A Topic Modeling Framework for Humpback Whale Song

Thomas Bergamaschi, UC Santa Cruz

Mentors: Danelle Cline, Ben Yair Raanan, Dr. John Ryan

Summer 2018

Keywords: humpback whale song, passive acoustic monitoring, unsupervised machine learning, natural language processing, topic modeling

ABSTRACT

Humpback whales produce intricate hierarchical patterns of repeated vocal elements called songs. Studying these songs requires consistent classification of song components and measurement of variation over time and geographic location. Performing such classification and measurements manually is time consuming and can produce subjective results that differ from study to study, which makes it difficult to perform large scale analyses and impedes comparability between results of different analyses. For these reasons, efforts have been made to develop automated methods for performing such analyses. These analyses have operated at the level of song units, using a manual or automated detection pass as a precursor to modeling, and using the units isolated by the detection pass as the model's input. This paper explores the application of a probabilistic topic modeling framework to modeling humpback whale song, which does not require a detection pass as a precursor, and is a mixed membership model, allowing the model to adapt the evolving nature of humpback whale song. The findings presented in this paper indicate that the framework is capable of differentiating song units and background noise, and to a certain extent, between song units themselves.

1 INTRODUCTION

Humpback whale song has inspired and mystified both cetologists and the world at large for decades. These complex vocalizations were first described by Roger Payne and Scott McVay in their 1971 paper *Songs of Humpback Whales*. In this seminal paper, they showed that humpbacks produce hierarchical patterns of repeated vocal elements so intricate that the two felt these vocalizations were deserving of a name to distinguish them, and called them songs, a term from avian literature [1]. They called the smallest aurally distinguishable song element a *unit*, and found that one or more of these units are repeated to create a *phrase*. Phrases are in turn repeated to make up *themes* and finally, themes are repeated to make up a song. Songs, which in general range from 7 to 30 minutes are repeated during *song sessions*, which can last several hours [1], [2]. The discovery of this phenomenon sparked a subsequent flurry of research and investigation to better understand these songs, their purpose, and their role in the lives of humpback whales.

The discovery was soon made that the individuals producing songs were male and that peak song production occurs during mating season [2]. It was also shown that singers in acoustic contact incorporate elements from each other's song and thus produce similar songs, but that the songs evolve over time within these acoustically connected communities as a result of singers modifying spectral and temporal characteristics of song elements [2]. These findings lead to the hypotheses that song plays a role in both attracting a mate and in facilitating interaction between males [2]. However, testing these hypotheses and conducting further study of the song of humpback whales requires consistent classification of song components and measurement of variation over time and geographic location.

Performing such classification and measurements manually is time consuming and can produce subjective results that differ from study to study, which makes it difficult to perform large scale analyses and impedes comparability between results of different analyses. Moreover, a massive amount of acoustic data has been generated by hydrophones around the world that record humpback song, and the objective, systematic

analysis of such a large dataset is unmanageable to perform manually. For these reasons, efforts have been made to develop automated methods for performing such analyses.

The development of automated methods that are capable of fully capturing the complexity of humpback whale song is naturally a complex task. The model must distinguish song units from background noise, identify groups of similar song units, and also adapt as these song units evolve over time and geographic location. Methods such as self-organizing maps (SOMs), hierarchical clustering, and hidden Markov models (HMMs) have been applied to modeling humpback whale song [3]–[5]. Such analyses have operated at the level of song units, using a manual or automated detection pass as a precursor to modeling, and using the units isolated by the detection pass as the model's input.

This paper explores the application of a probabilistic topic modeling framework to modeling humpback whale song. This framework does not require a detection pass as a precursor. It takes as input unsegmented acoustic data, and the model itself separates background noise from signal by identifying separate topics for each. Moreover, it is a mixed membership model, allowing the model to adapt the evolving nature of humpback whale song.

Probabilistic topic models are a suite of algorithms that originate from natural language processing (NLP). They are typically used to model the underlying topics in large collections of textual data [6]. While this is the most common application of probabilistic topic models, there are several examples of successful applications outside the realm of textual analysis that inspired the exploration of their use in this paper [7][8].

This paper is organized as follows: Section 2 provides an explanation of the topic modeling framework and our approach to applying the framework to humpback whale song. Section 3 presents the results attained from our application of the topic modeling framework to humpback whale song recorded in the Monterey Bay, California. Section 4 discusses and interprets the presented results. Section 5 concludes and suggests future work regarding this model and its applications to humpback whale song analysis.

