations are more clearly clusterable than PAFs, knowing how explicit features of data (e.g. fundamental frequency) are related to variability can be more useful than being able to capture variability in the feature space without an intuitive understanding of what those features represent. These different representations may capture different components of the signals. To highlight this, we show how two PAFs (Mean Fundamental Frequency and Pitch Saliency) vary within spectrographic UMAP clusters (Fig 2.3C,D), by overlaying the color-coded PAFs onto the UMAP projections of the spectrographic representations from Fig 2.3B). The relationships between PAFs and UMAP spectrogram projections exemplifies the variability of different PAFs within clusters, as well as the non-linear relationships learned by UMAP projections. Additional PAFs overlaid on UMAP projections are shown in Supporting information.

2.2.3 Discrete latent projections of animal vocalizations

To explore the broad utility of latent models in capturing features of vocal repertoires, we analyzed nineteen datasets consisting of 400 hours of vocalizations and over 3,000,000 discrete vocal units from 29 unique species (Supporting information). Each vocalization dataset was temporally segmented into discrete units (e.g. syllables, notes), either based upon segmentation boundaries provided by the dataset (where available), or using a novel dynamic-thresholding segmentation algorithm that segments syllables of vocalizations between detected pauses in the vocal stream (See Segmentation). Each dataset was chosen because it contains large repertoires of vocalizations from relatively acoustically isolated individuals that can be cleanly separated into temporally-discrete vocal units. With each temporally discrete vocal unit we computed a spectrographic representation (Supporting information; See Spectrogramming). We then

projected the spectrograms into latent feature spaces using UMAP (Figs 2.4, 2.5, 2.7, 2.8). From these latent feature spaces, we analyzed datasets for classic vocal features of animal communication signals, speech features, stereotypy/clusterability, and sequential organization.

Vocal features

Latent non-linear projections often untangle complex features of data in human interpretable ways. For example, the latent spaces of some neural networks linearize the presence of a beard in an image of a face without being trained on beards in any explicit way [386, 358]. Complex features of vocalizations are similarly captured in intuitive ways in latent projections [381, 150, 382, 430]. Depending on the organization of the dataset projected into a latent space, these features can extend over biologically or psychologically relevant scales. Accordingly, we used our latent models to look at spectro-temporal structure within the vocal repertoires of individual's, and across individuals, populations, and phylogeny. These latent projections capture a range of complex features, including individual identity (Fig 2.4), species identity (Fig 2.5A,B), linguistic features (Fig 2.7, Supporting information), syllabic categories (Figs 2.12, 2.10, 2.8, 2.13), and geographical variability (Fig 2.5C). We discuss each of these complex features in more detail below.

Individual identity

Many species produce caller-specific vocalizations that facilitate the identification of individuals when other sensory cues, such as sight, are not available. The features of vocalizations facilitating individual identification vary between species. We projected identity call datasets (i.e., sets of calls thought to carry individual identity information) from four different species into UMAP latent spaces (one per species) to observe whether individual identity falls out naturally within the latent space.

We looked at four datasets where both caller and call-type are available. Caller identity is evident in latent projections of all four datasets (Fig 2.4). The first dataset is comprised of macaque coo calls, where identity information is thought to be distributed across multiple

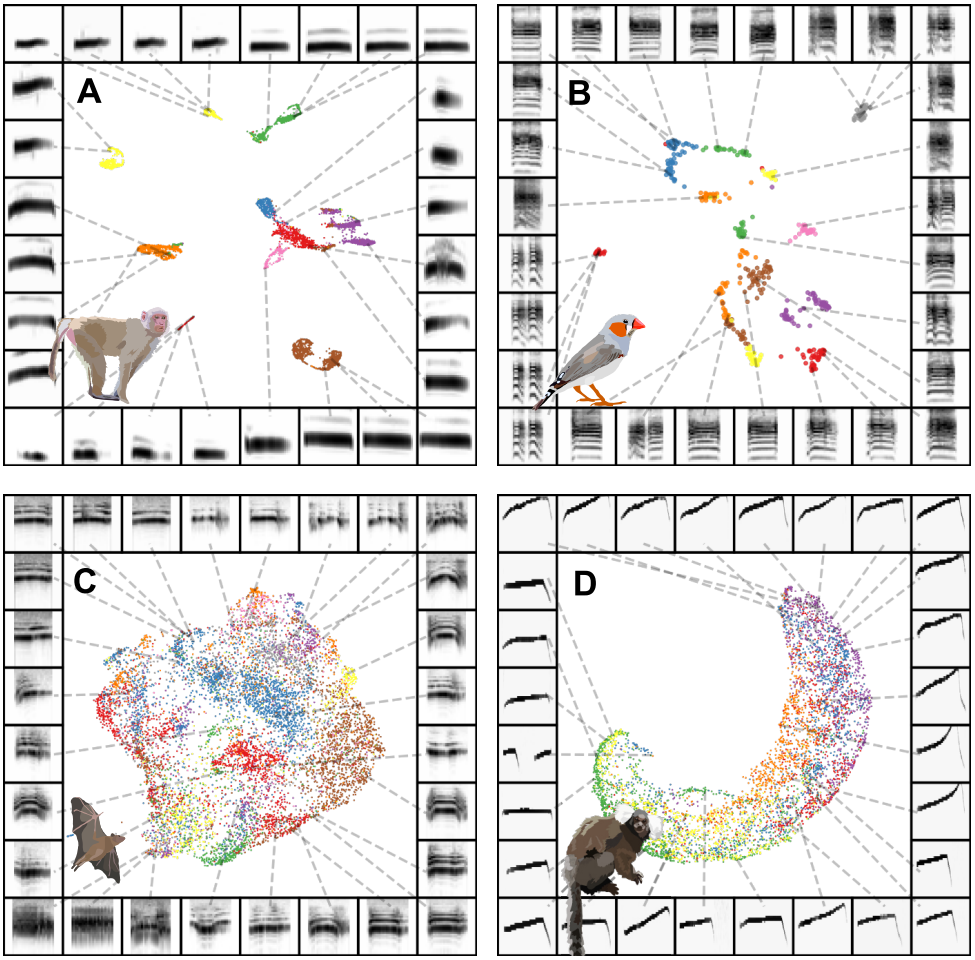


Figure 2.4. Individual identity is captured in projections for some datasets. Each plot shows vocal elements discretized, spectrogrammed, and then embedded into a 2D UMAP space, where each point in the scatterplot represents a single element (e.g. syllable of birdsong). Scatterplots are colored by individual identity. The borders around each plot are example spectrograms pointing toward different regions of the scatterplot. (A) Rhesus macaque coo calls. (B) Zebra finch distance calls. (C) Fruit bat infant isolation calls. (D) Marmoset phee calls.

features including fundamental frequency, duration, and Weiner entropy [130]. Indeed, the latent projection of coo calls clustered tightly by individual identity (silhouette score = 0.378; Fig 2.4A). The same is true for zebra finch distance calls [109] (silhouette score = 0.615; Fig 2.4B). Egyptian fruit bat pup isolation calls, which in other bat species are discriminable by adult females [36, 113, 36] clearly show regions of UMAP space densely occupied by single individual's vocalizations, but no clear clusters (silhouette score = -0.078; Fig 2.4C). In the marmoset phee call dataset [288] it is perhaps interesting that given the range of potential features thought to carry individual identity [130], phee calls appear to lie along a single continuum where each individual's calls occupy overlapping regions of the continuum (silhouette score = -0.062; Fig 2.4D). The silhouette score for each species was well above chance ($H(2) ¿ 20$, $p ¡ 10^{-5}$). These patterns predict that some calls, such as macaque *coo* calls, would be more easily discriminable by conspecifics than other calls, such as marmoset *phee* calls.

The latent projections of these datasets demonstrate that individual identity can be obtained from all these vocalizations. Importantly, this information is available without *a priori* knowledge of specific spectro-temporal features, which is likely also the case for the animals attempting to use it. Because no caller identity information is used in learning the latent projections, the emergence of this information indicates that the similarity of within-caller vocalizations contains enough statistical power to overcome variability between callers. This within-caller structure likely facilitates conspecific learning of individual identity without *a priori* expectations for the distribution of relevant features [26], in the same way that developing sensory systems adapt to natural environmental statistics [33].

Cross species comparisons

Classical comparative studies of vocalizations across species rely on experience with multiple species' vocal repertoires. This constrains comparisons to those species whose vocalizations are understood in similar feature spaces, or forces the choice of common feature spaces that may obscure relevant variation differently in different species. Because latent models learn

arbitrary complex features of datasets, they can yield less biased comparisons between vocal repertoires where the relevant axes are unknown, and where the surface structures are either very different, for example canary and starling song, or very similar, like the echolocation clicks of two closely related beaked whales.

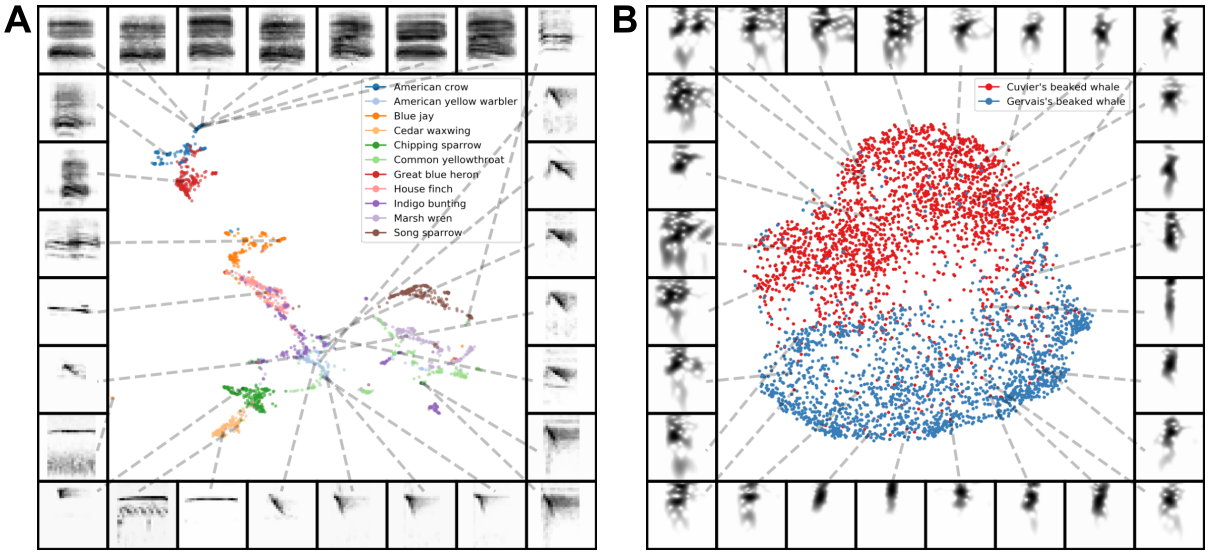


Figure 2.5. Comparing species with latent projections. (A) Calls from eleven species of North American birds are projected into the same UMAP latent space. (B) Cuvier's and Gervais's beaked whale echolocation clicks are projected into UMAP latent space and fall into two discrete clusters.

To explore how well latent projections capture vocal repertoire variation across species, we projected a dataset containing monosyllabic vocalizations [470] from eleven different species of North American birds into UMAP latent space (silhouette score = 0.377), well above chance ($H(2) = 1396$, p ¡ 10^{-10}). Similar "calls", like those from the American crow *caw* and great blue heron *roh* are closer together in latent space, while more distinct vocalizations, like chipping sparrow notes, are further apart (Fig 2.5A). Latent projections like this have the potential power to enable comparisons across broad phylogenies without requiring decisions about which acoustic features to compare.

At the other extreme is the common challenge in bioacoustics research to differentiate between species with very similar vocal repertoires. For example, Cuvier's and Gervais' beaked

whales, two sympatric species recorded in the Gulf of Mexico, have echolocation clicks with highly overlapping power spectra that are generally differentiated using supervised learning approaches (c.f. [172, 127]). We projected a dataset containing Cuvier's and Gervais' beaked whale echolocation clicks into UMAP latent space. Species-identity again falls out nicely, with clicks assorting into distinct clusters that correspond to species (Fig 2.5B). The silhouette score of UMAP on the spectrogram (shown in Fig 2.5B) was 0.401, higher than the silhouette score of UMAP on the power spectra (0.171; H(2) = 2411; p ¡ 10^{-10}) which is in turn higher than the silhouette score of the power spectra alone (0.066; H(2) = 769; p ¡ 10^{-10}). Each silhouette score is also well above chance (H(2) ¿ 500; p ¡ 10^{-10}). The utility of an approach such as UMAP to clustering echolocation clicks is perhaps unsurprising; recent work [127] has shown that graph-based methods are successful for representing and clustering echolocation clicks of a larger dataset of cetacean echolocation clicks.

Population geography

Some vocal learning species produce different vocal repertoires (regiolects) across populations occupying different geographic regions. Differences in regiolects between populations are borne out in the categorical perception of notes [233, 315, 350], much the same as cross-linguistic differences in the categorical perception of phonemes in human speech [181]. To compare vocalizations across geographical populations in the swamp sparrow, which produces regionally distinct trill-like songs [234], we projected individual notes into a UMAP latent space. Although the macro-structure of clusters suggest common note-types across multiple populations, most of the larger clusters show multiple clear sub-regions that are tied to vocal differences between geographical populations (Fig 2.6). We further explore how these projections of notes relate to vocal clusters in traditional feature spaces later in the manuscript.

Phonological features

The sound segments that make up spoken human language can be described by distinctive phonological features that are grouped according to articulation place and manner, glottal state,

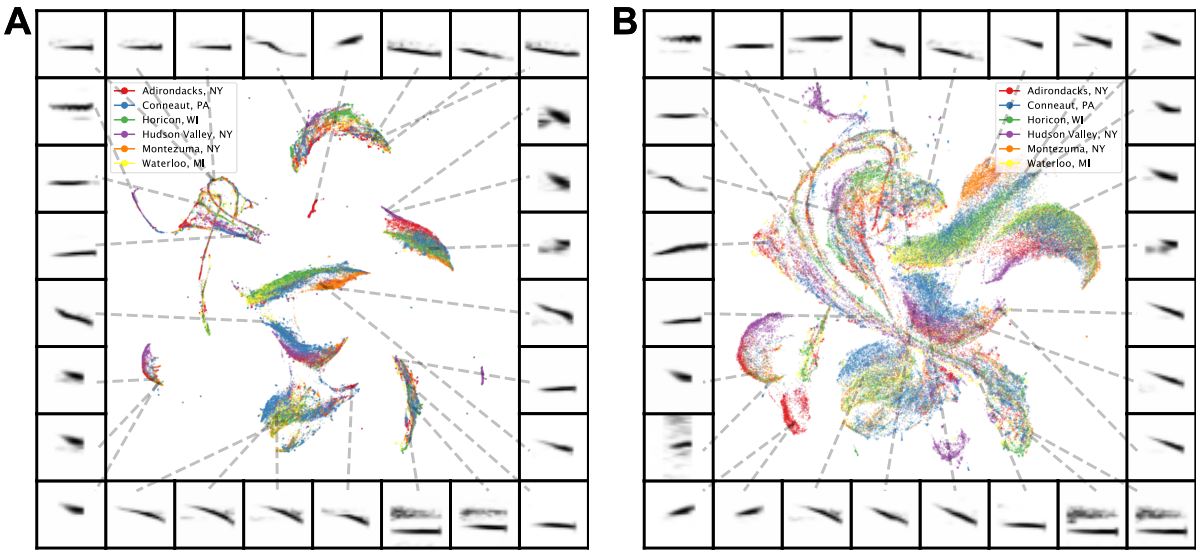


Figure 2.6. Comparing notes of swamp sparrow song across different geographic populations. (A) Notes of swamp sparrow song from six different geographical populations projected into a 2D UMAP feature space. (B) The same dataset from (A) projected into a 2D UMAP feature space where the parameter `min_dist` is set at 0.25 to visualize more spread in the projections.

and vowel space. A natural way to look more closely at variation in phoneme production is to look at variation between phonemes that comprise the same phonological features. As an example, we projected sets of consonants that shared individual phonological features into UMAP latent space (Figs 2.7, Supporting information). In most cases, individual phonemes tended to project to distinct regions of latent space based upon phonetic category, and consistent with their perceptual categorization. At the same time, we note that latent projections vary smoothly from one category to the next, rather than falling into discrete clusters. This provides a framework that could be used in future work to characterize the distributional properties of speech sounds in an unbiased manner. Likewise, it would be interesting to contrast projections of phonemes from multiple languages, in a similar manner as the swamp sparrow (Fig 2.6), to visualize and characterize variation in phonetic categories across languages [181].

Variation in discrete distributions and stereotypy

In species as phylogenetically diverse as songbirds and rock hyraxes, analyzing the sequential organization of communication relies upon similar methods of segmentation and

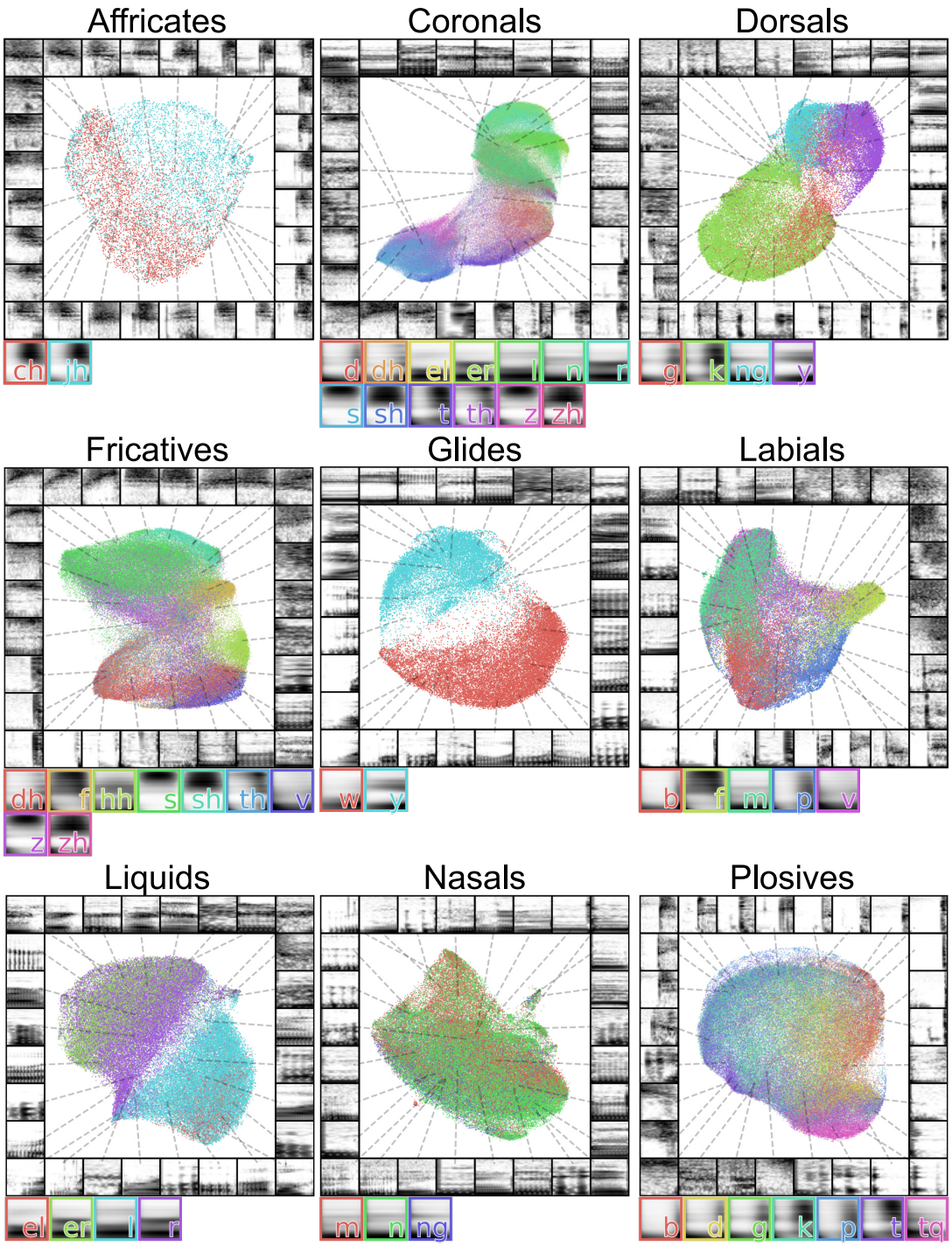


Figure 2.7. Latent projections of consonants. Each plot shows a different set of consonants grouped by phonetic features. The average spectrogram for each consonant is shown to the right of each plot.

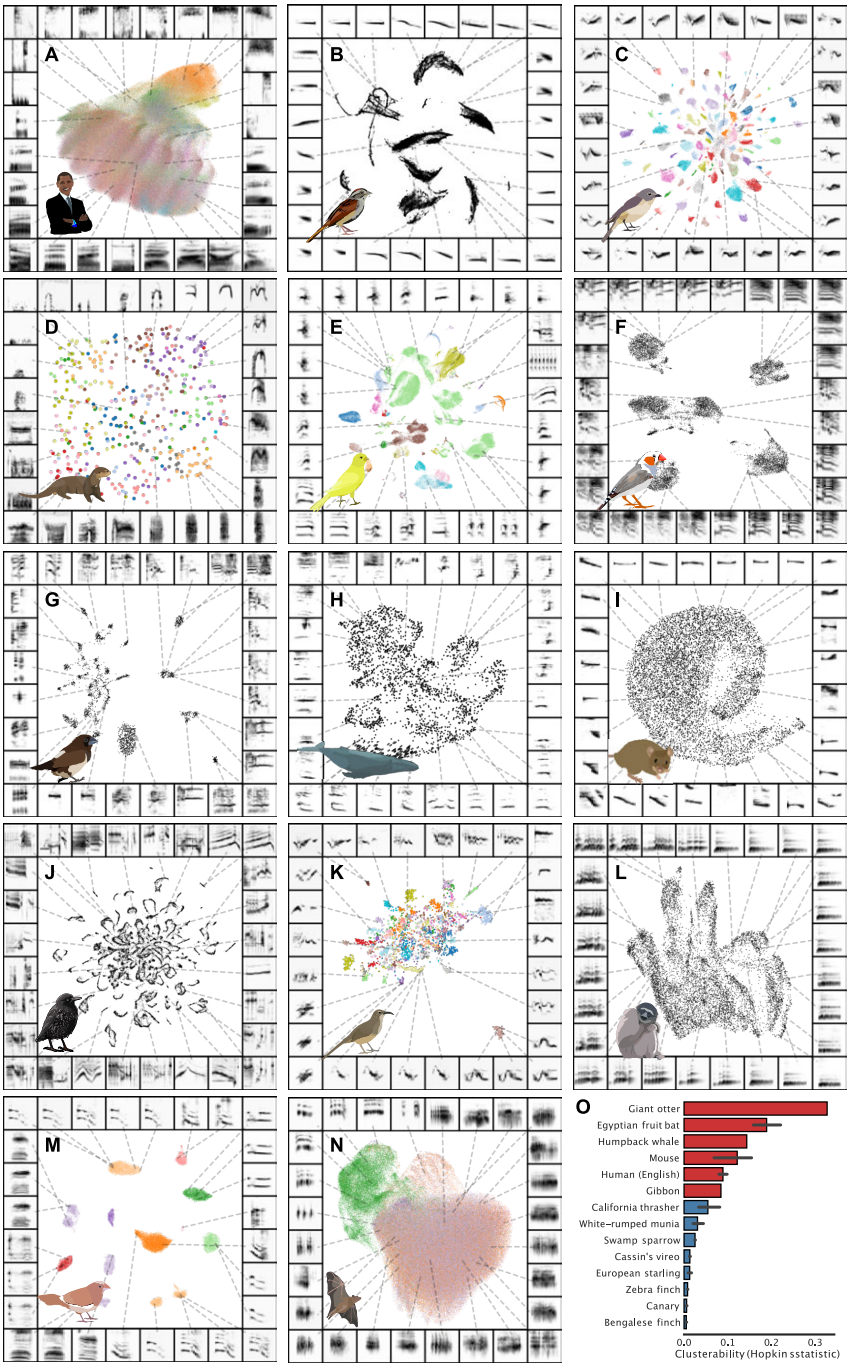


Figure 2.8. UMAP projections of vocal repertoires across diverse species.

categorization of discrete vocal elements [200]. In species such as the Bengalese finch, where syllables are highly stereotyped, clustering syllables into discrete categories is a natural way to abstract song. The utility of clustering song elements in other species, however, is more contentious because discrete category boundaries are not as easily discerned [437, 419, 150, 171].

To compare broad structural characteristics across a wide sampling of species, we projected vocalizations from 14 datasets of different species vocalizations, ranging across songbirds, cetaceans, primates, and rodents into UMAP space (Fig 2.8). To do so, we sampled from a diverse range of datasets, each of which was recorded from a different species in a different setting (Supporting information). Some datasets were recorded from single isolated individuals in a sound isolated chamber in a laboratory setting, while others were recorded from large numbers of freely behaving individuals in the wild. In addition, the units of vocalization from each dataset are variable. We used the smallest units of each vocalization that could be easily segmented, for example, syllables, notes, and phonemes. Thus, this comparison across species is not well-controlled. Still, such a dataset enabling a broad comparison in a well-controlled manner does not exist. Latent projections of such diverse recordings, while limited in a number of ways, have the potential to provide a glimpse into broad structure into vocal repertoires, yielding novel insights into broad trends in animal communication. For each dataset, we computed spectrograms of isolated elements, and projected those spectrograms into UMAP space (Fig 2.8). Where putative element labels are available, we plot them in color over each dataset.

Visually inspecting the latent projections of vocalizations reveals appreciable variability in how the repertoires of different species cluster in latent space. For example, mouse USVs appear as a single cluster (Fig 2.8I), while zebra finch syllables appear as multiple discrete clusters (Fig 2.8M,F), and gibbon song sits somewhere in between (Fig 2.8L). This suggests that the spectro-temporal acoustic diversity of vocal repertoires fall along a continuum ranging from unclustered and uni-modal to highly clustered.

We quantified this effect using a linear mixed-effects model comparing the Hopkin's

statistic across UMAP projections of vocalizations from single individuals (n = 289), controlling for the number of vocalizations produced by each individual as well as random variability in clusterability at the level of species. We included each of the species in Fig 2.8 except giant otter and gibbon vocalizations, as individual identity was not available for those datasets. We find that songbird vocalizations are significantly more clustered than mammalian vocalizations ($\chi^2(1) = 20$, p ¡ 10^{-5}; See Methods).

The stereotypy of songbird (and other avian) vocal elements is well documented [460, 408] and at least in zebra finches is related to the high temporal precision in the singing-related neural activity of vocal-motor brain regions [161, 117, 65]. The observed differences in stereotypy between songbirds and mammals should be interpreted with consideration of the broad variability underlying the datasets, however.

Clustering vocal element categories

UMAP projections of birdsongs largely fall more neatly into discriminable clusters (Fig 2.8). If clusters in latent space are highly similar to experimenter-labeled element categories, unsupervised latent clustering could provide an automated and less time-intensive alternative to hand-labeling elements of vocalizations. To examine this, we compared how well clusters in latent space correspond to experimenter-labeled categories in three human-labeled datasets: two separate Bengalese finch datasets [322, 219], and one Cassin's vireo dataset [165]. We compared four different labeling techniques: a hierarchical density-based clustering algorithm (HDBSCAN; [58, 278]) applied to UMAP projections of spectrograms, HDBSCAN applied to PCA projections of spectrograms[1], k-means [338] clustering applied over UMAP, and k-means clustering applied over spectrograms (Fig 2.10; Table 2.1).

Like the contrast between MDS and UMAP, the k-means clustering algorithm works directly on the Euclidean distances between data points, whereas HDBSCAN operates on a graph-based transform of the input data (Fig 2.9). Briefly, HDBSCAN first defines a 'mutual

[1]HDBSCAN is applied to 100-dimensional PCA projections rather than spectrograms directly because HDBSCAN does not perform well in high-dimensional spaces [278].

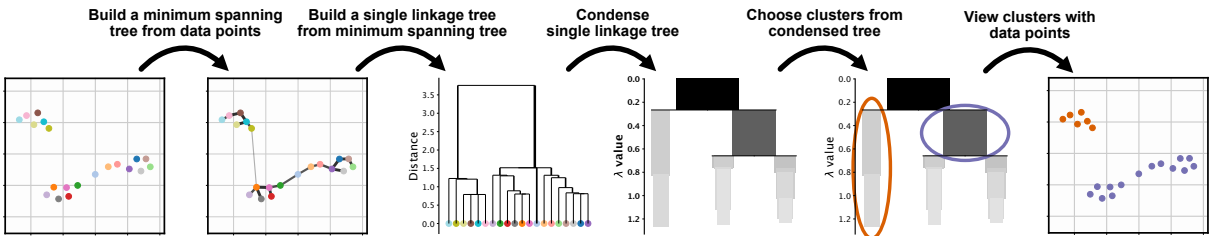


Figure 2.9. HDBSCAN density-based clustering. Clusters are found by generating a graphical representation of data, and then clustering on the graph. The data shown in this figure are from the latent projections from Fig 2.1. Notably, the three clusters in Fig 1. are clustered into only two clusters using HDBSCAN, exhibiting a potential shortcoming of the HDBSCAN algorithm. The grey colormap in the condensed trees represent the number of points in the branch of the tree. Λ is a value used to compute the persistence of clusters in the condensed trees.

reachability' distance between elements, a measure of the distance between points in the dataset weighted by the local sparsity/density of each point (measured as the distance to a *k*th nearest neighbor). HDBSCAN then builds a graph, where each edge between vertices (points in the dataset) is the mutual reachability between those points, and then prunes the edges to construct a minimum spanning tree (a graph containing the minimum set of edges needed to connect all of the vertices). The minimum spanning tree is converted into a hierarchy of clusters of points sorted by mutual reachability distance, and then condensed iteratively into a smaller hierarchy of putative clusters. Finally, clusters are chosen as those that persist and are stable over the greatest range in the hierarchy.

To make the k-means algorithm more competitive with HDBSCAN, we set the number of clusters in k-means equal to the number of clusters in the hand-clustered dataset, while HDBSCAN was not parameterized at all. We computed the similarity between hand and algorithmically labeled datasets using three related metrics, homogeneity, completeness, and V-measure ([372]; see Methods section). Homogeneity measures the extent to which algorithmic clusters fall into the same hand-labeled syllable category while completeness measures the extent to which hand-labeled categories belong to the same algorithmic cluster. V-measure is the harmonic mean between the homogeneity and completeness, which is equal to the mutual information between the algorithmic clusters and the hand-labels, normalized by the mean of

their marginal entropy [372].

Table 2.1. Cluster similarity to hand labels for two Bengalese finch and one Cassin's vireo dataset. Four clustering methods were used: (1) KMeans on spectrograms (2) KMeans on UMAP projections (3) HDBSCAN on first 100 principal components of spectrograms (4) HDBSCAN clustering of UMAP projections. With KMeans 'K' was set to the correct number of clusters to make it more competitive with HDBSCAN clustering. Standard deviation across individual birds is shown for the finch datasets. Best performing method for each metric is bolded.

	Homogeneity	Completeness	V-measure
B. Finch (Koumura)			
KMeans	0.911±0.044	0.85±0.064	0.879±0.051
KMeans/UMAP	0.842±0.116	0.796±0.145	0.817±0.132
HDBSCAN/PCA	0.968±0.036	**0.86±0.14**	**0.902±0.086**
HDBSCAN/UMAP	**0.99±0.006**	0.74±0.122	0.841±0.088
B. Finch (Nicholson)			
KMeans	0.954±0.024	0.707±0.101	0.809±0.074
KMeans/UMAP	**0.967±0.018**	0.688±0.098	0.801±0.072
HDBSCAN/PCA	0.901±0.067	0.837±0.027	0.866±0.034
HDBSCAN/UMAP	0.963±0.022	**0.855±0.076**	**0.903±0.042**
Cassin's vireo			
KMeans	0.894	0.808	0.849
KMeans/UMAP	0.928	0.829	0.875
HDBSCAN/PCA	0.849	0.906	0.877
HDBSCAN/UMAP	**0.936**	**0.94**	**0.938**

For all three datasets, the HDBSCAN clusters most closely match those of humans as is indicated by the V-measure (Table 2.1). In both the Nicholson [322] Bengalese finch dataset and the Cassin's vireo dataset, the closest match to human clustering is achieved by HDBSCAN on the UMAP projections. In the Koumura dataset [219], HDBSCAN on the PCA projections gives the closest match to human clustering, where homogeneity is higher with HDBSCAN/UMAP and completeness is higher with HDBSCAN/PCA. A high homogeneity and low completeness score indicates that algorithmic clusters tend to fall into the same hand-labeled category, but multiple sub-clusters are found within each hand labeled category. As we show in Abstracting and visualizing sequential organization, this difference between algorithmically found labels often reflects real structure in the dataset that human labeling ignores. More broadly, our clustering results show that latent projections facilitate unsupervised clustering of vocal

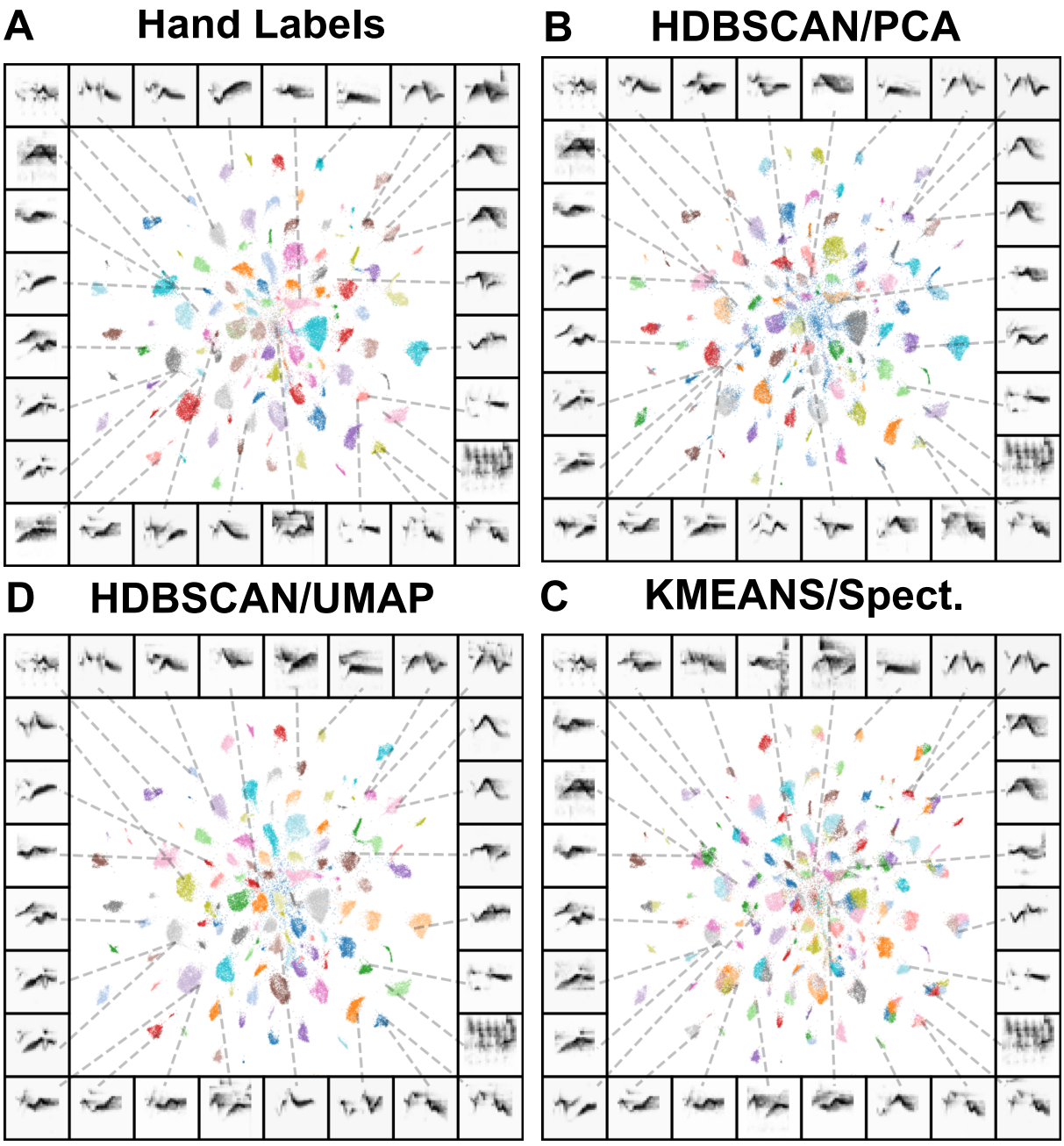


Figure 2.10. Clustered UMAP projections of Cassin's vireo syllable spectrograms. Panels (A-D) show the same scatterplot, where each point corresponds to a single syllable spectrogram projected into two UMAP dimensions. Points are colored by their hand-labeled categories (A), which generally fall into discrete clusters in UMAP space. Remaining panels show the same data colored according to cluster labels produced by (B) HDBSCAN over PCA projections (100 dimensions), (C) HDBSCAN on UMAP projections, and (D) k-means directly on syllable spectrograms.

elements into human-like syllable categories better than spectrographic representations alone. At the same time, unsupervised latent clustering is not strictly equivalent to hand labeling, and the two methods may yield different results.

Comparing latent features and clusters to known feature spaces

When the features underlying behaviorally relevant vocal variability in a species are known *a priori*, latent feature spaces learned directly from the data may be unnecessary to infer the underlying structure of a vocal repertoire. Although sets of behaviorally relevant features are not known for most species, Swamp sparrows are an exception, as their vocalizations have a relatively long history of careful characterization [267, 233]. Swamp sparrows produce songs that are hierarchically organized into syllables made up of shorter notes, which in turn can be well-described by only a few simple features. This set of known *a priori* features provides a useful comparison for the latent features learned by UMAP.

We compared the features learned by UMAP with the known feature-space of swamp sparrow notes using a dataset of songs recorded in the wild. In Fig 2.11 we show UMAP and known-feature spaces for notes from a population of swamp sparrows recorded in Conneaut Marsh, Pennsylvania. We compare the spectrogram of each note projected into UMAP space to the same note projected onto three features known to describe much of the behaviorally relevant variance in swamp sparrow song [233, 267]: peak frequency at the beginning and ending of the note (Fig 2.11A), note length (Fig 2.11B), and the overall change in peak frequency (Fig 2.11B). We then clustered the UMAP projections (Fig 2.11C) using HDBSCAN and the known feature space using a Gaussian Mixture Model (GMM; see Clustering vocalizations). For comparison, we also visualize the known features projected into UMAP (Fig 2.11D).

HDBSCAN found 12 unique clusters, as opposed to the normal 6-10 note categories typically used to define swamp sparrow song [233]. The GMM was set to find 10 clusters, as was used in the same dataset in prior work [233]. Between the GMM and HDBSCAN clustering, we find a degree of overlap well above chance (homogeneity = 0.633; completeness = 0.715,

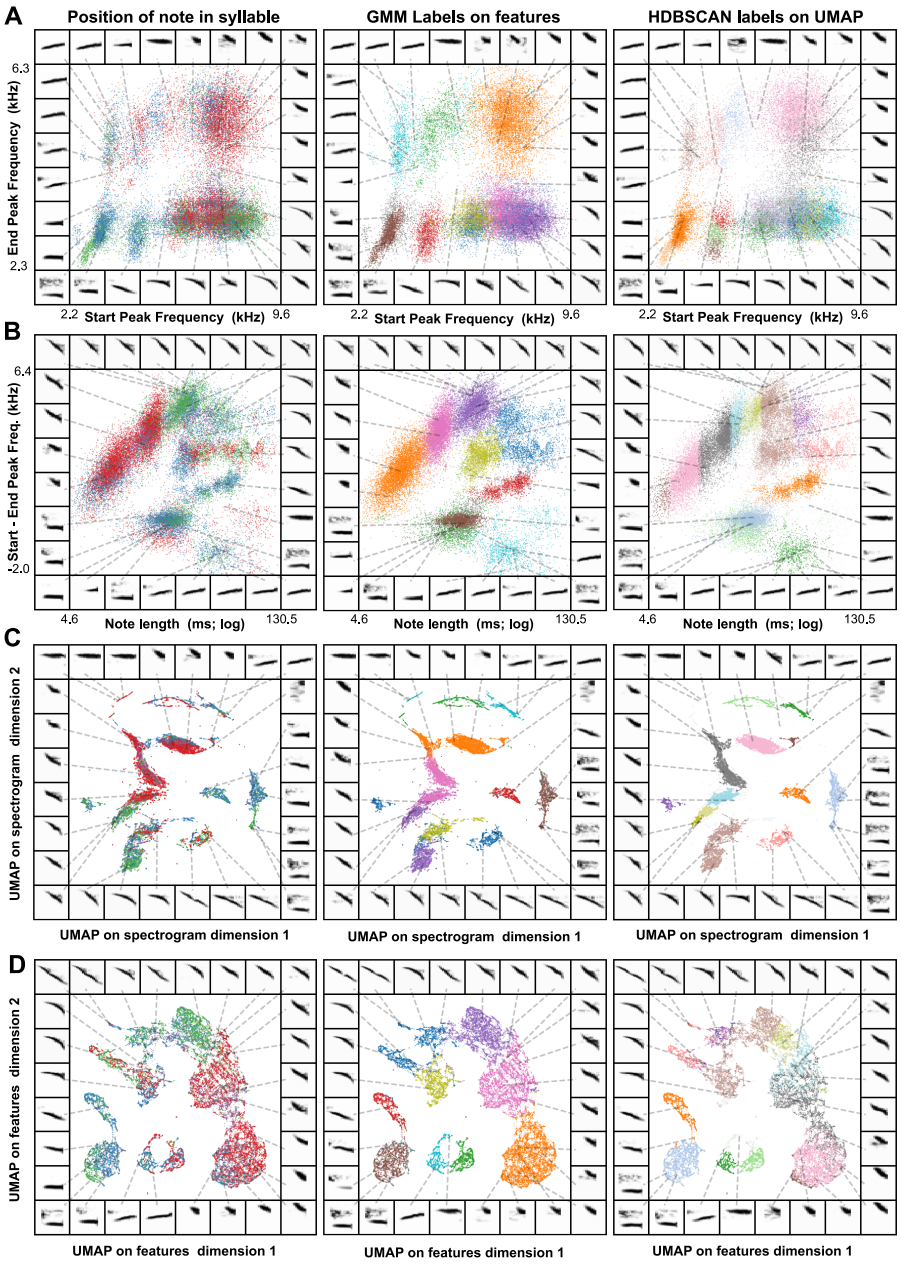


Figure 2.11. Comparing latent and known features in swamp sparrow song. (A) A scatterplot of the start and end peak frequencies of the notes produced by birds recorded in Conneaut Marsh, PA. The left panel shows notes colored by the position of each note in the syllable (red = first, blue = second, green = third). The center panel shows the sample scatterplot colored by a Gaussian Mixture Model labels (fit to the start and end peak frequencies and the note duration). The right panel shows the scatterplot colored by HDBSCAN labels over a UMAP projection of the spectrograms of notes. (B) The same notes, plotting the change in peak frequency over the note against the note's duration. (C) The same notes plotted as a UMAP projection over note-spectrograms. (D) The features from (A) and (B) projected together into a 2D UMAP space.

V-measure = 0.672; chance V-measure = 0.001; bootstrapped p ¡ 10^{-4}; Fig 2.11A,B). Using the position of the note within each syllable as a common reference (most syllables were comprised of 3 or fewer notes), we compared the overlap between the two clustering methods. Both labeling schemes were similarly related to the position of notes within a syllable (e.g. first, second, third; v-measure GMM = 0.162; V-measure HDBSCAN = 0.144), and both were well above chance (bootstrapped p ¡ 10^{-4}). We repeated the same analysis on a second population of swamp sparrow recorded in Hudson Valley, NY (Supporting information), and found a similar overlap between the two clustering schemes (homogeneity = 0.643; completeness = 0.815, V-measure = 0.719; chance V-measure = 0.002; bootstrapped p ¡ 10^{-4}) and a similar level of overlap with the position of notes (V-measure GMM = 0.133; V-measure HDBSCAN = 0.144).

Given this pattern of results, it is unlikely that one would want to substitute the unsupervised latent features for the known features when trying to describe swamp sparrow song in the most efficient low-dimensional space. Still, both feature sets yield surprisingly similar compressed representations. Thus, in the absence of known features, the unsupervised methods can provide either (1) a useful starting point for more refined analyses to discover ”known” features, or (2) a functional analysis space that likely captures much (but not all) of the behaviorally relevant signal variation.

Abstracting and visualizing sequential organization

As acoustic signals, animal vocalizations have an inherent temporal structure that can extend across time scales from short easily discretized elements such as notes, to longer duration syllables, phrases, songs, bouts, etc. The latent projection methods described above can be used to abstract corpora of song elements well-suited to temporal pattern analyses [381], and to make more direct measures of continuous vocalization time series. Moreover, their automaticity enables the high throughput necessary to satisfy intensive data requirements for most quantitative sequence models.

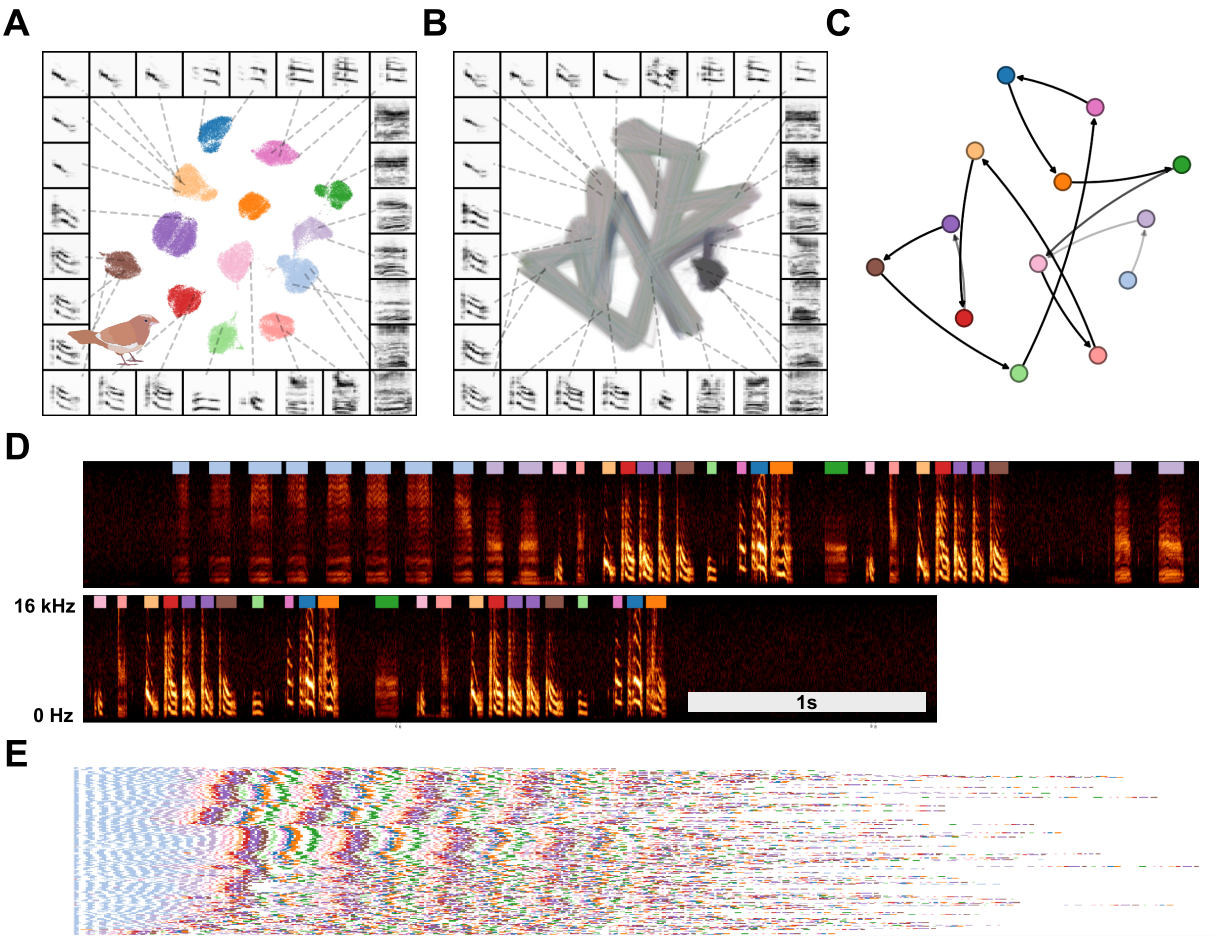


Figure 2.12. Latent visualizations of Bengalese finch song sequences. (A) Syllables of Bengalese finch songs from one individual are projected into 2D UMAP latent space and clustered using HDBSCAN. (B) Transitions between elements of song are visualized as line segments, where the color of the line segment represents its position within a bout. (C) The syllable categories and transitions in (A) and (B) can be abstracted to transition probabilities between syllable categories, as in a Markov model. (D) An example vocalization from the same individual, with syllable clusters from (A) shown above each syllable. (E) A series of song bouts. Each row is one bout, showing overlapping structure in syllable sequences. Bouts are sorted by similarity to help show structure in song.

In practice, modeling sequential organization can be applied to any discrete dataset of vocal elements, whether labeled by hand or algorithmically. Latent projections of vocal elements have the added benefit of allowing visualization of the sequential organization that can be compared to abstracted models. As an example of this, we derived a corpus of symbolically segmented vocalizations from a dataset of Bengalese finch song using latent projections and

clustering (Fig 2.12). Bengalese finch song bouts comprise a small number (˜5-15) of highly stereotyped syllables produced in well-defined temporal sequences a few dozen syllables long [194]. We first projected syllables from a single Bengalese finch into UMAP latent space, then visualized transitions between vocal elements in latent space as line segments between points (Fig 2.12B), revealing highly regular patterns. To abstract this organization to a grammatical model, we clustered latent projections into discrete categories using HDBSCAN. Each bout is then treated as a sequence of symbolically labeled syllables (e.g. $B \rightarrow B \rightarrow C \rightarrow A$; Fig 2.12D) and the entire dataset rendered as a corpus of transcribed song (Fig 2.12E). Using the transcribed corpus, one can abstract statistical and grammatical models of song, such as the Markov model shown in Fig 2.12C or the information-theoretic analysis in Sainburg et al., [381].

Sequential organization is tied to labeling method

As noted previously, hand labels and latent cluster labels of birdsong syllables generally overlap (e.g. Fig 2.10), but may disagree for a sizable minority of syllables (Table 2.1). Similarly, in mice, different algorithmic methods for abstracting and transcribing mouse vocal units (USVs) can result in substantial differences between syntactic descriptions of sequential organization [171]. We were interested in the differences between the abstracted sequential organization of birdsong when syllables were labeled by hand versus clustered in latent space. Because we have Bengalese finch datasets that are hand transcribed from two different research groups [322, 220], these datasets are ideal for comparing the sequential structure of algorithmic versus hand-transcribed song.

To contrast the two labeling methods, we first took the two Bengalese finch song datasets, projected syllables into UMAP latent space, and visualized them using the hand transcriptions provided by the datasets (Fig 2.13A,H). We then took the syllable projections and clustered them using HDBSCAN. In both datasets, we find that many individual hand-transcribed syllable categories are comprised of multiple HDBSCAN-labelled clusters in latent space (Fig 2.13A,B,H,I). To compare the different sequential abstractions of the algorithmically transcribed labels and

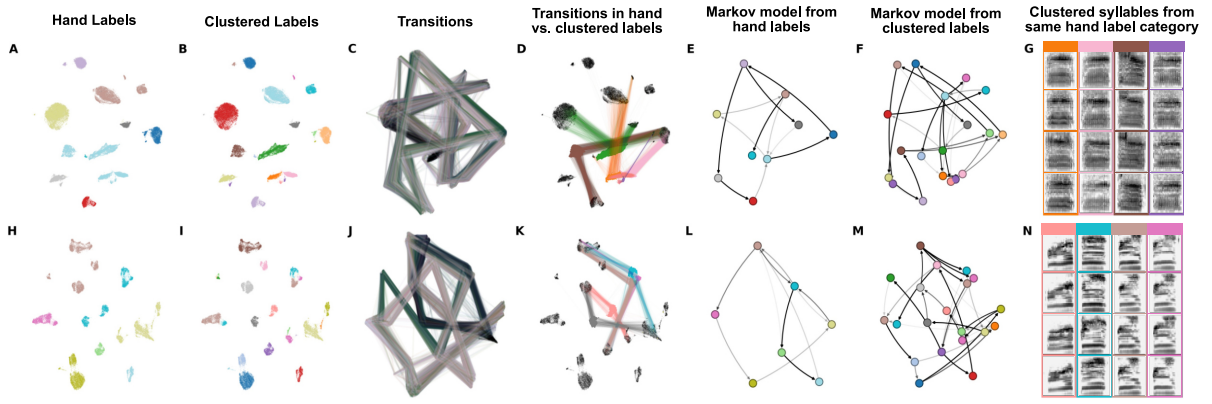


Figure 2.13. Latent comparisons of hand- and algorithmically-clustered Bengalese finch song. A-G are from a dataset produced by Nicholson et al., [9] and H-N are from a dataset produced by Koumura et al., [10] (A,H) UMAP projections of syllables of Bengalese finch song, colored by hand labels. (B,I) Algorithmic labels (UMAP/HDBSCAN). (C, J) Transitions between syllables, where color represents time within a bout of song. (D,K) Comparing the transitions between elements from a single hand-labeled category that comprises multiple algorithmically labeled clusters. Each algorithmically labeled cluster and the corresponding incoming and outgoing transitions are colored. Transitions to different regions of the UMAP projections demonstrate that the algorithmic clustering method finds clusters with different syntactic roles within hand-labeled categories. (E,L) Markov model from hand labels colored the same as in (A,H) (F,M) Markov model from clustered labels, colored the same as in (B,I). (G,H) Examples of syllables from multiple algorithmic clusters falling under a single hand-labeled cluster. Colored bounding boxes around each syllable denotes the color category from (D,K).

the hand transcribed labels, we visualized the transitions between syllables in latent space (Fig 2.13C,J). These visualizations reveal that different algorithmically-transcribed clusters belonging to the same hand-transcribed label often transition to and from separate clusters in latent space. That is, the sub-category acoustics of the elements predict and are predicted by specific transitions. We visualize this effect more explicitly in Fig 2.13D and K, showing the first-order (incoming and outgoing) transitions between one hand-labeled syllable category (from Fig 2.13A and H), colored by the multiple HDBSCAN clusters that it comprises (from Fig 2.13B and I). Thus, different HDBSCAN labels that belong to the same hand-labeled category can play a different role in song-syntax, having different incoming and outgoing transitions. In Fig 2.13E,F,L,M, this complexity plays out in an abstracted Markov model, where the HDBSCAN-derived model reflects the latent transitions observed in Fig 2.13C,J more explicitly than the model abstracted

from hand-labeled syllables. To further understand why these clusters are labeled as the same category by hand but different categories using HDBSCAN clustering, we show example syllables from each cluster Fig 2.13G,N. Although syllables from different HDBSCAN clusters look very similar, they are differentiated by subtle yet systematic variation. Conversely, different subsets of the same experimenter-labeled category can play different syntactic roles in song sequences. The syntactic organization in Bengalese finch song is often described using Partially Observable Markov Models (POMMs) or Hidden Markov Models (HMMs), where the same syllable category plays different syntactic roles dependent on its current position in song syntax [194]. In so far as the sequential organization abstracted from hand labels obscures some of the sequential structure captured by algorithmic transcriptions, our results suggest that these different syntactic roles may be explained by the presence of different syllable categories.

To compare the difference in sequential organization captured by hand labels versus HDBSCAN labels quantitatively, we treated both HDBSCAN and hand labels as hidden states in separate HMMs and compared their ability to accurately model song sequences. An HMM is a finite-state model for a sequence of visible states (e.g song syllables), that is assumed to emerge from set of unobserved ('hidden') states, inferred algorithmically. To make our HMMs directly comparable, we use the hand labels as visible states, and infer hidden states from either the hand labels (e.g. Fig 2.14A) or the HDBSCAN labels (e.g. Fig 2.14B). By design, the hidden states of these two HMMs are explicitly constrained to either the hand or HDBSCAN labels, and thus ignore higher-order transitions that might carry useful sequence information. For comparison, we also trained an HMM where hidden states were inferred using the Baum-Welch algorithm and allowed to incorporate higher-order syllable sequences (e.g. Fig 2.14C; see Methods). For example, in the sequence of visible states $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e$, there might be a hidden state representing $d|a,b,c$. HMMs allowing high-order latent representations have been used to model sequential organization in birdsong [195] and have a long history of modeling human speech.

We compared each model on its ability to predict the sequence of hand labels using the Akaike Information Criterion (AIC), which normalizes model likelihood by the number

of parameters in the model [147]. Because models are compared on their ability to predict hand-labeled sequences, our comparison is biased toward sequential models based upon the hand-labels. Nonetheless, in 13 of 15 birds, the HDBSCAN clustered latent states better captured sequential dynamics (ΔAIC ¿ 2.0; Fig 2.14D). As expected, the Baum-Welch trained HMM is better able to explain the sequential organization in Bengalese finch song than either HMM constrained to use the hand or HDBSCAN labels in each bird (ΔAIC ¿ 2.0; Fig 2.14D). This indicates that second-order (or higher) transitions also contribute to the sequential structure of song in Bengalese finches. In Fig 2.14C, we overlay the hidden states learned by the complete HMM on the UMAP syllable projections of a single Bengalese finch from the Koumura dataset (an example bird from the Nicholson dataset is shown in Supporting information). This reveals several clusters with clear, uniformly colored subregions, indicating HMM hidden states that are not captured by the hand labels or HDBSCAN but still reflect non-random acoustic differences (Fig 2.14A,B).

2.2.4 Temporally continuous latent trajectories

Not all vocal repertoires are made up of elements that fall into highly discrete clusters in latent space (Fig 2.8). For several of the datasets we analysed, categorically discrete elements are not readily apparent, making analyses such as the cluster-based analyses performed in Figure 2.12 more difficult. In addition, many vocalizations are difficult to segment temporally, and determining what features to use for segmentation requires careful consideration [200]. In many bird songs, for example, clear pauses exist between song elements that enable one to distinguish syllables. In other vocalizations, however, experimenters must rely on less well-defined physical features for segmentation [182, 200], which may in turn invoke a range of biases and unwarranted assumptions. At the same time, much of the research on animal vocal production, perception, and sequential organization relies on identifying ”units” of a vocal repertoire [200]. To better understand the effects of temporal discretization and categorical segmentation in our analyses, we considered vocalizations as continuous trajectories in latent space and compared the resulting

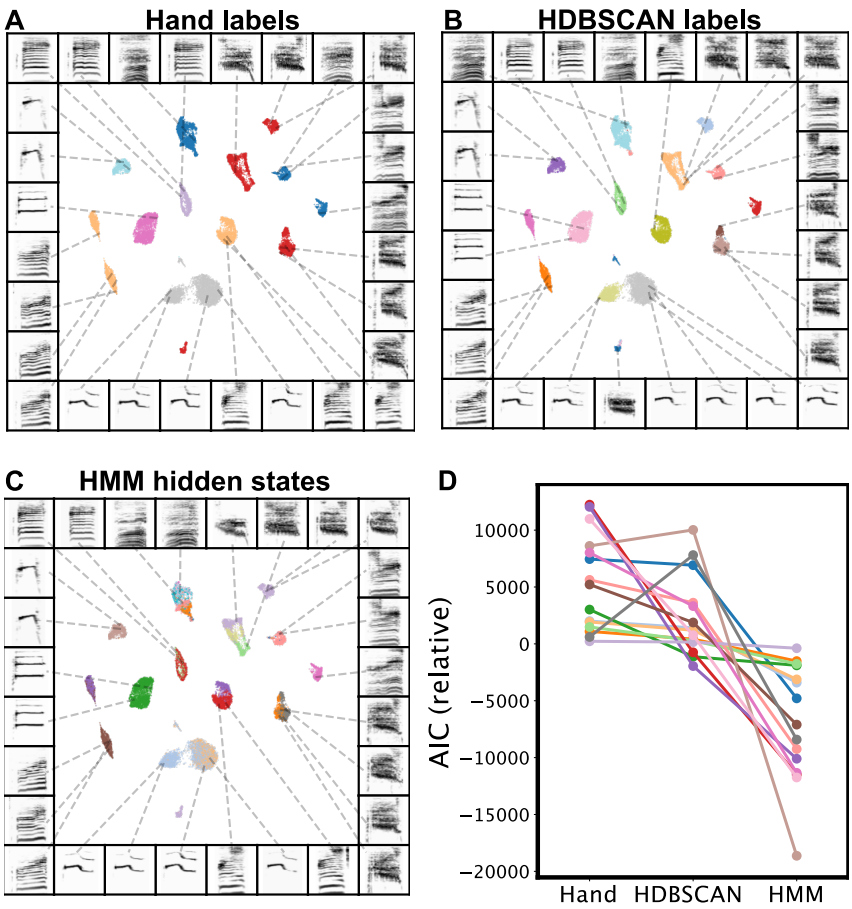
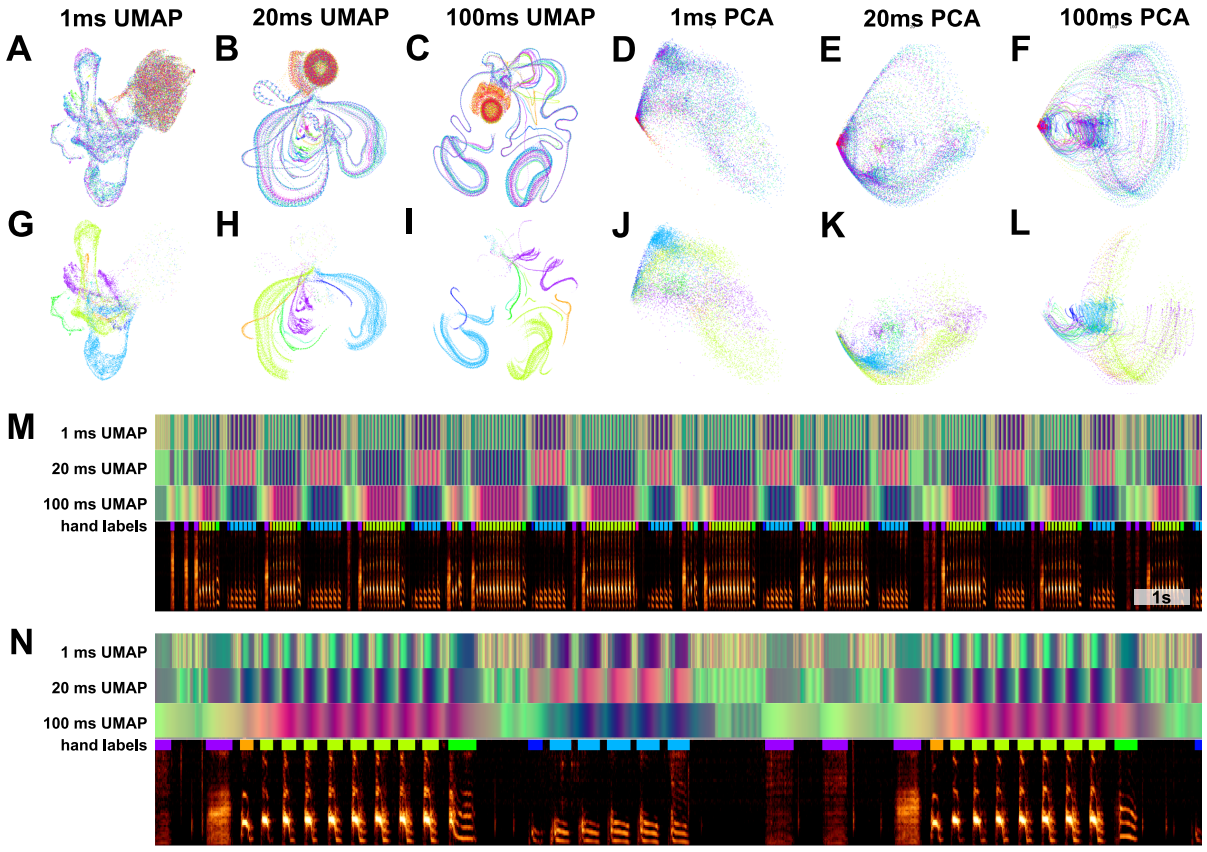


Figure 2.14. Comparison of Hidden Markov Model performance using different hidden states. Projections are shown for a single example bird from the Koumura dataset [219]. UMAP projections are labeled by three labeling schemes: (A) Hand labels (B) HDBSCAN labels on UMAP, and (C) Trained Hidden Markov Model (HMM) labels. (D) Models are compared across individual birds (points) on the basis of AIC. Each line depicts the relative (centered at zero) AIC scores for each bird for each model. Lower relative AIC equates to better model fit.

representations to those that treat vocal segments as single points (as in the previous Bengalese finch example in Fig 2.12). We explored four datasets, ranging from highly discrete clusters of vocal elements (Bengalese finch, Fig 2.15), to relatively discrete clustering (European starlings, Fig 2.16) to low clusterability (Mouse USV, Fig 2.17; Human speech, Fig 2.18). In each dataset, we find that continuous latent trajectories capture short and long timescale structure in vocal sequences without requiring vocal elements to be segmented or labeled.



(N) a subset of the bout shown in (M). In G-L, unlabeled points (points that are in between syllables) are not shown for visual clarity.

Figure 2.15. Continuous UMAP projections of Bengalese finch song from a single bout produced by one individual. (A-C) Bengalese finch song is segmented into either 1ms (A), 20ms (B), or 100ms (C) rolling windows of song, which are projected into UMAP. Color represents time within the bout of song 2(red marks the beginning and ending of the bout, corresponding to silence). (D-F) The same plots as in (A-C), projected into PCA instead of UMAP. (G-I) The same plots as (A-C) colored by hand-labeled element categories (unlabelled points are not shown). (J-L) The same plots as (D-F) colored by hand-labeled syllable categories. (M) UMAP projections represented in colorspace over a bout spectrogram. The top three rows are the UMAP projections from (A-C) projected into RGB colorspace to show the position within UMAP space over time as over the underlying spectrogram data. The fourth row are the hand labels. The final row is a bout spectrogram.

Bengalese finch song provides a relatively easy visual comparison between the discrete and continuous treatments of song, because it consists of a small number of unique highly stereotyped syllables (Fig 2.15). With a single bout of Bengalese finch song, which contains

several dozen syllables, we generated a latent trajectory of song as UMAP projections of temporally-rolling windows of the bout spectrogram (See Projections section). To explore this latent space, we varied the window length between 1 and 100ms (Fig 2.15A-L). At each window size, we compared UMAP projections (Fig 2.15A-C) to PCA projections (Fig 2.15D-F). In both PCA and UMAP, trajectories are more clearly visible as window size increases across the range tested, and overall the UMAP trajectories show more well-defined structure than the PCA trajectories. To compare continuous projections to discrete syllables, we re-colored the continuous trajectories by the discrete syllable labels obtained from the dataset. Again, as the window size increases, each syllable converges to a more distinct trajectory in UMAP space (Fig 2.15G-I). To visualize the discrete syllable labels and the continuous latent projections in relation to song, we converted the 2D projections into colorspace and show them as a continuous trajectory alongside the song spectrograms and discrete labels in Fig 2.15M,N. Colorspace representations of the 2D projections consist of treating the two UMAP dimensions as either a red, green, or blue channel in RGB (3D) colorspace, and holding the third channel constant. This creates a colormap projection of the two UMAP dimensions.

