Monterey Bay Aquarium
Research Institute

Exploring Coherence Metrics for Optimizing Topic Models of Humpback Song

Madison Pickett, Massachusetts Institute of Technology

Mentors: Danelle Cline and John Ryan

Summer 2020

Keywords: humpback whale song, unsupervised machine learning, topic modeling, coherence, perplexity, embedding features, topic probability

ABSTRACT

Humpback whales produce collections of intricate patterns and unique calls known as songs. In order to study these complex songs, consistent classification and measurement of humpback whale song features are necessary. Currently, there is no standard classification to apply to humpback songs. A well-established lexicon, or terminology, of humpback calls will allow scientists to compare and share their work across the world which could open a window into the culture and communication of humpback whales. Topic modeling, an unsupervised machine learning technique, can be used to consistently and objectively automate the labeling of humpback units. This paper specifically explores coherence as an evaluation metric for topic modeling. Optimizing the topic model will allow for more consistent and reliable results, and coherence has the potential to further improve the topic model.

INTRODUCTION

The broad and complex variety of vocalizations of the humpback song has continually captured the interest of researchers and scientists. With the expansion of acoustic technology, researchers have been able to further study the intricacies of the humpback song [1]. Figure 1 displays a spectrogram (a time, energy, and frequency plot) of a single male humpback whale song.

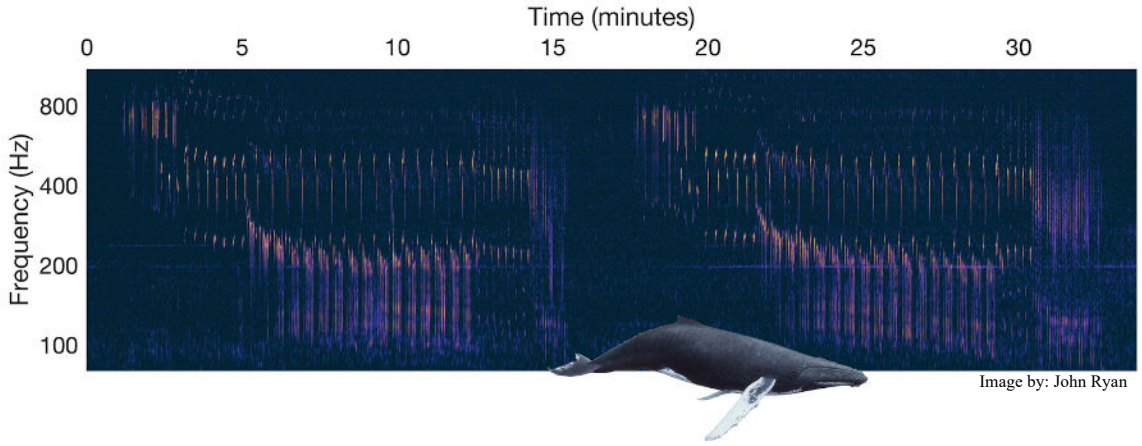


Figure 1: A spectrogram of a single male humpback song demonstrates a 15 minute long song that is repeated twice in this song session. A spectrogram is a time, frequency, and energy plot.

The smallest distinguishable element of a humpback song is defined as a *unit*. Multiple units together create a *phrase*. The next level of categorization is a *theme* which consists of a sequence of phrases. Lastly, multiple themes describe an individual *song* [2]. Individual humpback songs can range from 7 minutes to 30 minutes each, and songs can repeat multiple times to create a *song session* that can last up to 24 hours [3]. By studying years of recordings of humpback whales, researchers have discovered that humpback songs evolve over time as humpbacks change their own songs by introducing new song units or modifying the order and repetition of their own song [4].

These discoveries have resulted in an interest in understanding how humpback songs have changed over time. In order to answer this question, humpback whale songs need to be analyzed and calls need to be classified in a consistent, well-

established method. If we can learn how humpback songs have evolved over time, years' worth of humpback song data could expand and further progress our understanding of humpback whale culture and communication. However, there is no universal method to systematically distinguish one call from another [1]. The current method of hand labeling and classifying data is time consuming and subjective. Scientists across the globe are using their own classification methods which have resulted in inconsistent identification of distinct calls [5]. The lack of a standard lexicon, or a catalog of unique call types, inhibits the comparison of humpback data across different studies and locations. Developing a way in which to consistently distinguish unique calls will allow humpback research to expand and flourish.

