Evaluating the Performance of Topic Modeling for Humpback Whale Song Phrase Segmentation

Franny Oppenheimer

Mentors: John Ryan, Danelle Cline, Duane Edgington

Summer 2025

Keywords: Humpback whale song, passive acoustic monitoring, unsupervised machine learning, bioacoustics

ABSTRACT

Each population of humpback whales produces a complex and unique pattern of calls known as song, of which components can be shared between populations via cultural transmission. Connectivity between populations can be quantified by cataloguing the song of each population over a number of years and looking for shared components, though this work requires individuals to extract, classify, and compare song components, and is therefore both subjective and extremely time-intensive. Topic modeling is an unsupervised machine learning technique which, when applied to whale song, is able to consistently delineate song units - this paper describes both efforts to coerce topic modeling into classification at the more commonly-used phrase level and to evaluate Adjusted Rand Index as a metric for model optimization. The findings presented here indicate the potential capability of the model to delineate song phrases, though presently with a lack of accuracy, and the subsequent need for more exploration of alternative clustering and discretization techniques.

1. INTRODUCTION

Male humpback whales are known to sing complex songs at both their low-latitude breeding areas and their high-latitude feeding grounds [1], [2], which are hypothesized and widely accepted to be a form of lekking behavior, as juvenile males do not begin to sing until they reach reproductive maturity [3]. There are 14 distinct population segments of humpback whales worldwide, each of which sings its own unique song [4]. These songs are known to change over time, a process known as cultural transmission, both via slow-growing "evolutionary" change thought to be caused by individual variation becoming incorporated into the group's repertoire, and via quick "revolutionary" change thought to be caused by interaction between populations and adoption of song fragments of the other population's song [5], [6]. This revolutionary change, involving huge amounts of song being replaced over short periods of time, can potentially be used to track the predicted climate-induced increase in migration route overlap and subsequently increased population connectivity, particularly between populations within the same ocean basin [6], [1].

There has been much work done in efforts to categorize and track song change within and between populations of humpback whales, to examine both connectivity between populations and the response of singers to environmental change (e.g. [5], [1]). Song has historically been split into groups of repeating patterns, with an entire song cycle made up of repeating themes, themes made up of repeating phrases, and phrases made up of repeating units [7]. Because units lack standardized labeling across studies but are encapsulated within phrase designations and themes typically consist of a single repeating phrase [6], most comparison between and within populations is done at the phrase level. There is currently a large-scale effort being made by the WhaleTrust organization to categorize the phrase repertoire of all humpback populations within the North Pacific basin, deemed the North Pacific Songs Project (NPSP). As part of the NPSP, recordings from Monterey Bay have been parsed through for phrase identification, and a particular phrase of some interest has been identified. Known as Olive, this phrase first emerged in the Hawaiian population during the 2022 season, and has since been recorded from the 2023 season in sites to the east, including Monterey Bay [Oppenheimer pers. comm . 2025]. This supports some degree of connectivity across the populations of

the North Pacific, but raises questions regarding both the origin of this specific phrase, and the unique qualities that have allowed for its rapid transmission.

