CHARACTERIZING STRUCTURAL VARIATION IN THE NOTES OF ADELAIDE'S WARBLER (*SETOPHAGA ADELAIDAE*) SONGS

SAMANTHA Y. HUANG
Bachelor of Science, Cornell University, 2020

A thesis submitted
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

PSYCHOLOGY

Department of Psychology
University of Lethbridge
LETHBRIDGE, ALBERTA, CANADA

CHARACTERIZING STRUCTURAL VARIATION IN THE NOTES OF ADELAIDE'S WARBLER (*SETOPHAGA ADELAIDAE*) SONGS

SAMANTHA Y. HUANG

Date of Defense: July 24, 2023

Dr. D. M. Logue Thesis Supervisor	Associate Professor	Ph.D.
Dr. J. B. Leca Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. J. Z. Zhang Thesis Examination Committee Member	Associate Professor	Ph.D.
Dr. F. Li Chair, Thesis Examination Committee	Professor	Ph.D.

ABSTRACT

Acoustic structure varies at multiple levels in birdsong. Note-level analyses are beneficial because they can enable scientists to characterize song structures, define song types, and more precisely measure vocal performance. By-eye analysis of spectrograms is sometimes adequate for note classification. However, Adelaide's warbler (*Setophaga adelaidae*) notes are difficult to classify visually because they contain a blend of discrete and continuous variation. We ran a novel, automated, image-based analysis called latent visualization to determine if discrete note types exist and to explore variation among notes. Notes differed primarily in general frequency characteristics, and note clusters (i.e., types) were not discrete. Contrary to our expectations, cluster labels disregarded note contour, which is the most important aspect of note structure for by-eye classification. Overall, latent visualization may not be ideal for species that produce frequency-modulated notes. We recommend running future image-based analyses on note contours alongside acoustic feature analysis for validation.

PREFACE

Contributions of authors

David M. Logue and Samantha Y. Huang designed the study. D.M.L. led data collection and annotation. Peter C. Mower built the song-type exemplar dataset. Brayden L. Carlson developed the workflow pipeline for latent visualization and wrote additional code to clean spectrograms, apply segmentation and clustering, and visualize the latent-space representations and other figures. B.L.C. and S.Y.H. analyzed the data. S.Y.H. wrote all chapters. D.M.L. advised on the data analysis and writing. All copyrighted materials are reprinted under license.

Ethics

This study adheres to the guidelines from the Institutional Animal Care and Use Committee at the University of Puerto Rico, Mayaguez (17 September 2010) and the Animal Welfare Committee at the University of Lethbridge (protocol #1605). It also follows the ASAB/ABS Guidelines for the use of animals in research. Fieldwork was conducted with permission from the U.S. Fish and Wildlife Service (permit 41521-2016-11) and the Departamento de Recursos Naturales y Ambientales (permit 2016-IC-068-1). Bird handling was conducted under D.M.L.'s master bird banding license (no. 2399).

ACKNOWLEDGEMENTS

I thank the University of Lethbridge for providing me the opportunity to conduct this master's research. I also thank my committee members, Dr. David Logue, Dr. Jean-Baptiste Leca, and Dr. Zhang, for their support throughout my degree.

I am utterly grateful to David, my supervisor. David, thank you for being a steadfast mentor and friend who has challenged me in more ways than I thought possible during my time here. You have taught me so much about being a scientist through your research, lab meetings, and candid one-on-ones. Yet I am also grateful for the community you have fostered, coming into lab with casual banter and making sure to check in with every student before you leave. Like many others, you always encourage me to speak. You have taken the time both to understand how I work best and to invest in other areas of my life, like music. I have fond memories of our 2022 field season, and I will always be thankful that we got to share life so closely despite the demands of fieldwork. I am constantly amazed by your level of teaching and leadership, and I look forward to seeing the Birdsong Lab grow.

I thank my fellow graduate students and the professors in the Psychology Department for providing support and walking alongside me throughout my master's research. (Juan, I see you, too.) I also thank my roommates and my friends at Amazing Grace Community Church and in the Lethbridge Symphony Orchestra for helping me settle into this small Canadian city I have called home for the past two years.

I am blessed to have found community in the on-campus InterVarsity Christian Fellowship. Friends, thank you for keeping me sane as I moved forward in my thesis one small step at a time. You always remind me of the bigger picture, and I am still touched that you are all constantly glad to have my company.

To my lab mates: you are a part of my story, Sam, Peter, Juley, Heath, Brayden, Angela, Brigham, Guianne, Jade, and Tosin. My research family picture is incomplete without each of you. I am quietly excited to see the discoveries we are making together about this small tropical bird, the Adelaide's warbler.

I could never have come this far without the immense work of my undergraduate assistant, Brayden. You, sir, shattered expectations long ago and have gone light years beyond any research partner I could have asked for. Yet I am awestruck by your perpetual humility. You are constantly down to earth and prefer to downplay your skills in computer science, instead deferring to my judgment. You are wont to brush off my accolades, so I will just say that I have appreciated your friendship and constant attention to my messages, whether data-pull requests or simply catching up on life.

Peter, Juley, and Heath, you are much more than lab mates to me. You were my first friends here in Lethbridge, and I savor our hangouts, whether cooking, crafting, chilling, or tagging along to anything outside the lab. Thank you for never giving up on me. I find so much joy and comfort with you all.

I would also like to show my appreciation to my mentor, Jennifer. Thank you for your steadfast love and support throughout the ups and downs of my thesis journey. In the depths of emotional turmoil, you remind me how to hold onto my faith. SDG.

Finally, thank you to my family. I thank my parents for letting me call them every day, one small expression of their fierce commitment to supporting me. I thank my grandparents, Ama and Agong, for the vast love they share across an ocean. I love you both, and I am always excited to see you call. Last but not least, I am grateful to my sister, Claudia. Thank you for being there whenever I needed to rant or sulk or just get a virtual pat on the head. I am grateful

we have been able to navigate the hardships of graduate school together, and I look forward to whatever lies ahead for both of us.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AVS	Animal vocalization segmentation
CV	Coefficient of variation
dB	Decibels
DTS	Dynamic threshold segmentation
DTW	Dynamic time warping
FLDA	Fisher's linear discriminant analysis
HDBSCAN	Hierarchical density-based spatial clustering of applications with noise
Hz	Hertz
ISOMAP	Isometric feature mapping
kHz	Kilohertz
k-NN	*k*-nearest neighbor
MDS	Multidimensional scaling
MFCCs	Mel frequency cepstral coefficients
MPS	Modulation power spectrum
ms	Milliseconds
NMI	Normalized mutual information
PAFs	Predefined acoustic features
PCA	Principal component analysis
s	Seconds
SNR	Signal-to-noise ratio
SPCC	Spectrogram cross-correlation
t-SNE	t-distributed stochastic neighborhood embedding
UMAP	Uniform manifold approximation and projection
UPGMA	Unweighted pair-group method of arithmetic averages
USVs	Ultrasonic vocalizations
VAEs	Variational autoencoders
WCSS	Within-cluster sum of squares
2D	Two-dimensional
3D	Three-dimensional

CHAPTER 1: A REVIEW OF ANALYSIS METHODS IN BIOACOUSTICS

Many animals use sound to communicate. Vocalizations are an important class of acoustic signal that animals produce by forcing air through the respiratory system to generate vibrations. The acoustic structure of animal vocalizations varies widely and is closely related to signal function. Birdsong is a well-studied acoustic signal used for mate attraction and territory defense. Quantifying structural variation in birdsong at both the song and note levels is an essential step toward understanding song function. In this review, we walk through the importance of acoustic variation in birdsong, factors to consider when quantifying song and note structure, and methods to quantify acoustic variation.

1.1 ACOUSTIC VARIATION IN BIRDSONG

Acoustic structure varies widely in birdsong. We can examine structural variation at different levels based on the acoustic unit of interest. Acoustic units are any unit of sound and include, from smallest to largest, notes, syllables, songs, and song bouts. The definitions of these units vary among studies (Odom et al., 2021). For the sake of this review, we define a note as a continuous trace on a spectrogram and a song as a series of notes. A syllable is a group of notes within a song that is repeated in the bird's repertoire with the same consecutive series of notes. Finally, a song bout is a series of songs. Within these acoustic units, structural variation can range from discrete to continuous. For example, song sparrows (*Melospiza melodia*) sing discrete song types, but black-capped chickadees (*Poecile atricapillus*) sing continuous song types (Christie et al., 2004; Podos et al., 1992).

Structural variation comes in part from song learning in birds and is important because it reveals clues about evolutionary history, affects receiver responses, and is bound by physiological constraints (Ballentine et al., 2003; dos Santos et al., 2018; DuBois et al., 2011;

Geberzahn & Aubin, 2014; Lachlan et al., 2013; Leitão et al., 2006; Martins, 1996; Stoddard et al., 1988). For instance, analyses at the note level can enable scientists to characterize song structures, define song types, and more precisely measure vocal performance (Geberzahn & Aubin, 2014; Leitão et al., 2006; Stoddard et al., 1988). Navigating analyses of acoustic variation, however, can be difficult.

1.2 WORKFLOW PIPELINE FOR ANALYZING ACOUSTIC VARIATION

Recently, roadmaps have come out for quantifying acoustic variation in animal vocalizations (Kershenbaum et al., 2016; Odom et al., 2021). They lay out a workflow that includes data collection, unit identification, measurement extraction, data reduction (if necessary), and classification. In short, these workflows measure acoustic variation in one or more dimensions and then classify acoustic units into groups. Each step aims to address challenges inherent to the measurement of acoustic variation, including the multidimensional nature of sound, a lack of standard terminology regarding acoustic units and variation, and navigating the vast array of analysis techniques available. These techniques range from manual to automated and can be applied to both feature- and image-based analyses. Accordingly, I describe these dichotomies below.

1.2.1 MANUAL V. AUTOMATED METHODS

Manual and automated methods are available for analyzing acoustic variation in animal vocalizations. Manual methods involve the collection and analysis of acoustic measurements by humans, and automated methods rely on machines for measurement collection and analysis. Both types of methods have pros and cons. Scientists can identify, compare, and classify notes by eye using spectrograms, but doing so is time-consuming and requires expert knowledge. Manual analyses can introduce human bias in how notes or song types are defined. For example,

annotators may disagree when labeling note types. In addition, manual analyses are limited by the size of the dataset that scientists can analyze and the number of acoustic features they can measure. Earlier studies in bioacoustics have mourned technological limits to understanding song structure (Clark et al., 1987; Podos et al., 1992).

Automated analyses are quicker than manual analyses and can handle larger datasets but have their own shortcomings. Many automated analyses are still novel, requiring external knowledge in computer science. They also lack the accuracy of human experts. For example, automated analyses are more consistent and repeatable than manual analyses but may struggle to distinguish focal sounds from unwanted acoustic material (Wadewitz et al., 2015). Automated approaches have also revealed new knowledge, like a more detailed description of song types or finer-scale acoustic structure (Clark et al., 1987; Goffinet et al., 2021; Keen et al., 2021; Wadewitz et al., 2015). Manual and automated methods can be applied to different types of analyses based on the research question of interest.

1.2.2 FEATURE- V. IMAGE-BASED ANALYSIS

Scientists can run either a feature- or image-based analysis based on their research question. Feature-based analysis uses measured acoustic features (e.g., frequency, duration, amplitude) to compare acoustic signals. Odom et al. (2021) lay out a recommended framework for feature-based analysis. First, the research question determines the acoustic unit(s) to compare. Then, scientists choose acoustic metrics that apply to that acoustic unit. Finally, the chosen metrics will suggest which type of feature-based analysis to run. Acoustic features are useful for describing how groups differ. However, analyses tend to take a long time, and scientists have to choose which features are important. Therefore, acoustic features may not always be the best metric to use when quantifying acoustic variation.

Image-based analysis is an alternative approach to quantifying acoustic variation that does not involve the collection of acoustic features. Instead, it quantifies variation in visual images of animal vocalizations (e.g., spectrograms). Image-based analysis can pick up on differences that scientists may not have considered and tends to be fast once it is set up. The setup, however, can take time, and analysis ultimately does not reveal how groups differ, requiring scientists to run additional analyses. Furthermore, image-based analysis can generate unanticipated artifacts by, for example, sorting images based on noise rather than the signal. Finally, scientists cannot choose or weight acoustic variables in this type of analysis. In short, acoustic feature and image-based analyses both have advantages and disadvantages (Clark et al., 1987; Sainburg et al., 2020). Below, I walk through manual and automated methods for acoustic feature analysis and image-based analysis, beginning with the earliest forms of data collection.

1.2.3 DATA COLLECTION AND UNIT IDENTIFICATION

Modern bioacousticians use acoustic recorders to collect animal vocalizations. In contrast, the earliest analysis of animal vocalizations came in the form of dictation. Scientists would listen to an animal vocalize and then write down what they heard. For example, Saunders (1915) began with musical notation but found he was limited by the musical staff's fixed half-step pitches and metered time along with other mechanical rules. Saunders attempted to address these limitations by developing a novel graphic method to display a song's pitch, duration, and intensity (Figure 1.1). Despite these advancements, the spontaneity of handwritten dictation inevitably leads to imprecision.

1.2.3.1 SPECTROGRAMS

Collection of animal vocalizations moved beyond handwritten dictation with the invention of recording technology and the development of the sound spectrogram. The earliest

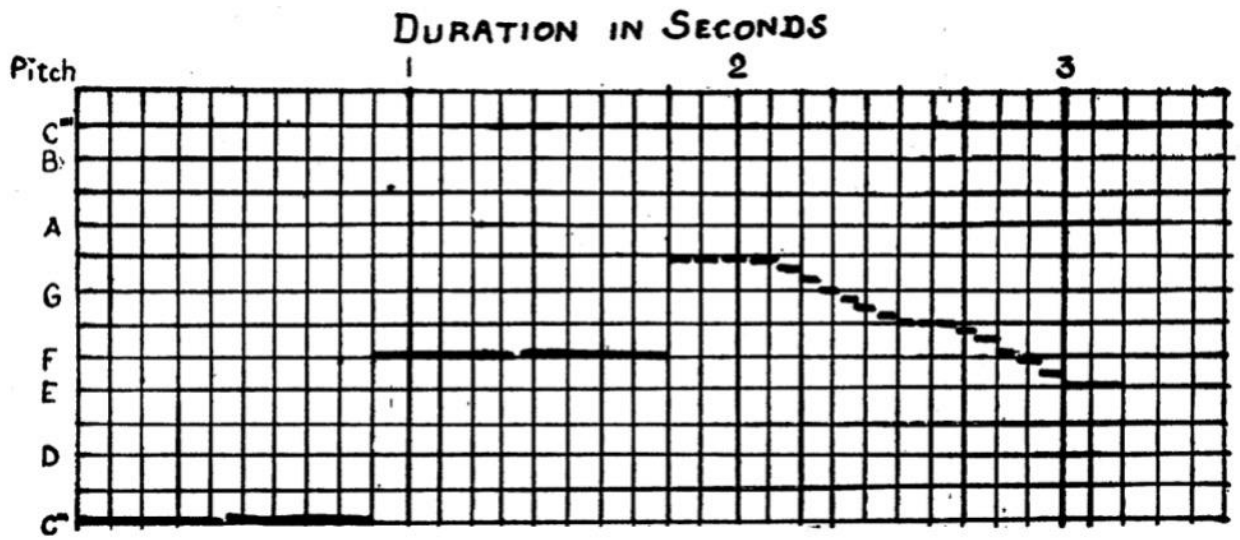


Figure 1.1. A song of the vesper sparrow (*Pooecetes gramineus*), dictated by Saunders using the graphic method (1915). Reprint licensed by the Copyright Clearance Center on behalf of Oxford University Press - Journals (license ID 1371347-1).

spectrograms were analog sonograms, physically generated on paper by spectrographs (Pieplow, 2009). This generation was based on Fourier's theorem and was later digitized in the form of the Fourier transform, which converts soundwaves to a sound spectrum (Amador & Mindlin, 2023). Today, spectrograms are a type of acoustic feature space built by running fast Fourier transforms (FFTs) on the sound pressure waveform. This type of space represents sound in a way that is easier for humans to understand than the sound pressure waveform. Similar to Saunder's graphic method, which measures pitch on the y-axis and duration on the x-axis, spectrograms measure time on the x-axis and frequency on the y-axis. Spectrograms represent intensity (amplitude) with color or shade. Spectrograms allow for more accurate and repeatable measurements of pitch (called "frequency" from here on), duration, intensity (called "amplitude" from here on), and other acoustic features than handwritten dictation (Figure 1.2). Spectrograms have enabled scientists to identify acoustic units within songs like notes and syllables, count the number of notes or syllables in each song, and measure durations, frequencies, and more.

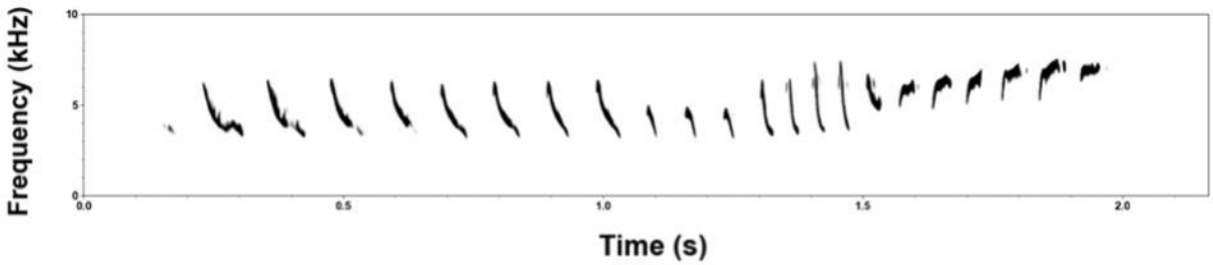


Figure 1.2. Spectrogram of an Adelaide's warbler (*Setophaga adelaidae*) song. Frequency is measured on the y-axis in kilohertz, and time is measured on the x-axis in seconds. Amplitude is measured in the form of darkness (e.g., louder sounds are darker on the spectrogram).

Scientists have used spectrograms to visually identify acoustic units in the songs of many species, including blue grosbeaks (*Guiraca caerulea*), southern house wrens (*Troglodytes aedon chilensis*), willow warblers (*Phylloscopus trochilus*), serins (*Serinus serinus*), and song sparrows (Ballentine et al., 2003; dos Santos et al., 2018; Gil & Slater, 2000; Mota & Cardoso, 2001; Podos et al., 1992). The units we discuss in this review are syllables and notes because we are focused on the building blocks of songs.

Once scientists identify these notes or syllables, they can build repertoires at the note, syllable, or song level. For example, scientists may find that one blue grosbeak sings ten types of notes, whereas another may only sing eight types (Ballentine et al., 2003). Scientists can also identify song types by listing the order in which notes are sung. Each song type has a different order of notes and can be used to build individual song repertoires. In some cases, technology can be used to separate the notes within a song for further analysis (dos Santos et al., 2016; Sainburg et al., 2020).

1.2.3.2 SPECTROGRAM SEGMENTATION

Recently developed methods can automatically separate notes through a process called dynamic threshold segmentation (DTS). Segmentation looks for silent gaps between notes. Scientists define silence by setting a threshold based on background noise, taking into consideration the expected lengths of a note and the period of silence between notes (Figure 1.3;

Sainburg et al., 2020). Segmentation is useful because it enables scientists to save each note as a separate file and run subsequent analyses on individual notes. In other words, segmentation is useful for running automated classifier analyses at the note level. Before classification, scientists need to extract measurements (if running an acoustic feature analysis), which can occur both in spectrograms and other acoustic feature spaces.

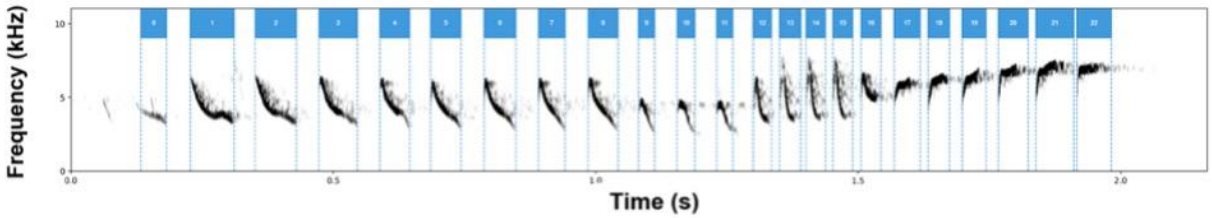


Figure 1.3. DTS applied to an Adelaide's warbler song.

1.2.4 MEASUREMENT EXTRACTION

1.2.4.1 ACOUSTIC FEATURES

Like spectrograms, other acoustic feature spaces perform nonlinear operations on the sound pressure waveform to aid in the quantification of sound. Elie & Theunissen (2016) tested an alternative acoustic feature space to spectrograms that included the fundamental and temporal and spectral envelopes of zebra finch (*Taeniopygia guttata*) vocalizations. They extracted acoustic variables called "predefined acoustic features (PAFs)," which are properties that humans can easily perceive (e.g., mean time, maximum amplitude, fundamental frequency). PAFs are useful for gaining an intuitive understanding of a song's acoustic structure but are limited in number, meaning they may not cover all the variation present in a song's acoustic structure.

Other acoustic feature spaces include the modulation power spectrum (MPS), which focuses on time and amplitude, and Mel frequency cepstral coefficients (MFCCs), which focus on formants used in speech (Elie & Theunissen, 2016). The MPS enables scientists to study spectro-temporal modulation, which has proven useful in discriminating natural sounds

(Fukushima et al., 2015; Singh & Theunissen, 2003; Woolley et al., 2005). MFCCs are useful in studying the vocalizations of mammals, who perceive frequencies along a roughly logarithmic scale (Prat et al., 2016; Reby et al., 2006).

Once scientists choose an acoustic feature space, they can extract feature measurements either by eye using a cursor on the screen or automatically by telling the computer where the signal is (e.g., in the time-frequency space of a spectrogram) and letting it extract features. After acoustic features have been collected, scientists can measure the similarity among these features.

1.2.4.2 SIMILARITY MEASURES

Scientists can use similarity measures to better define song variants and song types or to compare song structure among individuals. Similarity measures are one way for scientists to obtain distance measures, which they can use to array sounds in acoustic space. Examples of similarity measures include the coefficient of variation (CV) and the Levenshtein distance. The CV measures global variation, or variation among all songs in a group, by dividing the standard deviation (e.g., of the number of different notes produced in a given song) by the mean. In contrast, the Levenshtein distance measures pairwise similarity, a form of local variation that compares two songs at a time. The Levenshtein distance is a linguistic metric that counts the number of "edits" required to change the order of notes in one song to the order of notes in the next successive song (dos Santos et al., 2018; Gil & Slater, 2000). In other words, the Levenshtein distance uses note order to measure differences between songs.