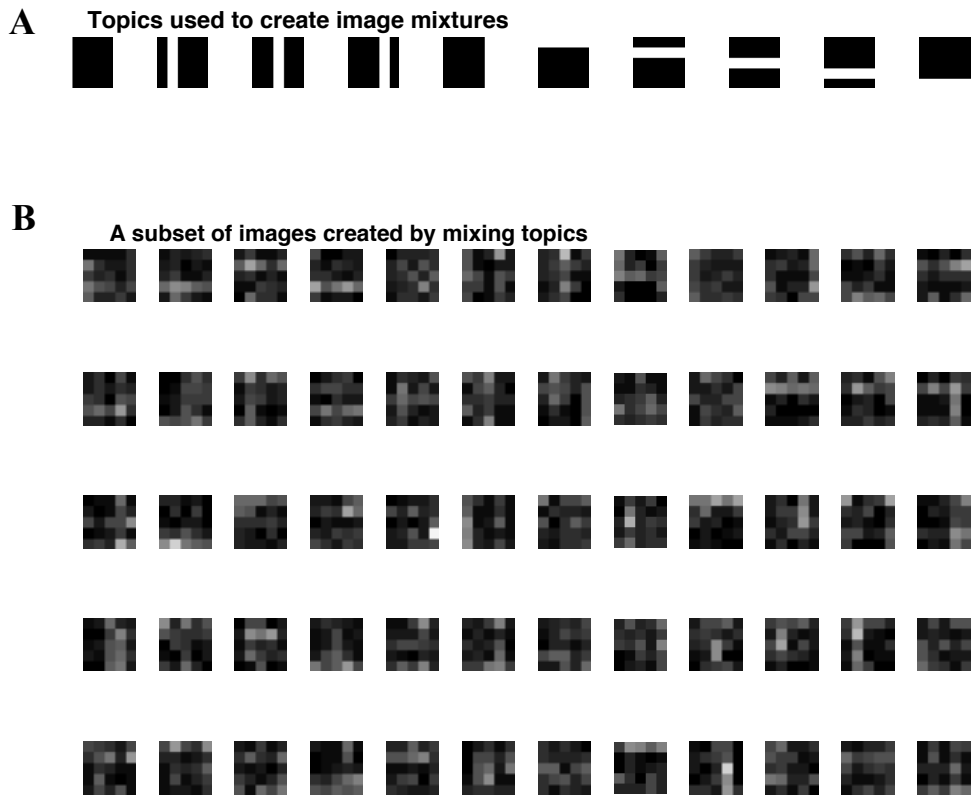
Machine learning presents a solution to analyze humpback song in a more efficient and less subjective manner that will further the development of a common lexicon. There are two types of machine learning: supervised and unsupervised. Supervised machine learning requires labeled data to train and learn while unsupervised machine learning does not [6]. This paper focuses on topic modeling, a class of unsupervised machine learning, in order to help create an objective and efficient method to analyze humpback song. Specifically, this paper demonstrates how coherence can be a useful metric to optimize the topic model.

DEFINITIONS AND METHODS

TOPIC MODELING

To increase both the efficiency and objectivity of labeling humpback data, topic modeling can be used. This unsupervised machine learning method analyzes data into clusters or patterns that can be more easily recognizable and interpreted. Topic modeling originated from natural language processing in the human language and is typically used to organize massive collections of textual data [7]. Specifically, a type of probabilistic topic modeling known as Latent Dirichlet Allocation (LDA) was used. This generative, imaginary random process assumes that documents contain multiple topics [8]. Figure 2 demonstrates the generative random process

of topic modeling through black and white images [9]. In this example, each image represents a document and each pixel represents a word. The intensity of the image is the frequency. Figure 2a displays the set of 10 defined topics, and Figure 2b consists of images created by randomly mixing the topics. The pixels in Figure 2c are constructed by sampling the LDA distribution of the topics for each pixel and assigning it a random pixel from the selected topic. Figure 2c also introduces the parameter of number of iterations. The clarity and distinguishability of the topics improve with a greater number of iterations, but levels out at approximately 50 iterations.