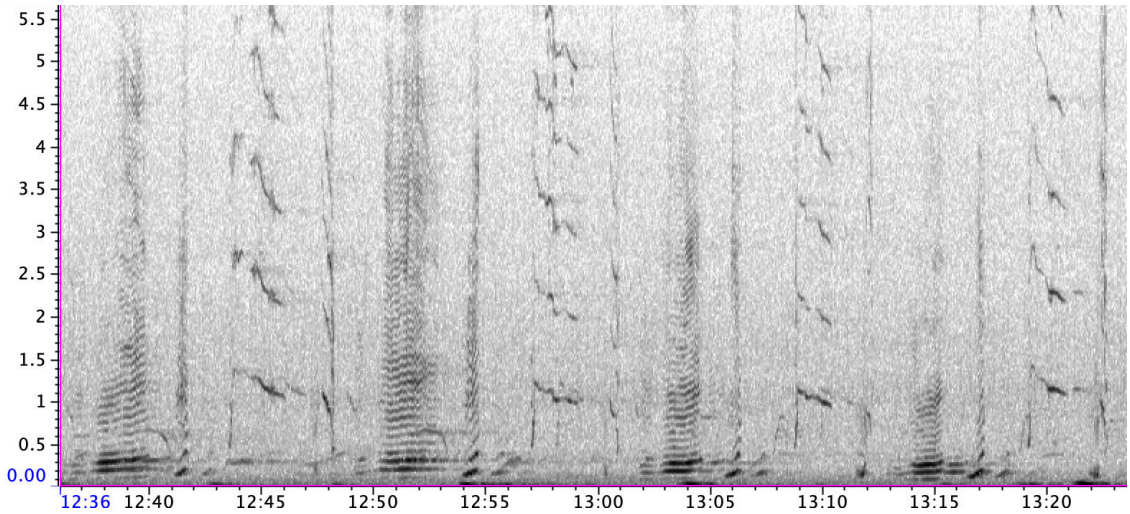

Figure 1. Spectrogram of Olive as it appeared in a recording from November 28, 2023 from the MARS cabled observatory in Monterey, California (original image courtesy of Jim Darling).

In order to address the questions around Olive, and for the NPSP as a whole to be completed, researchers have to sift through hundreds of hours of sound files and identify phrases by hand. This involves not only a significant level of subjectivity in the classification itself, but also a vast energy and time cost. Though efforts have been made to reduce subjectivity (e.g. [8], [9]) and to automate song detection (e.g. [10]), the need for an automated method of classification when it comes to long-term studies with high volumes of data has become apparent. Probabilistic topic modeling, a form of unsupervised natural language processing (NLP) used to identify and assign underlying topics in suites of textual data [11], is a promising frontier in the field of automated classification. By depending solely on a model for data processing and grouping, one eliminates both human subjectivity in situations of ambiguity and the extensive time input requirement associated. The application of topic modeling to humpback whale song has been explored in the past, at the level of unit classification [13], but has not yet been applied to the more pressing matter of phrase classification. This paper describes an

attempt at modeling at the phrase level, including the integration of two metrics for parameter optimization and the eventual application of the model to data recorded in Monterey Bay, California.

2. MATERIALS AND METHODS

2.1 METRICS

Previous work has introduced perplexity, a measure of model confidence, and coherence, a measure of intra-topic similarity as potential metrics for quantifying the efficacy of topic modeling for whale song (e.g. [13], [14]) . While these techniques are valuable in evaluating model function, they do not assess the accuracy of the model in assigning the correct topic groupings in the widely-accepted format of song structure (i.e. at the cycle, theme, and phrase levels). The Adjusted Rand Index (ARI), utilized in this study to provide a metric for model accuracy against a groundtruth, is a modification of the Rand Index, which was developed in 1971 for cluster validation [15]. ARI allows for the comparison of two sets of clusters (in this case one experimental and one groundtruth) without requiring the same number of clusters, and without the need for the values within the clusters to match [16]. ARI ranges between -1 and 1, with anything below 0 indicating a worse-than-random performance, and 0 as the expected value (random) [16]. ARI can be calculated using the following equation:

$$ARI = \frac{(\frac{n}{2})\,(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{(\frac{n}{2})^2 - [(a+b)(a+c)+(c+d)(b+d)]}$$

Where $\frac{n}{2}$ is the total number of possible combinations in pairs, and a - d are combinations of objects in all possible paired formations (within or between clusters) [16].