2 METHODS AND APPROACH

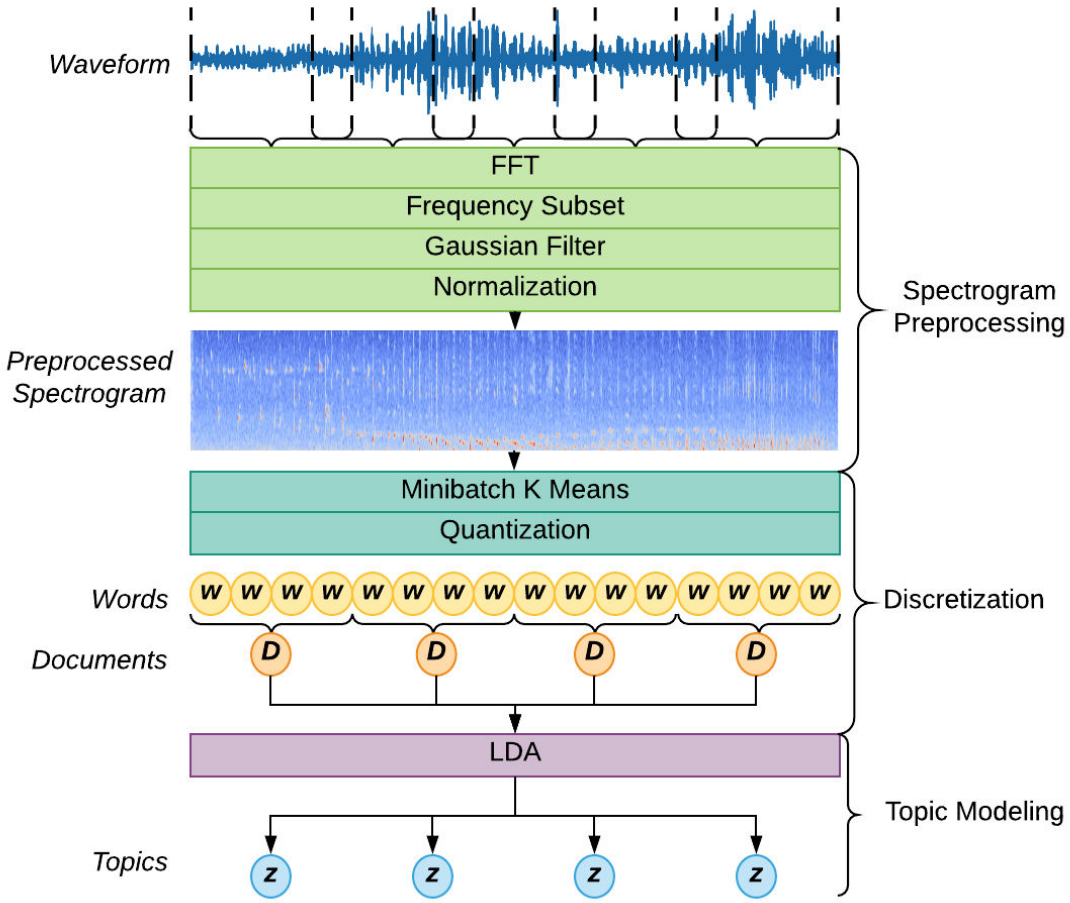

Figure 2-1: A depiction of the model framework. Data is fed in as a waveform. Spectrogram preprocessing divides the waveform into its frequency components, and performs subsequent operations to optimize it for quantization. Quantization converts the preprocessed spectrogram to a series of integers (words) from a fixed set (vocabulary) and forms these words into groups (documents). Documents are fed into the topic model which outputs documents represented as mixtures of topics.

This section first provides background on the dataset used in this analysis, and then an explanation of the topic modeling, which consists of two steps: preprocessing and topic modeling. The preprocessing step is further decomposed into two steps: spectrogram preprocessing and discretization. Figure 2-1 offers a depiction of these steps and the dataflow through them.

2.1 DATASET

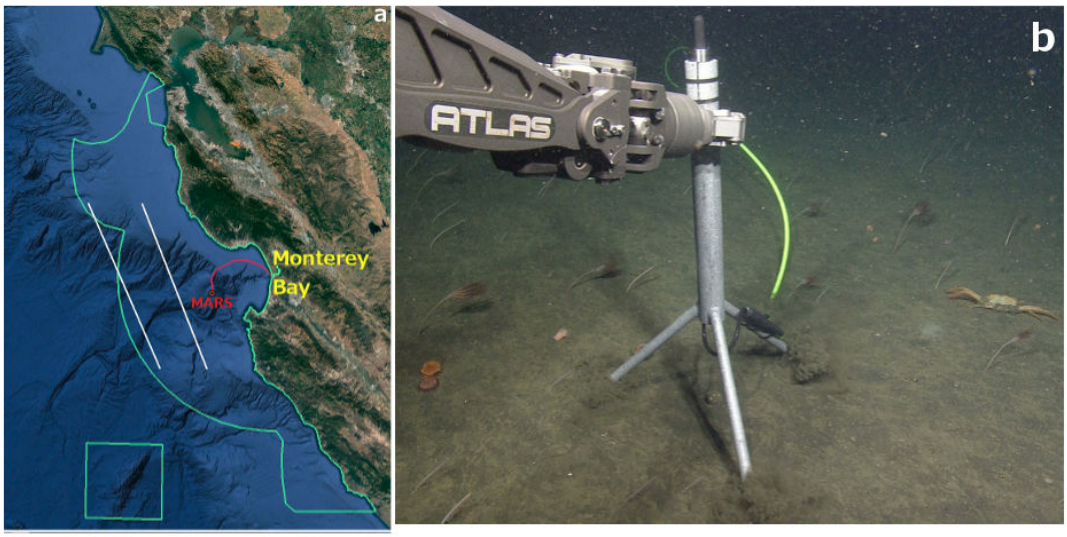

Figure 2-2: Left: the location of the MARS cabled observatory in the greater Monterey Bay. Right: an image of the hydrophone used to collect the data.