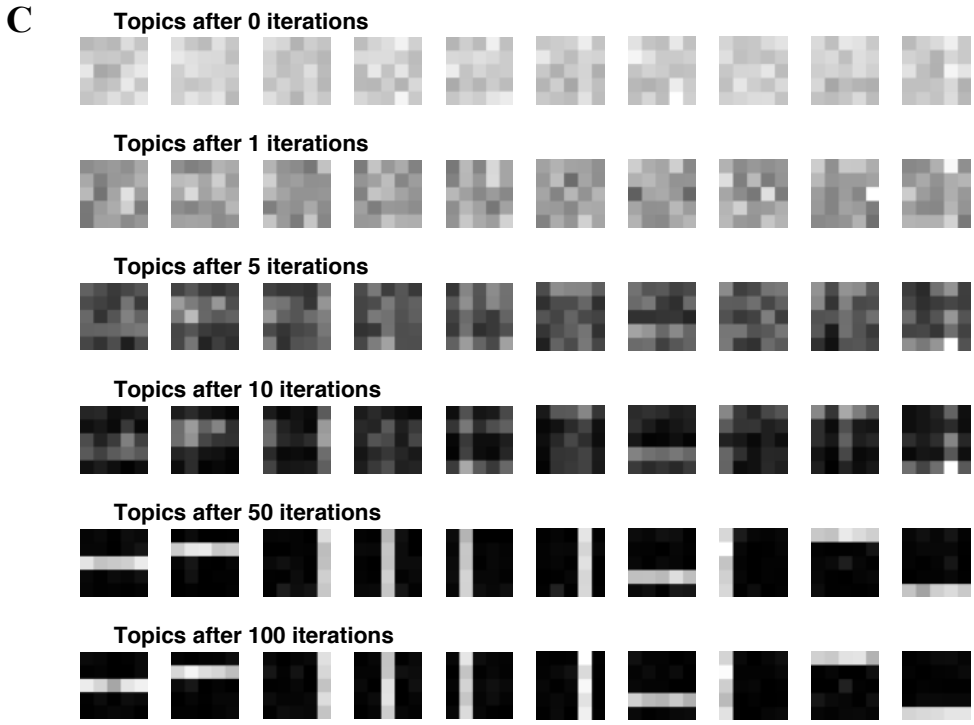


Figure 2: Images created by the generative random process of topic modeling in which an image is a document, a pixel is a word, and the intensity is the frequency. **(a)** The horizonal and vertical lines in the images represent the 10 defined topics. **(b)** A set of images that were created by randomly mixing the topics. **(c)** Topics, or pixels, were randomly selected for each document. Iterations of this random sampling demonstrate how the clarity can improve with a greater number of iterations, but only up to 50 iterations at which time the clarity seems to remain constant [9].

The input for a probabilistic topic model is a set of documents with multiple topics in each. The output consists of two probability distributions. The first output is a probability distribution of words over topic which is the probability that a specific word is in a particular topic. The second output of the topic model is the probability distribution of topics over documents which is the probability that a particular topic is in a document. This paper focuses on the second probability distribution comprised of topics over documents. The inputs and outputs of the topic modeling process are shown in Figure 3.

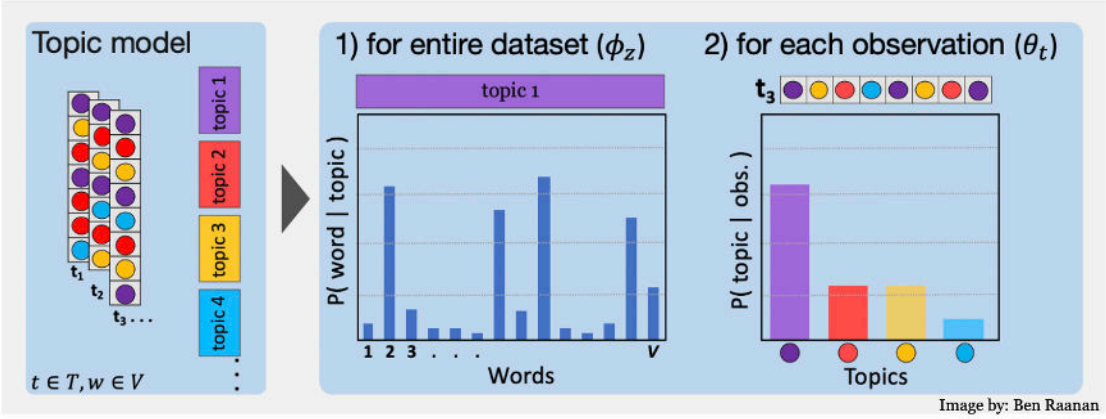


Figure 3: A diagram of the process of topic modeling. The input is a set of documents with different topics. The output is two probability distributions. The first (1), represented by ϕ, is the probability that a word is in a specific topic. The second (2), represented by θ, is the probability that a topic is in a specific document.