Latent trajectories of European starling song

European starling song provides an interesting case study for exploring the sequential organization of song using continuous latent projections because starling song is more sequentially complex than Bengalese finch song, but is still highly stereotyped and has well-characterized temporal structure. European starling song is comprised of a large number of individual song elements, usually transcribed as 'motifs', that are produced within a bout of singing. Song bouts last several tens of seconds and contain many unique motifs grouped into three broad classes: introductory whistles, variable motifs, and high-frequency terminal motifs [107]. Motifs are variable within classes, and variability is affected by the presence of potential mates and seasonality [365, 3]. Although sequentially ordered motifs are usually segmentable by gaps of silence occurring when starlings are taking breaths, segmenting motifs using silence alone can be

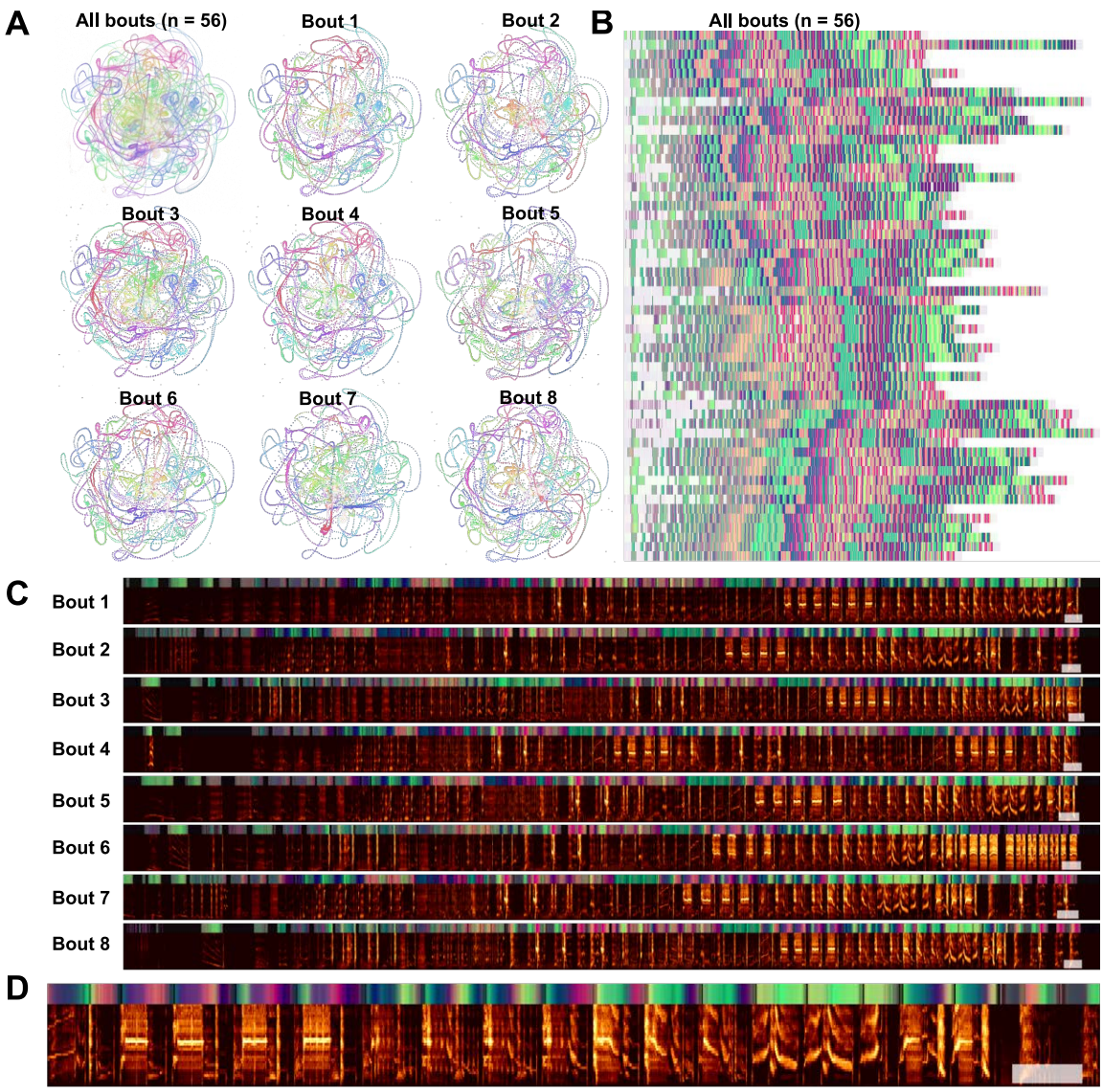


Figure 2.16. Starling bouts projected into continuous UMAP space. (A) The top left panel is each of 56 bouts of starling song projected into UMAP with a rolling window length of 200ms, color represents time within the bout. Each of the other 8 panels is a single bout, demonstrating the high similarity across bouts. (B) Latent UMAP projections of the 56 bouts of song projected into colorspace in the same manner as Fig 2.15M. Although the exact structure of a bout of song is variable from rendition to rendition, similar elements tend to occur at similar regions of song and the overall structure is preserved. (C) The eight example bouts from (A) with UMAP colorspace projections above. The white box at the end of each plot corresponds to one second. (D) A zoomed-in section of the first spectrogram in C.

difficult because pauses are often short and bleed into surrounding syllables [438]. When syllables are temporally discretized, they are relatively clusterable (Fig 2.8), however syllables tend to

vary somewhat continuously (Fig 2.16D). To analyze starling song independent of assumptions about segment (motif) boundaries and element categories, we projected bouts of song from a single male European starling into UMAP trajectories using the same methods as with Bengalese finch song in Fig 2.15. We used a 200ms time window for these projections, around the order of a shorter syllable of starling song and longer than the pause in between syllables, resulting our projections capturing information about transitions between syllables. Time windows of different lengths reveal structure at different timescales, for example windows shorter than the length of a pause between syllables will return to the region of latent space corresponding to silence (e.g. 2.15A,B) and capture within syllable structure but not the transitions between syllables.

We find that the broad structure of song bouts are highly repetitive across renditions, but contain elements within each bout that are variable across bout renditions. For example, in Fig 2.16A, the top left plot is an overlay showing the trajectories of 56 bouts performed by a single bird, with color representing time within each bout. The eight plots surrounding it are single bout renditions. Different song elements are well time-locked as indicated by a similar hue present in the same regions of each plot. Additionally, most parts of the song occur in each rendition. However, certain song elements are produced or repeated in some renditions but not others. To illustrate this better, in Fig 2.16B, we show the same 56 bouts projected into colorspace in the same manner as Fig 2.15M,N, where each row is one bout rendition. We observe that, while each rendition contains most of the same patterns at relatively similar times, some patterns occur more variably. In Fig 2.16C and D we show example spectrograms corresponding to latent projections in Fig 2.16A, showing how the latent projections map onto spectrograms.

Quantifying and visualizing the sequential structure of song using continuous trajectories rather than discrete element labels is robust to errors and biases in segmenting and categorizing syllables of song. Our results show the potential utility of continuous latent trajectories as a viable alternative to discrete methods for analyzing song structure even with highly complex, many-element, song.

Latent trajectories and clusterability of mouse USVs

House mice produce ultrasonic vocalizations (USVs) comprising temporally discrete syllable-like elements that are hierarchically organized and produced over long timescales, generally lasting seconds to minutes [62]. When analyzed for temporal structure, mouse vocalizations are typically segmented into temporally-discrete USVs and then categorized into discrete clusters [171, 62, 443, 73, 200] in a manner similar to syllables of birdsong. As was observed on the basis of the Hopkin's statistic (Fig 2.8), however, USVs do not cluster into discrete distributions in the same manner as birdsong. Choosing different arbitrary clustering heuristics will therefore have profound impacts on downstream analyses of sequential organization [171].

We sought to better understand the continuous variation present in mouse USVs, and explore the sequential organization of mouse vocalizations without having to categorize USVs. To do this, we represented mouse USVs as continuous trajectories (Fig 2.17E) in UMAP latent space using similar methods as with starlings (Fig 2.16) and finches (Fig 2.15). In Fig 2.17, we use a single recording of one individual producing 1,590 (Fig 2.17G) USVs over 205 seconds as a case study to examine the categorical and sequential organization of USVs. We projected every USV produced in that sequence as a trajectory in UMAP latent space (Fig 2.17A,C,D). Similar to our observations in Fig 2.8I using discrete segments, we do not observe clear element categories within continuous trajectories, as observed for Bengalese finch song (e.g. Fig 2.15I).

To explore the categorical structure of USVs further, we reordered all of the USVs in Fig 2.17G by the similarity of their latent trajectories (measured by the Euclidean distance between latent projection vectors; Fig 2.17F) and plotted them side-by-side (Fig 2.17H). Both the similarity matrix of the latent trajectories (Fig 2.17F) and the similarity-reordered spectrograms (Fig 2.17H) show that while some USVs are similar to their neighbors, no highly stereotyped USV categories are observable.

Although USVs do not aggregate into clearly discernible, discrete clusters, the temporal organization of USVs within the vocal sequence is not random. Some latent trajectories are more

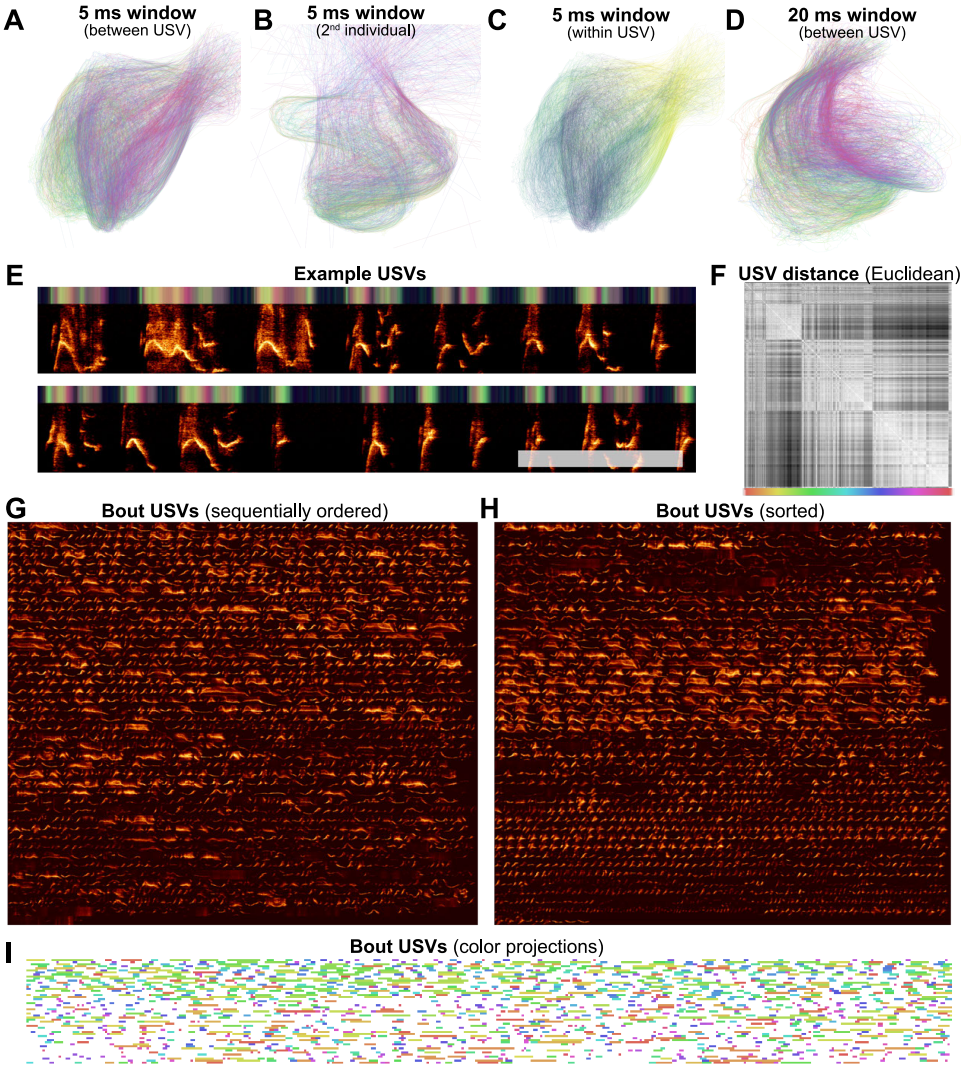


Figure 2.17. USV patterns revealed through latent projections of a single mouse vocal sequence. (A) Each USV is plotted as a line and colored by its position within the sequence. Projections are sampled from a 5ms rolling window. (B) Projections from a different recording from a second individual using the same method as in (A). (C) The same plot as in A, where color represents time within a USV. (D) The same plot as in (A) but with a 20ms rolling window. (E) An example section of the USVs from (A), where the bar on the top of the plot shows the UMAP projections in colorspace (the first and second UMAP dimensions are plotted as color dimensions). 2The white scale bar corresponds to 250ms. (F) A distance matrix between each of 1,590 USVs produced in the sequence visualized in (A), reordered so that similar USVs are closer to one another. (G) Each of the 1,590 USVs produced in the sequence from (A), in order (left to right, top to bottom). (H) The same USVs as in (G), reordered based upon the distance matrix in (F). (I) The entire sequence from (A) where USVs are color-coded based upon their position in the distance matrix in (F).

frequent at different parts of the vocalization. In Fig 2.17A, we color-coded USV trajectories according to each USV's position within the sequence. The local similarities in coloring (e.g., the purple and green hues) indicate that specific USV trajectories tend to occur in distinct parts of the sequence. Arranging all of the USVs in order (Fig 2.17G) makes this organization more evident, where one can see that shorter and lower amplitude USVs tend to occur more frequently at the end of the sequence. To visualize the vocalizations as a sequence of discrete elements, we plotted the entire sequence of USVs (Fig 2.17I), with colored labels representing the USV's position in the reordered similarity matrix (in a similar manner as the discrete category labels in Fig 2.15E). In this visualization, one can see that different colors dominate different parts of the sequence, again reflecting that shorter and quieter USVs tend to occur at the end of the sequence.

Latent trajectories of human speech

Discrete elements of human speech (i.e. phonemes) are not spoken in isolation and their acoustics are influenced by neighboring sounds, a process termed co-articulation. For example, when producing the words 'day', 'say', or 'way', the position of the tongue, lips, and teeth differ dramatically at the beginning of the phoneme 'ey' due to the preceding 'd', 's', or 'w' phonemes, respectively. This results in differences in the pronunciation of 'ey' across words (Fig 2.18F). Co-articulation explains much of the acoustic variation observed within phonetic categories. Abstracting to phonetic categories therefore discounts much of this context-dependent acoustic variance.

We explored co-articulation in speech, by projecting sets of words differing by a single phoneme (i.e. minimal pairs) into continuous latent spaces, then extracted trajectories of words and phonemes that capture sub-phonetic context-dependency (Fig 2.18). We obtained the words from the same Buckeye corpus of conversational English used in Figs 2.8, 2.7, and Supporting information. We computed spectrograms over all examples of each target word, then projected sliding 4-ms windows from each spectrogram into UMAP latent space to yield a continuous vocal trajectory over each word (Fig 2.18). We visualized trajectories by their corresponding

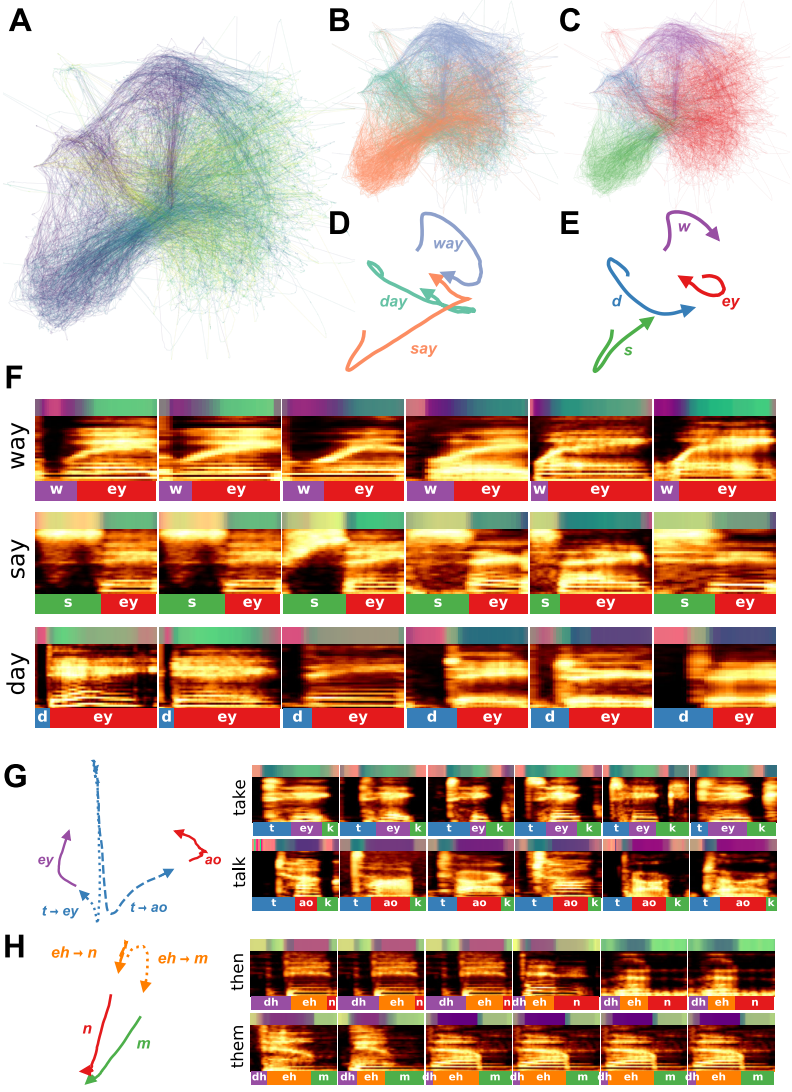


Figure 2.18. Speech trajectories showing coarticulation in minimal pairs. (A) Utterances of the words 'day', 'say', and 'way' are projected into a continuous UMAP latent space with a window size of 4ms. Color represents time, where darker is earlier in the word. (B) The same projections as in (A) but color-coded by the corresponding word. (C) The same projections are colored by the corresponding phonemes. (D) The average latent trajectory for each word. (E) The average trajectory for each phoneme. (F) Example spectrograms of words, with latent trajectories above spectrograms and phoneme labels below spectrograms. (G) Average trajectories and corresponding spectrograms for the words 'take' and 'talk' showing the different trajectories for 't' in each word. (H) Average trajectories and the corresponding spectrograms for the words 'then' and 'them' showing the different trajectories for 'eh' in each word.

word and phoneme labels (Fig 2.18B,C) and computed the average latent trajectory for each word and phoneme (Fig 2.18D,E). The average trajectories reveal context-dependent variation

within phonemes caused by coarticulation. For example, the words 'way', 'day', and 'say' each end in the same phoneme ('ey'; Fig 2.18A-F), which appears as an overlapping region in the latent space (the red region in Fig 2.18C). The endings of each average word trajectory vary, however, indicating that the production of 'ey' differs based on its specific context (Fig 2.18D). The difference between the production of 'ey' can be observed in the average latent trajectory over each word, where the trajectories for 'day' and 'say' end in a sharp transition, while the trajectory for 'way' is smoother (Fig 2.18D). These differences are apparent in Fig 2.18F which shows examples of each word's spectrogram accompanied by its corresponding phoneme labels and color-coded latent trajectory. In the production of 'say' and 'day' a more abrupt transition occurs in latent space between 's'/'d' and 'ey', as indicated by the yellow to blue-green transitions above spectrograms in 'say' and the pink to blue-green transition above 'day'. For 'way', in contrast, a smoother transition occurs from the purple region of latent space corresponding to 'w' to the blue-green region of latent space corresponding to 'ey'.

Latent space trajectories can reveal other co-articulations as well. In Fig 2.18G, we show the different trajectories characterizing the phoneme 't' in the context of the word 'take' versus 'talk'. In this case, the 't' phoneme follows a similar trajectory for both words until it nears the next phoneme ('ey' vs. 'ao'), at which point the production of 't' diverges for the different words. A similar example can be seen for co-articulation of the phoneme 'eh' in the words 'them' versus 'then' (Fig 2.18H). These examples show the utility of latent trajectories in describing sub-phonemic variation in speech signals in a continuous manner rather than as discrete units.

2.3 Discussion

We have presented a set of computational methods for projecting vocal communication signals into low-dimensional latent representational spaces, learned directly from the spectrograms of the signals. We demonstrate the flexibility and power of these methods by applying them to a wide sample of animal vocal communication signals, including songbirds, primates,

rodents, bats, and cetaceans (Fig 2.8). Deployed over short timescales of a few hundred milliseconds, our methods capture significant behaviorally-relevant structure in the spectro-temporal acoustics of these diverse species' vocalizations. We find that complex attributes of vocal signals, such as individual identity (Fig 2.4), species identity (Fig 2.5A,B), geographic population variability (Fig 2.5C), phonetics (Fig 2.7, Supporting information), and similarity-based clusters (Fig 2.10) can all be captured by the unsupervised latent space representations we present. We also show that songbirds tend to produce signals that cluster discretely in latent space, whereas mammalian vocalizations are more uniformly distributed, an observation that deserves much closer investigation in more species. Applied to longer timescales, spanning seconds or minutes, the same methods allowed us to visualize sequential organization and test models of vocal sequencing (Fig 2.12). We demonstrated that in some cases latent approaches confer advantages over hand labeling or supervised learning (Fig 2.13, 2.14). Finally, we visualized vocalizations as continuous trajectories in latent space (Figs 2.15, 2.16, 2.17, 2.18), providing a powerful method for studying sequential organization without discretization [200].

Latent models have shown increasing utility in the biological sciences over the past several years. As machine learning algorithms improve, so will their utility in characterizing the complex patterns present in biological systems like animal communication. In neuroscience, latent models already play an important role in characterizing complex neural population dynamics [79]. Similarly, latent models are playing an increasingly important role in computational ethology [46], where characterizations of animal movements and behaviors have uncovered complex sequential organization [268, 29, 462]. In animal communication, pattern recognition using various machine learning techniques has been used to characterize vocalizations and label auditory objects [381, 76, 73, 443, 150, 215, 171]. Our work furthers this emerging research area by demonstrating the utility of unsupervised latent models for both systematically visualizing and abstracting structure from animal vocalizations across a wide range of species.

Latent and known features

Our methods show that *a priori* feature-based compression is not a prerequisite to progress in understanding behaviorally relevant acoustic diversity. The methods we describe are not meant, however, as a wholesale replacement of more traditional analyses based on compression of vocal signals into known behaviorally-relevant feature spaces. In most cases, these known feature spaces are the result of careful exploration, experimentation, and testing, and therefore encapsulate an invaluable pool of knowledge. When available, this knowledge should be used. Our methods are most useful when this knowledge is either unavailable, or may not hold for all of the species one wishes to investigate. For comparison across species, they provide a common space that is unbiased by the features of any one species (Fig 2.5), and within species they can reveal behaviorally relevant structure (Figs 2.4, 2.6, 2.7). In the cases where we compared latent features to the representations of signals based on known features (Figs 2.3, 2.11), it is clear that the known features captured aspects of the signals that the latent representations missed. At the same time, however, the latent representations capture much of the same variance, albeit without reference to intuitive features. Thus, when possible the distributional properties of signals revealed by our unsupervised methods can (and should be) linked to specific physical features of the signals.

In light of the observation that latent features can provide a close approximation to feature-based representations, it is interesting to ask why UMAP works as well as it does in the myriad ways we have shown. Like other compression algorithms, UMAP relies on statistical regularities in the input data to find the low dimensional manifold that best captures a combination of global and local structure. Thus, the co-variance between behaviorally relevant features and those revealed in UMAP indicates that behaviorally relevant dimensions of signals contain reliable acoustic variance. In other words, the statistical structure of the signals reflect their function. While these may not be the dimensions of maximal variance over the whole signal set (which is why PCA can miss them), the local variance is reliable enough to be captured by UMAP. While it may be less surprising to note that animal communication relies upon reliable signal variance

to convey information, it is noteworthy that our signal analysis methods have advanced to the point where we can directly measure that variance without prior knowledge.

Discrete and continuous representations of vocalizations

Studies of animal communication classically rely on segmenting vocalizations into discrete temporal units. In many species, this temporal segmentation is a natural step in representing and analyzing vocal data. In birdsong, for example, temporally distinct syllables are often well defined by clear pauses between highly stereotyped syllable categories (Fig 2.8O). We showed that the syllables labeled through unsupervised clustering account for sequential organization in Bengalese finch song better than experimenter-defined hand labels, even when hand-labels are treated as ground-truth (Fig 2.14D). Using an HMM that was free to define states based on higher-order sequential dynamics revealed even finer sub-classes of elements with reliable acoustic structure (Fig 2.14D). Thus, one strategy to improve syllable labeling algorithms going forward is to include models for the sequential dynamics of vocalizations (e.g [235]). Such models should take into account recent findings that Markovian assumptions do not fully account for the long-range dynamics in all bird songs [381] or other signals with long-range organization such as human speech [249]. Lastly, neither density-based clustering, hand clustering, nor sequence-based clustering, model the animal's categorical perception directly. Therefore, making perceptual inferences based upon these labels is limited without behavioral or physiological investigations with the animal.

Another strategy for studying vocal sequences is to avoid the problem of segmentation/discretization altogether. Indeed, in many non-avian species, vocal elements are either not clearly stereotyped or temporally distinct (Fig 2.8), and methods for segmentation can vary based upon changes in a range of acoustic properties, similar sounds, or higher-order organization [200]. These constraints force experimenters to make decisions that can affect downstream analyses [150, 171]. We projected continuous latent representations of vocalizations ranging from the highly stereotyped song of Bengalese finches, to highly variable mouse USVs, and found that

continuous latent projections effectively described useful aspects of spectro-temporal structure and sequential organization (Figs 2.15, 2.16, and 2.17). Continuous latent variable projections of human speech capture sub-phoneme temporal dynamics that correspond to co-articulation (Fig 2.18). Collectively, our results show that continuous latent representations of vocalizations provide an alternative to discrete segment-based representations while remaining agnostic to segment boundaries, and without the need to segment vocalizations into discrete elements or symbolic categories. Of course, where elements can be clustered into clear and discrete element categories, it may be valuable to do so. The link from temporally continuous vocalization to symbolically discrete sequences will be an important target for future investigations.

2.3.1 Limitations

Throughout this manuscript, we have discussed and applied a set of tools for analyzing vocal communication signals. Although we have spent much of the manuscript focusing on the utility of these tools, there are limits to their application. Here, we discuss a few of the drawbacks and challenges.

Unsupervised learning with noisy vocal signals

Supervised learning algorithms are trained explicitly to learn what features of a dataset are relevant to mapping input data to a set of labels. For example, a neural network trained to classify birdsong syllables based upon hand labels is a supervised algorithm. Such algorithms learn what parts of the data are relevant to this mapping (signal), and what parts of data should be ignored (noise). Conversely, unsupervised learning algorithms model structure in data without external reference to what is signal and what is noise. The datasets we used ranged from relatively noisy signals recorded in the wild, to recordings in a laboratory setting from a single individual in a sound-isolated chamber. Algorithms like PCA or UMAP can ignore some level of noise in data. In PCA, for example, when signal explains more variance in the data than noise, the first few principal components will capture primarily signal. Similarly, UMAP embeddings rely on

the construction of a nearest neighbor graph. So long as noise does not substantially influence the construction of this graph, some degree of noise can be ignored. Still, high background noise in recordings can impact the quality of latent projections. Thus, signal-aware methods for reducing noise before projecting the data would be ideal. In the methods, we discuss one method we used to decrease background time-domain noise in some of the noisier signals using a technique called ”spectral gating”. Considerations of how to reduce the noise in data are crucial to modeling structure in animal communication signals, especially using unsupervised learning algorithms.

Unsupervised learning with small vocalization datasets

Contemporary machine learning algorithms often rely on very large datasets. To learn the structure of a complex vocal communicative repertoire, having more coverage over the vocal repertoire is better; it would be difficult to find clusters of vocalizations when only a few exemplars are available for each vocalization. In contrast, when features of a dataset are already known, less data is needed to make a comparison. For example, it might take a machine learning algorithm many exemplars to untangle data in such a way that important features like fundamental frequency are learned. As such, when datasets are small, methods like UMAP are less useful in modeling data, and carefully selecting features is generally a more appropriate method for making comparisons.

Representing data and distance across vocalizations

Graph-based models like UMAP find structure in data by building graphical representations of datasets and then embedding those graphs into low-dimensional spaces. Building a graphical representation of a dataset is predicated on determining a notion of distance between points. Deciding how to measure the distance between two elements of animal communication requires careful thought. Throughout this manuscript, we computed spectrograms of vocalizations, and computed distance as the Euclidean distance between those spectrograms. This measure of distance, while easy to compute, is one of many ways to measure the distance between points.

The use of both PAFs and spectrograms should be considered carefully when making comparisons in vocal datasets. Descriptive statistics can be overly reductive and may not capture all of the relevant characteristics of the signal, while spectrogram representations can be overcomplete, and require further dimensionality reduction to reveal relevant features in statistical analyses [307, 109, 132]. Treating time-frequency bins of spectrograms as independent features inaccurately reflects the perceptual space of animals, who are sensitive to relative relationships between time varying components and spectral shape (e.g., [44]) less than absolute power at specific at a specific time or frequency. For example, the spectrograms of two identically shaped vocalizations shifted in frequency by a quarter octave may appear completely uncorrelated when each time-frequency coefficient is treated as an independent dimension. Yet, those same vocalizations might be treated as effectively the same by a receiver. Topological methods such as UMAP or t-SNE partially resolve this issue, because their graph-based representations rely on the relationships between neighboring data points as inputs. As a result, vocalizations that are distant in Euclidean space (i.e. whose spectrograms are uncorrelated) can be close in latent space. Even when using spectrograms, determining the parameters of the spectrogram is an important consideration, and can impact the result of downstream machine learning tasks for bioacoustics [207, 109].

Constructing a graph in UMAP relies on computing the distances between some representation of the data (here, vocalizations). Representing vocal elements as spectrograms or PAFs, and constructing a graph on the basis of the Euclidean distance between those features are two ways of constructing that graph. In principle, any distance metric could be used in place of Euclidean distance to build the graph in UMAP. For example, the distance between two spectrograms can be computed using Dynamic Time Warping (DTW) [212, 80], Dynamic Frequency Warping (DFW) [410], or peaks in cross-correlations [198] to add invariance to shifts in time and frequency between vocal elements[2]. Determining what notion of distance is most

[2]In the code [377], we show an example of how DTW can be used as the distance metric in UMAP instead of Euclidean distance.

reasonable to compare two vocalizations requires consideration. When acoustic features are known to capture the structure of an animal's communication, either by careful study or explicitly probing an animal's perceptual representations of their vocal repertoire, the distance can be computed on the basis of those acoustic features. Here, we use Euclidean distance between spectrograms to build UMAP graphs, which we find is effective to capture structure in many vocal signals.

Parameterization and understanding structure in latent projections

It has been well documented that the structure found using graph-based dimensionality reduction algorithms like t-SNE and UMAP can be heavily biased by the parameterization used in the algorithm [454, 72]. Generally, the default parameters used in UMAP are good starting points. In this manuscript, we used the default parameters in all of our projections, except where otherwise noted. Still, exploring the persistence of structure across parameterizations is an important consideration when making inferences based upon structure in latent space.

2.3.2 Future work

Synthesizing animal vocalization signals

The present work discusses latent models from the angle of dimensionality reduction, learning a low dimensional descriptive representation of the structure of the signal. Here, we left a second important aspect of latent models unexplored: generativity. One aspect of machine learning that is largely under-utilized in animal communication research, and psychophysics more generally, is using generative latent models that jointly model the probability in data space and latent space, to generate vocalizations directly from samples in latent space. Generative techniques enable the synthesis of complex, high-dimensional data such as animal communication signals by sampling from low-dimensional latent spaces. Preliminary work has already been done in this area, for example, generating syllables of birdsong as stimuli for psychophysical and neurophysiological probes [430, 382] using deep neural networks. HMMs have also been used to synthesize vocalizations [38]. With the recent advancements in machine learning, especially

in areas such as generative modeling and text-to-speech, the synthesis of high-fidelity animal vocal signals is likely to become an important avenue for studying the full spectrum of vocal communication in more biologically realistic ways.

Local and global structure

The methods we present in this paper center around the graph-based dimensionality reduction algorithm UMAP. Graph-based dimensionality reduction algorithms like t-SNE and UMAP favor the preservation of local structure of global structure, as opposed to PCA and MDS, which favor the preservation of global structure. Capturing local structure means mapping nearby points in data-space to nearby points in the low-dimensional embedding space, while capturing global structure means preserving relationships at all scales; both local and more distant [402]. UMAP and t-SNE capture much more local structure than PCA, but less global structure. However, UMAP is an improvement over t-SNE in that it captures more global structure [280]. The current deficit in capturing global structure with graph-based dimensionality reduction algorithms is not necessarily a fundamental issue, however. Future advancements in non-linear graph-based dimensionality reduction algorithms will likely better capturing global structure. Capturing the density of distributions (like clusters of birdsong elements) is also likely an important feature of dimensionality reduction algorithms. At present, neither UMAP nor t-SNE embeddings are designed to capture local density (the distances between points) in data space (they are explicitly designed not to). Recent improvements on this front [313], for example, might aid in finding structural differences between directed birdsong which is more highly stereotyped and undirected birdsong, which is more exploratory. Advances in non-linear graph-based dimensionality reduction algorithms are likely to have important impacts on quantifying latent structure in vocal data.

Further directions

The work presented here is a first step in exploring the potential power of latent modeling in animal communication. We touch only briefly on a number of questions that we find interesting

and think important within the field of animal communication. Other researchers may certainly want to target other questions, and we hope that some of these techniques (and the provided code) may be adapted in that service. Our analyses were taken from a diverse range of animals, sampled in diverse conditions both in the wild and in the laboratory, and are thus not well controlled for variability between species. Certainly, as bioacoustic data becomes more open and readily available, testing large, cross-species, hypotheses will become more plausible. We introduced several areas in which latent models can act as a powerful tool to visually and quantitatively explore complex variation in vocal data. These methods are not restricted to bioacoustic data, however. We hope that the work presented here will encourage a larger incorporation of latent and unsupervised modeling as a means to represent, understand, and experiment with animal communication signals in general. At present, our work exhibits the utility of latent modeling on a small sampling of the many directions that can be taken in the characterization of animal communication.

2.4 Methods

2.4.1 Datasets

The Buckeye [346] dataset of conversational English was used for human speech. The swamp sparrow dataset is from [234] and was acquired from [232]. The California thrasher dataset is from [71] and was acquired from BirdDB [19]. The Cassin's vireo dataset is from [165] and was also acquired from BirdDB. The giant otter dataset was acquired from [310]. The canary song dataset is from [266] and was acquired via personal correspondence. Two zebra finch datasets were used. The first is a dataset comprised of a large number of motifs produced by several individuals from [337]. The second is a smaller library of vocalizations with more diverse vocalization types and a greater number of individuals than the motif dataset. It correspond to data from [109] and [110] and was acquired via personal correspondence. The white-rumped munia dataset is from [194]. The humpback whale dataset was acquired from Mobysound [283]. The house mice USV dataset was acquired from [62]. An additional higher SNR dataset of mouse USVs was sent from the same group via personal correspondence. The European starling dataset is from [381] and was acquired from [16]. The gibbon song is from [303]. The marmoset dataset was received via personal correspondence and was recorded similarly to [288]. The fruit bat data is from [348] and was acquired from [349]. The macaque data is from [130] and was acquired from [131]. The beaked whale dataset is from [172] and was acquired from [128]. The North American birds dataset is from [471] and was acquired from [470]. We used two Bengalese finch datasets. The first is from [220] and was acquired from [219]. The second is from [322].

2.4.2 Reducing noise in audio

One issue with automated analyses over animal communication is the requirement for signals to have relatively low background noise in their recordings. In part, background noise reduction can be performed algorithmically. At the same time, noisier data requires a greater

degree of human intervention to tell the algorithm what to consider signal, and what to consider noise. While some of the datasets used in our analyses were recorded in sound-isolated chambers in a laboratory, others were recorded in nature. The datasets we ultimately used for this paper were either relatively low noise or had some hand-annotations that were necessary to determine where syllables started and ended. For example, many of the datasets had hand segmented vocal element boundaries that were used instead of algorithmic segmentation. We show a comparison of the silhouette score of the Cassin's dataset used in Fig 2.2 for different signal-to-noise ratios (SNR) in Supporting information.

To reduce the background noise in acoustic signals, we wrote a spectral gating noise reduction algorithm [375]. The algorithm is inspired by the noise reduction algorithm used in the Audacity(R) sound editing software [425].

Given a waveform of audio with both signal and background noise (S_n), and a sample audio clip from the same or a similar waveform with only background noise (N). An outline of the algorithm is as follows:

1. Compute the short-time Fourier transform over N ($spec_n$).
2. Compute the mean and standard deviation of $spec_n$ for each frequency component over time.
3. Compute the short-time Fourier transform over S_n ($spec_s$).
4. For each frequency component, compute a threshold noise level based upon the mean and standard deviation of $spec_n$
5. Generate a mask over $spec_s$ based upon the power of $spec_s$ and the thresholds determined from ($spec_n$)
6. Smooth the mask over frequency and time.
7. Apply the mask to $spec_s$ to remove noise.

8. Compute the inverse short-time Fourier transform over $spec_s$ to generate a denoised time-domain signal.

We made a Python package of this algorithm called `noisereduce` available on GitHub [375]. In addition to the spectral gating noise reduction algorithm, segmentation was performed by a dynamic thresholding algorithm, which is described in the Segmentation section. We also show the fidelity of UMAP projections over different levels of noise in Supporting information, where we observe that UMAP is robust to relatively high noise in comparison to spectrograms.

2.4.3 Segmentation

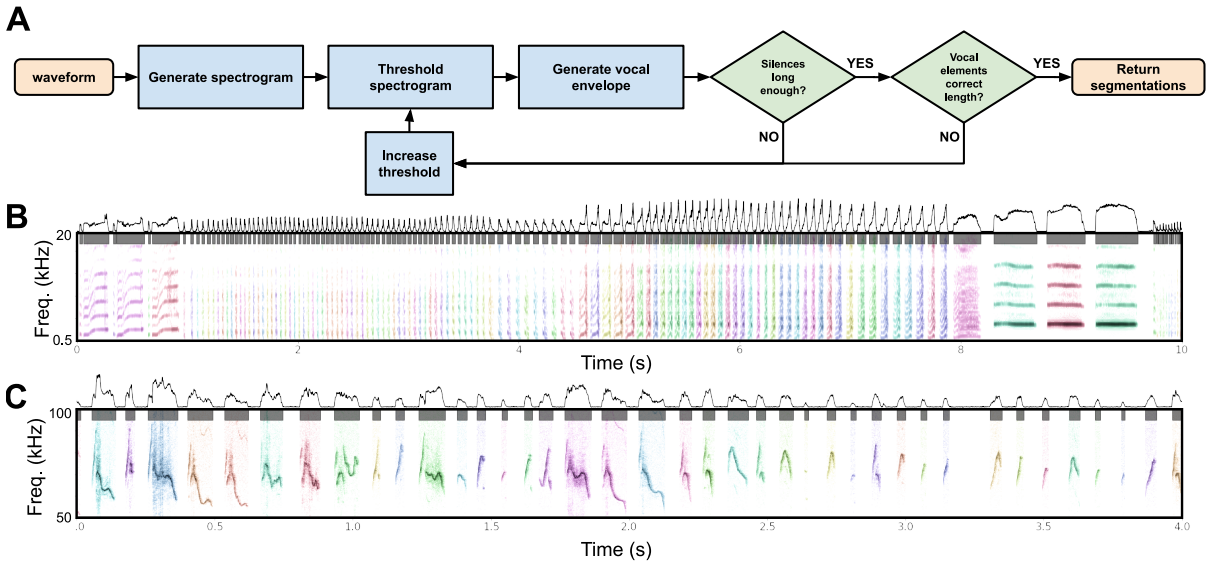


Figure 2.19. Segmentation algorithm (A) The dynamic threshold segmentation algorithm. The algorithm dynamically defines a noise threshold based upon the expected amount of silence in a clip of vocal behavior. Syllables are then returned as continuous vocal behavior separated by noise. (B) The segmentation method from (A) applied to canary syllables. (C) The segmentation method from (A) applied to mouse USVs.

Many datasets were made available with vocalizations already segmented either manually or algorithmically into units. When datasets were pre-segmented, we used the segment boundaries defined by the dataset authors. For all other datasets, we used a segmentation algorithm we call dynamic threshold segmentation (Fig 2.19A). The goal of the algorithm is to segment vocalization

waveforms into discrete elements (e.g. syllables) that are defined as regions of continuous vocalization surrounded by silent pauses. Because vocal data often sits atop background noise, the definition for silence versus vocal behavior was set as some threshold in the vocal envelope of the waveform. The purpose of the dynamic thresholding algorithm is to set that noise threshold dynamically based upon assumptions about the underlying signal, such as the expected length of a syllable or a period of silence. The algorithm first generates a spectrogram, thresholding power in the spectrogram below a set level to zero. It then generates a vocal envelope from the power of the spectrogram, which is the maximum power over the frequency components times the square root of the average power over the frequency components for each time bin over the spectrogram:

$$\mu_S(t) = \frac{1}{n}\sum_f S(t,f) \tag{2.1}$$

$$E(t) = \sqrt{\mu_S(t)}\max_f S(t,f) \tag{2.2}$$

Where E is the envelope, S is the spectrogram, t is the time bin in the spectrogram, f is the frequency bin in the spectrogram, and n is the total number of frequency bins.

The lengths of each continuous period of putative silence and vocal behavior are then computed. If lengths of vocalizations and silences meet a set of thresholds (e.g. minimum length of silence and maximum length of continuous vocalization) the algorithm completes and returns the spectrogram and segment boundaries. If the expected thresholds are not met, the algorithm repeats, either until the waveform is determined to have too low of a signal to noise ratio and discarded, or until the conditions are met and the segment boundaries are returned. The output of the algorithm, color coded by segment boundaries, are shown for a sample of canary song in Fig 2.19B and a sample of mouse USVs in Fig 2.19C. The code for this algorithm is available on Github [376].

2.4.4 Spectrogramming

Spectrograms are created by taking the absolute value of the one-sided short-time Fourier transformation of the Butterworth band-pass filtered waveform. Power is log-scaled and thresholded using the dynamic thresholding method described in the Segmentation section. Frequency ranges and scales are based upon the frequency ranges occupied by each dataset and species. Frequency is logarithmically scaled over a frequency range using a Mel filterbank (a filterbank logarithmically scaled to match human frequency perception). Mel frequency scaling was used as it has previously proven useful in extracting features of animal vocalizations [416], although in some cases linear frequency scaling can perform better in bioacoustics [207]. All of the spectrograms we computed had a total of 32 frequency bins, scaled across frequency ranges relevant to vocalizations in the species. None of these parameters were rigorously compared across each of the datasets, although we would recommend such comparisons in more detailed analyses. Each of the transformations done to data (e.g. downsampling, Mel scaling) reduce the dimensionality and impose *a priori* assumptions on the data. Performing analyses on a complete or invertible representation of the sound pressure waveform would make fewer assumptions [109], but is more computationally costly.

To create a syllable spectrogram dataset (e.g. for projecting into Fig 2.8), syllables are segmented from the vocalization spectrogram. To pad each syllable spectrogram to the same time length size, syllable spectrograms are log-rescaled in time (i.e. resampled in time relative to the log-duration of the syllable) then zero-padded to the length of the longest log-rescaled syllable.

2.4.5 Projections

Latent projections are either performed over discrete units (e.g. syllables) or as trajectories over continuously varying sequences. For discrete units, syllables are segmented from spectrograms of entire vocalizations, rescaled, and zero-padded to a uniform size (usually 32

frequency and 32 time components). These syllables are then projected into UMAP, where each time-frequency bin is treated as an independent dimension.

Trajectories in latent space are projected from rolling windows taken over a spectrogram of the entire vocal sequence (e.g. a bout of birdsong; Fig 2.20). The rolling window is a set length in milliseconds (e.g. 5ms) and each window is treated as a single point to be projected into latent space. The window then rolls one frame (in the spectrogram) at a time across the entire spectrogram, such that the number of samples in a bout trajectory is equal to the number of time-frames in the spectrogram. These time bins are then projected into UMAP latent space.

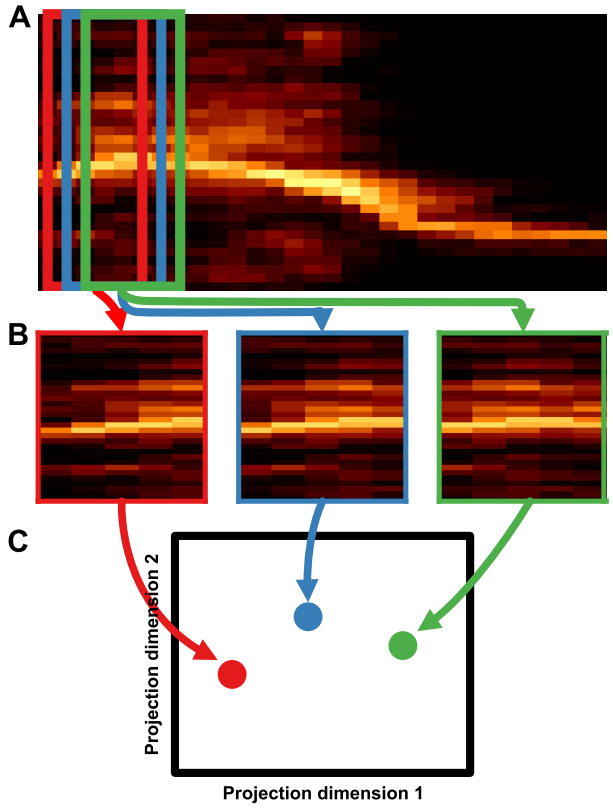


Figure 2.20. Continuous projections from vocalizations. (A) A spectrogram of each vocalization is computed. (B) Rolling windows are taken from each spectrogram at a set window length (here 5ms), and a step size of one time-frame of the short-time Fourier transform (STFT). (C) Windows are projected into latent space (e.g. UMAP or PCA).

2.4.6 Clusterability

Hopkin's statistic

We used the Hopkin's statistic [177] as a measure of the clusterability of datasets in UMAP space. In our case, the Hopkin's statistic was preferable over other metrics for determining clusterability, such as the Silhouette score [373] because the Hopkin's statistic does not require labeled datasets or make any assumptions about what cluster a data point should belong to. The Hopkin's statistic is part of at least one birdsong analysis toolkit [232].

The Hopkin's statistic compares the distance between nearest neighbors in a dataset (e.g. syllables projected into UMAP), to the distance between points from a randomly sampled dataset and their nearest neighbors. The statistic computes clusterability based upon the assumption that if the real dataset is more clustered than the randomly sampled dataset, points will be closer together than in the randomly sampled dataset. The Hopkin's statistic is computed over a set X of n data points (e.g. latent projections of syllables of birdsong), where the set X is compared with a baseline set Y of m data points sampled from either a uniform or normal distribution. We chose to sample Y from a uniform distribution over the convex subspace of X. The Hopkin's metric is then computed as:

$$\text{Hopkin's statistic} = \frac{\sum_{i=1}^{m} w_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d} \tag{2.3}$$

Where u_i is the distance of $y_i \in Y$ from its nearest neighbor in X and w_i is the distance of $x_i \in X$ from its nearest neighbor in X. Thus if the real dataset is more clustered than the sampled dataset, the Hopkin's statistic will approach 0, and if the dataset is less clustered than the randomly sampled dataset, the Hopkin's statistic will sit near 0.5. Note that the Hopkin's statistic is also commonly computed with $\sum_{i=1}^{m} u_i^d$ in the numerator rather than $\sum_{i=1}^{m} w_i^d$, where Hopkin's statistics closer to 1 would be higher clusterability, and closer to 0.5 would be closer to chance. We chose the former method because the range of Hopkin's statistics across datasets were more easily visible when log transformed.

To compare the clusterability across songbirds and mammals, we used a likelihood ratio test between linear mixed-effects models predicting the Hopkin's statistic for each individual. Each model controlled for the number of vocalizations produced by individuals, and random variation in clusterability at the species level. In addition, we included only individuals that had recordings consisting of at least 200 vocalizations. The likelihood ratio test was performed between a model with, and without including class (i.e. songbird versus mammal) as a category.

Silhouette score

As opposed to the Hopkin's statistic, which measures the general clusterability of a projection without regard to cluster labels, the silhouette score measures the clusterability of datasets when cluster labels are known or have already been inferred [373]. In other words, the Hopkin's statistic measures how clusterable a projection is, and the silhouette score measures how well fit a clustering is to a projection.

The silhouette score, S is computed as the mean of the silhouette coefficients for each data point. For each data point (i), the silhouette coefficient s_i is the mean distance between the data point and all other data points in the same cluster (a_i), minus the distance to that points nearest neighbor belonging to a different cluster (b_i), divided by the maximum of a_i and b_i, which can be written as:

$$s_i = \begin{cases} 1 - a_i/b_i, & \text{if } a_i < b_i \\ 0, & \text{if } a_i = b_i \\ b_i/a_i - 1, & \text{if } a_i > b_i \end{cases} \tag{2.4}$$

$$S = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{2.5}$$

This value is therefore bounded between 1, when the average distance to other points within-cluster (a_i) is very large relative to the distance to the points nearest neighbor (b_i), and -1 when the average distance to points within-cluster (a_i) is very small relative to the distance to

the nearest neighbor (b_i). Silhouette scores were compared across projections using a Kruskal-Wallis H-test over silhouette coefficients. Silhouette scores were compared to chance using a Kruskal-Wallis H-test over silhouette coefficients versus silhouette coefficients where labels are randomly permuted.

2.4.7 Clustering vocalizations

HDBSCAN

HDBSCAN clustering was performed on PCA and UMAP projections of Cassin's vireo syllables and Bengalese finch syllables (Table 2.1), as well as UMAP projections of swamp sparrow syllables (Fig 2.11). Each clustering used the default parameterization of UMAP and HDBSCAN, setting the minimum cluster size at 1% of the number of syllables/notes in the dataset.

K-means

We used k-means as a comparison to hierarchical density-based labeling when clustering Bengalese finch and Cassin's vireo syllables (Table 2.1). K-means clustering partitions a set of data points into k-clusters by tiling data space with a set of cluster centroids, and clustering each data point with the nearest centroid. We set the number of clusters (k) to the ground truth number of clusters in each dataset, to make clustering more competitive with HDBSCAN. We used the k-means implementation in Scikit-learn [338] to fit the models.

Gaussian Mixture model

We clustered the known feature space of swamp sparrow song using a Gaussian Mixture Model (GMM). GMMs assume that data are generated from a mixture of finite Gaussian distributions. Our GMM fit the parameters of the distribution to the data using expectation-maximization. In both the Conneaut Marsh, PA and Hudson Valley, NY swamp sparrow datasets, we set the number of distributions to be equal to the numbers used in the same populations of swamp sparrows in [233]. As opposed to [233], we clustered only on the duration of the notes,

and the start and end peak frequencies of the notes, without the mean peak frequency or vibrato amplitude. Still, these clusterings were similar to the clusterings presented in [233]. A direct comparison between our clustering using GMM and the clustering in [233] can be made by comparing Fig 2.11 and Supporting information with Lachlan and Nowicki [233] Fig S2A,B. We used the GMM implementation in Scikit-learn [338] to fit the models.

2.4.8 Comparing algorithmic and hand-transcriptions

Several different metrics can be used to measure the overlap between two separate labeling schemes. We used three metrics that capture different aspects of similarity to compare hand labeling to algorithmic clustering methods ([338, 372]; Table 2.1). Homogeneity measures whether all clusters fall into the same hand-labeled class in the labeled dataset.

$$\text{homogeneity(clusters, classes)} = 1 - \frac{H(\text{classes}|\text{clusters})}{H(\text{classes})} \tag{2.6}$$

Where $H(\text{classes}|\text{clusters})$ is the conditional entropy of the ground truth classes given the cluster labels, and $H(\text{classes}$ is the entropy of the classes.

Completeness measures the extent to which members belonging to the same hand-labeled class fall into the same cluster:

$$\text{completeness(clusters, classes)} = 1 - \frac{H(\text{clusters}|\text{classes})}{H(\text{clusters})} \tag{2.7}$$

V-measure is the harmonic mean between homogeneity and completeness.

$$\text{V-Measure} = 2 * \frac{\text{homogeneity} \cdot \text{completeness}}{\text{homogeneity} + \text{completeness}} \tag{2.8}$$

V-measure is also equivalent to the normalized mutual information between distributions [338]. In the swamp sparrow datasets, we compared the probability of overlap in clustering of labels (e.g. HDBSCAN and GMM) to chance by comparing V-measure of the true overlap to the

bootstrapped V-measure permuting the clusterings (10,000 times).

2.4.9 Hidden Markov Models (HMMs)

We used HMMs as a basis for comparing hand labels versus UMAP/HDBSCAN clustering in representing sequential organization in Bengalese finch song. Specifically, we treated hand labels as ground truth "visible" states in a discrete emission HMM, and generated several HMMs with different hidden states: hand labels, HDBSCAN labels, and hidden states learned using the Baum-Welch algorithm. HMMs were generated using the Python package Pomegranate [395]. Each model was compared on the basis of the log-likelihood of the data given the model. This log-likelihood score is treated as equal to the likelihood of the model given the data, and is also used as the basis computing AIC [147].

2.4.10 Data Availability

All of the vocalization datasets used in this study were acquired from external sources, most of them hosted publicly online (See Supporting information). The data needed to reproduce our results can be found on Zenodo (10.5281/zenodo.3775893).

2.4.11 Code Availability

The python code written specifically for this paper is available at Github.com/timsainb/AVGN_paper. A cleaner and more maintained code base is additionally available at Github.com/timsainb/AVGN.

2.4.12 Ethics statement

Procedures and methods comply with all relevant ethical regulations for animal testing and research and were carried out in accordance with the guidelines of the Institutional Animal Care and Use Committee at the University of California, San Diego (S05383).

2.5 Supporting information

Table 2.2. Overview of the species and datasets used in this paper.

Species	# Indv.	# Elements	Median len. (s)	Total length (s)	# Rec.	References
American crow	Unk.	syllables: 252	syllables: 0.37	100.5	252	[471, 470]
Bengalese finch	4	syllables: 215480	syllables: 0.065	40205.6	2663	[322]
Bengalese finch	11	notes: 214915	notes: 0.089	35365.9	2964	[220, 219]
Blue jay	Unk.	syllables: 250	syllables: 0.47	141.2	250	[471, 470]
California thrasher	18	syllables: 15328	syllables: 0.146	19958.9	92	[71, 19]
Canary	5	phrases: 22167 syllables: 497338	phrases: 1.319 syllables: 0.04	36986.9	2320	[266]
Cassin's vireo	48	syllables: 67316	syllables: 0.332	434782.4	422	[165, 19]
Cedar waxwind	Unk.	syllables: 245	syllables: 0.425	116.0	245	[471, 470]
Chipping sparrow	Unk.	syllables: 252	syllables: 0.09	24.9	252	[471, 470]
Common marmoset	33	calls: 14289	calls: 1.084	76400.7	768	[288]
Common yellowthroat	Unk.	syllables: 255	syllables: 0.1	35.4	255	[471, 470]
Cuvier's beajed whale	Unk.	clicks: 2237	clicks: 0.001	2.3	2237	[172, 128]
Egyptian fruit bat	83	syllables: 423043	syllables: 0.042	166676.8	83823	[348, 349]
European starling	7	syllables: 164230	syllables: 0.577	194529.9	3805	[16]
Gervais's beaked whale	Unk.	clicks: 1936	clicks: 0.001	2.0	1936	[172, 128]
Giant otter	Unk.	syllables: 452	syllables: 0.68	390.4	452	[310]
Gibbon	Unk.	syllables: 10333	syllables: 2.96	230400.0	128	[303]
Great blue heron	Unk.	syllables: 246	syllables: 0.138	44.1	246	[471, 470]
House finch	Unk.	syllables: 248	syllables: 0.093	25.9	248	[471, 470]
Human (English)	40	words: 283721 phones: 837896	words: 0.205 phones: 0.069	135708.4	254	[346]
Humpback whale	Unk.	syllables: 2006	syllables: 1.65	6730.8	13	[283]
Indigo bunting	Unk.	syllables: 251	syllables: 0.135	36.0	251	[471, 470]
Macaque	8	coos: 7284	coos: 0.324	2550.9	7284	[130, 131]
Marsh wren	Unk.	syllables: 248	syllables: 0.09	23.8	248	[471, 470]
Mouse	4	syllables: 34124	syllables: 0.018	25277.0	133	[62]
Song sparrow	Unk.	syllables: 258	syllables: 0.105	32.8	258	[471, 470]
Swamp sparrow	616	elements: 97513	elements: 0.021	4571.1	1867	[234, 232]
White-rumped munia	44	syllables: 109851	syllables: 0.05	17118.5	169	[194]
Yellow warbler	Unk.	syllables: 246	syllables: 0.078	21.4	246	[471, 470]
Zebra finch	6	motifs: 18028 syllables: 65892	motifs: 0.443 syllables: 0.105	8799.9	18028	[337]
Zebra finch	46	elements: 3347	elements: 0.153	1365.0	3347	[109, 110]

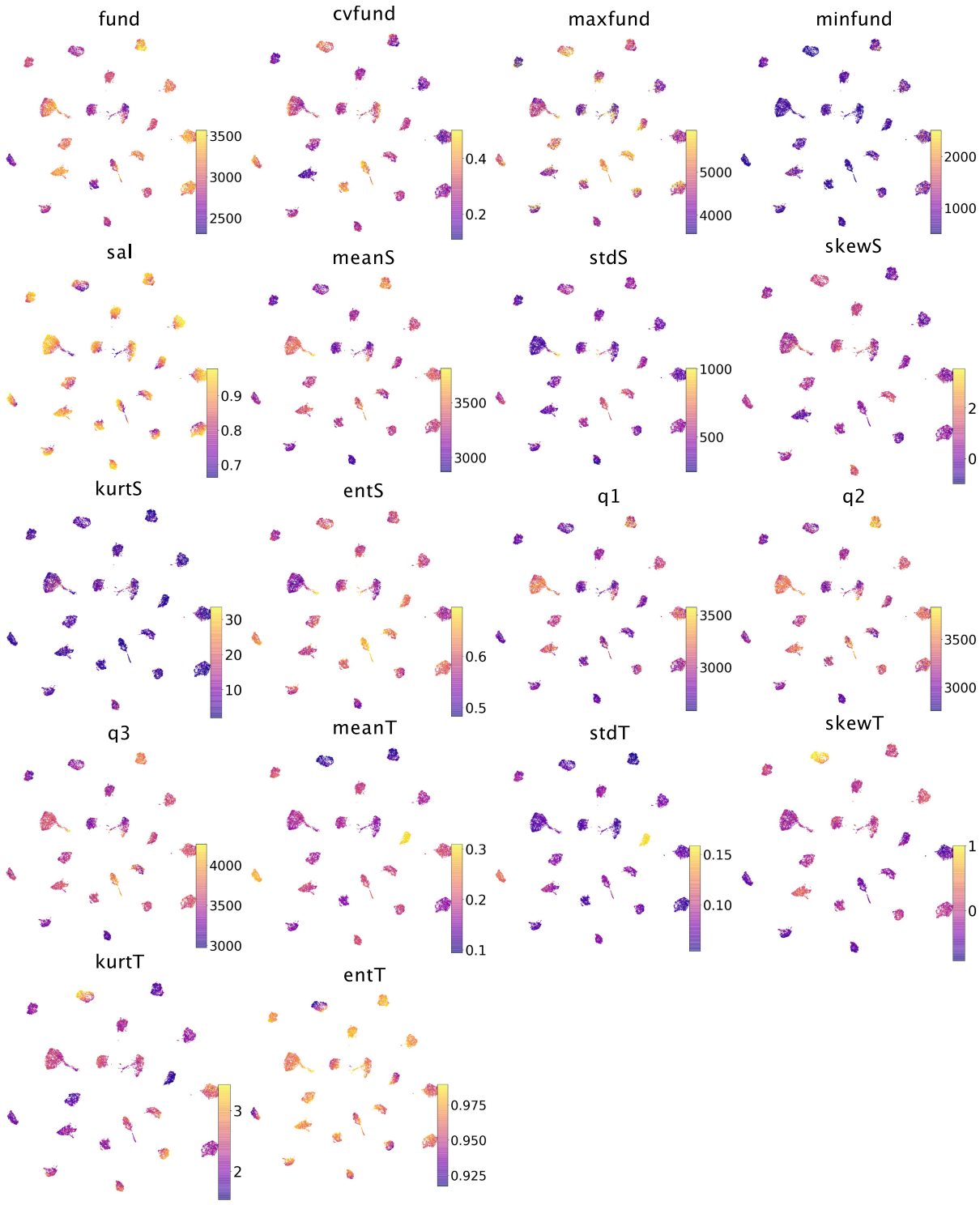


Figure 2.21. UMAP projections Cassin's vireo syllables with syllable features overlaid generated from the BioSound [110] python package. (A) More information regarding each feature can be found in Supporting information and Elie et al. [109, 110].

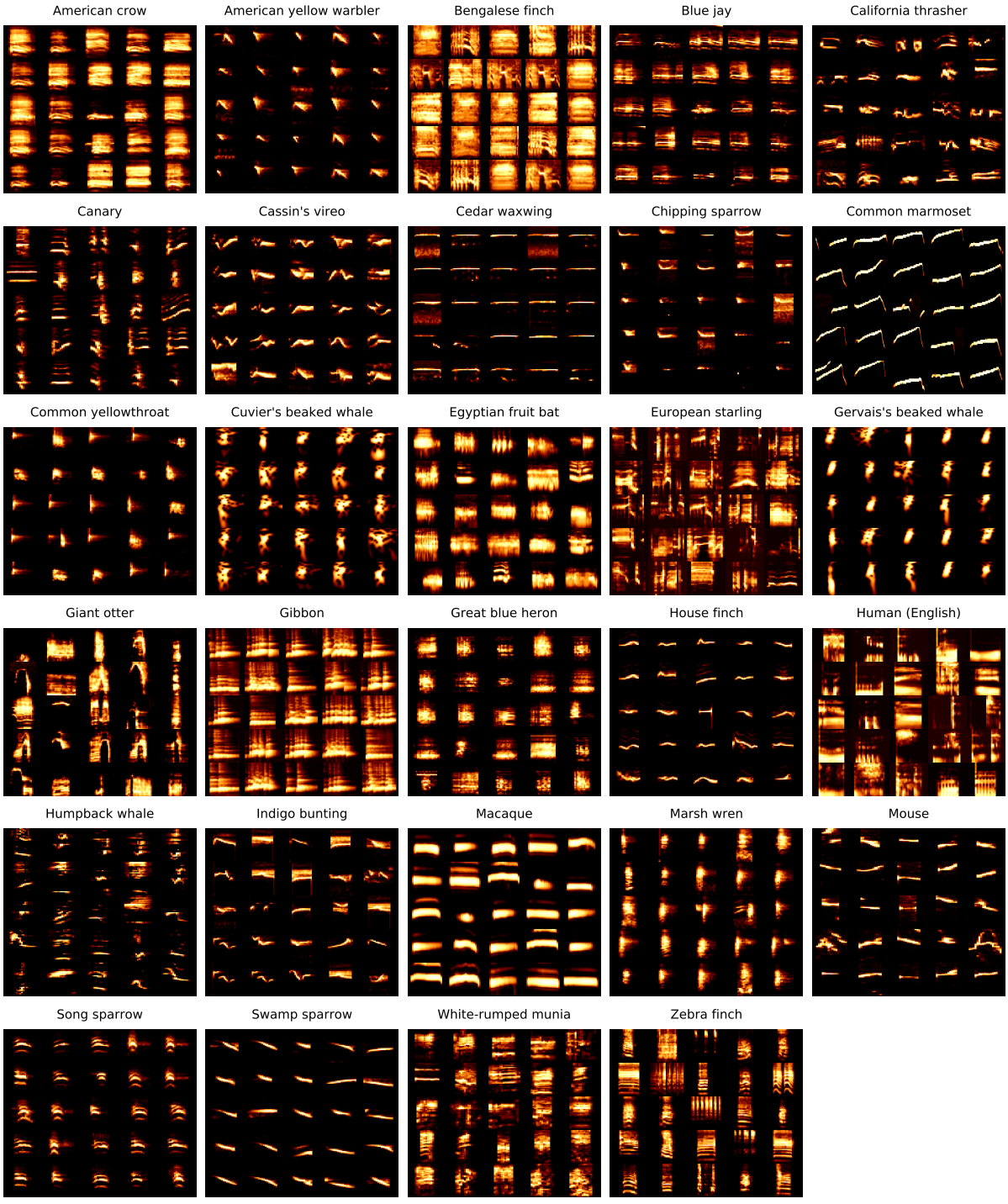


Figure 2.22. Example vocal elements from each of the species used in this paper.

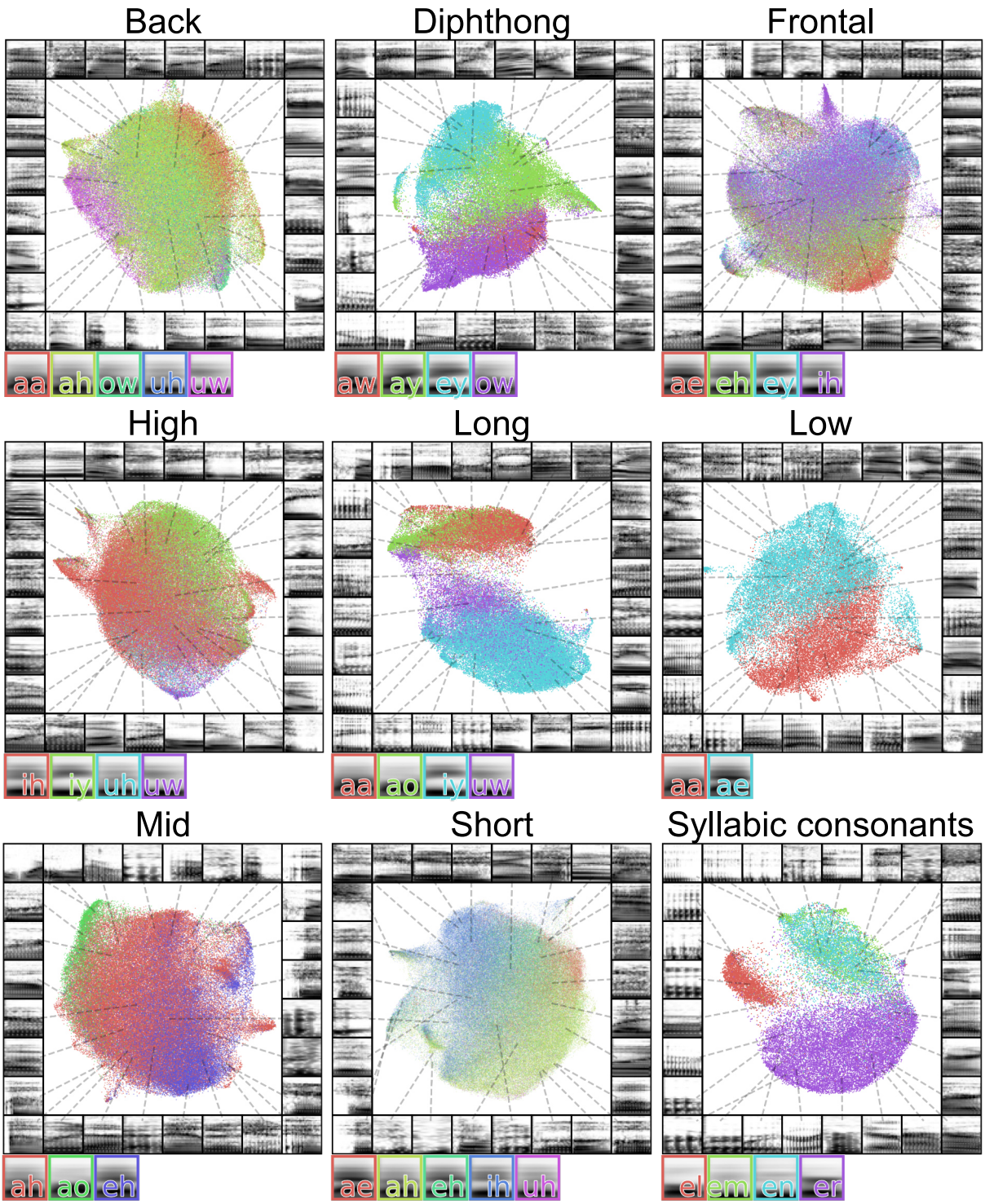


Figure 2.23. Latent projections of vowels. Each plot shows a different set of vowels grouped by phonetic features. The average spectrogram for each vowel is shown to the right of each plot.

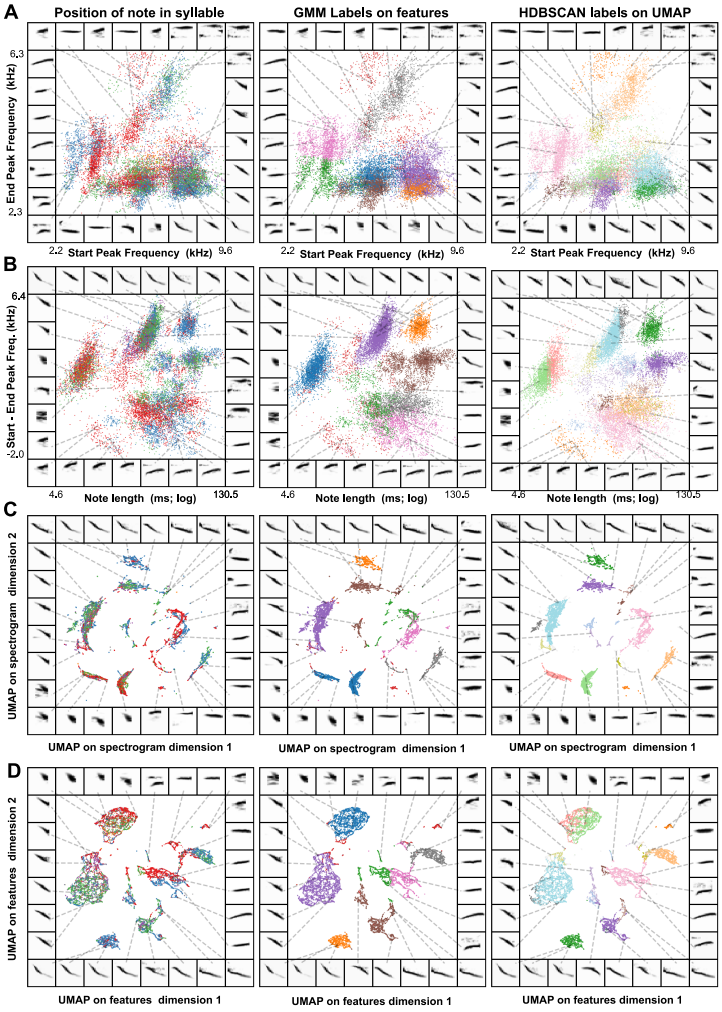


Figure 2.24. Comparing latent and known features in swamp sparrow song. (A) A scatterplot of the start and end peak frequencies of the notes produced by birds recorded in Hudson Valley, NY. The left panel shows notes colored by the position of each note in the syllable (red = first, blue = second, green = third). The center panel shows the sample scatterplot colored by a Gaussian Mixture Model labels (fit to the start and end peak frequencies and the note duration). The right panel shows the scatterplot colored by HDBSCAN labels over a UMAP projection of the spectrograms of notes. (B) The same notes, plotting the change in peak frequency over the note against the note's duration. (C) The same notes plotted as a UMAP projection over note-spectrograms. (D) The features from (A) and (B) projected together into a 2D UMAP space.