Like the Levenshtein distance, Jaccard's coefficient of correlation also measures pairwise similarity between songs. The coefficient is a function of the number of notes present in both songs and the number of notes unique to each song (Baulieu, 1989; Sneath & Sokal, 1973). In song sparrows, Podos et al. (1992) used Jaccard's coefficient to calculate pairwise similarities

between song variants, which enabled them to group similar song variants into song types. All these measures can quantify similarity but only apply when the acoustic unit of interest (e.g., song) has subunits that scientists can easily differentiate. Therefore, these measures can never be used to quantify similarity among notes because notes do not contain subunits. In the same way, similarity measures can only sometimes be applied to syllables because syllables do not always contain discrete notes. When differences between notes are more difficult to discern, image-based similarity measures may be helpful.

Similarity measures also exist in image-based analysis. Clark et al. (1987, p. 103) introduced a "sound comparative method," now called spectrogram cross-correlation (SPCC). SPCC overlays the frequency-time matrices of two sounds and determines their similarity by calculating the peak value of the resulting correlation function. Dynamic time warping (DTW) calculates similarity between sounds in the same manner as SPCC but allows the expansion or compression of time to maximize similarity (Keen et al., 2014; Lachlan et al., 2013). DTW may therefore be more useful due to its time flexibility. For example, DTW allows for notes with similar contours (i.e., shapes) but different durations to be scored as similar. More advanced forms of visual analysis take sounds beyond the acoustic feature space into latent feature space.

1.2.5 DIMENSIONALITY REDUCTION

Dimensionality reduction is a form of data compression in which "high-dimensional data [is compressed] into a smaller number of dimensions while retaining the structure and variance present in the original…data" (Sainburg et al., 2020, p. 2). When multiple parameters are extracted from an acoustic feature space, dimensionality reduction can help by reducing the number of components in play for a more intuitive understanding of acoustic variation.

Dimensionality reduction can be applied to both feature and visual analyses, compressing either a set of features or a similarity matrix.

1.2.5.1 PCA

Principal component analysis (PCA) is a feature-based analysis that uses dimensionality reduction to look for the factors that best summarize trends in the data, called principal components. In the case of bioacoustics, PCA combines acoustic features into principal components that serve as axes of variation. It is useful for data in which multiple acoustic features are correlated. The output of a PCA is a set of synthetic variables called "principal components." Each principal component is associated with a set of weightings for the univariate acoustic features. Researchers can interpret these weightings in meaningful ways. For example, in Table 1.1, *different syllable types per song* and *rate of syllable type production* are the only variables with high weightings for PC1, so we might conclude that component characterizes short-term variation within songs. The first principal component accounts for the greatest amount of variance in the data, followed by the second principal component, and so on (Frey & Pimentel, 1978; Gil & Slater, 2000). Because each principal component is orthogonal to all others, principal components are statistically independent of one another.

Table 1.1. Example weightings of variables in three principal components found in a hypothetical dataset of southern house wren songs. High weightings are italicized. Based on Table 1 in dos Santos et al., (2018) and Table 2 in Gil & Slater (2000).

Variables	PC1	PC2	PC3
Different syllable types per song	*0.73*	0.22	0.18
Rate of syllable type production	*0.81*	0.37	0.10
Coefficient of variation in number of different syllable types per song	0.23	*0.79*	0.21
Coefficient of variation in rate of syllable type production	0.15	*0.72*	0.26
Levenshtein distance	0.02	0.11	*0.77*
Song repertoire	0.01	0.09	0.07

PCA is useful for reducing over-fitting and eliminating intercorrelation among acoustic features (Crouch & Mason-Gamer, 2019; Hedwig et al., 2014). PCA is also a useful application for visual cluster analysis. (We will touch more on clustering later.) Other forms of dimensionality reduction use image-based analysis to look for acoustic features less commonly included in PCA.

1.2.5.2 VAEs

Variational autoencoders (VAEs) are a type of image-based analysis that uses neural networks to build a latent representation of the data. This representation may find acoustic features that better define acoustic variation than PCA because latent features discovered by VAEs have a higher representational capacity than principal components (i.e., span a higher-dimensional space). VAEs learn a probabilistic map between vocalizations and a latent feature space through an encoder and decoder. VAEs first compress the sound's spectrogram into a vector of latent dimensions, where "latent" means existing but not yet made manifest. The VAE then decodes the latent representation of the data to reconstruct the sound spectrogram as a way of ensuring that information has been preserved as much as possible. The encoded latent representations can then undergo dimensionality reduction to visualize variation in the sound and infer features that contribute to it (Figure 1.4).

Goffinet et al. (2021) used VAEs to analyze ultrasonic vocalizations (USVs) of laboratory mice and zebra finches. They found that the VAE's learned acoustic features outperform handpicked acoustic features because the VAEs carry unique information, have higher representational capacity, are more sensitive in comparison problems, and can investigate the diversity and stability of syllable repertoires. This unique approach offers a form of validation through the decoder but may still require additional analysis to interpret the latent

dimensions. More recently developed forms of dimensionality reduction also use latent dimensions to visualize acoustic variation.

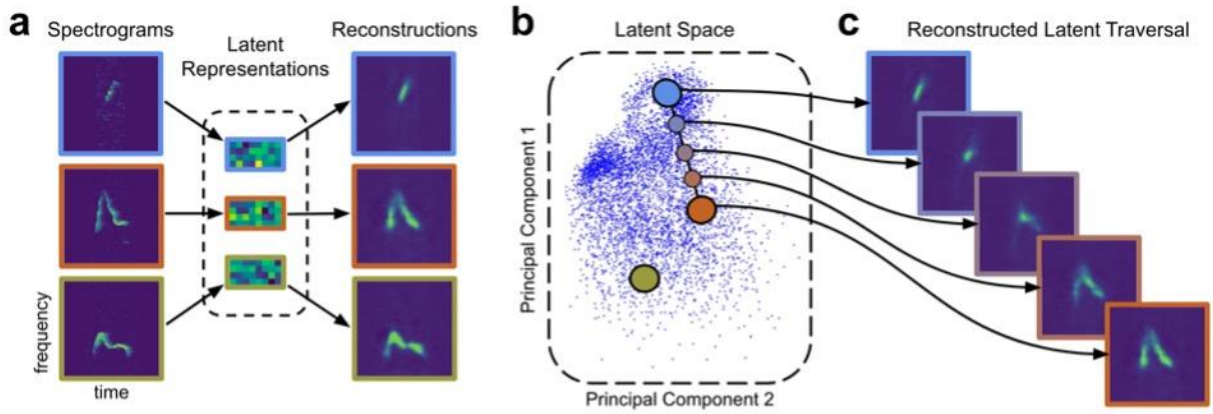


Figure 1.4. A visual workflow pipeline for VAEs run on laboratory mouse USVs. a. The VAE encodes spectrograms into probabilistic maps of latent representations and decodes them by building spectrographic reconstructions. b. The latent representations can then be visualized. c. Any point in the visualization can undergo spectrographic reconstruction via the decoder to infer acoustic features that contribute to acoustic variation (Goffinet et al., 2021). Reprinted under the Creative Commons Attribution 4.0 International Public License.

1.2.5.3 MDS AND ISOMAP

Dimensionality reduction techniques to project animal vocalizations into a latent feature space have developed rapidly within the last few years. PCA projects data onto a lower-dimensional surface that maximizes the variance of the projected data. Multidimensional scaling (MDS) builds off of PCA but plots data based on the distance between individual points instead of their correlation and aims to preserve pairwise Euclidean distance as much as possible. The Euclidean distance is the length of a line segment drawn between two points in a smooth, *n*-dimensional plane (Borchia, 2023). MDS is the precursor to complete isometric feature mapping (ISOMAP; Sainburg et al., 2020). ISOMAP was one of the first topological non-linear dimensionality reduction algorithms. It attempts to represent data graphically and then uses MDS in graphical, instead of Euclidean, space (Tenenbaum et al., 2000). Graph-based methods are

advantageous for preserving local data structure but disadvantageous because ISOMAP dimensions have no meaning.

1.2.5.4 t-SNE AND UMAP

Both t-distributed stochastic neighborhood embedding (t-SNE) and uniform manifold approximation and projection (UMAP) are built off of ISOMAP. These two methods maintain topological structure but also include probabilistic weighting in the graph before embedding the graph in a low-dimensional embedding space. This weighting helps to preserve local structure. UMAP does not preserve global distance, meaning that the algorithm assumes that the high-dimensional data space is warped. Sainburg et al. (2020) and Thomas et al. (2022) chose UMAP over t-SNE in their analyses due to the former's lower computational time and its preservation of local structure. Latent-space representations can be useful for analysis of animal vocal repertoires in which variation is difficult to quantify. These dimensionality reduction algorithms include a higher number of acoustic features in their analyses and create an intuitive two-dimensional (2D) visualization of acoustic variation. However, the latent dimensions of variation they produce can be difficult to interpret. Furthermore, these algorithms can be difficult to implement without expertise in computer science. Classifier analyses may therefore be helpful in interpreting these latent dimensions.

1.2.6 CLASSIFIER ANALYSES

Once the relevant measurements have been extracted from the chosen acoustic feature or visual space and reduced in dimensionality, classifiers can organize acoustic units into classes, or groups. There are two types of classifiers: supervised and unsupervised. Supervised classifiers use labeled data to predict the labels of unclassified data and iteratively adjust for the correct answer. Unsupervised classifiers reveal structure in unlabeled data.

1.2.6.1 SUPERVISED CLASSIFIERS

k-nearest neighbor (*k*-NN) algorithm and random forest

In supervised classification, humans train the classifier to organize data into the correct classes with a manually labeled dataset. In other words, humans supervise the classifier's learning. The *k*-nearest neighbor (*k*-NN) algorithm, Fisher's linear discriminant analysis (FLDA), and random forests are examples of supervised classifiers (Elie & Theunissen, 2016; Vaca-Castaño & Rodriguez, 2010). The *k*-NN algorithm uses Euclidean distance to classify data. The *k*-NN algorithm uses Euclidean distance to find a number *k* of the nearest neighboring points, called neighbors, to a data point. The nearest neighbors are taken from a training dataset that has already been classified. The data point is then assigned to the class to which most of its nearest neighbors belong (Figure 1.5). When assigning classes, the *k*-NN algorithm makes two assumptions. First, it assumes that similar data points are found close to each other. Second, it assumes that the classification in the training dataset is based on the data's features (Bhardwaj, 2023; Dietrich et al., 2004; Khatib, 2016). Other supervised classifiers, like random forests, take features of the data into greater consideration when assigning their classes.

The random forest algorithm aims to split the data into classes based on their features. The algorithm generates a set of decision trees in which each tree has a node that seeks to further split the data into random subsets (called feature randomness). Each new decision tree samples the original dataset with replacement, a process called "bagging," which means that the data subsets in different decision trees may overlap (IBM Cloud Education, 2020; Keen et al., 2014). By combining bagging and feature randomness, the resulting random forest of decision trees remains uncorrelated, which avoids over-fitting and allows for better exploration of potential

splits in the feature space (Breiman, 2001). We focus on supervised random forest analysis; however, random forests can also be applied in an unsupervised manner (Keen et al., 2021).

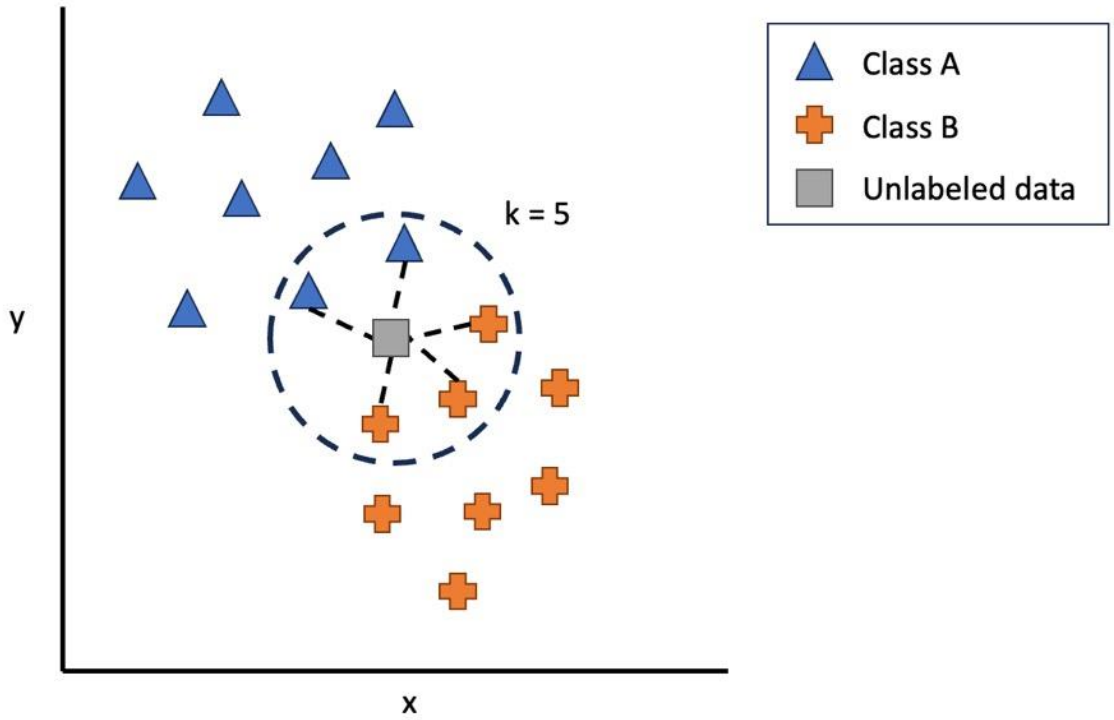


Figure 1.5. Visual workflow of the *k*-nearest neighbor algorithm. The algorithm finds the nearest neighbors of an unlabeled data point and assigns it to the class to which the majority of its nearest neighbors belong. In this case, the new example will be labeled as "Class B" because the majority of the example's surrounding neighbors (k = 5) are classified as "Class B." Based on Bhardwaj (2023).

To summarize, the random forest algorithm builds multiple decision trees and combines all their outputs to reach one final result (Figure 1.6). The high number of decision trees makes the model robust to outliers and noise, reduces the risk of overfitting the model, and enables scientists to test the contribution of different acoustic features to the model's predictive accuracy (Breiman, 2001; IBM Cloud Education, 2020). This classifier may therefore be useful for scientists who wish to compare the utility of different input variables (e.g., acoustic features) for predicting category membership (e.g., note types). However, a forest of decision trees is more difficult to interpret than a single decision tree. Furthermore, building the forest takes time and demands high computing power.

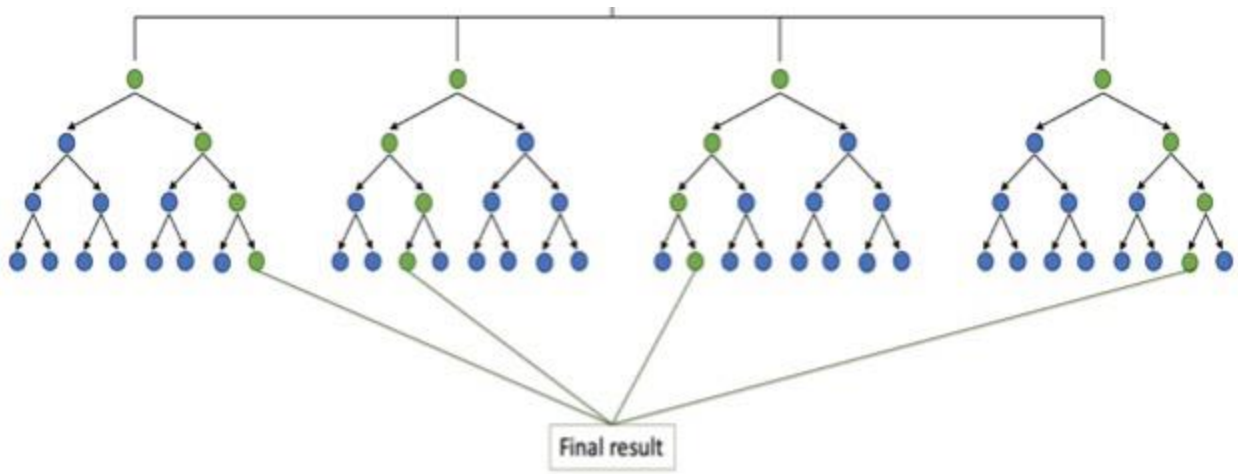


Figure 1.6. A visual of the random forest algorithm. The algorithm builds multiple unique decision trees and combines their output to reach one final result. Based on IBM Cloud Education (2020).

FLDA

Fisher's linear discriminant analysis (FLDA) aims to maximize the separability among known classes. In technical terms, FLDA classifies continuous data into discrete groups. FLDA projects data onto vectors that simultaneously maximize the distances among the means of each class and minimize the within-class variance, or scatter, of the classes (Starmer, 2016; Unzueta, 2021). FLDA reduces the dimensionality of data by comparing the distance among means in a hyperplane. In the case of FLDA, a hyperplane is a subspace that has one less dimension than the number of known classes (Yang, 2020). FLDA can sometimes outperform PCA when the number of input variables is much greater than the number of classes. This classifier could therefore be useful to scientists studying acoustic signals that have fewer categories than the number of acoustic features measured. This classifier is also advantageous over the random forest algorithm because scientists can easily identify which acoustic features to use to discriminate between classes (Elie & Theunissen, 2016). Unsupervised classifier analyses are like an unsupervised version of FLDA.

1.2.6.2 UNSUPERVISED CLASSIFIERS

Unsupervised classifiers, also known as clustering algorithms, assign unlabeled data to clusters, or groups. They are useful when the number or type of groups is unknown. Clustering algorithms can be either “hard” or “soft.” Hard clustering algorithms associate each data point with only one cluster, whereas soft algorithms estimate the probability that a data point is associated with each of several clusters (Carrasco, 2020; Wadewitz et al., 2015).

UPGMA

The unweighted pair-group method of arithmetic averages (UPGMA) is a hard clustering algorithm that uses agglomerative (i.e., bottom-up) hierarchical clustering to construct dendrograms, which are tree diagrams that show relationships between groups (Gil & Slater, 2000; Podos et al., 1992). UPGMA builds a dendrogram by calculating the pairwise distance among clusters and combining the two clusters with the lowest pairwise distance into a higher-level cluster. Pairwise distance comes from averaging the distances between every data point in one cluster and every data point in another cluster (i.e., average linkage clustering). After combining the two nearest clusters, the algorithm then calculates the new pairwise distances and combines the next two nearest clusters. These steps repeat until one cluster remains and the dendrogram is complete (Goldwasser, 2019; Sharma, 2019). Moat indices, which measure how isolated clusters are from each other, can then be applied to find the best level of clustering (Wirth et al., 1966). Since the advent of UPGMA, clustering algorithms have moved beyond numerical classification to capture the inherent structure of datasets in more sensitive ways.

Ward's clustering

Ward's clustering is another agglomerative hierarchical clustering algorithm that uses a different method to measure the distance between clusters. It assumes that the greatest amount of information about a given dataset is present when the data are ungrouped, such that each data point comprises its own group. The algorithm first links one data point with its nearest neighbor to minimize the amount of information lost (called the error sum of squares) and reduce the number of groups by one. The algorithm then considers whether to link a third data point with the first pair or to create a new pairing to continue minimizing the error sum of squares. This "hierarchical grouping" can continue until all the data points are in one group (Figure 1.7; Wadewitz et al., 2015; Ward, 1963). Both UPGMA and Ward's clustering use hierarchical grouping, but not all clustering algorithms use a hierarchical approach to group data.

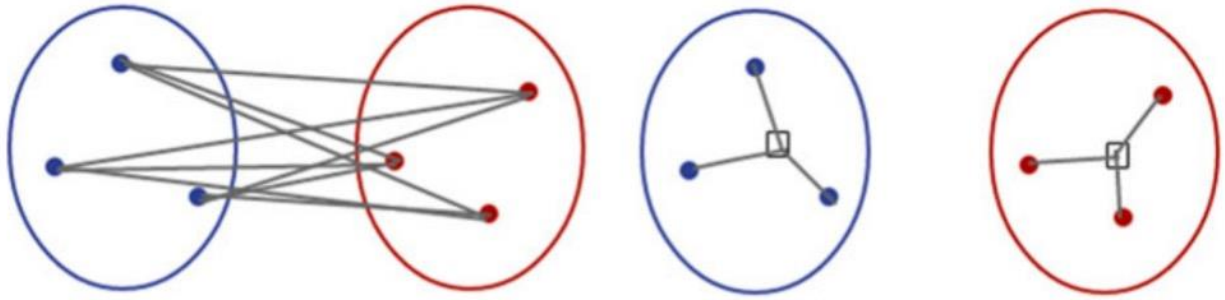


Figure 1.7. An example of the average linkage method from UPGMA (left) versus the method from Ward's clustering (right) for measuring distance between clusters. The average linkage method averages the pairwise distances between the data points of two clusters. In contrast, the method from Ward's clustering determines clusters by minimizing the error sum of squares (i.e., within-cluster variance; Reutterer & Dan, 2022). Reprint licensed by the Copyright Clearance Center on behalf of Springer Nature (license ID 5580380623842).