We also continued to use perplexity as a metric of model performance in conjunction with ARI, opting for a per-word perplexity score averaged over all documents (*D)*, for

each file in each run where W_a is the number of words in document d [15]. As used here, perplexity can be defined as:

$$Perplexity = \frac{\Sigma_{d \in D} exp(-\frac{\Sigma_{w \in d} logP(w|d)}{W_d})}{D}$$

2.2 PREPROCESSING

A high-quality recording from November 28, 2023 containing the olive phrase was identified from a dataset recorded by the hydrophone mounted on the MARS cabled observatory, stationed 900m deep in Monterey Bay. The recording was decimated from its original 256 kHz sampling rate to 16 kHz, given the sampling rate of the eventual target WhaleTrust files. Each file was converted to spectrogram form using 50%-overlap fast Fourier transforms (FFTs), and each spectrogram was then truncated to a frequency subset of 50-8000 Hz. Because we believed it necessary for the model to be able to differentiate between various units in order to correctly group phrases, parameters related to spectrogram preprocessing and discretization were not changed from the optimized (lowest perplexity) values presented in Bergamaschi (2018), with the notable exception of the replacement of a Gaussian filter with per-channel energy normalization (PCEN) for better noise suppression and cleaner spectrogram quality [17]. The spectrogram was then normalized in time and frequency by subtracting the mean and dividing by the standard deviation. The output FFT frames from preprocessing were then clustered using the mini-batch k-means algorithm [18] with the value of k determined in Bergamaschi 2018. The topic model itself utilized latent Dirichlet allocation (LDA), which assumes each document (d $\in$ D) to be composed of words (wi $\in$ d) which themselves belong to a mixture of topics (zi) from a distribution thereof (θd). This can be represented by the following equation, where α controls the sparsity of topics per document and β controls the sparsity of words per document [20]:

$$P(w,\ z,\ \theta,\ \phi \mid a,\ \beta)\ =\ P(\phi|\beta)P(\theta|a)P(z|\theta)P(w|\phi_z$$

A temporally smoothed variant of this method was used to include consideration of a temporal neighborhood to account for the continuous time-series nature of the data [20], and a collapsed Gibbs sampler was used to approximate the end of the file, which is difficult to compute directly [19].

2.3 PARAMETER SWEEPS

Groundtruths were created by splitting up the recording of interest by song cycle in RavenPro 1.6 (with each sub-file containing one song cycle) [21], creating a dataset of five files for the model to run on, and manually assigning phrase and background identifications. An initial sweep of Dirischlet parameters, including the words per document (α), words per topic (β), topic growth parameter (ɣ), and the depth of the temporal neighborhood in cells (g time) were performed to isolate the optimal ranges of each parameter (0-0.2 for alpha, beta, and gamma, and 1-20 for g time), and then another sweep of 30 values for each parameter within its optimal range with three trials per parameter set was performed on the five dataset files. ARI was calculated for each file in each parameter set and averaged across the three trials, using the associated groundtruth file. Perplexity was also recorded for each parameter set, averaged across each trial and each file.

RESULTS

The sweeps of alpha, beta, gamma, and g time yielded an ARI range of 0.0474 to 0.0817 and a perplexity range of 7.7877 to 20.3599. There was no significant relationship between ARI and perplexity (Spearman's p-value 0.12379) (Figure 2), nor significant difference in average ARI between individual files across all trials of all runs (Figure 3). The maximized ARI resulted from an alpha value of 0.0012106, a beta value of 0.0130550464885756, a gamma value of 0.01539759656, and a g time value of 4 (Figure

4). The minimized perplexity results resulted from an alpha value of 0.03832093, a beta value of 0.00111734, a gamma value of 0.00240823, and a g time value of 4 (Figure 5). The only parameter with a significant relationship to ARI was beta, with a Spearman's rho of -0.42334 and a p-value of 0.02056 (Figure 6).