The original dataset consisted of 429 isolated humpback whale songs, recorded between October 2015 and April 2017 by the hydrophone connected to the MARS cabled observatory. The hydrophone and its location in Monterey Bay are shown in Figure 2-2. The recordings were decimated from their original sample rate of 256 kHz to 32 kHz, the mean value was removed to correct for DC offset, and they were normalized by the magnitude of the maximum value, so that values fell within the range -1 to 1.

Of the 429 songs in the dataset, 10 songs with particularly high signal to noise ratios were chosen and all units within the songs were labeled by my fellow intern Miriam Hauer-Jenson with start time, end time, a manual classification of the unit type, and several other spectral measurements. The dataset used in the analysis summarized in this paper was scaled down to these 10 songs. This choice was made for the purpose of having a ground truth to which results could be compared, and for the purpose of allowing for rapid iteration over various parameterizations of the model and preprocessing pipeline.

2.2 SPECTROGRAM PREPROCESSING

Each file in the dataset was first converted from amplitude over time (waveform) to frequency over time (spectrogram) using a series of overlapping fast Fourier transforms

(FFTs). The spectrograms were subsetted in the frequency domain, between 50 Hz and 2 kHz, to include only the frequency band in which song units occurred. A Gaussian filter was then applied to the spectrogram to remove Gaussian noise and increase comparability between FFT frames. Finally, the spectrogram was normalized in the time and frequency domains by subtracting the mean and dividing by the standard deviation to give the songs units of standard deviation and to increase comparability between songs.

2.3 DISCRETIZATION

Discretization is the key step that allowed us to apply probabilistic topic models, which are designed to model discrete data, to our continuous acoustic data. Probabilistic topic models are most commonly used to analyze collections of textual documents, called corpora. In this use-case, words are represented as integers and take the value of $1 \ldots V$ where V is the size of the vocabulary [9]. Documents in the corpus are then represented as a sequence of these integers. These sequences are what probabilistic topic models take as input. In order to format our continuous acoustic data in this manner, we wished to treat each preprocessed FFT frame as a word in a vocabulary of fixed size, and represent it with an integer.

FFT frames resulting from the spectrogram preprocessing step were clustered using the mini-batch k-means algorithm [10], which was chosen over k-means because of the size of our dataset required a faster clustering algorithm in order to rapidly test different model configurations and parameterizations. The value of k was chosen using an inertia analysis. The set of centroids resulting from the mini-batch k-means (vocabulary) was used to quantize the preprocessed spectrograms as follows. For each FFT frame in the preprocessed spectrogram, the centroid that minimized the Euclidian distance between it and the FFT frame was determined, and the FFT frame was replaced by the index of that centroid (word). Temporally adjacent words were then grouped to create non-overlapping documents, which served as the input to the probabilistic topic model.

2.4 LATENT DIRICHLET ALLOCATION (LDA)

Latent Dirichlet allocation (LDA), the topic model used in this analysis, is a generative, mixed membership model intended for application to collections of discrete data such as textual corpora [9]. Our acoustic data was prepared for ingestion by LDA through the

discretization process described in Section 2.2. LDA assumes the observed data is the result of a generative process that involves latent, or unobserved variables called topics. Topics $\phi_{1:K}$ are distributions over words in the vocabulary $w_{1:V}$. LDA assumes that a document $d \in D$ is generated by first selecting a distribution over topics, θ_d. Then, to generate each word in the document $w_i \in d$, a topic z_i is sampled from the selected topic distribution θ_d, and a word w_i is sampled from the topic z_i. Therefore, documents are modeled as mixtures of topics, which can be interpreted as a lower dimensional representations that carry semantic meaning. This process can be described as the following joint probability distribution, where α and β are Dirichlet hyperparameters [11]:

$$P(\boldsymbol{w}, z, \theta, \phi \mid \alpha, \beta) = P(\phi \mid \beta) P(\theta \mid \alpha) P(z \mid \theta) P(\boldsymbol{w} \mid \phi_z)$$

Here, α controls the sparsity of θ, and β controls the sparsity of ϕ. A lower α will yield a model that characterizes documents using fewer topics, whereas a lower β will yield a model that characterizes topics using fewer words. Bayes theorem can be applied to learn the latent variables z, θ and ϕ from the joint probability distribution above by expressing their probability as the conditional posterior distribution given the observed data.

$$P(z, \theta, \phi \mid \alpha, \beta) = \frac{P(\boldsymbol{w}, z, \theta, \phi \mid \alpha, \beta)}{P(\boldsymbol{w} \mid \alpha, \beta)}$$

We applied a temporally smoothed variant of the above method to our discretized acoustic data in order to account for its time-series nature, where instead of computing the topic mixture for a single document, the topic mixture is computed for a temporal neighborhood of documents, as explained in Girdhar et al. [12]. We used a collapsed Gibbs sampler to approximate the posterior, as it is intractable to compute directly [11].