***k*-means clustering**

K-means clustering aims to group data by finding a fixed number of clusters and optimizing their centroids (i.e., means; Education Ecosystem (LEDU), 2018). This algorithm enables the user to choose an arbitrary number *k* of cluster centroids, then assigns each data point to the cluster with the nearest mean, minimizing the within-cluster sum of squares (WCSS). The

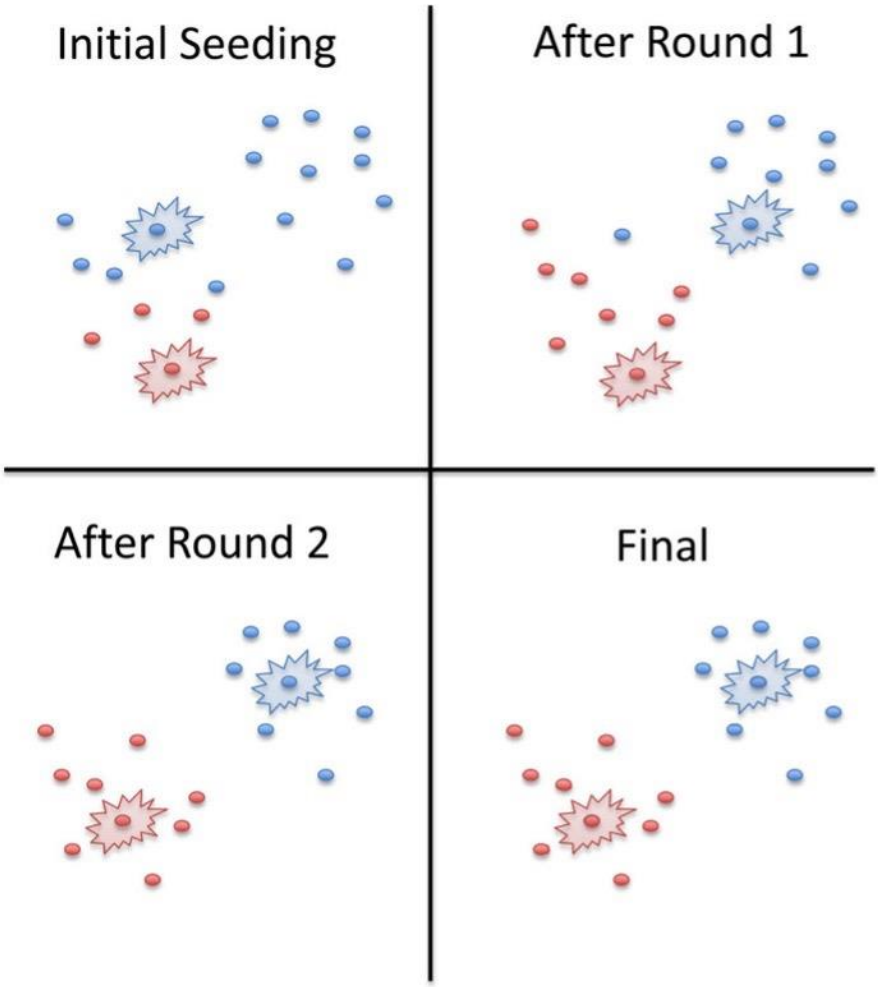


Figure 1.8. A workflow of *k*-means clustering iteratively optimizing cluster centroids. In the "Initial Seeding" panel, two points are randomly chosen as seeds (i.e., centroids) and are represented by starburst outlines. Each point is assigned the label (represented by color) of its closest centroid. New centroids are then chosen based on the average of all points in each cluster. In this example, after two rounds, clusters have reached a steady state (Page et al., 2014). Reprinted under the Creative Commons Attribution 4.0 International Public License.

mean of the cluster is updated to include the new data point, and then the process repeats until the WCSS can no longer be improved (Figure 1.8; MacQueen, 1967; Wadewitz et al., 2015).

The constant reassignment of data points can be advantageous, but finding the optimal number of clusters *k* can be difficult. Silhouette values use Euclidean distance to measure the tightness of data points within a given cluster and the degree of separation between different clusters. Silhouette values range from -1 to 1 such that more positive values indicate stronger clustering (Kaluthota et al., 2019; Wadewitz et al., 2015). Users can vary the *k*-value to look for

a peak silhouette value. A shortcoming, however, of *k*-means clustering is its inability to group data that has irregularly shaped clusters. In contrast, Ward's clustering is more robust to outliers and noise due to its linkage method, but its lack of iterative optimization may cause mistakes in its agglomerative approach.

When scientists use multiple clustering algorithms on their datasets, normalized mutual information (NMI) can measure how well the results of two clustering algorithms match. NMI ranges from 0 to 1, where higher values indicate a better match (Wadewitz et al., 2015). Other recently developed clustering algorithms build on Ward's and *k*-means clustering.

HDBSCAN

Hierarchical density-based spatial clustering of applications with noise (HDBSCAN) is a novel hard clustering algorithm that adds a density factor to hierarchical grouping. HDBSCAN estimates the density of data points, uses a global threshold to identify possible clusters based on the density, and determines final clusters hierarchically. Unlike *k*-means clustering, HDBSCAN can find clusters even if the data include noise and the clusters are irregularly shaped (Figure 1.9; Berba, 2020). However, acoustic units do not always fall into discrete types. When this is the case, soft clustering algorithms may be useful in classifying data.

Mixture-of-Gaussians and fuzzy *c*-means clustering

In contrast to hard clustering algorithms, which simply assign data points to clusters, soft clustering algorithms estimate the probability that a data point is associated with a given cluster. The mixture-of-Gaussians algorithm assumes that clusters come from different Gaussian distributions and tests the likelihood of a data point coming from a given distribution (Carrasco, 2020). Another soft clustering algorithm, fuzzy *c*-means clustering, determines the optimal

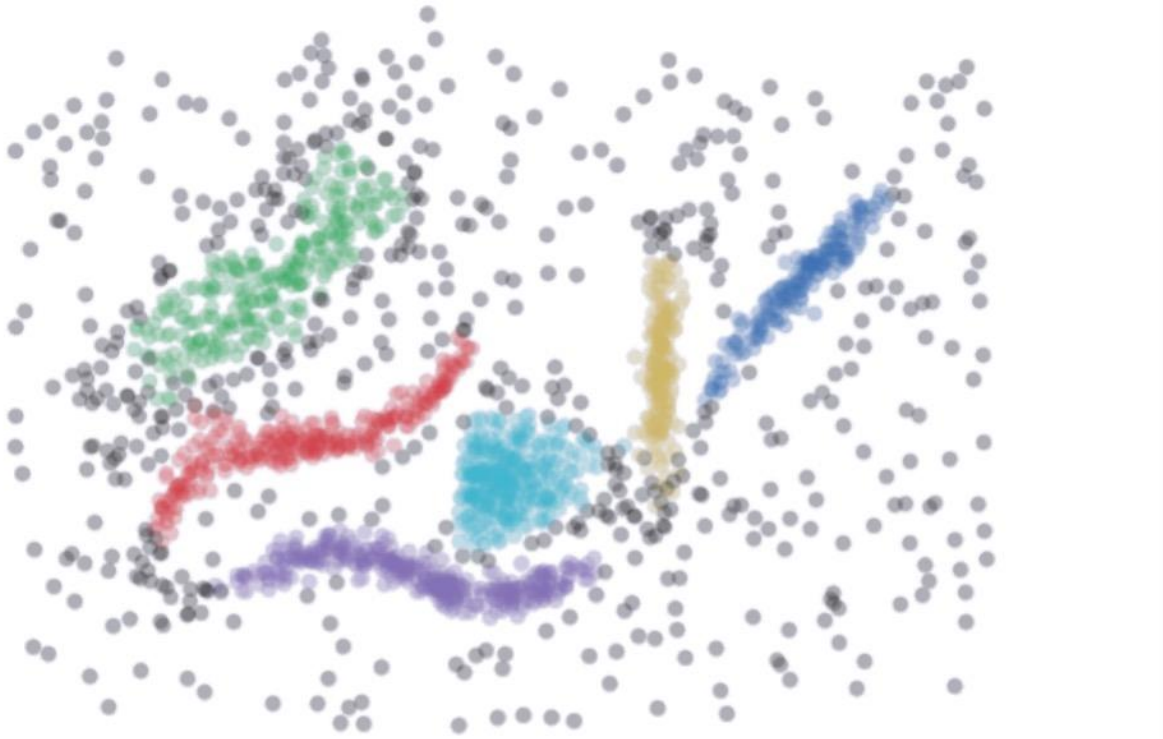


Figure 1.9. An example of HDBSCAN. Here, the algorithm found six clusters. Data points in black have been classified as noise (McInnes et al., 2017). Reprinted under the Creative Commons Attribution 4.0 International Public License.

number *c* of clusters for a given dataset based on the maximum number of clusters allowed and the fuzziness parameter μ, where $\mu = 1$ indicates clusters as crisp as those built by *k*-means clustering, and higher values of μ indicate increasingly fuzzy clusters. The algorithm assigns each data point a membership value *m* for each of the clusters, where *m* represents the probability that a data point belongs to a given cluster. The value *m* ranges from 0 to 1, where $m = 0$ means the data point has no association with the cluster, and $m = 1$ means the data point fully belongs to the cluster. Like *k*-means clustering, fuzzy *c*-means clustering arbitrarily assigns a number *c* of cluster centroids and then optimizes those centroids by taking weighted membership values into consideration over multiple iterations and adjusting accordingly (Figure 1.10; Wadewitz et al., 2015). Soft clustering may have unique advantages in analyzing acoustic variation due to its flexibility with membership values.

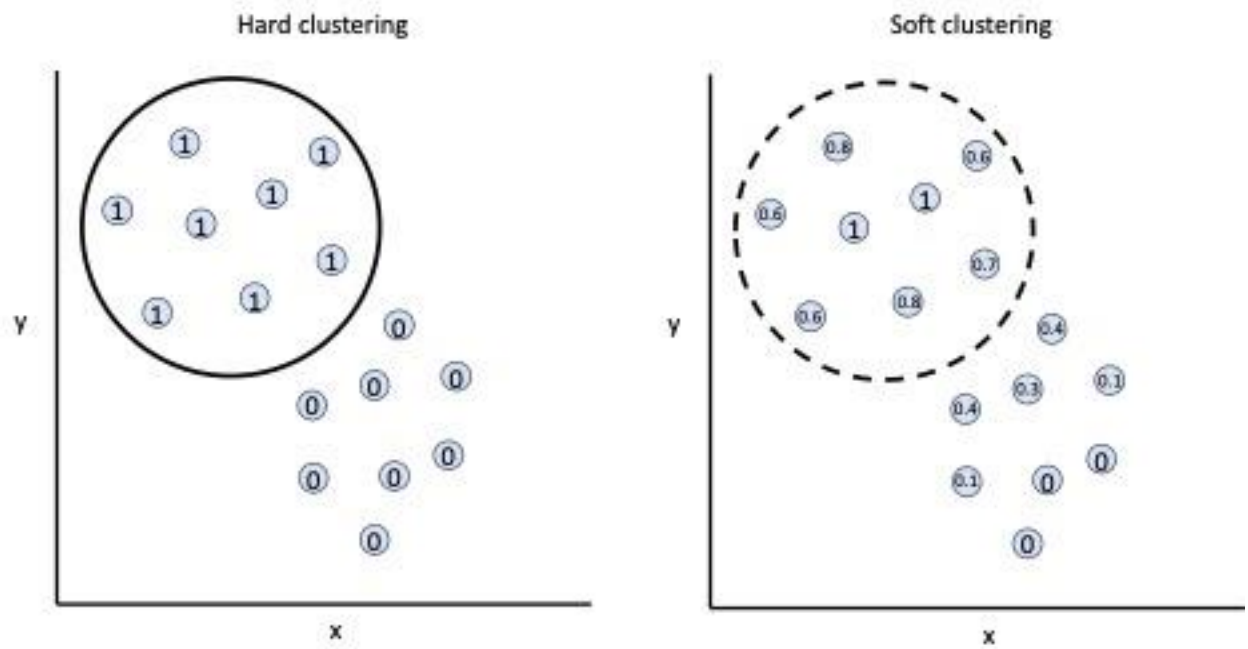


Figure 1.10. A comparison between hard and soft clustering. Membership values for each data point are in black. Membership values for hard clustering are only 0 or 1, but membership values for soft clustering vary between 0 and 1. Based on Yufeng (2021).

Wadewitz et al. (2015) tested Ward's, *k*-means, and fuzzy *c*-means clustering on baboon calls with four datasets, each containing a different number of acoustic features. They found that datasets with a larger number of acoustic features produce better clustering results. Furthermore, fuzzy *c*-means clustering best described the graded structure of baboon call types, in which acoustic variation fluctuates between discrete and continuous. In other words, call types could vary to different degrees such that they form a continuum in acoustic space. This led the authors to suggest the use of fuzzy *c*-means clustering in future studies over hard clustering algorithms due to its ability to reveal details in the graded structure of animal vocalizations.

1.3 SUMMARY

The acoustic structure of animal vocalizations varies widely. Scientists can study this acoustic variation to better understand song function in birds. A plethora of analysis methods exists for quantifying acoustic variation (Table 1.2). Scientists can apply manual or automated approaches to a feature- or image-based analysis. Regardless of approach, quantifying acoustic

variation requires data collection, unit identification, measurement extraction (of features or similarity), and possibly dimensionality reduction before running a classifier analysis. Novel automated methods enable scientists to better capture multidimensional variation in sound, whether using a neural network to reveal new axes of acoustic similarity or running a soft clustering algorithm to reveal the detail in graded acoustic structure. Scientists are still testing these methods, but recently published studies show that automated techniques are already opening new doors in the field of bioacoustics (Goffinet et al., 2021; Sainburg et al., 2020; Wadewitz et al., 2015).

Automated methods can enable bioacousticians to better understand structural variation, but their novelty and technical complexity can make implementation difficult. Thus, the applicability of latent visualization to bioacoustics analysis has not been widely explored. Specifically, latent visualization has never been applied to sounds characterized by a unique blend of discrete and continuous structural variation. In the next chapter, my co-authors and I address this gap in the literature with a novel workflow pipeline that uses DTS, UMAP, and fuzzy *c*-means clustering to analyze structural variation in the song of a tropical bird.

Table 1.2. Workflow pipeline of analysis for acoustic variation in animal vocalizations. All algorithms have been compiled here. *Note that *k*-NN and *k*-means clustering both use the variable *k* but are different algorithms.

Workflow Pipeline	Approaches	Sub-categories	Algorithms
Data collection and unit identification			Graphic method (obsolete), spectrogram, DTS
Measurement extraction	Acoustic feature spaces		Spectrogram, PAFs, MPS, MFCCs
	Similarity measures	Derived metrics	CV, Levenshtein distance, Jaccard's coefficient of correlation
		Visual comparison	SPCC, DTW, VAEs
Dimensionality reduction			PCA, MDS, ISOMAP, t-SNE, UMAP
Classifier analyses	Supervised classifiers		*k*-NN*, random forest, FLDA
	Unsupervised classifiers	Hard clustering	UPGMA, Ward's clustering, *k*-means clustering*, HDBSCAN
		Soft clustering	Mixture-of-Gaussians, fuzzy *c*-means clustering
		Clustering validation	Moat indices, silhouette values, NMI

CHAPTER 2: CHARACTERIZING STRUCTURAL VARIATION IN THE NOTES OF ADELAIDE'S WARBLER (*SETOPHAGA ADELAIDAE*) SONGS

2.1 PROBLEM STATEMENT

Numerous methods are available for analyzing structural variation at the note level. Methods range from manual to automated and can be applied to both acoustic feature- and image-based analyses (Chapter 1). Manual (i.e., by-eye) analysis of spectrograms is adequate for note classification in some species (Ballentine et al., 2003; Gil & Slater, 2000). However, by-eye classification of notes in other species has proven challenging.

The Adelaide's warbler (*Setophaga adelaidae*) is one species whose notes are difficult to classify. Notes in male Adelaide's warbler songs are structurally simple and short in duration. Virtually all their acoustic energy is in the fundamental frequency. Notes vary in duration, frequency, and patterns of frequency modulation. This variation is difficult to quantify by eye due to the presence of structural gradation among notes. For example, in Figure 2.1, the contour, or shape, of the note on the spectrogram gradually "morphs" over the first second of the song.

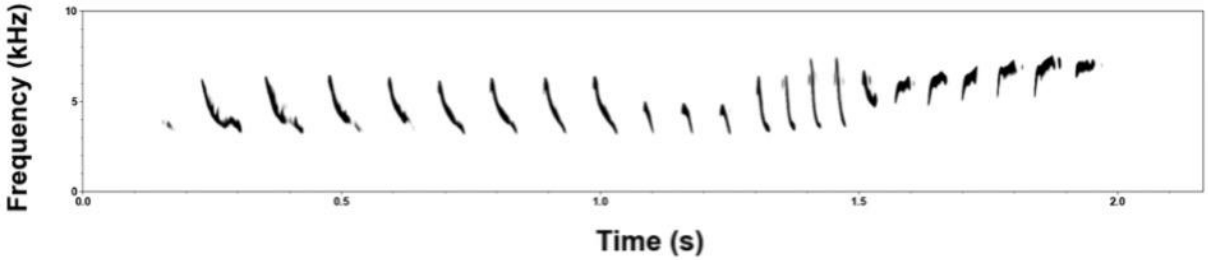


Figure 2.1. Spectrogram of an Adelaide's warbler song visualized in Luscinia 2.16.10.29.01 (max. freq. = 10 kHz, frame length = 5 ms,time step = 1 ms, dynamic range = 35 dB, dynamic equalization = 100 ms, de-reverberation = 100%, dereverberation range = 100 ms, high pass threshold = 1.0 kHz, noise removal = 10 dB; Lachlan, 2007).

This continuous variation in note structure (i.e., contour) differs from the discrete change that occurs at approximately 1.5 seconds. This blend of continuous and discrete variation characterizes note repertoires in Adelaide's warblers, making it difficult to differentiate note contours by eye, much less classify them into types. We hypothesized that automated methods

might help us characterize variation in note structure. Once we characterize this variation, we will be able to infer the presence or absence of discrete note types and better quantify the factors that differentiate notes.

Over the past few years, scientists have begun turning from manual to automated methods to classify animal vocalizations (Elie & Theunissen, 2016; Keen et al., 2021; Sainburg et al., 2020; Thomas et al., 2022). Manual classification requires expert knowledge and consumes significant time and energy. Furthermore, manual classification is inconsistent both within and among observers. In contrast, automated classification has the potential to be quicker, more consistent, and less subjective than manual classification. Automated methods are particularly advantageous because they can handle large datasets and quantify many dimensions of variation. Of course, automated classification has its own limitations.

Automated classification are not as good as human experts at ignoring background noise (Keen et al., 2014). For example, automated classification cannot distinguish notes sung simultaneously by the focal individual and other individuals (Swiston & Mennill, 2009). Perhaps the biggest hurdle for automated methods is their requirement of knowledge in computer science, especially for recently developed methods.

Sainburg et al. (2020) introduced an automated image-based analysis called latent visualization. Contrary to acoustic feature analyses, which use acoustic measurements to quantify differences among notes, latent visualization quantifies differences among notes by comparing the visual image of each note on a sound spectrogram. This novel methodology uses unsupervised machine learning techniques to visualize acoustic variation. Unsupervised techniques reduce dimensionality and cluster datapoints to help users identify patterns in the data (e.g., note types). We explain these techniques further in the methods section.

We applied both dimensionality reduction and clustering to our dataset by following the guide on latent visualization in Sainburg et al. (2020) and Thomas et al. (2022). We were interested to see how this automated, unsupervised image-based analysis would handle the blend of discrete and continuous variation present in the acoustic structure of our species' notes. The goals of our study were to confirm whether discrete note types exist in male Adelaide's warbler songs, explore the patterns that the algorithms revealed in their note structure, and evaluate the usefulness of these algorithms in analyzing the vocalizations of our study species. Based on our visual assessment of sound spectrograms, we hypothesized that we would find discrete note types and that these types would be most differentiated by their note contours (i.e., patterns of frequency modulation over time).

2.1.1 CONTRIBUTION OF AUTHORS

David M. Logue and Samantha Y. Huang designed the study. D.M.L. led data collection and annotation. Peter C. Mower built the song-type exemplar dataset. Brayden L. Carlson developed the workflow pipeline for latent visualization and wrote additional code to clean spectrograms, apply segmentation and clustering, and visualize the latent-space representations and other figures. B.L.C. and S.Y.H. analyzed the data. S.Y.H. wrote the chapter. D.M.L. advised on the data analysis and writing.

2.2 METHODS

2.2.1 STUDY SPECIES

Adelaide's warblers are New World warblers (family: Parulidae) endemic to Puerto Rico and Vieques. They are territorial year-round, socially monogamous, and insectivorous (Toms, 2020). Males sing frequency-modulated trills and have a repertoire size of 22.6 ± 2.6 song types

(Staicer, 1991). Each song comprises 23.5 ± 1.8 notes (*n* = 68,602 notes from 9 males; Logue et al., 2020).

2.2.2 RECORDING AND ANNOTATION

We recorded 20 male, mated Adelaide's warblers at the Cabo Rojo National Wildlife Refuge (U.S. Fish and Wildlife Service; 17.98º N, 67.17º W) during the breeding season from April 13 to May 6, 2017. Each male was continuously recorded from 45 minutes before sunrise to 2 hours 45 minutes after sunrise for four days. All birds were banded with unique combinations of colored bands for identification. Recordings were made with Marantz PMD661 digital recorders and Sennheiser ME 67 shotgun microphones (file format = wav, sampling rate = 44.1 kHz, bit depth = 16 bits).

Recordings were annotated in Raven Pro 1.6.1 (K. Lisa Yang Center for Conservation Bioacoustics, 2019). Each recorded song was assessed for recording quality and assigned a song type by comparing the song spectrogram to spectrograms of the known song repertoire of the individual. It is straightforward to assign songs to song types within an individual's repertoire (Kaluthota et al., 2019). Recording quality was ranked on a scale from one (1) (lowest quality) to three (3) (highest quality). All recordings with a quality ranking of one (1) were excluded. We then built a subset of song-type exemplars for each male. Our goal was to sample the full breadth of acoustic structures without compromising recording quality. We included all song-type variants with a high signal-to-noise ratio (SNR) in our analysis based on visual assessment of spectrograms (*n* = 13,192 notes from 612 songs sung by 20 males).

2.2.3 WORKFLOW FOR LATENT VISUALIZATION

We ran latent visualization in Python 3.11.2 (Python Software Foundation, 2023). Our steps were spectrogram pre-processing, unsupervised dimensionality reduction, and clustering.

2.2.3.1 SPECTROGRAM PRE-PROCESSING

Our workflow for spectrogram pre-processing consisted of spectrogram generation, acoustic filtering, segmentation, and image filtering. First, we generated spectrograms of each recorded song in our exemplar subset (n_fft = 512, hop_length_ms = 1 ms, win_length_ms = 5, ref_level_db = 20 dB, preemphasis = 0.97, min_level_db = -20 dB, min_level_db_floor = -10 dB, db_delta = 5 dB, silence_threshold = 0.01, min_silence_for_spec = 0.01, max_vocal_for_spec = 0.5, min_syllable_length_s = 0.01 s, num_mel_bins = 64, spectral_range = [1,500 Hz, 10,000 Hz], mel_lower_edge_hertz = 1,500 Hz, mel_upper_edge_hertz = 10,000 Hz, butter_lowcut = 1,500 Hz, butter_highcut = 10,000 Hz, sample_rate = 44,100 Hz, bandpass_filter = true, reduce_noise = false, normalize = true, dereverberate = true, mask_spec = false, realtime = false, exclude = [], power = 1.5, noise_reduce_kwargs = {}, mask_spec_kwargs = {spec_thresh = 0.9, offset = 1e-10}; based on default parameters in Sainburg et al., 2020). We applied acoustic filters to isolate the image of each note and normalize the amplitude among notes. These acoustic filters were a bandpass filter (set manually by cutting off the frequencies above the highest frequency and below the lowest frequency sung by a male in every song), amplitude normalization, and dereverberation. Amplitude was normalized using the *librosa.util.normalize()* function in the librosa package (norm = inf, axis = 0, threshold = None, fill = None; McFee et al., 2023). Dereverberation was applied using the *wpe()*, *istft()*, and *stft()* functions in the nara_wpe package (size = 512 samples, shift = 128 samples, channels = 2, sampling rate = 44,100 Hz, delay = 1, iterations = 20, taps = 10, alpha = 0.9999; Drude et al., 2018). Notes within each song were then separated with a segmentation algorithm.