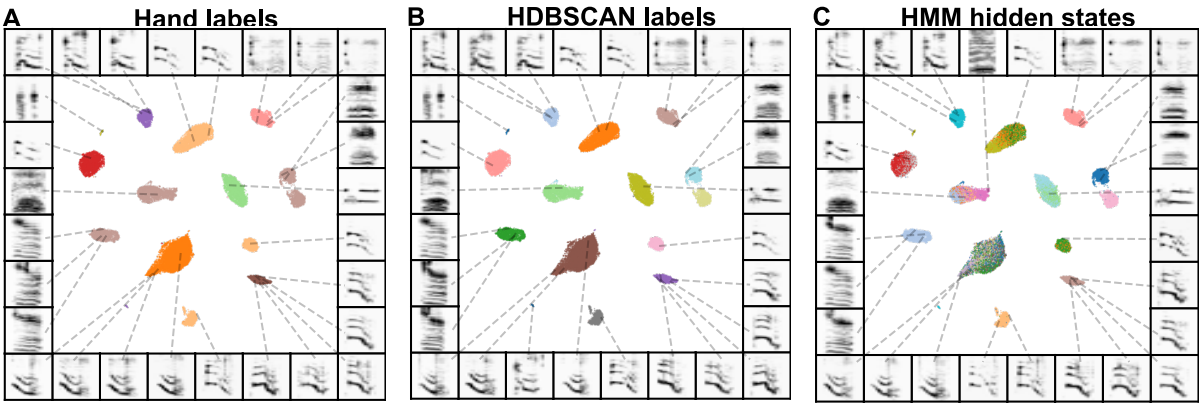


Figure 2.25. Comparison of Hidden Markov Model performance using different latent states. Projections are shown for a single example bird from the Nicholson dataset [322]. UMAP projections are labeled by three labeling schemes: (A) Hand labels, (B) HDBSCAN labels on UMAP, and (C) Trained Hidden Markov Model (HMM) labels.

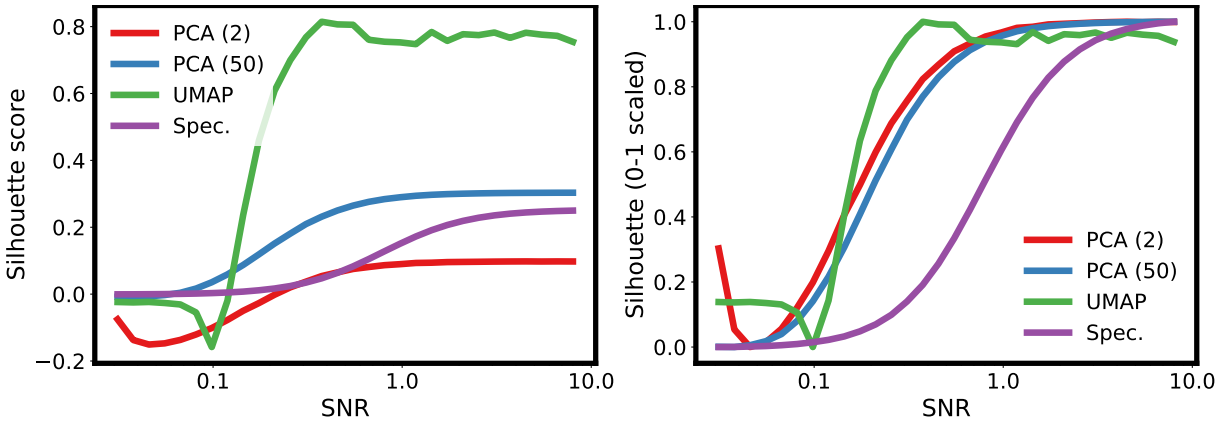


Figure 2.26. Silhouette score of UMAP projections with different levels background noise added to spectrogram. White noise is added to the spectrogram to modulate signal to noise ratio (SNR). The different projections (2-dimensional PCA, 50-dimensional PCA, 2-dimensional UMAP) and the spectrogram are compared on the basis of silhouette score for the labels of each Cassin's vireo syllable. The left panel shows the silhouette score, and the right panel shows the silhouette score scaled between 0 and 1 to more easily compare change as a function of SNR.

Table 2.3. BioSound features used in feature statistics analysis.

For more information see Elie et al. [109, 110].

variable name	feature	feature type
fund	Mean F0 (Hz)	f0 features
cvfund	Coeff. var. F0 (0-1)	f0 features
maxfund	Min. F0 (Hz)	f0 features
minfund	Max. F0 (Hz)	f0 features
sal	Pitch saliency	f0 features
meanS	Spectral mean (Hz)	spectral features
stdS	Spectral std. (Hz)	spectral features
skewS	Spectral skewness	spectral features
kurtS	Spectral Kurtosis	spectral features
entS	Spectral entropy (0-1)	spectral features
q1	Spectral Q1 (Hz)	spectral features
q2	Spectral Q2 (Hz)	spectral features
q3	Spectral Q3 (Hz)	spectral features
meanT	Mean time (ms)	temporal features
stdT	Time Std. (ms)	temporal features
skewT	Time Skewness	temporal features
kurtT	Time Kurtosis	temporal features
entT	Time entropy (0-1)	temporal features

2.6 Acknowledgments

Work supported by NSF GRF 2017216247 and an Annette Merle-Smith Fellowship to T.S. and NIH DC0164081 and DC018055 to T.Q.G. We additionally would like to thank Kyle McDonald and his colleagues for motivating some of our visualization techniques with their work on humpback whale song [277].

Chapter 2, in full, is a reprint of the material as it appears in PLOS Computational Biology, 2020, Sainburg, Tim, Thielk, Marvin, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Parametric UMAP

Abstract

UMAP is a non-parametric graph-based dimensionality reduction algorithm using applied Riemannian geometry and algebraic topology to find low-dimensional embeddings of structured data. The UMAP algorithm consists of two steps: (1) Compute a graphical representation of a dataset (fuzzy simplicial complex), and (2) Through stochastic gradient descent, optimize a low-dimensional embedding of the graph. Here, we extend the second step of UMAP to a parametric optimization over neural network weights, learning a parametric relationship between data and embedding. We first demonstrate that Parametric UMAP performs comparably to its non-parametric counterpart while conferring the benefit of a learned parametric mapping (e.g. fast online embeddings for new data). We then explore UMAP as a regularization, constraining the latent distribution of autoencoders, parametrically varying global structure preservation, and improving classifier accuracy for semi-supervised learning by capturing structure in unlabeled data.

3.1 Introduction

Current non-linear dimensionality reduction algorithms can be divided broadly into non-parametric algorithms which rely on the efficient computation of probabilistic relationships from neighborhood graphs to extract structure in large datasets (e.g. UMAP [280], t-SNE

[442], LargeVis [422]), and parametric algorithms, which, driven by advances in deep-learning, optimize an objective function related to capturing structure in a dataset over neural network weights (e.g. [173, 93, 94, 420, 205]).

In recent years, a number of parametric dimensionality reduction algorithms have been developed to wed these two classes of methods, learning a structured graphical representation of the data and using a deep neural network to capture that structure (discussed in 3.3). In particular, over the past decade, several variants of the t-SNE algorithm have proposed parameterized forms of t-SNE [440, 145, 53, 144]. Parametric t-SNE [440] for example, trains a deep neural network to minimize loss over a t-SNE graph. However, the t-SNE loss function itself is not well-suited to neural network training paradigms. In particular, t-SNE's optimization requires normalization over the entire dataset at each step of optimization, making batch-based optimization and online learning of large datasets difficult. In contrast, UMAP is optimized using negative sampling [287, 422] to sparsely sample edges during optimization, making it, in principle, more well-suited to batch-wise training as is common deep learning applications. Our proposed method, Parametric UMAP, brings the non-parametric graph-based dimensionality reduction algorithm UMAP into an emerging class of parametric topologically-inspired embedding algorithms.

In the following section, we broadly outline the algorithm underlying UMAP to explain why our proposed algorithm, Parametric UMAP, is particularly well suited to deep learning applications. We contextualize our discussion of UMAP in t-SNE, to outline the advantages that UMAP confers over t-SNE in the domain of parametric neural-network-based embedding. We then perform experiments comparing our algorithm, Parametric UMAP, to parametric and non-parametric algorithms. Finally, we show a novel extension of Parametric UMAP to semi-supervised learning.

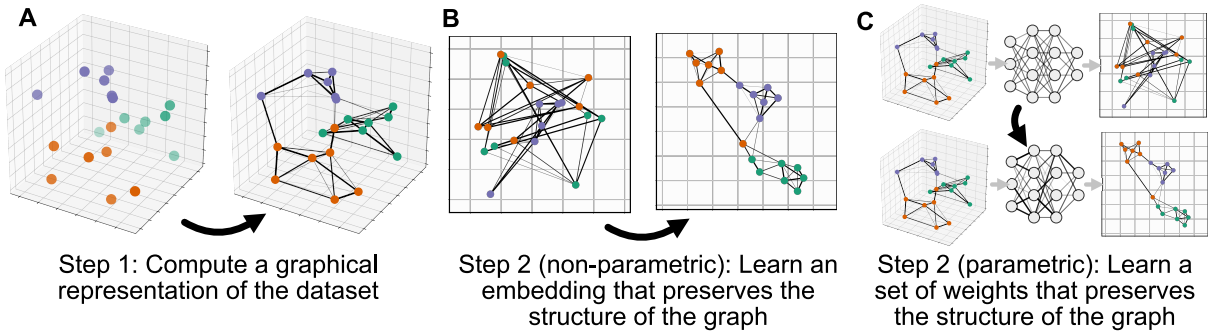


Figure 3.1. Overview of UMAP (A $\rightarrow$ B) and Parametric UMAP (A $\rightarrow$ C).

3.2 Parametric and non-parametric UMAP

UMAP and t-SNE have the same goal: Given a D-dimensional data set $\mathbf{X} \in \mathbb{R}^D$, produce a d dimensional embedding $Z \in \mathbb{R}^d$ such that points that are close together in X (e.g. x_i and x_j) are also close together in Z (z_i and z_j).

Both algorithms are comprised of the same two broad steps: first construct a graph of local relationships between datasets (Figure 3.1A), then optimize an embedding in low dimensional space which preserves the structure of the graph (Figure 3.1B). The parametric approach replaces the second step of this process with an optimization of the parameters of a deep neural network over batches (Figure 3.1C). To understand how Parametric UMAP is optimized, it is necessary to understand these steps.

3.2.1 Graph Construction

Computing probabilities in X

The first step in both UMAP and t-SNE is to compute a distribution of probabilities P between pairs of points in X based upon the distances between points in data space. Probabilities are initially computed as local, one-directional, probabilities between a point and its neighbors in data-space, then symmetrized to yield a final probability representing the relationship between pairs of points.

In t-SNE, these probabilities are treated as conditional probabilities of neighborhood

($p^{\text{t-SNE}}_{i|j}$) computed using a Gaussian distribution centered at x_i.

$$p^{\text{t-SNE}}_{j|i} = \frac{\exp\left(-\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)/2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\mathrm{d}(\mathbf{x}_i, \mathbf{x}_k)/2\sigma_i^2\right)} \tag{3.1}$$

Where $\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)$ represents the distance between x_i an x_j (e.g. Euclidean distance) and σ_i is the standard deviation for the Gaussian distribution, set based upon the a perplexity parameter such that one standard deviation of the Gaussian kernel fits a a set number of nearest-neighbors in X.

In UMAP, local, one-directional, probabilities ($P^{\text{UMAP}}_{i|j}$) are computed between a point and its neighbors to determine the probability with which an edge (or simplex) exists, based upon an assumption that data is uniformly distributed across a manifold in a warped dataspace. Under this assumption, a local notion of distance is set by the distance to the k^{th} nearest neighbor and the local probability is scaled by that local notion of distance.

$$p^{\text{UMAP}}_{j|i} = \exp(-(\mathrm{d}(\mathbf{x}_i, \mathbf{x}_j) - \rho_i)/\sigma_i) \tag{3.2}$$

Where ρ_i is a local connectivity parameter set to the distance from x_i to its nearest neighbor, and σ_i is a local connectivity parameter set to match the local distance around x_i upon its k nearest neighbors (where k is a hyperparameter).

After computing the one-directional edge probabilities for each datapoint, UMAP computes a global probability as the probability of either of the two local, one-directional, probabilities occurring:

$$p^{\text{UMAP}}_{ij} = \left(p_{j|i} + p_{i|j}\right) - p_{j|i}p_{i|j} \tag{3.3}$$

In contrast, t-SNE symmetrizes the conditional probabilities as

$$p^{\text{t-SNE}}_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{3.4}$$

3.2.2 Graph Embedding

After constructing a distribution of probabilistically weighted edges between points in X, UMAP and t-SNE initialize an embedding in Z corresponding to each data point, where a probability distribution (Q) is computed between points as was done with the distribution (P) in the input space. The objective of UMAP and t-SNE is then to optimize that embedding to minimize the difference between P and Q.

Computing probabilities in Z

In embedding space, the pairwise probabilities are computed directly without first computing local, one-directional probabilities.

In the t-SNE embedding space, the pairwise probability between two points $q_{i|j}^{\text{t-SNE}}$ is computed in a similar manner to $p_{i|j}^{\text{t-SNE}}$, but where the Gaussian distribution is replaced with the fatter-tailed Student's t-distribution (with one degree of freedom), which is used to overcome the 'crowding problem' [442] in translating volume differences in high-dimensional spaces to low-dimensional spaces:

$$q_{ij}^{\text{t-SNE}} = \frac{\left(1 + \left\|\mathbf{z}_i - \mathbf{z}_j\right\|^2\right)^{-1}}{\sum_{k \neq l}\left(1 + \left\|\mathbf{z}_k - \mathbf{z}_l\right\|^2\right)^{-1}} \tag{3.5}$$

UMAP's computation of the pairwise probability q_{ij}^{UMAP} between points in the embedding space Z uses a different family of functions:

$$q_{ij}^{\text{UMAP}} = \left(1 + a\left\|z_i - z_j\right\|^{2b}\right)^{-1} \tag{3.6}$$

Where a and b are hyperparameters set based upon a desired minimum distance between points in embedding space. Notably, the UMAP probability distribution in embedding space is not normalized, while the t-SNE distribution is normalized across the entire distribution of probabilities, meaning that the entire distribution of probabilities needs to be calculated before

each optimization step of t-SNE.

Cost function

Finally, the distribution of embeddings in Z is optimized to minimize the difference between Q and P.

In t-SNE, a Kullback-Leibler divergence between the two probability distributions is used, and gradient descent in t-SNE is computed over the embeddings:

$$C_{\text{t-SNE}} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.7}$$

In UMAP, the cost function is cross-entropy, also optimized using gradient descent:

$$C_{\text{UMAP}} = \sum_{i \neq j} p_{ij} \log \left(\frac{p_{ij}}{q_{ij}} \right) + \left(1 - p_{ij}\right) \log \left(\frac{1 - p_{ij}}{1 - q_{ij}} \right) \tag{3.8}$$

3.2.3 Attraction and repulsion

Minimizing the cost function over every possible pair of points in the dataset would be computationally expensive. UMAP and more recent variants of t-SNE both use shortcuts to bypass much of that computation. In UMAP, those shortcuts are directly advantageous to batch-wise training in a neural network.

The primary intuition behind these shortcuts is that the cost function of both t-SNE and UMAP can both be broken out into a mixture of attractive forces between locally connected embeddings and repulsive forces between non-locally connected embeddings.

Attractive forces

Both UMAP and t-SNE utilize a similar strategy in minimizing the computational cost over attractive forces: they rely on an approximate nearest neighbors graph[1]. The intuition

[1]UMAP requires substantially fewer nearest neighbors than t-SNE, which generally requires 3 times the perplexity hyperparameter (defaulted at 30 here), whereas UMAP computes only 15 neighbors by default, which is computationally less costly.

for this approach is that elements that are further apart in data space have very small edge probabilities, which can be treated as effectively zero. Thus, edge probabilities and attractive forces only need to be computed over the nearest neighbors, non-nearest neighbors can be treated as having an edge probability of zero. Because nearest-neighbor graphs are themselves computationally expensive, approximate nearest neighbors (e.g. [97]) produce effectively similar results.

Repulsive forces

Because most datapoints are not locally connected, we do not need to waste computation on most pairs of embeddings.

UMAP takes a shortcut motivated by the language model word2vec [287] and performs negative sampling over embeddings. Each training step iterates over positive, locally connected, edges and randomly samples edges from the remainder of the dataset treating their edge probabilities as zero to compute cross-entropy. Because most datapoints are not locally connected and have a very low edge probability, these negative samples are, on average, correct, allowing UMAP to sample only sparsely over edges in the dataset.

In t-SNE, repulsion is derived from the normalization of Q. A few methods for minimizing the amount of computation needed for repulsion have been developed. The first is the Barnes-Hut tree algorithm [441], which bins the embedding space into cells and where repulsive forces can be computed over cells rather than individual datapoints within those cells. Similarly, the more recent interpolation-based t-SNE (FIt-SNE; [251, 252]) divides the embedding space up into a grid and computes repulsive forces over the grid, rather than the full set of embeddings.

3.2.4 Parametric UMAP

To summarize, both t-SNE and UMAP rely on the construction of a graph, and a subsequent embedding that preserves the structure of that graph (Fig. 3.1). UMAP learns an embedding by minimizing cross-entropy sampled over positively weighted edges (attraction)

and using negative sampling randomly over the dataset (repulsion), allowing minimization to occur over sampled batches of the dataset. t-SNE, meanwhile, minimizes a KL divergence loss function normalized over the entire set of embeddings in the dataset using different approximation techniques to compute attractive and repulsive forces.

Because t-SNE optimization requires normalization over the distribution of embedding in projection space, gradient descent can only be performed after computing edge probabilities over the entire dataset. Projecting an entire dataset into a neural network between each gradient descent step would be too computationally expensive to optimize, however. The trick that Parametric t-SNE proposes for this problem is to split the dataset up into large batches (e.g. 5000 data points in the original paper) that are used to compute separate graphs that are independently normalized over and used constantly throughout training, meaning that relationships between elements in different batches are not explicitly preserved. Conversely, a parametric form of UMAP, by using negative sampling, can in principle be trained on batch sizes as small as a single edge, making it suitable for minibatch training needed for memory-expensive neural networks trained on the full graph over large datasets as well as online learning.

Given these design features, UMAP loss can be applied as a regularization in typical stochastic gradient descent deep learning paradigms, without requiring the batching trick that Parametric T-SNE relies upon. Despite this, a parametric extension to the UMAP learning algorithm has not yet been explored. Here, we explore the performance of a parametric extension to UMAP relative to current embedding algorithms and perform several experiments further extending Parametric UMAP to novel applications [2].

3.3 Related Work

Beyond Parametric t-SNE and Parametric UMAP, a number of recent parametric dimensionality reduction algorithms utilizing structure-preserving constraints exist which were

[2] See code implementations: Experiments https://github.com/timsainb/ParametricUMAP_paper Python package https://github.com/lmcinnes/umap

not compared here. This work is relevant to ours and is mentioned here to provide clarity on the current state of parametric topologically motivated and structure-preserving dimensionality reduction algorithms.

Moor et al., (topological autoencoders; [301]) and Hoffer et al. (Connectivity-Optimized Representation Learning; [174]) apply an additional topological structure-preserving loss using persistent homology over mini-batches to the latent space of an autoencoder. Jia et al., (Laplacian Autoencoders; [185]) similarly defines an autoencoder with a local structure preserving regularization. Mishne et al., (Diffusion Nets; [291]) define an autoencoder extension based upon diffusion maps that constrains the latent space of the autoencoder. Ding et al., (scvis; [93]) and Graving and Couzin (VAE-SNE; [154]) describe VAE-derived dimensionality reduction algorithms based upon the ELBO objective. Duque et al (geometry-regularized autoencoders; [104]) regularize an autoencoder with the PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding) embedding algorithm [300]. Szubert et al (ivis; [420]) and Robinson (Differential Embedding Networks; [367]) make use of Siamese neural network architectures with structure-preserving loss functions to learn embeddings. Pai et al., (DIMAL; [331]) similarly uses Siamese networks constrained to preserve geodesic distances for dimensionality reduction. Several of these parametric approaches indirectly condition neural networks (e.g. autoencoders) on non-parametric embeddings rather than directly upon the loss of the algorithm, which can be applied to arbitrary embedding algorithms. We contrast indirect and direct parametric embeddings in 3.5.6.

3.4 UMAP as a regularization

In machine learning, regularization refers to the modification of a learning algorithm to improve generalization to new data. Here, we consider both regularizing neural networks with UMAP loss, as well as using additional loss functions to regularize the embedding that UMAP learns. While non-parametric UMAP optimizes UMAP loss directly over embeddings

(Figure 3.2A), our proposed algorithm, Parametric UMAP, applies the same cost function over an encoder network (Figure 2B). By applying additional losses, we can use both regularize UMAP with, as well as use UMAP to, regularize additional training objectives, which we outline below.

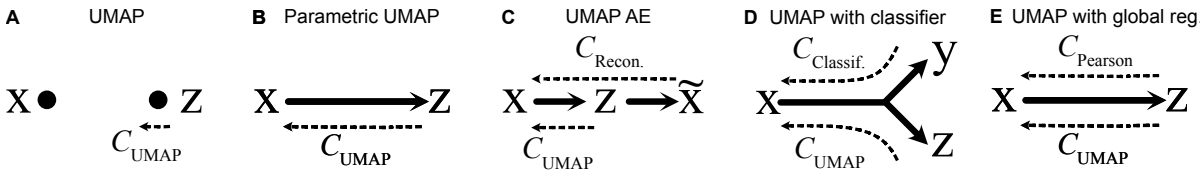


Figure 3.2. Varients of UMAP used in this paper. Solid lines represent neural networks. Dashed lines represent error gradients.

3.4.1 Autoencoding with UMAP

AEs are by themselves a powerful dimensionality reduction algorithm [173]. Thus, combining them with UMAP may yield additional benefits in capturing latent structure. We used an autoencoder as an additional regularization to Parametric UMAP (Figure 3.2C). A UMAP/AE hybrid is simply the combination of the UMAP loss and a reconstruction loss, both applied over the network. VAEs have similarly been used in conjunction with Parametric t-SNE for capturing structure in animal behavioral data [154] and combining t-SNE, which similarly emphasizes local structure, with AEs aids in capturing more global structure over the dataset [442, 154].

3.4.2 Semi-supervised learning

Parametric UMAP can be used to regularize supervised classifier networks, training the network on a combination of labeled data with the classifier loss and unlabeled data with UMAP loss (Figure 3.2D). Semi-supervised learning refers to the use of unlabeled data to jointly learn the structure of a dataset while labeled data is used to optimize the supervised objective function, such as classifying images. Here, we explore how UMAP can be jointly trained as an objective function in a deep neural network alongside a classifier.

In the example in Figure 3.3, we show an intuitive example of semi-supervised learning using UMAP over the Moons dataset [339]. By training a Y-shaped network (Figure 3.2D) both

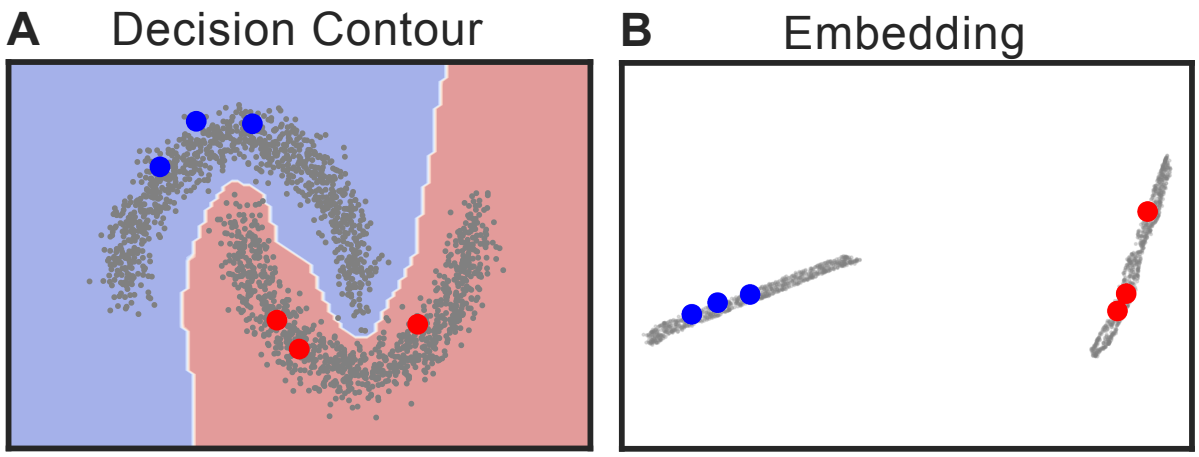


Figure 3.3. An example of semi-supervised learning with UMAP on the moons dataset.

on the classifier loss over labeled datapoints (Figure 3.3A, red and blue) and the UMAP loss over unlabeled datapoints (Figure 3.3A, grey) jointly, the shared latent space between the UMAP and classifier network pulls apart the two moons (Figure 3.3B), resulting in a decision boundary that divides cleanly between the two distributions in dataspace.

3.4.3 Preserving global structure

An open issue in dimensionality reduction is how to balance local and global structure preservation [88, 24, 209]. Algorithms that rely on sparse nearest neighbor graphs, like UMAP, focus on capturing the local structure present between points and their nearest neighbors, while global algorithms, like Multi-Dimensional Scaling (MDS), attempt to preserve all relationships during embedding. Local algorithms are both computationally more efficient and capture structure that is lost in global algorithms (e.g. the clusters corresponding to numbers found when projecting MNIST into UMAP). While local structure preservation captures more application-relevant structure in many datasets, however, the ability to additionally capture global structure is still often desirable. The approach used by non-parametric t-SNE and UMAP is to initialize embeddings with global structure-preserving embeddings such as PCA or Laplacian eigenmaps embeddings. In Parametric UMAP, we explore a different tactic, imposing global structure by jointly training on a global structure preservation loss directly.

3.5 Experiments

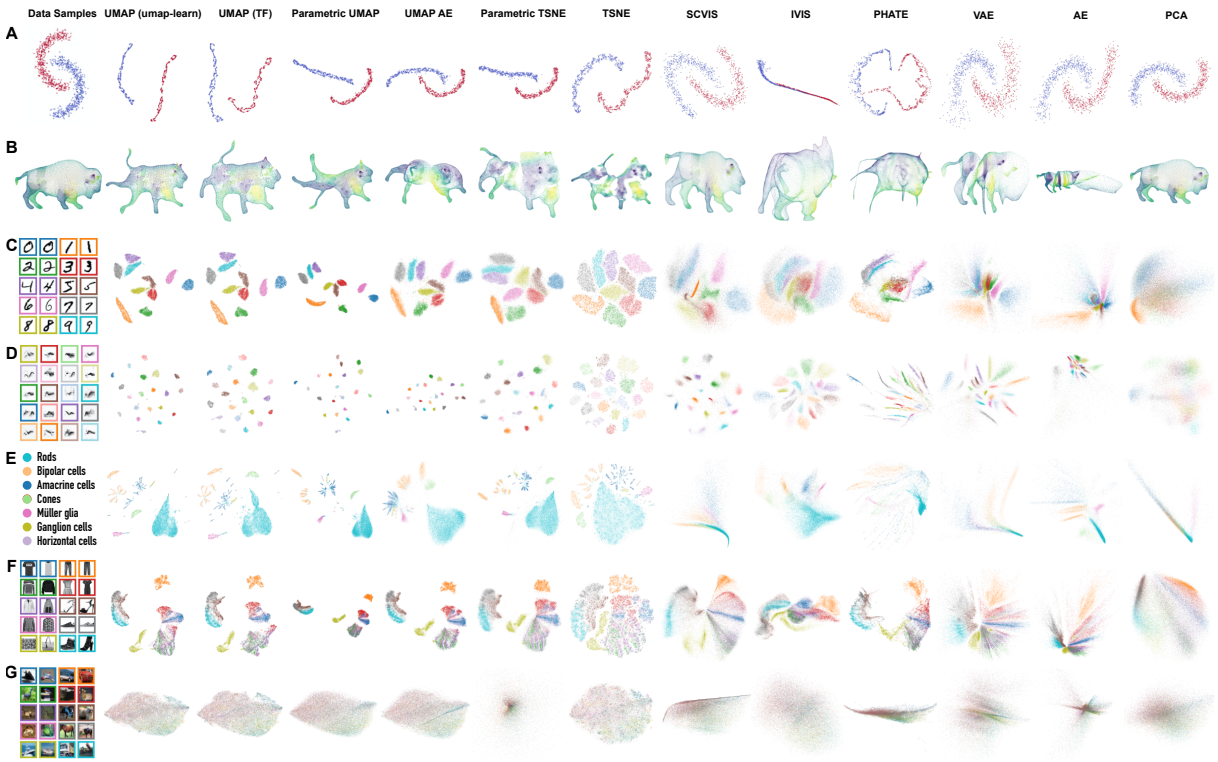


Figure 3.4. Comparison of projections from multiple datasets using UMAP, UMAP in Tensorflow, Parametric UMAP, Parametric UMAP with an Autoencoder loss, Parametric t-SNE, t-SNE, SCVIS, IVIS, PHATE, a VAE, an AE, and PCA. (a) Moons. (B) 3D buffalo. (c) MNIST (d) Cassin's vireo song segments (e) Mouse retina single-cell transcriptomes. (f) Fashion MNIST (g) CIFAR10. The Cassin's vireo dataset uses a dynamic time warping loss and an LSTM network for the encoder and decoder for the neural networks. The image datasets use a convnet for the encoder and decoder for the neural networks. The bison examples use a t-SNE perplexity of 500 and 150 nearest neighbors in UMAP to capture more global structure.

Experiments were performed comparing Parametric UMAP and a UMAP/AE hybrid, to several baselines: nonparametric UMAP, nonparametric t-SNE (FIt-SNE) [252, 347], Parametric t-SNE, an AE, a VAE, and PCA projections. As additional baselines, we compared PHATE (non-parametric), SCVIS (parametric), and IVIS (parametric) which are described in the related works section (3.3). We also compare a second non-parametric UMAP implementation that has the same underlying code as Parametric UMAP, but where optimization is performed over embeddings directly, rather than neural network weights. This comparison is made to provide

a bridge between the UMAP-learn implementation and parametric UMAP, to control for any potential implementation differences. Parametric t-SNE, Parametric UMAP, the AE, VAE, and the UMAP/AE hybrid use the same neural network architectures and optimizers within each dataset (described in Supplemental Materials).

We used the common machine learning benchmark datasets MNIST, FMNIST, and CIFAR10 alongside two real-world datasets in areas where UMAP has proven a useful tool for dimensionality reduction: a single-cell retinal transcriptome dataset [262], and a bioacoustic dataset of Cassin's vireo song, recorded in the Sierra Nevada mountains [165, 166].

3.5.1 Embeddings

We first confirm that Parametric UMAP produces embeddings that are of a similar quality to non-parametric UMAP. To quantitatively measure the quality of embeddings we compared embedding algorithms on several metrics across datasets. We compared each method/dataset on 2D and 64D projections[3]. Each metric is explained in detail in the Supplemental Materials. The 2D projection of each dataset/method is shown in Figure 3.4. The results are given in Supplementary Figures 1-6 and Supplementary Tables 2-7, and summarized below.

Trustworthiness

Trustworthiness (Supplementary Equation 1, [445]) is a measure of how much of the local structure of a dataset is preserved in a set of embeddings. In 2D, we observe each of the UMAP algorithms performs similarly in Trustworthiness, with t-SNE being slightly more trustworthy in each dataset (Figure 3.5, Supplementary Figure 1; Supplementary Table 2. At 64D, PCA, AE, VAE, and Parametric t-SNE are most trustworthy in comparison to each UMAP implementation, possibly reflecting the more approximate repulsion (negative sampling) used by UMAP.

[3]Where possible. In contrast with UMAP, Parametric UMAP, and Parametric t-SNE, Barnes Huts t-SNE can only embed in two or three dimensions [441] and while FIt-SNE can in principle scale to higher dimensions [252], embedding in more than 2 dimensions is unsupported in both the official implementation [206] and openTSNE [347]

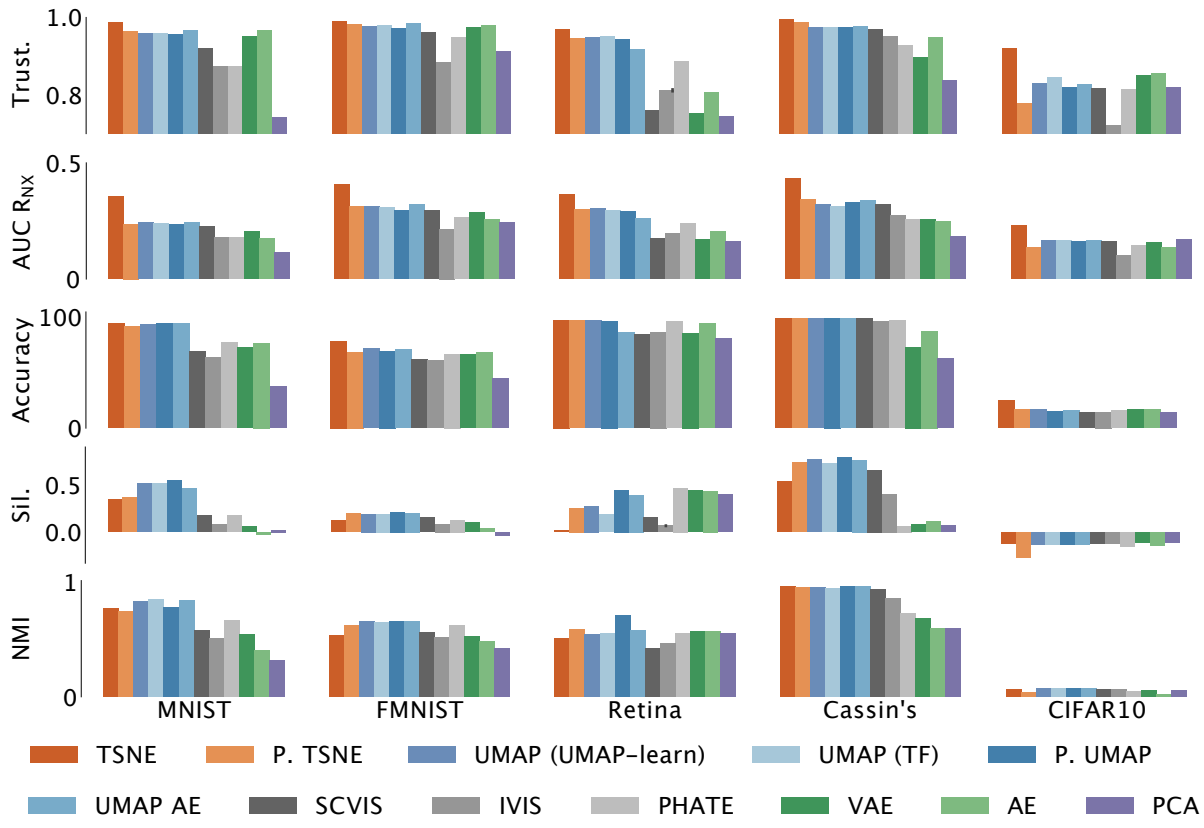


Figure 3.5. Embedding metrics for 2D projections. Full results are given in the Appendix. Accuracy is shown for KNN (k=1).

Area Under the Curve (AUC) of R_{NX}

To compare embeddings across scales (both local and global neighborhoods), we computed the AUC of R_{NX} for each embedding [241], which captures the agreement across K-ary neighborhoods, weighting nearest-neighbors as more important than further neighbors. In 2D we find that Parametric and non-parametric UMAP perform similarly while t-SNE has the highest AUC. At 64D, Parametric and non-parametric UMAP again perform similarly, with PCA having the highest AUC.

KNN-Classifier

A KNN-classifier is used as a baseline to measure supervised classification accuracy based upon local relationships in embeddings. We find KNN-classifier performance largely reflects Trustworthiness (Figure 3.5, Supplementary Figures 3,4; Supplementary Tables 4,5). In 2D, we observe a broadly similar performance between UMAP and t-SNE variants, each of which is substantially better than the PCA, AE, or VAE projections. At 64 dimensions UMAP projections are similar but in some datasets (FMNIST, CIFAR10) slightly under-performs PCA, AE, VAE, and Parametric t-SNE.

Silhouette score

Silhouette score measures how clustered a set of embeddings are given ground truth labels. In 2D, across datasets, we tend to see a better silhouette score for UMAP and Parametric UMAP projections than t-SNE and Parametric t-SNE, which are in turn more clustered than PCA in all cases but CIFAR10, which shows little difference from PCA (Figure 3.5, Supplementary Figure 5; Supplementary Table 5). The clustering of each dataset can also be observed in Figure 3.4, where t-SNE and Parametric t-SNE are more spread out within-cluster than UMAP. In 64D projections, we find the silhouette score of Parametric t-SNE is near or below that of PCA, which is lower than UMAP-based methods. We note, however, that the poor performance of Parametric t-SNE may reflect setting the degrees-of-freedom (α) at $d-1$ which is only one of three parameterization schemes that [440] suggests. A learned degrees-of-freedom parameter

might improve performance for parametric t-SNE at higher dimensions.

Clustering

To compare clustering directly across embeddings, we performed k-Means clustering over each latent projection and compared each embedding's clustering on the basis of the normalized mutual information (NMI) between clustering schemes (Figure 3.5, Supplemental Figure 6; Supplementary Table 6. In both the 2D and 64D projections, we find that NMI corresponds closely to the silhouette score. UMAP and t-SNE show comparable clustering in 2D, both well above PCA in most datasets. At 64D, each UMAP approach shows superior performance over t-SNE.

3.5.2 Training and embedding speed

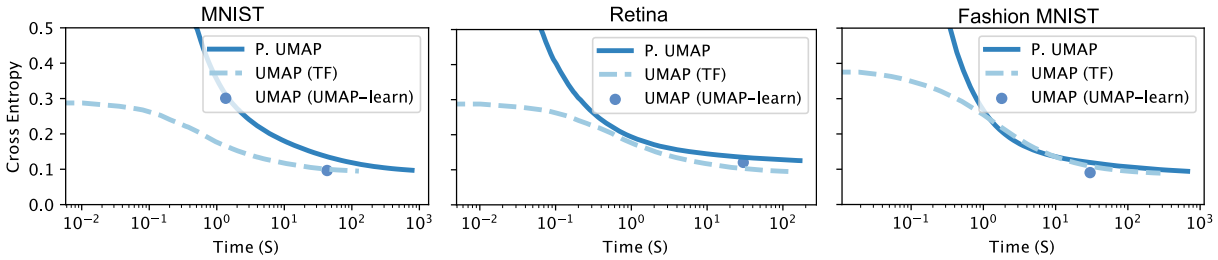


Figure 3.6. Training times comparison between UMAP and Parametric UMAP. All results were obtained with up to 32 threads on a machine with 2 AMD EPYC Rome 7252 8-Core CPU running at 3.1 GHz and a Quadro RTX 6000.

Training speed

Optimization in non-parametric UMAP is not influenced by the dimensionality of the original dataset; the dimensionality of the dataset only comes into play in computing the nearest-neighbors graph. In contrast, training speeds for Parametric UMAP are variable based upon the dimensionality of data and the architecture of the neural network used. The dimensionality of the embedding does not have a substantial effect on speed. In Figure 3.6, we show the cross-entropy loss over time for Parametric and non-parametric UMAP, for the MNIST, Fashion MNIST, and Retina datasets. Across each dataset, we find that non-parametric UMAP reaches a lower loss

more quickly than Parametric UMAP, but that Parametric UMAP reaches a similar cross-entropy within an order of magnitude of time. Thus, Parametric UMAP can train more slowly than non-parametric UMAP, but training times remain within a similar range making Parametric UMAP a reasonable alternative to non-parametric UMAP in terms of training time.

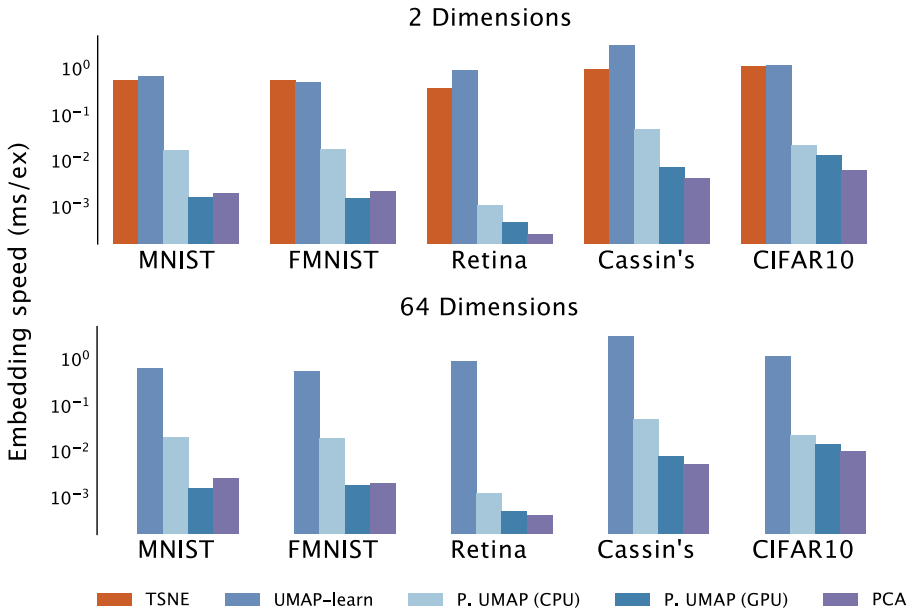


Figure 3.7. Comparison of embedding speeds using parametric UMAP and other embedding algorithms on a held-out testing dataset. Embeddings were performed on the same machine as Figure 3.6. Values shown are the median times over 10 runs.

Embedding and reconstruction speed

A parametric mapping allows embeddings to be inferred directly from data, resulting in a quicker embedding than non-parametric methods. The speed of embedding is especially important in signal processing paradigms where near-real-time embedding speeds are necessary. For example in brain-machine interfacing, bioacoustics, and computational ethology, fast embedding methods like PCA or deep neural networks are necessary for real-time analyses and manipulations, thus deep neural networks are increasingly being used (e.g. [333, 46, 383]). Here, we compare the embedding speed of a held-out test sample for each dataset, as well as the speed of reconstruction of the same held-out test samples.

Broadly, we observe similar embedding times for the non-parametric t-SNE and UMAP methods, which are several orders of magnitude slower than the parametric methods, where embeddings are direct projections into the learned networks (Figure 3.7). Because the same neural networks are used across the different parametric UMAP and t-SNE methods, we show only Parametric UMAP in Figure 3.7, which is only slightly slower than PCA, making it a viable candidate for fast embedding where PCA is currently used. Similarly, we compared parametric and non-parametric UMAP reconstruction speeds (Supplemental Figure 7). With the network architectures we used, reconstructions of Parametric UMAP are orders of magnitude faster than non-parametric UMAP, and slightly slower, but within the same order of magnitude, as PCA.

3.5.3 Capturing additional global structure in data

To capture additional global structure we added a naïve global structure preservation loss to Parametric UMAP, maximizing the Pearson correlation within batches between pairwise distances in embedding and data spaces:

$$C_{\text{Pearson}} = -\frac{\text{cov}(d_X, d_Z)}{\sigma_{d_X}\sigma_{d_Z}} \tag{3.9}$$

Where $\text{cov}(X,Y)$ is the covariance of data and embeddings, and σ_X and σ_Z are the standard deviations of the data and embeddings. The same notion of pairwise distance correlation has previously been used directly as a metric for global structure preservation [209, 24].

The weight of this additional loss can be used to dictate the balance between capturing global and local structure in the dataset. In Figure 3.8, we apply this loss at four different weights, ranging from only UMAP (left) to primarily global correlation (right). As expected, we observe that as we weight C_{Pearson} more heavily, the global correlation (measured as the correlation of the distance between pairs of points in embedding and data space) increases (indicated in each panel). Notably, when a small weight is used with each dataset, local structure is largely preserved while substantially improving global correlation.

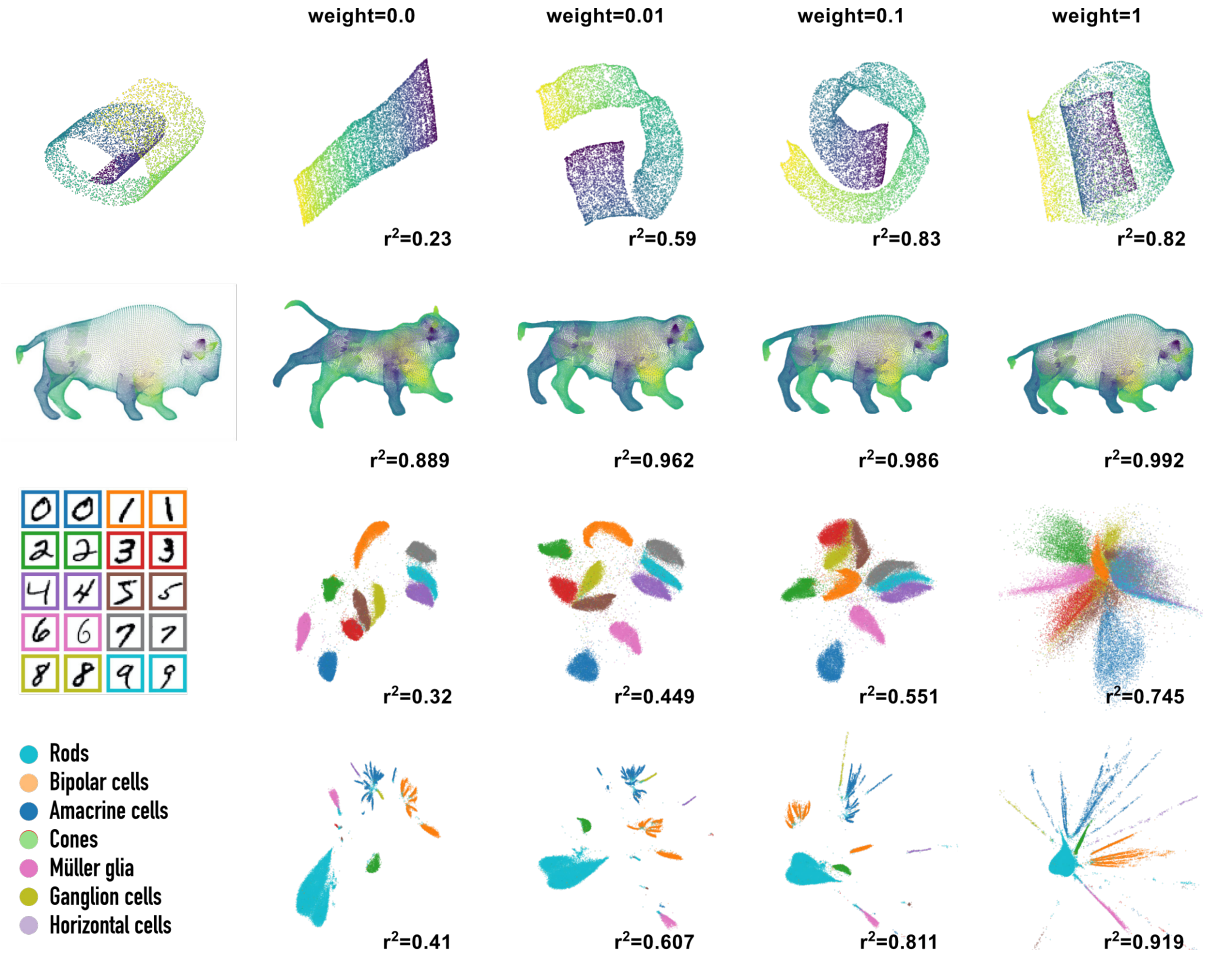


Figure 3.8. Global loss applied to Parametric UMAP embeddings with different weights. r^2 is the correlation between pairwise distances in data space and embedding space.

In Figure 3.9, we show the global distance correlation plotted against two local structure metrics (Silhouette score and Trustworthiness) for the MNIST and Macosko et al., [262] datasets corresponding to the projections shown in Figure 3.8 in relation to each embedding from Figure 3.4. In addition, we compared TriMap [8], a triplet-loss-based embedding algorithm designed to capture additional global structure by preserving the relative distances of triplets of data samples. We also compared Minimum Distortion Embedding (MDE), which comprises two separate embedding functions: a local embedding algorithm that preserves relationships between neighbors similar to UMAP and t-SNE, and a global embedding algorithm that preserves pairwise distances similar to MDS.

Broadly, with Parametric UMAP, we can observe the tradeoff between captured global and local structure with the weight of C_{Pearson} (light blue line in each panel of Figure 3.9). We observe that adding this loss can increase the amount of global structure captured while preserving much of the local structure, as indicated by the distance to the top right corner of each panel in Figure 3.9, which reflects the simultaneous capture of global and local relationships, relative to each other embedding algorithm.

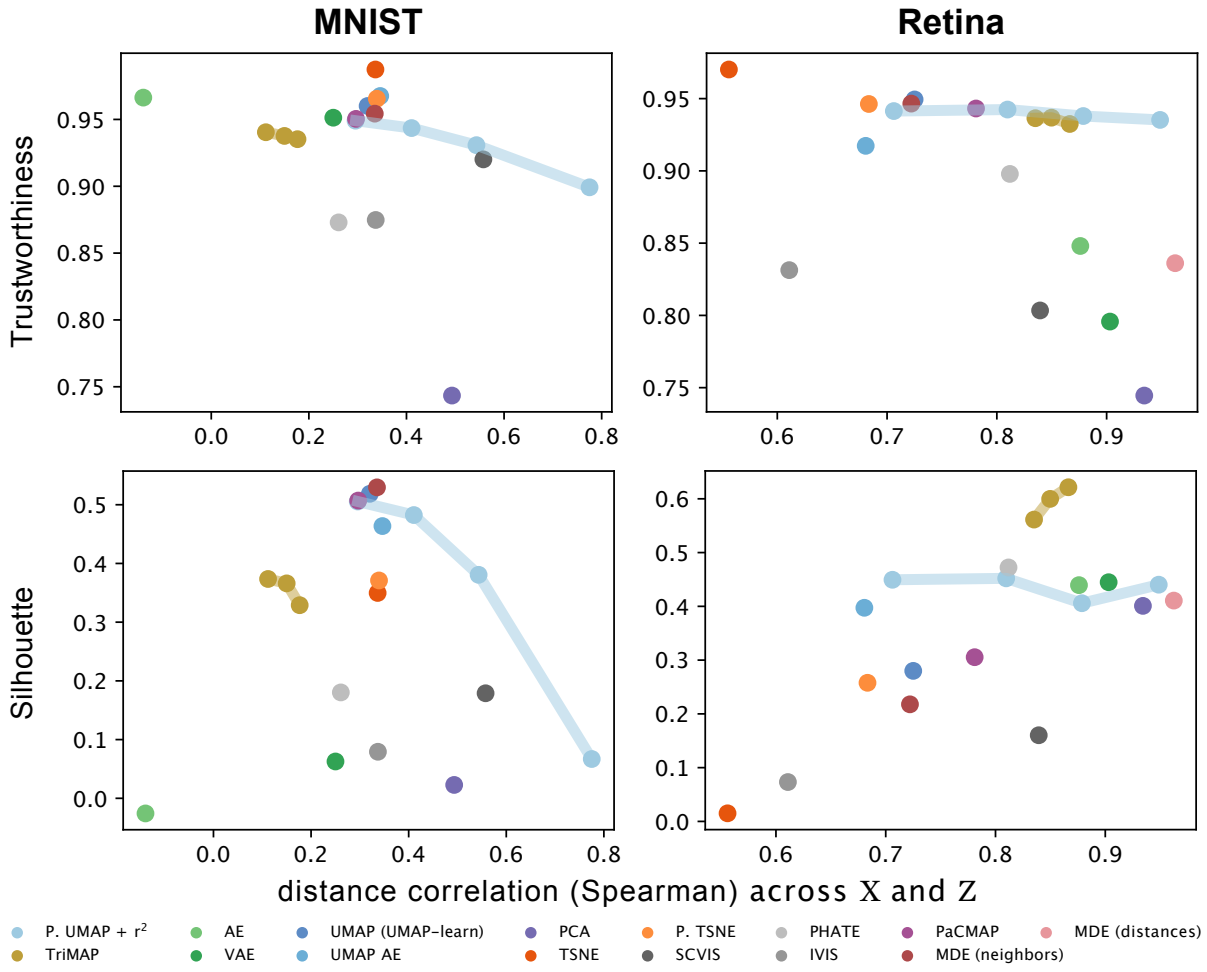


Figure 3.9. Comparison of pairwise global and local relationship preservation across embeddings for MNIST and Macosko [262]. Local structure metrics are silhouette score and Trustworthiness. Global structure metric is Spearman correlation of distances in X and Z. Connected lines are for the different weights of the correlation loss from Figure 3.8 in Parametric UMAP, and the λ parameter in TriMap (50, 500 and 5000). MDE (distances) is not given for MNIST because of memory issues (on 512Gb RAM).

3.5.4 Autoencoding with UMAP

The ability to reconstruct data from embeddings can both aid in understanding the structure of non-linear embeddings, as well as allow for manipulation and synthesis of data based on the learned features of the dataset. We compared the reconstruction accuracy across each method which had inverse-transform capabilities (i.e. $Z \rightarrow X$), as well as the reconstruction speed across the neural network-based implementations to non-parametric implementations and PCA. In addition, we performed latent algebra on Parametric UMAP embeddings both with and without an autoencoder regularization and found that reconstructed data can be linearly manipulated in complex feature space.

Reconstruction accuracy

We measured reconstruction accuracy as Mean Squared Error (MSE) across each dataset (Figure 3.10; Supplementary Table 7. In two dimensions, we find that Parametric UMAP typically reconstructs better than non-parametric UMAP, which in turn performs better than PCA. In addition, the autoencoder regularization slightly improves reconstruction performance. At 64 dimensions, the AE regularized Parametric UMAP is generally comparable to the AE and VAE and performs better than Parametric UMAP without autoencoder regularization. The non-parametric UMAP reconstruction algorithm is not compared at 64 dimensions because it relies on an estimation of Delaunay triangulation, which does not scale well with higher dimensions.

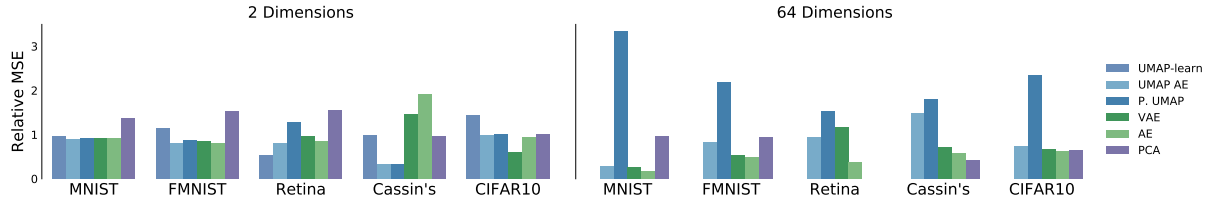


Figure 3.10. Reconstruction accuracy measured as mean squared error (MSE). MSE is shown relative to each dataset (setting mean at 1).

Latent features

Previous work shows that parametric embedding algorithms such as AEs (e.g. Variational Autoencoders) linearize complex data features in latent-space, for example, the presence of a pair of sunglasses in pictures of faces (e.g. [358, 457, 386]). Here, we performed latent-space algebra and reconstructed manipulations on Parametric UMAP latent-space to explore whether UMAP does the same.

To do so, we use the CelebAMask-HQ dataset, which contains annotations for 40 different facial features over a highly structured dataset of human faces. We projected the dataset of faces into a CNN autoencoder architecture based upon the architecture defined in [179]. We trained the network first using UMAP loss alone (Parametric UMAP), and second using the joint UMAP and AE loss (Fig 3.11). We then fit an OLS regression to predict the latent projections of the entire dataset using the 40 annotated features (e.g. hair color, presence of beard, smiling, etc). The vectors corresponding to each feature learned by the linear model were then treated as feature vectors in latent space and added and subtracted from projected images, then passed through the decoder to observe the resulting image (as in [386]).

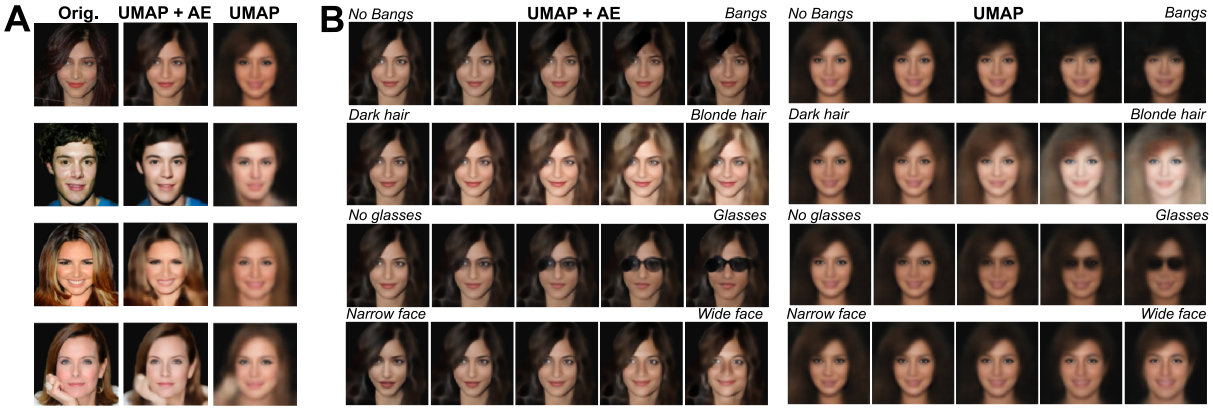


Figure 3.11. Reconstruction and interpolation. (A) Parametric UMAP reconstructions of faces from a holdout testing dataset. (B) The same networks, adding latent vectors corresponding to image features.

We find that complex latent features are linearized in latent space, both when the network is trained with UMAP loss alone as well as when the network is trained with AE loss. For

example, in the third set of images in Figure 3.10, a pair of glasses can be added or removed from the projected image by adding or subtracting its corresponding latent vector.

3.5.5 Semi-supervised learning

Real-word datasets are often comprised of a small number of labeled data, and a large number of unlabeled data. semi-supervised learning (SSL) aims to use the unlabeled data to learn the structure of the dataset, aiding a supervised learning algorithm in making decisions about the data. Current SOTA approaches in many areas of supervised learning such as computer vision rely on deep neural networks. Likewise, semi-supervised learning approaches modify supervised networks with structure-learning loss using unlabeled data. Parametric UMAP, being a neural network that learns structure from unlabeled data, is well suited to semi-supervised applications. Here, we determine the efficacy of UMAP for semi-supervised learning by comparing a neural network jointly trained on classification and UMAP (Figure 3.2D) with a network trained on classification alone using datasets with varying numbers of labeled data.

We compared datasets ranging from highly-structured (MNIST) to unstructured (CIFAR10) in UMAP using a naïve distance metric in data space (e.g. Euclidean distance over images). For image datasets, we used a deep convolutional neural network (CNN) which performs with relatively high accuracy for CNN classification on the fully supervised networks (see Supplementary Table 8 based upon the CNN13 architecture commonly used in SSL [326]. For the birdsong dataset, we used a BLSTM network, and for the retina dataset, we used a densely connected network.

Naïve UMAP embedding

For datasets where structure is learned in UMAP (e.g. MNIST, FMNIST) we expect that regularizing a classifier network with UMAP loss will aid the network in labeling data by learning the structure of the dataset from unlabeled data. To test this, we compared a baseline classifier to a network jointly trained on classifier loss and UMAP loss. We first trained the

baseline classifier to asymptotic performance on the validation dataset, then using the pretrained-weights from the baseline classifier, trained a Y-shaped network (Figure 3.2D) jointly on UMAP over Euclidean distances and a classifier loss over the dataset. We find that for each dataset where categorically-relevant structure is found in latent projections of the datasets (MNIST, FMNIST, birdsong, retina), classifications are improved in the semi-supervised network over the supervised network alone, especially with smaller numbers of training examples (Figure 3.12; Supplementary Table 8. In contrast, for CIFAR10, the additional UMAP loss impairs performance in the classifier.

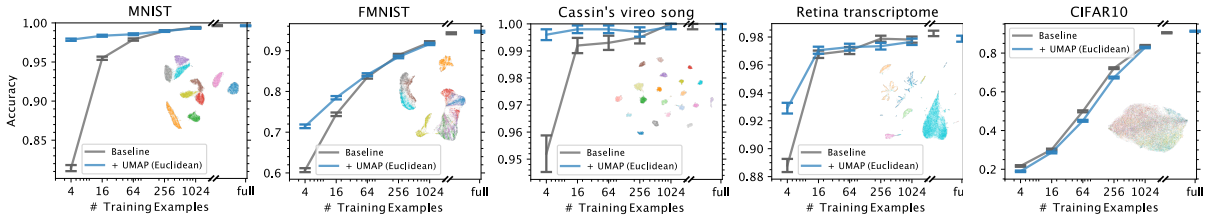


Figure 3.12. Baseline classifier with an additional UMAP loss with different numbers of labeled training examples. Non-parametric UMAP projections of the UMAP graph being jointly trained are shown in the bottom right of each panel. Error bars show SEM.

Consistency regularization and learned invariance using data augmentation

Several current SOTA SSL approaches employ a technique called consistency regularization [388]; training a classifier to produce the same predictions with unlabeled data which have been augmented and data that have not been augmented [409, 30]. In a similar vein, for each image dataset, we train the network to preserve the structure of the UMAP graph when data have been augmented. We computed a UMAP graph over un-augmented data and, using augmented data, trained the network jointly using classifier and UMAP loss, teaching the network to learn to optimize the same UMAP graph, invariant to augmentations in the data. We observe a further improvement in network accuracy for MNIST and FMNIST over the baseline, and the augmented baseline (Figure 3.13 left; Supplementary Table 8. For the CIFAR10 dataset, the addition of the UMAP loss, even over augmented data, reduces classification accuracy.

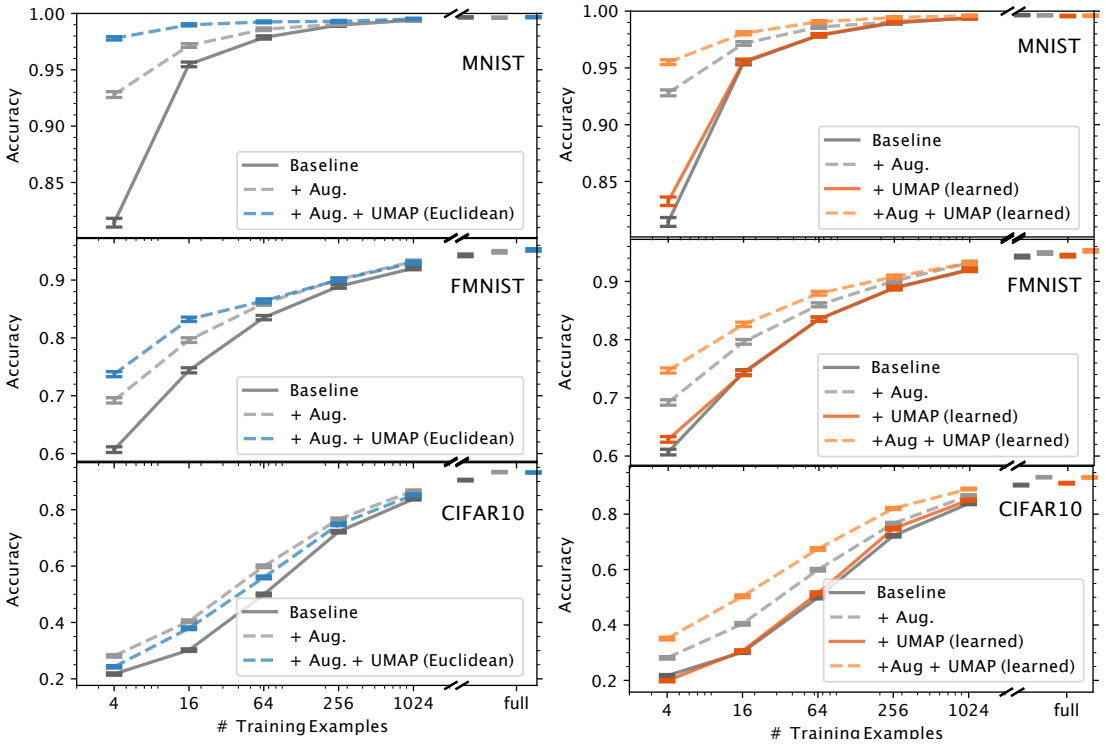


Figure 3.13. Comparison of baseline classifier, augmentation, and augmentation with an additional UMAP loss (left). SSL using UMAP over the learned latent graph, computed over latent activations in the classifier (right).

Learning a categorically-relevant UMAP metric using a supervised network

It is unsurprising that UMAP confers no improvement for the CIFAR10 dataset, as UMAP computed over the pixel-wise Euclidean distance between images in the CIFAR10 dataset does not capture very much categorically-relevant structure in the dataset. Because no common distance metric over CIFAR10 images is likely to capture such structure, we consider using supervision to learn a categorically-relevant distance metric for UMAP. We do so by training on a UMAP graph computed using distance over latent activations in the classifier network (as in, e.g. [60]), where categorical structure can be seen in UMAP projection (Figure 3.14). The intuition being that training the network with unlabeled data to capture distributional structure within the network's learned categorically-relevant space will aid in labeling new data.

We find that in all three datasets, without augmentation, the addition of the learned UMAP loss confers little to no improvement in classification accuracy over the data (Figure 3.13 right; Supplementary Table 8. When we look at non-parametric projections of the graph over latent activations, we see that the learned graph largely conforms to the network's already-present categorical decision making (e.g. Figure 3.14 predictions vs. ground truth). In contrast, with augmentation, the addition of the UMAP loss improves performance in each dataset, including CIFAR10. This contrast in improvement demonstrates that training the network to learn a distribution in a categorically-relevant space that is already intrinsic to the network does not confer any additional information that the network can use in classification. Training the network to be invariant toward augmentations in the data, however, does aid in regularizing the classifier, more in-line with directly training the network on consistency in classifications [388].

3.5.6 Comparisons with indirect parametric embeddings

In principle, any embedding technique can be implemented parametrically by training a parametric model (e.g. a neural network) to predict embeddings from the original high-dimensional data (as in Duque et al [104]). However, such a parametric embedding is limited in comparison to directly optimizing the algorithm's loss function. Parametric UMAP optimizes

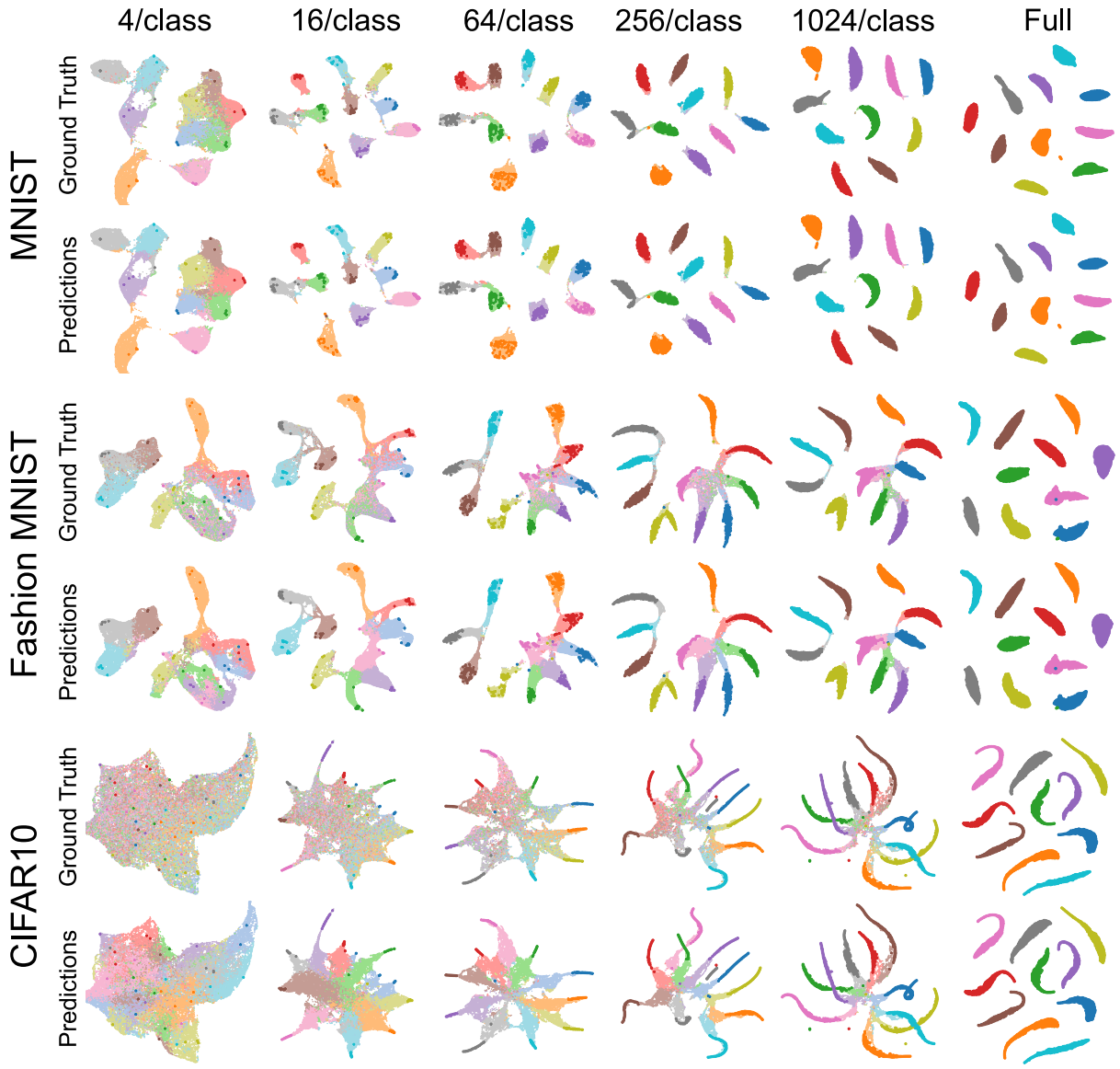


Figure 3.14. Non-parametric UMAP projections of activations in the last layer of a trained classifier for MNIST, FMNIST, and CIFAR10. For each dataset, the top row shows the ground truth labels on above, and the model's predictions below, in a light colormap. On top of each projection, the labeled datapoints used for training are shown in a darker colormap.

directly over the structure of the graph, with respect to the architecture of the network as well as additional constraints (e.g. additional losses). In contrast, training a neural network to predict non-parametric embeddings does not take additional constraints into account.