2.5 PERFORMANCE EVALUATION

Different parameterizations of the model must be compared to each other in order to decide on final values for α, β, and number of topics K. Therefore, a performance metric is needed to provide a basis for the comparison of different model parameterizations. For our performance metric, we used per-word perplexity. Per-word perplexity can be thought of as the uncertainty in recreating the word labels in a document given that

document's topic mixture. We used per-word perplexity score averaged over all documents, defined as:

$$Perplexity = \sum_{d \in D} \exp\left(-\frac{\sum_{w \in d} \log P(w|d)}{W_d}\right) \Big/ D$$

Where W_d is the number of words in document *d* and *D* is the total number of documents. Computing this metric for several different model parameterizations allowed us to decide on the optimal parameterization by choosing the set of parameters that minimized per-word perplexity.

3 RESULTS

This section demonstrates the application of our topic modeling framework, presented Section 2, to the dataset of 10 songs recorded in Monterey Bay. It first explains the reasoning behind the choices of spectrogram parameters and provides an example of the spectrogram preprocessing step in the framework being applied to a song in the dataset. Next, the inertia analysis used to determine vocabulary size is presented. Then the results grid search of α and β, and the number of topics *K* are presented with final selections for these parameters. Finally, the model output is visualized, and compared to the ground truth.

3.1 SPECTROGRAM PARAMETERS

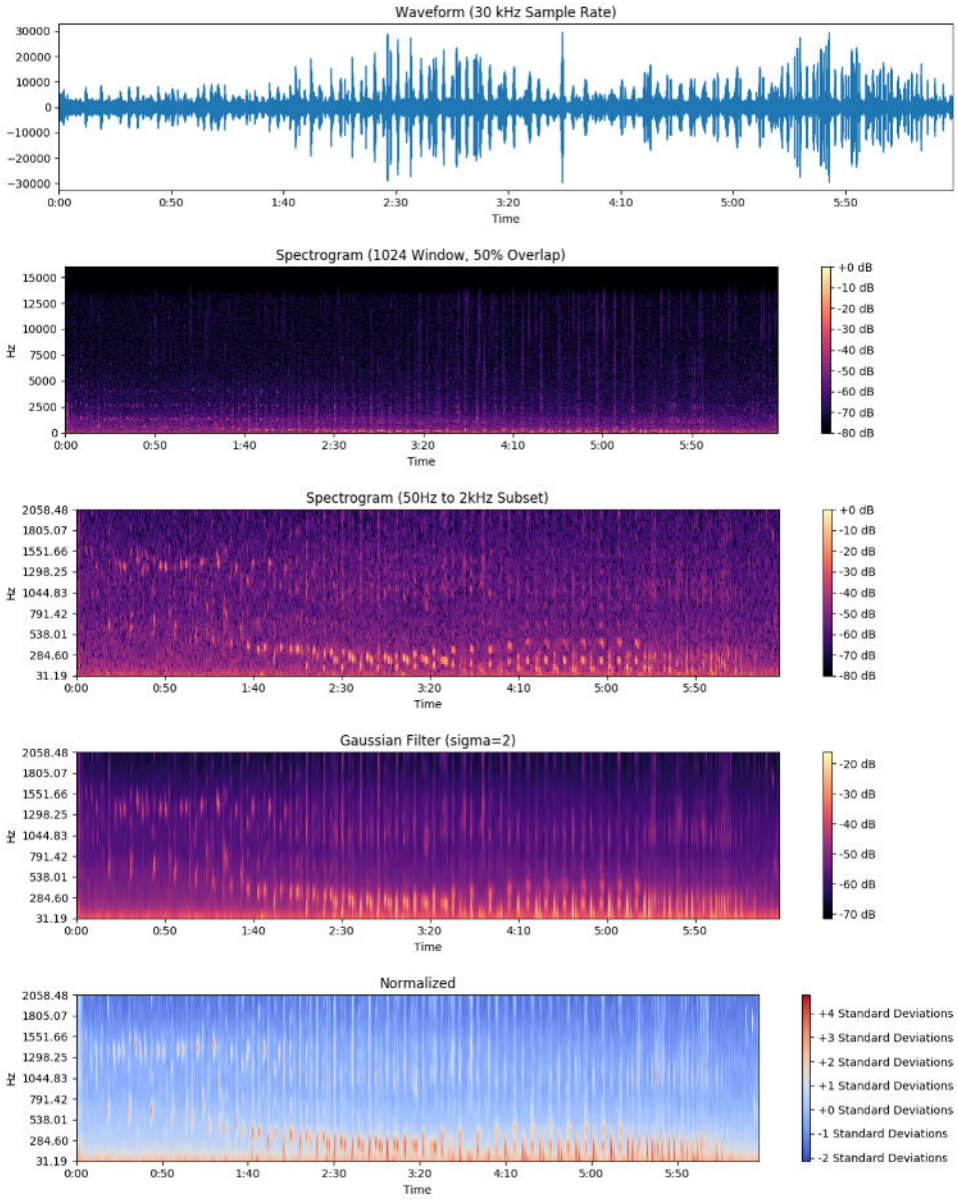

Figure 3-1: The spectrogram preprocessing pipeline applied to a humpback whale song recorded on December 7, 2015. From top down: its waveform, its spectrogram, its subsetted spectrogram, its filtered spectrogram, and its normalized spectrogram.