Dynamic threshold segmentation (DTS) looks for silent gaps between notes. Users define silence by setting a threshold based on background noise, the expected lengths of a note, and the

period of silence between notes (Sainburg et al., 2020). We developed an interface called animal vocalization segmentation (AVS) to visualize our application of DTS (Appendix 1). We started with the parameters we used for spectrogram generation and then manually adjusted the parameters for each song to optimize segmentation. We manually excluded segments that included partial notes that were generated when DTS separated the notes incorrectly and segments with loud overlapping noise (e.g., sound from other birds that overlapped the frequency range of the focal note, Appendix 2). We also collected metadata for each note spectrogram (Table 2.1). Note spectrograms were then zero-padded to the same time length and converted to images.

Table 2.1. Measurements collected for each note.

Measurement	Meaning
Individual	The bird that sang the note
Song recording	The song that the note is a part of
Minimum frequency	The lowest frequency of the note
Maximum frequency	The highest frequency of the note
Frequency midpoint	The average of the minimum and maximum frequencies
Frequency bandwidth	Maximum frequency minus minimum frequency
Duration	The length of the note in seconds
Position	The position of the note within the song (e.g., position = 1 means the first note of the song)

After generating individual note images, we applied image filtering because the images still contained noise. We applied denoising, median blur, brightness/contrast, and unsharp mask using visual inspection. Denoising removes the residual noise around a note using non-local means, which replaces the color of a pixel with the average of the colors of similar pixels. We used the *fastNIMeansDenoising()* function in the opencv-python package (h = 35, templateWindowSize = 7 pixels, searchWindowSize = 7 pixels; Buades et al., 2011) to remove the reverberation around notes that the dereverberation algorithm could not remove. Median blur

applies a filter to replace each pixel of a note image with the median of its neighboring pixels. We used the *medianBlur()* function in the opencv-python package (ksize = 1; Bradski, 2000) to smooth note images and remove imperfections. Brightness/Contrast adjusts the brightness of the note image or the contrast between the note and its background. We adjusted the contrast using the *convertScaleAbs()* function in the opencv-python package (alpha = 1.5, beta = 0; Bradski, 2000). This function calculates absolute values and converts them to 8-bit, which we used to maximize the crispness of the note against its background. Unsharp mask combines an image with a Gaussian smoothing filter (i.e., blurred version) of the image to sharpen the original image. We applied the *unsharp_mask()* function in the scikit-image package (radius = 3.5, amount = 3.5, channel_axis = False, preserve_range = False; Walt et al., 2014) to counteract any blurring from other filters and increase note crispness.

Finally, all note images were converted into feature vectors of equal length so we could apply dimensionality reduction (Thomas et al., 2022). We converted each image into a 64x50 array of numbers. These numbers correspond with the pixels in the image such that each number represents the intensity of a pixel in grayscale color space. Intensity ranged from 0 to 255, where 0 represents black and 255 represents white. Arrays were then row-wise concatenated to form feature vectors such that each number was a feature (i.e., 64 * 50 = 3200 features per note image). We then applied dimensionality reduction to these feature vectors.

2.2.3.2 UNSUPERVISED DIMENSIONALITY REDUCTION

Dimensionality reduction compresses high-dimensional data into a low-dimensional latent space while maintaining the structure and variance of the original data (Sainburg et al., 2020). Animal vocalizations like birdsong work well with dimensionality reduction. Although sound has multi-dimensional temporal and spectral variation, birdsong can usually be explained

in terms of lower-dimensional properties (e.g., Elie & Theunissen, 2016; Gil & Slater, 2000; Mota & Cardoso, 2001; Podos et al., 1992). In other words, birdsong can usually be explained in terms of properties intuitive to humans like frequency and duration. In a latent space, each dimension represents features of the dataset, and each datapoint represents, in the case of our study, a note from a song. The distance between two datapoints in the latent space represents their dissimilarity: as distance increases, similarity between the two datapoints decreases.

We applied a novel unsupervised dimensionality reduction algorithm called uniform manifold approximation and projection (UMAP). UMAP maps data points in latent space by building a weighted nearest-neighbor graph of the high-dimensional data. The algorithm then embeds the graph in a lower-dimensional latent space while optimizing preservation of the topological structure and probabilistic weighting of the graph (McInnes et al., 2020). We used UMAP to generate a two-dimensional (2D) and three-dimensional (3D) latent-space representation of our data.

2.2.3.3 CLUSTERING

Clustering algorithms assign unlabeled data to different groups (i.e., clusters) based on the inherent structure of the data (Xie & Beni, 1991). In the case of our study, we expected each cluster to represent a different note type. We applied fuzzy c-means clustering, which is a soft clustering algorithm, to our data. Soft clustering algorithms estimate the probability that each datapoint is associated with a given cluster. We chose fuzzy c-means clustering because we were interested in the gradations between note types. This algorithm determines the optimal number c of clusters for a given dataset based on the maximum number of clusters allowed and the fuzziness parameter μ, where $\mu = 1$ indicates hard clusters (i.e., each data point is associated with only one cluster), and higher values of μ indicate increasingly fuzzy clusters. The algorithm

assigns each data point a membership value *m* for each cluster, which represents how well the data point matches the properties of the cluster. The value *m* ranges from 0 to 1, where *m* = 0 means the data point shows no cluster properties, and *m* = 1 means the data point fully shows cluster properties. Each data point is then assigned to the cluster with the highest membership value. Fuzzy *c*-means clustering arbitrarily assigns a number *c* of cluster centroids and then optimizes those centroids by taking weighted membership values into consideration over multiple iterations and adjusting accordingly (Bezdek et al., 1984; Wadewitz et al., 2015; Zadeh, 1965). The code for our latent-space representation is available at https://github.com/braycarlson/warbler.py (includes segmentation, visualization, and clustering).

We verified our clustering solution with a hyperparameter called the Xie-Beni index and bootstrap validation. A hyperparameter optimizes the parameters of an algorithm. The Xie-Beni index divides the compactness of clusters by the separation among clusters (Hsu, 2018; Xie & Beni, 1991). We determined the best clustering solution by calculating the configuration of parameter settings that produced the lowest Xie-Beni index for the fuzzy *c*-means clustering algorithm in Python (Dias, 2019; Pal & Bezdek, 1995). The Xie-Beni index uses a validity function to identify overall compact and separate fuzzy *c*-partitions without assuming the number of substructures inherently present in the data. The index depends on the dataset, geometric distance measure, distance between cluster centroids, and fuzzy partition generated by any fuzzy algorithm used. The Xie-Beni index is mathematically justified through its relationship with the separation index (Xie & Beni, 1991).

We validated the 14-cluster solution for our dataset by bootstrapping our full dataset and data subsets. We sampled subsets of our data 1,000 times with replacement and clustered each subset with the parameters that minimized the Xie-Beni index. Each subset contained 800 data

points from each cluster (11,200 points total; Appendix 3). We also sampled smaller subsets of our data (800 points total) 1,000 times with replacement. and clustered each resampled subset with the parameters that minimized the Xie-Beni index (Appendix 4).

2.2.3.4 FURTHER EXPLORATION

We explored the 3D latent-space representation with an interactive visualization tool developed in Jupyter Notebook (Kluyver et al., 2016; Thomas, 2021). We also used R 4.2.1 (R Core Team, 2022) to generate song traces of all songs in our dataset. Each song trace was a subset of the 2D latent-space representation consisting of the notes in a recorded song and connected in the order they were sung. We generated song traces to visually compare manually labeled song types, especially those shared among individuals. Our goal was to evaluate the utility of the latent-space representation in assigning song types.

2.3 RESULTS

2.3.1 LATENT-SPACE REPRESENTATIONS

2.3.1.1 UNCLUSTERED LATENT-SPACE REPRESENTATIONS

The latent visualization arrayed most of the notes in one largely continuous, crescent-like shape. Without any clustering algorithm, there were a couple of breaks in the crescent and a few outlying groups but no clearly defined groups or separation between points (Figure 2.2). When viewed in three-dimensional space, the ends of the crescent extended in opposite directions along the *z*-axis, with the left end turning to extend along the *x*-axis. The top half of the crescent was broader than either end of the crescent, and the broad half descended along the *x*-axis (Figure 2.3).

We observed several patterns in the latent-space representation. Minimum frequency, maximum frequency, and frequency midpoint all increased starting at the bottom right corner of

the representation and traveling counterclockwise to the other point of the crescent (Figure 2.4a–c). Position also followed this counterclockwise increase albeit less strongly (Figure 2.4d). Frequency bandwidth increased along the *y*-axis such that the top half of the crescent had a greater frequency bandwidth than the bottom half (Figure 2.4e, top half highlighted). Finally, duration was greatest left of center in the latent-space representation (Figure 2.4f). We used the 3D representation to better visualize these trends, many of which increased or decreased along the line $z = x$ (Figure 2.5). A summary of values for each measurement is in Table 2.2.

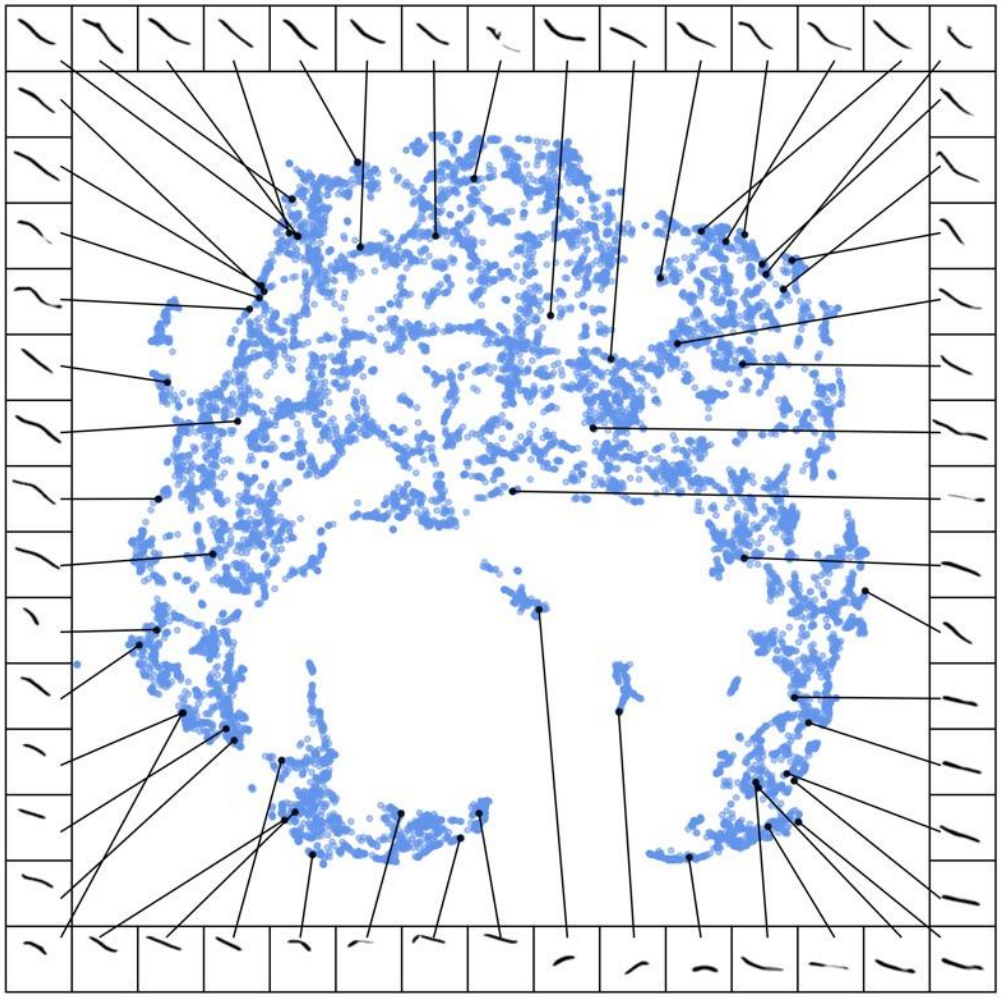


Figure 2.2. Unclustered 2D latent-space representation of Adelaide's warbler notes.

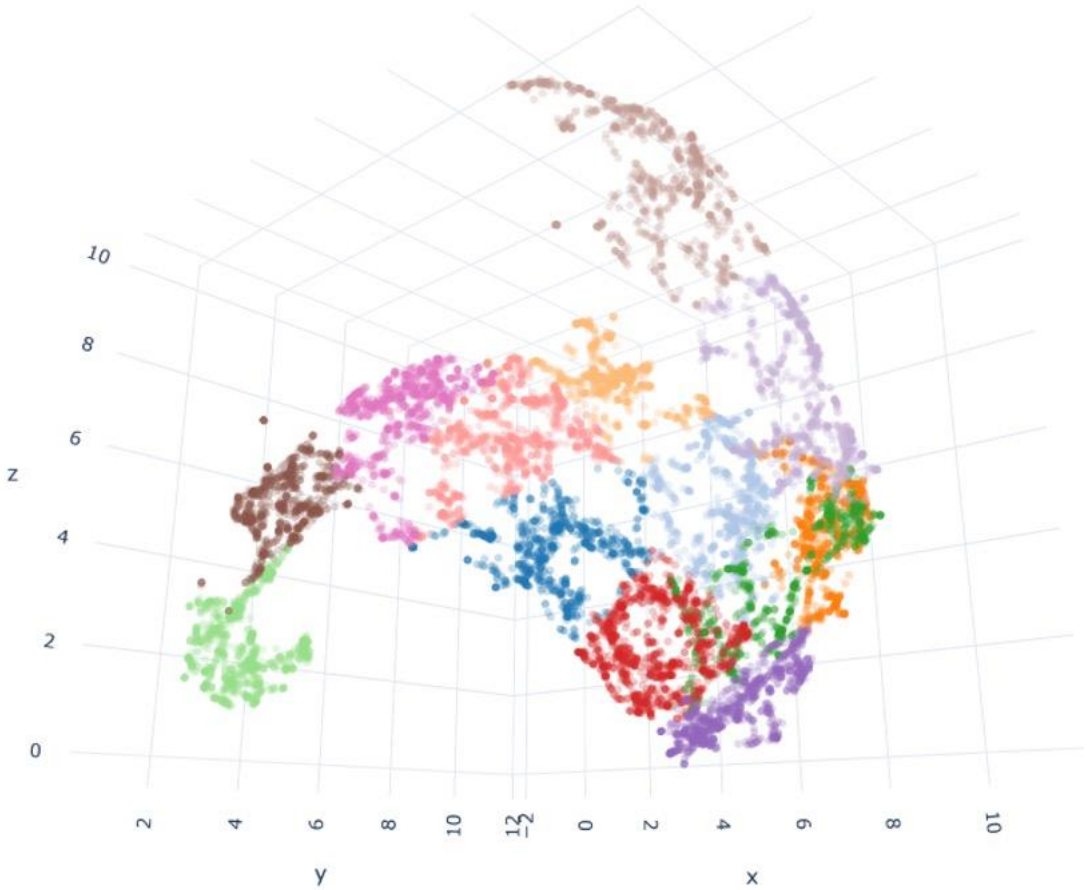


Figure 2.3. Clustered 3D latent-space representation of Adelaide's warbler notes. Colors represent clusters derived from fuzzy c-means clustering. We colored clusters in this figure to help viewers interpret the 3D graph. The 3D representation has one fewer cluster than the 2D representation because the clustering algorithm is sensitive to the number of dimensions. Additional images of the 3D latent-space representation at different angles are in Appendix 5.

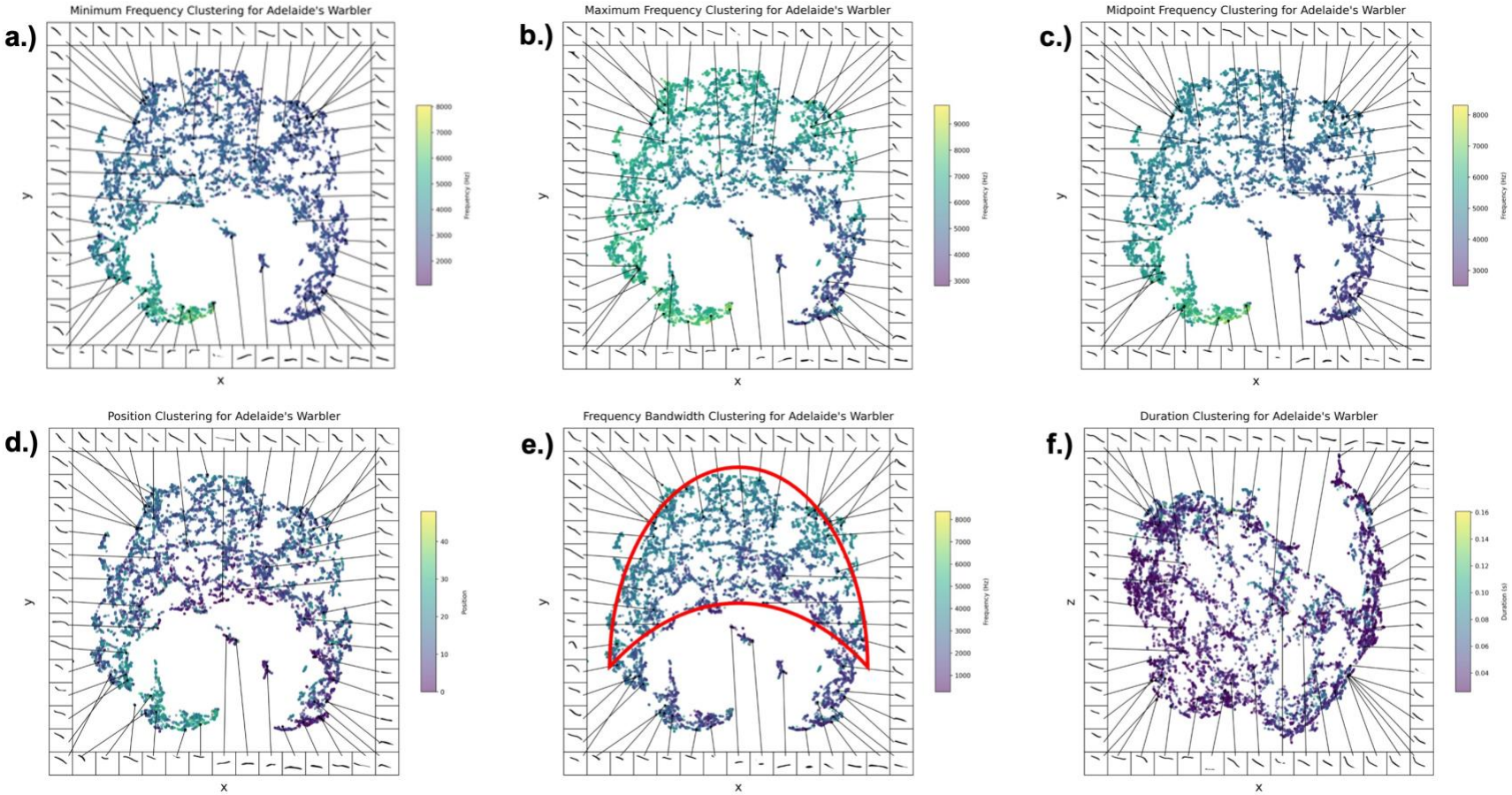


Figure 2.4. 2D latent-space representations of Adelaide's warbler notes labeled by different measures. The red outline in *e* marks a region with relatively high frequency bandwidth.

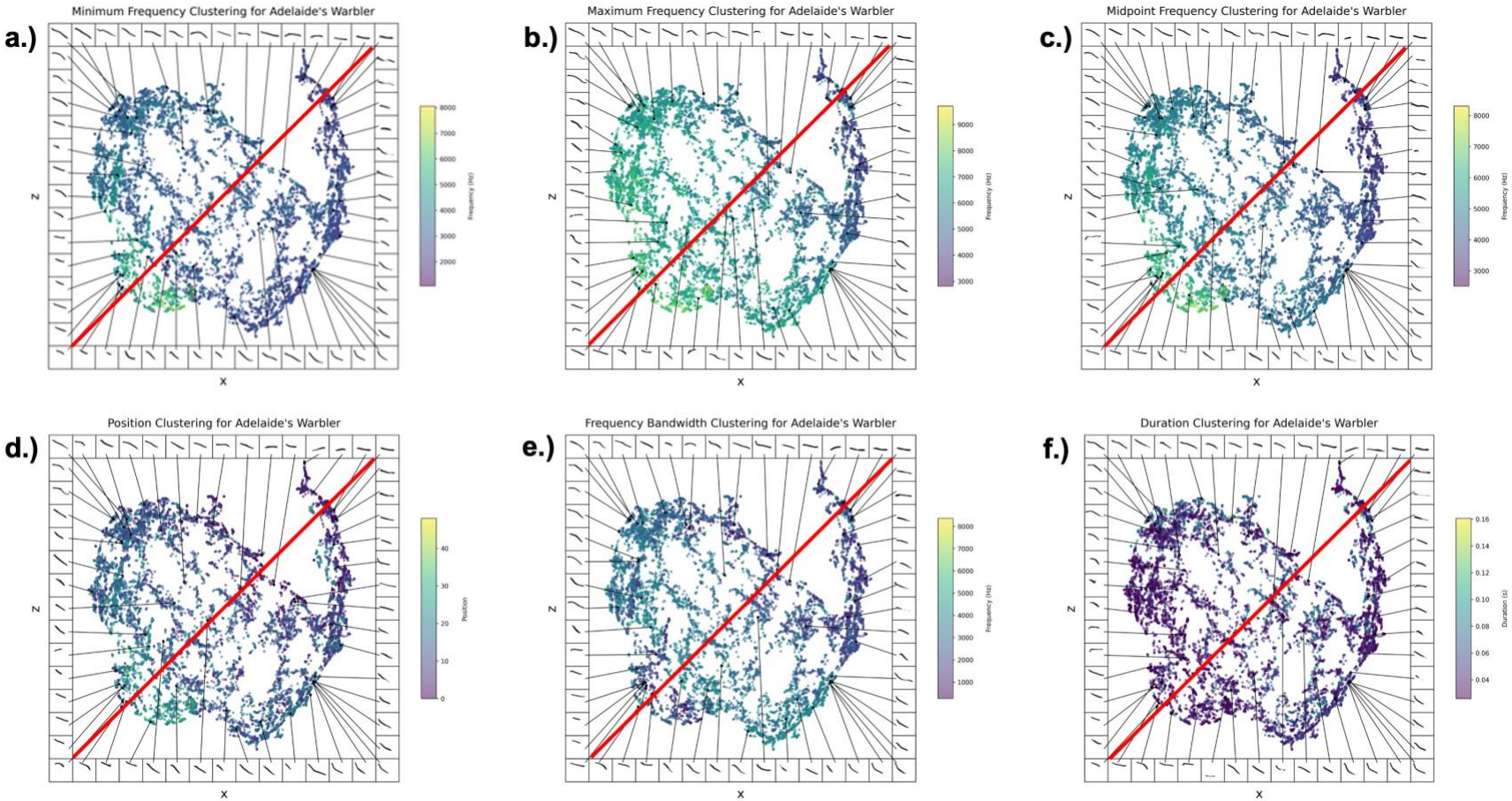


Figure 2.5. 2D latent-space representations of Adelaide's warbler notes rotated to show the *x*- and *z*-axes. In all representations, a major axis of variance is $z = x$ (red line).