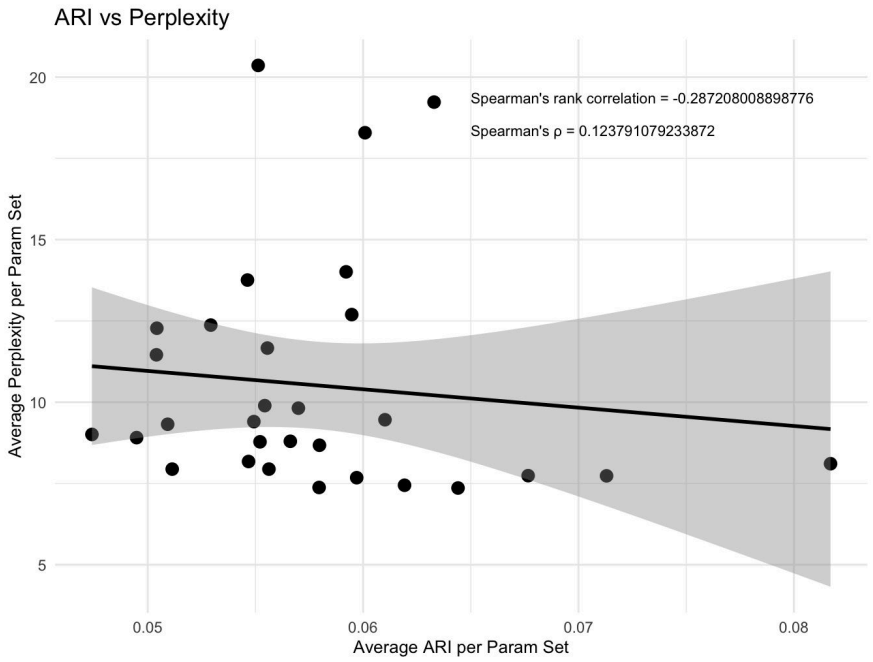

Figure 2. Plotted relationship between Adjusted Rand Index (ARI) and perplexity. No significant relationship was observed, with a Spearman's rho of -0.28 and a Spearman's p of 0.1238.

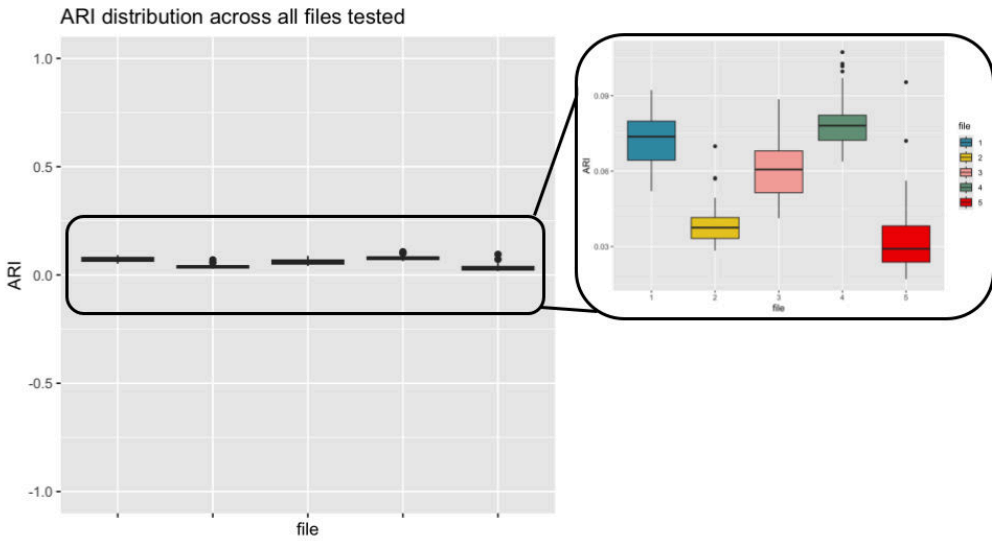

Figure 3. Adjusted Rand Index (ARI) distribution across the five files tested, averaged across 30 runs of varying parameter values with 3 trials per run. All files had an ARI above 0, with a maximum of 0.0817. No significant difference in ARI distribution was noted between the files.