The number of samples over which the discrete Fourier transform is computed (window size) and the percent overlap of these windows (overlap) greatly affect the spectral and temporal resolution of a spectrogram. Higher spectral and temporal resolution represent song units more clearly, but use more data to do so which increases computational cost of

downstream processing. Therefore, window size and overlap parameters were chosen to adequately represent individual song units using the least amount of data necessary to do so. It was determined that a window size of 1024 frames and an overlap of 50% provided the best trade-off between these two factors. Once the spectrograms were generated, data above 2 kHz and below 50 Hz was discarded, leaving data only from the frequency band in which song units occurred. The Gaussian filter was then applied with sigma equal to 2. Lastly, the spectrograms were normalized in both the time and frequency domains, putting the spectrogram in units of standard deviation. Figure 3-1: The spectrogram preprocessing pipeline applied to a humpback whale song recorded on December 7, 2015. From top down: its waveform, its spectrogram, its subsetted spectrogram, its filtered spectrogram, and its normalized spectrogram.shows an example of these steps applied to a recording in the dataset.

3.2 VOCABULARY SIZE

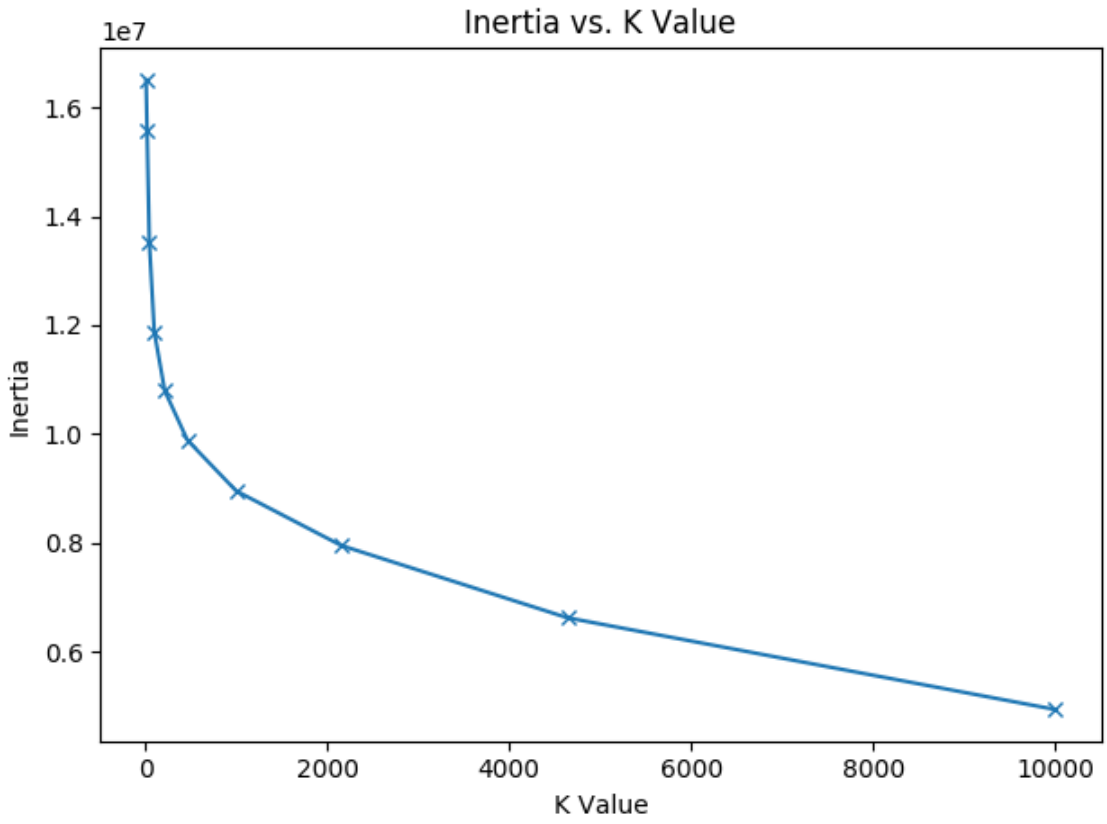

Figure 3-2: Inertia score for k values of 10, 21, 46, 100, 215, 464, 1000, 2154, 4641, and 10000. The optimal k value is found at the elbow of the curve. The figure shows that the optimal k value is 1000 according to this inertia analysis.