To exemplify this, in Figure 3.15 we compare Parametric UMAP to a neural network trained to predict non-parametric embeddings by minimizing MSE when the number of neurons in the network is limited. In the case of Parametric UMAP, the objective of the network is to come up with the best embedding of the UMAP graph that it can, given the constraints of the architecture of the network. In the indirect/MSE case, information about the structure of the graph is only available through an intermediary, the non-parametric embedding, thus this method cannot be optimized to learn an embedding of the data that best preserves the structure of the graph. In other words, the indirect method is not optimizing the embedding of the graph with respect to additional constraints. Instead, it is minimizing the distance between two sets of embeddings. The weighted graph is an intermediate topological representation (notably of no specific dimensionality) and is the best representation of the data under UMAP's assumptions. The process of embedding the data in a fixed dimensional space is necessarily a lossy one. Optimizing over the graph directly avoids this loss. This issue also applies when incorporating additional losses (e.g. a classifier loss, or autoencoder loss) to indirect embeddings.

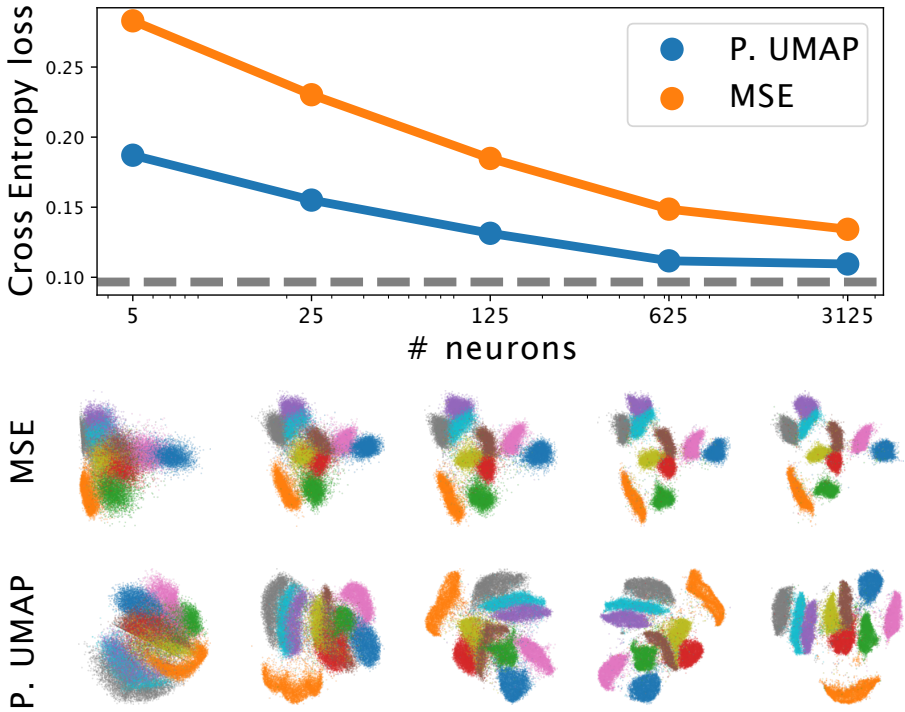


Figure 3.15. (top) Cross entropy loss for one-hidden-layer instance of Parametric UMAP versus a neural network trained to predict non-parametric embeddings using MSE on MNIST. The same network architectures are used in each case. The$x$-axis varies the number of neurons in the network's single hidden layer layer. The dashed grey line is the loss for the non-parametric embedding. (bottom) Projections corresponding to the losses shown in the panel above.

3.6 Discussion

In this paper, we propose a novel parametric extension to UMAP. This parametric form of UMAP produces similar embeddings to non-parametric UMAP, with the added benefit of a learned mapping between data space and embedding space. We demonstrated the utility of this learned mapping on several downstream tasks. We showed that parametric relationships can be used to improve inference times for embeddings and reconstructions by orders of magnitude while maintaining similar embedding quality to non-parametric UMAP. Combined with a global structure preservation loss, Parametric UMAP captures additional global relationships in data, outperforming methods where global structure is only imposed upon initialization (e.g. initializing with PCA embeddings). Combined with an autoencoder, UMAP improves reconstruction quality and allows for the reconstruction of high-dimensional UMAP projections. We also show that Parametric UMAP projections linearize complex features in latent space. Parametric UMAP can be used for semi-supervised learning, improving training accuracy on datasets where small numbers of training exemplars are available. We showed that UMAP loss applied to a classifier improves semi-supervised learning in real-world cases where UMAP projections carry categorically-relevant information (such as stereotyped birdsongs or single-cell transcriptomes), but not in cases where categorically-relevant structure is not present (such as CIFAR10). We devised two downstream approaches based around learned categorically-relevant distances, and consistency regularization, that show improvements on these more complex datasets. Parametric embedding also makes UMAP feasible in fields where dimensionality reduction of continuously generated signals plays an important role in real-time analysis and experimental control.

A number of future directions and extensions to our approach have the potential to further improve upon our results in dimensionality reduction and its various applications. For example, to improve global structure preservation, we jointly optimized over the Pearson correlation between data and embeddings. Using notions of global structure beyond pairwise distances in data space (such as global UMAP relationships or higher-dimensional simplices) may capture

additional structure in data. Similarly, one approach we used to improve classifier accuracy relied on obtaining a 'categorically relevant' metric, defined as the Euclidean distance between activation states of the final layer of a classifier. Recent works (e.g. as discussed and proposed in [397]) have explored methods for more directly capturing class information in the computation of distance, such as using the Fisher metric to capture category- and decision-relevant structure in classifier networks. Similar metrics may prove to further improve semi-supervised classifications with Parametric UMAP.

3.7 Acknowledgments

Work supported by NIH 5T32MH020002-20 to TS and 5R01DC018055-02 to TQG. We would also like to thank Kyle McDonald for making available his translation of Parametric t-SNE to Tensorflow/Keras, which we used as a basis for our own implementation.

Chapter 3, in full, is a reprint of the material as it appears in Neural Computation, 2021, Sainburg, Tim, McInnes, Leland, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 4

Parallels in the sequential organization of birdsong and human speech

Abstract

Human speech possesses a rich hierarchical structure that allows for meaning to be altered by words spaced far apart in time. Conversely, the sequential structure of nonhuman communication is thought to follow non-hierarchical Markovian dynamics operating over only short distances. Here, we show that human speech and birdsong share a similar sequential structure indicative of both hierarchical and Markovian organization. We analyze the sequential dynamics of song from multiple songbird species and speech from multiple languages by modeling the information content of signals as a function of the sequential distance between vocal elements. Across short sequence-distances, an exponential decay dominates the information in speech and birdsong, consistent with underlying Markovian processes. At longer sequence-distances, the decay in information follows a power law, consistent with underlying hierarchical processes. Thus, the sequential organization of acoustic elements in two learned vocal communication signals (speech and birdsong) shows functionally equivalent dynamics, governed by similar processes.

4.1 Introduction

Human language is unique among animal communication systems in its extensive capacity to convey infinite meaning through a finite set of linguistic units and rules[66]. The evolutionary origin of this capacity is not well understood, but it appears closely tied to the rich hierarchical structure of language which enables words to alter meanings across long distances (i.e. over the span of many intervening words or sentences) and timescales. For example, in the sentence, ”Mary, who went to my university, often said that she was an avid birder”, the pronoun ”she” references ”Mary”, which occurs nine words earlier. As the separation between words (within or between sentences) increases, the strength of these long-range dependencies decays following a power law[246, 250]. The dependencies between words are thought to derive from syntactic hierarchies[126, 67], but the hierarchical organization of language encompasses more than word- or phrase-level syntax. Indeed, similar power-law relationships exist for the long-range dependencies between characters in texts[6, 105], and are thought to reflect the general hierarchical organization of natural language, where higher levels of abstraction (e.g., semantic meaning, syntax, words) govern organization in lower level components (e.g., parts-of-speech, words, characters)[246, 250, 6, 105]. Using mutual information (MI) to quantify the strength of the relationship between elements (e.g. words or characters) in a sequence (i.e., the predictability of one element revealed by knowing another element), the power-law decay characteristic of natural languages[250, 247, 6, 105] has also been observed in other hierarchically organized sequences, such as music[244, 250] and DNA codons[340, 250]. Language is not, however, strictly hierarchical. The rules that govern the patterning of sounds in words (i.e., phonology) are explained by simpler Markovian processes[192, 167, 168], where each sound is dependent on only the sounds that immediately precede it. Rather than following a power law, sequences generated by Markovian processes are characterized by MI that decays exponentially as the sequential distance between any pair of elements increases[250, 245]. How Markovian and hierarchical processes combine to govern the sequential structure of speech over different timescales

is not well understood.

In contrast to the complexity of natural languages, non-human animal communication is thought to be dictated purely by Markovian dynamics confined to relatively short-distance relationships between vocal elements in a sequence[66, 162, 25]. Evidence from a variety of sources suggests, however, that other processes may be required to fully explain some non-human vocal communication systems[129, 201, 368, 266, 166, 390, 434, 419, 186, 50]. For example, non-Markovian long-range relationships across several hundred vocal units (extending over 7.5 to 16.5 minutes) have been reported in humpback whale song[419]. Hierarchically-organized dynamics, proposed as fundamental to sequential motor behaviors[237], could provide an alternate (or additional) structure for non-human vocal communication signals. Evidence supporting this hypothesis remains scarce[66, 25]. The present study examines how Markovian and hierarchical processes combine to govern the sequential structure of birdsong and speech. Our results indicate that these two learned vocal communication signals are governed by similar underlying processes.

4.2 Results

4.2.1 Modeling

To determine whether hierarchical, Markovian, or some combination of these two processes better explain sequential dependencies in vocal communication signals, we measured the sequential dependencies between vocal elements in birdsong and human speech. Birdsong (i.e., the learned vocalizations of Oscine birds) is an attractive system to investigate common characteristics of communication signals because birds are phylogenetically diverse and distant from humans, but their songs are spectrally and temporally complex like speech, with acoustic units (notes, motifs, phrases, and bouts) spanning multiple timescales[32]. A number of complex sequential relationships have been observed in the songs of different

species[129, 201, 368, 266, 166, 390, 434, 76]. Most theories of birdsong sequential organization assume purely short timescale dynamics[25, 141, 187, 195], however, and rely typically on far smaller corpora than those available for written-language. Because non-human species with complex vocal repertoires often produce hundreds of different vocal elements that may occur with exceptional rarity[166], fully capturing the long-timescale dynamics in these signals is data intensive.

To compare sequential dynamics in the vocal communication signals of birds and humans, we used large-scale data sets of song from four oscine species whose songs exhibit complex sequential organization (European starlings, Bengalese finches[322], Cassin's vireos[166, 165], and California thrashers[71, 390]). We compared these to large-scale data sets of phonetically-transcribed spontaneous speech from four languages (English[346], German[398], Italian[222], and Japanese[264]). To overcome the sparsity in the availability of large-scale transcribed birdsong data sets, we used a combination of hand-labeled corpora from Bengalese finches, Cassin's vireos, and California thrasher, and algorithmically transcribed data sets from European starlings (Methods; Figure 4.1). The full songbird data set comprises 86 birds totaling 668,617 song syllables recorded in over 195 hours of total singing (Supplementary Table 4.1). The Bengalese finch data was collected from laboratory-reared individuals. The European starling song was collected from wild-caught individuals recorded in a laboratory setting. The Cassin's vireo and California thrasher song were collected in the wild[166, 165, 71, 19]. The diversity of individual vocal elements (syllables; a unit of song surrounded by a pause in singing) for an example bird for each species are shown through UMAP[280] projections in Figure 4.1a-d, and sequential organization is shown in Figure 4.1e-i. For the human speech data sets, we used the Buckeye data set of spontaneous phonetically-transcribed American-English speech[346], the GECO data set of phonetically transcribed spontaneous German speech[398], the AsiCA corpus of ortho-phonetically transcribed spontaneous Italian (Calabrian) speech[222], and the CJS corpus of phonetically transcribed spontaneous Japanese speech[264] totaling 4,379,552 phones from 394 speakers over 150 hours of speaking (Supplementary Table 4.2).

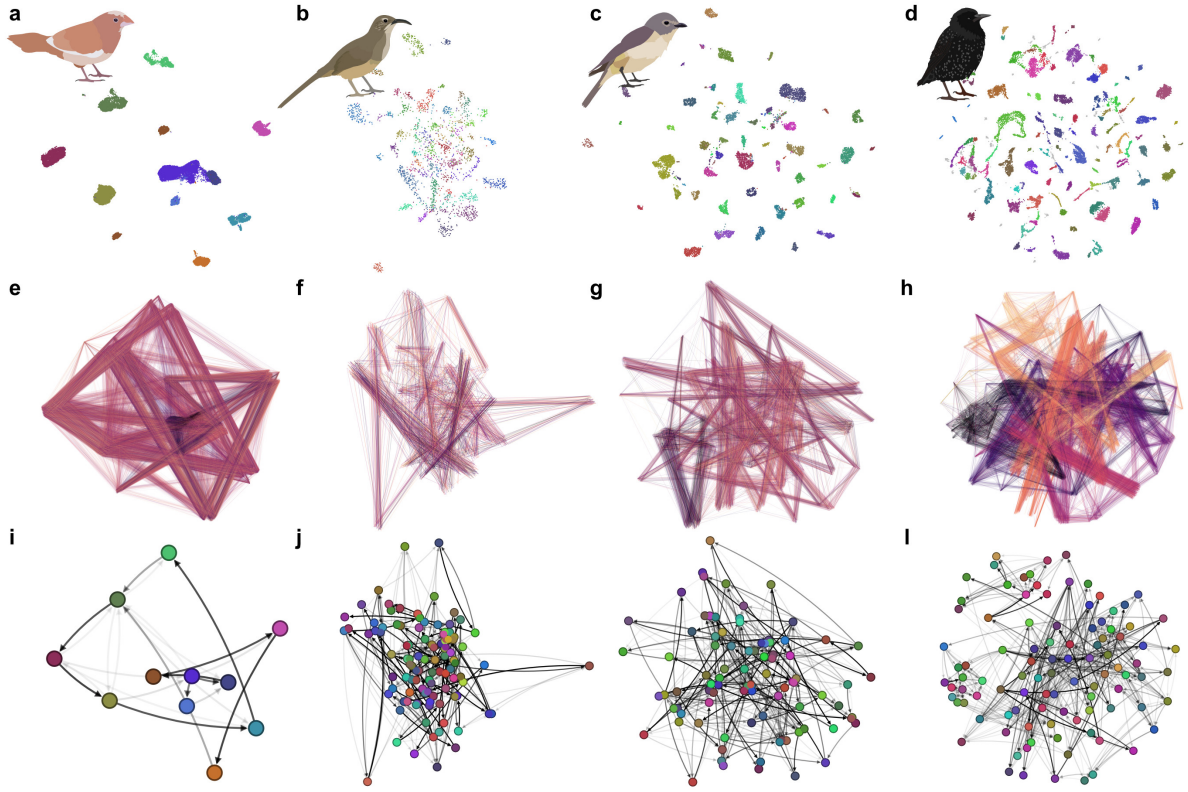


Figure 4.1. Latent and graphical representations of songbird vocalizations. Panels a-d show UMAP[280] reduced spectrographic representations of syllables from the songs of single birds projected into two-dimensions. Each point in the scatterplot represents a single syllable, where color is the syllable category. Syllable categories for Bengalese finch (a), California thrasher (b), and Cassin's vireo (c) are hand-labeled. European starlings (d) are labeled using a hierarchical density-based clustering algorithm[279]. Each column in the figure corresponds to the same animal. Transitions between syllables (e-h) in the same 2D space as a-d, where color represents the temporal position of a transition in a song and stronger hues show transitions that occur at the same position; weaker hues indicate syllable transitions that occur in multiple positions. Transitions between syllable categories (i-l), where colored circles represents a state or category corresponding to the scatterplots in panels a-d, and lines represent state transitions with opacity increasing in proportion to transition probability. For clarity, low-probability transitions ($\leq$5%) are not shown.

For each data set, we computed MI between pairs of syllables or phones, in birdsong or speech respectively, as a function of the sequential distance between elements (Equation (4.4)). For example, in the sequence $A \rightarrow B \rightarrow C \rightarrow D$, where letters denote syllable (or phone) categories, A and B have a sequential distance of 1, while A and D have a distance of 3. In

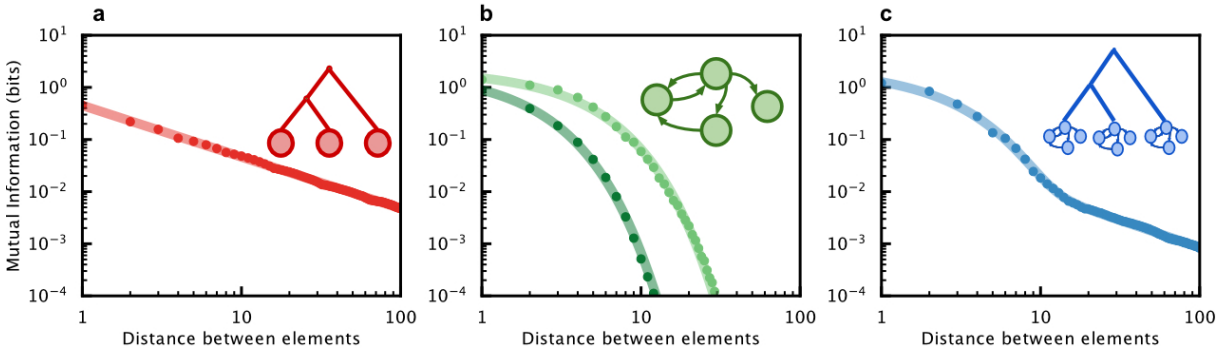


Figure 4.2. MI decay of sequences generated by three classes of models. (a) MI decay of sequences generated by the hierarchically organized model proposed by Lin and Tegmark[250] (red points) is best fit by a power-law decay (red line). (b) MI decay of sequences generated by Markov models of Bengalese finch song from Jin et al.[187] and Katahira et al.[195] (green points) are best fit by an exponential decay model (green lines). (c) MI decay of sequences generated by a composite model (blue points) that combines the hierarchical model (a) and the exponential model (b) is best fit by a composite model (blue line) with both power-law and exponential decays.

general, MI should decay as sequential distance between elements increases and the strength of their dependency drops, because elements separated by large sequential distances are less dependent (on average) than those separated by small sequential distances. To understand the relationship between MI decay and sequential distance in the context of existing theories, we modeled the long-range information content of sequences generated from three different classes of models: a recursive hierarchical model[250], Markov models of birdsong[187, 195], and a model combining hierarchical and Markovian processes by setting Markov generated sequences as the end states of the hierarchical model (Figure 4.2). We then compared three models on their fit to the MI decay: a three-parameter exponential decay model (Equation (4.5)), a three-parameter power-law decay model (Equation (4.6)), and a five-parameter model which linearly combined the exponential and power-law decay models (composite model; Equation (4.7)). Comparisons of model fits were made using the Akaike Information Criterion (AICc) and the corresponding relative probabilities of each model[54] (see Methods) to determine the best-fit model while accounting for the different number of parameters in each model. Consistent

with prior work[250, 247, 245, 246], the MI decay of sequences generated by the Markov models is best fit by an exponential decay, while the MI decay of the sequences generated from the hierarchical model is best fit by a power-law decay. For sequences generated by the combined hierarchical and Markovian dynamics, MI decay is best explained by the composite model, that linearly combines exponential and power-law decay (relative probability > 0.999). Because separate aspects of natural language can be explained by Markovian and non-Markovian dynamics, we hypothesized the MI decay observed in human language would be best explained by a pattern of MI decay similar to that observed in the composite model which combines both Markovian and hierarchical processes. Likewise, we hypothesized that Markovian dynamics alone would not provide a full explanation of the MI decay in birdsong.

4.2.2 Speech

In all four phonetically transcribed speech data sets, MI decay as a function of inter-phone distance is best fit by a composite model that combines a power-law and exponential decay (Figure 4.3, relative probabilities > 0.999, Supplementary Table 4.3). To understand the relative contributions of the exponential and power-law components more precisely, we measured the curvature of the fit of the log-transformed MI decay (Figure 4.3d). The minimum of the curvature corresponds to a downward elbow in the exponential component of the decay, and the maximum in the curvature corresponds to the point at which the contribution of the power law begins to outweigh that of the exponential. The minimum of the curvature for speech ($\sim$3-6 phones for each language or $\sim$0.21-0.31 seconds) aligns roughly with median word length (3-4 phones) in each language data set (Figure 4.3e), while the maximum curvature ($\sim$8-13 phones for each language) captures most ($\sim$89-99%) of the distribution of word lengths (in phones) in each data set. Thus, the exponential component contributes most strongly at short distances between phones, at the scale of words, while the power law primarily governs longer distances between phones, presumably reflecting relationships between words. The observed exponential decay at

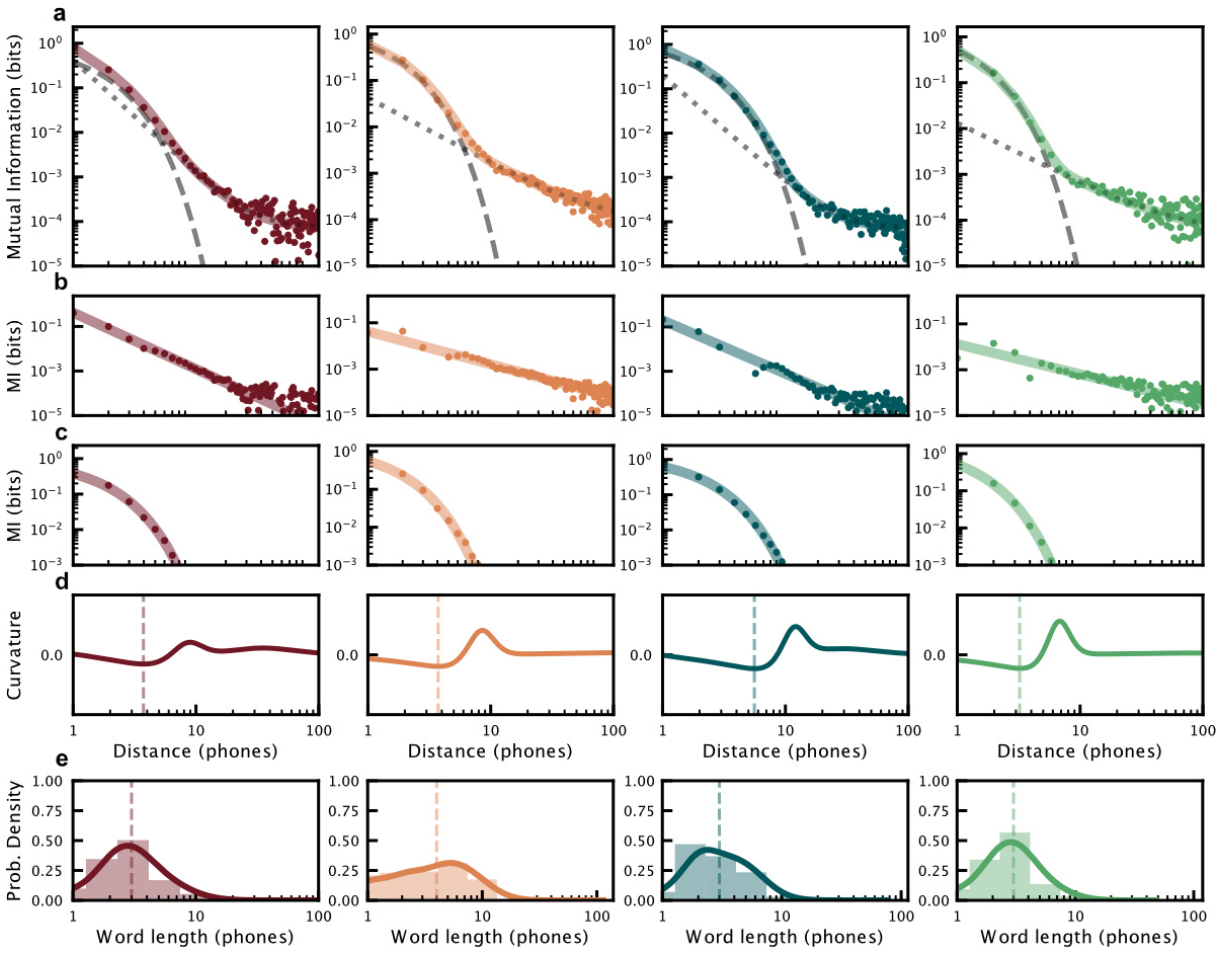


Figure 4.3. Mutual information decay in human speech. (a) MI decay in human speech for four languages (maroon: German, orange: Italian, blue-green: Japanese, green: English) as a function of the sequential distance between phones. MI decay in each language is best fit by a composite model (colored lines) with exponential and power-law decays, shown as a dashed and dotted grey lines, respectively. (b) The MI decay (as in (a)) with the exponential component of the fit model subtracted to show the power-law component of the decay. (c) The same as in (b), but with the power-law component subtracted to show exponential component of the decay. (d) Curvature of the fitted composite decay model showing the distance (in phones) at which the dominant portion of the decay transitions from exponential to power law. The dashed line is drawn at the minimum curvature for each language (English: 3.37, German: 3.57, Italian: 3.72, Japanese: 5.74) (e)Histograms showing the distribution of word lengths in phones, fit with a smoothed Gaussian kernel (colored line). The dashed vertical line shows the median word length (German: 3, Italian: 4, Japanese: 3, English: 3).

inter-word distances agrees with the longstanding consensus that phonological organization is governed by regular (or subregular) grammars with Markovian dynamics[192]. The emphasis

of a power-law decay at intra-word distances, likewise, agrees with the prior observations of hierarchical long-range organization in language[167, 168].

To more closely examine the language-relevant timescales over which Markovian and hierarchical processes operate in speech, we performed shuffling analyses that isolate the information carried within and between words and utterances in the phone data sets. We defined utterances in English and Japanese as periods of continuous speech broken by pauses in the speech stream (Supplementary Figure 4.5; median utterance length in Japanese: 19 phones, English: 21 phones; the German and Italian data sets were not transcribed by utterance). To isolate within-sequence (word or utterance) information, we shuffled the order of sequences within a transcript, while preserving the natural order of phones within each sequence. Isolating within-word information in this way yields MI decay in all four languages that is best fit by an exponential model (Supplementary Figure 4.6a-d). Isolating within-utterance information in the same way yields MI decay best fit by a composite model (Supplementary Figure 4.6i,j), much like the unshuffled data (Figure 4.3a). Thus, only Markovian dynamics appear to govern phone-to-phone dependencies within words. Using a similar strategy, we also isolated information between phones at longer timescales by shuffling the order of phones within each word or utterance, while preserving the order of words (or utterances). Removing within-word information in this way yields MI decay in English, Italian, and Japanese that is best fit by a composite model and MI decay in German that is best fit by a power-law model (Supplementary Figure 4.6e-h). Removing within-utterance information yields MI decay that is best fit by a power-law model (English; Supplementary Figure 4.6k) or a composite model (Japanese; Supplementary Figure 4.6l). Thus, phone-to-phone dependencies within utterances can be governed by both Markov and/or hierarchical processes. The strength of any Markovian dynamics between phones in different words or utterances weakens as sequence size increase, from words to utterances, eventually disappearing altogether in two of the four languages examined here. The same processes that govern phone-to-phone dependencies also appear to shape dependencies between other levels of organization in speech. We analyzed MI decay in the different speech data sets between words,

parts-of-speech, mora, and syllables (depending on transcription availability in each language, see Supplementary Table 4.2). The MI decay between words was similar to that between phones when within-word order was shuffled. Likewise, the MI decay between parts-of-speech paralleled that between words, and the MI decay between mora and syllables (Supplementary Figure 4.7) was similar to that between phones (Figure 4.3a). This supports the notion that long-range relationships in language are inter-related at multiple levels of organization[6].

4.2.3 Birdsong

As with speech, we analyzed the MI decay of birdsong as a function of inter-element distance (using song-syllables rather than phones) for the vocalizations of each of the four songbird species. In all four species, a composite model best fit the MI decay across syllable sequences. (Figure 4.4, relative probabilities > 0.999; Supplementary Table 4.4). The relative contributions of the exponential and power-law components mirrored those observed for phones in speech. That is, the exponential component of the decay is stronger at short syllable-distances, while the power-law component of the decay dominates longer-distance syllable relationships. The transition from exponential to power-law decay (minimum curvature of the fit), was much more variable between songbird species than between languages (Bengalese finch: $\sim$24 syllables or 2.64 seconds, European starlings $\sim$26 syllables or 19.13 seconds, Cassin's vireo: $\sim$21 syllables or 48.94 seconds, California thrasher: $\sim$2 syllables or 0.64 seconds).

To examine more closely the timescales over which Markovian and hierarchical processes operate in birdsong, we performed shuffling analyses (similar to those performed on speech data sets) that isolate the information carried within and between song bouts. We defined song bouts operationally by inter-syllable pauses based upon the species (see Methods). To isolate within-bout information, we shuffled the order of song bouts within a day, while preserving the natural order of syllables within each bout. This yields a syllable-to-syllable MI decay that is best fit by a composite model in each species (Supplementary Figure 4.8a-d), similar to that

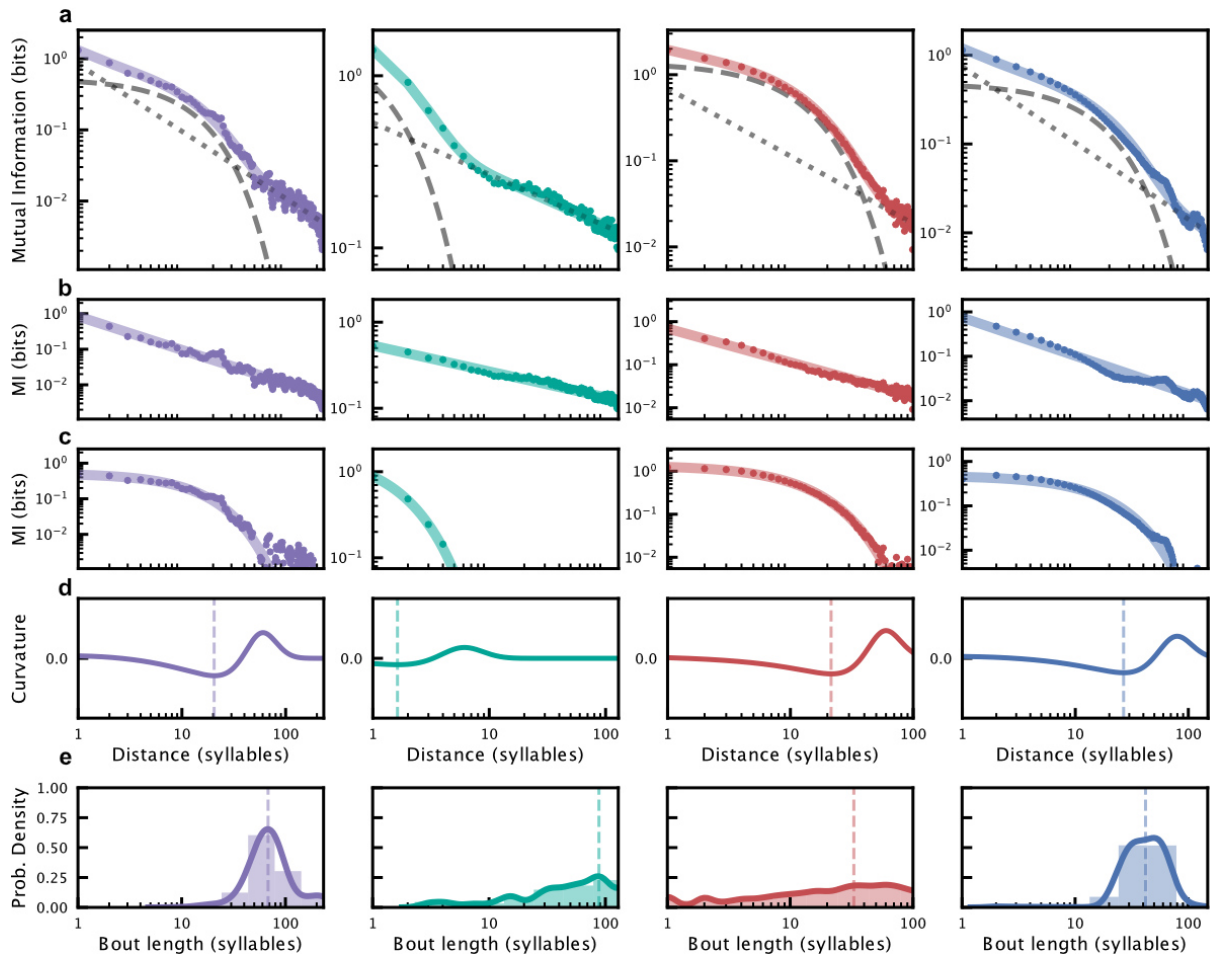


Figure 4.4. Mutual information decay in birdsong. (a) MI decay in song from four songbird species (purple: Bengalese finch, teal: Calofirnia thrasher, red: Cassin's vireo, blue: European starling) as a function of the sequential distance between syllables. MI decay in each species is best fit by a composite model (colored lines) with exponential and power-law decays, shown as a dashed and dotted grey lines, respectively. (b) The MI decay (as in (a)) with the exponential component of the fit model subtracted to show the power-law component of the decay. (c) The same as in (b), but with the power-law component subtracted to show exponential component of the decay. (d) Curvature of the fitted composite decay model showing the distance (in syllables) at which the dominant portion of the decay transitions from exponential to power law. The dashed line is drawn at the minimum curvature for each species (Bengalese finch: $\sim$24, California thrasher: $\sim$2, Cassin's vireo: $\sim$21, European starling: $\sim$26) (e) Histograms showing the distribution of bout lengths in syllables, fit with a smoothed Gaussian kernel (colored line). The dashed line shows the median bout length (Bengalese finch: 68, California thrasher: 88, Cassin's vireo: 33, European starling: 42).

observed in the unshuffled data (Figure 4.4). Thus, both Markovian and hierarchical processes operate at within-bout timescales. To confirm this, we also isolated within-bout relationships by computing the MI decay only over syllables pairs that occur within the same bout (as opposed to pairs occurring over an entire day of singing). Similar to the bout shuffling analysis, MI decay in each species was best fit by the composite model (Supplementary Figure 4.9). To isolate information between syllables at long timescales, we shuffled the order of syllables within bouts while preserving the order of bouts within a day. Removing within-bout information in this way yields MI decay that is best fit by an exponential decay alone (Supplementary Figure 4.8e-h). This contrasts with the results of similar shuffles of phones within words or within utterances in human speech (Supplementary Figure 4.6e-i), and suggests that the hierarchical dependencies in birdsong do not extend across song bouts. This may reflect important differences in how hierarchical processes shape the statistics of both communication signals. Alternatively, this may be an uninteresting artifact of the relatively small number of bouts produced by most birds each day (median bouts per day; finch: 117, starling: 13, thrasher: 1, vireo: 3; see discussion).

To understand how the syntactic organization of song might vary between individual songbirds, even those within the same species, we performed our MI analysis on the data from individuals (Supplementary Figures 10, 4.11). One important source of variability is the size of the data set for each individual. In general, the ability of the composite model to explain additional variance in the MI decay over the exponential model alone correlates positively with the total number of syllables in the data set (Supplementary Figure 4.11a; Pearson's correlation between (log) data set size and ΔAICc: $r = 0.57$, $p < 0.001$, $n = 66$). That is, for smaller data sets it is relatively more difficult to detect the hierarchical relationships in syllable-to-syllable dependencies. In general, repeating the within-bout and bout-order shuffling analyses on individual songbirds yields results consistent with analyses on the full species data sets (Supplementary Figure 4.11b-d). Even in larger data sets containing thousands of syllables, however, there are a number of individual songbirds for whom the composite decay model does not explain any additional variance beyond the exponential model alone (Supplementary Figure

4.11). In a subset of the data where it was possible, we also analyzed MI decay between syllables within a single-day recording session, looking at the longest available recordings in our data set, which were produced by Cassin's vireos and California thrashers and contained over 1000 syllables in some cases (Supplementary Figure 4.12). These single recording sessions show some variability even within individuals, exhibiting decay that in some cases appears to be purely dictated by a power law, and in other cases decay best-fit by the composite model.

4.3 Discussion

Collectively, our results reveal a common structure in both the short- and long-range sequential dependencies between vocal elements in birdsong and speech. For short timescale dependencies, information decay is predominantly exponential, indicating sequential structure that is governed largely by Markovian processes. Throughout vocal sequences, however, and especially for long timescale dependencies, a power law, indicative of non-Markovian hierarchical processes, governs information decay in both birdsong and speech.

These results change our understanding of how speech and birdsong are related. For speech, our observations of non-Markovian processes are not unexpected. For birdsong, they explain a variety of complex sequential dynamics observed in prior studies, including long-range organization[266], music-like structure[368], renewal processes[201, 129], and multiple timescales of organization[434, 76]. In addition, the dominance of Markovian dynamics at shorter timescales may explain why such models have seemed appealing in past descriptions of birdsong[32, 141] and language[188] which have relied on relatively small data sets parsed into short bouts (or smaller segments) where the non-Markovian structure is hard to detect (Supplementary Figure 4.11). Because the longer-range dependencies in birdsong and speech cannot be fully explained by Markov models, our observations rule out the notion that either birdsong or speech are fully defined by regular grammars[32]. Instead, we suggest that the organizing principles of birdsong[434], speech[66], and perhaps sequentially patterned behaviors

in general[237, 85], are better explained by models that incorporate hierarchical organization. The composite structure of the sequential dependencies in these signals helps explain why Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs) have been used successfully to model sequential dynamics in speech[42, 392, 153, 250, 327, 401, 357] and (to a lesser extent) animal communication[12, 76, 321, 194, 195, 282, 361, 456, 462]. HMMs are a class of Markov model which can represent hidden states that underlie observed data, allowing more complex (but still Markovian) sequential dynamics to be captured. HMMs have historically played an important role in speech and language modeling tasks such as speech synthesis[435] and speech recognition[357], but have recently been overtaken by RNNs[392, 389, 153, 327, 401], which model long-range dependencies better than the Markovian assumptions underlying HMMs. A similar shift to incorporate RNNs, or other methods to model hierarchical dynamics, will aid our understanding of at least some non-human vocal communication signals.

The structure of dependencies between vocal elements in birdsong and human speech are best described by both hierarchical and Markovian processes, but the relative contributions of these processes show some differences across languages and species. In speech, information between phones within words decays exponentially (Supplementary Figure 4.6a-d), while the information within utterances follows a combination of exponential and power-law decay (Supplementary Figure 4.6i,j). When this within-word and within-utterance structure is removed (Supplementary Figure 4.6), a strong power law still governs dependencies between phones, indicating a hierarchical organization that extends over very long timescales. Like speech, information between syllables within bouts of birdsong are best described by a combination of power-law and exponential decay (Supplementary Figures 4.9, 4.11a,b). In contrast to speech, however, we did not observe a significant power-law decay beyond that in the bout-level structure (Supplementary Figure 4.11c). The absence of a power law governing syllable dependencies between bouts must be confirmed in future work, as our failure to find it may reflect the fact that we had far fewer bouts per analysis window in the birdsong data sets than we had utterances in the speech data sets. If confirmed, however, it would indicate an upper-bound for the hierarchical

organization of birdsong. It may also suggest that a clearer delineation exists between the hierarchical and Markovian processes underlying speech than those underlying birdsong. In speech the exponential component of the decay is overtaken by the power-law decay at timescales less than one second (0.48-0.72 seconds; Figure 4.3a), whereas in birdsong the exponential component remains prominent for, in some cases, over two minutes (2.43-136.82 seconds; Figure 4.4a). In addition to upward pressures that may push the reach of hierarchical processes to shape longer and longer dependencies in speech, there may also be downward pressures that limit the operational range of Markovian dynamics. In any case, words, utterances, and bouts are only a small subset of the many possible levels of transcription in both signals (e.g. note $<$ syllable $<$ motif $<$ phrase $<$ bout $<$ song; phone $<$ syllable $<$ word $<$ phrase $<$ sentence). Understanding how the component processes that shape sequence statistics are blended and/or separated in different languages and species, and at different levels of organization is a topic for future work. It is also important to note that many individual songbirds produced songs that could be fully captured by Markov processes (Supplementary Figure 4.11). In so far as both the Markovian and hierarchical dynamics capture the output of underlying biological production mechanisms, it is tempting to postulate that variation in signal dynamics across individuals and species may reflect the pliability of these underlying mechanisms, and their capacity to serve as a target (in some species) for selective pressure. The songbird species sampled here are only a tiny subset of the many songbirds and nonhuman animals that produce sequentially patterned communication signals, let alone other sequentially organized behaviors and biological processes. It will be important for future work to document variation in hierarchical organization in a phylogenetically controlled manner and in the context of ontogenic experience (i.e., learning). Our sampling of songbird species was based on available large-scale corpora of songbird vocalizations, and most likely does not capture the full diversity of long- and short-range organizational patterns across birdsong and nonhuman communication. The same may hold true for our incomplete sampling of languages.

Our observations provide evidence that the sequential dynamics of human speech and

birdsong are governed by both Markovian and hierarchical processes. Importantly, this result does not speak to the presence of any specific formal grammar underlying the structure of birdsong, especially as it relates to the various hierarchical grammars thought to support the phrasal syntax of language. It is possible that the mechanisms governing syntax are distinct from those governing other levels of hierarchical organization. One parsimonious conclusion is that the non-Markovian dynamics seen here are epiphenomena of a class of hierarchical processes used to construct complex signals or behaviors from smaller parts, as have been observed in other organisms including fruit flies[29, 83]. These processes might reasonably be co-opted for speech and language production[261]. Regardless of variability in mechanisms, however, the power-law decay in information content between vocal elements is not unique to human language. It can and does occur in other temporally sequenced vocal communication signals including those that lack a well-defined (perhaps any) hierarchical syntactic organization through which meaning is conveyed.

4.4 Methods

4.4.1 Birdsong data sets.

We analyzed song recordings from four different species: European starling (*Sturnus vulgaris*), Bengalese finch (*Lonchura striata domestica*), Cassin's vireo (*Vireo cassinii*), and California thrasher (*Toxostoma redivivum*). As the four data sets were each hand-segmented or algorithmically segmented by different research groups, the segmentation methodology varies between species. The choice of the acoustic unit used in our analyses are somewhat arbitrary and the choice of the term syllable is used synonymously across all four species in this text, however the units that are referred to here as syllables for the California thrasher and Cassin's vireo are sometimes referred to as phrases in other work[166, 165, 71, 390]. Information about the length and diversity of each syllable repertoire is provided in Extended Data Table 4.1.

The Bengalese finch data set[322, 321] was recorded from sound-isolated individuals

and was hand-labeled. The Cassin's vireo[166, 164, 165] and the California thrasher[71] data sets were acquired from the Bird-DB[19] database of wild recordings, and were recorded from the Sierra Nevada and Santa Monica mountains respectively. Both data sets are hand-labeled. The European starling song[17] was collected from wild-caught male starlings (sexed by morphological characteristics) one year of age or older. Starling song was recorded at either 44.1 kHz or 48 kHz over the course of several days to weeks, at various points throughout the year in sound isolated chambers. Some European starlings were administered with testosterone before audio recordings to increase singing behavior. The methods for annotating the European starling data set are detailed in the *Corpus annotation for European starlings* section.

Procedures and methods comply with all relevant ethical regulations for animal testing and research and were carried out in accordance with the guidelines of the Institutional Animal Care and Use Committee at the University of California, San Diego.

4.4.2 Speech corpora.

Phone transcripts were taken from four different data sets: the Buckeye corpus of spontaneous conversational American-English speech[346], the IMS GECO corpus of spontaneous German speech[398], the AsiCA corpus of spontaneous Italian speech of the Calabrian dialect[222] (south Italian), and the CSJ corpus of spontaneous Japanese speech[264].

The American-English speech corpus (Buckeye) consists of conversational speech taken from 40 speakers in Columbus, Ohio. Alongside the recordings, the corpus includes transcripts of the speech and time aligned segmentation into words and phones. Phonetic alignment was performed in two steps: first using Hidden Markov Model (HMM) automatic alignment, followed by hand adjustment and relabeling to be consistent with the trained human labeler. The Buckeye data set also transcribes pauses, which are used as the basis for boundaries in an utterance in our analyses.

The German speech corpus (GECO) consists of 46 dialogs approximately 25 minutes in length each, in which previously unacquainted female subjects are recorded conversing with one

another. The GECO corpus is automatically aligned at the phoneme and word level using forced-alignment[360] from manually generated orthographic transcriptions. A second algorithmic step is then used to segment the data set into syllables[360].

The Italian speech data (AsiCA) consists of directed, informative, and spontaneous recordings. Only the spontaneous subset of the data set was used for our analysis to remain consistent with the other data sets. The spontaneous subset of the data set consists of 61 transcripts each lasting an average of 35 minutes. The AsiCA data set is transcribed using a hybrid orthographic/phonetic transcription method where certain phonetic features were noted with International Phonetic Alphabet (IPA) labels.

The Japanese speech corpus (CSJ) consists of spontaneous speech from either monologues or conversations which are hand transcribed. We use the core subset of the corpus, both because it is the fully annotated subset of the data set, and because it is similar in size to the other data sets used. The core subset of the corpus contains over 500,000 words annotated for phonemes and several other speech features and consists primarily of spontaneously spoken monologues. CSJ is also annotated at the level of *mora*, a syllable-like unit consisting of one or more phonemes and serving as the basis of the 5-7-5 structure of the Haiku[328]. In addition, CSJ is transcribed at the level of Inter-Pausal Units (IPUs) which are periods of continuous speech surrounded by an at-least 200ms pause. We refer here to IPUs as utterances to remain consistent with the Buckeye data set.

As each of the data sets was transcribed using a different methodology, this disparity between the transcription methods may account for some differences in the observed MI decay. The impact of using different transcription methods are at present unknown. The same disparity is true of the birdsong data sets.

4.4.3 Corpus annotation for European starlings.

The European starling corpus was annotated using a novel unsupervised segmentation and annotation algorithm being maintained at GitHub.com/timsainb/AVGN. An outline of the

algorithm is given here.

Spectrograms of each song bout were created by taking the absolute value of the one-sided short-time Fourier transformation of the bandpass filtered waveform. The resulting power was normalized from 0-1, log-scaled, and thresholded to remove low-amplitude background noise in each spectrogram. The threshold for each spectrogram was set dynamically. Beginning at a base power-threshold, all power in the spectrogram below that threshold was set to zero. We then estimated the periods of silence in the spectrogram as stretches of spectrogram where the sum of the power over all frequency channels at a given time-point was equal to zero. If there were no stretches of silence for at least n seconds (described below), the power threshold was increased and the process was repeated until our criteria for minimum length silence was met or the maximum threshold was exceeded. Song bouts for which the maximum threshold was exceeded in our algorithm were excluded as too noisy. This method also filtered out putative bouts that were composed of non-vocal sounds. Thresholded spectrograms were convolved with a Mel-filter, with 32 equally spaced frequency bands between the high and low cutoffs of the Butterworth bandpass filter, then rescaled between 0-255.

To segment song bouts into syllables, we computed the spectral envelope of each song spectrogram, as the sum power across the Mel-scaled frequency channels at every time-sample in the spectrogram. We defined syllables operationally as periods of continuous vocalization bracketed by silence. To find syllables, we first marked silences by minima in the spectral envelope and considered the signal between each silence as a putative syllable. We then compared the duration of the putative syllable to an upper bound on the expected syllable length for each species. If the putative syllable was longer than the expected syllable length, it was assumed to be a concatenation of two or more syllables which had not yet been segmented, and the threshold for silence was raised to find the boundary between those syllables. This processes repeated iteratively for each putative syllable until it was either segmented into multiple syllables or a maximum threshold was reached, at which point it was accepted as a long syllable. This dynamic segmentation algorithm is important for capturing certain introductory whistles in the European

starling song, which can be several times longer than any other syllable in a bout.

Several hyperparameters were used in the segmentation algorithm. The minimum and maximum expected lengths of a syllable in seconds (`ebr_min, ebr_max`) was set to 0.25s/0.75s. The minimum number of syllables (`min_num_sylls`) expected in a bout was set to 20. The maximum threshold for silence (`max_thresh`), relative to the maximum of the spectral envelope) was set to 2%. To threshold out overly noisy song, a minimum length of silence threshold was expected in each bout (`min_silence_for_spec`), set at 0.5s. The base spectrogram (log) threshold for power considered to be spectral background noise (`spec_thresh`) was set at 4.0. This threshold value was set dynamically, where the minimum spectral background noise (`spec_thresh_min`) was set to be 3.5.

We reshaped the syllable spectrograms to create uniformly sized inputs for the dimensionality reduction algorithm. Syllable time-axes were resized using spline interpolation to match a sampling rate of 32 frames equaling the upper limit of the length of a syllable for each species (e.g. a starling's longest syllables are $\sim$1 second, so all syllables are reshaped to a sampling rate of 32 samples/second). Syllables that were shorter than the set syllabic rate were zero-padded on either side to equal 32-time samples, and syllables that were longer than the upper bound were resized to 32-time samples to fit into the network.

Multiple algorithms exist to transcribe birdsong corpora into discrete elements. Our method is unique in that it does not rely on supervised (experimenter) element labeling, or hand-engineered acoustic features specific to individual species beyond syllable length. The method consists of two steps: (1) project the complex features of each birdsong data set onto a 2-dimensional space using the UMAP dimensionality reduction algorithm[280] and (2) apply a clustering algorithm to determine element boundaries[279]. Necessary parameters (e.g. the minimum cluster size) were set based upon visual inspection of the distributions of categories in the 2D latent space. We demonstrate the output of this method in Figure 4.1 both on a European starling data set using our automated transcription, and on the Cassin's vireo, California thrasher, and Bengalese finch data sets. The dimensionality reduction procedure was used for the Cassin's

vireo, Bengalese finch, and California thrasher data sets, but using hand-segmentations rather than algorithmic segmentations of boundaries. The hand-labels are also used rather than UMAP labels for these three species.

4.4.4 Song bouts.

Data sets were either were made available segmented into bouts by the authors of each data set, as in the case of the Bengalese finches, or were segmented into bouts based upon inter-syllable-gaps of greater than 60 seconds in the case of Cassin's vireo and California thrashers, and 10 seconds in the case of European starlings. These thresholds were set based upon the distribution of inter-syllable gaps for each species (Supplementary Figure 4.13).

4.4.5 Mutual information estimation.

We calculated Mutual Information (MI) using distributions of pairs of syllables (or phones) separated by some distance within the vocal sequence. For example, in the sequence "$a \rightarrow b \rightarrow c \rightarrow d \rightarrow e$", where letters denote exemplars of specific syllable or phones categories, the distribution of pairs at a distance of '2' would be $((a,c),(b,d),(c,e))$. We calculate MI between these pairs of elements as:

$$\hat{I}(X,Y) = \hat{S}(X) + \hat{S}(Y) - \hat{S}(X,Y) \tag{4.1}$$

where X is the distribution of single elements (a,b,c) in the example, and Y is the distribution of single elements (c,d,e). $\hat{S}(X)$ and $\hat{S}(Y)$ are the marginal entropies of the distributions of X and Y, respectively, and $\hat{S}(X,Y)$ is the entropy of the joint distribution of X and Y, $((a,c),(b,d),(c,e))$. We employ the Grassberger[152] method for entropy estimation used by Lin and Tegmark[250] which accounts for under-sampling true entropy from finite samples:

$$\hat{S} = \log_2(N) - \frac{1}{N}\sum_{i=1}^{K} N_i \psi(N_i) \tag{4.2}$$

where ψ is the digamma function, K is the number of categories (e.g. syllables or phones) and N is the total number of elements in each distribution. We account for the lower bound of mutual information by calculating the mutual information on the same data set, where the syllable sequence order is shuffled:

$$\hat{I}_{sh}(X,Y) = \hat{S}(X_{sh}) + \hat{S}(Y_{sh}) - \hat{S}(X_{sh},Y_{sh}) \tag{4.3}$$

Where X_{sh} and Y_{sh} refer to the same distributions as X and Y described above, taken from shuffled sequences. This shuffling consists of a permutation of each individual sequence being used in the analysis, which differs depending on the type of analysis (e.g. a bout of song in the analysis shown in Supplementary Figure 4.9 versus an entire day of song in Figure 4.4).

Finally, we subtract out the estimated lower bound of the mutual information from the original mutual information measure.

$$MI = \hat{I} - \hat{I}_{sh} \tag{4.4}$$

4.4.6 Mutual information decay fitting.

To determine the shape of the MI decay, we fit three decay models to the MI as a function of element distance: an exponential decay model, a power-law decay model, and a composite model of both, termed the composite decay:

$$\text{exponential decay} = a * e^{-x*b} + c \tag{4.5}$$

$$\text{power-law decay} = a * x^{b} + c \tag{4.6}$$

$$\text{composite decay} = a * e^{-x*b} + c * x^{d} + f \tag{4.7}$$

where x represents the inter-element distance between units (e.g. phones or syllables). To fit the model on a logarithmic scale, we computed the residuals between the log of the MI and of the model's estimation of the log of the MI. Because our distances were necessarily sampled linearly as integers, we scaled the residuals during fitting by the log of the distance between elements. This was done to emphasize fitting the decay in log-scale. The models were fit using the lmfit Python package[319].

4.4.7 Model selection.

We used the Akaike Information Criterion (AIC) to compare the relative quality of the exponential, composite, and power-law models. AIC takes into account goodness-of-fit and model simplicity, by penalizing larger numbers of parameters in each model (3 for the exponential and power-law models, 5 for the composite model). All comparisons use the AICc[54] estimator, which imposes an additional penalty (beyond the penalty imposed by AIC) to correct for higher-parameter models overfitting on smaller data sets. We choose the best-fit model for the MI decay of each bird's song and the human speech phone data sets using the difference in AICc between models[54]. In the text, we report the relative probability of a given model (in comparison to other models), which is computed directly from the AICc[54] (see Supplementary Information). We report the results using log-transformed data in the main text (Extended Data Tables. 4.4, 4.3).

To determine a reasonable range of element-to-element distances for all the birdsong and speech data sets, we analyzed the relative goodness-of-fit (AICc) and proportion of variance explained (r^2) for each model on decays over distances ranging from 15 to 1000 phones/syllables apart. The composite model provides the best fit for distances up to at least 1000 phones in each language (Supplementary Figure 4.14) and at least the first 100 syllables for all songbird species (Supplementary Figure 4.15). To keep analyses consistent across languages and songbird species we report on analyses using distances up to 100 elements (syllables in birdsong and phones in speech). Figures 4.3 and 4.4 show a longer range of decay in each language and songbird

species, plotted up to element-distances where the coefficient of determination (r^2) remained within 99.9% of its value when fit to 100-element distances.

4.4.8 Curvature of decay fits.

We calculated the curvature for those signals best fit by a composite model in log space (log-distance and log-MI).

$$\kappa = \frac{|y''|}{(1+y'^2)^{\frac{3}{2}}} \tag{4.8}$$

Where y is the log-scaled MI. We then found the local minima and the following local maxima of the curvature function, which corresponds to the 'knee' of the exponential portion of the decay function, and the transition between a primary contribution on the exponential decay to a primary contribution of the power-law decay.

4.4.9 Sequence analyses.

Our primary analysis was performed on sequences of syllables that were produced within the same day to allow for both within-bout and between-bout dynamics to be present. To do so, we considered all syllables produced within the same day as a single sequence and computed MI over pairs of syllables that crossed bouts, regardless of the delay in time between the pairs of syllables. In addition to the primary within-day analysis, we performed three controls to observe whether the observed MI decay was due purely to within-bout, or between-bout organization. The first control was to compute the MI between only syllables that occur within the same bout (as defined by a 10s gap between syllables). Similar to the primary analysis (Figure 4.4), the best-fit model for within-bout MI decay is the composite model (Supplementary Figures 4.11b, 4.9). To more directly dissociate within-bout and between-bout syllable dependencies in songbirds, we computed the MI decay after removing either within- or between-bout structure. To do this, we shuffled the ordering of bouts within a day while retaining the order of syllables within each bout (Supplementary Figure 4.11c), or shuffled the order of syllables within each bout while retaining the ordering of bouts (Supplementary Figure 4.11d). Analyses were performed on

individual songbirds with at least 150 syllables in their data set (Supplementary Figure 4.11), and on the full data set of all birds in a given species. We performed similar shuffling analysis on the speech data sets (Supplementary Figure 4.6). For speech, we shuffled the order of phones within-words (while preserving word order) to remove within-word information, and shuffled word order (while preserving within-word phone ordering) to remove between-word information. We used a similar shuffling strategy at the utterance level remove within- and between-utterance information. The speech data sets were not broken down into individuals due to limitations in data set size at the individual level, and because language is clearly shared between individuals in each speech data set.

To address the possibility that repeating syllables might account for long-range order, we performed separate analyses on both the original syllable sequences (as produced by the bird) and compressed sequences in which all sequentially repeated syllables were counted as a single syllable. The original and compressed sequences show similar MI decay shapes (Supplementary Figure 4.16). We also assessed how our results relate to the timescale of segmentation and discretization of syllables or phones by computing the decay in MI between discretized spectrograms of speech and birdsong at different temporal resolutions (Supplementary Figure 4.17) for a subset of the data. Long-range relationships are present throughout both speech and birdsong regardless of segmentation, but the pattern of MI decay does not follow the hypothesized decay models as closely as that observed when the signals are discretized to phones or syllables, supporting the non-arbitrariness of these low-level production units.

4.4.10 Computational models.

We compared the MI decay of sequences produced by three different artificial grammars: (1) Markov models used to describe the song of two Bengalese finches[187, 195], (2) The hierarchical model proposed by Lin and Tegmark[250], and (3) a model composed of both the hierarchical model advocated by Lin and Tegmark and a Markov model. While these models do not capture the full array of possible sequential models and their signatures in MI decay, they

well-capture the predictions made based upon the discussed literature[250, 245, 246, 6, 105] and provide an illustration of what would be expected given our competing hypotheses. With each model, we generate corpora of sequences, then compute the MI decay of the sequences using the same methods as with the birdsong and speech data. We also fit a power-law, exponential, and composite model to the MI decay, in the same manner (Figure 4.2).

A Markov model is a sequential model in which the probability of transitioning to a state (x_n) is dependant solely on the previous state (x_{n-1}). Sequences are generated from a Markov model by sampling an initial state, x_0 from the set of possible states S. x_0 is then followed by a new state from from the probability distribution $P(x_n|x_{n-1})$. Markov models can thus be captured by a Matrix M of conditional probabilities $M_{ab} = P(x_n = a|x_{n-1} = b)$, where $a \in S$ and $b \in S$. In the example (Figure 4.2b) we produce a set of 65,536 (2^{16}) sequences from Markov models describing two Bengalese finches[187, 195].

The hierarchical model from Lin and Tegmark[250] samples sequences recursively in a similar manner to how the Markov model samples sequences sequentially. Specifically, a state x_0 is drawn probabilistically from the set of possible states S as in the Markov model. The initial state x_0 is then replaced (rather than followed by, as in the Markov model) by q new states (rather than a single state as in the Markov model), which are similarly sampled probabilistically as $P(x_i|x_0)$, where x_i is any of the new q states replacing x_0. The hierarchical grammar can therefore similarly be captured by a conditional probability matrix $M_{ab} = P(x_{l+1} = a|x_l = b)$. The difference between the two models is that the sampled states are replaced recursively in the hierarchical model, whereas in the Markov model they are appended sequentially to the initial state. In the example (Figure 4.2a) we produce a set of 1000 sequences from a model parameterized with an alphabet of 5 states recursively subsampled 12 times, with 2 states replacing the initial state at each subsampling (generating sequences of length 4096).

The final model combines both the Markov model and the hierarchical model by using Markov-generated sequences as the end states of the hierarchical model. Specifically, the combined model is generated in a three-step process: (1) A Markov model is used to generate

sequences equal to the number of possible states of the hierarchical model (S). (2) The combined model is sampled in the exact same manner as the hierarchical model to produce sequences. (3) The end states of the hierarchical model are replaced with their corresponding Markov-generated states from (1). In the example (Figure 4.2c) we produce sequences in the same manner as the hierarchical model. Each state of these sequences is then replaced with sequences between 2 and 5 states long generated by a Markov model with an alphabet of 25 states.

Neither the hierarchical model nor the combined model is meant to exhaustively sample the potential ways in which hierarchical signals can be formed or combined with Markovian processes. Instead, both models are meant to illustrate the theory proposed by prior work and to act as a baseline for comparison for our analyses on real-world signals.

4.5 appendix

4.5.1 AICc

To calculate the AICc[54] for each competing model, we first calculated the (log scaled) residual sum of squares as:

$$RSS(MI, MI_{model}) = (MI - MI_{model})^2 \tag{4.9}$$

The log-likelihood of the model can then be calculated as:

$$\log \mathscr{L} = -\frac{n}{2} \log \left(\frac{RSS}{n} \right) \tag{4.10}$$

where n is the sample size. AIC can then be calculated as:

$$AIC = -2 \log \mathscr{L} + 2K \tag{4.11}$$

where K is the total number of parameters in the model that can be estimated. To be conservative we used the sample bias corrected AIC, AICc, for all reported results, calculated as:

$$AICc = AIC + \frac{2K(K+1)}{n-K-1} \tag{4.12}$$

although the correction made no difference in the results. We computed the ΔAIC as the difference between the best-fit model and each other model:

$$\Delta AIC_i = AICc_i - \min(AICc) \tag{4.13}$$

Using ΔAIC for each model, we calculate the relative likelihood of that model given the data as:

$$\ell_i = \mathscr{L}(model_i \mid data) = e^{-\frac{1}{2}\Delta AIC_i} \tag{4.14}$$

Then the relative probability of each model given the data is computed as the likelihood of each model over the sum of the likelihood of all competing models:

$$P(model_i \mid data) = \frac{\ell_i}{\sum_j \ell_j} \tag{4.15}$$

Finally, the evidence ratio for the best model versus any other given model is the ratio of probabilities of any two given models.

Supplementary Figures

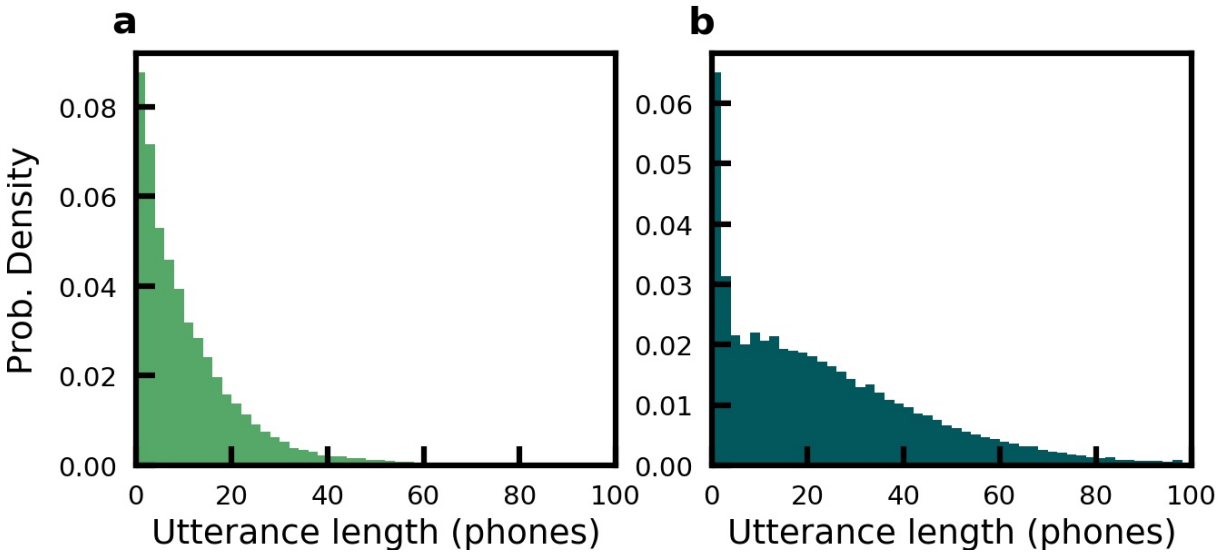


Figure 4.5. Utterance length in phones for English (a) and Japanese (b). The median utterance length in Japanese is 19 phones and in English is 21 phones. The German and Italian data sets were not transcribed by utterance.

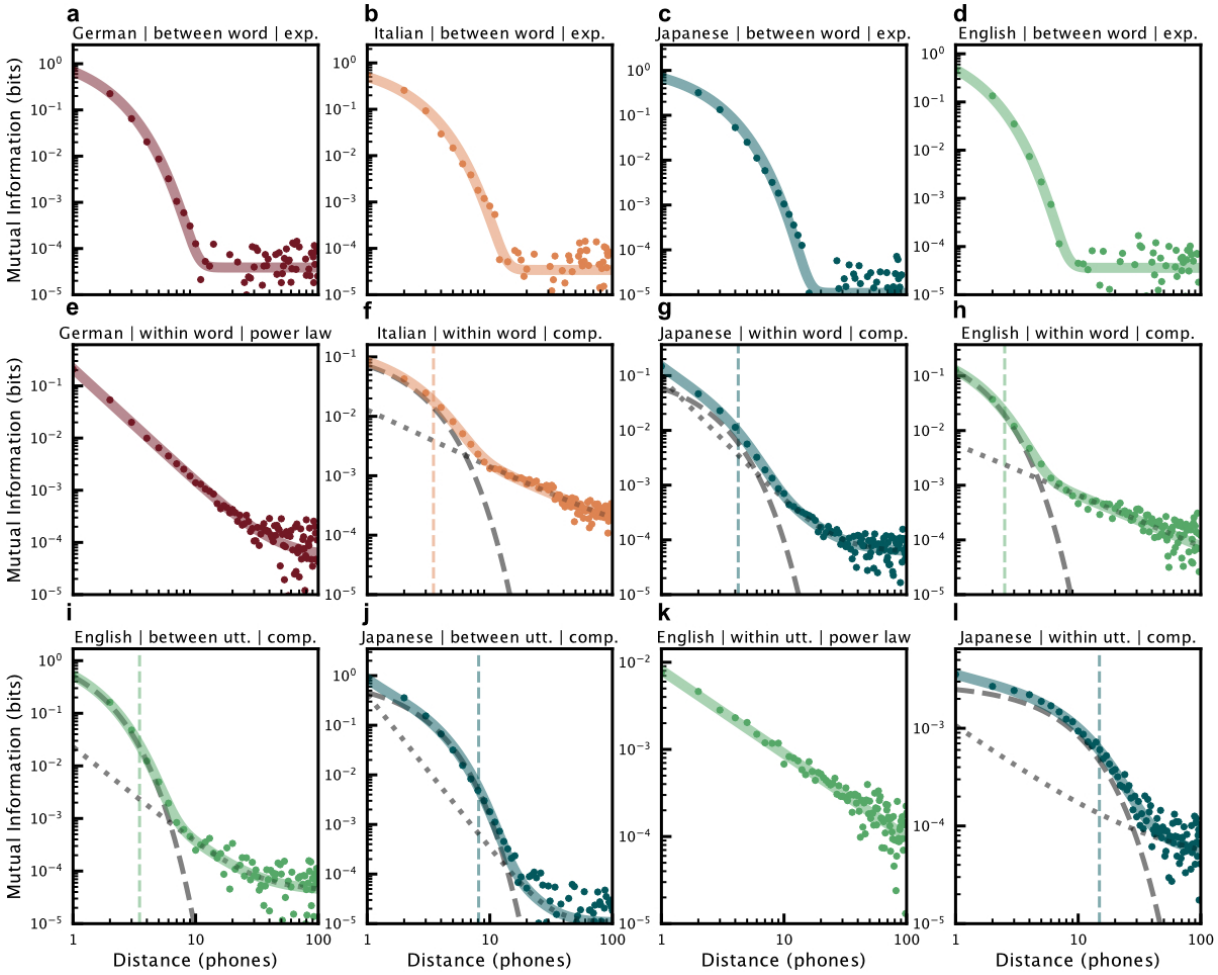


Figure 4.6. MI decay between phones in shuffled speech for different languages (maroon: German, blue-green: Japanese, orange: Italian, green: English). All plots show the MI between phones plotted as a function of sequential distance between phones (as in Figure 4.3). Panels (a-d) show MI when word order is shuffled while phone order within words is preserved. In all cases, decay is best fit by an exponential model (colored lines). Panels (e-h) show MI when phone order within words is shuffled and word order is preserved. Italian (f), Japanese (g), and English (h) are best fit by a composite model, whereas German (e) is best fit by a power-law model. Panels (i) and (j) show MI when the order of utterances are shuffled and phone order within each utterance is preserved. Both English (i) and Japanese (j) are best fit by a composite decay model. Panels k and l show MI when phone order within utterances is shuffled, and utterance order is preserved. English (k) is best-fit by a power-law model while Japanese (l) is best fit by a composite model.