In LDA, there are two important variables that describe the outputs as shown in Figure 3. ϕ is the distribution of words in the vocabulary and θ is the distribution of topics in the documents. To obtain these outputs, the Dirichlet hyperparameters, α and β, are critical in the following equation:

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

In this equation, α and β regulate the sparsity, or the infrequency, of θ and ϕ respectively. A small α will yield an output of a sparser θ in which documents that have fewer topics, and a small β will yield an output of ϕ that characterizes topics with fewer words, a sparser ϕ [7]. The parameters used for the topic modeling in this paper were defined as $\alpha = 0.01$ and $\beta = 0.1$. In addition, the output of θ was most important in calculating coherence.

For the goal of this topic model, the terms *document*, *topic*, and *word* represent different components of humpback song classification. A document is a *humpback*

song. A topic is a *label*. Lastly, a word is now a *slice of time of the spectrogram*. Figure 4 displays an example of the input and output of the topic model we used for humpback whale song. The input is a spectrogram of an 18 second clip of humpback song recorded in 2016. The output is a plot that represents the topic distribution over the entire 18 second clip – it is the probability that a time slice of the spectrogram is a specific topic. The different colors in Figure 4b represent the different potential topics.

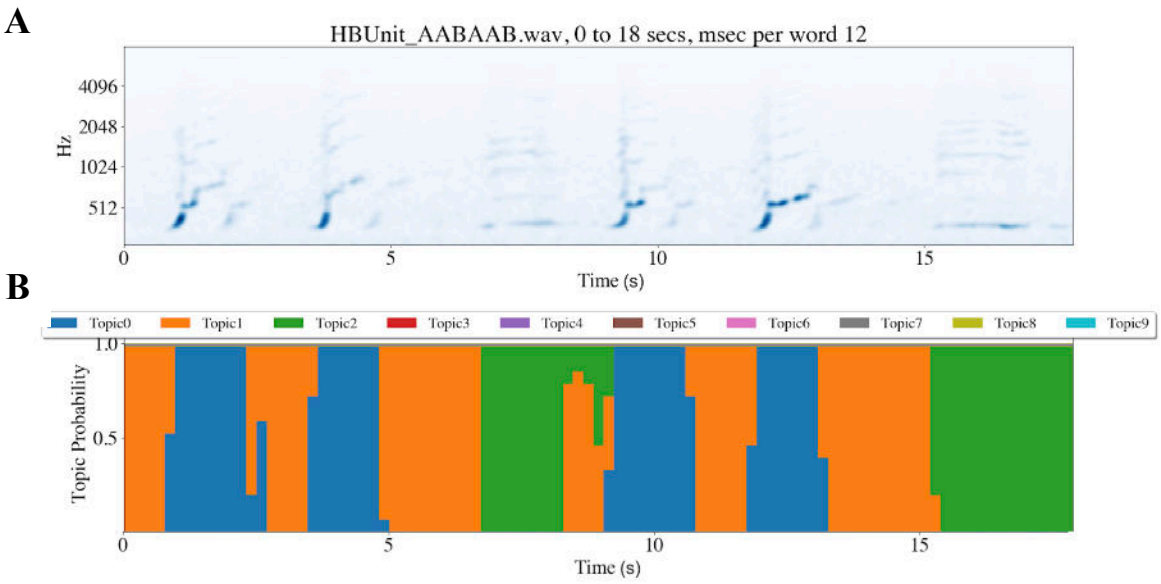


Figure 4: **(a)** A spectrogram of an 18 second clip of a humpback song recorded in 2016. This was the input into the topic model used. **(b)** A topic probability plot in which the colors represent different topics. Specifically, Topic 0 (blue) and Topic 2 (green) are the humpback whale calls and Topic 2 (orange) is background noise. The plot represents the probability that the specific time slice of the spectrogram is the designated topic.

This example clip of a humpback whale song contained the phrase "AAB" and was repeated twice. The topic probability output found the 4 "A" calls as Topic 0 (blue) and the 2 "B" calls as Topic 2 (green). Topic 1 (orange) is the background noise. By finding the different calls originally labeled as "AABAAB," the topic model for this example song clip performed as expected.