Table 2.2. Summary of acoustic parameters of male Adelaide's warbler notes (n = 13,192).

Measurement	Mean	Standard deviation	Minimum	Maximum	Median
Minimum frequency (Hz)	3197	903	1057	8051	3007
Maximum frequency (Hz)	6192	1053	2814	9710	6334
Frequency midpoint (Hz)	4694	811	2506	8308	4660
Frequency bandwidth (Hz)	2994	1105	259	8364	3033
Duration (s)	0.05	0.02	0.03	0.16	0.05
Position	13	8	1	49	12

2.3.1.2 CLUSTERED LATENT-SPACE REPRESENTATIONS

General frequency characteristics, position, duration

We found 14 clusters in our data (Figure 2.6). Clusters were of similar size and evenly spaced. Our clustering solution had a fuzziness parameter μ = 2.9. Maximum membership values (i.e., membership values for the clusters to which data points were assigned) averaged m = 0.34 and skewed right for the whole dataset (Figure 2.7) and for each cluster (Table 2.3).

The clusters at the ends of the crescent differed the most in terms of frequency midpoint, minimum frequency, and maximum frequency. For example, cluster B (in the bottom left corner of the representation) had the highest median values of frequency midpoint, minimum frequency, and maximum frequency among all the clusters. In contrast, cluster M (bottom right) had the lowest median values of minimum frequency, maximum frequency, and frequency midpoint (Figure 2.8a–c). Cluster B also had the highest median position value, and cluster K had the lowest median position value (Figure 2.8d). In the top half of the crescent, the clusters along the inner edge (clusters K, J, N) and outer edge (clusters H, C, E) differed in frequency bandwidth and maximum frequency. The clusters along the inner edge had lower frequency bandwidth and

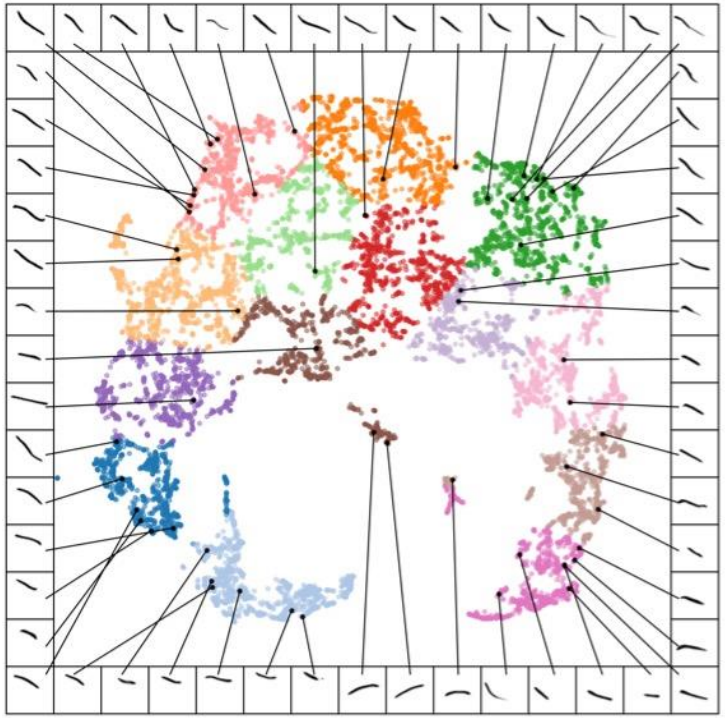


Figure 2.6. 2D clustered latent-space representation of Adelaide's warbler notes. We found 14 clusters in our data.

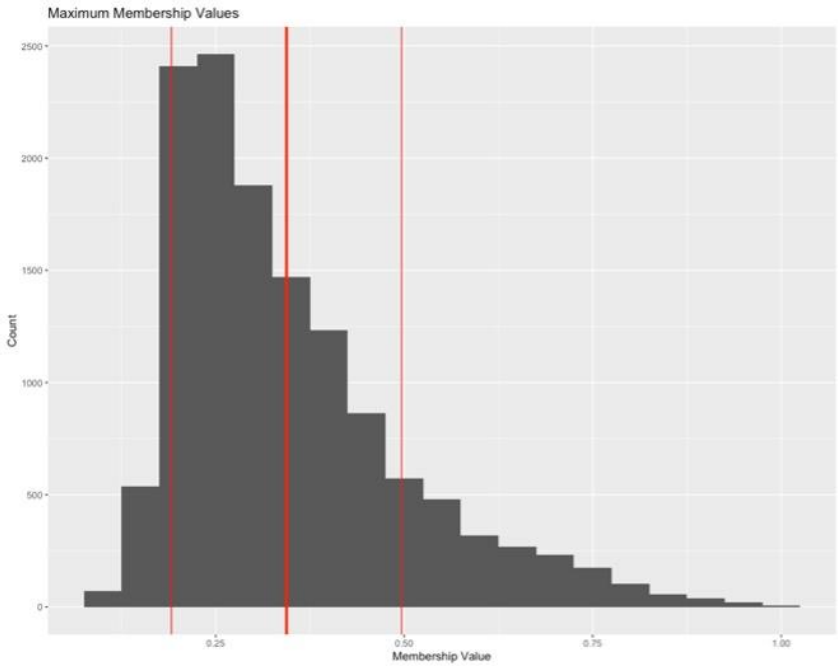


Figure 2.7. Histogram of maximum membership values for our dataset. Red lines represent mean (0.34, center bolded) and standard deviation (± 0.15).

Table 2.3. Maximum membership values for each cluster. Maximum membership values for the whole dataset are bolded.

Cluster	Mean	Standard Deviation	Minimum	Maximum	Median
A	0.40	0.14	0.19	0.98	0.37
B	0.38	0.16	0.17	0.99	0.35
C	0.32	0.14	0.16	0.83	0.28
D	0.30	0.12	0.16	0.82	0.25
E	0.34	0.13	0.16	0.93	0.30
F	0.29	0.15	0.15	0.98	0.23
G	0.29	0.12	0.15	0.88	0.24
H	0.37	0.15	0.16	0.95	0.34
I	0.34	0.13	0.18	0.96	0.30
J	0.34	0.17	0.16	0.98	0.25
K	0.30	0.17	0.12	0.99	0.24
L	0.39	0.16	0.12	0.95	0.35
M	0.41	0.15	0.12	0.82	0.43
N	0.35	0.17	0.17	0.99	0.29
Total	**0.34**	**0.15**	**0.12**	**0.99**	**0.30**

maximum frequency than the clusters along the outer edge (Figure 2.8e). Finally, median duration values increased from both ends of the crescent towards the inner left edge (clusters I, K, G). The clusters in the bottom left corner and at the top of the representation (clusters B, H, A, C) had the lowest median duration values (Figure 2.8f). We inspected the latent-space representation separately for each individual and found that all individuals sang notes from every cluster (Figure 2.9).

Note contours

We did not find strong associations between note contours and cluster assignments. Most notes decreased in frequency, but a minority of notes increased in frequency (Table 2.4). Within clusters in the top half of the crescent (clusters D, H, F, C, G, E), notes varied little with respect to slope. All clusters, however, contained variation in the number, location, and type of frequency inflections. Notes contained 0–3 inflection points and ranged from lines to different

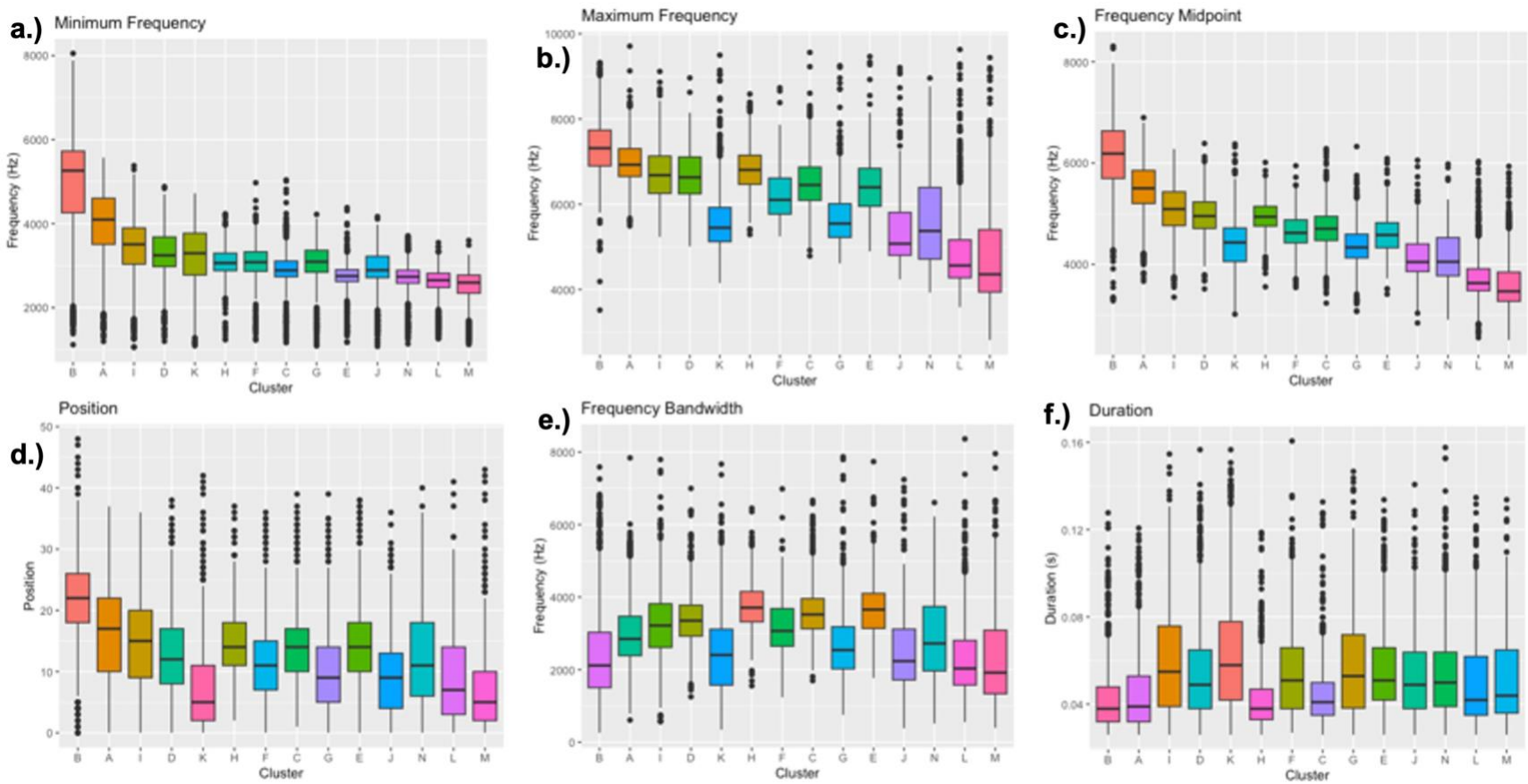


Figure 2.8. Boxplots summarizing attributes of note clusters. The *x*-axis is organized by clusters going left to right in the 2D latent-space representation.

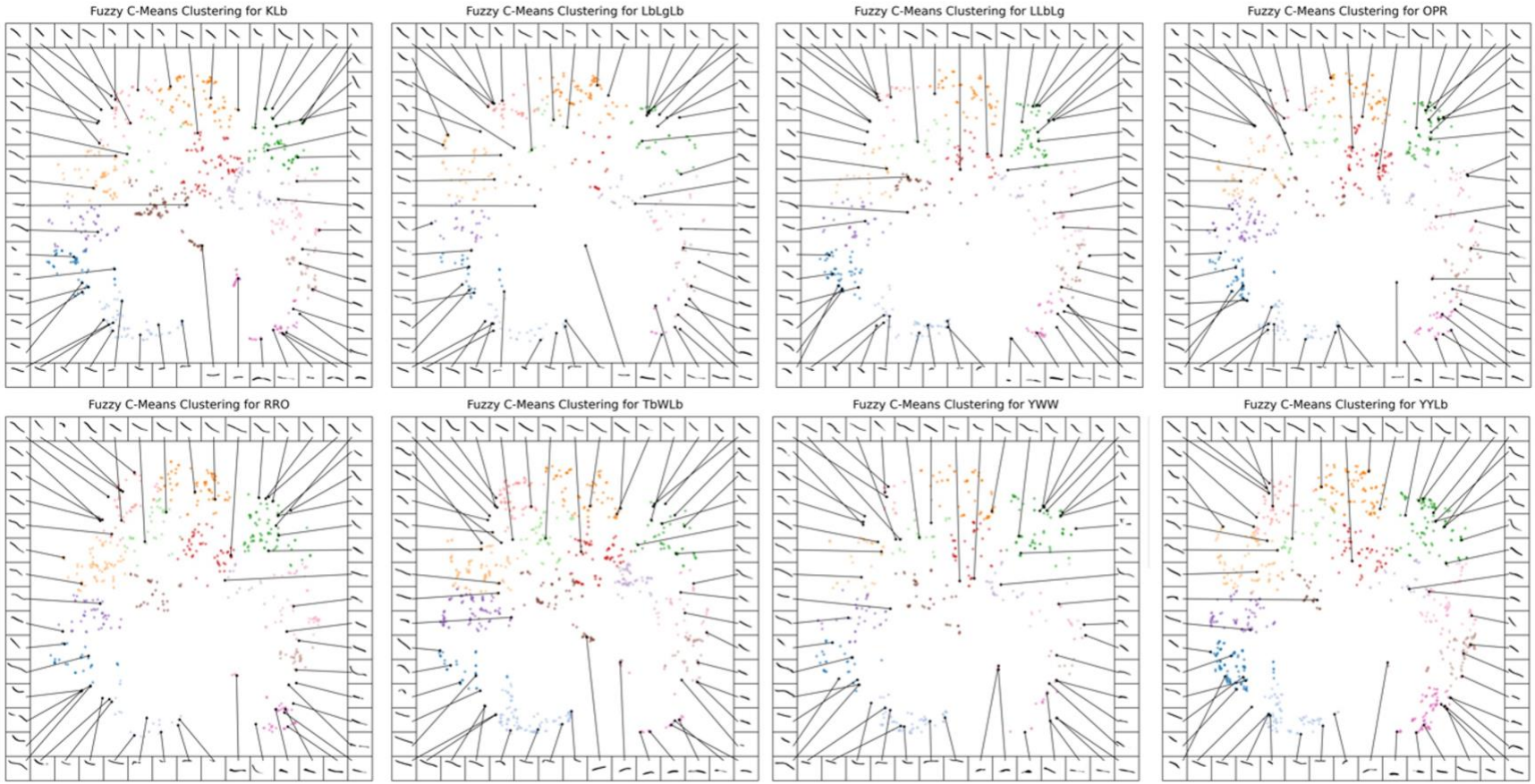


Figure 2.9. 2D latent-space representations of individual repertoires. All individuals represented in our dataset sang notes from every cluster.

types of curves (e.g., concave up or concave down curves, reverse sigmoidal curves). Some notes contained an inflection point roughly halfway through the note or multiple inflection points spread out equally throughout their duration. Other notes contained a sharp inflection at the beginning of the note, sometimes adding a shallower inflection point later in the note (Table 2.5). The examples in Tables 2.4 and 2.5 are not special cases. Variation in overall slope and note shape were ubiquitous in the dataset.

2.3.1.3 SONG TRACES

Song traces followed the contour of the crescent and included notes in multiple clusters. Song traces traveled counterclockwise, clockwise, or in both directions. Most song traces we examined went through the entire crescent, and some reversed course and went back through the crescent in the opposite direction. Within some song traces, consecutive notes went back and forth from one area of the crescent to another, but these areas did not always correspond to the labeled clusters.

Song traces of the same song type among individuals often matched but were sometimes inconsistent (Figures 2.10–2.12). Some song traces contained notes from different clusters. Other song traces followed different paths despite representing the same song type.

Table 2.4. Sample notes from clusters containing notes that have positive and negative frequency slopes.

Slope	Cluster B	Cluster A	Cluster K	Cluster L	Cluster M
Positive					
Flat					
Negative					

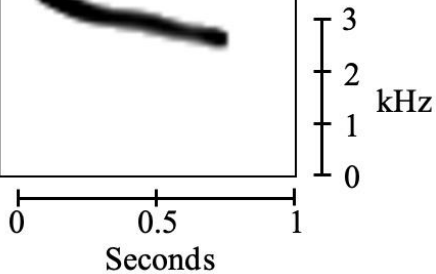


Table 2.5. Sample notes from clusters in the top half of the 2D latent-space representation. Multiple note contours exist in each cluster, varying in duration, slope, and inflection points.

Inflection point(s)	Cluster D	Cluster H	Cluster F	Cluster C	Cluster G	Cluster E
0						
1						
2						
3						

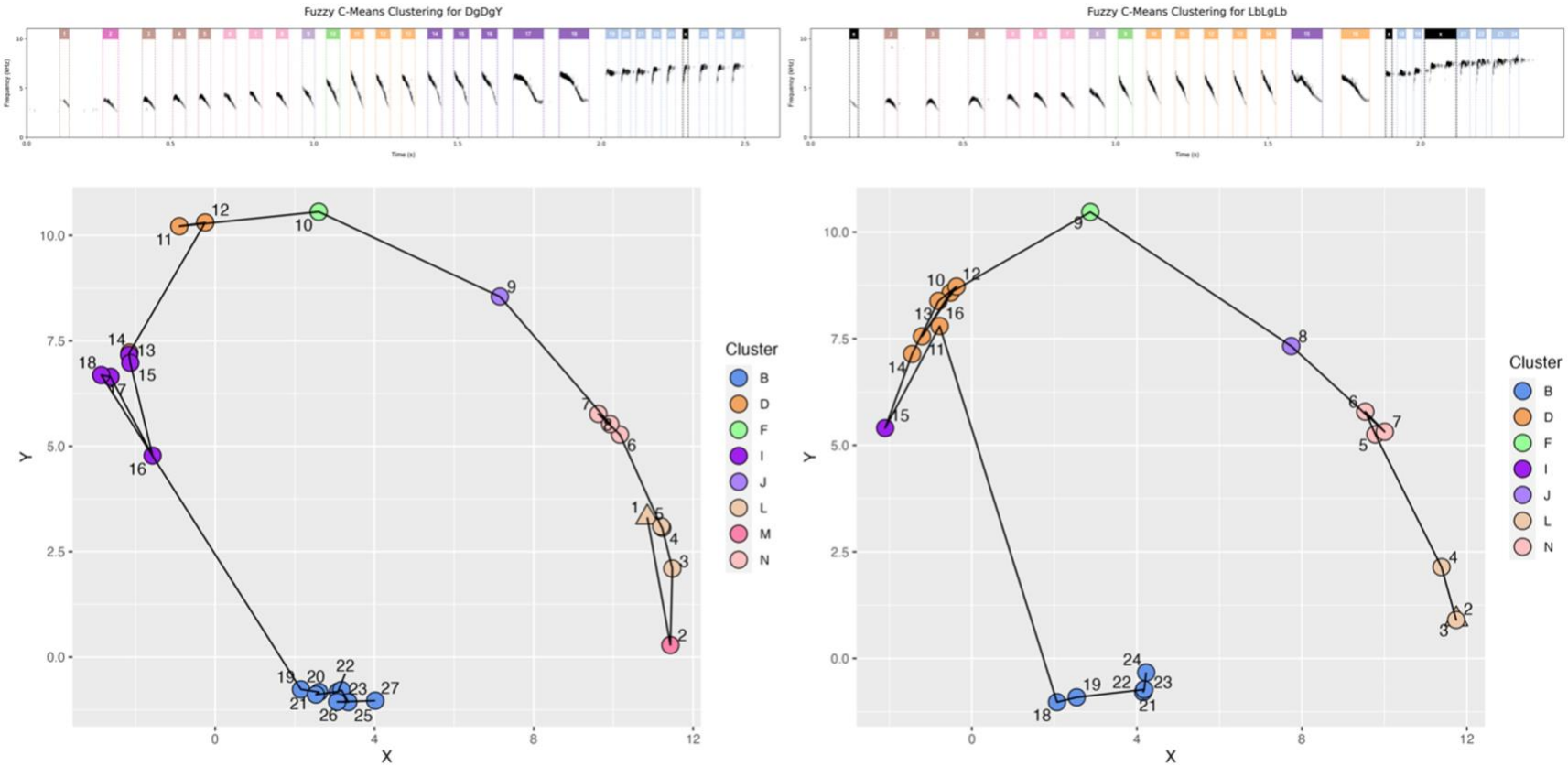


Figure 2.10. Song traces of the same song type for individuals DgDgY and LbLgLb with corresponding spectrograms. Number labels indicate note position. Clusters are color-coded. In the spectrograms, black bars with an "x" indicate segments that were excluded due to poor segmentation. Additional song traces of this same song type for other individuals are in Appendix 6.