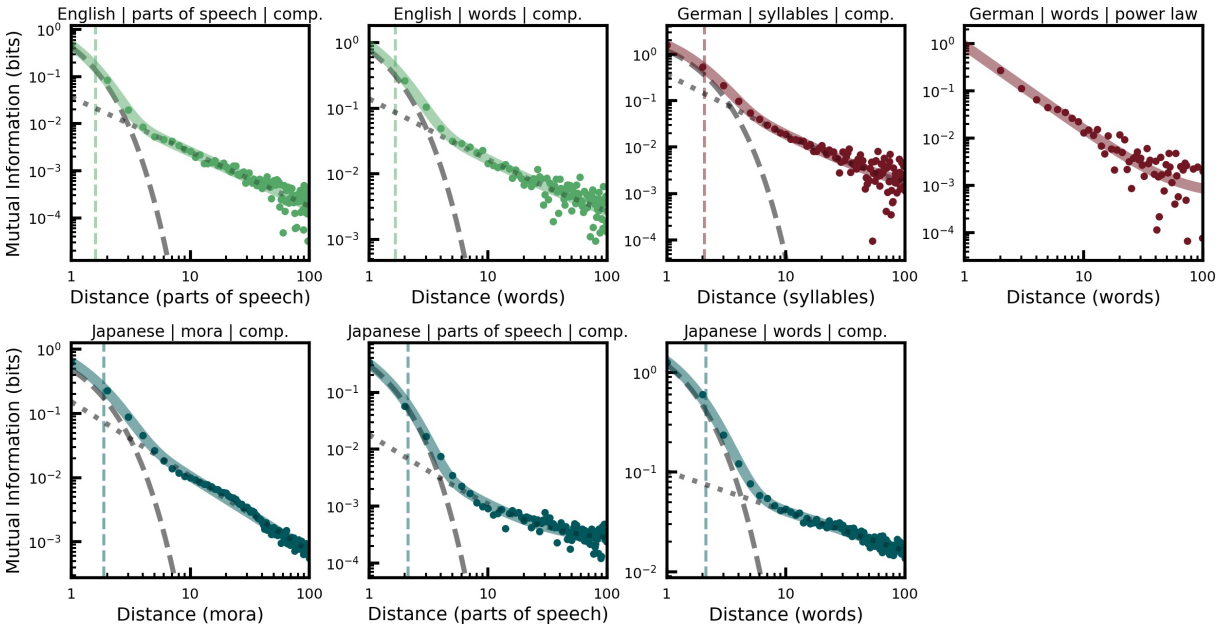


Figure 4.7. MI decay between words, syllables, mora, and parts-of-speech plotted as a function of sequential distances between each of these elements in three languages (green: English, maroon: German, blue-green: Japanese). Not all element categories are available for all languages. For all cases but German words, MI decay is best fit by a composite model (colored lines) with exponential and power-law decays, shown as a dashed and dotted grey lines, respectively. The MI decay between German words is best fit by a power-law. The minima in curvature (colored vertical dashed lines) for words, part-of-speech, and syllables are shorter (in their respective units) than the minima for phones in each language. For English, the minimum curvature is at 1.7 for words, and at 1.6 for parts-of-speech. For German syllables the minimum curvature is at 2.1. For Japanese, the minimum curvature is at 1.9 for mora, 2.1 for words, and 2.1 for parts-of-speech. A minimum curvature is not given for German words because the decay is best fit by a power-law model alone.

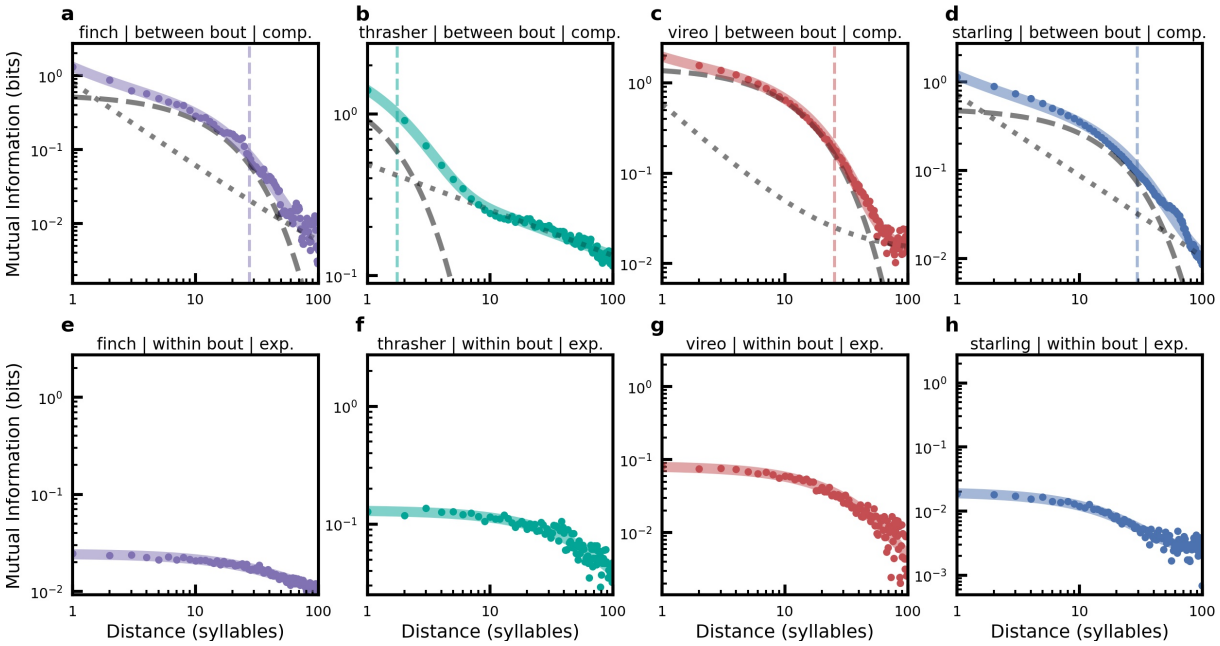


Figure 4.8. MI decay between syllables in shuffled songs from four songbird species (purple: Bengalese finch, teal: California thrasher, red: Cassin's vireo, blue: European starling). All plots show the MI between syllables plotted as a function of sequential distance between syllables (as in Figure 4.4). Panels (a-d) show MI when bout order is shuffled and syllable order within bouts is preserved. Decay in all species is best fit by a composite model. Panels (e-h) show MI when syllable order within each bout is shuffled, and the order of bouts is preserved. Decay in all species is best fit by an exponential model.

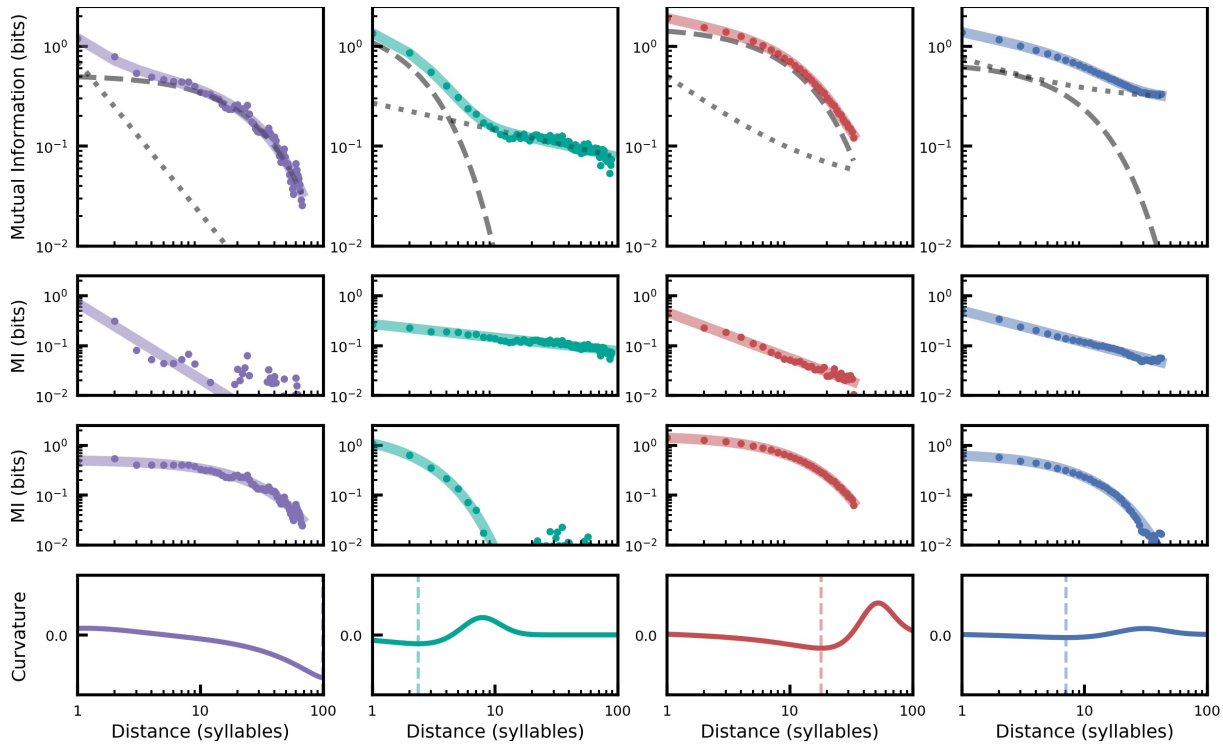
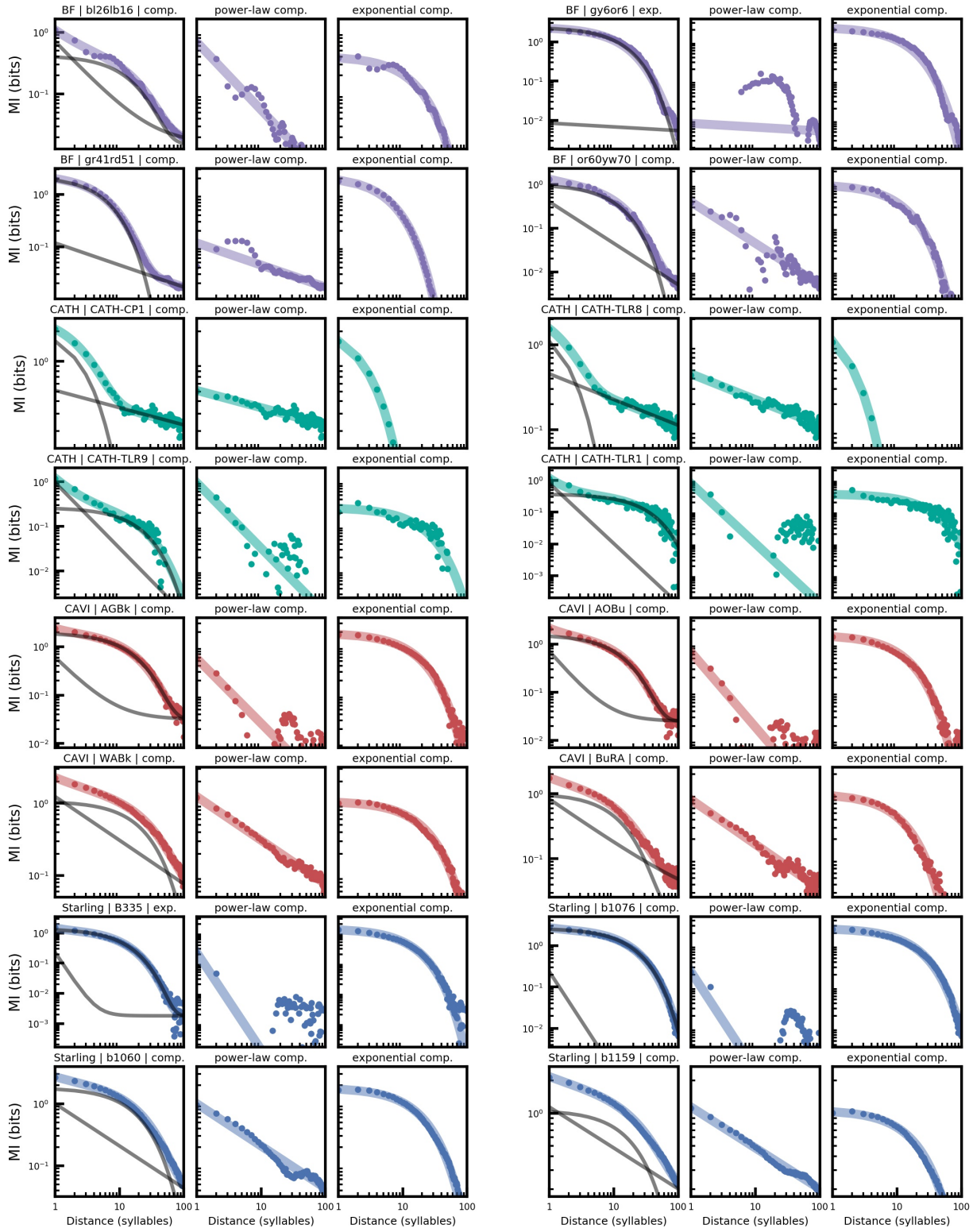


Figure 4.9. Mutual information decay between syllables in the songs of four songbird species (as in Figure 4.4; purple: Bengalese finch, teal: California thrasher, red: Cassin's vireo, blue: European starling), but when the analysis is restricted to syllable pairs that do not span multiple song bouts. MI is plotted from a distance of 1 syllable to the median song length in syllables, to allow a sufficient number of examples for the MI calculation.

Figure 4.10. MI decay in the four largest data sets from individual songbirds in each species. Plots are grouped into sets of three (in a row), corresponding to the data from a one individual songbird (purple: Bengalese finch, teal: California thrasher, red: Cassin's vireo, blue: European starling). For a given bird, the three subplots from left to right show (1) the full MI decay with the fitted model (colored line) and the individual model components (grey lines), (2) the power-law fit to the MI when the exponential component is subtracted, and (3) exponential fit to the MI when the power-law component is subtracted. The species, individual ID, and best-fit model is shown in the title of the leftmost subplot.


BF | bl26lb16 | comp.
power-law comp.
exponential comp.
BF | gy6or6 | exp.
BF | gr41rd51 | comp.
BF | or60yw70 | comp.
CATH | CATH-CP1 | comp.
CATH | CATH-TLR8 | comp.
CATH | CATH-TLR9 | comp.
CATH | CATH-TLR1 | comp.
CAVI | AGBk | comp.
CAVI | AOBu | comp.
CAVI | WABk | comp.
CAVI | BuRA | comp.
Starling | B335 | exp.
Starling | b1076 | comp.
Starling | b1060 | comp.
Starling | b1159 | comp.
MI (bits)
Distance (syllables)

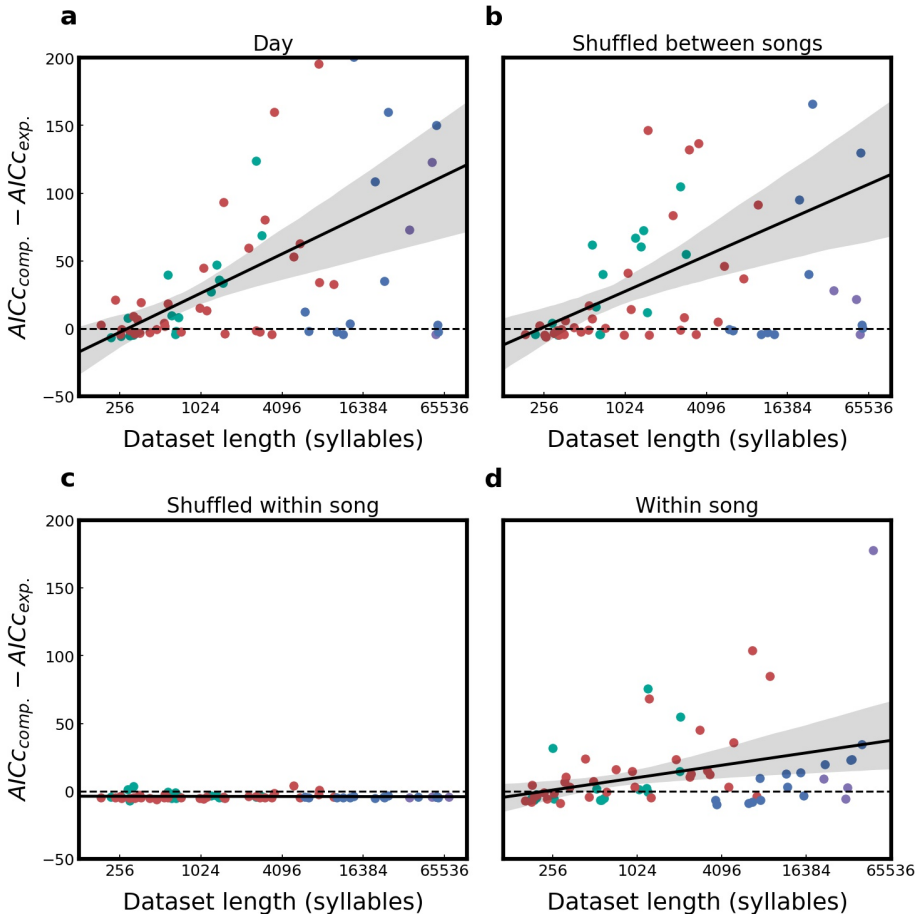
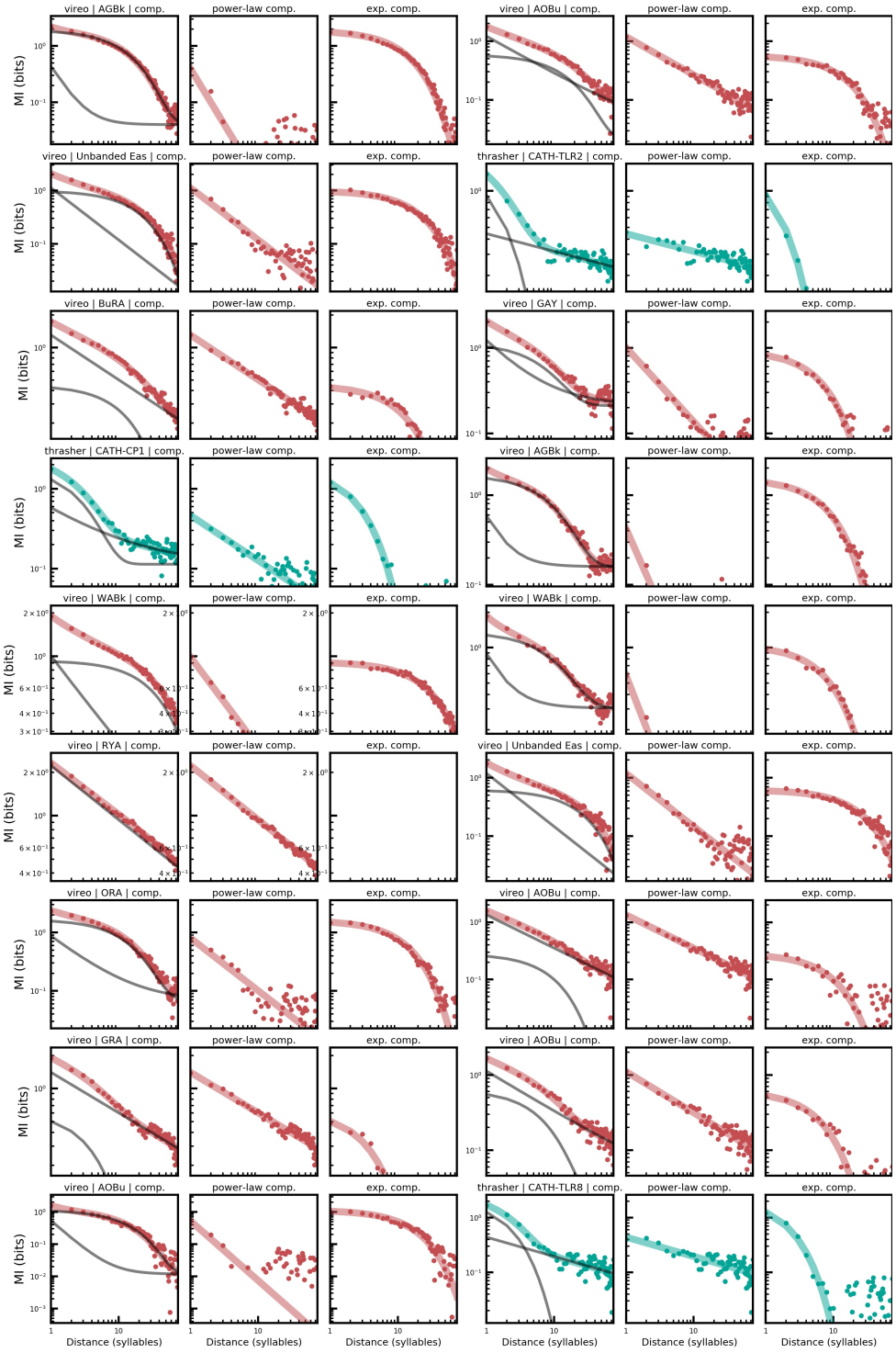


Figure 4.11. Relative decay model fits. Scatterplots (a-d) showing the difference in model fit (ΔAICc) for the composite model versus the exponential model of MI decay for individual songbirds (purple: Bengalese finch, teal: California thrasher, red: Cassin's vireo, blue: European starling) plotted as a function of the number of syllables in each individual songbird's data set. The black line shows a linear regression model with 95% confidence interval fit to the positive relationship between the improvement of the composite model over the exponential model as a function of (log) data set size. Points above zero (dashed line) are better fit by the composite model, while points below are better fit by the exponential model. (a) MI decay for each bird computed across all bouts within a day. (b) The same plot as in (a), but shuffling the ordering of bouts to remove between-bout structure. (c) The same plot as in (a), but shuffling syllable order within bouts, to remove within-bout structure. (d) The same plot as in (a), but where the analysis is restricted to only those syllable pairs within the same song bout.

Figure 4.12. The decay in MI between syllables in the 18 individual songbirds with the longest available recordings in all data sets. Vireo and thrasher decay. Each set of three plots is from either a Cassin's vireo (red) or California thrasher (teal). The three subplots for each bird are organized as in Supplementary Figure 4.10, showing (from left to right) the full MI decay (colored line, with individual model components in grey), power-law fit after the exponential component is subtracted, and exponential fit after the power-law component subtracted. The species, individual ID, and best-fit model is given in the title of the left-most subplot.


vireo | AGBk | comp.
power-law comp.
exp. comp.
vireo | AOBu | comp.
power-law comp.
exp. comp.
vireo | Unbanded Eas | comp.
power-law comp.
exp. comp.
thrasher | CATH-TLR2 | comp.
power-law comp.
exp. comp.
vireo | BuRA | comp.
power-law comp.
exp. comp.
vireo | GAY | comp.
power-law comp.
exp. comp.
thrasher | CATH-CP1 | comp.
power-law comp.
exp. comp.
vireo | AGBk | comp.
power-law comp.
exp. comp.
vireo | WABk | comp.
power-law comp.
exp. comp.
vireo | WABk | comp.
power-law comp.
exp. comp.
vireo | RYA | comp.
power-law comp.
exp. comp.
vireo | Unbanded Eas | comp.
power-law comp.
exp. comp.
vireo | ORA | comp.
power-law comp.
exp. comp.
vireo | AOBu | comp.
power-law comp.
exp. comp.
vireo | GRA | comp.
power-law comp.
exp. comp.
vireo | AOBu | comp.
power-law comp.
exp. comp.
vireo | AOBu | comp.
power-law comp.
exp. comp.
thrasher | CATH-TLR8 | comp.
power-law comp.
exp. comp.
MI (bits)
Distance (syllables)

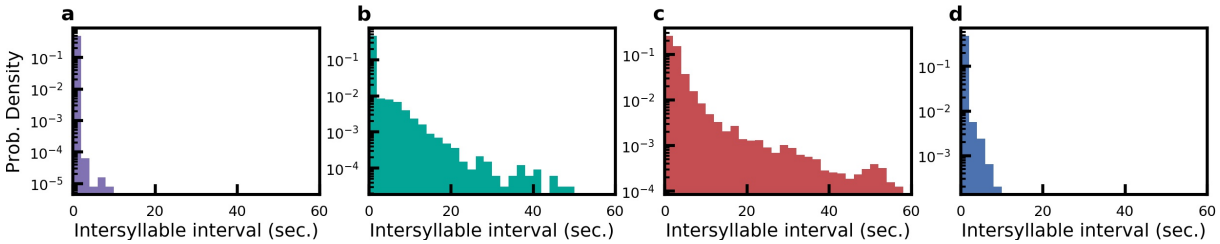


Figure 4.13. The intersyllable interval time in seconds for each songbird species. (a) Bengalese finch (b) California thrasher (c) Cassin's vireo (d) European starling.

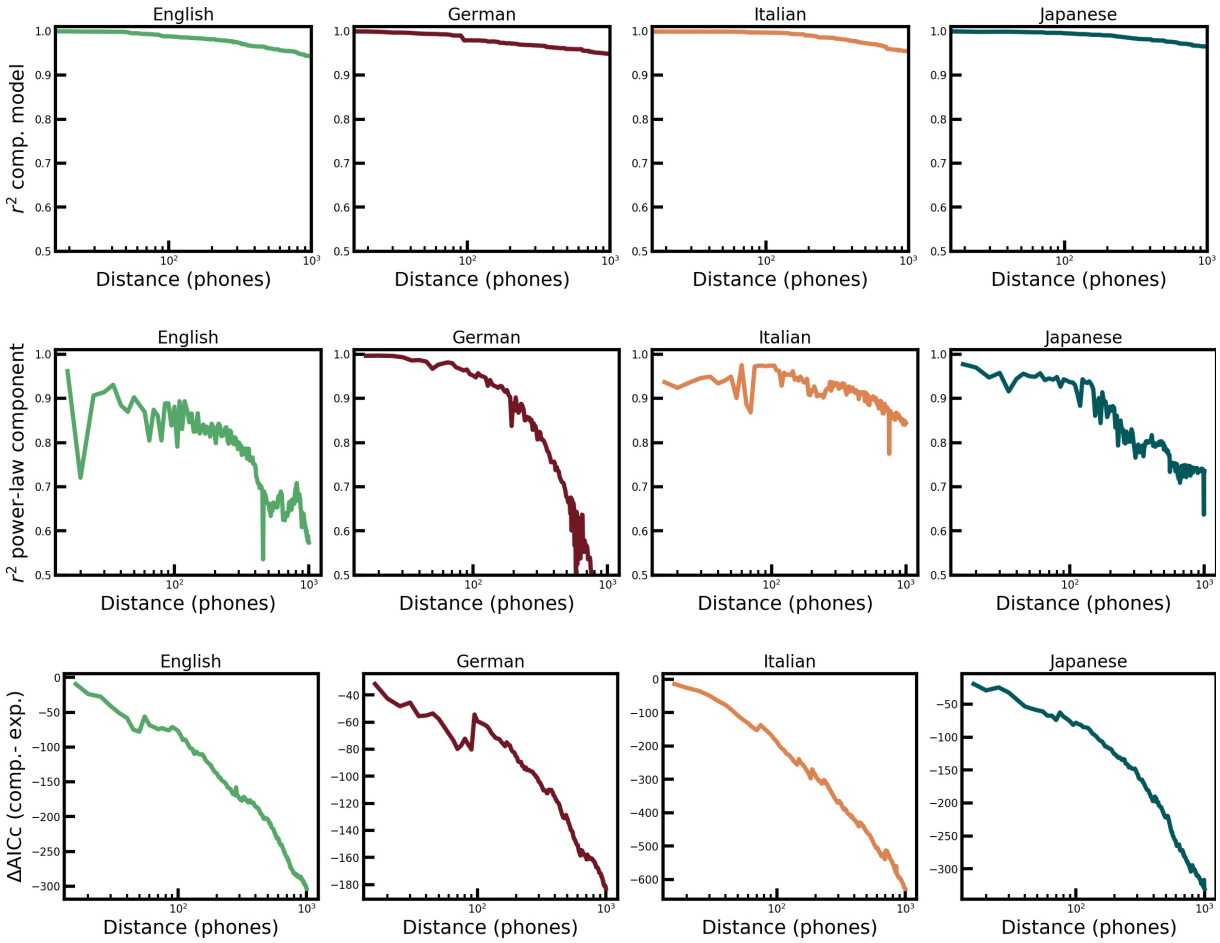


Figure 4.14. The goodness of fit of the composite decay model for each language as a function of the MI analysis length. The coefficient of determination (r^2) for the full composite model (top) and the power-law component of the composite model (center). r^2 is computed for fits of the composite model on MI decay at distances of 15-1000 phones (x-axis). (bottom) ΔAICc between composite and exponential decay models for each language as a function of the maximum phone-to-phone distance computed (green: English, maroon: German, orange: Italian, blue-green: Japanese).

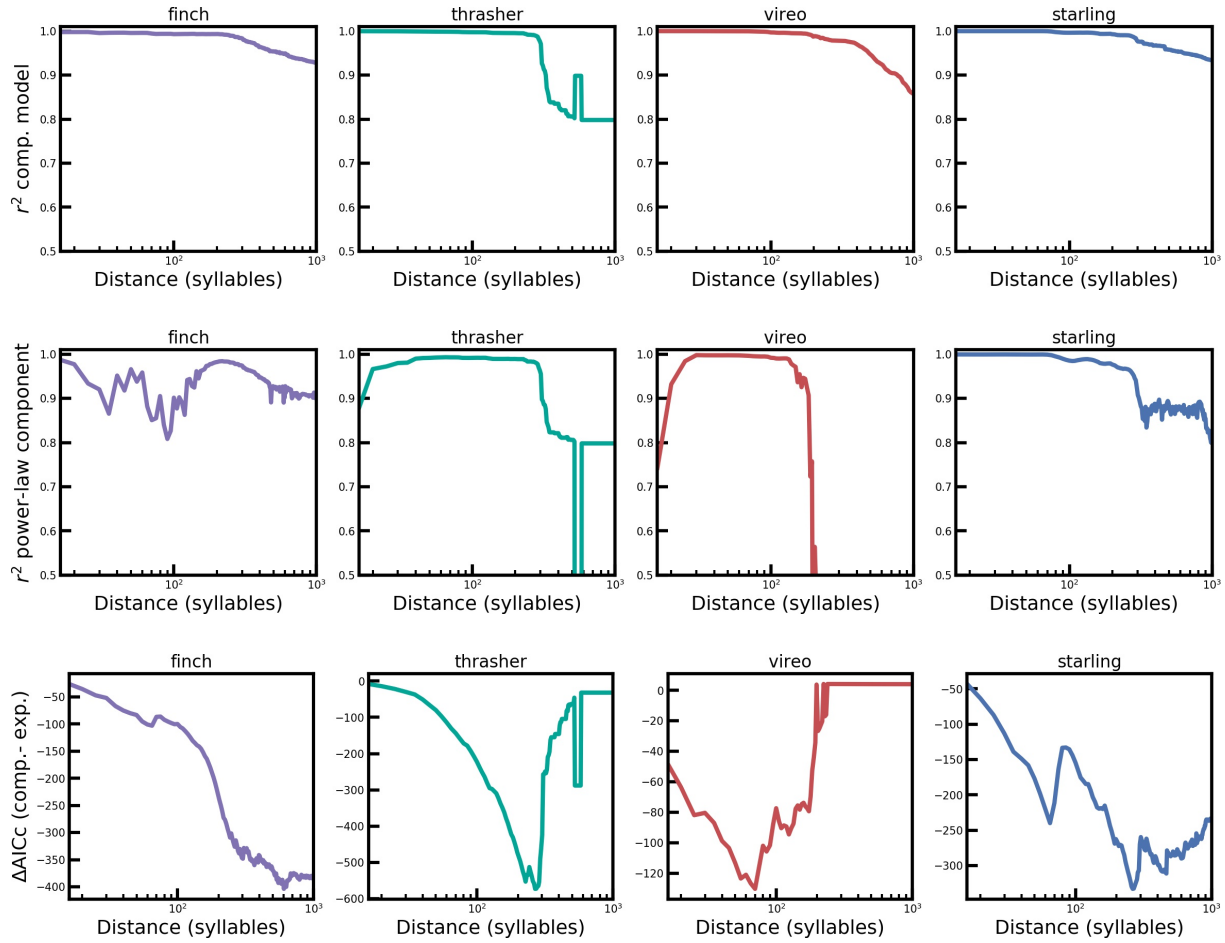


Figure 4.15. The goodness of fit of the composite decay model for each songbird species as a function of the MI analysis length. The coefficient of determination (r^2) for the full composite decay model (top) and the power-law component of the composite model (center). r^2 is computed for fits of the composite model on MI decay of distances of 15-1000 syllables (x-axis). (bottom) ΔAICc between composite and exponential decay models for each species as a function of maximum syllable-to-syllable distance computed (purple: Bengalese finch, teal: California thrasher, red: Cassin's vireo, blue: European starling).

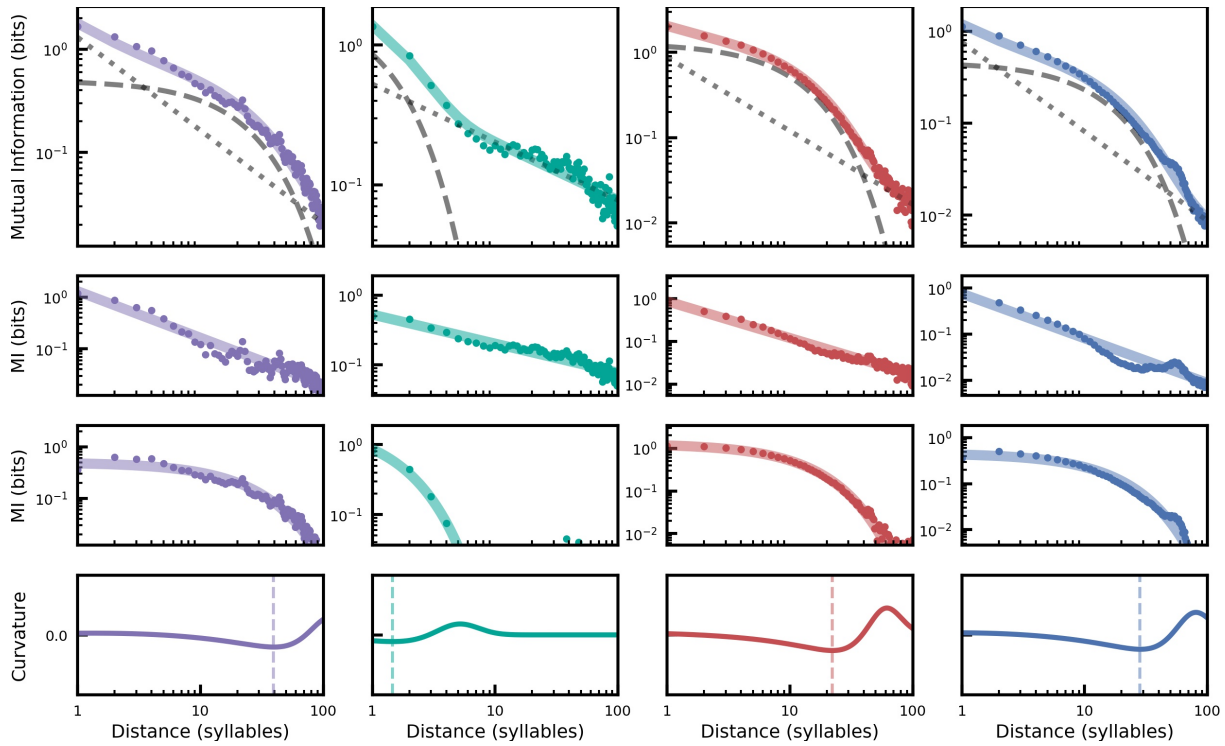


Figure 4.16. Decay of MI between syllables in the birdsong data sets after removing sequentially repeated syllables. Data follow those in Figure 4.4 (purple: Bengalese finch, teal: California thrasher, red: Cassin's vireo, blue: European starling). The decay of Cassin's vireo, California thrasher, and European starling song is largely unaffected, whereas exponential portion of the decay of Bengalese finch song is shifted.

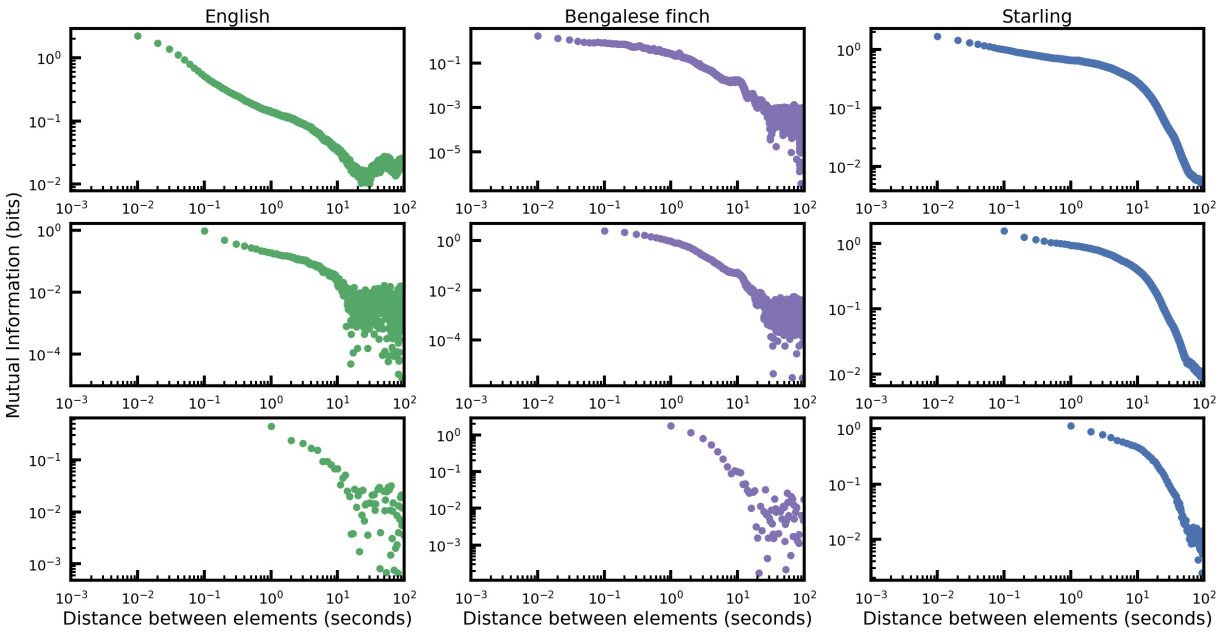


Figure 4.17. Decay in MI between song and speech signal components arbitrarily parsed at multiple timescales. Raw waveforms are split into discrete units at three different timescales (0.01, 0.1, and 1 second), and classified using k-means clustering. Each set of three plots in a column shows the MI between units at one of three timescales (0.001 to 1 second, top to bottom) as a function of the distance between units. Each column shows data for vocalizations of a single individual (green: English, purple: Bengalese finch, blue: European starling). The analysis was only performed on a subset of individuals/data due to the length of time required to segment, cluster and calculate MI on small timescales with large data sets.

Supplementary Tables

Table 4.1. Birdsong dataset statistics.

Species	**Ben. finch**	**Eur. starling**	**Cass. vireo**	**Cal. thrasher**
Origin	Laboratory	Wild-caught	Wild	Wild
# individuals	4	14	50	18
# syllables	215,740	368,956	68,157	15,764
Duration (hrs.)	7.44	89.14	94.07	4.5
Hand Labelled	Yes	No	Yes	Yes
Unique syllables (median)	18.5	151.5	48	51.5
Syllables in bout (median)	68	42	7	21
Syllable length (s; median)	0.07	0.68	0.33	0.15

Table 4.2. Language dataset statistics.

Dataset	**Buckeye**	**GECO**	**AsiCA**	**CSJ**
Language	English	German	Italian	Japanese
Transcripts	40	92	61	201
Duration (Hrs.)	37.9	39.9	35.4	37.6
# Phones	841,266	839,543	1,065,084	1,633,659
Unique phone labels	45	70	90	49
Transcription				
Phone	Yes	Yes	Ortho-phonetic	Yes
Mora	No	No	No	Yes
Words	Yes	Yes	Ortho-phonetic	Yes
Syllables	No	Yes	No	No
Part of speech	Yes	No	No	Yes
Utterance	Yes	No	No	Yes

Table 4.3. Language corpus model fit results at 100 phones of distance.

		German	Italian	English	Japanese
AICc	**exp**	-261.645	-355.721	-255.784	-401.333
	composite	-343.68	-566.454	-311.642	-509.903
	power-law	-326.137	-435.96	-279.64	-348.479
r^2	**exp**	0.966	0.977	0.954	0.991
	composite	0.986	0.997	0.975	0.997
	power-law	0.983	0.99	0.964	0.985
Relative likelihood	**exp**	$<$0.001	$<$0.001	$<$0.001	$<$0.001
	composite	$>$0.999	$>$0.999	$>$0.999	$>$0.999
	power-law	$<$0.001	$<$0.001	$<$0.001	$<$0.001
Relative probability	**exp**	$<$0.001	$<$0.001	$<$0.001	$<$0.001
	composite	$>$0.999	$>$0.999	$>$0.999	$>$0.999
	power-law	$<$0.001	$<$0.001	$<$0.001	$<$0.001

Table 4.4. Birdsong dataset model fit results at 100 syllables of distance.

		Ben. finch	Cal. thrasher	Cass. vireo	Eur. starling
AICc	**exp**	-489.251	-582.14	-637.678	-520.903
	composite	-586.509	-797.431	-763.787	-676.984
	power-law	-390.009	-698.559	-354.734	-405.85
r^2	**exp**	0.98	0.975	0.995	0.981
	composite	0.993	0.997	0.999	0.996
	power-law	0.945	0.992	0.92	0.942
Relative likelihood	**exp**	$<$0.001	$<$0.001	$<$0.001	$<$0.001
	composite	$>$0.999	$>$0.999	$>$0.999	$>$0.999
	power-law	$<$0.001	$<$0.001	$<$0.001	$<$0.001
Relative probability	**exp**	$<$0.001	$<$0.001	$<$0.001	$<$0.001
	composite	$>$0.999	$>$0.999	$>$0.999	$>$0.999
	power-law	$<$0.001	$<$0.001	$<$0.001	$<$0.001

4.6 Acknowledgments

Chapter 4, in full, is a reprint of the material as it appears in Nature Communications, 2019, Sainburg, Tim, Theilman, Brad, Thielk, Marvin, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Long-range sequential dependencies precede complex syntactic production in language acquisition

Abstract

To convey meaning, human language relies on hierarchically organized, long-range relationships spanning words, phrases, sentences, and discourse. As the distances between elements (e.g., phonemes, characters, words) in human language sequences increase, the strength of the long-range relationships between those elements decays following a power law. This power-law relationship has been attributed variously to long-range sequential organization present in human language syntax, semantics, and discourse structure. However, non-linguistic behaviors in numerous phylogenetically distant species, ranging from humpback whale song to fruit fly motility, also demonstrate similar long-range statistical dependencies. Therefore, we hypothesized that long-range statistical dependencies in human speech may occur independently of linguistic structure. To test this hypothesis, we measured long-range dependencies in several speech corpora from children (aged 6 months – 12 years). We find that adult-like power-law statistical dependencies are present in human vocalizations at the earliest detectable ages, prior to the production of complex linguistic structure. These linguistic structures cannot, therefore, be the sole cause of long-range statistical dependencies in language.

5.1 Introduction

Since Shannon's original work characterizing the sequential dependencies present in language, the structure underlying long-range information in language has been the subject of a great deal of interest in linguistics, statistical physics, cognitive science, and psychology [399, 381, 7, 6, 249, 151, 391, 106, 4, 284, 298, 299, 297, 246, 317, 56, 316, 78, 400, 135]. Long-range information content refers to the dependencies between discrete elements (e.g., units of spoken or written language) that persist over long sequential distances spanning words, phrases, sentences, and discourse. For example, in Shannon's original work, participants were given a series of letters from an English text and were asked to predict the letter that would occur next. Using the responses of these participants, Shannon derived an upper bound on the information added by including each preceding letter in the sequence. More recent investigations compute statistical dependencies directly from language corpora using either correlation functions [299, 297, 7, 6, 391, 106, 284] or mutual information (MI) functions [381, 246, 249, 151] between elements in a sequence. In both cases, dependencies are calculated as a function of the sequential distance between pairs of elements. For example, in the sequence $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow f$, at a distance of three elements, relationships would be calculated over the pairs a and d, b and e, and c and f.

On average, as the distance between elements increases, statistical dependencies weaken. Across many different sequence types, including phonemes, syllables, and words in both text and speech, the decay of long-range correlations and MI in language follows a power law (Eq. 5.6) [381, 7, 6, 249, 151, 391, 106, 4, 284, 298, 299, 297, 246, 78, 400]. This power-law relationship is thought to derive at least in part from the hierarchical organization of language, and has been variously attributed to hierarchical structure in human language syntax [249], semantics [7], and discourse structure [6].

To understand the link between hierarchical sequential organization in language and long-range sequential dependencies, it is helpful to consider both the latent and surface structure

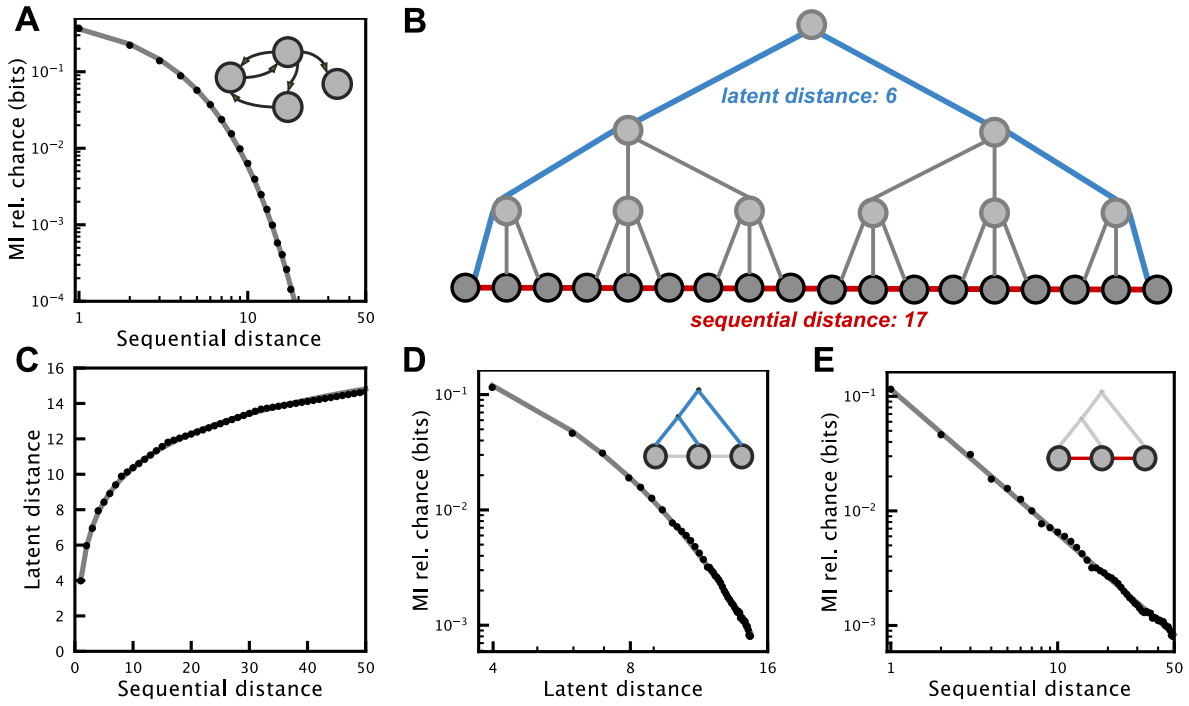


Figure 5.1. Comparison of long-range statistical dependencies between sequences with and without deep latent relationships. (A) The MI between elements in an iteratively (Markov model) generated sequence decays exponentially as a function of sequential distance. (B) An example sequence with hierarchical latent structure. The latent distance between the two end elements in the sequence is 6 (blue), while the sequential distance is 17 (red). (C) In sequences with hierarchical latent structure, the sequential distance between elements is logarithmically related to the latent distance (fit model: $a * log_{x*b} + c$ where x is sequential distance). (D) Like sequential distance in (A), The MI between elements in a hierarchically generated sequence decays exponentially as a function of latent distance. (E) The MI between elements in a hierarchically generated sequence decays following a power law as a function of sequential distance, which is related to the exponential MI decay seen in (D) and the logarithmic relationship between sequential and latent distance seen in (C). In (A), the probabilistic Markov model used to generate the empirical data has 2 states with a self-transition probability of 0.1. In (C-E) a probabilistic context-free grammar [249] with the same transition probability is used.

of a sequence (Fig. 5.1). When only the surface structure of a sequence is available, as it is for language corpora, a power-law decay in the MI between sequence elements gives evidence of an underlying hierarchical latent structure. This phenomenon can be demonstrated by comparing the MI between elements in a sequence generated from a hierarchically-structured language model, such as a probabilistic context-free grammar (PCFG), to the MI between elements in a sequence generated by a non-hierarchical model, such as a Markov process (Fig. 5.1). For sequences generated by a Markov process, the strength of the relationship between elements decays exponentially (Eq. 5.5) as sequential distance increases [249, 245] (Fig. 5.1A). In the PCFG model, however, linear distances in the sequence are coupled to logarithmic distances in the latent structure of the hierarchy (Fig. 5.1B-C). While information continues to decay exponentially as a function of the distance in the latent hierarchy (Fig. 5.1D), this log-scaling results in a power-law decay when MI is computed over corresponding sequential distances (Fig. 5.1E).

The thought that human language syntax is generated by CFGs [68] has led many to speculate that the long-range dependencies observed in language corpora are the product of abstract linguistic structure [7, 6, 249, 381]. Although the long-range statistical dependencies in language corpora are clearly tied to linguistic structure [6, 7], it does not follow that this structure is necessarily the only source for long-range dependencies in language. Indeed, hierarchical organization is unique to neither CFGs nor human language, and diverse classes of mechanisms, many of which are decidedly not language-like [318, 417, 125, 34, 311, 146], are capable of generating power-law relationships. Many non-linguistic human behaviors [77, 124, 458, 40, 237, 221], animal behaviors [83, 355, 57, 242, 29, 243], animal vocalizations [201, 368, 266, 166, 390, 419, 186, 50, 121, 305], and other biologically-generated processes [341, 412, 247, 340, 449, 210, 146, 59] are organized hierarchically. Likewise, long-range, power-law distributed dependencies are observed in sequential behaviors, including whale song [419], birdsong [381, 259], and Drosophila [29] and Zebrafish motility [142]. Instead, long-range dependencies in language and other human behavior [85, 403, 237] may reflect more

general biological processes inherited from the organization of underlying neurophysiological mechanisms [434, 211, 41, 439] that are, in turn, characterized by power-law relationships in temporal sequencing [253, 163, 332]. When viewed as an instance of this more general class of sequentially organized behavior, one might reasonably predict that human speech should display long-range statistical dependencies independent of linguistic structure.

To test whether long-range statistical dependencies occur independently of complex linguistic structure in speech we used MI decay as a measure of long-range dependencies over several speech corpora from children ranging from 6 months of age to adults [371, 263, 82, 335, 468, 48, 61, 43, 90, 271, 294, 143, 308, 320]. Because complex linguistic productions emerge during language acquisition, we use these corpora to determine whether long-range relationships are present in human vocalizations prior to the production of linguistically complex speech, or whether they emerge alongside linguistically complex productions. If long-range dependencies were to emerge over the course of development alongside complex utterances, we could conclude that abstract linguistic structure plays a dominant role in the sequential statistical structure of speech. However, if long-range statistical dependencies are observed in infant speech prior to the production of structurally complex utterances, then it is likely that the long-range dependencies observed in adult speech are not solely governed by abstract linguistic structure. Indeed, we find that human speech exhibits long-range power-law statistical dependencies like those observed in mature human language early in development, at 6 to 12 months of age, while children are still in the ”babbling” stage of language development.

5.2 Methods

5.2.1 Datasets

We examined MI decay in sequences of words over nine datasets of natural speech from English speaking children included in the CHILDES repository [263, 61, 43, 90, 271, 294, 143, 308, 320] and three datasets of sequences of phonemes from the PhonBank repository

[371, 82, 335, 468], both of which are part of the TalkBank repository [263]. Each dataset within CHILDES and PhonBank was collected in a slightly different manner. In our analyses, we included only transcripts of spontaneous speech that were collected from typically-developing children (usually at an in-home setting with family or an experimenter). The subset of CHILDES we used includes word-level transcripts of speech from children aged 12 months to 12 years of age. The subset of PhonBank we used includes phonetic transcriptions of speech given in the International Phonetic Alphabet (IPA) from children aged 6 months to four years of age. Between the phoneme and word-level datasets, a large range of speech and language development is covered.

For the MI analysis on phonemes, we binned transcripts into five 6-month age groups (6-12, 12-18, 18-24, 24-30, 30-36) and one age group from 3 years to 4 years. Each transcript was analyzed as sequences of phonemes, where phoneme distributions for each transcript are treated independently to account for variation in acquired vocabulary across individuals during development. Because transcript lengths varied between age groups (Fig. S1), we analyzed MI at sequential distances up to the median transcript length for each age group. For the MI analysis on words, we binned transcripts into four 6-month age groups (12-18, 18-24, 24-30, 30-36) and one age group from 3 years to 12 years. We analyzed words in the same manner as phonemes[1].

5.2.2 Mutual information

For each dataset, we calculate the sequential MI over the elements of the sequence dataset (e.g. words produced by a child). Each element in each sequence is treated as unique to that transcript to account for different distributions of behaviors across different transcripts within datasets.

Given a sequence of discrete elements $a \to b \to c \to d \to e$ We calculate mutual information as:

[1]No 6-12 month age group was used in word-level analyses because of the sparsity of word-level productions at that age

$$I(X,Y) = S(X) + S(Y) - S(X,Y) \tag{5.1}$$

Where X and Y are the distributions of single elements at a given distance. For example, at a distance of two, X is the distribution $[a,b,c]$ and Y is $[c,d,e]$ from the set of element-pairs ($a-c$, $b-d$, and $c-e$). $\hat{S}(X)$ and $\hat{S}(Y)$ are the marginal entropies of the distributions of X and Y, respectively, and $\hat{S}(X,Y)$ is the entropy of the joint distribution of X and Y.

To estimate entropy, we employ the Grassberger [152] method which accounts for under-sampling true entropy from finite samples:

$$\hat{S} = \log_2(N) - \frac{1}{N}\sum_{i=1}^{K} N_i \psi(N_i) \tag{5.2}$$

where ψ is the digamma function, K is the number of categories of elements (e.g. words or phones) and N is the total number of elements in each distribution.

We then adjust the estimated MI to account for chance. To do so, we subtract a lower bound estimate of chance MI ($\hat{I}_{sh}$):

$$MI = \hat{I} - \hat{I}_{sh} \tag{5.3}$$

This sets chance MI at zero. We estimate MI at chance ($\hat{I}_{sh}$) by calculating MI on permuted distributions of labels X and Y:

$$\hat{I}_{sh}(X,Y) = \hat{S}(X_{sh}) + \hat{S}(Y_{sh}) + \hat{S}(X_{sh},Y_{sh}) \tag{5.4}$$

X_{sh} and Y_{sh} refer to random permutations of the distributions X and Y described above. Permuting X and Y effects the joint entropy $S(X_{sh},Y_{sh})$ in I_{sh}, but not the marginal entropies $S(X_{sh})$ and $S(Y_{sh})$. $\hat{I}_{sh}$ is related to the Expected Mutual Information [448, 180, 447] which accounts for chance using a hypergeometric model of randomness.

Importantly, MI calculated over a sequence as a function of distance is referred to as a

"mutual information function", to distinguish it as the functional form of mutual information, which measures the dependency between two random variables [246]. In the mutual information function, samples from the distributions X and Y are taken from the same sequence, thus they are not independent. MI as a function of distance acts as a generalized form of the correlation function that can be computed over symbolic sequences and captures non-linear relationships [246].

5.2.3 Fitting mutual information decay

We fit the three following models:

An exponential decay model:

$$MI = a * e^{-x*b} + f \tag{5.5}$$

A power-law model:

$$MI = c * x^{d} + f \tag{5.6}$$

A composite model of the power-law and exponential models:

$$MI = a * e^{-x*b} + c * x^{d} + f \tag{5.7}$$

where x represents the inter-element distance between units (e.g. phones or syllables).

To fit the model on a logarithmic scale, we computed the residuals between the log of the MI and the log of the model's estimation of the MI. We scaled the residuals during fitting by the log of the distance between elements to emphasize fitting the decay in log-scale because distance was necessarily sampled linearly as integers. Models were fit using the lmfit Python package [319] using Nelder-Mead minimization. We compared model fits on the basis of AICc and report the relative probability of each model fit to the MI decay [55, 381].

5.2.4 Controls

Datasets are organized hierarchically into transcripts, utterances, words, and phonemes allowing us to shuffle the dataset at multiple levels of organization. To ensure that our MI decay results are a direct result of the sequential organization of each dataset, we performed a control in each dataset in which we shuffled behavioral elements within each individual transcript at each hierarchical level. In addition, to ensure that long-range relationships were not due to trivial repetitions of behaviors, we looked in each dataset at MI decay over sequences in which repeated elements were removed. Finally, we analyzed transcripts from a subset of the longest individual transcripts to confirm that our results were not the product of mixing together multiple datasets and transcripts.

5.3 Results

Although much work has explored the information content and long-range sequential organization of human language, relatively few studies have examined these properties in speech [381] or language development directly. Here we investigate the long-range information present in speech during language development using datasets from the TalkBank project [371, 263].

We first examined MI decay in sequences of phones over three datasets of natural speech from English-speaking children included in the PhonBank repository. Across all age groups, starting at 6-12 months of age, the decay in MI over sequences of phonemes is best fit by a composite power-law and exponential decay model (Fig. 5.2A-C; relative probabilities 0.897 to ¿0.999; Table S2). In each age group, we observe both a clear power law prominent over long distances (Fig. 5.2B) and a clear exponential decay at short, word-length distances (Fig. 5.2C), consistent with prior results adult speech [381]. We then examined MI decay in sequences of words over nine datasets of natural speech from English-speaking children included in the CHILDES repository. As with phonemes, the MI decay between words is best fit by a composite model of power-law and exponential decay (Eq. 5.7; relative probability = 0.989 for 12-18

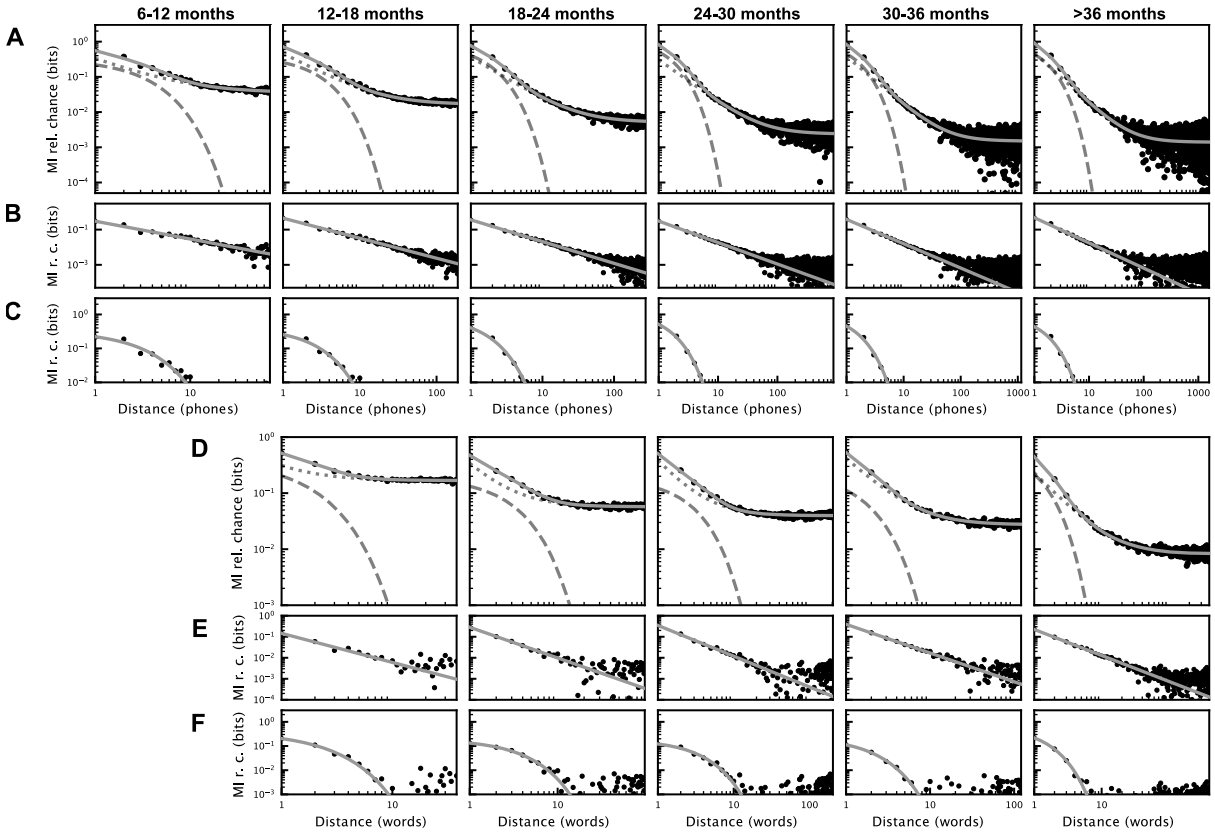


Figure 5.2. Mutual Information decay over words and phonemes during development. (A) MI decay over phonemes for each age group. MI decay is best fit by a composite model (solid grey line) for all age groups across phonemes and words. Exponential and power-law decays are shown as a dashed and dotted grey lines, respectively. (B) The MI decay (as in (A)) with the exponential component of the fit model subtracted to show the power-law component of the decay. (C) The same as in (B), but with the power-law component subtracted to show the exponential component of the decay. (D-F) The same analyses as A-C, but for words.

months and ¿ 0.999 for all other age groups; Fig. 5.2D-F; Table S1).

As controls, we also computed the MI decay over sequences of words and phonemes that had been shuffled to isolate sequential relationships at different levels of organization (e.g. phoneme, word, utterance, transcript; Figs. 5.3, 5.5, 5.6). A subset of these controls over the PhonBank dataset are shown in Fig. 5.3 while the remainder are given in Figs. 5.5 and 5.6. Consistent with Sainburg et al., [381], we observe that short-range relationships captured by exponential decay are largely carried within words and utterances, while long-range relationships captured by a power-law decay are carried across longer timescales between words

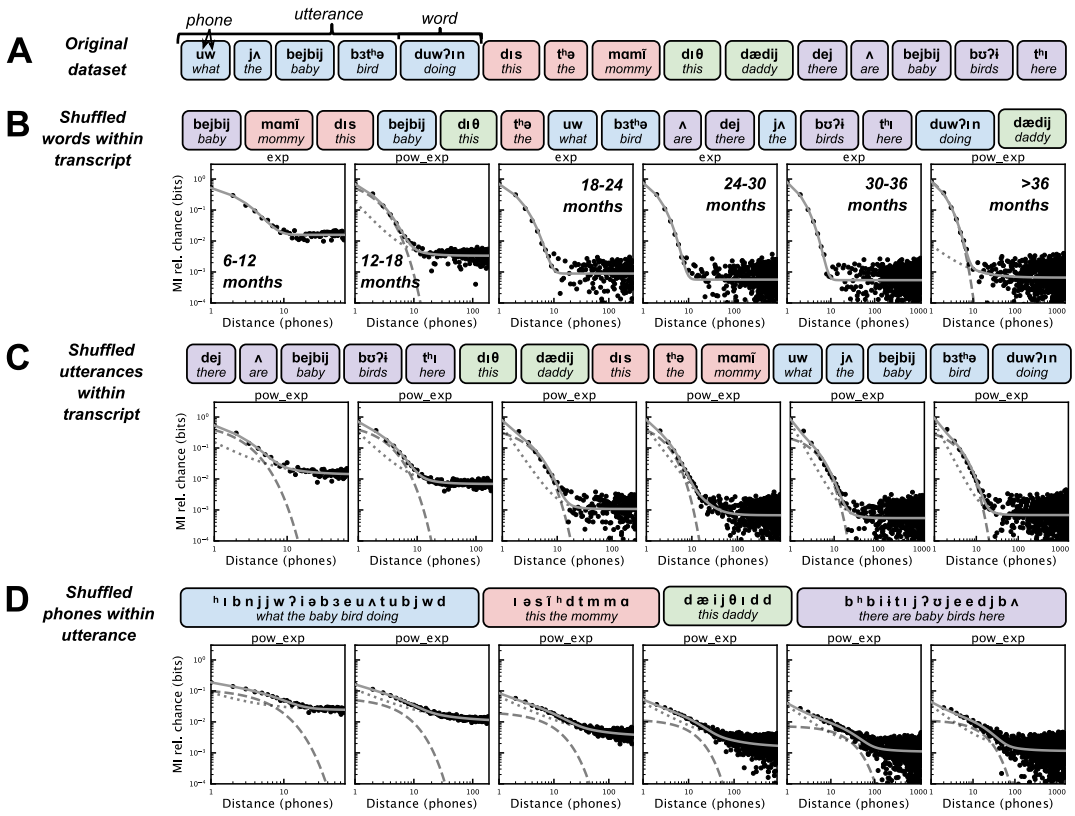


Figure 5.3. MI decay between phones under different shuffling conditions. (A) An example sequence of utterances from the PhonBank dataset. Utterances are grouped by color, words are grouped by rounded rectangles, and phones are displayed in bold above orthographic transcriptions. (B) MI decay, as in Fig 5.2 when words are shuffled within each transcript. (C) MI decay when utterances are shuffled within each transcript. (D) MI decay when phones are shuffled within each utterance. The best fit model is printed above each plot and is plotted as grey lies alongside the data.

and utterances. In particular, long-range relationships are eliminated when between-utterance structure is removed by randomly shuffling the order of words or utterances within a transcript (Figs. 5.3B-C, 5.6C) while short timescale exponential relationships are preserved. In contrast, long-timescale relationships are retained when within-utterance structure is removed by shuffling words phonemes or words within utterances (Figs. 5.3D, 5.6B) or phonemes within words (Fig. 5.5C), while short-timescale relationships are largely eliminated. When MI decay is computed over part-of-speech labels for the words in CHILDES transcripts, we find a transition from MI decay that is best-fit by a power-law decay alone at 12-24 months of age, to MI decay

that is best fit by a composite model of power-law and exponential decay after 24 months (Fig 5.6D). Shuffling phoneme order within transcripts removes all sequential relationships (Figs. 5.5F). Across each shuffle analysis, we observe that short-range information content captured by exponential decay is largely captured within words and utterances, while long-range information is carried between utterances, even during early language production.

As an additional control, to ensure that the observed MI decay patterns are not the product of mixing datasets from multiple individuals, we also computed the MI decay of the longest individual transcripts comprising each age cohort across both phonemes and words. The decay of the longest individual transcripts parallels the results across transcripts shown in Fig. 5.2 (Figs. S5, S6). We also analyzed the MI decay of transcripts when repeated elements were removed to ensure long-range relationships were not the product of behavioral repetitions. Removing repeats does not qualitatively alter the pattern of long-range relationships between elements (Fig. 5.7).

One reasonable hypothesis is that these long-range relationships in child speech are driven by interaction. Child speech is produced in an interactive context with adults, thus, adult speech could be driving the long-range relationships observed in child speech. If this were the case, one could argue that the complex hierarchical structure underpinning the adult's speech was driving the long-range dependencies found in infant speech. To test whether this is the case, for each corpus where adult speech was transcribed ($n_{CHILDES}$ =1630, $n_{PhonBank}$ =309) we tested the effect of non-subject engagement on the improvement in model fit (ΔAICc) of a power-law model over exponential model alone. In both datasets, we observe that adult involvement (the proportion of speech not produced by the child) provides no additional predictive information about the improvement in fit of the power-law model over the exponential model, when controlling for the dataset, child's age, and length of the transcript (CHILDES: F(1,1620)=1.49, p=0.22; PhonBank: F(1,306)=0.21, p=0.65). Although our results do not provide irrefutable evidence that the long-range range relationships observed are driven by adult speech, these results do not rule out the possibility. Our analyses were based on the natural variability in adult speech across corpora and are not explicitly controlled.

5.4 Discussion

We analyzed the long-range sequential information present in speech during child development. We observed adult-like long-range statistical relationships [381] present as early as 6 to 12 months in phoneme sequences, and at 12-18 months in word sequences, preceding the production of complex linguistic structure [156]. Thus, long-range statistical relationships in speech cannot be the unique product of complex linguistic productions.