To determine our vocabulary size, or k value, we used an inertia analysis. To perform this analysis, we used mini-batch k means to cluster the preprocessed FFT frames from all recordings in the 10 song dataset, with k equal to 10, 21, 46, 100, 215, 464, 1000, 2154, 4641, and 10000. We plotted the sum of squared distances from each FFT frame to its cluster's centroid (inertia) for each of these k values, and produced the plot shown in Figure 3-2. We used this plot to choose a k value of 1000, the k value at the elbow of the curve where a marginal increase in k yields a substantially diminished improvement in inertia.

3.3 MODEL PARAMETERS

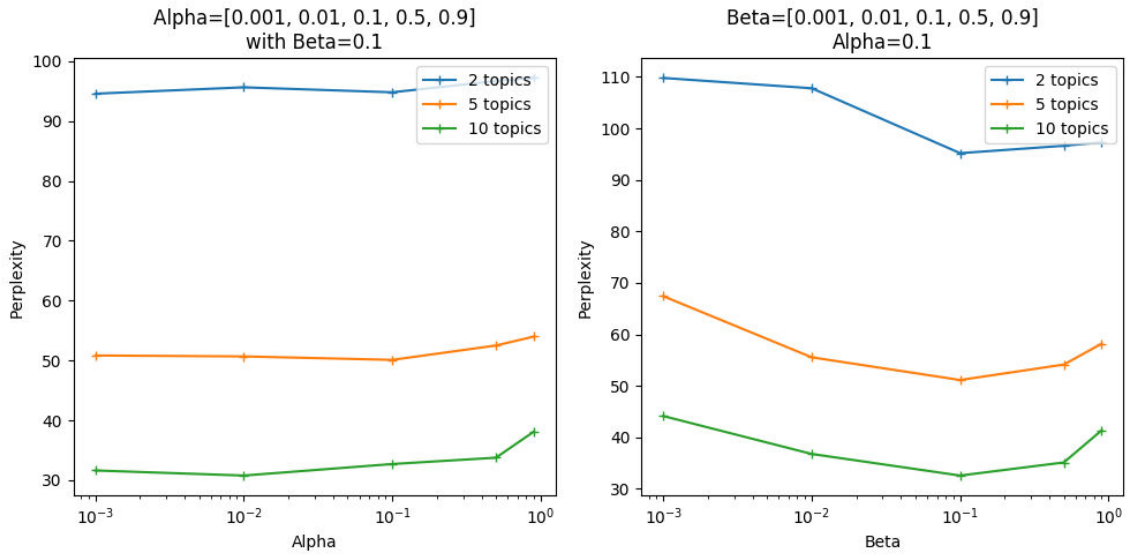

Figure 3-3: Grid search of model parameters α, β, and number of topics K plotted versus average per-document perplexity. The plot shows that the minimum average per-document perplexity is achieved with α = 0.01, β = 0.1 and K = 10.

A grid search of the Dirichlet hyperparameters α and β, and the number of topics *K* was performed to find the combination of model parameters that minimized perplexity. The model was run with α equal to 0.001, 0.01, 0.1, 0.5, and 0.9 with β fixed at 0.1, and again with β equal to 0.001, 0.01, 0.1, 0.5, and 0.9, and α fixed at 0.1. This was repeated for *K* equal to 2, 5, 10, 15, and 20. The model was run 5 times for each set of α, β and *K* and the average per-document perplexity score was averaged over the 5 runs. Figure 3-3 shows the plotted results, and from it one can see that minimum average per-document perplexity is achieved with α = 0.01, β = 0.1 and K = 10. Note that *K* = 15 and *K* = 20

are not shown on this plot as their average per-document perplexities were orders of magnitude higher than that of K = 2, 5, and 10.