PERPLEXITY

To evaluate the performance of topic modeling, the metric perplexity was used. Perplexity is a predictive likelihood that specifically measures the probability that new data occurs given what was already learned by the model. In other words, perplexity characterizes how surprised a model is with new, unseen data [10]. Perplexity is calculated as:

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

In which D is the number of documents and W_d is the number of words in the specific document d [7]. However, perplexity does not depict the consistency of the topics but rather describes the presence of a new topic. This paper focuses on a different optimization metric to see if it can help improve the topic model.

COHERENCE

Coherence is an evaluation metric that can be used to assess the performance of the topic model. Coherence is typically used to analyze the relationship between two sets of data or the similarity between data sets. In topic modeling, topic coherence measures the quality of the data by comparing the semantic similarity between highly repetitive words in a topic [10]. Coherence score is a scale from 0 to 1 in which a good coherence (high similarity) has a score of 1, and a bad coherence (low similarity) has a score of 0 [11]. In other words, a good coherence is when two signals or data sets are perfectly related and identical, whereas a bad coherence is defined as having no association between data sets.

The MATLAB function `mscohere` was used to calculate coherence. `mscohere` calculates the magnitude squared coherence using the power spectral densities and the cross power spectral density of the input [12]. The coherence output of MATLAB function uses the coherence scale in which a score close to 1

demonstrates that the x and y (rows and columns of the input matrix) correspond to each other at the respective time value. The `mscohere` function is defined as:

$$C_{xy}(f) = \frac{|P_{xy}(f)|^2}{P_{xx}(f)P_{yy}(f)}.$$

Here, P_{xy} is the m-dimensional vector of cross power spectral densities between the inputs and y. P_{xx} is the m-by-m matrix of power spectral densities and cross power spectral densities of the input. P_{yy} is the power spectral density of the output [12].

The coherence function has many parameters that affect the outcome of the coherence calculations. The two most important parameters for topic coherence are window and overlap. The window divides the input into segments of the specified length. The default window, Hamming window, was found to output a smoother plot than the Blackman window, so the Hamming default window was selected. The overlap parameter defines the number of overlaps in the input. Ultimately, a window of 4 and an overlap of 3 led to the best results for the shorter humpback clips between 9 seconds and 18 seconds. However, these are important parameters to modify depending on the length of the clips.

TESTING COHERENCE

By creating and utilizing two simple .wav files, the application of coherence could be assessed and tested. Each file consists of 3 different tones. As demonstrated in Figure 5a and 5b, the first tones in both files are exactly the same. The third tones in each file are also the same. However, note that the middle tone is different between the two files.