These results compel reconsideration of the mechanisms that shape long-range statistical relationships in human language. Traditionally, the power-law decay in information between the elements of language (phonemes, words, etc.) has been thought to be imposed by the hierarchical linguistic structure of syntax, semantics, and discourse [7, 249, 6]. Early speech development provides a natural experiment in which one can examine human vocal communication absent the production of complex syntactic and semantic structures. Remarkably, even at a very early age, prior to the production of mature syntactic structures, vocal sequences show adult-like long-range dependencies. This does not rule out the possibility that long-range dependencies in adult language are driven in part by linguistic structures, but the absence of these organizational components in the youngest children indicates that other mechanisms very likely shape the long-range structure of speech. Whether these early mechanisms are replaced by more classical hierarchical linguistic structures over the course of development or remain important throughout life remains a topic for future research. It is possible that multiple mechanisms impose long-range dependencies on human speech and language, and that these operate on different developmental timescales. The observation of similar power laws in diverse non-linguistic behaviors reinforces the idea that multiple mechanisms can shape the sequential dynamics of behavior, including speech. At the same time, we note that there are many potential sources for long-range correlations in biological and physical systems that do not guarantee an underlying hierarchical structure [318, 417, 125, 34, 311, 146]. While our results are consistent with the notion that linguistic structure is overlaid on a more general, hierarchically organized, motor

control structure, it is possible that the long-range dependencies observed in young children reflect other forms of underlying processes, and that the only dependencies relevant to language are those that emerge later with adult-like linguistic structures. Although possible, this latter idea seems less parsimonious as it would involve a reduplication of dependencies that already exist in the signal.

It is also possible that the long-range structure we observe is driven by external and environmental factors, such as long-range statistics in the child's linguistic environment. For example, in animal behavior, long-range statistical relationships between behavioral states can be affected by variables such as lighting environment [142]. While we did not observe that long-range relationships were driven by interactions with adult speakers, we do not rule out language interaction or other exogenous variables as a possible driver for the observed long-range relationships.

Regardless of any further understanding of the specific mechanisms that underlie the sequential dependencies in speech, clear patterns in the information conveyed across time exist in the non-linguistic human vocal behavior. In principle, this structure is available to listeners and can provide predictive information to any cognitive agent that engages with it. Humans are necessarily sensitive to long-range relationships in language, and although more sparse, evidence for long-range sensitivities outside language has also been reported, such as scale invariance in retrospective memory tasks [274] and attention to power-law timescales in anticipation of future events in cognitive tasks [413]. Among non-human animals the evidence supporting sensitivity to the long-range dynamics (power-law or otherwise) of information in the environment is not well studied, especially at long intervals. If non-human animals can perceive the long-range statistical dependencies present in their environment, this capacity may constitute a broad faculty of language [162], that is, a necessary, but not uniquely human, component of language. Indeed, the presence of long-range statistical dependencies in non-linguistic behaviors and a generalized perceptual sensitivity to them could provide a scaffold for language to evolve, and where hierarchical syntax and semantics can be understood as later additions that exploit existing

long-range structures and sensitivities.

5.4.1 Data Availability

The datasets can be acquired from the TalkBank repository [263] and PhonBank repository [371]. We performed analyses over these transcripts without any modification. Example transcripts for each dataset are displayed in the Supplementary Information. The distribution of sequence lengths of each dataset is shown in Fig. S1. The code necessary for reproducing our results is available on GitHub [377].

5.4.2 Author contributions

T.S., A.M., and T.Q.G. developed the study concept, contributed to the study design, and wrote the manuscript. Data analysis was performed by T.S. All authors approved the final version of the manuscript for submission.

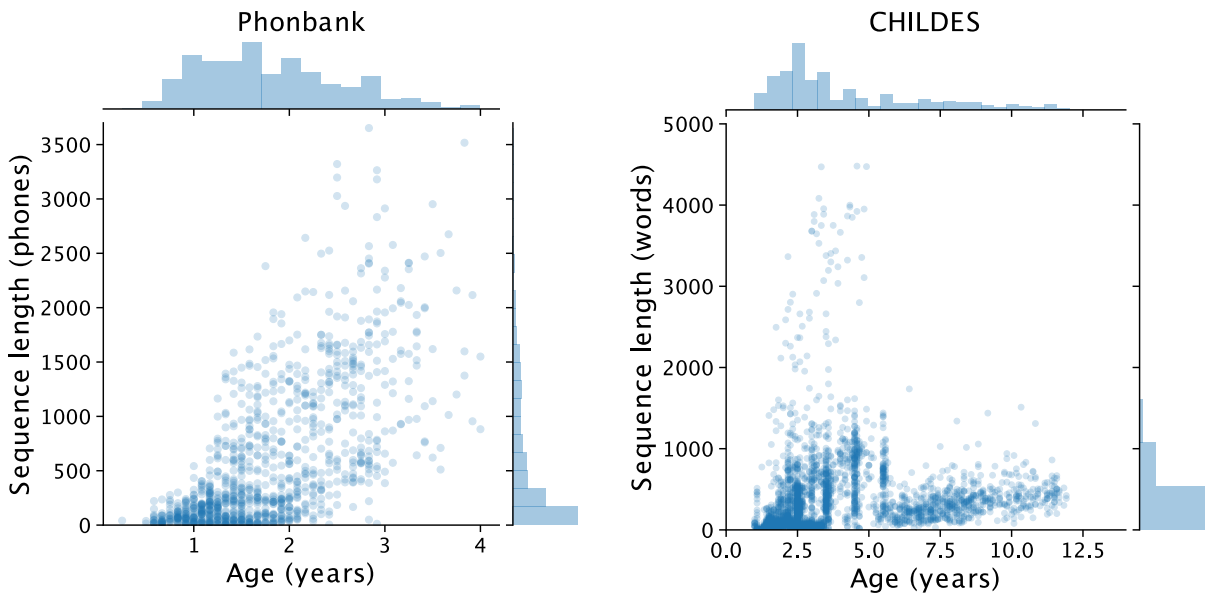


Figure 5.4. Distribution of sequence lengths for each dataset.

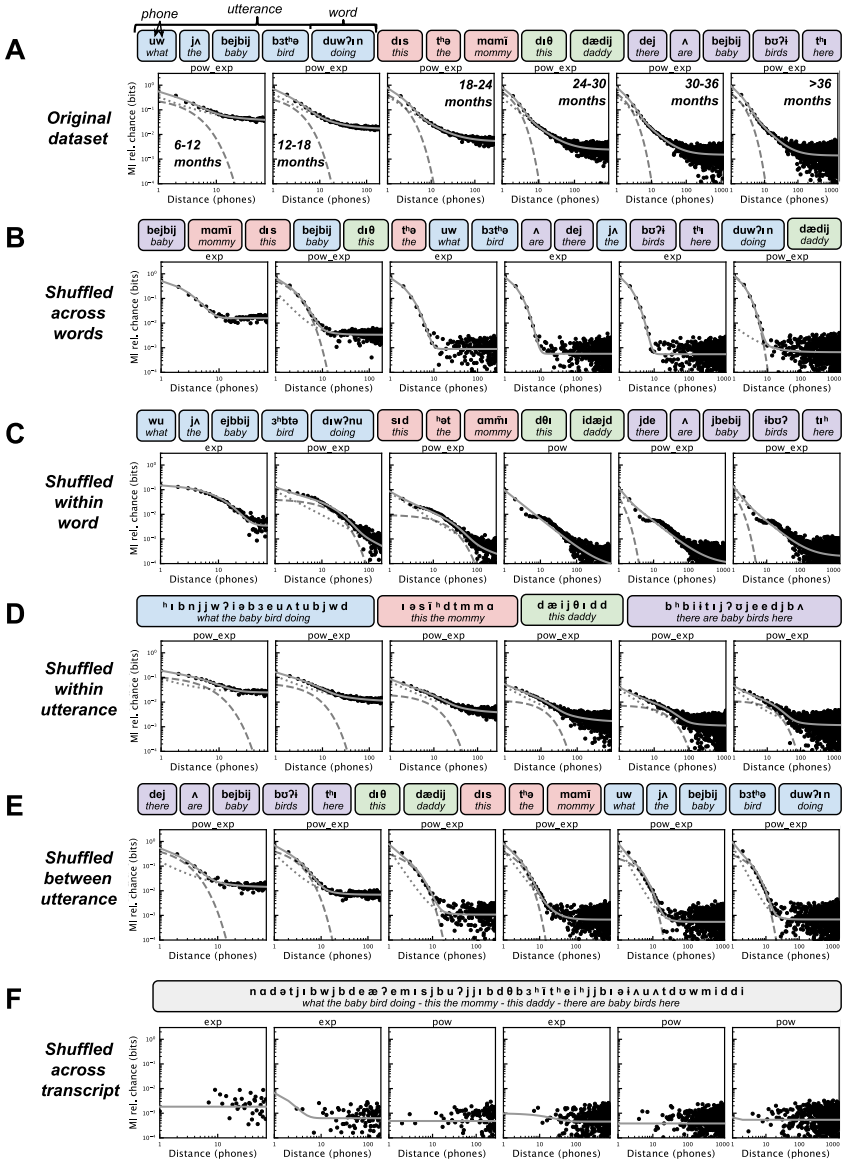


Figure 5.5. MI decay between phones under different shuffling conditions. (A) MI decay for each age group from the entire dataset, as in Fig. 2A. The sequence above the MI decay shows an example set of utterances of the corpus to illustrate the shuffling conditions. Utterances are grouped by color, words are grouped by rounded rectangles, and phones are displayed in bold above orthographic transcriptions. (B) Words are shuffled within each transcript. (C) Phones are shuffled within words. (D) Phones are shuffled within utterances. (E) Utterances are shuffled within each transcript. (F) Phones are shuffled within each transcript. The best fit model is printed above each plot, and is plotted as grey lies alongside the data.

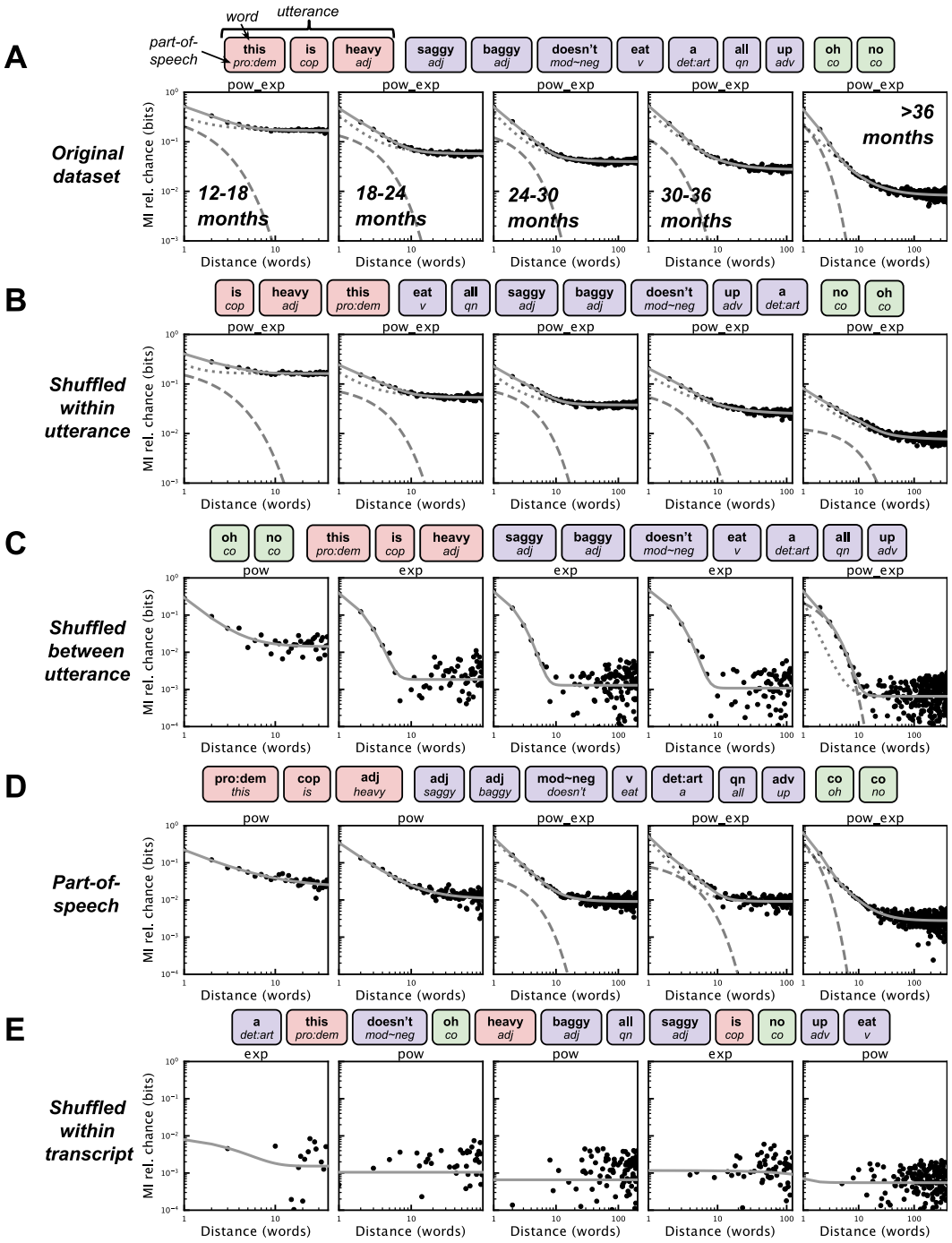


Figure 5.6. MI decay between words under different shuffling conditions. (A) MI decay for each age group from the entire dataset, as in Fig. 2D. (B) Words are shuffled within each utterance. (C) Utterances are shuffled within each transcript. (D) MI is calculated over part-of-speech transcriptions of words. (E) Words are shuffled within each transcript. (F) Words are shuffled within each transcript. The best fit model is printed above each plot, and is plotted as grey lies alongside the data.

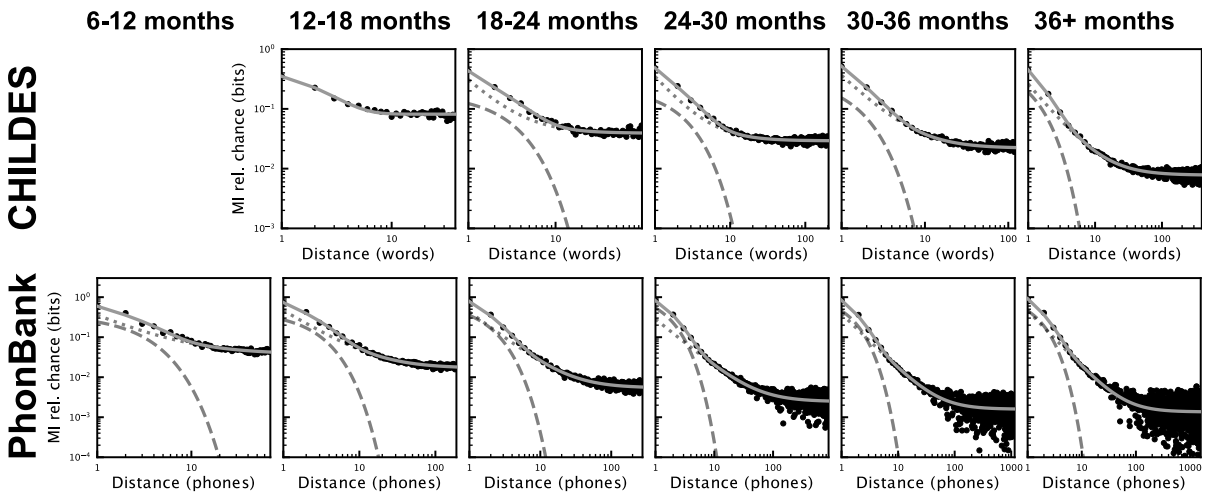


Figure 5.7. MI decay with repeated elements removed across each dataset.

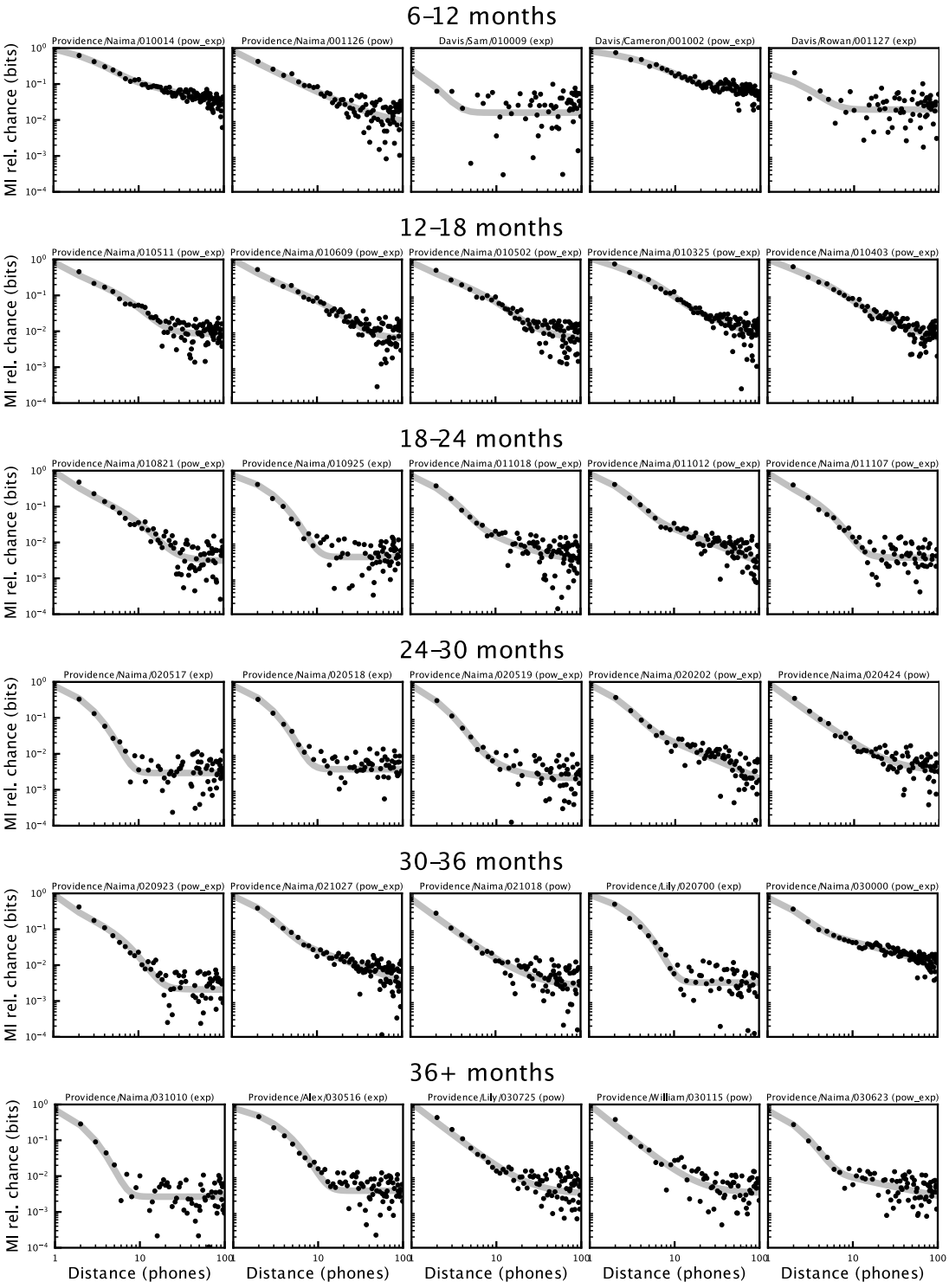


Figure 5.8. MI decay and best fit model of five largest transcripts for each age group across PhonBank. Transcript identity and best fit model are displayed above each plot.

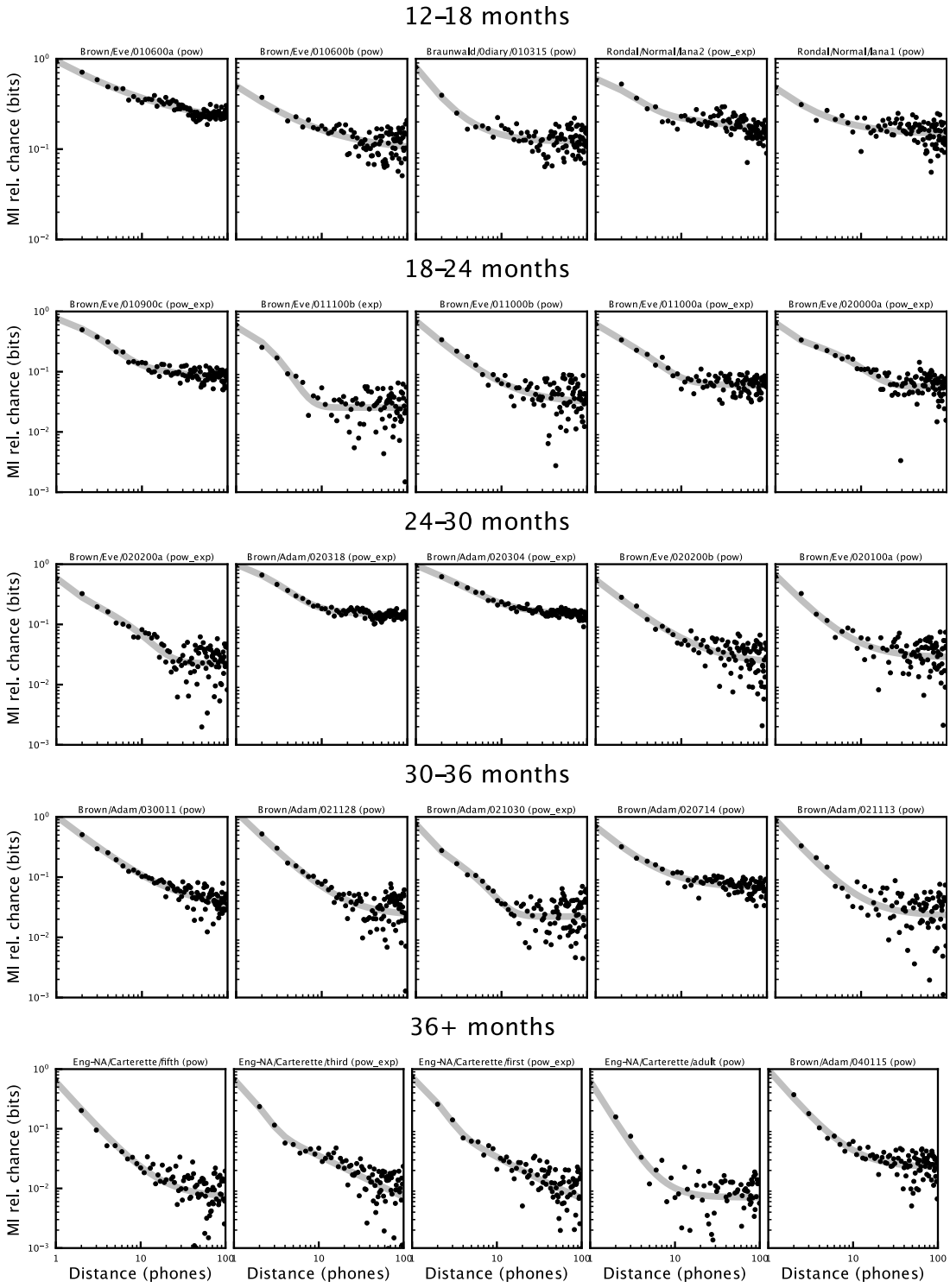


Figure 5.9. MI decay and best fit model of five largest transcripts for each age group across CHILDES. Transcript identity and best fit model are displayed above each plot.

5.5 Acknowledgments

Chapter 5, in full, is a reprint of a manuscript under review. Sainburg, Tim, Mai, Anna, Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Chapter 6

Prediction and probabilistic integration underlie learned context-dependent categorical vocal sequence perception and sensory physiology

Abstract

To distinguish between vocal elements in communication, both songbirds and humans rely on categorical perception of smoothly varying acoustic spaces which can be biased by sequential context. The cognitive and physiological mechanisms by which this bias occurs are not well understood. We developed a behavioral task that modulates the predictive probability of birdsong sequences, training European starlings to classify ambiguous syllables synthesized from samples in the latent space of a neural network, in the context of varying predictive sequential information. We find that song predictability biases perceptual classification of syllables, following a Bayesian model of information integration. Using the same behavioral task, we then chronically recorded from populations of auditory neurons while birds were engaged in the task. We find that sensory neurons capture the uncertainty, or likelihood, in perceptual decision-making and are modulated by the predictive information present in the syllable sequences. Rather than integrating the prior and likelihood, as is seen in the animal's behavior, the modulation of neurons is consistent with an increase in perceptual acuity in the likelihood in higher probability regions of acoustic space.

6.1 Introduction

Categorical perception (CP), the grouping of smoothly varying signals into discrete classes, plays an important role in organizing complex experiences into a shared representational space by enabling the abstraction and generalization of individual instances of a signal to other instances. In humans, auditory categorical perception is fundamental to communication through speech. The acquisition of categorical perceptual boundaries in speech is learned and differs across languages. Infants, for example, can discriminate phonetic boundaries in speech that adults cannot [225]. Our perception of phonemes does not occur in isolation, however. Phonemes occur in the context of words, utterances, discourse, and environmental contexts which provide important additional cues enabling the speech to be correctly perceived. When sensory information is ambiguous, predictions and prior knowledge bias perception towards more likely scenarios. For example, the Ganong effect [136] describes the tendency to shift categorical perception of ambiguous phonemes based upon our expectations about the words they belong within. The same ambiguous phoneme between '/b/' and '/p/' is more likely to be perceived as 'peace' than 'beace' but less likely to be perceived as 'peef' than 'beef', because 'peace' and 'beef' are words, and 'beace' and 'peef' are not. Context-dependent categorical perception in speech is also driven by the sequential organization of speech elements. For example, the categorical perception of phonemes is modulated by their position within words and relative to other phonemes [270].

Categorical perception is not unique to human speech and has been observed in a number of sensory modalities and species [155, 225, 270]. Songbirds perceive some elements of song categorically [350, 270] and categorical perception of song elements can be biased by sequential context. For example, Lachlan and Nowicki [233] used playback experiments with swamp sparrows to demonstrate that, like speech perception, categorical perceptual boundaries of notes are modulated by their position within a song. Swamp sparrow songs are learned and preserved across populations for generations [234], suggesting that perceptual categories and the

modulation of categorical perception may also be learned.

In speech, it has been proposed that phoneme categorical perception can be understood Bayesian inference over acoustic distributions [119, 324, 229]. Under this framework, biasing of speech perception through prediction, top-down influences, and prior expectations are modeled as probabilistic integration. The specific cognitive and neural mechanisms underlying how predictive information and context-dependency modulate categorical perception of speech are not well-understood, however [324].

In songbirds, less is known about the mechanisms underlying categorical perception. For example, whether predictive information biases perception and whether perceptual biases can similarly be explained through Bayesian integration, is unknown. Physiological investigations into how predictive information in the songbird sensory system could provide insight into the physiological mechanisms underlying predictive and context-dependent categorical perception. For example, a neural correlate for categorical perception has previously been described in the sensorimotor system of the songbird [350]. Prather et al., [350] found that in swamp sparrows, categorical neural responses measured as action potential per stimulus, reflect natural vocal boundaries in neurons in the auditory-motor nuclei HVC, which project to the striatal nuclei Area X.

In this manuscript, we developed a paradigm to explicitly impose probabilistic predictive information in a sequence of birdsong syllables and trained European starlings to classify ambiguous syllables under varying sequential predictive information. We hypothesized that songbird perception would follow a Bayesian account of information integration, modulating perceptual boundaries as a function of predictive information.

We found that songbirds do integrate prior sequential information in their perception of vocal signals and that the bias caused by this information integration is modulated by the strength of the context and the ambiguity of the signal. Response characteristics of the birds reflect all aspects of a Bayesian account of sequential information integration, including the prior probabilities, likelihoods, and posterior.

Using this behavioral paradigm, we then explored the physiological underpinnings of categorical perception and context integration. We chronically recorded spiking activity from populations of neurons in the primary auditory nuclei Field L, two secondary auditory nuclei that bidirectionally project with Field L, CM, NCM, as well as NCL, a broad region lateral to NCM that has bidirectional projections with Field L, and has variously been shown to be involved in visual and multi-modal working memory [160, 223]. We found that neural responses reflected variation in the behavioral uncertainty underlying decision making (i.e. the likelihood of the Bayesian model) and that neural responses to the categorized syllable are modulated and biased by the predictive information imposed by the preceding syllable. This bias was consistent with a Bayesian model exhibiting increased acuity in the likelihood in high-probability regions of acoustic space.

6.2 Results

6.2.1 Paradigm

We developed a context-dependent categorical perception paradigm in which birds classified smoothly varying morphs of syllables of birdsong in the context of sequential predictive information. We modeled this behavioral paradigm using a Bayesian perceptual decision-making framework and used this model as the basis of our behavioral and physiological hypotheses.

To implement this context-related CP shift in a natural stimulus environment, we created a two-alternative choice (2AC) category learning task in which songbirds were trained to classify stimuli on a single dimension, represented by a smoothly varying syllable of European starling song generated from a linear interpolation in the latent space of a deep convolutional Variational Autoencoder (Fig 6.1A) [205]. We chose to use a neural-network-based approach to synthesize songs rather than a single-dimensional feature like voice-onset-time, pitch-shifts, tones, or duration, because most vocal signals, including the song of European starlings, do not vary along single, linear, dimensions.

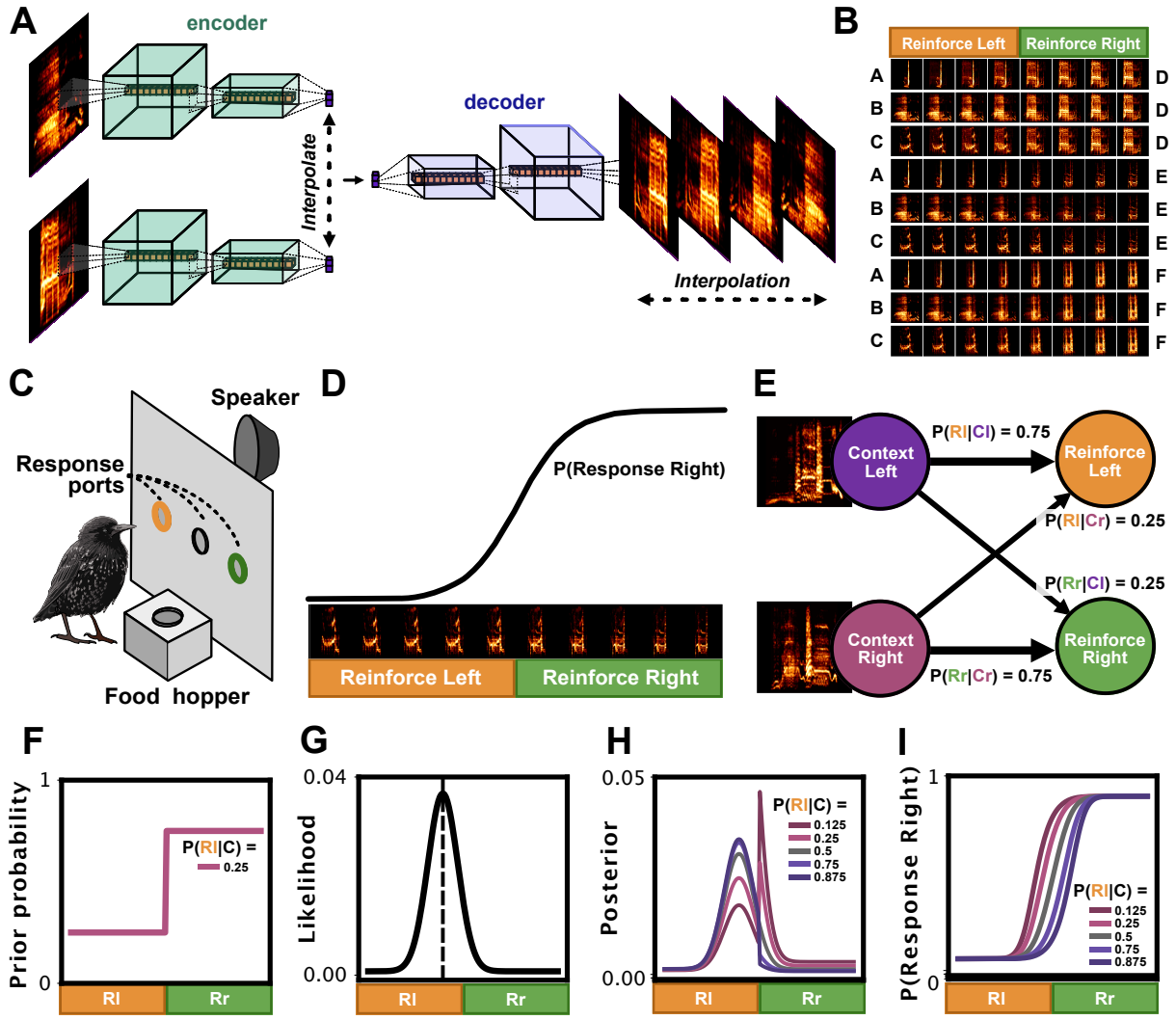


Figure 6.1. Overview of behavior and hypothesis. (A) Stimuli morphs are generated as interpolations projections of two song syllables in the latent space of a neural network. (B) Samples from the 9 morphs (rows) used the birds are trained on. The reinforced category is shown above the morphs, and the endpoints are labeled on the left and right of the morphs. (C) The behavioral apparatus used for this experiment. The green and orange response ports correspond to the stimuli classes in (B). (D) A psychometric curve depicting stimuli classification over one morph. (E) Two example context cue syllables precede the reinforced syllables, holding predictive information about the class they belong to. (F-I) A Bayesian model depicting our hypothesis. (F) An example prior probability represents the probability of a morph stimulus given by the preceding cue syllable (here the cue predicts a right stimulus). (G) the likelihood is given by a Gaussian distribution centered around the true syllable presented. (H) the posterior probability under the five cue probabilities used in this study. (I) The predicted behavior response under the Bayesian model, depicting a shift in categorical perceptual decision making as a function of the cue probability.

The morphs generated from this interpolation (Fig 6.1B) were divided into at the halfway point in the interpolation, with the first half of the morph being reinforced with a food reward after pecking into the left response port and the second half of the morph being reinforced after a peck to the right response port (Figure 6.1C bottom).

After training the birds on the initial classification task, yielding a psychometric function of classifications over each morph (Fig 6.1D), a cue syllable was added preceding the target classified syllable (i.e. the morph). Each cue syllable provides predictive information about the category of the target stimulus (Fig 6.1E)).

We modeled our hypothesis of the effect of the cue syllable on the psychometric curve as Bayesian integration (Fig 6.1F-I). By treating this cue stimulus as a prior probability over the morph (Fig 6.1F) and representing the uncertainty over the stimulus in the morph as Gaussian probability distribution centered around the true stimulus (Fig 6.1G), we predicted that the posterior probability given sensory information and the cue stimulus (Fig 6.1H) would shift the classification of stimuli near the boundary between the two classes in the direction predicted by the cue stimulus (Fig 6.1I). If observed, this shift in classification based upon cue information represents a shift in categorical perceptual decision-making through the integration of temporal contextual information.

We trained a total of 20 European starlings on our behavioral task, performing a total of 4.8 million behavioral trials. Each subject learned the task to at least 75% accuracy (Table 6.1).

6.2.2 Context dependent shift in perceptual decision making

For each bird and morph we fit a psychometric 4-parameter logistic function (Fig. 6.2A) to each subject's classifications. We used the parameters of the fit psychometric model to test the Bayesian model's cue-dependent perceptual shift prediction. We contrast this model of behavior with an alternative decision-making strategy, in which information is not temporally integrated with the cue (Fig 6.2B). Under the Bayesian hypothesis, classification of the reinforced syllable will be modulated by integrating the likelihood imposed by the stimulus with the prior imposed by

the sequential cue (Fig 6.2B top). Under the alternative non-integration hypothesis, Information from the cue and reinforced syllable will not be integrated, but treated as independent, resulting in an overall shift in the probability of a left or right classification, but not a shift in the decision boundary (inflection point; Fig 6.2B bottom). Across subjects, we observe both a shift in the inflection point, indicative of Bayesian integration of cue and reinforced syllable, as well as an overall shift in decision probability, indicative of a mixed reliance on cue versus reinforced syllable (Fig 6.2C). Across each morph for each bird, we find s robust shift in the inflection point (Fig. 6.2D; t=18.5, p=6.9e-54, n=350). This shift increases as the cue's prediction probability increases (Fig. 6.2E; r^2=0.446, p = 1-52, n=1050).

The observed cue-related shift is predicted by the Bayesian model. For each bird's classification of each morph, we fit the Bayesian model to the behavioral data and predicted the inflection point shift given each cue probability. The red dashed line in Fig. 6.2E depicts a linear regression between the observed psychometric shift and the predicted inflection point shift from the Bayesian model's predictions.

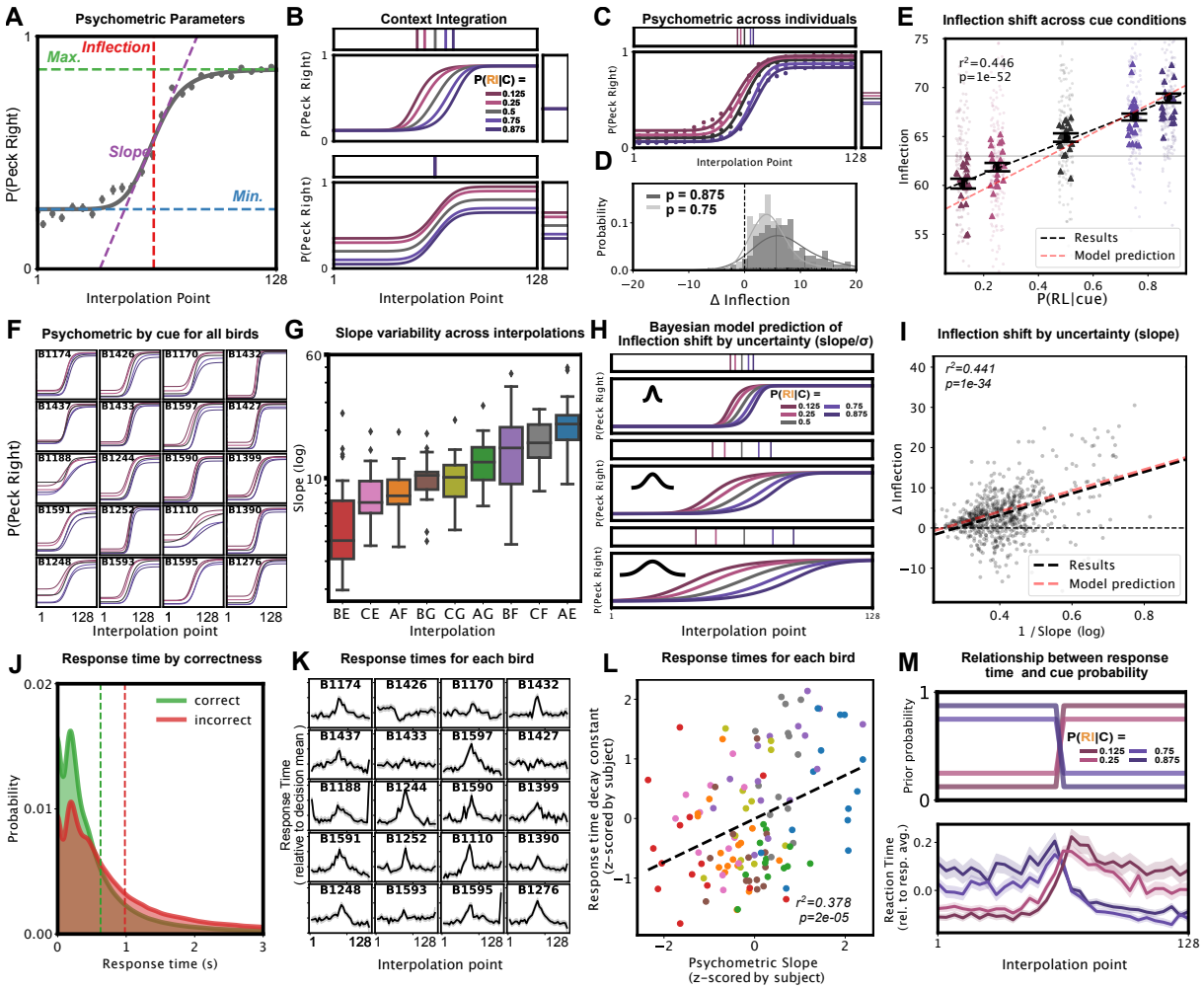
6.2.3 Context dependent perceptual shift increases with uncertainty

Across subjects, we observe substantial variation in the slope of the psychometric function fit to the bird's behavior. Some individuals drew a much sharper categorical boundary than others, as exhibited by a greater slope in the fit psychometric (e.g. B1432 vs B1110 in Fig. 6.2F). The slope of the psychometric also varies from morph to morph (Fig. 6.2G). The slope of the psychometric reflects uncertainty in the Bayesian model. Under greater uncertainty about the reinforced syllable, the Bayesian model predicts that integration of the cue stimulus will result in a greater shift in categorical perception (i.e. the inflection point; Fig. 6.2H [35]). Our empirical results support that hypothesis, with a smaller inflection point shift in the direction of the cue as the slope of the psychometric steepens (Fig. 6.2I, r^2=0.441, p=1e-34, n=700). The red dashed line in Fig. 6.2I dashed line represents the Bayesian model's prediction of the relationship between the cue shift and the slope of the psychometric.

6.2.4 Reaction time represent likelihood and prior probability

In addition to the animal's decisions, we observe that the response time in making a decision reflects both the uncertainty in decision making (the slope of the psychometric) as well as the prior probability given by the cue. We observe that the response time is greater in incorrect trials than correct trials (Fig. 6.2J) and that in most of the subjects (17/20) the response time increases with proximity to the categorical boundary, indicating the increased difficulty in classification (Fig. 6.2J). For each bird and morph, we fit an exponential decay model of reaction time as a function of distance from the categorical boundary. In morphs where a decay was observed (set at an r2 $>$ 0.001 and decay range $>$ 0.1 standard deviations) we found a strong relationship between the exponential decay constant, and the psychometric slope (Fig. 6.2L; r^2 = 0.378, p=2e-5, n=129). Finally, we observe that the response timing is related to the prior probability imposed by the task (Fig. 6.2M). Across subjects, response times are fastest when the reinforced stimulus is cued, and slowest when the cue is strongest in the opposite direction.

Figure 6.2. Overview of behavioral results. (A) An example psychometric fit with parameters. (B; top) An example of the context-dependent category shift as a function of cue information hypothesis, as predicted by the Bayesian model. (B; bottom) An example of an alternative hypothesis, in which decisions are made either using the cue or the categorical stimuli, without integration of the two sources of information results in no category boundary shift. The corresponding lines in the connected horizontal and vertical boxes indicate the shift in the inflection point (vertical lines) as well as the and midpoint between mid and max in the psychometric function (vertical lines). Colors indicate the cue probabilities given in Figure 6.1. (C) The results across birds and morph indicate that both strategies from (B) are present in behavior. (D) Cue shift between left and right cues for each morph and bird at p=0.875 and 0.75. (E) The categorical boundary (inflection point) shifts as a function of the strength of the cue. The Bayesian model, predicts a similar shift from the uncued data. (F) Psychometric fits for cued conditions for each of the subjects. (G) Morphs (interpolations) vary on the slope of the fit psychometric function, indicating variation in uncertainty in decision making by morph. (H) The Bayesian model predicts a greater shift in categorical boundary as a function of the uncertainty of the categorical stimulus (σ of the likelihood and slope of the psychometric model). (I) As predicted by the Bayesian model, the shift in the categorical boundary increases as a function of uncertainty.) (J) Response time across birds for correct versus incorrect trials. (K) Response time over the morph for each bird. (L) Decay constants of exponential decay fit to reaction time as a function of distance from decision boundary, in relation to the slope of the fit psychometric function, for each bird and morph. Point colors reflect the morph categories shown in panel (G). (M; top) The imposed prior probability in the task for each condition. (M; bottom) Reaction time over morph for each cue condition.


A
Psychometric Parameters
Max.
Inflection
Slope
Min.
P(Peck Right)
Interpolation Point
B
Context Integration
P(Peck Right)
P(R|C) =
0.125
0.25
0.5
0.75
0.875
Interpolation Point
C
Psychometric across individuals
P(Peck Right)
Interpolation Point
D
Probability
p = 0.875
p = 0.75
Δ Inflection
E
Inflection shift across cue conditions
r²=0.446
p=1e-52
Inflection
Results
Model prediction
P(RL|cue)
F
Psychometric by cue for all birds
P(Peck Right)
B1174
B1426
B1170
B1432
B1437
B1433
B1597
B1427
B1188
B1244
B1590
B1399
B1591
B1252
B1110
B1390
B1248
B1593
B1595
B1276
Interpolation point
G
Slope variability across interpolations
Slope (log)
BE CE AF BG CG AG BF CF AE
Interpolation
H
Bayesian model prediction of
Inflection shift by uncertainty (slope/σ)
P(Peck Right)
P(R|C) =
0.125
0.25
0.5
0.75
0.875
Interpolation point
I
Inflection shift by uncertainty (slope)
r²=0.441
p=1e-34
Δ Inflection
Results
Model prediction
1 / Slope (log)
J
Response time by correctness
correct
incorrect
Probability
Response time (s)
K
Response times for each bird
Response Time
(relative to decision mean)
B1174
B1426
B1170
B1432
B1437
B1433
B1597
B1427
B1188
B1244
B1590
B1399
B1591
B1252
B1110
B1390
B1248
B1593
B1595
B1276
Interpolation point
L
Response times for each bird
Response time decay constant
(z-scored by subject)
r²=0.378
p=2e-05
Psychometric Slope
(z-scored by subject)
M
Relationship between response
time and cue probability
Prior probability
P(R|C) =
0.125
0.25
0.75
0.875
Reaction Time
(rel. to resp. avg.)
Interpolation point

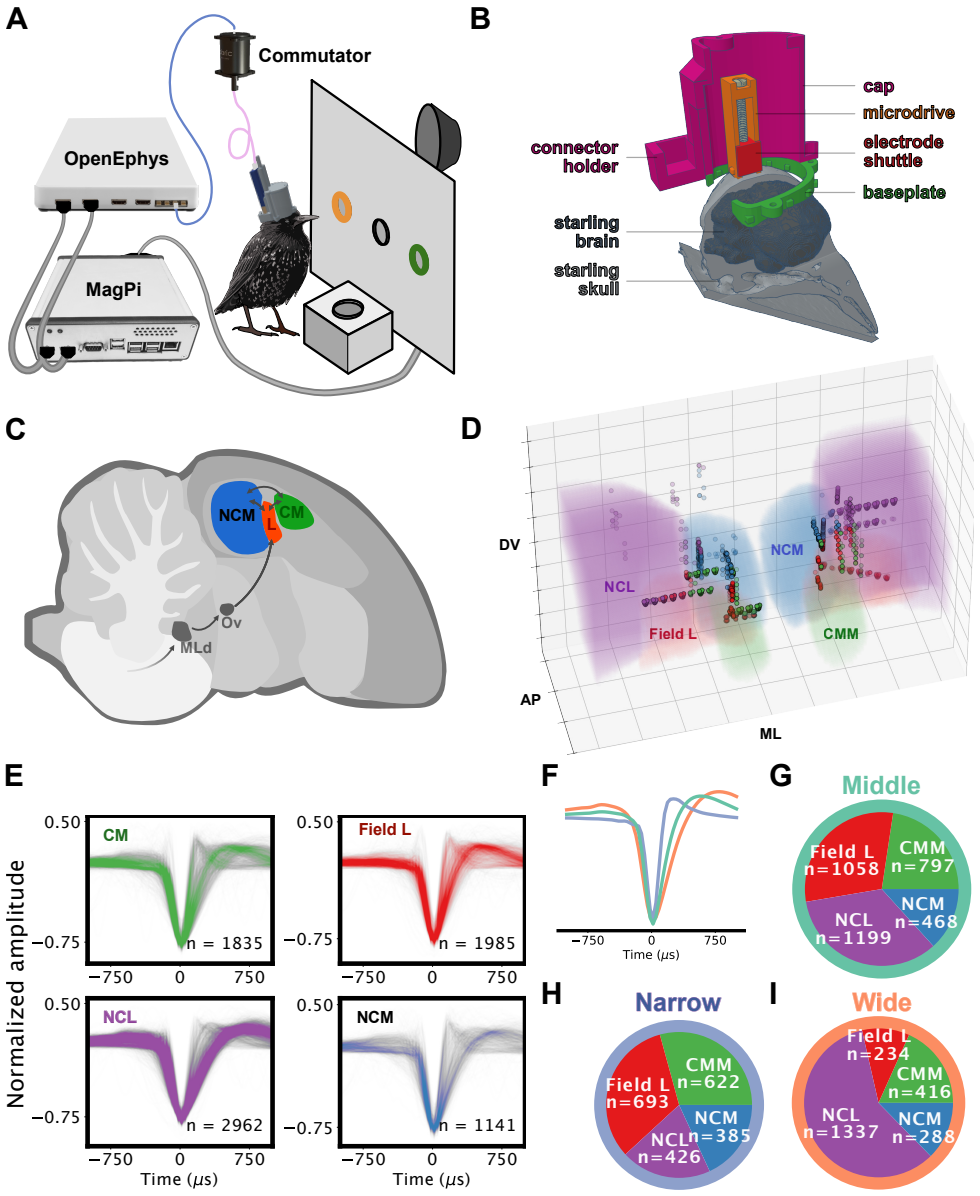


Figure 6.3. Overview of physiological paradigm and data set. (A) Continuous recordings are performed in free moving birds syncing physiology with operant conditioning behavior. (B) Birds are unilaterally or bilaterally implanted with 32-64 channel electrodes using 3D printed microdrives and protective head caps. (C) Nuclei OV projects to the primary auditory region Field L, which has bidirectionally projections with NCM and CMM. NCL, lateral to NCM, additionally exhibits bilateral projections with Field L (not pictured). (D) A visualization of recording sites, shown over top of the starling brain atlas [86]. (E) Amplitude-normalized voltage traces of peak channel activity for all categorical units (see Methods 6.4.32) used in analysis for each brain region. (F) Average trace of each unit-type cluster (G-I) Brain regions for each unit type.

6.2.5 Physiology paradigm

We developed a paradigm from which to record extracellular neural activity using 1-2 (unilaterally or bilaterally) implanted 32-64 channel 1-8 shank silicon electrodes from freely behaving subjects while they engaged with the behavioral apparatus. To this end, we designed MagPi, a Raspberry Pi-based interface between our behavioral panels and the OpenEphys neural acquisition device to run and record behaviors in sync with 24/7 neural recordings (Fig 6.3A). Birds were implanted with electrodes using a custom-designed 3D printed plastic microdrive and protective cap that enabled months-long recordings (Fig 6.3B). Electrodes were implanted in the secondary auditory regions CM (Caudal Mesopallium), NCM (Caudomedial Nidopallium), NCL (Caudolateral Nidopallium), and the primary auditory region Field L, three adjacent and bidirectionally connected regions of the songbird brain (Fig 6.3C,D).

We recorded from 10 subjects over a total of 222 days (5317 hours) of recordings. Chronically implanted subjects performed over 400,000 behavioral trials during recording. In addition, during the evening after birds had completed their behavioral trials for the day we turned the lights out in the behavior boxes and passively played back the same morph stimuli to the birds a total of 1.2 million times while recording.

In addition to the chronic recordings, we performed acute playback recordings of the same stimuli under light anesthesia (urethane) with 4 untrained (naive) birds totaling 38 recording sites across 2 bilaterally implanted 32-channel electrodes. In total, we performed over 59,000 passive playbacks with these naive birds.

Electrophysiology datasets were then spikesorted and aligned to behavioral data. Spike-sorted units (putative single-neurons) were merged across recording sessions in order to retain units over multiple days when possible (See 6.4.24). In total, we recorded from 14,406 units disctributed across the four brain regions (Fig 6.3E), which were further subset into 7,923 units used for this study on the basis of their auditory response properties (See next section). We clustered the templates of those units into three unit types (Fig 6.3E), which were predominantly

characterized by their spike width (Fig 6.8).

6.2.6 Quantifying response similarity and estimating a neurometric

Spike train data were analyzed as spike vectors over the categorical stimulus. For each trial, the time histogram (bin width=10ms) of the stimulus-aligned spike train is convolved with a Gaussian kernel (σ=25ms; Fig 6.10). Sample spike trains and trial-averaged spike vectors are shown for a sample unit for each morph in Fig 6.4E and F. From the trial spike vectors, a similarity matrix is computed as the cosine similarity between spike vectors (Fig 6.4I) which is used to compute a neurometric function (Fig 6.4J). We additionally used the cosine similarity matrix to compute a metric for a unit's categoricality (Fig 6.4K-L; see Methods 6.4.31) which reflects the similarity of unit responses within morph category versus between category. Using the categoricality metric we then extracted a set of categorically-relevant auditory units (See Methods 6.4.32). Across categorical units, we find that spike vector responses to all morphs to be smoothly varying on average but show variability across units in the degree of smoothness (Fig 6.4M-N).

6.2.7 Neurometric slope reflects psychometric uncertainty

Using the computed neurometric function for each behavioral unit, we find that neural responses to morphs reflect stimulus-to-stimulus variability in decision making (i.e. the likelihood of the Bayesian model).

We compared the slope of the neurometric function to the slope of the psychometric function for each bird and morph [1]. We find that the slope of the psychometric function reflects the likelihood in the Bayesian model, i.e. the uncertainty in decision making. Across subjects, we observe a significant positive correlation between the neurometric and psychometric slopes (r^2=0.115, p=$3e-126$, n=42551). This relationship is observed within-subject in nine of ten subjects (Fig 6.4P). When controlling for individual variation in neurometric and psychometric

[1] both neurometric and psychometric slopes were scaled to the range of the neurometric function as well as log scaled, see Methods

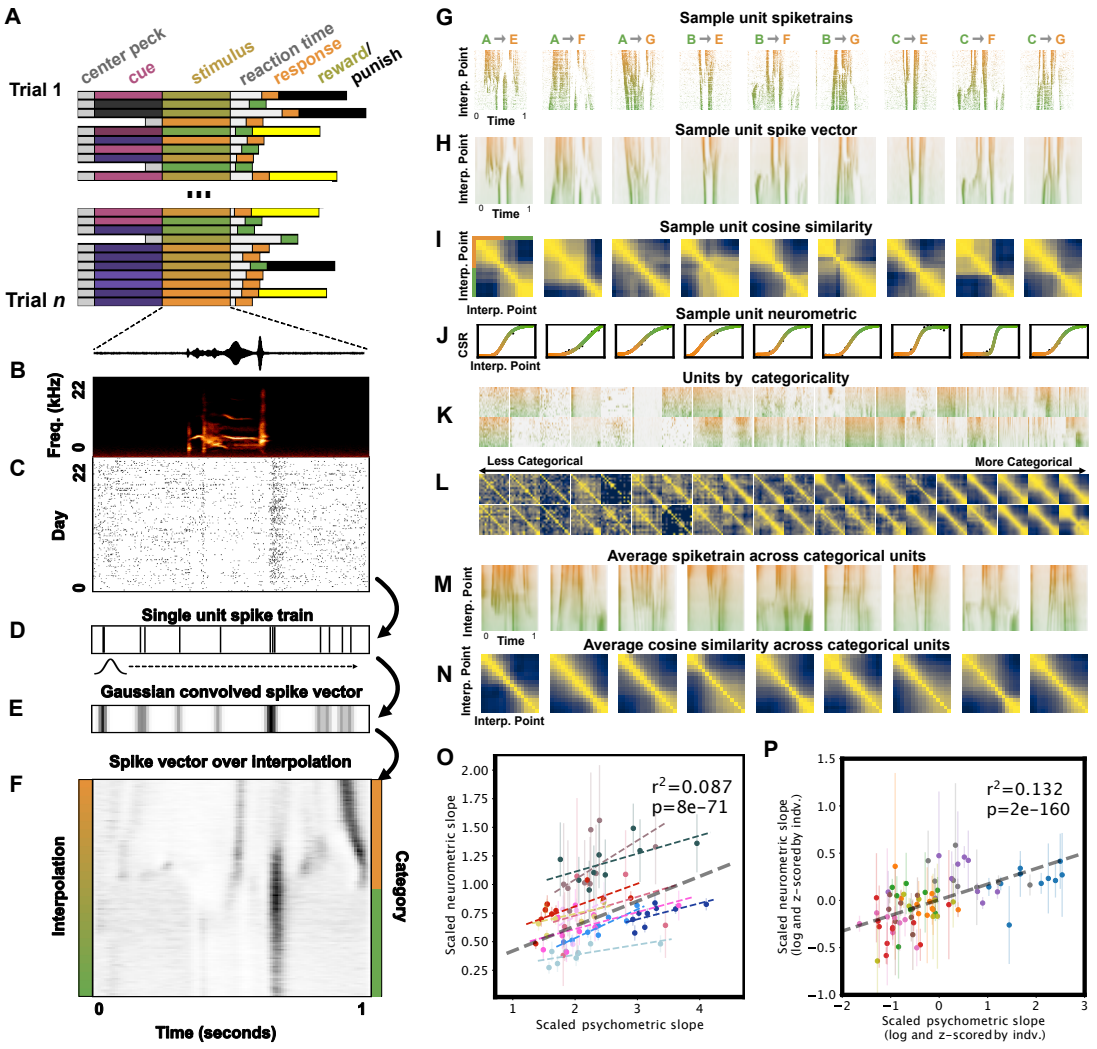


Figure 6.4. Neurometric functions of single units reflect psychometric functions of perceptual behavior. (A) Trial-by-trial behavioral data structure. (B) Spectrogram of categorical (morph) stimulus for a single trial. (C) Spike raster for a single unit across trials. (D) A single example spike train from (C). (E) A spike vector is computed as the spike train from (D) convolved with a Gaussian kernel. (F) The average spike vector for the unit in (C-D) for a single morph (AE). (G) Sample spike trains for one unit across 8 morphs. (H) Spike vector representations of the spike trains from (G). (I) Cosine similarity matrices computed from the spike trains in H. (J) Neurometric functions are computed from the similarity matrices in (I). CSR stands for Categorical Similarity Ratio (see Methods 6.4.30). (K) Sample morph spike vectors (as in (H) for units, sorted by unit categoricality. (L) Similarity matrices for the units in (K). (M) Average spike trains across each categorical unit for morphs. (N) Average cosine similarity matrices across all categorical units, for each morph. (O) Psychometric slope (logged and scaled by psychometric range) versus neurometric slope (also logged and scaled by psychometric range) for each subject and morph. Each subject is shown with a unique color and regression line. A regression line across subjects is shown in gray. (P) The same data as in (O) z-scored by subject, where color corresponds to morph (same color correspondence as Fig 6.2).

slopes within birds by z-scoring the psychometric and neurometric slopes by bird, we observe a stronger relationship (r^2=0.147, p=$3e-204$, n=42551), where across birds the same morphs occupy similar relative neurometric and psychometric positions 6.4Q).

6.2.8 Within subject perceptual variability is reflected in neural response

We additionally assessed whether subject-to-subject variability in behavior (i.e. the psychometric slope) was reflected in neural responses. We again compared the slope of the fit psychometric function to the slope of the fit neurometric function for each unit and morph. We performed a hierarchical regression comparing a prediction of the neurometric slope from the morph (`neurometric slope` $\sim$ `morph + subject`) against a prediction of the neurometric slope from the morph and the psychometric slope (`neurometric slope` $\sim$ `morph + psychometric slope + subject`). We find that the psychometric function explains more variance in the neurometric function than stimulus alone ($F(1,42533) = 6.65$, p=0.01), suggesting that neural responses reflect individual variability in behavior.

6.2.9 Context modulates neural response

Our behavioral results demonstrate a shift in perceptual classification of morph stimuli biased by contextual cue information. These results suggest that underlying physiological processes integrate information from the cue syllable with the sensory morph syllable to modulate perceptual behavior. Whether that modulation occurs early in sensory processing, or only later, in decision making and motor systems, is unknown. Previous work across sensory modalities, animal models, and behavioral paradigms has established that predictive information increases responses in motor and decision-making related brain regions [418]. In contrast, in sensory regions, the opposite effect is observed. Activity tends to be suppressed when expectation is greater [418]. For example, in humans, neural responses are suppressed in the primary visual cortex when events are more likely [214]. Similar expectation-related suppression is also

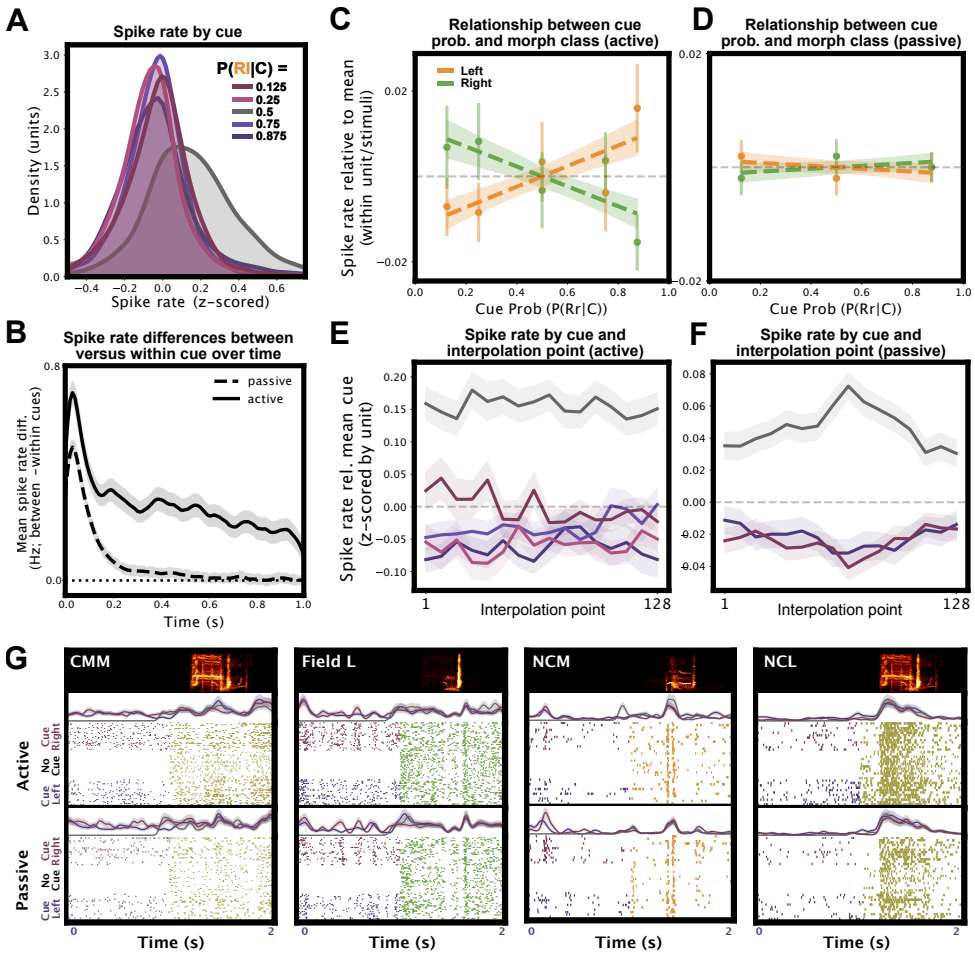


Figure 6.5. Predictive syllables modulate response to morph syllable. (A) Differences in spike rate for each predictive cue stimulus for active behavior playbacks, where uncued trials yield the highest spike rate. Spike rate is z-scored within each unit, and the difference is shown as the difference within stimulus across cueing conditions for each unit. (B) Mean difference in spike rate between cue conditions minus mean difference within cue conditions across time for passive and active trials, where zero seconds is the onset of the stimulus. A value above zero indicates greater similarity within cues than between cues. (C) Relationship between cue probability and morph class, within unit and stimulus, measured through spike rate across morph playback. For both the left and right morph stimuli, a regression line is shown with a 95% bootstrapped confidence interval. (D) Same as (E) for passive trials. (E) Spike rate differences within each unit across morph interpolation points for active trials. (F) Same as (C) for passive trials. (G) Sample unit responses for four units from the four regions recorded in this study (subjects B1170, B1597, B1248, and B1593 from left to right). The top of each panel shows a spectrogram of the morph stimulus played back. Below, a trace is shown for three cue conditions (No cue, P(Rl—C) = 0.125, and P(Rl—C) = 0.875) corresponding to the average Gaussian convolved spike vector and 95% CI for active trials. Below the trace are sample spike rasters for each cue condition, where each row is a child. Below the rasters, the sample trace and raster plots are repeated for the same unit in the passive trial condition.

observed in human audition [433, 432]. In mice, primary auditory cortex modulations are also task-related: single-unit neural responses are more categorically selective when animals are actively classifying stimuli than during passive listening [467]. Thus, it would be reasonable in our experiment to hypothesize that early sensory modulation will occur and be differentially modulated during active behavior and passive listening.

To assess whether the predictive information present in the cue modulates neural responses to the morph stimulus, we measured overall spike rate changes in units as a function of the predictive cue stimulus in active behavioral trials (Fig 6.5). Controlling for stimulus-to-stimulus spike rate variability within each unit, we find a main effect of cue identify on spike rate, with non-cued trials showing the highest spike rate (where spike rate is z-scored per unit; X^2(4, N = 851118) = 15162, p ¡1e-5; Fig 6.5A). To quantify the time-course of this cue-related modulation, we measured the difference in spike rate over time for pairs stimuli presentations that were either preceded by the same or different cue syllables (Fig 6.5B). We find that within-cue similarity in spike rate persists throughout the categorical morph stimulus presentation in active trials. In contrast, during passive playback, the cue-related similarity quickly decays to chance, at around 200-300ms.

6.2.10 Expectation suppresses spike rate in predicted stimuli

We next assessed whether cue-related modulation was associated with the predictive information present in the cue syllable. To this end, we measured the interaction between the cue probability and the morph stimulus class, controlling for the unit's overall response to the stimulus and differences in response strength to each cue (Fig 6.12). During active behavior trials, we find a significant interaction between cue probability and the morph stimulus class. As stimulus probability increases, spike rate decreases (X^2(1, N = 851118) = 392, p ¡1e-5; Fig. 6.5C). In the passive playback condition, we did not observe the same effect (Fig. 6.5D). In Fig. 6.5E,F, we plot the differences in spike rate as a function of the cue and point along the morph. In the passive condition, while no difference between predictive probability and stimulus class

exists, the spike rate increases with proximity to the decision boundary in the non-cued trials relative to the cued trials (Fig. 6.5F). We additionally tested for an interaction between stimulus probability and cue in each brain region, neuron class, subject, and morph class (Fig. 6.13). We find consistent results in Field L, NCL, and to a lesser extent CMM, while in NCM, we observe the opposite interaction, increased response to predicted stimuli. Results were broadly consistent across unit classes and morphs. In birds, the interaction between predictive probability and cue was consistent with the brain region they were recorded from.

6.2.11 Predictive response modulation is consistent with a shift in the likelihood of Bayesian model

Having established that neural signals are modulated by the predictive cue, we next explored how cue-related modulation reflects the similarity between neural responses.

We base our hypotheses about neural response modulation again on the Bayesian model presented in Figure 6.1F-I, which reflects similar Bayesian models of integration in speech perception. In particular, the perceptual magnet effect characterizes a phenomenon in categorical phoneme perception in which speech perception is warped around categorical boundaries [224]. The Bayesian account of the perceptual magnet effect suggests that this perceptual warping is due to a shift in the posterior probability of the stimulus toward higher probability regions of acoustic space, resulting from the integration of prior distributional information with a noisy representation of the acoustic stimulus [119] (Fig 6.6A). Under this model, in our task, as the predictive probability toward one side of the morph increases (i.e. in the context of a predictive cue), the within-category similarity of the posterior on the predicted side of the morph will increase and the within-category similarity of the low-probability side of the morph will decrease (Fig 6.6B). Thus, neural signals reflecting this posterior distribution will similarly increase in similarity as a function of predictivity (Fig 6.6C,D). This model accounts for changes to the posterior, reflecting perceptual decisions, but does not differentiate between modulations to the likelihood. It is well established that, while physiological modulations do occur in early

sensory regions of the brain under the context of predictive information, the manner in which those modulations occur is qualitatively different than in decision-making regions [418]. Thus, in Figure 6.6E we propose an extension to the Bayesian model of perceptual decision making, reflecting modulation in the likelihood due to predictive information. In this model, predictive information reorganizes sensory representation to increase acuity in regions of acoustic space where events are more likely to occur. We model this as a narrowing of the likelihood (here, σ of the Gaussian distribution) reflecting that neural resources are being redistributed to reduce perceptual noise in regions of acoustic space that are more likely to occur. In our task (Fig 6.6F), in contrast to the posterior distribution, this model predicts that increasing acuity in the likelihood in high-probability regions of stimulus space will result in a reduction of within-category similarity (6.6G,H). Thus, neural signals representing a likelihood with increased acuity would also decrease in similarity as a function of cue predictivity.