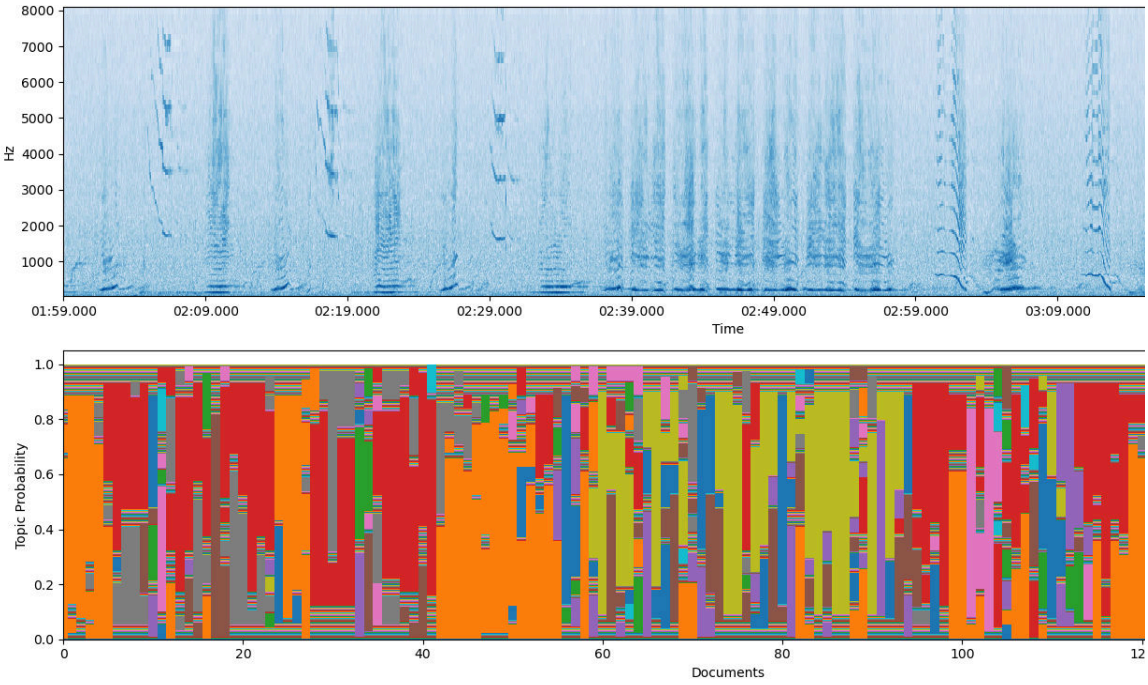

Figure 4. Visualization of one target file created from a run of the parameters demonstrating the maximum Adjusted Rand Index value (alpha value of 0.0012106, beta value of 0.0130550464885756, gamma value of 0.01539759656, g time value of 4).