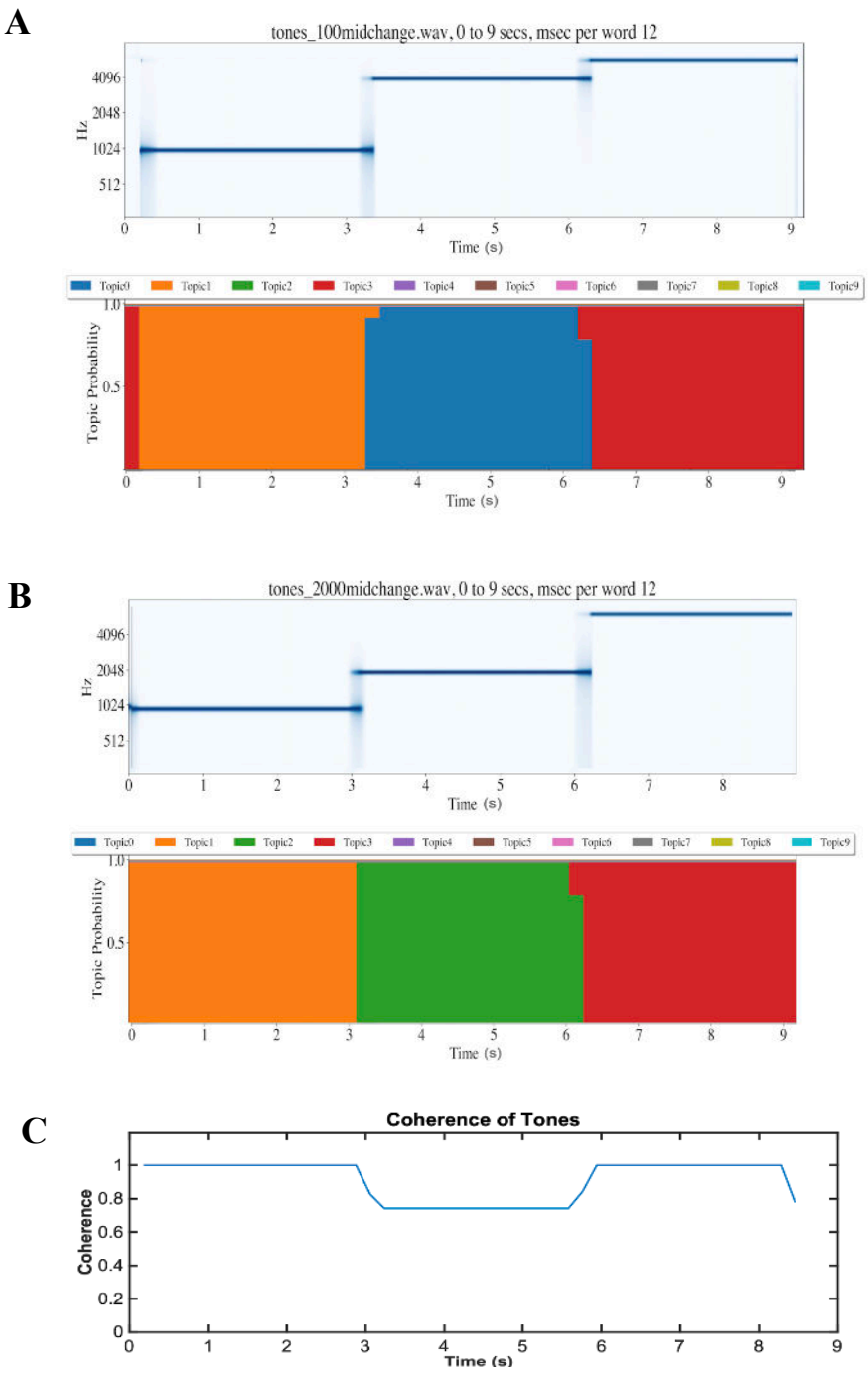


Figure 5: Simple .wav files each with 3 tones were created to test coherence. **(a)** The first file with 3 different tones, and the topic probability output displays 3 different topics. **(b)** The second file with 3 tones. The first tone and third tone are the same to their respective tones in the first file, but the second tone is different. The topic probability output displays the same results. **(c)** The coherence calculation between both .wav files. A high coherence of 1 for the first and third tones and a lower coherence for the second tone as expected.

The topic probability for both .wav files captured the different tones and represented the differences in the topics. Topic 1 (orange) is the first call in both files. Topic 2 (green) is only the second call in the second file, and Topic 0 (blue) is only the second call in the first file. Lastly, Topic 3 (red) is the third tone in both files. Using both topic probability outputs, coherence was calculated as shown in the bottom plot in Figure 5c. For the first set of tones in both files (same tones and same topics) the coherence score is 1. For the second set of tones (different tones and different topics) the coherence drops to 0.7 as there is a lack in similarity in the tones. Lastly, the third set of tones has a high coherence of 1 because they are the same tones and same topics. The calculation of coherence clearly demonstrates the similarities and differences between the topic probability of both files.

PROCESS

In order to calculate coherence, a number of steps must be completed. The entire process is illustrated in Figure 6. The input is a .wav file of a humpback song. The topic model then performs a few processes which includes: Gaussian filter, normalization, preprocessing of the spectrogram, and discretization [7]. Each of these steps have their own parameters that must be modified and optimized for best results. Then the data proceeds through topic modeling. The output of the topic model is θ: the topic probability distribution plot over the length of the document (time). After obtaining the θ values of one .wav file and the θ values of another .wav file (or from different clips of the same file), coherence can be calculated.