To assess whether the similarity of neural responses changed as a function of the cue, we compared the trial-to-trial cosine similarity of the spike vector response across morphs. Our empirical results are more consistent with the model of likelihood modulation than posterior integration. In particular, we find that in the presence of a predictive cue, the within-category similarity is higher in the non-predicted class than the predicted class (Fig 6.6I). The within cue similarity across units and morphs decreases as a function of the probability of the cue class (r^2 = -0.019, p=1.92e-10; n=107628; Fig 6.6J). This effect suggests that differences in predicted regions of acoustic space are accentuated between stimuli to improve acuity. If our neural signals had reflected the posterior, as in our behavioral results, we would have expected an effect in the opposite direction, in which similarity would have increased within stimulus class. As in our spike rate analyses, we repeated the similarity analysis over individual morphs, brain regions, subjects, and unit types (Figs 6.16, 6.17). Our results parallel those in the spike rate analyses. We observe results in CMM, Field L, and NCL, while in NCM we observe the opposite effect. In birds, the effect was consistent with the brain regions they were recorded from. Results were broadly consistent across morphs and unit types.

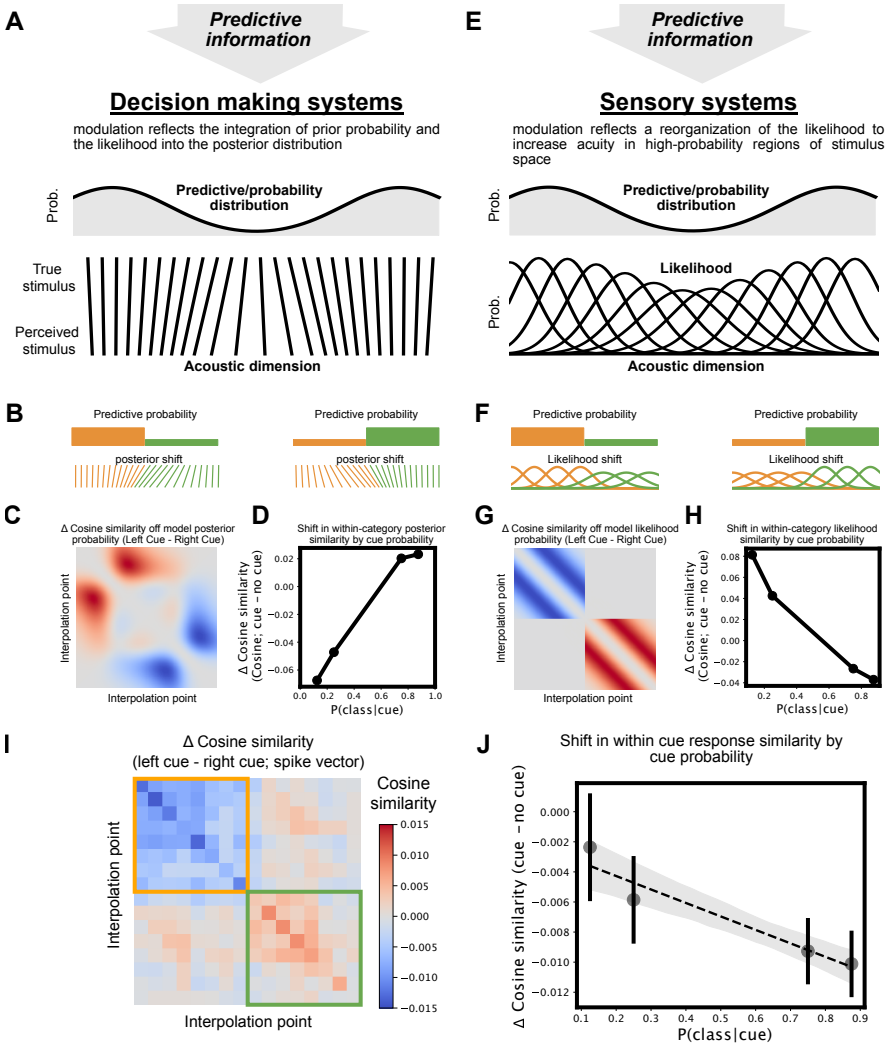


Figure 6.6. Modulation of response similarity as a function of predictive cue probability. (A) A summary of the perceptual magnet effect [224] and the corresponding Bayesian model [119]. (B) Visualizing the posterior shift from (A) in the context of this manuscript's behavioral experiment. (C) The shift in the Bayesian model's posterior distribution (measures as cosine similarity over the PDF) when cued. (D) The shift in the within-category posterior cosine similarity as a function of the probability of that class. (E) An extension to the model in (A) in which predictive information biases the likelihood to increase acuity in high probability regions of acoustic space. (F) Visualizing the likelihood shift from (E) in the context of this manuscript's behavioral experiment. (G) The shift in the Bayesian model's likelihood distribution (measures as cosine similarity over the PDF) when cued. (H) The shift in the within-category likelihood cosine similarity as a function of the probability of that class. (I) The observed shift in spike train vector cosine similarity for left-cued minus right-cued trials. The shift is depicted here averaged across units and morphs. (J) The relationship between the probability of the stimulus class and the shift in similarity from baseline (the non-cued condition).

6.3 Discussion

Categorical perception involves a non-linear mapping between physical sensory signals and their representation in perceptual space, a phenomenon whose ubiquity suggests a fundamental role in sensory integration. This warping of perceptual space is not fixed. Contextual information can bias categorical perception, a phenomenon observed both in speech perception [136] as well as in wild songbirds [233]. The neural and cognitive mechanisms underlying this bias are not well understood. In this work, we trained songbirds on a categorical perceptual decision-making task, in which we controlled the prior predictive contextual information present in the task. We found that songbirds use this information to bias their perceptual decisions. This bias is well-predicted by a Bayesian model of perceptual decision making, in which the extent to which bias occurs reflects the prior predictive information, as well as the uncertainty over the stimulus. These observations are consistent with Bayesian integration hypotheses of human context-dependent categorical speech perception [324, 119]. Using the same behavioral framework, we then chronically recorded from populations of sensory neurons while birds were making perceptual decisions. We found that sensory responses reflect the likelihood of behavioral responses to ambiguous syllables, indicating that the basis of categorical perception arises early in sensory processing. We then investigated whether predictive information from song-sequence context biases sensory processing. We found that cue probability does bias sensory representations, suppressing spike rate. Finally, we measured the similarity between neural responses as a function of their position along each morph for each cue condition. We found that neural responses to stimuli belonging to the expected class of the morph were more dissimilar, consistent with a modulation in the likelihood of our Bayesian perceptual model.

6.3.1 How is predictive information actively maintained and integrated?

The sensory populations we recorded from only tell one part of the story about prediction and cue integration. The populations we recorded from reflected a bias imposed by predictive

information, but not where that predictive information comes from, or how it is maintained or integrated. In speech, ongoing work seeks to uncover the neural systems underlying predictive information as they relate to lexical and pre-lexical feedback circuits [324]. Songbirds may provide a useful model for understanding how temporal predictive information biases vocal signal perception. Birdsong is underlied by both short-range acoustic structure that parallels linguistic phonology [381, 32] as well as longer-range structure occurring across timescales. The songbird basal ganglia, for example, is involved in the integration of sequential syntactic information [1] and may be a valuable future target for the study of sequential information integration. In parallel, recent work on visual working memory in corvids and pigeons has found that lateral portions of songbird NCL show similar active maintenance patterns as human prefrontal cortex [160, 223, 323]. Future work on songbird syntax and categorical perception have the potential to play a key role in uncovering the physiological systems that underlie how predictive and sequential information bias perception and recognition.

6.3.2 How do populations of neurons represent predictive information?

We recorded from a tiny fraction the neuron composing each of the brain regions used in this study and most of the neurons we recorded from were not recorded simultaneously. For example, out of an estimated 5 million neurons in starling NCM, we recorded from 1141 units. The analyses we performed in this study were performed over single units rather than populations of neurons. The Bayesian model we used to make predictions about behavior simplifies the decision-making process into a one-dimensional acoustic space. The physiological systems underlying these cognitive behaviors are vastly more complex, however. In future work, as our ability to record from larger populations of neurons improves, alongside our capacity to model high-dimensional neural populations (e.g. through machine learning [333] or topological [429] techniques) improves, a more full picture will be available of the physiological modulations in this experiment than can be observed across single neurons.

6.3.3 Is there a distinction between categorical perception and perceptual decision making?

In our experiment, we framed our task as explicit categorical perceptual decision-making, in which birds make decisions in order to receive a food reward. This task design differs from the context dependency present in categorical perception in communication because in our task an explicit perceptual decision is made. The boundary between perception and decision-making in speech and communication is not clear. For example, the degree to which perceptions are categorical can be dependent upon the conditions of the experimental framework [394]. Under a Bayesian framework, the results presented here suggest a physiological distinction between neural systems involved in representing the likelihood and prior in perception and decision making. Further physiological investigations may prove an important resource for determining the extent and location within processing circuits at which sensory signals are modulated to reflect behavioral decisions and, in the case of speech, comprehension.

6.3.4 How do prediction, attention, and integration differ?

Our data suggest that neural modulation occurs in the presence of predictive information in primary and secondary sensory regions of the songbird brain. These modulations are consistent with a modulation of the likelihood of a Bayesian model of perception. The mechanisms underlying that modulation are unclear. An important distinction exists between prediction, attention, perceptual integration, and modulation. It seems reasonable to the authors that neural resources are opportunistically dedicated to processing more likely events, resulting in greater acuity for predicted events than unpredicted events. Our observation is not the first example of a dynamic reorganization of the sensory brain to enhance perceptual acuity. Physiologically probing the mechanisms underlying such a resource dedication will be necessary to tease apart whether the observed phenomena are related to attentional mechanisms, or emerge from some other physiological substrate.

6.3.5 Are more natural stimulus spaces better poised for probing the complexities of vocal communication?

Our experiment differed from many prior studies on vocal perception and perceptual decision making in that stimuli presented varied along a complex and non-linear acoustic continuum derived from a linear interpolation in the latent space of a neural network projection. Prior studies on categorical perception and decision-making continuums in speech, animal communication, and auditory neuroscience have typically relied upon simplified stimulus spaces such as sine tones, voice onset time, or pitch-shifted stimulus. When available, simple stimulus spaces are useful, however, much of vocal communication cannot be well described with just one or two parameters. Our work demonstrates that ongoing advances in machine learning enable stimulus generation that better match the complexities in the acoustic repertoires of animals, which are often both more behaviorally relevant and physiologically salient.

6.3.6 Final note

Taken together, our results reveal novel information about the cognitive and physiological processes underlying perception in the context of varying predictive temporal information. Context dependency in song sequences can be modulated by predictive temporal context. That modulation parallels the probabilistic integration observed in human phonetic perception. Categorical perception of birdsong is physiologically reflected in early auditory regions of the songbird brain, which capture aspects of the likelihood of a probabilistic integration model. In the context of predictive sequential information, neural responses to vocal elements are modulated. That modulation differs during active behavior and passive listening. Finally, that modulation is consistent with a model of increasing perceptual acuity in high probability regions of acoustic space, rather than probabilistic integration, which likely occurs in downstream decision-making regions of the brain. In full, these results help bridge research into the cognitive underpinnings of language perception, animal communication, and the neuroscience of probabilistic information integration more generally.

6.4 Methods

6.4.1 Summary

Experiments consisted of a behavioral component and a chronic physiology component. The experimental protocol for the behavioral component was kept constant and underlied by the same software and hardware in both conditions, with the addition of chronic electrophysiological recording in the physiology component.

6.4.2 Subjects

Behavioral data was collected from 20 wild-caught European starlings of unknown sex. Before beginning experimental training, subjects were housed in a large mixed-sex aviary. Of the 20 starlings used for behavior, 10 individuals were used for chronic physiology. In addition, 4 behavior-naive subjects were used for acute physiological playback experiments.

6.4.3 Ethical note

All procedures were approved by the Institutional Animal Care and Use Committee of the University of California (S05383).

6.4.4 Datasets

Our final behavioral dataset was composed of 4.8 million behavioral trials from 20 birds. Our final chronic neural dataset was composed of 402,797 behavioral trials, with 365,360 responses, a total of 1,594,257 audio playbacks, occurring over 5,345 hours (222 days) of recording, across 10 birds. Our final acute dataset consisted of 59,533 passive playbacks under anesthesia across 4 birds.

Table 6.1. Behavioral datasets

Subject	Number of trials	Acc. (final 10k trials)
B1174	358106	0.8632
B1426	394956	0.8021
B1170	520304	0.8244
B1432	769387	0.9505
B1437	87702	0.8739
B1433	173287	0.9129
B1597	209867	0.9216
B1427	396727	0.8963
B1188	74724	0.8486
B1244	235092	0.866
B1590	169421	0.925
B1399	114908	0.9118
B1591	141700	0.8699
B1252	203956	0.9426
B1110	112694	0.7651
B1390	130094	0.8635
B1248	177069	0.8636
B1593	336716	0.8753
B1595	98607	0.8914
B1276	170278	0.9231
Total	**4875595**	**-**
Mean	**243779.75**	**0.87954**

Table 6.2. Neural datasets

Subject	Active trials	Playbacks	Behavioral responses	Recording hours	Units
Chronic					
B1188	54995	105956	54636	360	218
B1595	6817	39419	6743	81	47
B1276	2	14653	0	18	33
B1426	2823	10596	2777	31	64
B1432	64124	218032	63141	533	981
B1170	34909	151886	34387	442	435
B1597	44231	203396	41533	650	1796
B1244	2689	8423	2680	19	124
B1593	110335	503133	89397	1959	1964
B1248	81872	338763	70066	1252	1042
Acute					
B1239	-	8973	-	11	124
B1279	-	23463	-	22	744
B1459	-	18070	-	26	314
B1500	-	9027	-	11	37

6.4.5 Stimulus generation

Stimuli were syllables of European starling song synthesized from a Variational Autoencoder (VAE) trained on syllables extracted from a library of European starling song [17].

6.4.6 Training dataset

Syllables were syllabically segmented using the dynamic thresholding approach outlined in [384] and available in the vocalization segmentation python package (https://github.com/timsainb/vocalization-segmentation). Syllables were then zero-padded to be 1 second long. Spectrograms of each syllable were then computed of each syllable with 128 frequency bands spaced between 50 and 22050 Hz, and downsampled to 128 time-bins (128 Hz), resulting in a 128x128 spectrogram of each syllable, used to train the VAE.

Table 6.3. Variational autoencoder architecture outline

Encoder		Decoder	
Layer Type	**Dimensionality**	**Layer Type**	**Dimensionality**
input	[32, 128, 128, 1]	input	[32, 16]
convolutional	[32, 64, 64, 64]	fully connected	[32, 1024]
convolutional	[32, 32, 32, 128]	fully connected	[32, 16384]
convolutional	[32, 16, 16, 256]	reshape	[32, 4, 4, 1024]
convolutional	[32, 8, 8, 512]	upsample	[32, 8, 8, 1024]
convolutional	[32, 4, 4, 1024]	convolutional	[32, 8, 8, 1024]
fully connected	[32, 1024]	upsample	[32, 16, 16, 1024]
fully connected	[32, 16]	convolutional	[32, 16, 16, 512]
		upsample	[32, 32, 32, 512]
		convolutional	[32, 32, 32, 256]
		upsample	[32, 64, 64, 256]
		convolutional	[32, 64, 64, 128]
		upsample	[32, 128, 128, 128]
		convolutional	[32, 128, 128, 64]
		convolutional	[32, 128, 128, 1]

6.4.7 Neural network

The neural network architecture we used followed those in our AVGN repository (https://github.com/timsainb/avgn_paper). We used a convolutional VAE architecture with a 16-dimensional latent space. The network was trained on batches of 32 syllables at a time. Artificial neurons used a leaky ReLu non-linearity. The network was trained with the ADAM optimizer. The network was trained in Tensorflow.

6.4.8 Sampling and synthesis

Each syllable stimulus (used for cues and endpoints) was sampled from the original dataset and passed through the VAE. The stimuli were chosen to be diverse, well-reconstructed in the VAE, and roughly equidistant both in spectrogram space (both input and reconstruction) as well as the latent space of the VAE. It is not expected that distances in spectral or neural network latent space would have a 1:1 relationship with an animal's perception of similarity. Morph stimuli were then sampled as linear interpolations between the latent (16D) representations of the

chosen stimuli. 128 points were taken in each interpolation, and passed through the decoder of the VAE, producing the final 128 point morph. Waveform stimuli were then generated from the spectrogram output of the decoder of the VAE using the Griffin-Lim algorithm. These waveforms were used for stimuli playback to the birds.

6.4.9 Behavioral shaping

Shaping occurred over four stages. First, each subject was trained to obtain food from a solenoid-powered hopper underneath the food port (Fig. 1A) by pecking the center response port, while the bird was freely allowed to explore the behavioral apparatus. Once the bird had pecked into the center peck port, stage two of the shaping procedure began, where the bird was required to peck into the center of the peck port to continue receiving food. After pecking the center port 100 times, they were transferred to stage 3. In stage three, the birds were trained to peck the left or right peck ports, which were lit with LEDs. Finally, the birds were transferred to stage four, where they would initiate behaviors by pecking into the center port and then either the left or right response port (cued randomly with a flashing light) for a food reward. After completing this shaping procedure, birds were transferred to the full task.

6.4.10 Behavioral training paradigm

Birds were initially trained to differentiate between syllables generated via the two endpoints in a single morph. After several days of above-chance accuracy with one pair of morph endpoints, the number of morph endpoints was increased until the birds showed above accuracy classification of the endpoints of all 9 morph. After learning the correct response for endpoints in each morph, birds were transferred to the full stimulus set which included 128 stimuli (linearly sampled and equally spaced in latent space) spanning each of the 9 morphs (1152 stimuli total). After the birds were performing reliably above chance on each full morph stimulus set for several days, we added cue stimuli preceding the categorical stimuli to provide context-dependent information at p=0.125, p=0.25, p=0.5, p=0.75, and p=0.875.

6.4.11 Training parameters

Several behavioral parameters were used in behavioral training, given here for reproducibility. Trials were reinforced at a variable ratio of between 2-4, manually set to ensure the bird maximally produced trials, but did not lose more than 10 grams of weight from baseline when in the restricted feeding condition. Punishment was set at a 5-second lights-off period, in which new behavioral trials could not be initiated. A minimum of 1 second between trials, regardless of correctness, was imposed. Birds were given a 5-second response window after stimuli had played back. Lighting conditions were set to match the daylight in the experimental location (San Diego, California).

6.4.12 Cue stimulus

Like the morph stimuli, the cue stimuli are 1-second long syllables synthesized from reconstruction from the variational autoencoder. Behavioral trials were presented with one of 6 cue conditions: no cue P(L—No Cue)=0.5 (NC), cue with no predictive information P(L—Cue)=0.5 (CN), cue left at p=.875% P(L—Cue)=0.875 (CL1), cue left at p=0.75% P(L—Cue)=0.75 (CL0), cue right at p=.875% P(L—Cue)=0.125 (CR1), cue right at p=0.75% P(L—Cue)=0.25 (CR0). 16% of trials were presented in the no cue condition (NC). 4% of trials were presented with the uninformative cue condition (CN). The remaining 80% of trials were evenly split between the cue right and cue left conditions. Because the CN condition was sampled with a substantially lower probability than the other conditions, resulting in a low number of total trials in comparison to each other cue condition, it was not included in physiological analyses. In passive physiology playback conditions, due to time constraints in playing back the full stimulus set of 128 interpolation points for each of 9 morphs and 6 cue conditions, we played back only the 87.5% predictive cue conditions in the AE and BF morphs.

6.4.13 Psychometric fit

To assess the shift in categorical perception, in each of the birds (n=20) we fit a psychometric (four-parameter logistic) function both to the overall responses to stimuli in the left and right categories of the morph, as well as to each individual morph.

$$logistic(x) = max. + \frac{min. - max.}{1 + (\frac{x}{inflection})^{slope}}$$

6.4.14 Bayesian integration hypothesis

To formalize our hypothesis, when a stimulus varies upon a single dimension x, the perceived value of x as a function of the true value of x and contextual *cue* information can be described by Bayes' rule:

$$\underbrace{P(x_{true} \mid x_{sensed}, cue)}_{\text{posterior}} \propto \underbrace{P(x_{sensed} \mid x_{true}, cue)}_{\text{likelihood}} \underbrace{P(x_{true} \mid cue)}_{\text{prior}} \tag{6.1}$$

By modulating the prior distribution of the categorical stimuli (x) with a cue, we predict that the perception of x will shift.

Preceding each to-be-categorized target stimulus (x), the cue stimulus provides predictive information about the category of the target stimulus. By treating this cue stimulus as a prior probability over x, we predicted that the determined posterior probability of x given sensory information and the cue stimulus would shift the classification of stimuli near the boundary between the two classes in the direction predicted by the cue stimulus.

Explicitly, we treat the likelihood of a target being sensed $P(x_{sensed} \mid x_{true}, cue)$ as a Gaussian probability distribution around the true target x_{true} [218, 119]:

$$P(x_{sensed} \mid x_{true}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_{true} - x_{sensed}}{\sigma_{sensed}}\right)^2} \tag{6.2}$$

and shift the prior probability as a function of the cue

$$P(x_{true} \mid cue) = \begin{cases} cue_{prob} & x_{true} > \text{categorical boundary} \\ 1 - cue_{prob} & x_{true} < \text{categorical boundary} \end{cases} \tag{6.3}$$

where cue_{prob} represents the predictive probability of the cue stimulus. We predict that birds will make a categorical decision based upon the posterior,

$$decision(x_{true}, x_{sensed}) = P(x_{true} \mid x_{sensed}, cue) \cdot category(x_{true}) \tag{6.4}$$

where $category(x_{true})$ is simply the trained category label of x in the 2AC task:

$$category(x_{true}) = \begin{cases} 0 & x_{true} > \text{categorical boundary} \\ 1 & x_{true} < \text{categorical boundary} \end{cases} \tag{6.5}$$

Under this model, the categorical decision of the bird is modulated by the prior cue information, resulting in a shift in the categorical decision point along the stimulus dimension in the direction predicted by the cue (Figure 6.1I).

6.4.15 Bayesian fit

In addition to fitting a psychometric function capturing the shape of the behavioral responses, we fit a Bayesian model reflecting our probabilistic hypothesis described above. This model used five parameters: the shape of the Gaussian of the likelihood (σ_{sensed}), a parameter corresponding to side bias in the apparatus (γ), and parameters representing inattention to the cue stimulus (δ), the target stimulus (β), and overall inattention to the task (α).

$$bias_{side}(\gamma) = category(x_{true})(1 - 2(1 - \gamma)) + 1 - \gamma$$

$$likelihood = P(x_{sensed} \mid x_{true}, cue)(1 - \beta) + bias_{side}(\gamma)\beta$$

$$posterior \propto P(x_{true} \mid x_{sensed}, cue)(1-\alpha) + bias_{side}(\gamma)\alpha$$

$$prior = P(x_{true} \mid cue)(1-\delta) + bias_{side}(\gamma)\delta$$

6.4.16 Response time

For each behavioral trial, we measured the time between the end of a stimulus presentation and the time that a subject's beak was detected in a behavioral response port. In Figure 6.2J-M we found that the response time varied based upon stimuli and cue conditions. To account for side biases in decision making (e.g. the bird having a position preference when engaging with the behavioral apparatus that positions them further toward the left or right peckport, for each analysis we z-scored reaction time for each bird's responses to each interpolation and class.

To parameterize the decay in response time as a function of the distance in the morph from the decision/class boundary, we fit the decay in response time to an exponential decay function (Fig 6.7). We discluded three birds from analysis (B1426, B1433, B1427) who we observed did not exhibit the same decay in reaction time as a function of distance from the decision boundary (Fig 6.2K).

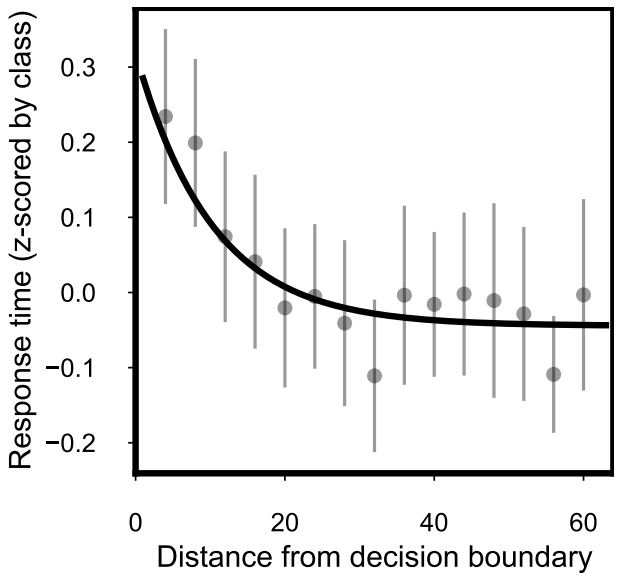


Figure 6.7. Sample fit of response time decay for a single morph (AE) for a single bird (B1174).

6.4.17 Chronic electrophysiology

We used 32 or 64-channel Neuronexus Si-Probes (A4x2-tet-7mm150-200-121, Buzsaki32, Buzsaki64, A1x32-Edge5mm-20-177) implanted either unilaterally or bilaterally. Probes were coated with PEDOT using an Intan RHD Electroplating Board no more than one week prior to implant. Probes were mounted on 3D-printed drives (described in Section 6.4.19), which were stereotactically implanted using the procedure outlined in Section 6.4.20. Extracellular voltages were amplified and digitized at 30kHz using an Intan RHD recording headstage, output through an SPI cable through an electrically assisted commutator to an Open Ephys recording system.

6.4.18 Behavioral neural acquisition interfacing with PiOperant

Behavioral and physiology were synced using a custom designed Raspberry-Pi-based system (PiOperant) for automating our behavioral paradigm and interfacing with the OpenEphys neural acquisition device. PiOperant interfaces with our behavioral panel using the Python software pyoperant (https://github.com/gentnerlab/pyoperant). Behavioral states and audio signals were input and synced with OpenEphys over two HDMI inputs (digital and analog) and a ZMQ interface containing additional information about behavioral trials.

6.4.19 Microdrives and head caps

Microdrives and head caps were custom-designed over the course of this experiment and were printed using a FormLabs Form3 3D printer using FormLabs standard grey resin printed at 25-50 micron resolution. Microdrives were comprised of a drive, a shuttle, and a MiniTaps 6/16" 00-90 gold screw, hand-tapped and fastened to the drive with a brass nut. The screw was used to raise and lower the shuttle manually, at a depth of 282 microns per full rotation. Head caps were designed to be removable enable moving probes further down, as well as easy explant and re-use of probes.

6.4.20 Electrode implant procedure

Subjects were given analgesia by means of a 5mg/kg dose of carprofen (Rimadyl) administered intramuscularly. Animals were then anesthetized with a gaseous mixture of Isoflurane/oxygen (1-2.5%, 0.7 lpm). The scalp and feathers around the scalp were then removed and part of the skull over the y-sinus (the stereotactic reference sinus between the cerebellum and the two hemispheres of the brain) was visible. A craniotomy was opened above the recording site. A second craniotomy for the ground was then performed several millimeters away from the primary craniotomy. A platinum-iridium ground wire was then inserted in the craniotomy above the dura and glued to the skull. The baseplate for the head cap was then cemented (Metabond) to the skull. The durotomy was then performed in the original craniotomy, and the electrode was stereotactically lowered, attached to the microdrive at a rate of no more than 100 microns per minute. Once the final site was reached, the microdrive was then cemented to the skull, and a silicone base was applied above the craniotomy to prevent infection. The head cap was then screwed into the baseplate, protecting the recording site and probe. The headstage was then attached to the outside of the head cap.

In some individuals, multiple implants were performed in serial when one probe failed by explanting and removing the first probe and microdrive, creating a new craniotomy in the opposite hemisphere and durotomy, and implanting a new probe/microdrive. In one individual, two probes/drives were implanted simultaneously one in each hemisphere.

6.4.21 Recordings and behavior blocks

Recordings were performed 24 hours per day in order to track individual neurons over days. Recordings consisted of (1) behavior blocks, in which subjects freely interacted with the behavioral apparatus, (2) a free feeding period, in which the behavioral apparatus presented food to the bird without requiring the bird to perform trials, (3) a passive playback block, in which lights were turned off and the birds were passively presented with stimuli, and (4) a sleep block,

in which the lights were left off and no stimuli were played back.

6.4.22 Chronic behavior blocks

Chronic behavior blocks were matched to behavior blocks without physiology. The behavioral apparatus was left on throughout the day, allowing subjects to initiate trials through a peck in the central peck port. Trials were intermittently reinforced with a food reward and punished with the lights briefly turning off on incorrect trials. Using this paradigm, subjects performed several thousand trials per day.

6.4.23 Chronic passive playback blocks

At a set time at the end of each day, we turned the lights out in the bird's operant conditioning block and passively played back the morph stimuli to the bird. The bird's activity and sleep state during this time was not monitored. The silence interval between stimuli was randomly sampled between 1.1 and 1.5 seconds.

6.4.24 Spikesorting and merging over long-term chronic recordings

Spikesoring was performed over each 12 hour block of recording using Kilosort 2-2.5 [329] and SpikeInterface [52]. LFP was bandpass filtered between 300 and 6000 Hz and further normalized using common median referencing. To retain units across days/sorts, we additionally used an overlapping procedure to merge each neighboring pair of recordings together. To do so, we took the last 30 minutes of the previous recording, and the first 30 minutes of the following recording, and separately sorted that hour-long recording, which overlapped with the two larger recordings. We then computed the overlap between units in the overlapping recording and each of the two full recordings. Units were then considered to be the same unit if their "agreement" score (SpikeInterface; the spike coincidence of the two units) was above a set threshold (set at 0.5). Units from each of the larger recordings that were merged with the same unit in the overlapped recording were then merged, allowing the same unit to be tracked over multiple days.

6.4.25 Stimulus alignment

Stimuli playback was aligned to neural data using a 1kHz sine way sent from the MagPi behavioral control device to the OpenEphys acquisition board collected simultaneously with neural data, alongside a binary switch indicating the onset and offset of playback. An additional message giving information about the specific trial was sent over the local network via ZMQ.

6.4.26 Acute recording sessions

Four naive birds were acutely recorded from while passively playing back morph stimuli under anesthesia. Birds were anesthetized with 20% urethan (7ml/kg, i.m). A metal pin was affixed to the skull with dental cement, and a window was removed from the top layer of skull, where a small craniotomy and durotomy were performed over the recording location. Birds were placed in a sound-attenuated chamber (Acoustic Systems). For the acute recordings, we used 32-channel 4-shank silicon microelectrodes with large spacing (200-400um) between sites to maximize spatial coverage across our recording sites. Electrodes were stereotactically inserted and lowered slowly until all sites were in brain tissue, and 15 minutes were taken before starting a recording block to allow the probe to stabilize. Auditory stimuli were presented from a speaker mounted 20cm above the center of the bird's head. Stimuli were played back with an inter-stimulus interval randomly sampled between 1.1 and 1.5 seconds. Recordings were then processed using the same pipeline as in chronic recordings.

6.4.27 Localizing units

Unit locations were defined as the location of the peak recording channel on which the unit was present. The recording channel was determined from its position within the shank, and the shank's position relative to the stereotactic implant. Stereotactic implant locations were recorded relative to the Y-sinus between the cerebellum and two hemispheres of the brain, and the depth relative to the surface of the brain. Implant locations relative to nuclei were then determined relative to voxel mapping of the European starling brain atlas [86], as shown in Fig

6.3D).

6.4.28 Clustering unit spike shapes

Unit spike shapes were clustered using the Birch clustering algorithm [469] fit to the template voltage trace of the peak channel of each unit into 5 clusters. The clusters found were primarily observed to vary upon their spike-width, and were (post hoc) manually labeled on the basis of spike width (i.e. wide, middle, narrow). Narrow spiking cells are generally believed to be inhibitory, while wide-spiking cells are thought to be excitatory [23, 184]. Spike width was negatively correlated with spike rate (logged; r^2 = -0.239, p=3.58e-103, n=7923).

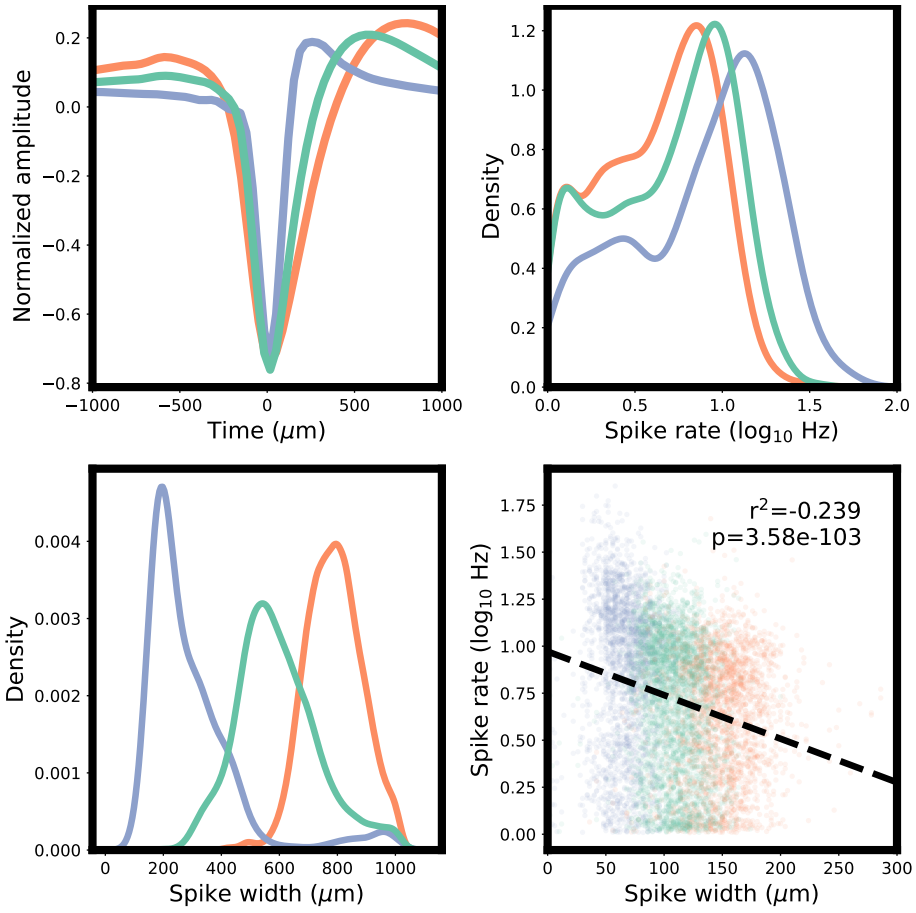


Figure 6.8. Spike widths and rates for each unit type. Spike width is given as the time between trough and peak.

6.4.29 Neural feature representation and response similarity

We represented spike trains as vectors using the methods outlined in Fig 6.4A-F). In particular, a PSTH of spike trains was computed with 10ms time-bins, which was then smoothed with a Gaussian kernel with a σ of 25ms. Morphs were sampled at a resolution of 128 points. For physiological analyses, we reduced the sampling resolution, binning the 128 interpolation points into 16 points along the morph, thus the neural response vectors and similarity matrices are 100 time-bins by 16 interpolation bins, and 16 interpolation bins by 16 interpolation bins, respectively.

We computed neural response similarity as the cosine similarity of the Gaussian convolved spike vectors, which has been effectively used to find similarity in spike trains in the past [120]. A number of different similarity metrics could have been used in its place, for example, correlation coefficients [396, 429] and Euclidean distance between Gaussian convolved spike trains. We compared the cosine similarity to the several other similarity metrics used in neural analyses including the correlation coefficient, Euclidean distance, and Manhattan distance, and found broadly similar results (Fig 6.9).

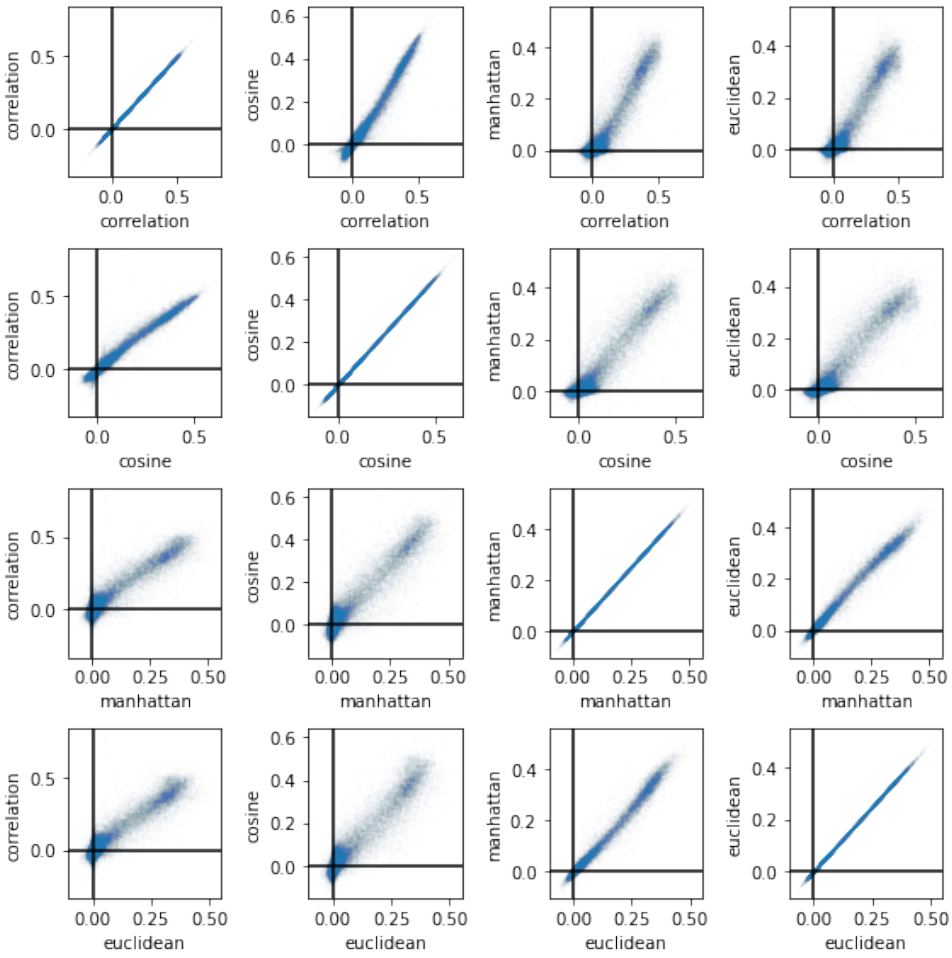


Figure 6.9. Comparison of unit categoricality using cosine similarity, correlation coefficient, Euclidean distance $(1/(1+D))$ and Manhattan distance $(1/(1+D))$.

6.4.30 Estimating a neurometric function from the similarity matrix

The neurometric function is computed on the basis of the similarity matrix and is detailed in Fig 6.10. For each interpolation point, we took the average of the within and between-category similarity (S_{C1} and S_{C2}) took the ratio ($\frac{S_{C1}}{S_{C1}+S_{C1}}$) as the categorical similarity ratio. We then fit the same four-parameter logistic function as used in the psychometric function to the categorical similarity ratio as a function of interpolation point.

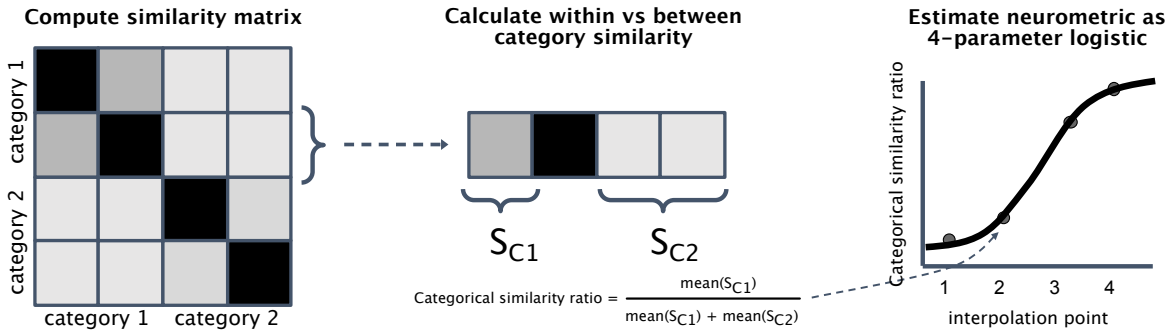


Figure 6.10. Method for computing a neurometric function from a similarity matrix.

6.4.31 Categoricality metric

Unit categoricality was computed using the similarity matrix (as seen in Fig 6.4). The similarity matrix used to compute a unit's categoricality was the mean cosine similarity matrix across interpolation responses, where the cosine similarity matrix was computed over average response vectors for each interpolation point.

Similarity matrices were divided into four quadrants, corresponding to the within-category similarities for each category, and the between-category similarities. Categoricality was computed as the mean similarity in the within-category quadrants of the similarity matrix (i.e. the top left and bottom right), minus the between category similarities.

6.4.32 Subsetting categorical units

We operationalized behaviorally relevant, categorical, units on the basis of their response characteristics to the morph stimuli. Categorical units were determined by a threshold set in the categoricality metric. This threshold was set at a categoricality metric value above 0.1. These thresholds were set based upon visual assessment of unit responses (Fig 6.11) and similarity matrices.

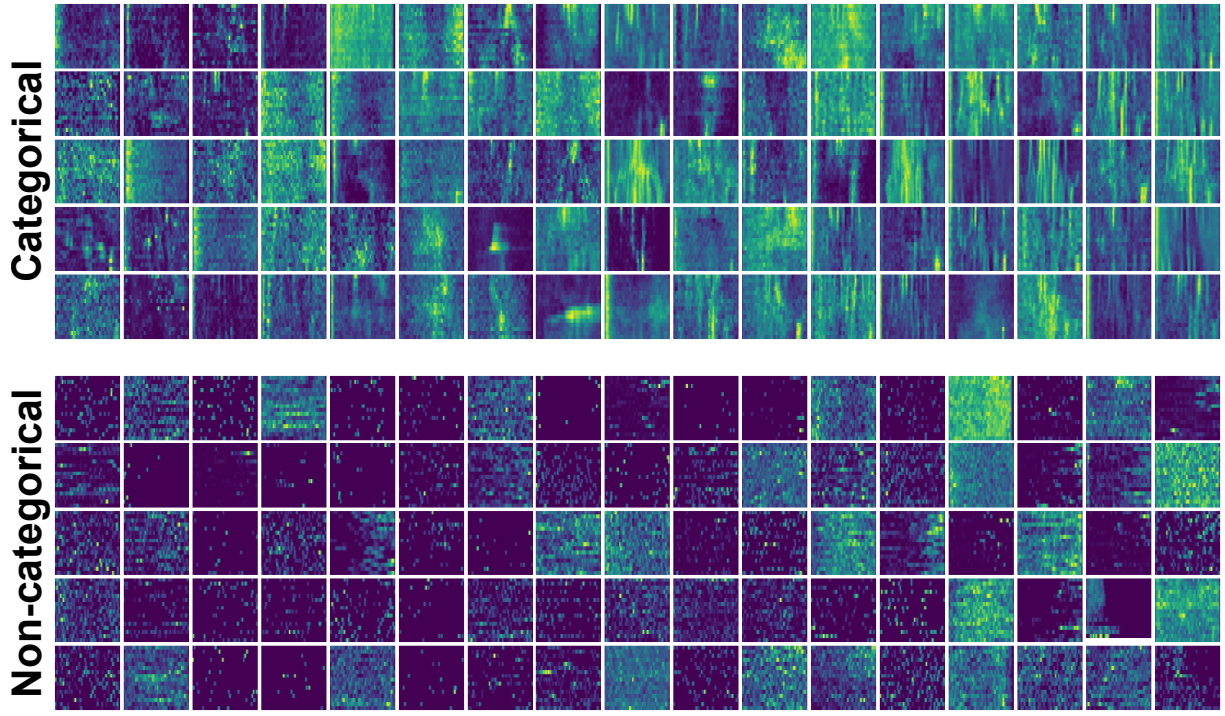


Figure 6.11. Categorical and non-categorical units, sorted by categoricality (right is greater).

6.4.33 Comparing spike rate across units, cues, and morphs

The analyses performed in section 6.2.5 were performed over unit spike-rates in response to the morph stimuli, where spike rate was z-scored over the unit's spike rates across all stimuli. A figure visualizing the main effect of cue and interactions between cue probability and stimulus class is shown in Figure 6.12. In addition, we shuffled the cue labels to ensure that our results were not due to inherent sampling biases present in the data (e.g. a left cue is more predictive of

a left morph point, thus more cue left to left morph point samples exist in the dataset).

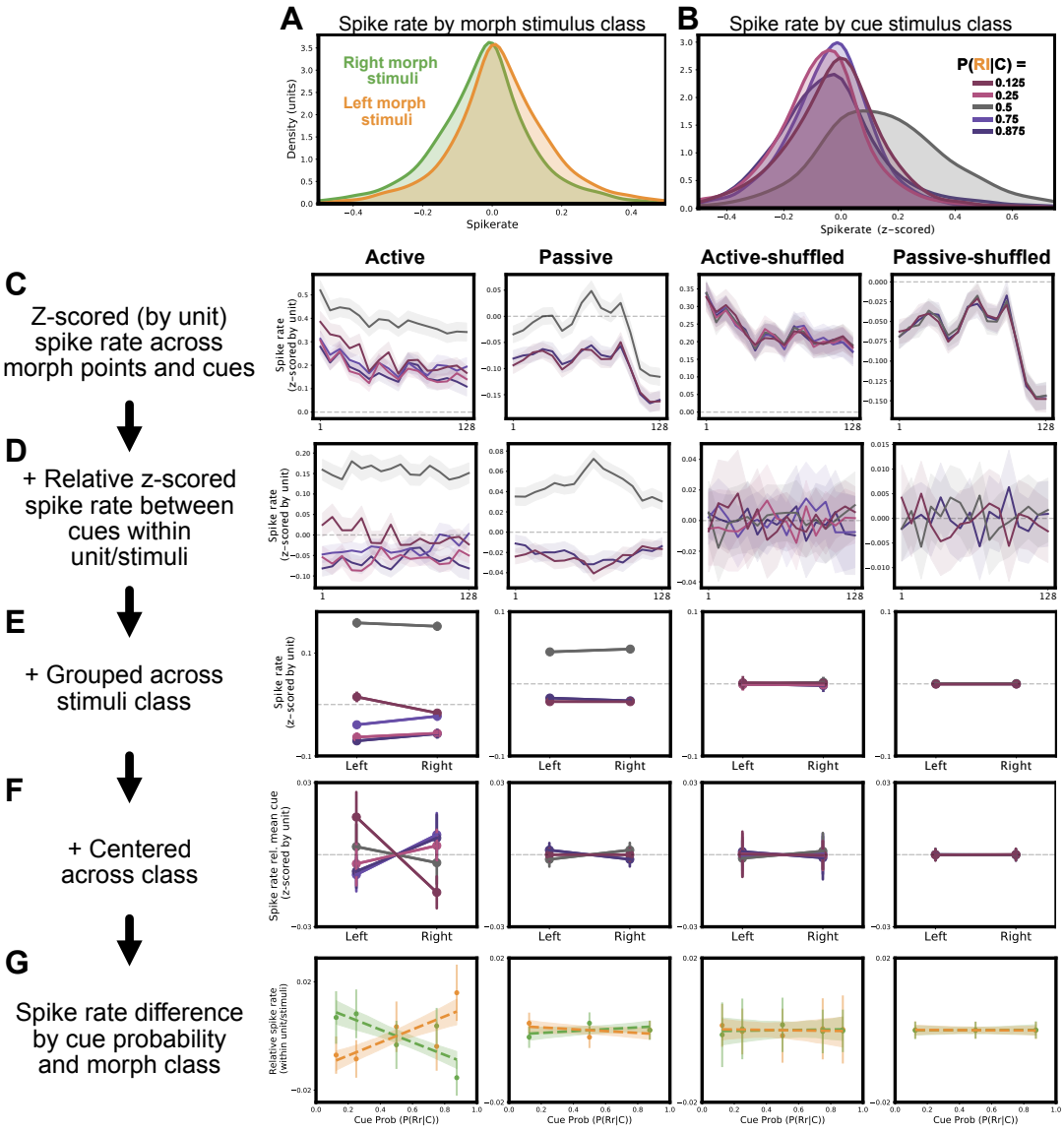


Figure 6.12. Spike rate modulation by cue. (A) Spike rate differs between stimuli. On average, across morphs, left stimuli elicit a greater spike rate than right stimuli. (B) Within stimuli, cues elicit differing spike rates for the same morph stimuli. Uncued trials yield the highest spike rate. (C) Across morph stimuli, average spike rates are shown for each cue condition in passive, and active conditions. Additionally, we perform the same analyses over data where cue labels are shuffled in the active and passive conditions (thus, no difference between spike rates across cues are observed). (D) The same data as the four panels from (C), subtracting the spike rate for each stimulus averaged across cue conditions. (E) The same data as in the four panels from (D), shown across morph sides (left and right) rather than morph interpolation points. (F) The same data as in (E) centered for each morph at zero, to show interactions between cue conditions and morph stimuli classes. (G) The same data as in (G) plotted for each morph stimuli class as a function of cue probability, exhibiting the relationship between cue probability and spike rate for each cue class.

6.4.34 Morph class and cue interactions by subject, brain region, unit type, and morph

In Figure 6.13 we plot the interactions between cue probability and morph stimulus class, controlling for the overall unit's spike rates to stimuli and a main effect of cue responses.

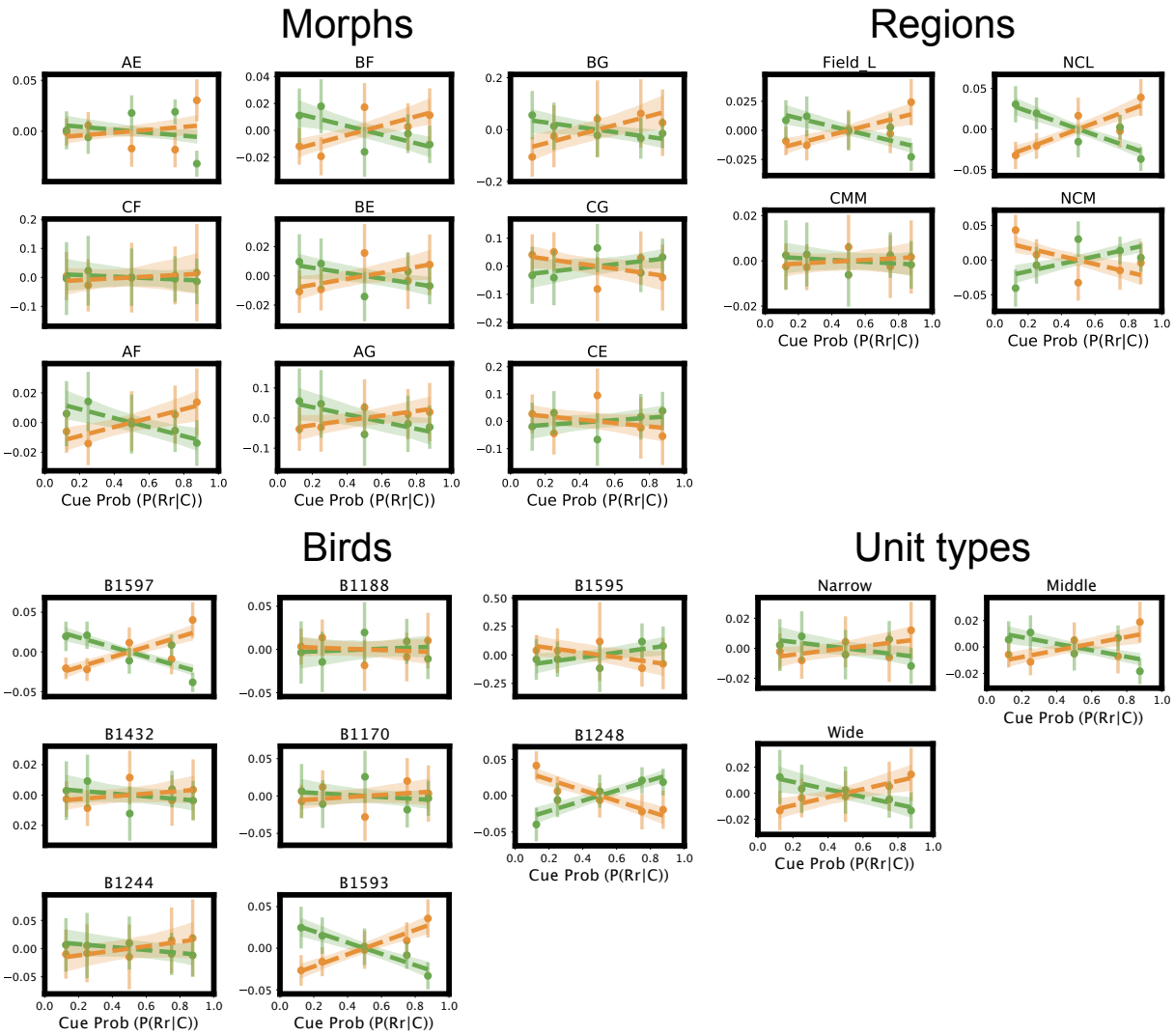


Figure 6.13. Interaction between cue probability and morph stimulus class on spike rate. Interactions are shown across four categories: Morph, brain region, subject, and unit class.

6.4.35 Differences in spike rate as a function of time

In section 6.2.5 we compared differences in spike rate as a function of their cue (i.e. within versus between cue).

To ensure that the effects of spike rate modulation occur between cue conditions, and not only between cue conditions and the uncued condition (where the main effect of cue on spike rate is greatest) we did not include the uncued condition in the spike rate differences between cue conditions in Fig 6.5B. We only included units and stimuli where we had active and passive behavioral trials (n units = 4668). We then, for each unit and stimulus, took the average difference in response vectors between trials for trials with the same cue, and trials with different cues. The difference between the average difference between cues, minus within cues, will equal zero when there is no difference between cue conditions. To ensure that no factors exogenous to between-cue differences are causing this effect, in Figure 6.14 we show the same analysis where cue labels have been shuffled within stimuli.

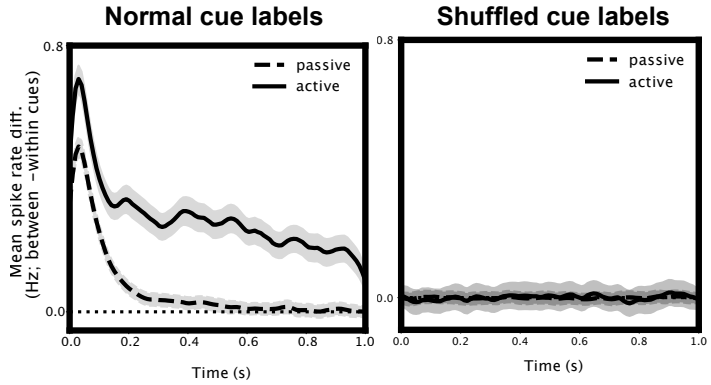


Figure 6.14. Spike rate differences between minus within cue categories over time. The right panel shows the same analysis with shuffled cue labels.

6.4.36 Within cue response similarity

For each unit and cue, we computed the cosine similarity matrix across each morph. Cosine similarity matrices were computed by taking the average cosine similarity across trials for each interpolation point (16) in the morph. Analyses were only performed over active behavioral

trials, where the subject provided a response. We then contrasted the cosine similarity matrices across different cue conditions. Figure 6.15 (top left) shows the average cosine similarity across left cues subtracted by the average cosine similarity across right cues. Blue in the top left of the plot (the orange bounding box) depicts less similarity in the predicted left class in left-cued trials. The reverse is true for the red in the bottom right. We measured this relationship showing that predicted morph classes are less similar within-class in Figure 6.15 (top right). Each point and confidence interval consists of the within-class similarity relative to the same unit's response to non-cued stimuli across trials. The negative relationship confirms that higher-probability cued trials exhibit less similar responses. In a similar manner as in Figures 6.12 and 6.14, we repeated this analysis over the same data in which cue labels had been shuffled within unit/interpolation. In the shuffled condition, we observe that the effect is removed. Finally, we broke out the analyses from Figure 6.15 in Figure 6.16 and 6.17, which are discussed in the main text.

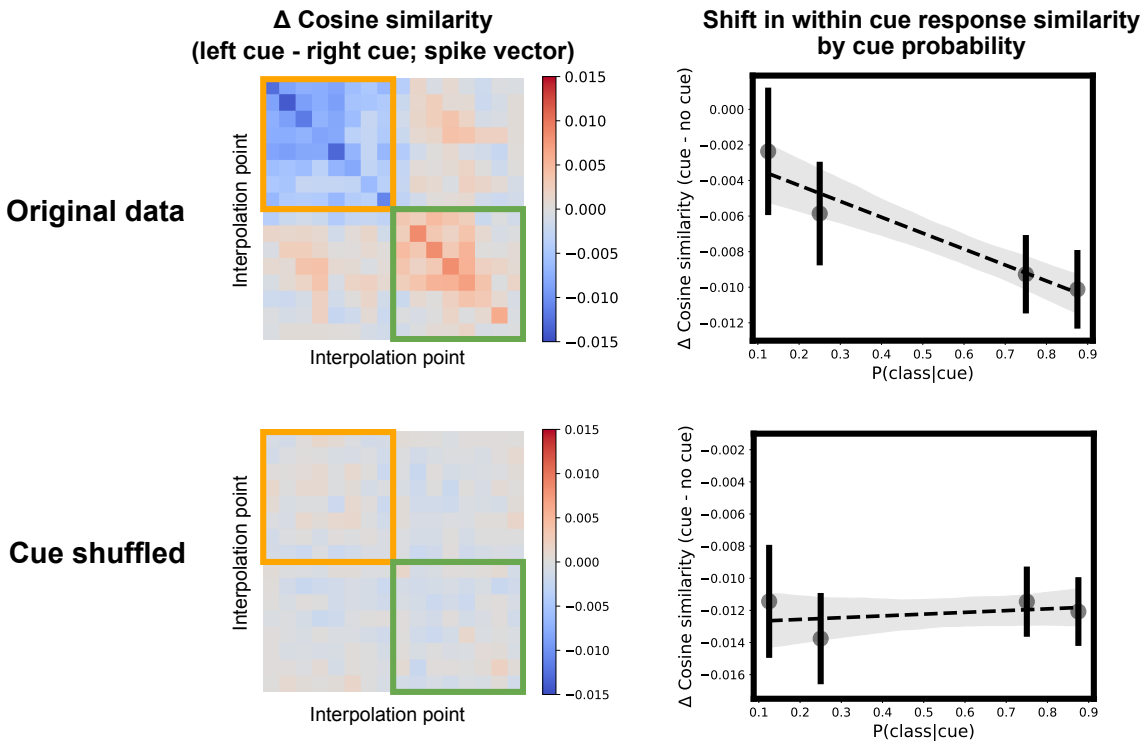


Figure 6.15. Spike vector cosine similarity and shift in similarity as a function of class probability as seen in Fig 6.6 I and J, compared with the same analysis performed over the same dataset where cue labels are shuffled.

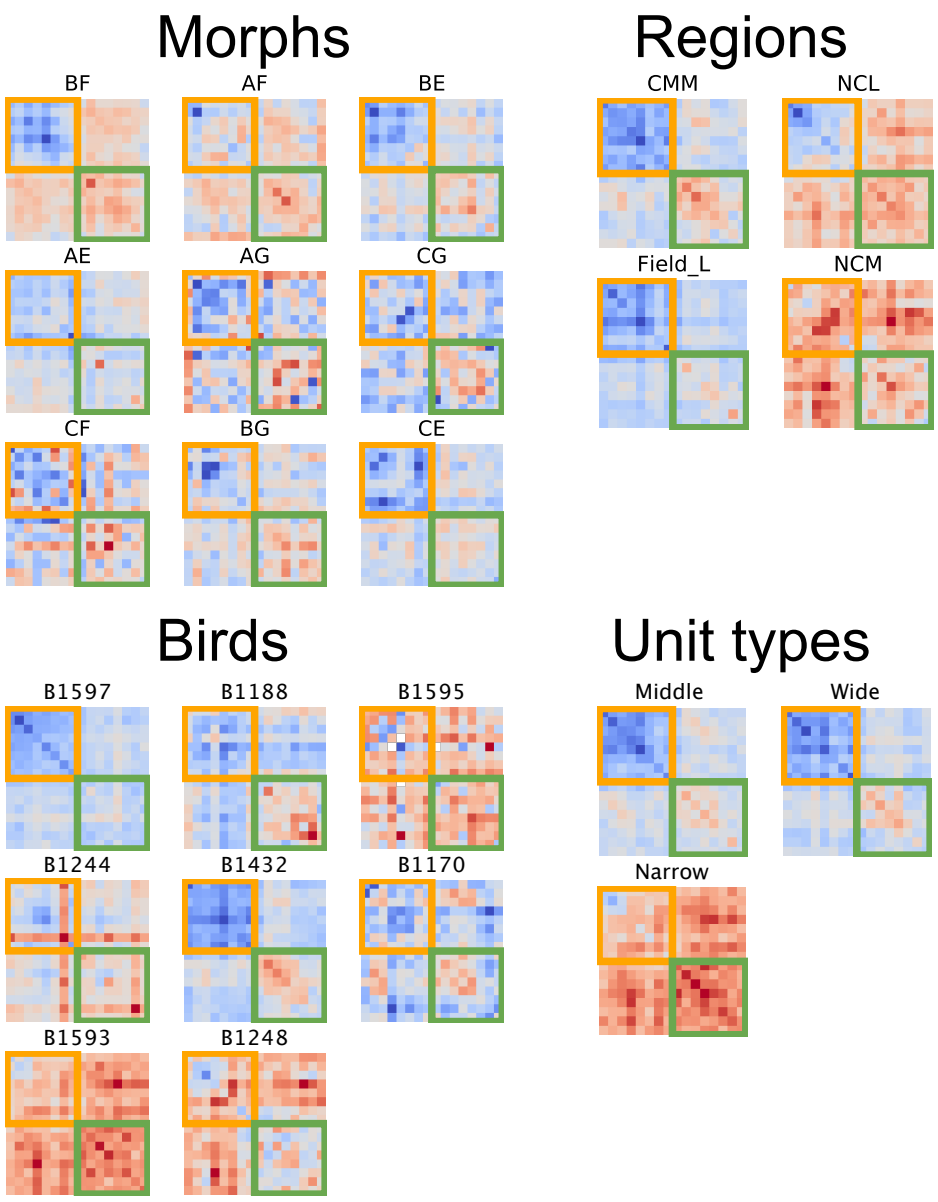


Figure 6.16. The similarity from Fig 6.6I broken out into individual morphs, brain regions, subjects, and unit types. The similarity matrix depicts the shift in spike train vector cosine similarity for left-cued minus right-cued trials. The shift is depicted here averaged is averaged across units.

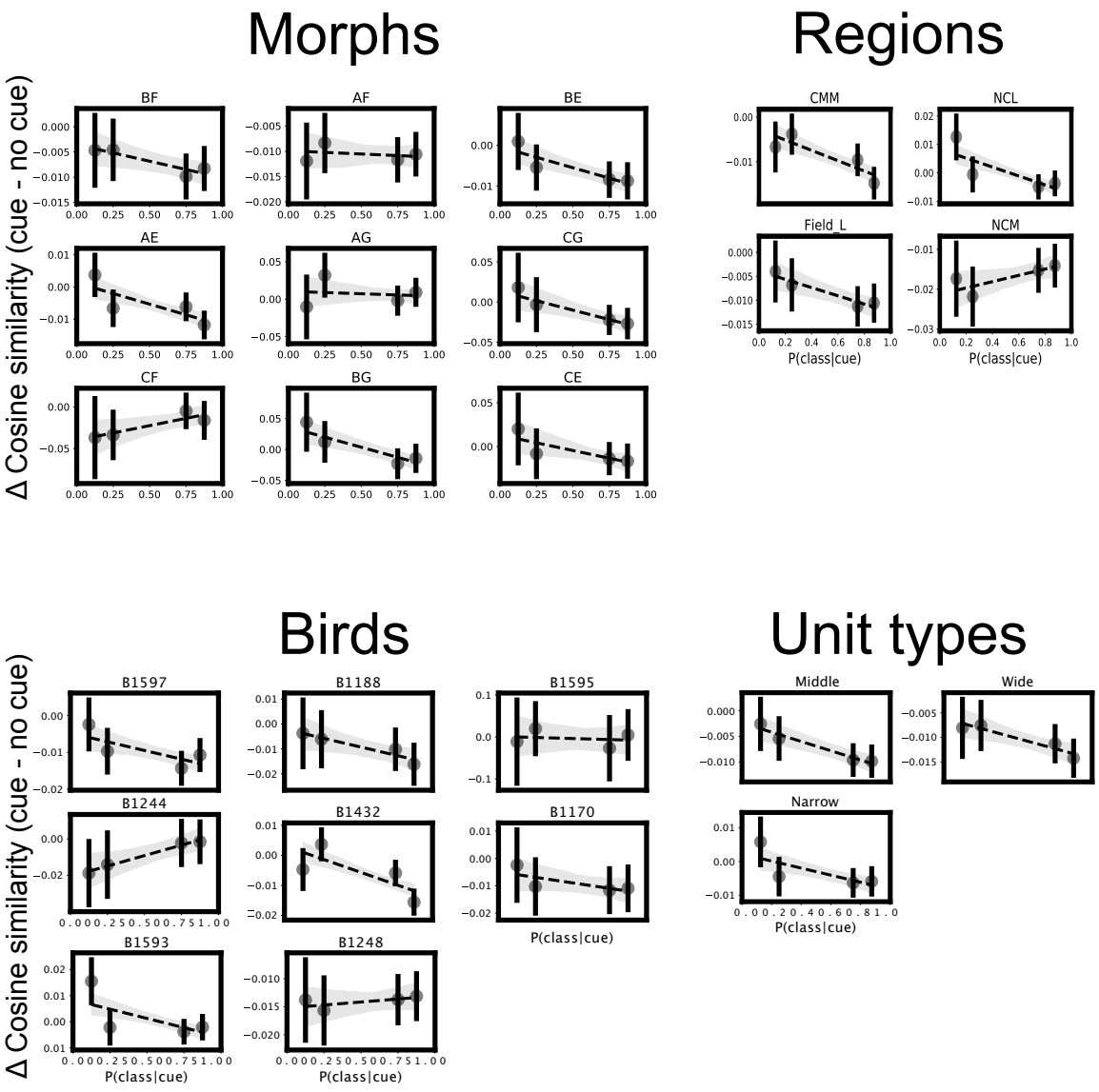


Figure 6.17. The relationship between the probability of the stimulus class and the shift in similarity from baseline (the non-cued condition) shown in Fig 6.6J, broken out into morphs, brain regions, subjects, and unit types.

6.4.37 Data and code availability

Data and code will be available upon publication.

6.4.38 Acknowledgments

This work was supported by a CARTA Fellowship to T.S., NIH 5T32MH020002-20 to T.S., and 5R01DC018055-02 to T.G.

Chapter 6, in full, is a reprint of a manuscript in preparation. Sainburg, Tim McPherson, Trevor S Arneodo, Ezequiel M. Rudraraju, Srihita Turvey, Michael Thielman, Brad Marcos, Pablo Tostado Thielk, Marvin Gentner, Timothy Q. The dissertation author was the primary investigator and author of this paper.

Bibliography

[1] Kentaro Abe and Dai Watanabe. Songbirds possess the spontaneous ability to discriminate syntactic rules. *Nature neuroscience*, 14(8):1067–1074, 2011.

[2] Kuntoro Adi, Michael T Johnson, and Tomasz S Osiejuk. Acoustic censusing using automatic vocalization classification and identity recognition. *The Journal of the Acoustical Society of America*, 127(2):874–883, 2010.

[3] M Adret-Hausberger and Peter F Jenkins. Complex organization of the warbling song in the european starling sturnus vulgaris. *Behaviour*, pages 138–156, 1988.

[4] Paolo Allegrini, Paolo Grigolini, and Luigi Palatella. Intermittency and scale-free networks: a dynamical model for human language complexity. *Chaos, Solitons & Fractals*, 20(1):95–105, 2004.

[5] Jesús B Alonso, Josué Cabrera, Rohit Shyamnani, Carlos M Travieso, Federico Bolaños, Adrián García, Alexander Villegas, and Mark Wainwright. Automatic anuran identification using noise removal and audio activity detection. *Expert Systems with Applications*, 72:83–92, 2017.

[6] Eduardo G Altmann, Giampaolo Cristadoro, and Mirko Degli Esposti. On the origin of long-range correlations in texts. *Proceedings of the National Academy of Sciences*, 109(29):11582–11587, 2012.

[7] Enrique Alvarez-Lacalle, Beate Dorow, J-P Eckmann, and Elisha Moses. Hierarchical structures induce long-range dynamical correlations in written texts. *Proceedings of the National Academy of Sciences*, 103(21):7956–7961, 2006.

[8] Ehsan Amid and Manfred K Warmuth. Trimap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.

[9] David J Anderson and Pietro Perona. Toward a science of computational ethology. *Neuron*, 84(1):18–31, 2014.

[10] Sven E Anderson, Amish S Dave, and Daniel Margoliash. Template-based automatic

recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, 100(2):1209–1219, 1996.

[11] Andrey Anikin. Soundgen: An open-source tool for synthesizing nonverbal vocalizations. *Behavior research methods*, 51(2):778–792, 2019.

[12] Ezequiel Arneodo, Shukai Chen, Vikash Gilja, and Timothy Q Gentner. A neural decoder for learned vocal behavior. *bioRxiv*, page 193987, 2017.

[13] Ezequiel M Arneodo, Shukai Chen, Daril E Brown II, Vikash Gilja, and Timothy Q Gentner. Neurally driven synthesis of learned, complex vocalizations. *Current Biology*, 2021.

[14] Ezequiel M Arneodo and Gabriel B Mindlin. Source-tract coupling in birdsong production. *Physical Review E*, 79(6):061921, 2009.