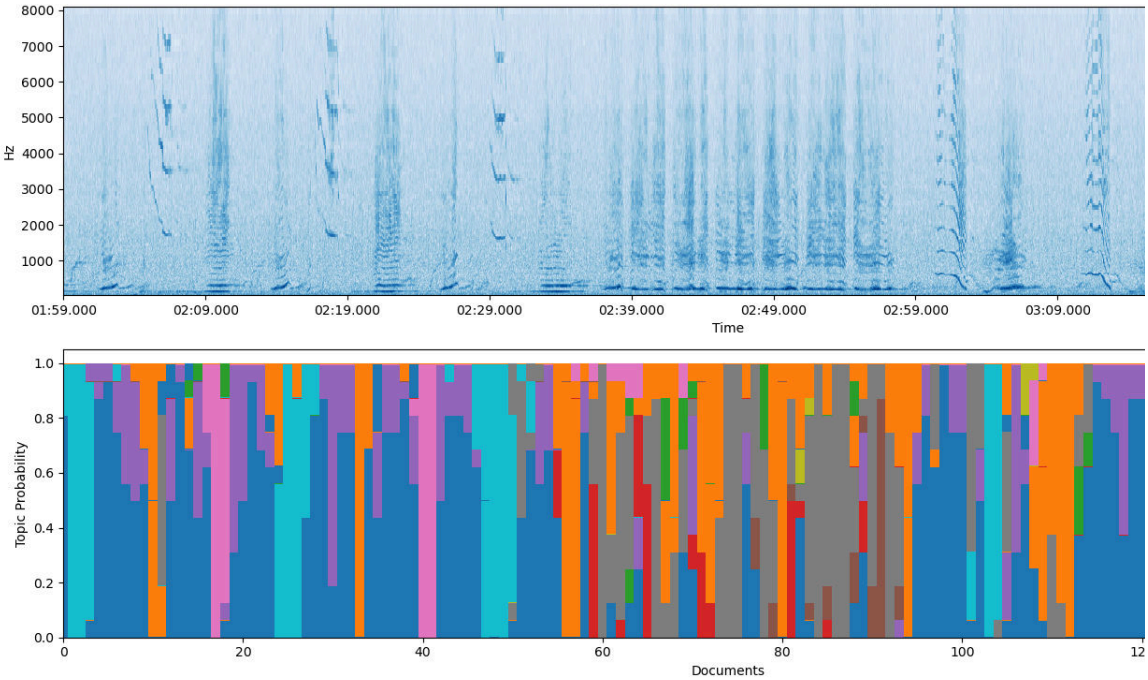

Figure 5. Visualization of one target file created from a run of the parameters demonstrating the minimum perplexity value (alpha value of 0.03832093, beta value of 0.00111734, gamma value of 0.00240823, g time value of 4).

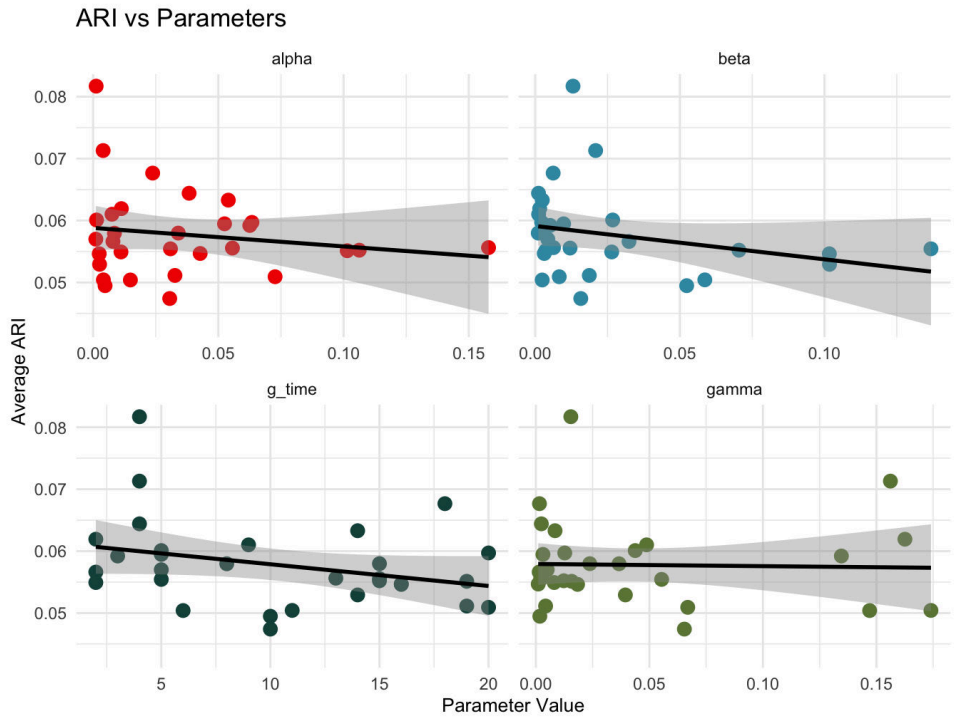

Figure 6. Words per document (alpha), words per topic (beta), topic growth parameter (gamma), and the depth of the temporal neighborhood in cells (g time) plotted against Adjusted Rand Index (ARI). No significant relationship was observed in any parameter except beta, with a Spearman's p-value of 0.02056.