3.4 MODEL VISUALIZATION

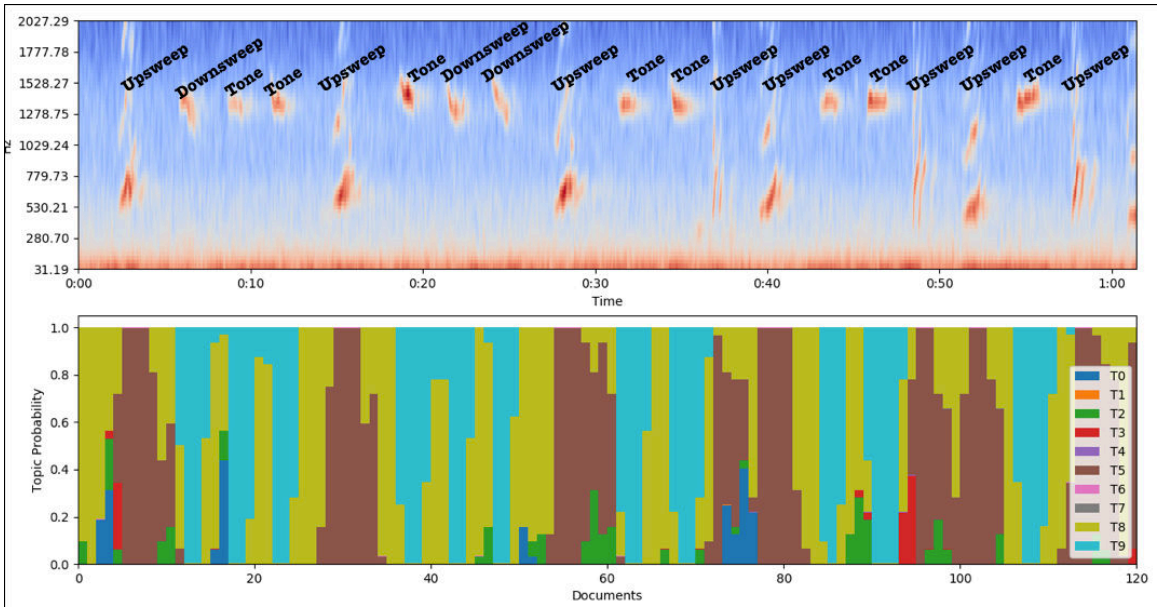

Figure 3-4: A visualization of the model's output. On top is the spectrogram of a 60 second portion of the humpback song recorded on December 7, 2015. Unit labels are included above each unit for reference. On bottom is a stacked bar plot of the topic mixtures for the documents of this 60 second portion of song, where the x axis is documents and the y axis is topic probability. Documents in which topics 0, 2, and 8 are in high probability correspond to segments of the recording that are background noise, and contain no vocalizations. Documents with topic 5 in high probability correspond to upsweep units. Documents with topic 9 in high probability correspond to downsweep or tone units.

Finally, the model was run with the parameterization that minimized perplexity: $\alpha = 0.01$, $\beta = 0.1$ and $K = 10$. The model's output was visualized as a stacked bar plot where each bar represents a document and each color represents a topic in that document's topic mixture. Figure 3-4 shows this bar plot for 60 second segment humpback whale song with a spectrogram containing song unit labels. Documents in which topics 0, 2, and 8 are in high probability correspond to segments of the recording that contain no vocalizations. Documents with topic 5 in high probability correspond to upsweep units. Documents with topic 9 in high probability correspond to downsweep or tone units.

3.5 GROUND TRUTH COMPARISON

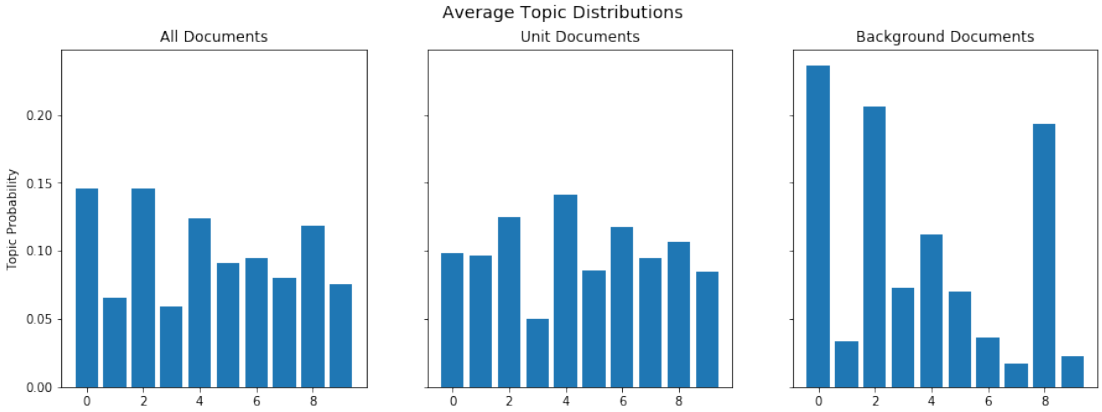

Figure 3-5: Left: the average topic mixtures of all documents in the song. Center: the average topic mixtures of all documents that contain a unit. Right: the average topic distribution of all documents that contain background noise.

The model was compared to the ground truth labels to determine what topics represented background noise and what topics represented units. Figure 3-5 shows the average topic mixtures of all documents on the left, the average topic mixtures of documents that contain units in the middle, and the average topic mixtures of documents containing background noise on the right. A comparison of the average topic mixture of documents containing background noise to the average topic mixture of all documents shows that topics 0, 2, and 8 are in significantly higher probability in documents containing background noise than in an average document. Therefore we can identify topics 0, 2, and 8 as the topics representing background noise.

4 DISCUSSION

The visualization of the model in Figure 3-4 and the comparison to ground truth in Figure 3-5 strongly suggest that the topic modeling framework presented in this paper may have the capability to learn topics that differentiate between song units and background noise. Topics 0, 2, and 6 consistently corresponded to background noise in the model visualization in Figure 3-4, and this is corroborated by comparing the average topic mixtures of documents containing song units and the average topic mixture of all documents in Figure 3-5.