[15] Ezequiel M Arneodo, Yonatan Sanz Perl, Franz Goller, and Gabriel B Mindlin. Prosthetic avian vocal organ controlled by a freely behaving bird based on a low dimensional model of the biomechanical periphery. *PLoS computational biology*, 8(6):e1002546, 2012.

[16] Zeke Arneodo, Tim Sainburg, James Jeanne, and Timothy Gentner. An acoustically isolated european starling song library, June 2019.

[17] Zeke Arneodo, Tim Sainburg, James Jeanne, and Timothy Gentner. An acoustically isolated European starling song library, June 2019.

[18] Gustavo Arriaga, Eric P Zhou, and Erich D Jarvis. Of mice, birds, and men: the mouse ultrasonic song system has some features similar to humans and song-learning birds. 2012.

[19] Julio G Arriaga, Martin L Cody, Edgar E Vallejo, and Charles E Taylor. Bird-db: A database for annotated bird song sequences. *Ecological Informatics*, 27:21–25, 2015.

[20] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.

[21] David J Bailey, Julia C Rosebush, and Juli Wade. The hippocampus and caudomedial neostriatum show selective responsiveness to conspecific song in the female zebra finch. *Journal of neurobiology*, 52(1):43–51, 2002.

[22] Alison J Barker, Grigorii Veviurko, Nigel C Bennett, Daniel W Hart, Lina Mograby, and Gary R Lewin. Cultural transmission of vocal dialect in the naked mole-rat. *Science*, 371(6528):503–507, 2021.

[23] Peter Barthó, Hajime Hirase, Lenaïc Monconduit, Michael Zugaro, Kenneth D Harris, and Gyorgy Buzsaki. Characterization of neocortical principal cells and interneurons by network interactions and extracellular features. *Journal of neurophysiology*, 92(1):600–608, 2004.

[24] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38, 2019.

[25] Gabriël JL Beckers, Johan J Bolhuis, Kazuo Okanoya, and Robert C Berwick. Birdsong neurolinguistics: Songbird context-free grammar claim is premature. *Neuroreport*, 23(3):139–145, 2012.

[26] Michael D Beecher. Signature systems and kin recognition. *American Zoologist*, 22(3):477–490, 1982.

[27] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[28] Gordon J Berman. Measuring behavior across scales. *BMC biology*, 16(1):1–11, 2018.

[29] Gordon J Berman, William Bialek, and Joshua W Shaevitz. Predictability and hierarchy in drosophila behavior. *Proceedings of the National Academy of Sciences*, 113(42):11943–11948, 2016.

[30] David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020.

[31] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019.

[32] Robert C Berwick, Kazuo Okanoya, Gabriel JL Beckers, and Johan J Bolhuis. Songs to syntax: the linguistics of birdsong. *Trends in cognitive sciences*, 15(3):113–121, 2011.

[33] Colin Blakemore and Grahame F Cooper. Development of the brain depends on the visual environment. *Nature*, 228(5270):477–478, 1970.

[34] Guido Boffetta, Vincenzo Carbone, Paolo Giuliani, Pierluigi Veltri, and Angelo Vulpiani. Power laws in solar flares: self-organized criticality or turbulence? *Physical review letters*, 83(22):4662, 1999.

[35] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700, 2006.

[36] Kirsten M Bohn, Gerald S Wilkinson, and Cynthia F Moss. Discrimination of infant isolation calls by female greater spear-nosed bats, phyllostomus hastatus. *Animal behaviour*, 73(3):423–432, 2007.

[37] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing*, 27(2):113–120, 1979.

[38] Jordi Bonada, Robert Lachlan, and Merlijn Blaauw. Bird song synthesis based on hidden markov models. *Interspeech 2016; 2016 Sep 08-12; San Francisco (CA).[Baixas]: ISCA; 2016. p. 2582-6.*, 2016.

[39] Sarah W Bottjer and Brie Altenau. Parallel pathways for vocal learning in basal ganglia of songbirds. *Nature neuroscience*, 13(2):153–155, 2010.

[40] Matthew M Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, 12(5):201–208, 2008.

[41] Matthew M Botvinick, Yael Niv, and Andrew C Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262–280, 2009.

[42] Herve A Bourlard and Nelson Morgan. *Connectionist speech recognition: a hybrid approach*, volume 247. Springer Science & Business Media, 2012.

[43] Susan R Braunwald. Mother-child communication: the function of maternal-language input. *Word*, 27(1-3):28–50, 1971.

[44] Micah R Bregman, Aniruddh D Patel, and Timothy Q Gentner. Songbirds use spectral shape, not pitch, for sound pattern recognition. *Proceedings of the National Academy of Sciences*, 113(6):1666–1671, 2016.

[45] Alexander Brown, Saurabh Garg, and James Montgomery. Automatic and efficient denoising of bioacoustics recordings using mmse stsa. *IEEE Access*, 6:5010–5022, 2017.

[46] André EX Brown and Benjamin De Bivort. Ethology as a physical science. *Nature Physics*, 14(7):653–657, 2018.

[47] Guy J Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

[48] Roger Brown. *A first language: The early stages.* Harvard U. Press, 1973.

[49] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[50] Julia Hyland Bruno and Ofer Tchernichovski. Regularities in zebra finch song beyond the repeated motif. *Behavioural Processes*, 2017.

[51] Julia Hyland Bruno and Ofer Tchernichovski. Regularities in zebra finch song beyond the repeated motif. *Behavioural processes*, 163:53–59, 2019.

[52] Alessio P Buccino, Cole L Hurwitz, Samuel Garcia, Jeremy Magland, Joshua H Siegle, Roger Hurwitz, and Matthias H Hennig. Spikeinterface, a unified framework for spike sorting. *Elife*, 9:e61834, 2020.

[53] Kerstin Bunte, Michael Biehl, and Barbara Hammer. A general framework for dimensionality-reducing data visualization mapping. *Neural Computation*, 24(3):771–804, 2012.

[54] Kenneth P Burnham, David R Anderson, and Kathryn P Huyvaert. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65(1):23–35, 2011.

[55] Kenneth P. Burnham, David R. Anderson, and Kathryn P. Huyvaert. Aic model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65(1):23–35, Jan 2011.

[56] N. G. Burton and C. R. Licklider. Long-range constraints in the statistical structure of printed english. *American Journal of Psychology*, 68(4):650, Dec 01 1955.

[57] Richard W Byrne and Jennifer ME Byrne. Complex leaf-gathering skills of mountain gorillas (gorilla g. beringei): variability and standardization. *American Journal of Primatology*, 31(4):241–261, 1993.

[58] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer, 2013.

[59] E Canessa and A Calmetta. Physics of a random biological process. *Physical Review E*, 50(1):R47, 1994.

[60] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. Activation atlas. *Distill*, 4(3):e15, 2019.

[61] Edward C Carterette and Margaret Hubbard Jones. *Informal speech: Alphabetic & phonemic texts with statistical analyses and tables*. Univ of California Press, 1974.

[62] Gregg A Castellucci, Daniel Calbick, and David McCormick. The temporal organization of mouse ultrasonic vocalizations. *PloS one*, 13(10):e0199929, 2018.

[63] Subhojit Chakladar, Nikos K Logothetis, and Christopher I Petkov. Morphing rhesus monkey vocalizations. *Journal of neuroscience methods*, 170(1):45–55, 2008.

[64] RA Charif, AM Waack, and LM Strickman. Raven pro 1.4 user's manual. *Cornell Lab of Ornithology, Ithaca, NY*, 25506974, 2010.

[65] Zhiyi Chi and Daniel Margoliash. Temporal precision and temporal drift in brain and behavior of zebra finch song. *Neuron*, 32(5):899–910, 2001.

[66] Noam Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956.

[67] Noam Chomsky. *Syntactic structures*. Mouton, 1957.

[68] Noam Chomsky. On certain formal properties of grammars. *Information and control*, 2(2):137–167, 1959.

[69] Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019.

[70] Christopher W Clark, Peter Marler, and Kim Beeman. Quantitative analysis of animal vocal phonology: an application to swamp sparrow song. *Ethology*, 76(2):101–115, 1987.

[71] Martin L Cody, Edward Stabler, Hector Manuel Sanchez Castellanos, and Charles E Taylor. Structure, syntax and "small-world" organization in the complex songs of california thrashers (toxostoma redivivum). *Bioacoustics*, 25(1):41–54, 2016.

[72] Andy Coenen and Adam Pearce. Understanding umap, Dec 2019.

[73] Kevin R Coffey, Russell G Marx, and John F Neumaier. Deepsqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations. *Neuropsychopharmacology*, 44(5):859, 2019.

[74] Yarden Cohen, David A Nicholson, Alexa Sanchioni, Emily K Mallaber, Viktoriya Skidanova, and Timothy J Gardner. Tweetynet: A neural network that enables high-throughput, automated annotation of birdsong. *bioRxiv*, 2020.

[75] Yarden Cohen, Jun Shen, Dawit Semu, Daniel P Leman, William A Liberti, L Nathan

Perkins, Derek C Liberti, Darrell N Kotton, and Timothy J Gardner. Hidden neural states underlie canary song syntax. *Nature*, 582(7813):539–544, 2020.

[76] Yarden Cohen, Jun Shen, Dawit Semu, Daniel P Leman, William A Liberti, Nathan L Perkins, Derek C Liberti, Darrell Kotton, and Timothy J Gardner. Hidden neural states underlie canary song syntax. *bioRxiv*, page 561761, 2019.

[77] Richard Cooper and Tim Shallice. Contention scheduling and the control of routine activities. *Cognitive neuropsychology*, 17(4):297–338, 2000.

[78] Thomas Cover and Roger King. A convergent gambling estimate of the entropy of english. *IEEE Transactions on Information Theory*, 24(4):413–421, 1978.

[79] John P Cunningham and M Yu Byron. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500, 2014.

[80] C Daniel Meliza, Sara C Keen, and Dustin R Rubenstein. Pitch-and spectral-based dynamic time warping methods for comparing field recordings of harmonic avian vocalizations. *The Journal of the Acoustical Society of America*, 134(2):1407–1415, 2013.

[81] Sandeep Robert Datta, David J Anderson, Kristin Branson, Pietro Perona, and Andrew Leifer. Computational neuroethology: a call to action. *Neuron*, 104(1):11–24, 2019.

[82] Barbara L Davis and Peter F MacNeilage. The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, 38(6):1199–1211, 1995.

[83] Marian Dawkins and Richard Dawkins. Hierachical organization and postural facilitation: Rules for grooming in flies. *Animal Behaviour*, 24(4):739–755, 1976.

[84] Richard Dawkins. Hierarchical organisation: A candidate principle for ethology. *Growing points in ethology*, 7:54, 1976.

[85] Richard Dawkins. Hierarchical organisation: A candidate principle for ethology. 1976.

[86] Geert De Groof, Isabelle George, Sara Touj, Martin Stacho, Elisabeth Jonckers, Hugo Cousillas, Martine Hausberger, Onur Güntürkün, and Annemie Van der Linden. A three-dimensional digital atlas of the starling brain. *Brain Structure and Function*, 221(4):1899–1909, 2016.

[87] Vin De Silva and Joshua B Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *NIPS*, volume 15, pages 705–712, 2002.

[88] Vin De Silva and Joshua B Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. *Advances in neural information processing systems*, pages

721–728, 2003.

[89] Stanislas Dehaene, Florent Meyniel, Catherine Wacongne, Liping Wang, and Christophe Pallier. The neural representation of sequences: from transition probabilities to algebraic patterns and linguistic trees. *Neuron*, 88(1):2–19, 2015.

[90] Marty J Demetras, Kathryn Nolan Post, and Catherine E Snow. Feedback to first language learners: The role of repetitions and clarification questions. *Journal of child language*, 13(2):275–292, 1986.

[91] Bettina Diekamp, Thomas Kalt, and Onur Güntürkün. Working memory neurons in pigeons. *Journal of Neuroscience*, 22(4):RC210–RC210, 2002.

[92] Christopher DiMattina and Xiaoqin Wang. Virtual vocalization stimuli for investigating neural representations of species-specific vocalizations. *Journal of neurophysiology*, 95(2):1244–1262, 2006.

[93] Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications*, 9(1):1–13, 2018.

[94] Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *BioRxiv*, page 853457, 2019.

[95] Helen M Ditz and Andreas Nieder. Neurons selective to the number of visual items in the corvid songbird endbrain. *Proceedings of the National Academy of Sciences*, 112(25):7827–7832, 2015.

[96] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.

[97] Wei Dong, Charikar Moses, and Kai Li. Efficient k-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th international conference on World wide web*, pages 577–586, 2011.

[98] Robert J Dooling, Thomas J Park, Susan D Brown, Kazuo Okanoya, and Sigfrid D Soli. Perceptual organization of acoustic stimuli by budgerigars (melopsittacus undulatus): Ii. vocal signals. *Journal of Comparative Psychology*, 101(4):367, 1987.

[99] Robert J Dooling and Nora H Prior. Do we hear what birds hear in birdsong? *Animal behaviour*, 124:283–289, 2017.

[100] Allison J Doupe and Patricia K Kuhl. Birdsong and human speech: common themes and mechanisms. *Annual review of neuroscience*, 22(1):567–631, 1999.

[101] Homer Dudley. Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169–177, 1939.

[102] Rebecca A Dunlop, Michael J Noad, Douglas H Cato, and Dale Stokes. The social vocalization repertoire of east australian migrating humpback whales (megaptera novaeangliae). *The Journal of the Acoustical Society of America*, 122(5):2893–2905, 2007.

[103] Timothy W Dunn, Jesse D Marshall, Kyle S Severson, Diego E Aldarondo, David GC Hildebrand, Selmaan N Chettih, William L Wang, Amanda J Gellis, David E Carlson, Dmitriy Aronov, et al. Geometric deep learning enables 3d kinematic profiling across species and environments. *Nature methods*, 18(5):564–573, 2021.

[104] Andrés F Duque, Sacha Morin, Guy Wolf, and Kevin R Moon. Extendable and invertible manifold learning with geometry regularized autoencoders. *arXiv preprint arXiv:2007.07142*, 2020.

[105] Werner Ebeling and Alexander Neiman. Long-range correlations between letters and sentences in texts. *Physica A: Statistical Mechanics and its Applications*, 215(3):233–241, 1995.

[106] Werner Ebeling and Thorsten Pöschel. Entropy and long-range correlations in literary english. *EPL (Europhysics Letters)*, 26(4):241, 1994.

[107] Marcel Eens, Rianne Pinxten, and Rudolf Frans Verheyen. Temporal and sequential organization of song bouts in the starling. *Ardea*, 77(6), 1989.

[108] Donald H Eldredge, James D Miller, and Barbara A Bohne. A frequency-position map for the chinchilla cochlea. *The Journal of the Acoustical Society of America*, 69(4):1091–1095, 1981.

[109] Julie E Elie and Frederic E Theunissen. The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal cognition*, 19(2):285–315, 2016.

[110] Julie E Elie and Frédéric E Theunissen. Zebra finches identify individuals using vocal signatures unique to each call type. *Nature communications*, 9(1):1–11, 2018.

[111] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

[112] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pages 1068–1077.

PMLR, 2017.

[113] Sina Engler, Andreas Rose, and Mirjam Knörnschild. Isolation call ontogeny in bat pups (glossophaga soricina). *Behaviour*, 154(3):267–286, 2017.

[114] Yariv Ephraim and David Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing*, 32(6):1109–1121, 1984.

[115] Aysu Ezen-Can. A comparison of lstm and bert for small corpus. *arXiv preprint arXiv:2009.05451*, 2020.

[116] Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems*, pages 10078–10089, 2019.

[117] Michale S Fee, AA Kozhevnikov, and RHR Hahnloser. Neural mechanisms of vocal sequence generation in the songbird. *Ann NY Acad Sci*, 1016(1), 2004.

[118] Michale S Fee, Boris Shraiman, Bijan Pesaran, and Partha P Mitra. The role of nonlinear dynamics of the syrinx in the vocalizations of a songbird. *Nature*, 395(6697):67–71, 1998.

[119] Naomi H Feldman, Thomas L Griffiths, and James L Morgan. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752, 2009.

[120] Jean-Marc Fellous, Paul HE Tiesinga, Peter J Thomas, and Terrence J Sejnowski. Discovering spike patterns in neuronal responses. *Journal of Neuroscience*, 24(12):2989–3001, 2004.

[121] Ramon Ferrer-i Cancho and Brenda McCowan. The span of correlations in dolphin whistle sequences. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(06):P06002, 2012.

[122] Julia Fischer and Kurt Hammerschmidt. Towards a new taxonomy of primate vocal production learning. *Philosophical Transactions of the Royal Society B*, 375(1789):20190045, 2020.

[123] W Tecumseh Fitch and Angela D Friederici. Artificial grammar learning meets formal language theory: an overview. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598):1933–1955, 2012.

[124] W Tecumseh Fitch and Mauricio D Martins. Hierarchical processing in music, language, and action: Lashley revisited. *Annals of the New York Academy of Sciences*, 1316(1):87–

104, 2014.

[125] Evelyn Fox Keller. Revisiting "scale-free" networks. *BioEssays*, 27(10):1060–1068, 2005.

[126] Stefan L Frank, Rens Bod, and Morten H Christiansen. How hierarchical is language use? *Proceedings of the Royal Society of London B: Biological Sciences*, 279(1747):4522–4531, 2012.

[127] Kaitlin E Frasier, Marie A Roch, Melissa S Soldevilla, Sean M Wiggins, Lance P Garrison, and John A Hildebrand. Automated classification of dolphin echolocation click types from the gulf of mexico. *PLoS computational biology*, 13(12):e1005823, 2017.

[128] Yoav Freund. Beakedwhaleclassification. https://github.com/yoavfreund/BeakedWhaleClassification, 2019.

[129] Hisataka Fujimoto, Taku Hasegawa, and Dai Watanabe. Neural coding of syntactic structure in learned vocalizations in the songbird. *Journal of Neuroscience*, 31(27):10023–10033, 2011.

[130] Makoto Fukushima, Alex M Doyle, Matthew P Mullarkey, Mortimer Mishkin, and Bruno B Averbeck. Distributed acoustic cues for caller identity in macaque vocalization. *Royal Society open science*, 2(12):150432, 2015.

[131] Makoto Fukushima, Alexandra Doyle, Matthew Mullarkey, Mortimer Mishkin, and Bruno Averbeck. macaque coo calls, 11 2016.

[132] James Lewis Fuller. The vocal repertoire of adult male blue monkeys (cercopithecus mitis stulmanni): a quantitative analysis of acoustic structure. *American journal of primatology*, 76(3):203–216, 2014.

[133] Sean A Fulop and Kelly Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *The Journal of the Acoustical Society of America*, 119(1):360–371, 2006.

[134] Takafumi Furuyama, Kohta I Kobayasi, and Hiroshi Riquimaroux. Acoustic characteristics used by japanese macaques for individual discrimination. *Journal of Experimental Biology*, 220(19):3571–3578, 2017.

[135] Richard Futrell, Kyle Mahowald, and Edward Gibson. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341, 2015.

[136] William F Ganong. Phonetic categorization in auditory word perception. *Journal of*

experimental psychology: Human perception and performance, 6(1):110, 1980.

[137] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

[138] Tim Gardner, G Cecchi, M Magnasco, R Laje, and Gabriel B Mindlin. Simple motor gestures for birdsongs. *Physical review letters*, 87(20):208101, 2001.

[139] Timothy J Gardner and Marcelo O Magnasco. Sparse time-frequency representations. *Proceedings of the National Academy of Sciences*, 103(16):6094–6099, 2006.

[140] Timothy Q Gentner, Kimberly M Fenn, Daniel Margoliash, and Howard C Nusbaum. Recursive syntactic pattern learning by songbirds. *Nature*, 440(7088):1204–1207, 2006.

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

[142] Marcus Ghosh and Jason Rihel. Hierarchical compression reveals sub-second to day-long structure in larval zebrafish behaviour. *bioRxiv*, page 694471, 2019.

[143] Ronald Bradley Gillam and Nils A Pearson. *TNL: test of narrative language*. Pro-ed Austin, TX, 2004.

[144] Andrej Gisbrecht, Wouter Lueks, Bassam Mokbel, and Barbara Hammer. Out-of-sample kernel extensions for nonparametric dimensionality reduction. In *ESANN*, 2012.

[145] Andrej Gisbrecht, Alexander Schulz, and Barbara Hammer. Parametric nonlinear dimensionality reduction using kernel t-sne. *Neurocomputing*, 147:71–82, 2015.

[146] T Gisiger. Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biological Reviews*, 76(2):161–209, 2001.

[147] Paolo Giudici, Tobias Ryden, and Pierre Vandekerkhove. Likelihood-ratio tests for hidden markov models. *Biometrics*, 56(3):742–747, 2000.

[148] Hervé Goëau, Hervé Glotin, Willem-Pier Vellinga, Robert Planqué, Andreas Rauber, and Alexis Joly. Lifeclef bird identification task 2014. In *CLEF: Conference and Labs of the Evaluation forum*, number 1180, pages 585–597, 2014.

[149] Jack Goffinet, Samuel Brudner, Richard Mooney, and John Pearson. Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *Elife*, 10:e67855, 2021.

[150] Jack Goffinet, Richard Mooney, and John Pearson. Inferring low-dimensional latent descriptions of animal vocalizations. *bioRxiv*, page 811661, 2019.

[151] Peter Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35(3):669–675, 1989.

[152] Peter Grassberger. Entropy estimates from insufficient samplings. *Preprint at https://arxiv.org/abs/physics/0307138*, 2003.

[153] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013.

[154] Jacob M Graving and Iain D Couzin. Vae-sne: a deep generative model for simultaneous dimensionality reduction and clustering. *BioRxiv*, 2020.

[155] Patrick A Green, Nicholas C Brandley, and Stephen Nowicki. Categorical perception in animal communication and decision-making. *Behavioral Ecology*, 31(4):859–867, 2020.

[156] Patricia M Greenfield. Language, tools and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and brain sciences*, 14(4):531–551, 1991.

[157] Donald D Greenwood. The mel scale's disqualifying bias and a consistency of pitch-difference equisections in 1956 with equal cochlear distances and equal frequency ratios. *Hearing research*, 103(1-2):199–224, 1997.

[158] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984.

[159] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[160] Onur Güntürkün. The avian 'prefrontal cortex'and cognition. *Current opinion in neurobiology*, 15(6):686–693, 2005.

[161] Richard HR Hahnloser, Alexay A Kozhevnikov, and Michale S Fee. An ultra-sparse code underliesthe generation of neural sequences in a songbird. *Nature*, 419(6902):65, 2002.

[162] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[163] Biyu J He, John M Zempel, Abraham Z Snyder, and Marcus E Raichle. The temporal

structures and functional significance of scale-free brain activity. *Neuron*, 66(3):353–369, 2010.

[164] Richard Hedley. Data used in PLoS One article "Complexity, Predictability and Time Homogeneity of Syntax in the Songs of Cassin's Vireo (Vireo cassini)" by Hedley (2016). 3 2016.

[165] Richard W Hedley. Complexity, predictability and time homogeneity of syntax in the songs of cassin's vireo (vireo cassinii). *PloS one*, 11(4):e0150822, 2016.

[166] Richard W Hedley. Composition and sequential organization of song repertoires in cassin's vireo (vireo cassinii). *Journal of Ornithology*, 157(1):13–22, 2016.

[167] Jeffrey Heinz and William Idsardi. Sentence and word complexity. *Science*, 333(6040):295–297, 2011.

[168] Jeffrey Heinz and William Idsardi. What complexity differences reveal about domains in language. *Topics in Cognitive Science*, 5(1):111–131, 2013.

[169] Sascha Helduser, Sen Cheng, and Onur Güntürkün. Identification of two forebrain structures that mediate execution of memorized sequences in the pigeon. *Journal of neurophysiology*, 109(4):958–968, 2013.

[170] Sascha Helduser and Onur Güntürkün. Neural substrates for serial reaction time tasks in pigeons. *Behavioural brain research*, 230(1):132–143, 2012.

[171] Stav Hertz, Benjamin Weiner, Nisim Perets, and Michael London. High order structure in mouse courtship vocalizations. *bioRxiv*, page 728477, 2019.

[172] John A Hildebrand, Simone Baumann-Pickering, Kaitlin E Frasier, Jennifer S Trickey, Karlina P Merkens, Sean M Wiggins, Mark A McDonald, Lance P Garrison, Danielle Harris, Tiago A Marques, et al. Passive acoustic monitoring of beaked whale densities in the gulf of mexico. *Scientific reports*, 5:16343, 2015.

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

[174] Christoph Hofer, Roland Kwitt, Marc Niethammer, and Mandar Dixit. Connectivity-optimized representation learning via persistent homology. In *International Conference on Machine Learning*, pages 2751–2760. PMLR, 2019.

[175] Marie-Jeanne Holveck, Ana Catarina Vieira de Castro, Robert F Lachlan, Carel ten Cate, and Katharina Riebel. Accuracy of song syntax learning and singing consistency signal early condition in zebra finches. *Behavioral Ecology*, 19(6):1267–1281, 2008.

[176] Timothy E Holy and Zhongsheng Guo. Ultrasonic songs of male mice. *PLoS biology*, 3(12):e386, 2005.

[177] Brian Hopkins and John Gordon Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, 1954.

[178] Sheng-Bin Hsu, Chang-Hsing Lee, Pei-Chun Chang, Chin-Chuan Han, and Kuo-Chin Fan. Local wavelet acoustic pattern: A novel time–frequency descriptor for birdsong recognition. *IEEE Transactions on Multimedia*, 20(12):3187–3199, 2018.

[179] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[180] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[181] Paul Iverson and Patricia K Kuhl. Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, 97(1):553–562, 1995.

[182] Vincent M Janik. Pitfalls in the categorization of behaviour: a comparison of dolphin whistle classification methods. *Animal Behaviour*, 57(1):133–143, 1999.

[183] Eathan Janney, Hollis Taylor, Constance Scharff, David Rothenberg, Lucas C Parra, and Ofer Tchernichovski. Temporal regularity increases with repertoire complexity in the australian pied butcherbird's song. *Royal Society open science*, 3(9):160357, 2016.

[184] James M Jeanne, Tatyana O Sharpee, and Timothy Q Gentner. Associative learning enhances population coding by inverting interneuronal correlation patterns. *Neuron*, 78(2):352–363, 2013.

[185] Kui Jia, Lin Sun, Shenghua Gao, Zhan Song, and Bertram E Shi. Laplacian auto-encoders: An explicit learning of nonlinear data manifold. *Neurocomputing*, 160:250–260, 2015.

[186] Xinjian Jiang, Tenghai Long, Weicong Cao, Junru Li, Stanislas Dehaene, and Liping Wang. Production of supra-regular spatial sequences by macaque monkeys. *Current Biology*, 28(12):1851–1859, 2018.

[187] Dezhe Z Jin and Alexay A Kozhevnikov. A compact statistical model of the song syntax in bengalese finch. *PLoS computational biology*, 7(3):e1001108, 2011.

[188] Dan Jurafsky and James H Martin. *NGrams*, chapter 4. Prentice Hall, Pearson Education International, 2009.

[189] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR, 2018.

[190] Tobias Kalenscher, Bettina Diekamp, and Onur GuÈntuÈrkuÈn. Neural architecture of choice behaviour in a concurrent interval schedule. *European Journal of Neuroscience*, 18(9):2627–2637, 2003.

[191] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[192] Ronald M Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, 1994.

[193] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE, 2019.

[194] Kentaro Katahira, Kenta Suzuki, Hiroko Kagawa, and Kazuo Okanoya. A simple explanation for the evolution of complex song syntax in bengalese finches. *Biology letters*, 9(6):20130842, 2013.

[195] Kentaro Katahira, Kenta Suzuki, Kazuo Okanoya, and Masato Okada. Complex sequencing rules of birdsong can be explained by simple hidden markov processes. *PloS one*, 6(9):e24516, 2011.

[196] Hideki Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[197] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain De Cheveigne. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, 27(3-4):187–207, 1999.

[198] Sara Keen, Jesse C Ross, Emily T Griffiths, Michael Lanzone, and Andrew Farnsworth. A comparison of similarity-based approaches in the classification of flight calls of four species of north american wood-warblers (parulidae). *Ecological Informatics*, 21:25–33, 2014.

[199] Sara C Keen, Karan J Odom, Michael S Webster, Gregory M Kohn, Timothy F Wright, and Marcelo Araya-Salas. A machine learning approach for classifying and quantifying acoustic diversity. *Methods in Ecology and Evolution*, 2021.

[200] Arik Kershenbaum, Daniel T Blumstein, Marie A Roch, Çağlar Akçay, Gregory Backus, Mark A Bee, Kirsten Bohn, Yan Cao, Gerald Carter, Cristiane Cäsar, et al. Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews*, 91(1):13–52, 2016.

[201] Arik Kershenbaum, Ann E Bowles, Todd M Freeberg, Dezhe Z Jin, Adriano R Lameira, and Kirsten Bohn. Animal vocal sequences: not the markov chains we thought they were. *Proceedings of the Royal Society B: Biological Sciences*, 281(1792):20141370, 2014.

[202] Davi Miara Kiapuchinski, Carlos Raimundo Erig Lima, and Celso Antônio Alves Kaestner. Spectral noise gate technique applied to birdsong preprocessing on embedded unit. In *2012 IEEE International Symposium on Multimedia*, pages 24–27. IEEE, 2012.

[203] Bongjun Kim and Bryan Pardo. A human-in-the-loop system for sound event detection and annotation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–23, 2018.

[204] Hyoung-Gook Kim, Klaus Obermayer, Mathias Bode, and Dietmar Ruwisch. Real-time noise canceling based on spectral minimum detection and diffusive gain factors. *The Journal of the Acoustical Society of America*, 108(5):2484–2484, 2000.

[205] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[206] KlugerLab. Open source survey. https://github.com/KlugerLab/FIt-SNE, commit = 4f57d6a0e4c030202a07a60bc1bb1ed1544bf679, 2020.

[207] Elly C Knight, Sergio Poo Hernandez, Erin M Bayne, Vadim Bulitko, and Benjamin V Tucker. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, pages 1–19, 2019.

[208] Elly C Knight, Sergio Poo Hernandez, Erin M Bayne, Vadim Bulitko, and Benjamin V Tucker. Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks. *Bioacoustics*, 29(3):337–355, 2020.

[209] Dmitry Kobak and George C Linderman. Initialization is critical for preserving global data structure in both t-sne and umap. *Nature Biotechnology*, pages 1–2, 2021.

[210] Masanori Kobayashi and Toshimitsu Musha. 1/f fluctuation of heartbeat period. *IEEE transactions on Biomedical Engineering*, (6):456–457, 1982.

[211] Etienne Koechlin, Chrystele Ody, and Frédérique Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181–1185, 2003.

[212] Joseph A Kogan and Daniel Margoliash. Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden markov models: A comparative study. *The Journal of the Acoustical Society of America*, 103(4):2185–2196, 1998.

[213] Daniel Kohlsdorf, Denise Herzing, and Thad Starner. An auto encoder for audio dolphin communication. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[214] Peter Kok, Janneke FM Jehee, and Floris P De Lange. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, 75(2):265–270, 2012.

[215] Sepp Kollmorgen, Richard Hahnloser, and Valerio Mante. Neighborhood-statistics reveal complex dynamics of song acquisition in the zebra finch. *bioRxiv*, page 595512, 2019.

[216] Sepp Kollmorgen, Richard HR Hahnloser, and Valerio Mante. Nearest neighbours reveal fast and slow components of motor learning. *Nature*, 577(7791):526–530, 2020.

[217] Qiuqiang Kong, Yong Xu, and Mark D Plumbley. Joint detection and classification convolutional neural network on weakly labelled bird audio detection. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 1749–1753. IEEE, 2017.

[218] Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244, 2004.

[219] Takuya Koumura. BirdsongRecognition. *Figshare*, 7 2016.

[220] Takuya Koumura and Kazuo Okanoya. Automatic recognition of element classes and boundaries in the birdsong with variable sequences. *PloS one*, 11(7):e0159188, 2016.

[221] Valeri Aleksandrovich Kozhevnikov and Liudmila Andreevna Chistovich. Speech: Articulation and perception. 1965.

[222] Thomas Krefeld and Stephen Lucke. ASICA-online: Profilo di un nuovo atlante sintattico della Calabria. pages 169–211, 2008.

[223] Sven Kröner and Onur Güntürkün. Afferent and efferent connections of the caudolateral neostriatum in the pigeon (columba livia): a retro-and anterograde pathway tracing study. *Journal of Comparative Neurology*, 407(2):228–260, 1999.

[224] Patricia K Kuhl. Human adults and human infants show a "perceptual magnet effect"

for the prototypes of speech categories, monkeys do not. *Perception & psychophysics*, 50(2):93–107, 1991.

[225] Patricia K Kuhl. Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, 5(11):831–843, 2004.

[226] Patricia K Kuhl and James D Miller. Speech perception by the chinchilla: Identification functions for synthetic vot stimuli. *The Journal of the Acoustical Society of America*, 63(3):905–917, 1978.

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

[228] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. *arXiv preprint arXiv:1910.06711*, 2019.

[229] Gina R Kuperberg and T Florian Jaeger. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience*, 31(1):32–59, 2016.

[230] RF Lachlan. Luscinia: a bioacoustics analysis computer program. *See luscinia. sourceforge. net*, 2007.

[231] RF Lachlan, L Verhagen, S Peters, and C ten Cate. Are there species-universal categories in bird song phonology and syntax? a comparative study of chaffinches (fringilla coelebs), zebra finches (taenopygia guttata), and swamp sparrows (melospiza georgiana). *Journal of Comparative Psychology*, 124(1):92, 2010.

[232] Robert Lachlan and Oliver Ratmann. Data-set for Lachlan et al. 2018. *Figshare*, 5 2018.

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

[234] Robert F Lachlan, Oliver Ratmann, and Stephen Nowicki. Cultural conformity generates extremely stable traditions in bird song. *Nature communications*, 9(1):2417, 2018.

[235] Robert F Lachlan, Machteld N Verzijden, Caroline S Bernard, Peter-Paul Jonker, Bram Koese, Shirley Jaarsma, Willemijn Spoor, Peter JB Slater, and Carel ten Cate. The progressive loss of syntactical structure in bird song along an island colonization chain. *Current Biology*, 23(19):1896–1901, 2013.

[236] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther.

Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, pages 1558–1566. PMLR, 2016.

[237] Karl Spencer Lashley. *The problem of serial order in behavior*, volume 21. Bobbs-Merrill, 1951.

[238] Mario Lasseck. Bird song classification in field recordings: winning solution for nips4b 2013 competition. In *Proc. of int. symp. Neural Information Scaled for Bioacoustics, sabiod. org/nips4b, joint to NIPS, Nevada*, pages 176–181, 2013.

[239] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[240] Geon Woo Lee and Hong Kook Kim. Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection. *Applied Sciences*, 10(9):3230, 2020.

[241] John A Lee, Diego H Peluffo-Ordóñez, and Michel Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[242] Louis Lefebvre. Grooming in crickets: timing and hierarchical organization. *Animal Behaviour*, 29(4):973–984, 1981.

[243] Louis Lefebvre. The organization of grooming in budgerigars. *Behavioural processes*, 7(2):93–106, 1982.

[244] Daniel J Levitin, Parag Chordia, and Vinod Menon. Musical rhythm spectra from Bach to Joplin obey a 1/f power law. *Proceedings of the National Academy of Sciences*, 109(10):3716–3720, 2012.

[245] Wentian Li. Power spectra of regular languages and cellular automata. *Complex Systems*, 1(1):107–130, 1987.

[246] Wentian Li. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60(5):823–837, 1990.

[247] Wentian Li and Kunihiko Kaneko. Long-range correlation and partial 1/fα spectrum in a noncoding DNA sequence. *EPL (Europhysics Letters)*, 17(7):655, 1992.

[248] Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358, 1957.

[249] Henry W Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

[250] Henry W Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7):299, 2017.

[251] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Efficient algorithms for t-distributed stochastic neighborhood embedding. *arXiv preprint arXiv:1712.09005*, 2017.

[252] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. *Nature methods*, 16(3):243–245, 2019.

[253] Klaus Linkenkaer-Hansen, Vadim V. Nikouline, J. Matias Palva, and Risto J. Ilmoniemi. Long-range temporal correlations and scaling behavior in human brain oscillations. *The Journal of Neuroscience*, 21(4):1370–1377, February 2001.

[254] Shi Tong Liu, Pilar Montes-Lourido, Xiaoqin Wang, and Srivatsun Sadagopan. Optimal features for auditory categorization. *Nature communications*, 10(1):1–14, 2019.

[255] Yen Yi Loo and Kristal E Cain. A call to expand avian vocal development research. *Frontiers in Ecology and Evolution*, page 772, 2021.

[256] Vincent Lostanlen, Kaitlin Palmer, Elly Knight, Christopher Clark, Holger Klinck, Andrew Farnsworth, Tina Wong, Jason Cramer, and Juan Pablo Bello. Long-distance detection of bioacoustic events with per-channel energy normalization. *arXiv preprint arXiv:1911.00417*, 2019.

[257] Vincent Lostanlen, Justin Salamon, Mark Cartwright, Brian McFee, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Per-channel energy normalization: Why and how. *IEEE Signal Processing Letters*, 26(1):39–43, 2018.

[258] Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello. Robust sound event detection in bioacoustic sensor networks. *PloS one*, 14(10):e0214168, 2019.

[259] Shouwen Ma, Andries Ter Maat, and Manfred Gahr. Power-law scaling of calling dynamics in zebra finches. *Scientific reports*, 7(1):8397, 2017.

[260] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[261] Maryellen C MacDonald. How language production shapes language form and compre-

hension. *Frontiers in Psychology*, 4:226, 2013.

[262] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[263] Brian MacWhinney. The childes project: Tools for analyzing talk: Volume i: Transcription format and programs, volume ii: The database, 2000.

[264] Kikuo Maekawa. Corpus of Spontaneous Japanese: Its design and evaluation. In *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[265] Linda Main and John Thornton. A cortically-inspired model for bioacoustics recognition. In *International Conference on Neural Information Processing*, pages 348–355. Springer, 2015.

[266] Jeffrey E Markowitz, Elizabeth Ivie, Laura Kligler, and Timothy J Gardner. Long-range order in canary song. *PLoS computational biology*, 9(5):e1003052, 2013.

[267] Peter Marler and Roberta Pickert. Species-universal microstructure in the learned song of the swamp sparrow (melospiza georgiana). *Animal Behaviour*, 32(3):673–689, 1984.

[268] João C Marques, Simone Lackner, Rita Félix, and Michael B Orger. Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering. *Current Biology*, 28(2):181–195, 2018.

[269] Jesse D Marshall, Diego E Aldarondo, Timothy W Dunn, William L Wang, Gordon J Berman, and Bence P Ölveczky. Continuous whole-body 3d kinematic recordings across the rodent behavioral repertoire. *Neuron*, 109(3):420–437, 2021.

[270] William D Marslen-Wilson and Alan Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1):29–63, 1978.

[271] Elise F Masur and Jean B Gleason. Parent–child interaction and the acquisition of lexical information during play. *Developmental Psychology*, 16(5):404, 1980.

[272] Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada. Deep griffin–lim iteration. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 61–65. IEEE, 2019.

[273] Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289,

2018.

[274] Elizabeth A Maylor, Nick Chater, and Gordon DA Brown. Scale invariance in the retrieval of retrospective and prospective memories. *Psychonomic Bulletin & Review*, 8(1):162–167, 2001.

[275] Josh H McDermott, Andrew J Oxenham, and Eero P Simoncelli. Sound texture synthesis via filter statistics. In *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 297–300. IEEE, 2009.

[276] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–940, 2011.

[277] Kyle McDonald. Data of the humpback whale, June 2019. [Online; posted 05-June-2019].

[278] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *J. Open Source Software*, 2(11):205, 2017.

[279] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), mar 2017.

[280] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[281] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. Samplernn: An unconditional end-to-end neural audio generation model. *arXiv preprint arXiv:1612.07837*, 2016.

[282] David K Mellinger and Christopher W Clark. Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, 107(6):3518–3529, 2000.

[283] David K Mellinger and Christopher W Clark. Mobysound: A reference archive for studying automatic recognition of marine mammal sounds. *Applied Acoustics*, 67(11-12):1226–1242, 2006.

[284] SS Melnyk, OV Usatenko, VA Yampol'skii, and VA Golick. Competition between two kinds of correlations in literary texts. *Physical Review E*, 72(2):026140, 2005.

[285] David G Mets and Michael S Brainard. An automated approach to the quantitation of vocalizations and vocal learning in the songbird. *PLoS computational biology*, 14(8):e1006437, 2018.

[286] Alexander Mielke and Klaus Zuberbühler. A method for automated individual, species

and call type recognition in free-ranging animals. *Animal Behaviour*, 86(2):475–482, 2013.

[287] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[288] Cory T Miller, Katherine Mandel, and Xiaoqin Wang. The communicative content of the common marmoset phee call during antiphonal calling. *American journal of primatology*, 72(11):974–980, 2010.

[289] Edward H Miller. An approach to the analysis of graded vocalizations of birds. *Behavioral and Neural Biology*, 27(1):25–38, 1979.

[290] Michelle Milmine, Arii Watanabe, and Michael Colombo. Neural correlates of directed forgetting in the avian prefrontal cortex. *Behavioral neuroscience*, 122(1):199, 2008.

[291] Gal Mishne, Uri Shaham, Alexander Cloninger, and Israel Cohen. Diffusion nets. *Applied and Computational Harmonic Analysis*, 47(2):259–285, 2019.

[292] Tomoko Mizuhara and Kazuo Okanoya. Do songbirds hear songs syllable by syllable? *Behavioural processes*, 174:104089, 2020.

[293] Vasile V Moca, Harald Bârzan, Adriana Nagy-Dăbâcan, and Raul C Mureșan. Time-frequency super-resolution with superlets. *Nature communications*, 12(1):1–18, 2021.

[294] Ernst Moerk. Factors of style and personality. *Journal of psycholinguistic research*, 1(3):257–268, 1972.

[295] Felix W Moll and Andreas Nieder. Cross-modal associative mnemonic signals in crow endbrain neurons. *Current Biology*, 25(16):2196–2201, 2015.

[296] Felix W Moll and Andreas Nieder. Modality-invariant audio-visual association coding in crow endbrain neurons. *Neurobiology of learning and memory*, 137:65–76, 2017.

[297] Marcelo A Montemurro and Pedro A Pury. Long-range fractal correlations in literary corpora. *Fractals*, 10(04):451–461, 2002.

[298] Marcelo A Montemurro and Damián H Zanette. Entropic analysis of the role of words in literary texts. *Advances in complex systems*, 5(01):7–17, 2002.

[299] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.

[300] Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

[301] Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. *ICML*, 2020.

[302] Veronica Morfi, Robert F Lachlan, and Dan Stowell. Deep perceptual embeddings for unlabelled animal sound events. *The Journal of the Acoustical Society of America*, 150(1):2–11, 2021.

[303] Takashi Morita and Hiroki Koda. Superregular grammars do not provide additional explanatory power but allow for a compact analysis of animal song. *Royal Society open science*, 6(7):190139, 2019.

[304] Takashi Morita, Hiroki Koda, Kazuo Okanoya, and Ryosuke O Tachibana. Birdsong sequence exhibits long context dependency comparable to human language syntax. *bioRxiv*, 2020.

[305] Takashi Morita, Hiroki Koda, Kazuo Okanoya, and Ryosuke O Tachibana. Measuring long context dependency in birdsong using an artificial neural network with a long-lasting working memory. *bioRxiv*, 2020.

[306] David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

[307] Solveig C Mouterde, Frédéric E Theunissen, Julie E Elie, Clémentine Vignal, and Nicolas Mathevon. Acoustic communication and sound degradation: how do the individual signatures of male and female zebra finch calls transmit over distance? *PloS one*, 9(7), 2014.

[308] Maura Jones Moyle, Susan Ellis Weismer, Julia L Evans, and Mary J Lindstrom. Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research*, 2007.

[309] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[310] Christina AS Mumm and Mirjam Knörnschild. The vocal repertoire of adult and neonate giant otters (pteronura brasiliensis). *PloS one*, 9(11):e112562, 2014.

[311] Miguel A Munoz. Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics*, 90(3):031001, 2018.

[312] Rafael Hernández Murcia and Víctor Suárez Paniagua. The icml 2013 bird challenge. 2013.

[313] Ashwin Narayan, Bonnie Berger, and Hyunghoon Cho. Density-preserving data visualization unveils dynamic patterns of single-cell transcriptomic variability. *bioRxiv*, 2020.

[314] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165, 2019.

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

[316] Edwin B. Newman. The pattern of vowels and consonants in various languages. *The American Journal of Psychology*, 64:369–379, 1951.

[317] Edwin B. Newman and Louis J. Gerstman. A new method for analyzing printed english. *Journal of experimental psychology*, 44(2):114–125, 08 1952.

[318] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[319] Matthew Newville, Till Stensitzki, Daniel B Allen, Michal Rawlik, Antonino Ingargiola, and Andrew Nelson. Lmfit: non-linear least-square minimization and curve-fitting for Python. *Astrophysics Source Code Library*, 2016.

[320] Johanna G Nicholas and Ann E Geers. Communication of oral deaf and normally hearing children at 36 months of age. *Journal of Speech, Language, and Hearing Research*, 40(6):1314–1327, 1997.

[321] David Nicholson. Comparison of machine learning methods applied to birdsong element classification. In *Proceedings of the 15th Python in Science Conference*, pages 57–61, 2016.

[322] David Nicholson, Jonah E. Queen, and Samuel J. Sober. Bengalese Finch song repository. *Figshare*, 10 2017.

[323] Andreas Nieder. Consciousness without cortex. *Current Opinion in Neurobiology*, 71:69–76, 2021.

[324] Dennis Norris, James M McQueen, and Anne Cutler. Prediction, bayesian inference and feedback in speech recognition. *Language, cognition and neuroscience*, 31(1):4–18, 2016.

[325] Daniel E Okobi, Arkarup Banerjee, Andrew MM Matheson, Steven M Phelps, and Michael A Long. Motor cortical control of vocal interaction in neotropical singing mice. *Science*, 363(6430):983–988, 2019.

[326] Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. In *Advances in neural information processing systems*, pages 3235–3246, 2018.

[327] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[328] Takashi Otake, Giyoo Hatano, Anne Cutler, and Jacques Mehler. Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, 32(2):258–278, 1993.

[329] Marius Pachitariu, Nicholas A Steinmetz, Shabnam N Kadir, Matteo Carandini, and Kenneth D Harris. Fast and accurate spike sorting of high-channel count probes with kilosort. *Advances in neural information processing systems*, 29:4448–4456, 2016.

[330] Silvia Pagliarini, Nathan Trouvain, Arthur Leblois, and Xavier Hinaut. What does the canary say? low-dimensional gan applied to birdsong. 2021.

[331] Gautam Pai, Ronen Talmon, Alex Bronstein, and Ron Kimmel. Dimal: Deep isometric manifold learning using sparse geodesic sampling. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 819–828. IEEE, 2019.

[332] J Matias Palva, Alexander Zhigalov, Jonni Hirvonen, Onerva Korhonen, Klaus Linkenkaer-Hansen, and Satu Palva. Neuronal long-range temporal correlations and avalanche dynamics are correlated with behavioral scaling laws. *Proceedings of the National Academy of Sciences*, 110(9):3585–3590, 2013.

[333] Chethan Pandarinath, Daniel J O'Shea, Jasmine Collins, Rafal Jozefowicz, Sergey D Stavisky, Jonathan C Kao, Eric M Trautmann, Matthew T Kaufman, Stephen I Ryu, Leigh R Hochberg, et al. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature methods*, 15(10):805–815, 2018.

[334] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019.

[335] Jennifer M Parsons. *Positional effects in phonological development: a case study*. PhD thesis, Memorial University of Newfoundland, 2006.

[336] Gail L Patricelli and Eileen A Hebets. New dimensions in animal communication: the case for complexity. *Current Opinion in Behavioral Sciences*, 12:80–89, 2016.

[337] Ben Pearre, L Nathan Perkins, Jeffrey E Markowitz, and Timothy J Gardner. A fast and accurate zebra finch syllable detector. *PloS one*, 12(7):e0181992, 2017.

[338] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[339] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[340] C-K Peng, Sergej V Buldyrev, Ary L Goldberger, Shlomo Havlin, Francesco Sciortino, Michael Simons, and HE Stanley. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168, 1992.

[341] C-K Peng, J Mietus, JM Hausdorff, Shlomo Havlin, H Eugene Stanley, and Ary L Goldberger. Long-range anticorrelations and non-gaussian behavior of the heartbeat. *Physical review letters*, 70(9):1343, 1993.

[342] Talmo D Pereira, Deigo E Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S-H Wang, Mala Murthy, and Joshua Shaevitz. Fast animal pose estimation using deep neural networks. *Nature Methods 16, 117–125*, 2019.

[343] Talmo D Pereira, Diego E Aldarondo, Lindsay Willmore, Mikhail Kislin, Samuel S-H Wang, Mala Murthy, and Joshua W Shaevitz. Fast animal pose estimation using deep neural networks. *Nature methods*, 16(1):117–125, 2019.

[344] Yonatan Sanz Perl, Ezequiel M Arneodo, Ana Amador, Franz Goller, and Gabriel B Mindlin. Reconstruction of physiological instructions from zebra finch song. *Physical Review E*, 84(5):051909, 2011.

[345] Christopher I Petkov and Erich Jarvis. Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Frontiers in evolutionary neuroscience*, 4:12, 2012.

[346] Mark A Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth

Hume, and Eric Fosler-Lussier. Buckeye corpus of conversational speech. *Ohio State University (Distributor)*, 2007.

[347] Pavlin G. Poličar, Martin Stražar, and Blaž Zupan. opentsne: a modular python library for t-sne dimensionality reduction and embedding. *bioRxiv*, 2019.

[348] Yosef Prat, Mor Taub, Ester Pratt, and Yossi Yovel. An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny. *Scientific data*, 4:170143, 2017.

[349] Yosef Prat, Mor Taub, Ester Pratt, and Yossi Yovel. An annotated dataset of egyptian fruit bat vocalizations across varying contexts and during vocal ontogeny, September 2017.

[350] Jonathan F Prather, Stephen Nowicki, Rindy C Anderson, Susan Peters, and Richard Mooney. Neural correlates of categorical perception in learned vocal communication. *Nature neuroscience*, 12(2):221, 2009.

[351] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 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, 2019.

[352] Nirosha Priyadarshani, Stephen Marsland, and Isabel Castro. Automated birdsong recognition in complex acoustic environments: a review. *Journal of Avian Biology*, 49(5):jav–01447, 2018.

[353] Nirosha Priyadarshani, Stephen Marsland, Isabel Castro, and Amal Punchihewa. Birdsong denoising using wavelets. *PloS one*, 11(1):e0146790, 2016.

[354] Nirosha Priyadarshani, Stephen Marsland, Julius Juodakis, Isabel Castro, and Virginia Listanti. Wavelet filters for automated recognition of birdsong in long-time field recordings. *Methods in Ecology and Evolution*, 11(3):403–417, 2020.

[355] Jill D Pruetz and Paco Bertolani. Savanna chimpanzees, pan troglodytes verus, hunt with tools. *Current biology*, 17(5):412–417, 2007.

[356] Zdenek Pruusa and Pavel Rajmic. Toward high-quality real-time signal reconstruction from stft magnitude. *IEEE signal processing letters*, 24(6):892–896, 2017.

[357] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[358] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[359] Monzilur Rahman, Ben DB Willmore, Andrew J King, and Nicol S Harper. Simple transformations capture auditory input to cortex. *Proceedings of the National Academy of Sciences*, 117(45):28442–28451, 2020.

[360] Stefan Rapp. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov models—An aligner for German. In *Proceedings of ELSNET Goes East and IMACS Workshop "Integration of Language and Speech in Academia and Industry" Moscow, Russia)*, 1995.

[361] David Reby, Régine André-Obrecht, Arnaud Galinier, Jerome Farinas, and Bruno Cargnelutti. Cepstral coefficients and hidden markov models reveal idiosyncratic voice characteristics in red deer (cervus elaphus) stags. *The Journal of the Acoustical Society of America*, 120(6):4080–4089, 2006.

[362] Yao Ren, Michael T Johnson, and Jidong Tao. Perceptually motivated wavelet packet transform for bioacoustic signal enhancement. *The Journal of the Acoustical Society of America*, 124(1):316–327, 2008.

[363] Santiago Renteria, Edgar Vallejo, and Charles Taylor. Birdsong phrase verification and classification using siamese neural networks. *bioRxiv*, 2021.

[364] Paul Rinnert, Maximilian E Kirschhock, and Andreas Nieder. Neuronal correlates of spatial working memory in the endbrain of crows. *Current Biology*, 29(16):2616–2624, 2019.

[365] Lauren V Riters, Marcel Eens, Rianne Pinxten, Deborah L Duffy, Jacques Balthazart, and Gregory F Ball. Seasonal changes in courtship song and the medial preoptic area in male european starlings (sturnus vulgaris). *Hormones and behavior*, 38(4):250–261, 2000.

[366] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. A hierarchical latent vector model for learning long-term structure in music. In *International conference on machine learning*, pages 4364–4373. PMLR, 2018.

[367] Isaac Robinson. Interpretable visualizations with differentiating embedding networks. *arXiv preprint arXiv:2006.06640*, 2020.

[368] Tina C Roeske, Damian Kelty-Stephen, and Sebastian Wallot. Multifractal analysis reveals music-like dynamic structure in songbird rhythms. *Scientific Reports*, 8(1):4570, 2018.

[369] Martin Rohrmeier, Willem Zuidema, Geraint A Wiggins, and Constance Scharff. Principles of structure building in music, language and animal song. *Philosophical transactions of the Royal Society B: Biological sciences*, 370(1664):20140097, 2015.

[370] Jonas Rose and Michael Colombo. Neural correlates of executive control in the avian

brain. *PLoS biology*, 3(6):e190, 2005.

[371] Yvan Rose and Brian MacWhinney. The phonbank project: Data and software-assisted methods for the study of phonology and phonological development. 2014.

[372] Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420, 2007.

[373] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[374] Brian E Russ, Yune-Sang Lee, and Yale E Cohen. Neural and behavioral correlates of auditory categorization. *Hearing research*, 229(1-2):204–212, 2007.

[375] Tim Sainburg. timsainb/noisereduce: v1.0.1. https://github.com/timsainb/noisereduce, June 2019.

[376] Tim Sainburg. Vocalseg. https://github.com/timsainb/vocalization-segmentation, 2019.

[377] Tim Sainburg. Code for "finding, visualizing, and quantifying latent structure across diverse animal vocal communication signals". https://github.com/timsainb/avgn_paper, 2020.

[378] Tim Sainburg. timsainb/noisereduce: v2.0.0, July 2021.

[379] Tim Sainburg, Anna Mai, and Timothy Q Gentner. Long-range sequential dependencies precede complex syntactic production in language acquisition. *bioRxiv*, 2020.

[380] Tim Sainburg, Leland McInnes, and Timothy Q Gentner. Parametric umap: learning embeddings with deep neural networks for representation and semi-supervised learning. *Neural Computation*, 2021.

[381] Tim Sainburg, Brad Theilman, Marvin Thielk, and Timothy Q Gentner. Parallels in the sequential organization of birdsong and human speech. *Nature communications*, 10(1):1–11, 2019.

[382] Tim Sainburg, Marvin Thielk, and Timothy Gentner. Learned context dependent categorical perception in a songbird. In *Conference on Cognitive Computational Neuroscience*, pages 1–4, 2018.

[383] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv*, 2019.

[384] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020.

[385] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Latent space visualization, characterization, and generation of diverse vocal communication signals. *bioRxiv*, page 870311, 2020.

[386] Tim Sainburg, Marvin Thielk, Brad Theilman, Benjamin Migliori, and Timothy Gentner. Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. *arXiv preprint arXiv:1807.06650*, 2018.

[387] Nicolas Saint-Arnaud and Kris Popat. Analysis and synthesis of sound textures. In *in Readings in Computational Auditory Scene Analysis*. Citeseer, 1995.

[388] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, pages 1163–1171, 2016.

[389] Haşim Sak, Andrew Senior, and Françoise Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *Fifteenth annual conference of the international speech communication association*, 2014.

[390] Kazutoshi Sasahara, Martin L Cody, David Cohen, and Charles E Taylor. Structural design principles of complex bird songs: a network-based approach. 2012.

[391] Alain Schenkel, Jun Zhang, and Yi-Cheng Zhang. Long range correlation in human writings. *Fractals*, 1(01):47–57, 1993.

[392] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[393] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

[394] Bert Schouten, Ellen Gerrits, and Arjan Van Hessen. The end of categorical perception as we know it. *Speech communication*, 41(1):71–80, 2003.

[395] Jacob Schreiber. Pomegranate: fast and flexible probabilistic modeling in python. *The Journal of Machine Learning Research*, 18(1):5992–5997, 2017.

[396] Susanne Schreiber, Jean-Marc Fellous, D Whitmer, Paul Tiesinga, and Terrence J Sejnowski. A new correlation-based measure of spike timing reliability. *Neurocomputing*,

52:925–931, 2003.

[397] Alexander Schulz, Fabian Hinder, and Barbara Hammer. Deepview: Visualizing classification boundaries of deep neural networks as scatter plots using discriminative dimensionality reduction. *arXiv preprint arXiv:1909.09154*, 2019.

[398] Antje Schweitzer and Natalie Lewandowski. Convergence of articulation rate in spontaneous speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech 2013, Lyon)*, pages 525–529, 2013.

[399] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64, 1951.

[400] Huitao Shen. Mutual information scaling and expressive power of sequence models. *arXiv preprint arXiv:1905.04271*, 2019.

[401] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[402] Vin D Silva and Joshua B Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in neural information processing systems*, pages 721–728, 2003.

[403] Herbert A Simon. The architecture of complexity. In *Facets of systems science*, pages 457–476. Springer, 1991.

[404] Jonnathan Singh Alvarado, Jack Goffinet, Valerie Michael, William Liberti, Jordan Hatfield, Timothy Gardner, John Pearson, and Richard Mooney. Neural dynamics underlying birdsong practice and performance. *Nature*, pages 1–5, 2021.

[405] JM Sinnott, MD Beecher, DB Moody, and WC Stebbins. Speech sound discrimination by monkeys and humans. *The Journal of the Acoustical Society of America*, 60(3):687–695, 1976.

[406] Jacobo D Sitt, Ezequiel M Arneodo, Franz Goller, and Gabriel B Mindlin. Physiologically driven avian vocal synthesizer. *Physical Review E*, 81(3):031927, 2010.

[407] JD Sitt, A Amador, F Goller, and GB Mindlin. Dynamical origin of spectrally rich vocalizations in birdsong. *Physical Review E*, 78(1):011905, 2008.

[408] G Troy Smith, Eliot A Brenowitz, Michael D Beecher, and John C Wingfield. Seasonal

changes in testosterone, neural attributes of song control nuclei, and song structure in wild songbirds. *Journal of Neuroscience*, 17(15):6001–6010, 1997.

[409] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[410] Panu Somervuo. Time–frequency warping of spectrograms applied to bird sound analyses. *Bioacoustics*, 28(3):257–268, 2019.

[411] Raimund Specht. Avisoft-saslab pro: sound analysis and synthesis laboratory. *Avisoft Bioacoustics, Berlin*, 2002.

[412] H Eugene Stanley, Viktor Afanasyev, Luis A Nunes Amaral, SV Buldyrev, AL Goldberger, Steve Havlin, Harry Leschhorn, P Maass, Rosario N Mantegna, C-K Peng, et al. Anomalous fluctuations in the dynamics of complex systems: from dna and physiology to econophysics. *Physica A: Statistical Mechanics and its Applications*, 224(1-2):302–321, 1996.

[413] Damian G Stephen, Nigel Stepp, James A Dixon, and MT Turvey. Strong anticipation: Sensitivity to long-range correlations in synchronization behavior. *Physica A: Statistical Mechanics and its Applications*, 387(21):5271–5278, 2008.

[414] Stanley Smith Stevens, John Volkmann, and Edwin Broomell Newman. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustical society of america*, 8(3):185–190, 1937.

[415] Dan Stowell, Tereza Petruskova, Martin Salek, and Pavel Linhart. Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. *Journal of the Royal Society Interface*, 16(153):20180940, 2019.

[416] Dan Stowell and Mark D Plumbley. Automatic large-scale classification of bird sounds is strongly improved by unsupervised feature learning. *PeerJ*, 2:e488, 2014.

[417] Michael PH Stumpf and Mason A Porter. Critical truths about power laws. *Science*, 335(6069):665–666, 2012.

[418] Christopher Summerfield and Floris P De Lange. Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, 15(11):745–756, 2014.

[419] Ryuji Suzuki, John R Buck, and Peter L Tyack. Information entropy of humpback whale songs. *The Journal of the Acoustical Society of America*, 119(3):1849–1866, 2006.

[420] Benjamin Szubert, Jennifer E Cole, Claudia Monaco, and Ignat Drozdov. Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific reports*, 9(1):1–10, 2019.

[421] Ryosuke O Tachibana, Naoya Oosugi, and Kazuo Okanoya. Semi-automatic classification of birdsong elements using a linear support vector machine. *PloS one*, 9(3):e92584, 2014.

[422] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th international conference on world wide web*, pages 287–297, 2016.

[423] Ofer Tchernichovski and Partha P Mitra. Sound analysis pro user manual. *CCNY, New York*, 2004.

[424] Ofer Tchernichovski, Fernando Nottebohm, Ching Elizabeth Ho, Bijan Pesaran, and Partha Pratim Mitra. A procedure for an automated measurement of song similarity. *Animal behaviour*, 59(6):1167–1176, 2000.

[425] Audacity Team. Audacity(r): Free audio editor and recorder [computer application]. https://www.audacityteam.org/, 1999-2019. Audacity(R) software is copyright (C) 1999-2019 Audacity Team. The name Audacity(R) is a registered trademark of Dominic Mazzoni.

[426] Carel ten Cate. On the phonetic and syntactic processing abilities of birds: From songs to speech and artificial grammars. *Current Opinion in Neurobiology*, 28:157–164, 2014.

[427] Carel ten Cate and Kazuo Okanoya. Revisiting the syntactic abilities of non-human animals: natural vocalizations and artificial grammar learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1598):1984–1994, 2012.

[428] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[429] Brad Theilman, Krista Perks, and Timothy Q Gentner. Spike train coactivity encodes learned natural stimulus invariances in songbird auditory cortex. *Journal of Neuroscience*, 41(1):73–88, 2021.

[430] Marvin Thielk, Tim Sainburg, Tatyana Sharpee, and Timothy Gentner. Combining biological and artificial approaches to understand perceptual spaces for categorizing natural acoustic signals. In *Conference on Cognitive Computational Neuroscience*, pages 1–4, 2018.

[431] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. *arXiv preprint arXiv:2005.11676*, 2020.

[432] Ana Todorovic, Jan-Mathijs Schoffelen, Freek Van Ede, Eric Maris, and Floris P De Lange. Temporal expectation and attention jointly modulate auditory oscillatory activity in the beta band. *PLoS One*, 10(3):e0120288, 2015.

[433] Ana Todorovic, Freek van Ede, Eric Maris, and Floris P de Lange. Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an meg study. *Journal of Neuroscience*, 31(25):9118–9123, 2011.

[434] Dietmar Todt and Henrike Hultsch. How songbirds deal with large amounts of serial information: retrieval rules suggest a hierarchical song memory. *Biological Cybernetics*, 79(6):487–500, 1998.

[435] Keiichi Tokuda, Takayoshi Yoshimura, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speech parameter generation algorithms for hmm-based speech synthesis. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1315–1318. IEEE, 2000.

[436] Ke Tran, Arianna Bisazza, and Christof Monz. The importance of being recurrent for modeling hierarchical structure. *arXiv preprint arXiv:1803.03585*, 2018.

[437] PL Tyack. Acoustic communication under the sea. In *Animal acoustic communication*, pages 163–220. Springer, 1998.

[438] AM Uchida, RA Meyers, BG Cooper, and F Goller. Fibre architecture and song activation rates of syringeal muscles are not lateralized in the european starling. *Journal of Experimental Biology*, 213(7):1069–1078, 2010.

[439] Michael T Ullman. A neurocognitive perspective on language: The declarative/procedural model. *Nature reviews neuroscience*, 2(10):717–726, 2001.

[440] Laurens Van Der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pages 384–391, 2009.

[441] Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.

[442] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[443] Maarten Van Segbroeck, Allison T Knoll, Pat Levitt, and Shrikanth Narayanan. Mupet—mouse ultrasonic profile extraction: a signal processing tool for rapid and unsupervised analysis of ultrasonic vocalizations. *Neuron*, 94(3):465–485, 2017.

[444] Lena Veit and Andreas Nieder. Abstract rule neurons in the endbrain support intelligent

behaviour in corvid songbirds. *Nature communications*, 4(1):1–11, 2013.

[445] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6-7):889–899, 2006.

[446] Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. The zero resource speech challenge 2015. In *Sixteenth annual conference of the international speech communication association*, 2015.

[447] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080, 2009.

[448] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[449] Gandhimohan M Viswanathan, V Afanasyev, SV Buldyrev, EJ Murphy, PA Prince, and H Eugene Stanley. Lévy flight search patterns of wandering albatrosses. *Nature*, 381(6581):413, 1996.

[450] Kaya von Eugen, Sepideh Tabrik, Onur Güntürkün, and Felix Ströckens. A comparative analysis of the dopaminergic innervation of the executive caudal nidopallium in pigeon, chicken, zebra finch, and carrion crow. *Journal of Comparative Neurology*, 528(17):2929–2955, 2020.

[451] Lysann Wagener, Maria Loconsole, Helen M Ditz, and Andreas Nieder. Neurons in the endbrain of numerically naive crows spontaneously encode visual numerosity. *Current Biology*, 28(7):1090–1094, 2018.

[452] DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[453] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard F Lyon, and Rif A Saurous. Trainable frontend for robust and far-field keyword spotting. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5670–5674. IEEE, 2017.

[454] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

[455] Michael Weiss, Henrike Hultsch, Iris Adam, Constance Scharff, and Silke Kipper. The use of network analysis to study complex animal communication systems: a study on nightin-

gale song. *Proceedings of the Royal Society B: Biological Sciences*, 281(1785):20140460, 2014.

[456] Felix Weninger and Björn Schuller. Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 337–340. IEEE, 2011.

[457] Tom White. Sampling generative networks. *arXiv preprint arXiv:1609.04468*, 2016.

[458] Andrew Whiten, Emma Flynn, Katy Brown, and Tanya Lee. Imitation of hierarchical action structure by young children. *Developmental science*, 9(6):574–582, 2006.

[459] Heather Williams. Choreography of song, dance and beak movements in the zebra finch (taeniopygia guttata). *Journal of Experimental Biology*, 204(20):3497–3506, 2001.

[460] Heather Williams. Birdsong and singing behavior. *ANNALS-NEW YORK ACADEMY OF SCIENCES*, pages 1–30, 2004.

[461] Benjamin Wilson, William D Marslen-Wilson, and Christopher I Petkov. Conserved sequence processing in primate frontal cortex. *Trends in neurosciences*, 40(2):72–82, 2017.

[462] Alexander B Wiltschko, Matthew J Johnson, Giuliano Iurilli, Ralph E Peterson, Jesse M Katon, Stan L Pashkovski, Victoria E Abraira, Ryan P Adams, and Sandeep Robert Datta. Mapping sub-second structure in mouse behavior. *Neuron*, 88(6):1121–1135, 2015.

[463] Alexander B Wiltschko, Tatsuya Tsukahara, Ayman Zeine, Rockwell Anyoha, Winthrop F Gillis, Jeffrey E Markowitz, Ralph E Peterson, Jesse Katon, Matthew J Johnson, and Sandeep Robert Datta. Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature neuroscience*, 23(11):1433–1443, 2020.

[464] Jason Wimmer, Michael Towsey, Birgit Planitz, Paul Roe, and Ian Williamson. Scaling acoustic data analysis through collaboration and automation. In *2010 IEEE Sixth International Conference on e-Science*, pages 308–315. IEEE, 2010.

[465] Jun Xiao and Patrick Flandrin. Multitaper time-frequency reassignment for nonstationary spectrum estimation and chirp enhancement. *IEEE Transactions on Signal Processing*, 55(6):2851–2860, 2007.

[466] Jie Xie, Juan G Colonna, and Jinglan Zhang. Bioacoustic signal denoising: a review. *Artificial Intelligence Review*, pages 1–23, 2020.

[467] Yu Xin, Lin Zhong, Yuan Zhang, Taotao Zhou, Jingwei Pan, and Ning-long Xu. Sensory-

to-category transformation via dynamic reorganization of ensemble structures in mouse auditory cortex. *Neuron*, 103(5):909–921, 2019.

[468] Jae Yung Song, Katherine Demuth, Karen Evans, and Stefanie Shattuck-Hufnagel. Durational cues to fricative codas in 2-year-olds' american english: Voicing and morphemic factors. *The Journal of the Acoustical Society of America*, 133(5):2931–2946, 2013.

[469] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182, 1997.

[470] Zhao Zhao. North american bird species. *Zenodo*, May 2018.

[471] Zhao Zhao, Sai-hua Zhang, Zhi-yong Xu, Kristen Bellisario, Nian-hua Dai, Hichem Omrani, and Bryan C Pijanowski. Automated bird acoustic event detection and robust species classification. *Ecological Informatics*, 39:99–108, 2017.

[472] Willem Zuidema, Robert M French, Raquel G Alhama, Kevin Ellis, Timothy J O'Donnell, Tim Sainburg, and Timothy Q Gentner. Five ways in which computational modeling can help advance cognitive science: Lessons from artificial grammar learning. *Topics in cognitive science*, 12(3):925–941, 2020.