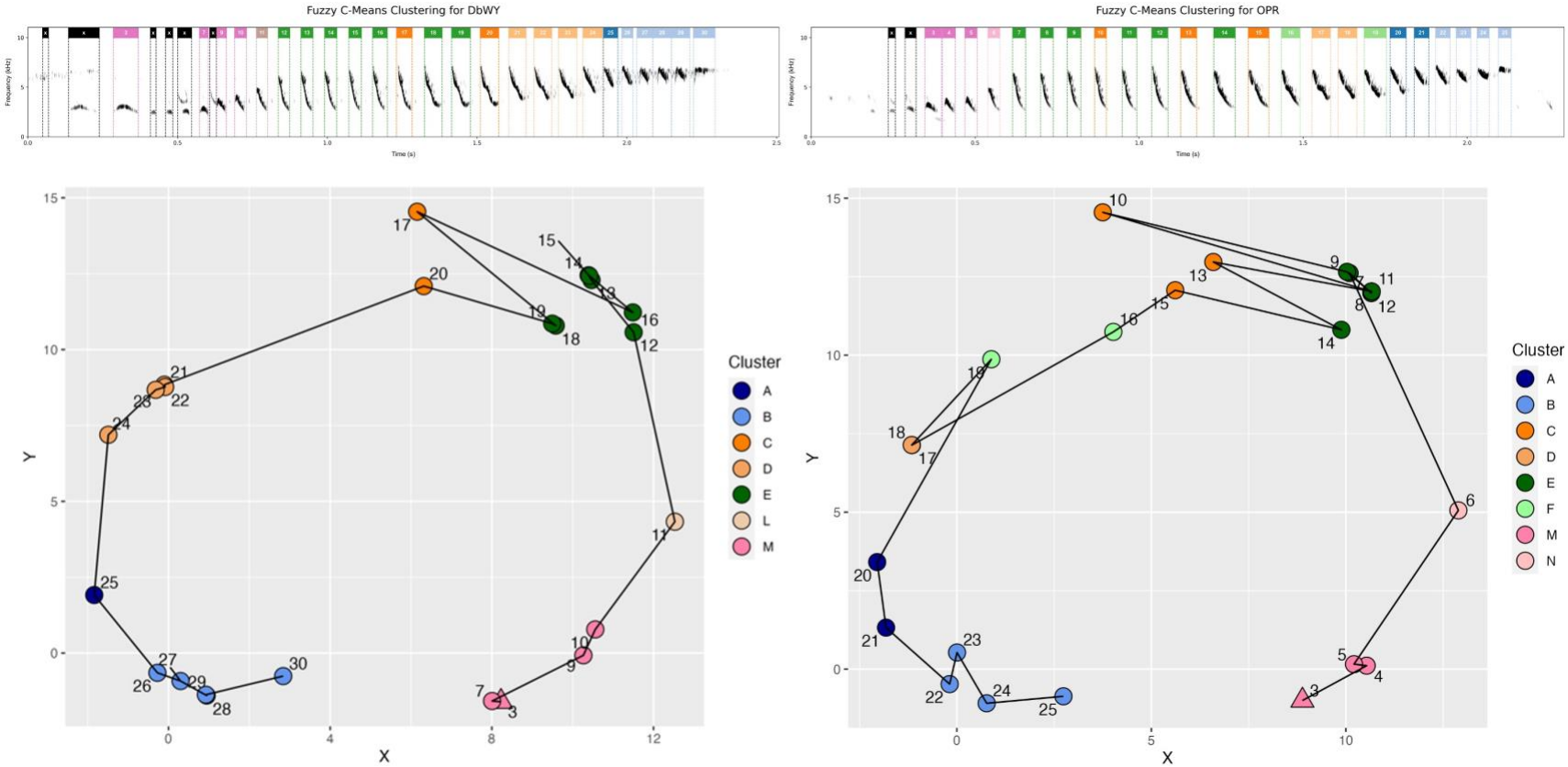


Figure 2.11. Song traces of the same song type for individuals DbWY and OPR with corresponding spectrograms. Number labels indicate note position. Clusters are color-coded. In the spectrograms, black bars with an "x" indicate segments that were excluded due to poor segmentation. Additional song traces of this same song type for other individuals are in Appendix 7.

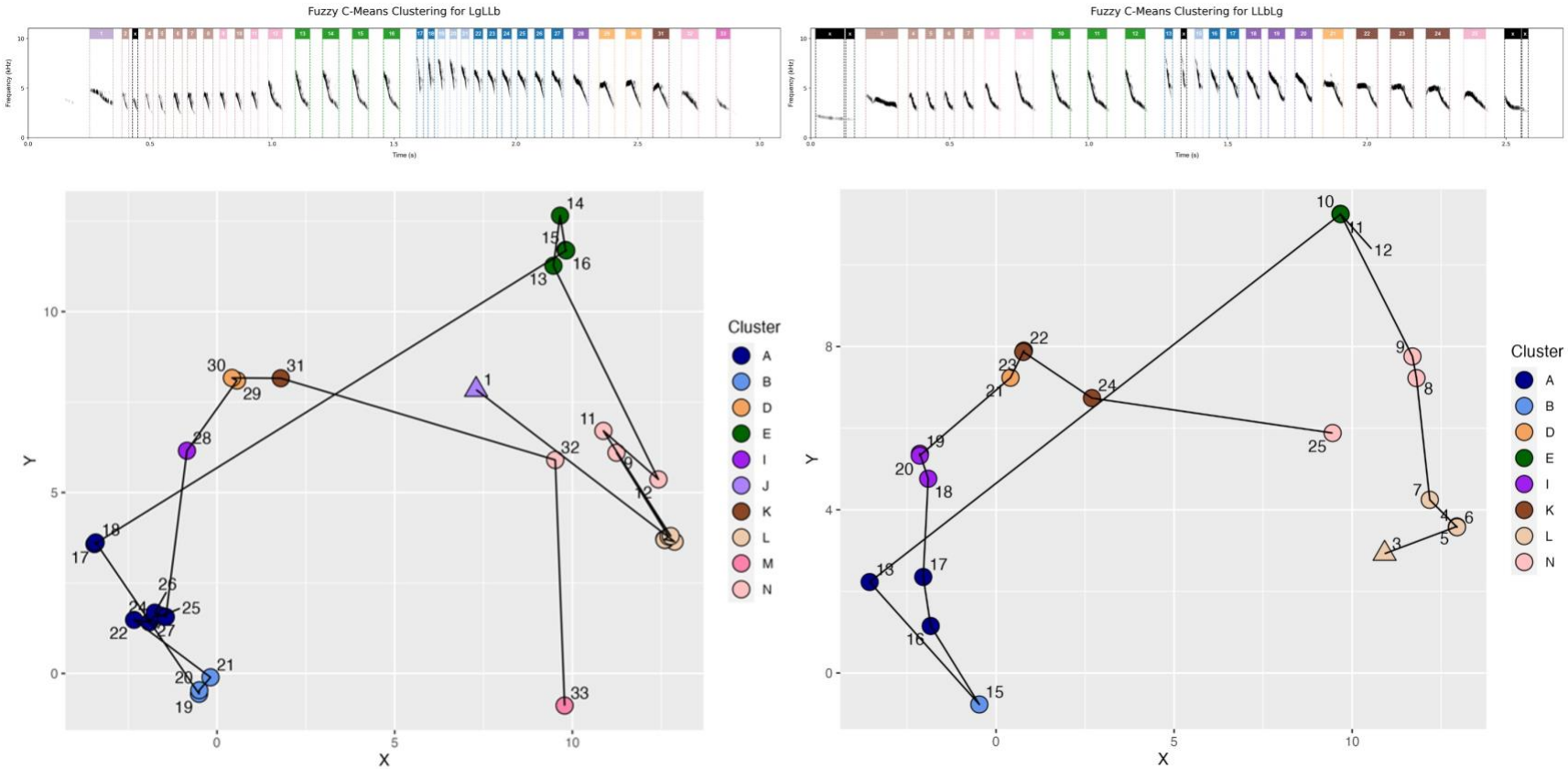


Figure 2.12. Song traces of the same song type for individuals LgLLb and LLbLg with corresponding spectrograms. Number labels indicate note position. Clusters are color-coded. In the spectrograms, black bars with an "x" indicate segments that were excluded due to poor segmentation. Additional song traces of this same song type for other individuals are in Appendix 8.

CHAPTER 3: DISCUSSION AND CONCLUSIONS

3.1 DISCUSSION

We ran an image-based analysis of acoustic note structure in Adelaide's warblers using novel automated methods. We used the resulting latent-space representation to explore the variation present among Adelaide's warbler notes along with the possibility of discrete note types.

The automated analysis projected notes into latent space in a single, continuous, crescent-like shape and found 14 fuzzy clusters (Figure 2.6). Most notes had low membership values and partially showed the properties of multiple clusters (Figure 2.7). Notes were projected primarily based on general frequency characteristics (Figure 2.4). Song traces reflected broad patterns of frequency modulation throughout songs (Figure 2.10–2.12). That general frequency characteristics made up a primary axis of variation was no surprise to us due to the frequency-modulated song structure of Adelaide's warblers. What we did not expect, however, was the wide variation of note contour within clusters.

3.1.1 NOTE CONTOURS

Adelaide's warbler note structures vary not just in general frequency characteristics but in patterns of frequency modulation. We thought the automated analysis would cluster notes based on their contours. We expected the cluster labels to differentiate note contours, and we expected the fuzzy nature of the clustering algorithm to accommodate the note morphs that make by-eye classification of note types difficult. Contrary to our expectations, the automated analysis prioritized general frequency characteristics over note contour and assigned the same cluster label to notes that had similar general frequency characteristics regardless of note contour (Tables 2.4–2.5). This was surprising to us because note contour is the most important aspect of note structure in by-eye classification.

Note contours are important because they seem to represent distinct vocal gestures. Each vocal gesture requires the singer to manipulate his vocal apparatus in a unique way. In swamp sparrows (*Melospiza georgiana*), note contours differentiate note types. These notes, in turn, define song types, in which a single note or syllable (i.e., series of notes) is repeated multiple times (Clark et al., 1987; Nowicki et al., 1992). Additionally, note contours are important for by-eye classification because their differences are easy to see. However, latent visualization did not differentiate Adelaide's warbler notes based on their contours. One possible explanation may lie in the workings behind this image-based analysis. When UMAP measures the distance between feature vectors, it compares the same pixel in each note spectrogram. UMAP does not consider shifting the values in feature vectors to normalize for changes in general frequency characteristics, therefore overlooking possible similarities in note contour. Disregarding note contour may have contributed to the lack of distinction among clusters.

3.1.2 CLUSTERS AND SONG TRACES

Clusters were not very distinct with respect to the acoustic feature measurements (Figure 2.8). This lack of distinction may be due to the continuous nature of the notes in the projected crescent. Clusters were not well-separated in space, and the ranges of acoustic features and other characteristics overlapped among clusters. This overlap may be responsible for the finding that some song traces reverse direction multiple times among clusters in the top half of the crescent. Unfortunately, due to the image-based nature of this analysis, we are unsure what may be contributing to such extensive structural variation both along the *x*-axis and in our song traces.

We generated song traces of all songs in our dataset to compare song types. We were especially interested in comparing song traces of the same song type sung by different individuals. Each song trace spanned the majority of the latent space, which we expected due to

the frequency modulation within songs. Some song traces of the same song type looked very similar (Figure 2.10–2.12). Other song traces had strong similarities but did not match completely (Appendices 6–8). The sensitivity of our clusters to general frequency characteristics meant that the equivalent notes (e.g., note 10 in song type 7) sung by different individuals were sometimes assigned different cluster labels. This sensitivity may also contribute to differences in the path that song traces took despite representing the same song type. The song traces further lack complete accuracy due to our exclusion of some note segments. We excluded segments that incorrectly separated notes or contained noise, causing some song traces to skip notes present within songs. When this occurred, song traces connected notes that were not sung consecutively. We believe, however, that song traces may still be useful in refining song repertoires. Instead of relying on by-eye classification of song spectrograms alone, we could generate corresponding song traces to support or challenge our classification of song types. Furthermore, song traces could also reveal variation within a song type that is not obvious from by-eye analysis of spectrograms.

In contrast, we are wary of using the cluster labels from this automated analysis to assign note types in our study population. Less than 15% of the data had a membership value of $m \geq 0.5$ (Figure 2.7). In other words, less than 15% of the data were strongly associated with the cluster to which they were assigned. Based on the unclustered projection alone, we infer that discrete note types may not exist in Adelaide's warblers, but we are hesitant to conclusively argue against the existence of discrete types without first exploring alternate options for quantifying acoustic note structure in this species.

3.2 CONCLUSIONS

3.2.1 IMPLEMENTATION OF LATENT VISUALIZATION

Proponents of latent visualization have argued for its ability to generate intuitive and quantifiable visuals of acoustic variation (Sainburg et al., 2020; Thomas et al., 2022). We did not find that to be the case for acoustic note structure in Adelaide's warblers. We found trends in general frequency characteristics but were unable to discern other axes of variation or differences among clusters. We had hoped that an image-based analysis would allow us to quantify the blend of discrete and continuous variation among note structures, but the method we used did not cluster notes by note contour (Tables 2.4–2.5). Furthermore, the novelty of this methodology made application to our dataset difficult. We required a computer science expert to write additional code, and the analysis had to be run on a supercomputer. This analysis was ultimately less onerous than a manual analysis of acoustic note structure, but we would recommend a more supervised analysis for notes that vary in frequency and patterns of frequency modulation (i.e., note contours).

Perhaps a shortcoming of this study is our lack of a previously labeled dataset for comparison. Because latent visualization is still a novel methodology in the field of bioacoustics, it may have been better to first build off of a supervised form of analysis. Other scientists have tested unsupervised techniques for analyzing acoustic variation by comparing them to other automated analyses, or to manual analyses (Goffinet et al., 2021; Keen et al., 2021; Wadewitz et al., 2015). These methods are so novel that they have not yet been widely applied in the field of bioacoustics. Our study is therefore also a test of the application of this methodology to acoustic variation in birdsong, specifically to birdsong where the presence of discrete note types is

unclear. Latent visualization has worked in some studies like magpie calls (Walsh et al., 2023). Application to Adelaide's warblers, however, has proven less useful.

Latent visualization may not have been the best methodology to help us better understand structural note variation in our species. Nevertheless, the latent-space representation sheds new light on this variation by showing us that patterns underlie note organization in Adelaide's warbler songs. All recorded individuals sang notes that cover the entire range of frequency in our study population (Figure 2.9), but the song traces showed us that variation exists among individuals for the same song type (Appendices 6–8). Adelaide's warblers may discriminate less between notes and may instead focus on broader acoustic gestures, like how individuals sweep through different frequency ranges. Further study on structural note variation will better inform us of its merit as an acoustic metric for our species.

3.2.2 FUTURE DIRECTIONS

Machine learning techniques like latent visualization have the potential to build intuitive note categories. Our analysis, however, failed to separate Adelaide's warbler notes based on their contours (i.e., patterns of frequency modulation). Alternate image-based analyses like DTW may be more useful for separating notes in this way. In contrast, latent visualization may be more appropriate for separating notes based on their frequency characteristics. For now, we recommend latent visualization for species that do not have frequency-modulated notes in their vocalizations.

We have shared our workflow pipeline on Github in the hopes of helping other bioacousticians to implement automated methods in their own studies. Specifically, we found DTS to be a useful, time-saving algorithm for breaking songs down into their component notes. We warn, however, that parameter optimization for all algorithms used (DTS, UMAP, and fuzzy

c-means clustering) requires the input of human experts because the optimal values for each differ based on the study species.

3.2.3 SUMMARY

In summary, we used novel automated methods to generate a latent visualization of acoustic note structure in Adelaide's warblers. The latent-space representation grouped notes into one crescent-like shape and applied 14 cluster labels. According to the projected crescent, Adelaide's warbler notes differ primarily in general frequency characteristics, and discrete note types are not present. However, cluster labels seemingly disregarded note contour, which is the most important aspect of note structure for by-eye classification. Overall, the image-based analysis we used here may not be ideal for species that produce frequency-modulated notes. For future image-based analyses, we recommend comparing note contours for a more intuitive understanding of acoustic variation. We also recommend running an acoustic feature analysis alongside the image-based analysis to further validate the results of the latter.

REFERENCES

Amador, A., & Mindlin, G. B. (2023). The science of birdsong and the spectrogram, the technique that changed it all [version 1; peer review: Awaiting peer review]. *Molecular Psychology*, *2*(9). https://doi.org/10.12688/molpsychol.17520.1

Ballentine, B., Badyaev, A., & Hill, G. E. (2003). Changes in Song Complexity Correspond to Periods of Female Fertility in Blue Grosbeaks (*Guiraca caerulea*). *Ethology*, *109*(1), 55–66. https://doi.org/10.1046/j.1439-0310.2003.00852.x

Baulieu, F. B. (1989). A classification of presence/absence based dissimilarity coefficients. *Journal of Classification*, *6*(1), 233–246. https://doi.org/10.1007/BF01908601

Berba, P. (2020, July 8). *A gentle introduction to HDBSCAN and density-based clustering*. Medium. https://towardsdatascience.com/a-gentle-introduction-to-hdbscan-and-density-based-clustering-5fd79329c1e8

Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, *10*(2), 191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Bhardwaj, M. (2023, January 1). Euclidean-Distance Classifier. *Medium*. https://bhardwajmanu.medium.com/euclidean-distance-classifier-1b59345e75da

Borchia, D. (2023, February 12). *Euclidean Distance Calculator*. Omni Calculator. https://www.omnicalculator.com/math/euclidean-distance

Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Buades, A., Coll, B., & Morel, J.-M. (2011). Non-Local Means Denoising. *Image Processing On Line*, *1*, 208–212. https://doi.org/10.5201/ipol.2011.bcm_nlm

Carrasco, O. C. (2020, February 21). *Gaussian Mixture Models Explained*. Medium. https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

Christie, P. J., Mennill, D. J., & Ratcliffe, L. M. (2004). Pitch shifts and song structure indicate male quality in the dawn chorus of black-capped chickadees. *Behavioral Ecology and Sociobiology*, *55*(4), 341–348. https://doi.org/10.1007/s00265-003-0711-3

Clark, C. W., Marler, P., & Beeman, K. (1987). Quantitative Analysis of Animal Vocal Phonology: An Application to Swamp Sparrow Song. *Ethology*, *76*(2), 101–115. https://doi.org/10.1111/j.1439-0310.1987.tb00676.x

Crouch, N. M. A., & Mason-Gamer, R. J. (2019). Identifying ecological drivers of interspecific variation in song complexity in songbirds (Passeriformes, Passeri). *Journal of Avian Biology*, *50*(3). https://doi.org/10.1111/jav.02020

Dias, M. L. D. (2019). *fuzzy-c-means: An implementation of the fuzzy* c-*means clustering algorithm* (1.6.3) [Python; Computer]. https://doi.org/10.5281/zenodo.3066222

Dietrich, C., Palm, G., Riede, K., & Schwenker, F. (2004). Classification of bioacoustic time series based on the combination of global and local decisions. *Pattern Recognition*, *37*(12), 2293–2305. https://doi.org/10.1016/j.patcog.2004.04.004

dos Santos, E. B., Llambías, P. E., & Rendall, D. (2016). The structure and organization of song in Southern House Wrens (*Troglodytes aedon chilensis*). *Journal of Ornithology*, *157*(1), 289–301. https://doi.org/10.1007/s10336-015-1277-3

dos Santos, E. B., Llambías, P. E., & Rendall, D. (2018). Male song diversity and its relation to breeding success in southern house wrens *Troglodytes aedon chilensis*. *Journal of Avian Biology*, *49*(6). https://doi.org/10.1111/jav.01606

Drude, L., Heymann, J., Boeddeker, C., & Haeb-Umbach, R. (2018). NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing. *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, 216–220.

DuBois, A. L., Nowicki, S., & Searcy, W. A. (2011). Discrimination of vocal performance by male swamp sparrows. *Behavioral Ecology and Sociobiology*, *65*(4), 717–726. https://doi.org/10.1007/s00265-010-1073-2

Education Ecosystem (LEDU). (2018, September 12). *Understanding K-means Clustering in Machine Learning*. Medium. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

Elie, J. E., & Theunissen, F. E. (2016). 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. https://doi.org/10.1007/s10071-015-0933-6

Frey, D. F., & Pimentel, R. A. (1978). Principal component analysis and factor analysis. In P. W. Colgan (Ed.), *Quantitative Ethology* (pp. 219–246). John Wiley & Sons.

Fukushima, M., Doyle, A. M., Mullarkey, M. P., Mishkin, M., & Averbeck, B. B. (2015). Distributed acoustic cues for caller identity in macaque vocalization. *Royal Society Open Science*, *2*(12), 150432. https://doi.org/10.1098/rsos.150432

Geberzahn, N., & Aubin, T. (2014). Assessing vocal performance in complex birdsong: A novel approach. *BMC Biology*, *12*(1), 58. https://doi.org/10.1186/s12915-014-0058-4

Gil, D., & Slater, P. J. B. (2000). Song Organisation and Singing Patterns of the Willow Warbler, *Phylloscopus trochilus*. *Behaviour*, *137*(6), 759–782.

Goffinet, J., Brudner, S., Mooney, R., & Pearson, J. (2021). *Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires*. 33. https://doi.org/10.7554/eLife.67855

Goldwasser, M. (2019, March 26). *Unweighted Pair Group Method with Arithmetic means (UPGMA) algorithm*. Introduction to Computer Science: Bioinformatics. https://cs.slu.edu/~goldwasser/courses/slu/csci1020/2019_Spring/lectures/UPGMA/

Hedwig, D., Hammerschmidt, K., Mundry, R., Robbins, M. M., & Boesch, C. (2014). Acoustic structure and variation in mountain and western gorilla close calls: A syntactic approach. *Behaviour*, *151*(8), 1091–1120. https://doi.org/10.1163/1568539X-00003175

Hsu, C.-L. (2018, May 28). Cluster Validaty and Cluster Number Selection. *An Explorer of Things*. https://chih-ling-hsu.github.io/2018/05/28/cluster-number

IBM Cloud Education. (2020, December 7). *What is Random Forest?* IBM Cloud. https://www.ibm.com/cloud/learn/random-forest

K. Lisa Yang Center for Conservation Bioacoustics. (2019). *Raven Pro: Interactive Sound Analysis Software* (1.6.1) [Computer]. The Cornell Lab of Ornithology. https://ravensoundsoftware.com

Kaluthota, C. D., Medina, O. J., & Logue, D. M. (2019). Quantifying song categories in Adelaide's Warbler (*Setophaga adelaidae*). *Journal of Ornithology*, *160*(2), 305–315. https://doi.org/10.1007/s10336-018-01623-w

Keen, S. C., Odom, K. J., Webster, M. S., Kohn, G. M., Wright, T. F., & Araya-Salas, M. (2021). A machine learning approach for classifying and quantifying acoustic diversity. *Methods in Ecology and Evolution*, *12*(7), 1213–1225. https://doi.org/10.1111/2041-210X.13599

Keen, S. C., Ross, J. C., Griffiths, E. T., Lanzone, M., & Farnsworth, A. (2014). 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. https://doi.org/10.1016/j.ecoinf.2014.01.001

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., Bohn, K., Cao, Y., Carter, G., Cäsar, C., Coen, M., DeRuiter, S. L., Doyle, L., Edelman, S., Ferrer-i-Cancho, R., Freeberg, T. M., Garland, E. C., Gustison, M., Harley, H. E., … Zamora-Gutierrez, V. (2016). Acoustic sequences in non-human animals: A tutorial review and prospectus. *Biological Reviews*, *91*(1), 13–52. https://doi.org/10.1111/brv.12160

Khatib, F. (2016, August 28). *Newbie's Guide to ML — Part 3*. Medium. https://medium.com/ml-for-newbies/newbies-guide-to-ml-part-3-6c9ee4b616d8

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., & Willing, C. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (pp. 87–90). IOS Press.

Lachlan, R. F. (2007). *Luscinia: A bioacoustics analysis computer program* (2.16.10.29.01) [Computer].