Moreover, the visualization of the model indicates that the topic modeling framework may be able to learn topics that differentiate different unit types. In Figure 3-4, topic 5

consistently corresponded to upsweep units, whereas topic 4 corresponded to both downsweeps and tones. The model's ability to distinguish downsweeps and tones from upsweeps but not from each other could be due to LDA's assumption of word exchangeability.

LDA assumes both document exchangeability and word exchangeability [9]. In other words, it assumes that neither the order of the documents nor the order of the words within them is relevant to the latent processes responsible for generating them, which the topic model aims to learn. While this assumption is acceptable for applications of LDA to textual corpora, it is an unreasonable assumption in the context of time series data. In the topic modeling framework presented in this paper, the assumption of document exchangeability is relaxed by computing topic mixtures for a temporal neighborhood of documents instead of a single document. However, this temporally smoothed variant of LDA still assumes word exchangeability within these temporal neighborhoods. Therefore, song units that occupy the same frequency band, and as such contain the same words in their corresponding documents, will likely be modeled with similar topic mixtures, regardless of how their energy changes over time within that frequency band. The spectrogram shows that most of the energy in the upsweeps occurs between 500 Hz and 1 kHz, whereas for both the downsweeps and tones that range is between around 1 kHz and 1.5 kHz. This difference in frequency band is a possible reason why downsweeps and tones are modeled as a single topic when upsweeps are modeled as a separate one.

In order to better understand the relationship between topics and unit types, further comparison to ground truth is required. Ideas on how this can be achieved, as well as other thoughts for future work on the model, are provided in Section 5.

5 CONCLUSIONS

The findings presented in this paper indicate that the framework is capable of differentiating song units and background noise, and to a certain extent, between song units themselves. These results warrant further exploration of this topic modeling framework and its application to humpback whale song.

The first step in such further exploration would be seek a better understanding of the relationship of topics, background noise, and unit types, through further comparison to the ground truth. The ground truth can be used to assign sematic meaning to the topics generated by computing the probability of a topic mixture given a class should be for a training subset of the dataset. Then, this conditional probability can be reversed using Bayes' theorem to find the probability of a class given a topic mixture. The resulting conditional probabilities would help to understand the relationships between topics, background noise, and unit types. They could also be used to add a detection or classification step on the end of the pipeline. Detection/classification accuracies may provide a better performance metric for tuning model parameters than perplexity because they incorporate information from the ground truth. Another next step could be exploring the use of other features. The framework presented in this paper used FFT frames as the feature vectors clustered to produce the vocabulary. The use of other feature vectors could be tested, such as Mel frequency cepstrum coefficients (MFCCs) which are commonly used in analysis of human speech. Finally, the model could be made non-parametric by using the Chinese restaurant process to automatically discover the number of topics, rather than providing number of topics as a parameter.

ACKNOWLEDGEMENTS

Thank you very much to my mentors Danelle Cline, Ben Yair Raanan, and Dr. John Ryan for their help and guidance. Thank you to the WARP Lab at Woods Hole Oceanographic Institute's for their Realtime Online Spatiotemporal Topic modeling (ROST) CLI used in this project. Thank you to the authors of libROSA, a Python package for audio analysis used in this project. Thank you to the Monterey Bay Aquarium Research Institute (MBARI) for welcoming me as an intern this summer, and to the David and Lucille Packard Foundation for funding the internship program.

References:

[1] R. S. Payne and S. McVay, "Songs of Humpback Whales," *Science (80-.).*, vol. 173, no. 3997, pp. 585–597, Aug. 1971.

[2] D. M. Cholewiak, R. S. Sousa-Lima, and S. Cerchio, "Humpback whale song hierarchical structure: Historical context and discussion of current classification issues," *Mar. Mammal Sci.*, vol. 29, no. 3, pp. 312–332, 2013.

[3] J. A. Allen, A. Murray, M. J. Noad, R. A. Dunlop, and E. C. Garland, "Using self-organizing maps to classify humpback whale song units and quantify their similarity," *J. Acoust. Soc. Am.*, vol. 142, no. 4, pp. 1943–1952, 2017.

[4] A. K. Stimpert, L. E. Peavey, A. S. Friedlaender, and D. P. Nowacek, "Humpback Whale Song and Foraging Behavior on an Antarctic Feeding Ground," *PLoS One*, vol. 7, no. 12, pp. 1–8, 2012.

[5] F. Pace, P. White, and O. Adam, "Hidden Markov Modeling for humpback whale (Megaptera Novaeanglie) call classification," no. October 2015, p. 070046, 2012.

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

[7] A. Kalmbach, Y. Girdhar, and G. Dudek, "Unsupervised environment recognition and modeling using sound sensing," *Proc. - IEEE Int. Conf. Robot. Autom.*, pp. 2699–2704, 2013.

[8] B.-Y. Raanan *et al.*, "Detection of unanticipated faults for autonomous underwater vehicles using online topic models," *J. F. Robot.*, vol. 35, no. 5, pp. 705–716, Aug. 2018.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

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

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

[12] 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.