DISCUSSION

The results of the Dirichlet parameter sweeps indicate that beta was the only parameter with a significant relationship to ARI, displaying a correlation of -0.42334 (Figure 5). Because beta represents words per document, this suggests an issue with the initial clustering. Work on this project was completed under the assumption that the clustering parameters from the lowest-perplexity parameter sweep (as generated in [16]) were optimal. This may not, in fact, be the case, and further work involving testing of different methods is warranted.

G time was consistent for both the maximized ARI parameter sweep and the minimized perplexity parameter sweep (value of 4), potentially indicating an optimized value for that parameter - however, with no significant relationship between g time and ARI, this is difficult to say with certainty. The visualization of the parameter set for maximized ARI

displays some recognition of multiple different units (both flat contours and upsweeps) within the same ground truth phrase as the same topic (Figure 4), though with high levels of uncertainty indicated by the sheer number of possible topic assignments, and some conflation with units assigned to different ground truth phrases. While this demonstrates potential for the model to begin clustering at the phrase level, it also indicates that topic designations are still not consistently or concretely recognizing ground truth phrases as topics.

CONCLUSIONS

While we began to see clustering at the desired hierarchical level, further work is necessary to achieve confidence in the model's ability. Due to the fact that there appeared to be no significant relationship between ARI and perplexity (further evidenced the high perplexity on the maximized ARI parameter sweep and the low ARI on the minimized perplexity parameter sweep), it will likely prove difficult to optimize both in future. Creating a weighted metric that encompasses both will be crucial to our understanding of model function. Additionally, the initial clustering parameters used in this study were assumed to be optimal due to their minimized perplexity score and visually-confirmed accuracy against ground truth unit selections. However, these parameters were not swept and compared using ARI as the key metric given the groundtruth unit designations, which would be a logical next step in the case that further work maintains the assumption that units must be recognized before phrases can be precisely delineated.

Because the Dirichlet parameters largely appeared to have insignificant effects on the ARI score of the model, another next step is to test the clustering parameters, including the vocabulary size for clustering and the clustering type - we are beginning to see some preliminary results with differing numbers of Mel bins [Oppenheimer pers. comm.]. We are also hoping to explore the use of other types of audio-recognition models, such as the Google multi-species model and/or the OpenAI Whisper model, to create clustered input for discretization.

ACKNOWLEDGEMENTS

I want to thank Dr. John Ryan, Dr. Duane Edgington, Dr. Carlos Rueda, and especially Danelle Cline for their endless support, without which this project would not have been possible. Thank you to the K. Lisa Yang Center for Conservation Bioacoustics for their RavenPro software, and to the WARP Lab at Woods Hole Oceanographic Institute for their Realtime Online Spatiotemporal Topic modeling (ROST) CLI, both of which were used extensively in this project. The MBARI Summer Internship Program is generously supported through a gift from the Dean and Helen Witter Family Fund and the Rentschler Family Fund in memory of former MBARI board member Frank Roberts (1920-2019) and by the David and Lucile Packard Foundation. Additional funding is provided by the Maxwell/Hanrahan Foundation.

References:

[1] H.E. Winn, T.J. Thompson, Cummings, W. C., Hain, J., Hudnall, J., Hays, H., & Steiner, W. W., "Song of the humpback whale—population comparisons," *Behavioral ecology and sociobiology*, 8, 41-46, 1981.

[2] A.K. Stimpert, L.E. Peavey, A.S. Friedlaender, & D.P. Nowacek. "Humpback whale song and foraging behavior on an Antarctic feeding ground," *PLoS One*, 7(12), e51214, 2012.

[3] L.M. Herman, "The multiple functions of male song within the humpback whale (Megaptera novaeangliae) mating system: review, evaluation, and synthesis," *Biological Reviews*, 92(3), 1795–1818, 2017.