Lachlan, R. F., Verzijden, M. N., Bernard, C. S., Jonker, P.-P., Koese, B., Jaarsma, S., Spoor, W., Slater, P. J. B., & ten Cate, C. (2013). The Progressive Loss of Syntactical Structure in Bird Song along an Island Colonization Chain. *Current Biology*, *23*(19), 1896–1901. https://doi.org/10.1016/j.cub.2013.07.057

Leitão, A., ten Cate, C., & Riebel, K. (2006). Within-song complexity in a songbird is meaningful to both male and female receivers. *Animal Behaviour*, *71*(6), 1289–1296. https://doi.org/10.1016/j.anbehav.2005.08.008

Logue, D. M., Sheppard, J. A., Walton, B., Brinkman, B. E., & Medina, O. J. (2020). An analysis of avian vocal performance at the note and song levels. *Bioacoustics*, *29*(6), 709–730. https://doi.org/10.1080/09524622.2019.1674693

MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, *1*, 281–297.

Martins, E. P. (Ed.). (1996). *Phylogenies and the Comparative Method in Animal Behavior*. Oxford University Press.

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., Niekirk, B. van, Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., … Pimenta, W. (2023). *Librosa* (0.10.0) [Python; Computer]. https://doi.org/10.5281/zenodo.7746972

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, *2*(11), 205. https://doi.org/10.21105/joss.00205

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426v3 [stat.ML]. https://doi.org/10.48550/arXiv.1802.03426

Mota, P. G., & Cardoso, G. C. (2001). Song organisation and patterns of variation in the serin (*Serinus serinus*). *Acta Ethologica*, *3*(2), 141–150. https://doi.org/10.1007/s102110000034

Nowicki, S., Westneat, M., & Hoese, W. (1992). Birdsong: Motor function and the evolution of communication. *Seminars in Neuroscience*, *4*, 385–390. https://doi.org/10.1016/1044-5765(92)90046-5

Odom, K. J., Araya-Salas, M., Morano, J. L., Ligon, R. A., Leighton, G. M., Taff, C. C., Dalziell, A. H., Billings, A. C., Germain, R. R., Pardo, M., Andrade, L. G., Hedwig, D., Keen, S. C., Shiu, Y., Charif, R. A., Webster, M. S., & Rice, A. N. (2021). Comparative bioacoustics: A roadmap for quantifying and comparing animal sounds across diverse taxa. *Biological Reviews*, *96*(4), 1135–1159. https://doi.org/10.1111/brv.12695

Page, J. T., Liechty, Z. S., Huynh, M. D., & Udall, J. A. (2014). BamBam: Genome sequence analysis tools for biologists. *BMC Research Notes*, *7*(1), 829. https://doi.org/10.1186/1756-0500-7-829

Pal, N. R., & Bezdek, J. C. (1995). On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems*, *3*(3), 370–379. https://doi.org/10.1109/91.413225

Pieplow, N. (2009, December 7). A Brief History of Spectrograms. *Earbirding*. http://earbirding.com/blog/archives/1229

Podos, J., Peters, S., Rudnicky, T., Marler, P., & Nowicki, S. (1992). The Organization of Song Repertoires in Song Sparrows: Themes and Variations. *Ethology*, *90*(2), 89–106. https://doi.org/10.1111/j.1439-0310.1992.tb00824.x

Prat, Y., Taub, M., & Yovel, Y. (2016). Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep39419

Python Software Foundation. (2023). *Python* (3.11.2) [Computer]. https://www.python.org

R Core Team. (2022). *R: A language and environment for statistical computing* (4.2.1) [Computer]. R Foundation for Statistical Computing. https://www.R-project.org

Reby, D., André-Obrecht, R., Galinier, A., Farinas, J., & Cargnelutti, B. (2006). 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. https://doi.org/10.1121/1.2358006

Reutterer, T., & Dan, D. (2022). Cluster Analysis in Marketing Research. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of Market Research* (pp. 221–249). Springer International Publishing. https://doi.org/10.1007/978-3-319-57413-4_11

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLOS Computational Biology*, *16*(10), e1008228. https://doi.org/10.1371/journal.pcbi.1008228

Saunders, A. A. (1915). Some Suggestions for Better Methods of Recording and Studying Bird Songs. *The Auk*, *32*(2), 173–183. https://doi.org/10.2307/4072426

Sharma, R. (2019, January 28). UPGMA Method: Designing a Phylogenetic Tree. *Medium*. https://medium.com/@sharma.ravit/upgma-method-designing-a-phylogenetic-tree-9a708de18419

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394–3411. https://doi.org/10.1121/1.1624067

Sneath, P. H. A., & Sokal, R. R. (1973). Numerical taxonomy. The principles and practice of numerical classification. *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* https://www.cabdirect.org/cabdirect/abstract/19730310919

Staicer, C. A. (1991). *The role of male song in the socioecology of the tropical resident Adelaide's warbler (*Dendroica adelaidae*)* [PhD dissertation]. University of Massachusetts Amherst.

Starmer, J. (Director). (2016, July 10). *StatQuest: Linear Discriminant Analysis (LDA) clearly explained*. StatQuest. https://www.youtube.com/watch?v=azXCzI57Yfc

Stoddard, P. K., Beecher, M. D., & Willis, M. S. (1988). Response of territorial male song sparrows to song types and variations. *Behavioral Ecology and Sociobiology*, *22*(2), 125–130. https://doi.org/10.1007/BF00303547

Swiston, K. A., & Mennill, D. J. (2009). Comparison of manual and automated methods for identifying target sounds in audio recordings of Pileated, Pale-billed, and putative Ivory-billed woodpeckers. *Journal of Field Ornithology*, *80*(1), 42–50. https://doi.org/10.1111/j.1557-9263.2009.00204.x

Tenenbaum, J. B., Silva, V. de, & Langford, J. C. (2000). A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, *290*(5500), 2319–2323. https://doi.org/10.1126/science.290.5500.2319

Thomas, M. (2021). *Tutorial for generating and evaluating latent-space representations of vocalizations using UMAP* (1.0) [Jupyter Notebook; Computer]. https://doi.org/10.5281/zenodo.5767842

Thomas, M., Jensen, F. H., Averly, B., Demartsev, V., Manser, M. B., Sainburg, T., Roch, M. A., & Strandburg-Peshkin, A. (2022). A practical guide for generating unsupervised, spectrogram-based latent space representations of animal vocalizations. *Journal of Animal Ecology*, *00*, 1–15. https://doi.org/10.1111/1365-2656.13754

Toms, J. D. (2020). Adelaide's Warbler (*Setophaga adelaidae*), version 1.0. In T. S. Schulenberg (Ed.), *Birds of the World*. Cornell Lab of Ornithology. https://doi.org/10.2173/bow.adewar1.01

Unzueta, D. (2021, December 22). *Fisher's Linear Discriminant: Intuitively Explained*. Medium. https://towardsdatascience.com/fishers-linear-discriminant-intuitively-explained-52a1ba79e1bb

Vaca-Castaño, G., & Rodriguez, D. (2010). Using syllabic Mel cepstrum features and k-nearest neighbors to identify anurans and birds species. *2010 IEEE Workshop On Signal Processing Systems*, 466–471. https://doi.org/10.1109/SIPS.2010.5624892

Wadewitz, P., Hammerschmidt, K., Battaglia, D., Witt, A., Wolf, F., & Fischer, J. (2015). Characterizing Vocal Repertoires—Hard vs. Soft Classification Approaches. *PLOS ONE*, *10*(4), e0125785. https://doi.org/10.1371/journal.pone.0125785

Walsh, S. L., Engesser, S., Townsend, S. W., & Ridley, A. R. (2023). Multi-level combinatoriality in magpie non-song vocalizations. *Journal of The Royal Society Interface*, *20*(199), 20220679. https://doi.org/10.1098/rsif.2022.0679

Walt, S. van der, Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., & Yu, T. (2014). scikit-image: Image processing in Python. *PeerJ*, *2*, e453. https://doi.org/10.7717/peerj.453

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wirth, M., Estabrook, G. F., & Rogers, D. J. (1966). A Graph Theory Model for Systematic Biology, with an Example for the Oncidiinae (Orchidaceae). *Systematic Biology*, *15*(1), 59–69. https://doi.org/10.2307/sysbio/15.1.59

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), Article 10. https://doi.org/10.1038/nn1536

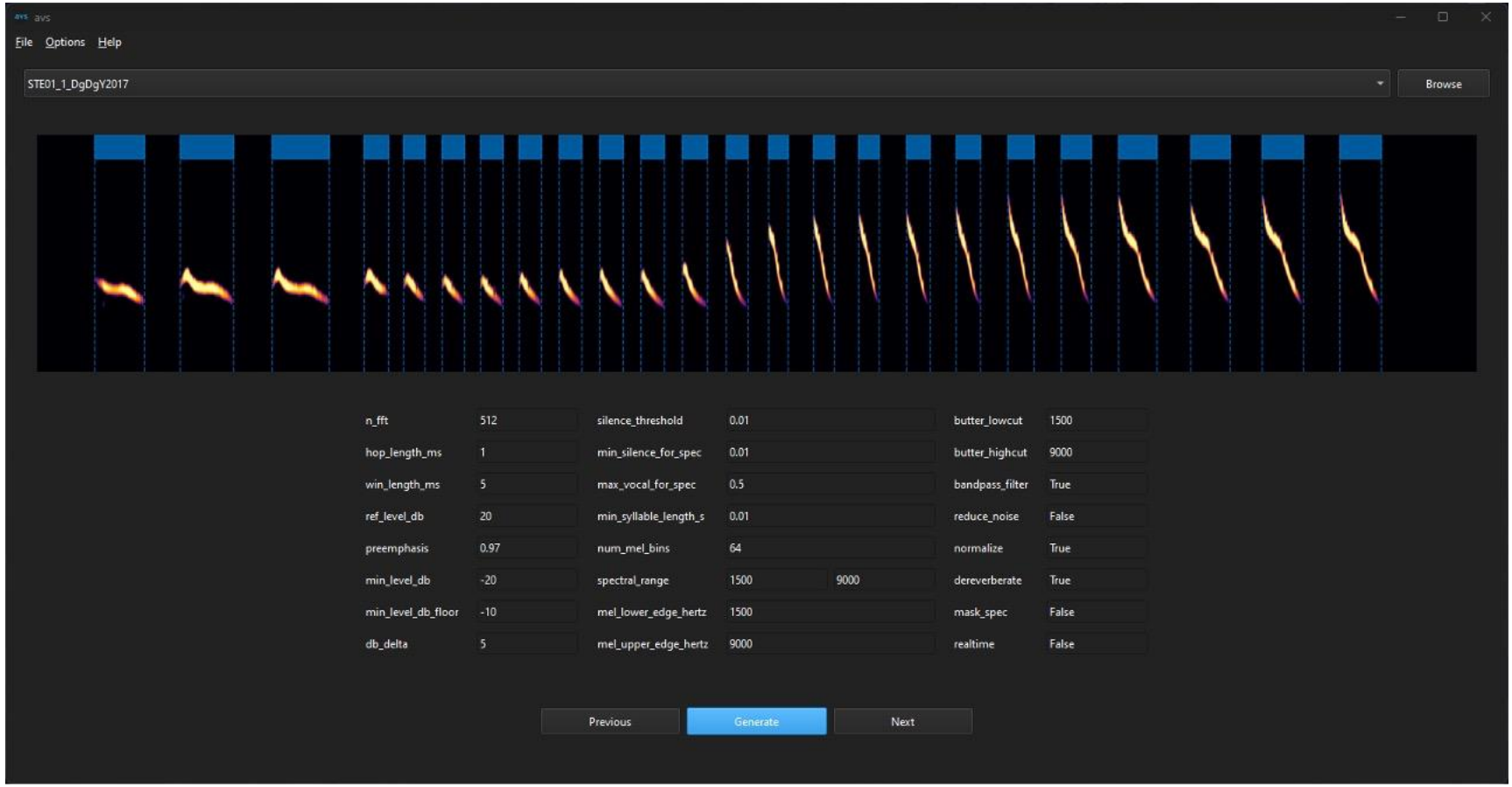
Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(8), 841–847. https://doi.org/10.1109/34.85677

Yang, X. (2020, May 9). *Linear Discriminant Analysis, Explained*. Medium. https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b

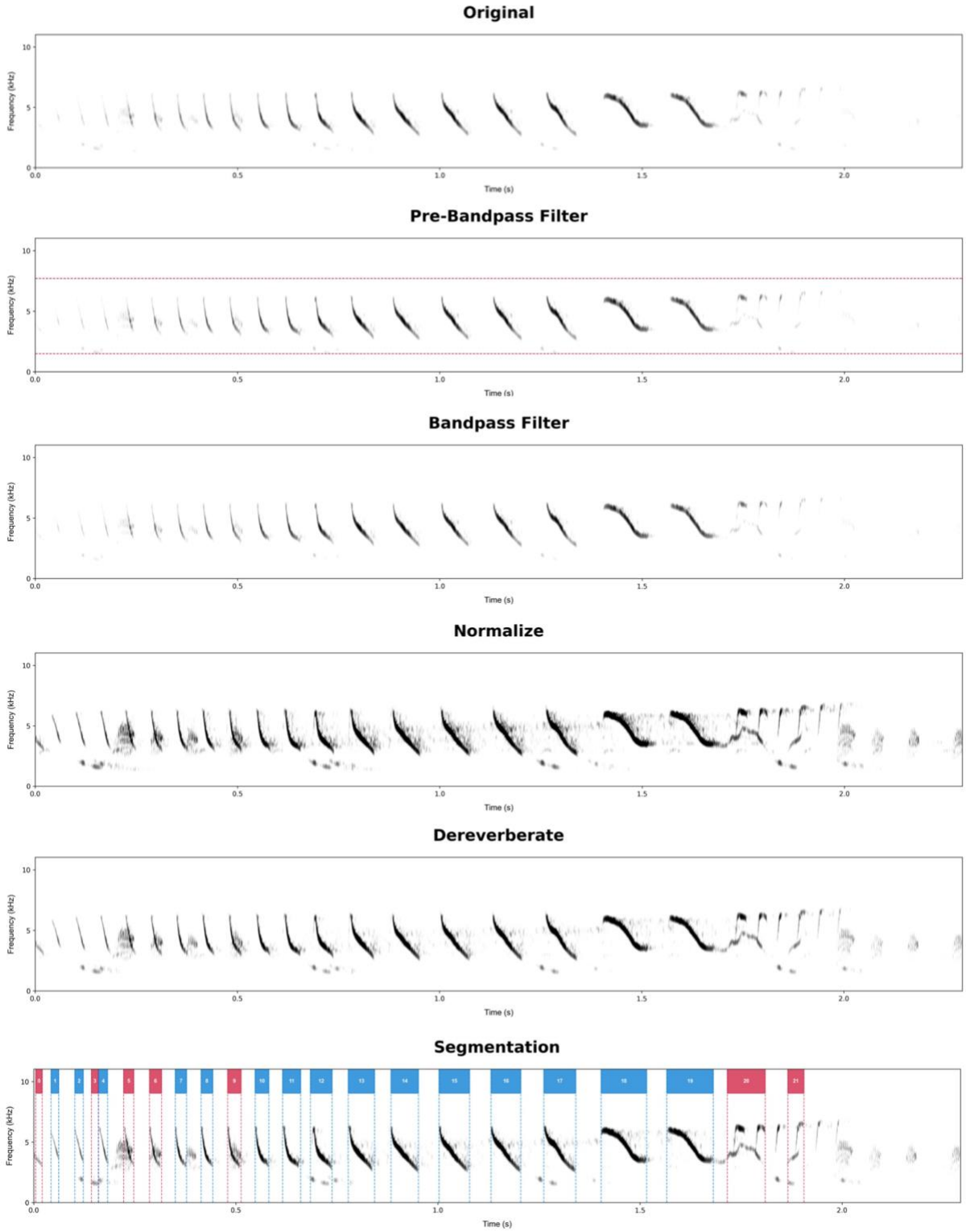
Yufeng. (2021, November 10). *Fuzzy C-Means Clustering with Python*. Medium. https://towardsdatascience.com/fuzzy-c-means-clustering-with-python-f4908c714081

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

APPENDIX 1. VISUAL OF ANIMAL VOCALIZATION SEGMENTATION (AVS)

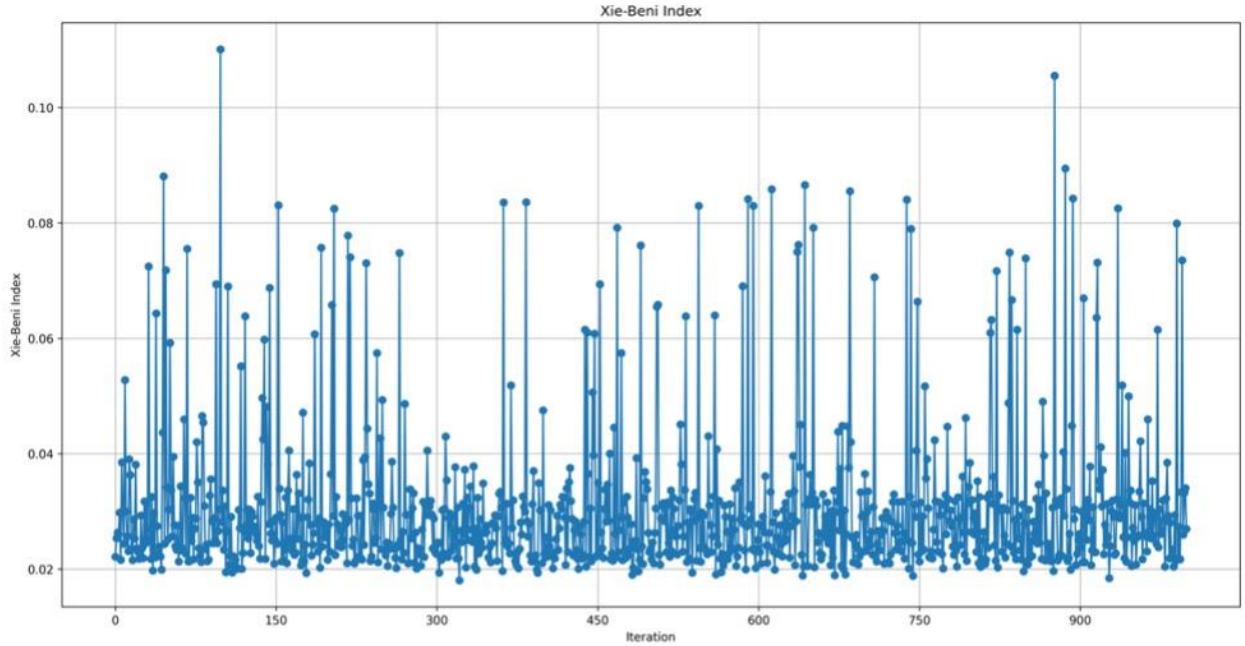


APPENDIX 2. SAMPLE WORKFLOW OF SPECTROGRAM PRE-PROCESSING



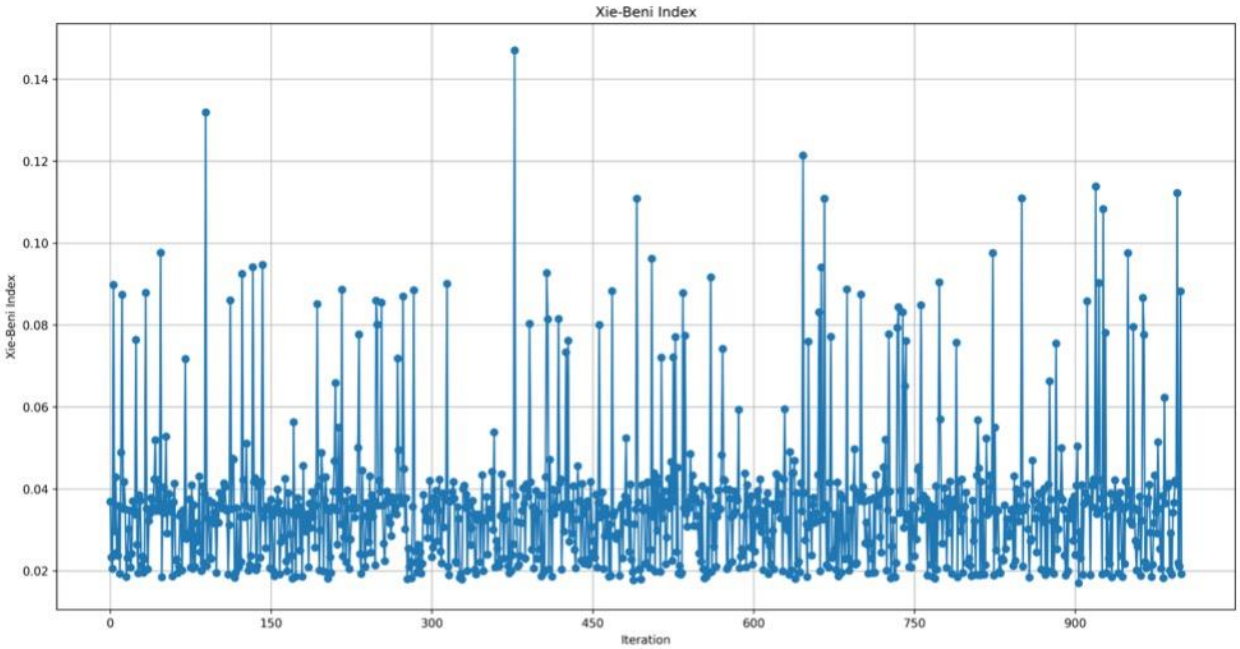
Appendix 2. Spectrograms are filtered and segmented in preparation for latent visualization. Excluded segments are in red.

APPENDIX 3. CLUSTER VALIDATION USING LARGE SUBSETS



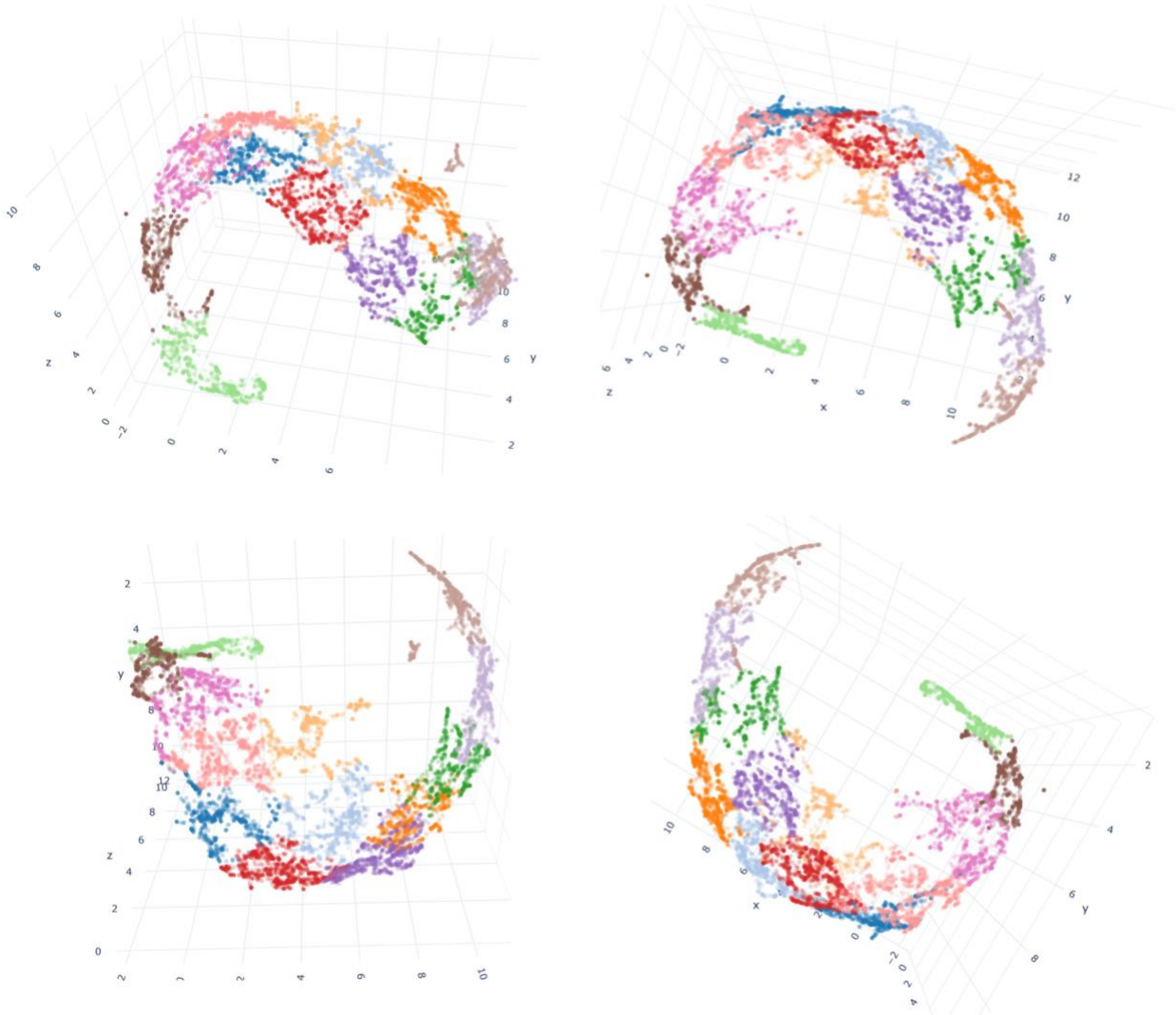
Appendix 3. The value of the Xie-Beni index remains fairly consistent throughout all iterations.

APPENDIX 4. CLUSTER VALIDATION USING SMALL SUBSETS

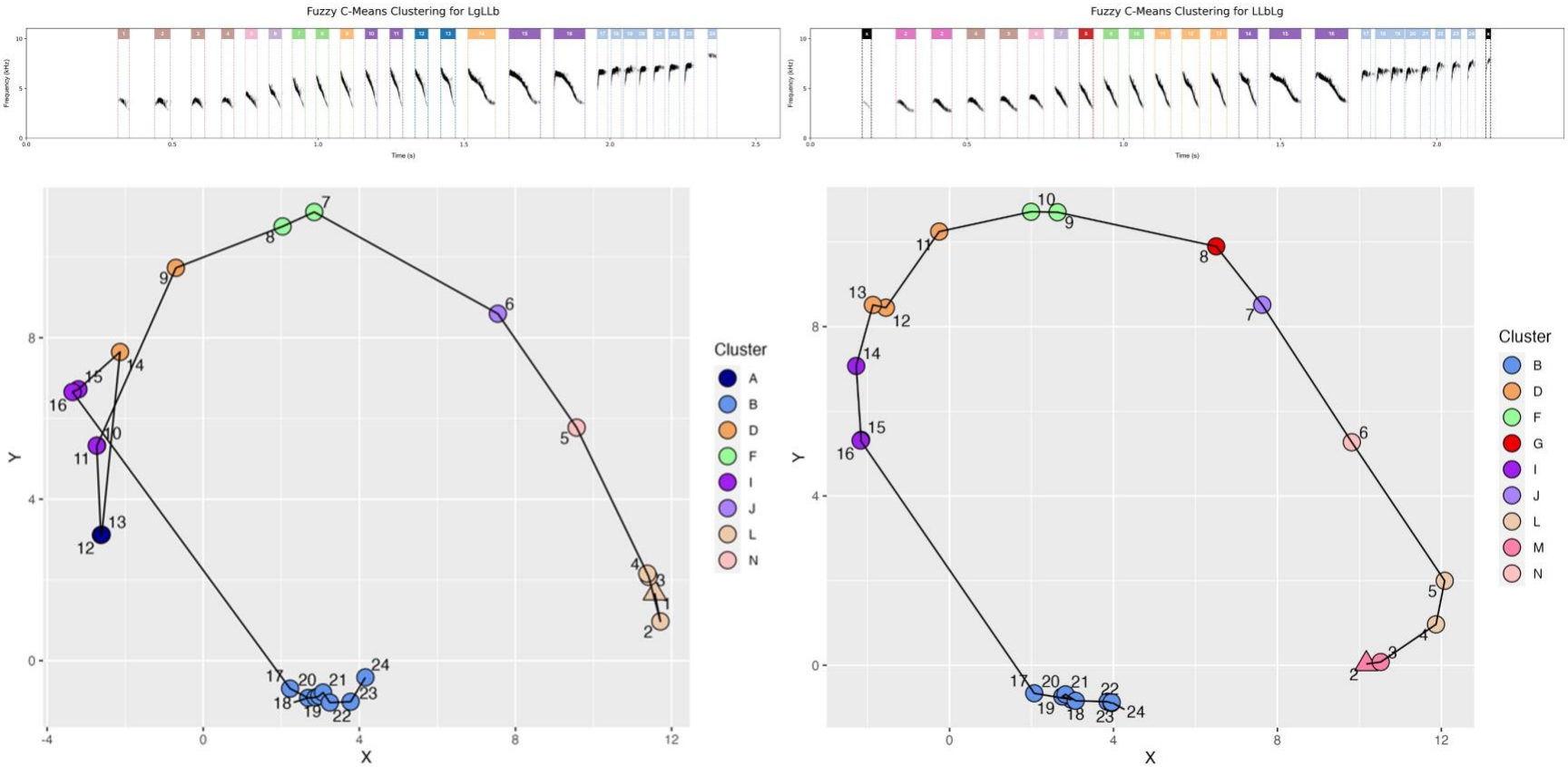


Appendix 4. The value of the Xie-Beni index remains fairly consistent throughout all iterations.

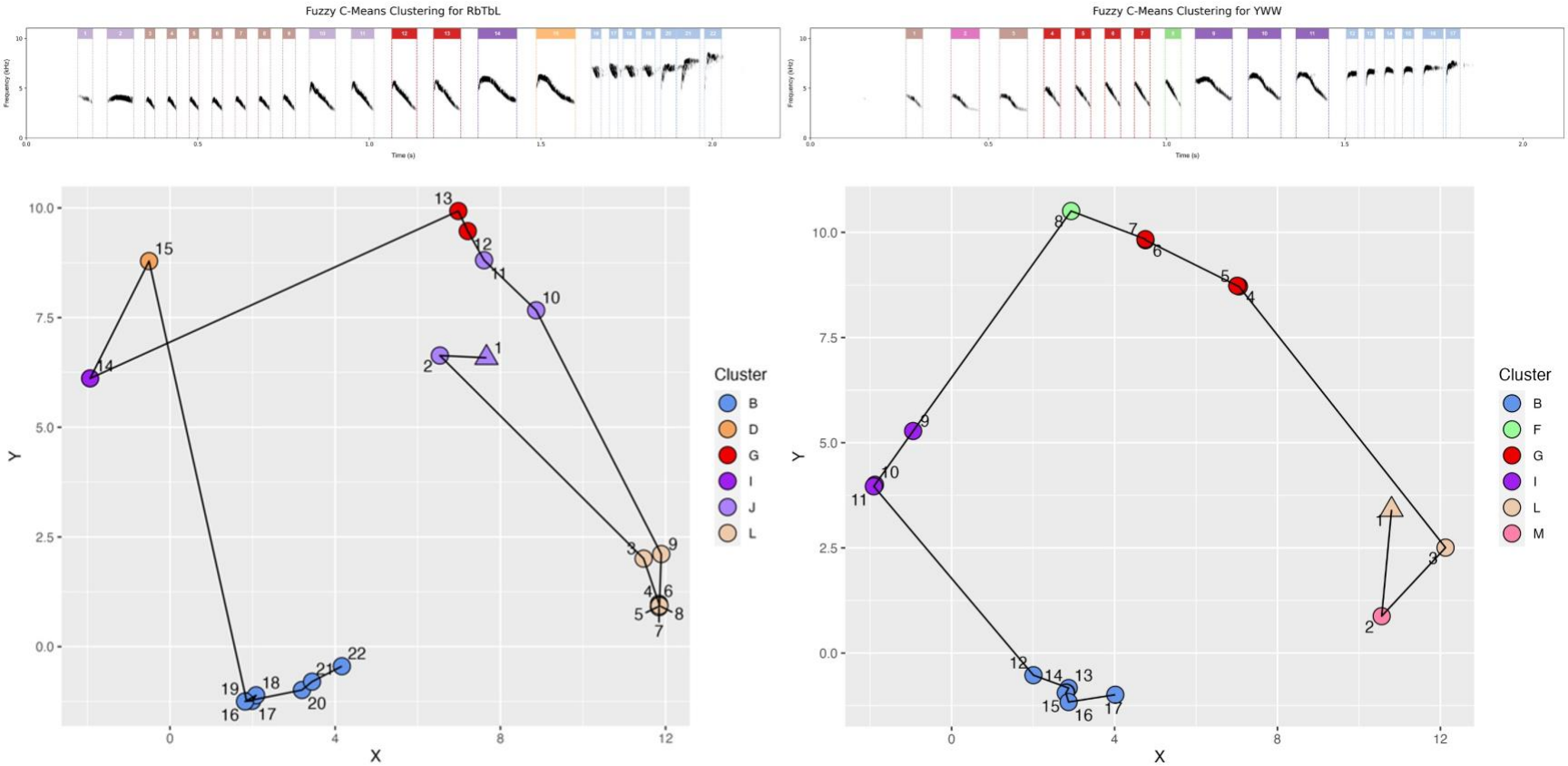
APPENDIX 5. IMAGES OF THE 3D CLUSTERED LATENT-SPACE REPRESENTATION FROM DIFFERENT ANGLES



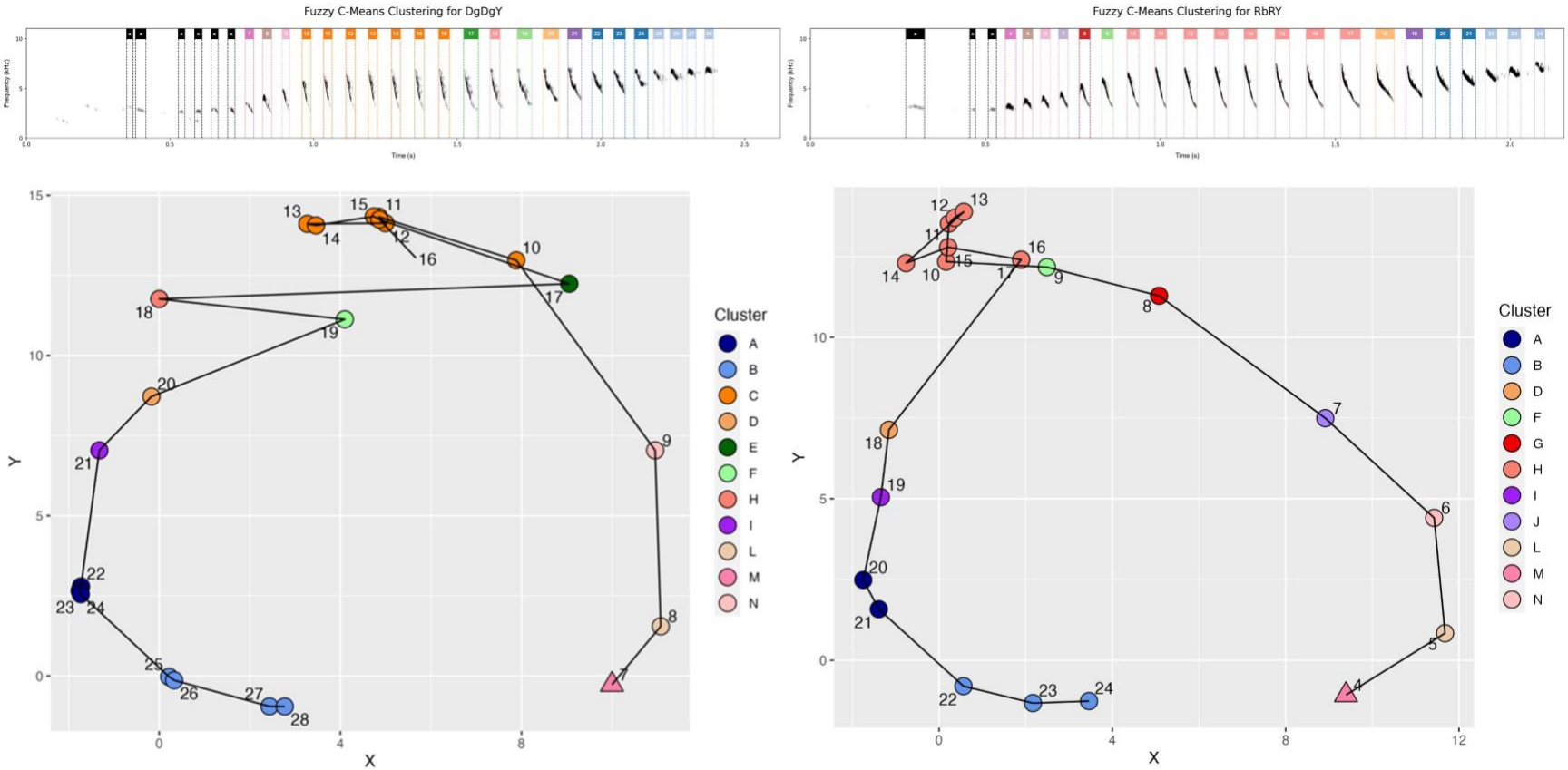
APPENDIX 6A. ADDITIONAL SONG TRACES FOR THE SONG TYPE IN FIGURE 2.10



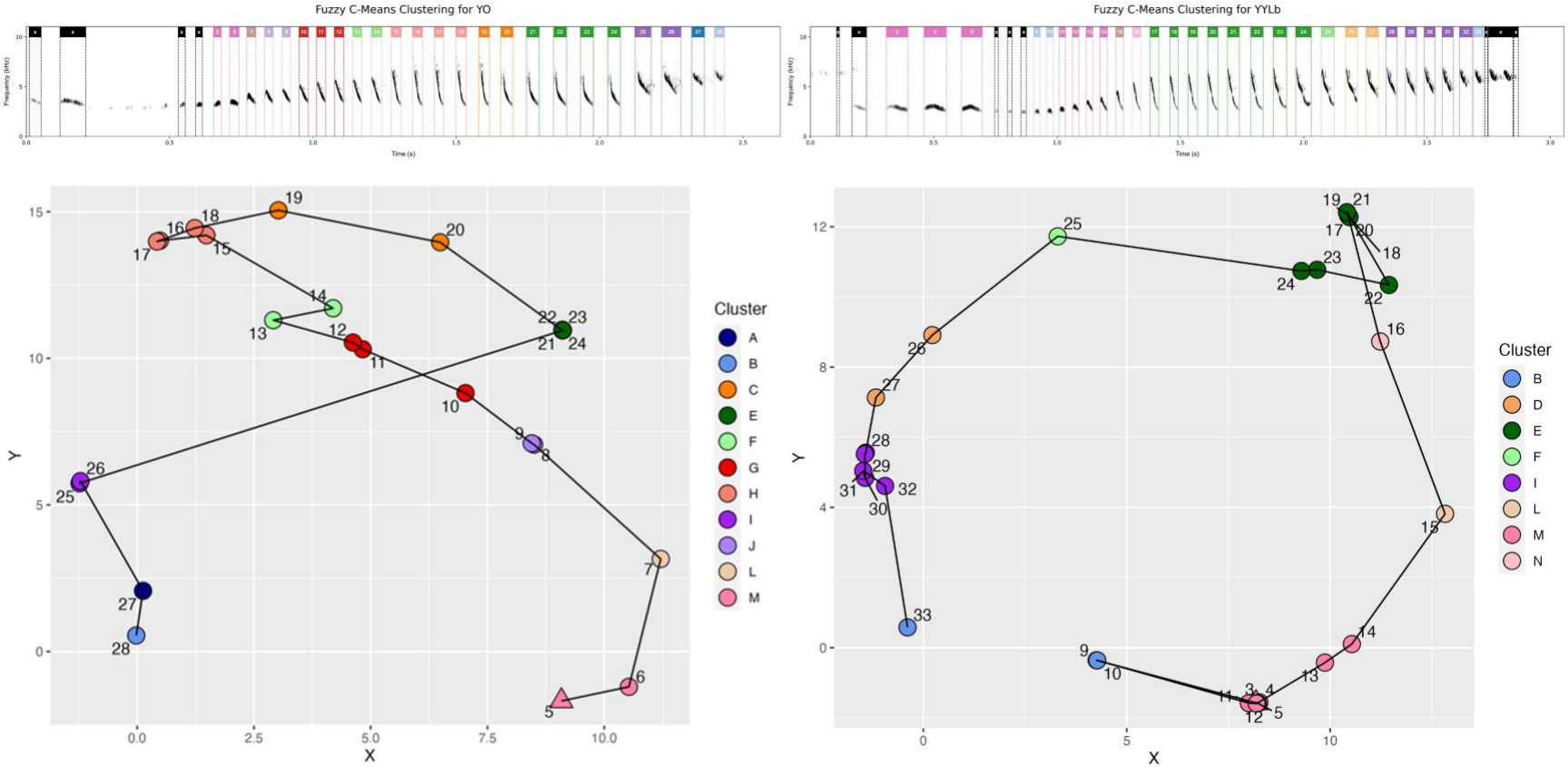
APPENDIX 6B. ADDITIONAL SONG TRACES FOR THE SONG TYPE IN FIGURE 2.10



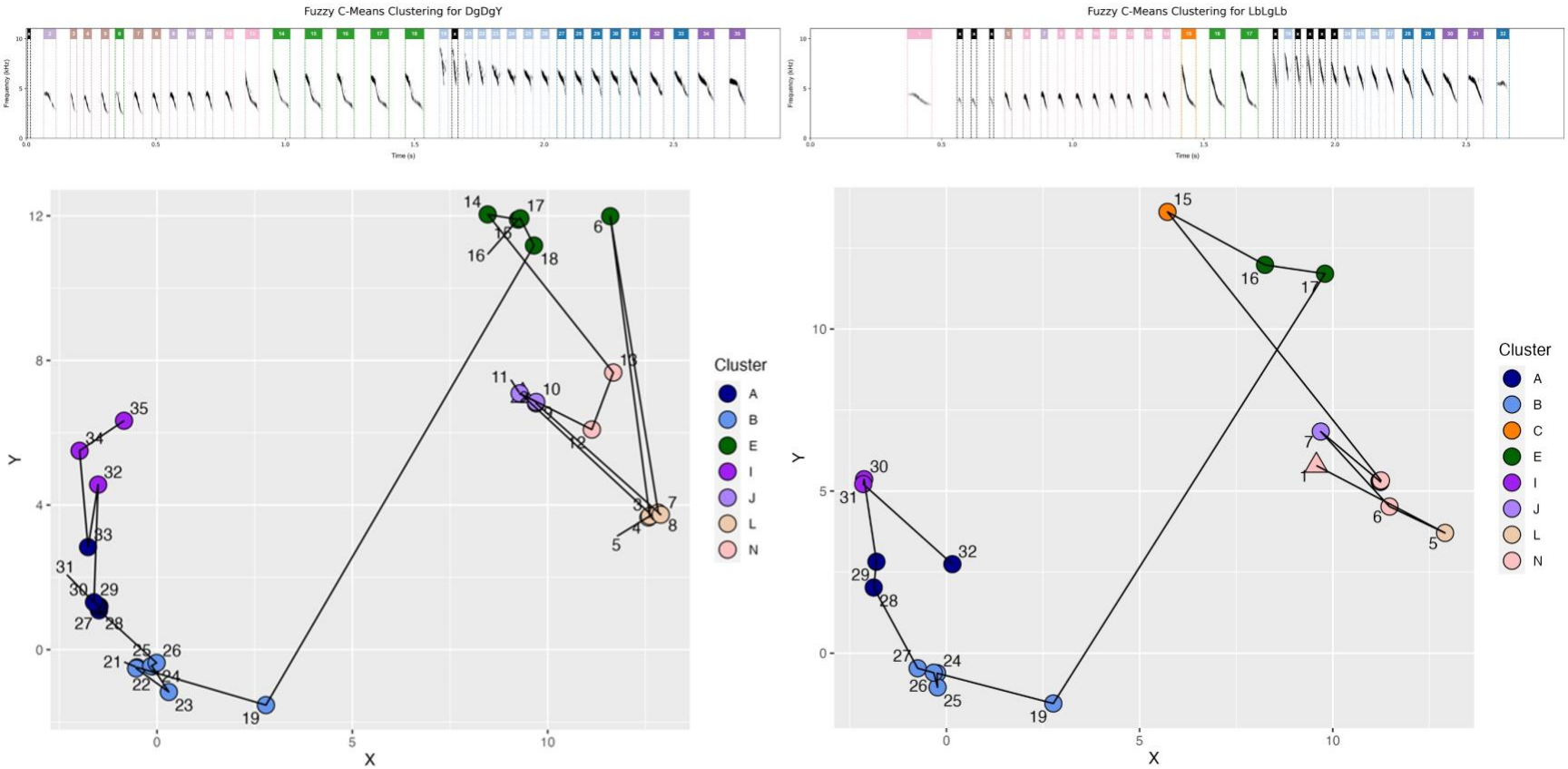
APPENDIX 7A. ADDITIONAL SONG TRACES FOR THE SONG TYPE IN FIGURE 2.11



APPENDIX 7B. ADDITIONAL SONG TRACES FOR THE SONG TYPE IN FIGURE 2.11



APPENDIX 8A. ADDITIONAL SONG TRACES FOR THE SONG TYPE IN FIGURE 2.12



APPENDIX 8B. ADDITIONAL SONG TRACES FOR THE SONG TYPE IN FIGURE 2.12