[4] S. Bettridge, C.S. Baker, J. Barlow, P.J. Clapham, M. Ford, D. Gouveia, D.K. Mattila, R.M. Pace III, P.E. Rosel, G.K. Silber, P.R. Wade, "Status review of the humpback whale (Megaptera novaeangliae) under the endangered species act," *NOAA* -TM-NMFS-SWFSC-540. U.S. Department of Commerce. 2015.

[5] J.D. Darling, J.M.V. Acebes, O. Frey., R. Jorge Urbán, & M. Yamaguchi, "Convergence and divergence of songs suggests ongoing, but annually variable, mixing of humpback whale populations throughout the North Pacific," *Scientific Reports*, 9(1), 7002, 2019.

[6] J.N. Schulze, J. Denkinger, J. Oña, M.M. Poole, & E.C. Garland, "Humpback whale song revolutions continue to spread from the central into the eastern South Pacific," *Royal Society Open Science*, 9(8), 220158, 2022.

[7] R.S. Payne & S. McVay, "Songs of Humpback Whales: Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans," *Science*, 173(3997), 585–597, 1971.

[8] F. Oppenheimer. "A Description of the Song Unit Repertoire and Rate of Change of Southeastern Pacific Humpback Whales at Their Breeding Area in the Gulf of Chiriquí, Panama," *Scholarworks*, 2024.

[9] F. Oppenheimer, "Estimating Change in Humpback Whale Song Repertoire Through Occupancy Modeling and Phrase Transition Analysis," *Scholarworks*, 2025.

[10] A.N. Allen, M. Harvey, L. Harrell, A. Jansen, K.P. Merkens, C.C. Wall, J. Cattiau, & E.M. Oleson, “A convolutional neural network for automated detection of humpback whale song in a diverse, long-term passive acoustic dataset,” *Frontiers in Marine Science*, 8, 607321, 2021

[11] V. Kather, F. Seipel, B. Berges, G. Davis, C. Gibson, M. Harvey, L. Henry, A. Stevenson, & D. Risch, “Development of a machine learning detector for North Atlantic humpback whale song,” *The Journal of the Acoustical Society of America*, 155(3), 2050-2064, 2024.

[12] D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[13] T. Bergamaschi, “A Topic Modeling Framework for Humpback Whale Song,” MBARI Summer Internship, 2018.

[14] M. Pickett, D. Cline, and J. Ryan, “Exploring Coherence Metrics for Optimizing Topic Models of Humpback Song,” Monterey Bay Aquarium Research Institute, 2020.

[15] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, 66:846–850, 1971.

[16] J. M. Santos & M. Embrechts, “On the use of the adjusted rand index as a metric for evaluating supervised classification,” In International conference on artificial neural networks (pp. 175-184). Berlin, Heidelberg: Springer Berlin Heidelberg, Sept. 2009.

[17] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, & J.P. Bello, “Per-channel energy normalization: Why and how,” *IEEE Signal Processing Letters*, 26(1), 39-43, 2018.

[18] D. Sculley, “Web-scale k-means clustering,” *in Proceedings of the 19th international conference on World wide web* - WWW ’10, p. 1177, 2010.

[19] T. L. Griffiths and M. Steyvers, “Finding scientific topics,” *PNAS*, vol. 101, no. 1, pp. 5228–5235, 2004.

[20] Y. Girdhar, P. Giguère, and G. Dudek, “Autonomous adaptive exploration using realtime online spatiotemporal topic modeling,” *Int. J. Rob. Res.*, vol. 33, no. 4, pp. 645–657, Apr. 2014.

[21] K. Lisa Yang Center for Conservation Bioacoustics at the Cornell Lab of Ornithology. Raven Pro: Interactive Sound Analysis Software (Version 1.6.5) [Computer software]. Ithaca, NY: The Cornell Lab of Ornithology. Available from https://www.ravensoundsoftware.com, 2024.