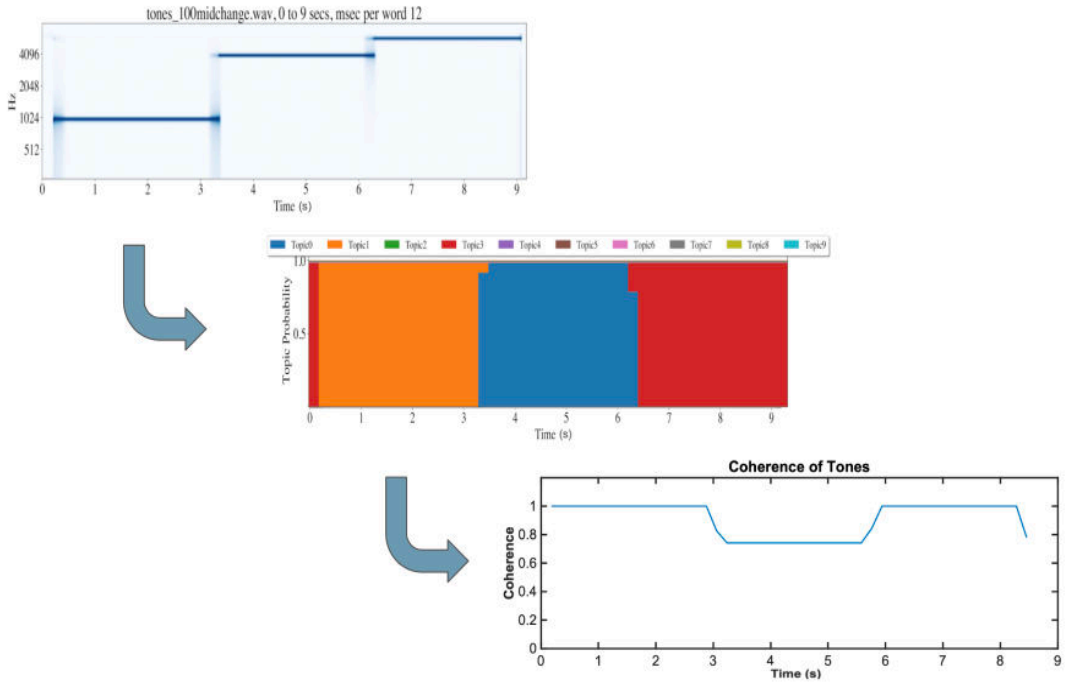


Figure 6: This flow chart represents the process of topic modeling that was used. The input into the topic model is a humpback whale song .wav file. The output is the topic probability distribution. To calculate coherence, two sets of data are required. The final outcome is a coherence score over time.

RESULTS

COHERENCE OF HUMPBACK SONG

With the deeper understanding of coherence from testing with the tone files, the metric was then applied to humpback whale song data. Two different portions of the humpback whale song recorded off Monterey Bay in 2016 were used. Each clip of the humpback whale song had 3 different units, or calls, which were previously labeled as “AAC” and “AAB” phrases as shown in Figure 7.

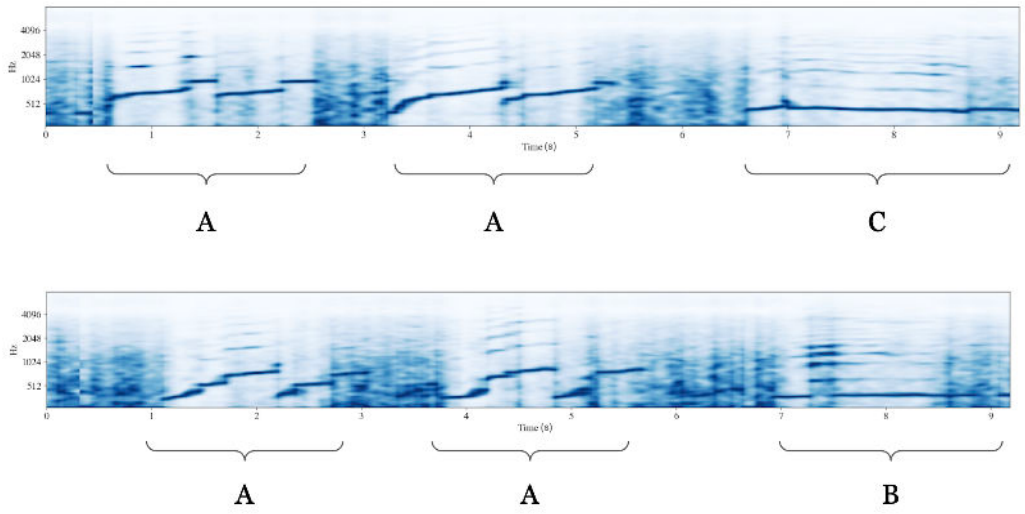


Figure 7: Two spectrograms from a humpback whale song. The two clips are different as they contain the phrases "AAC" and "AAB" respectively.

Both clips of the song were entered into the topic model. The topic probability output recognized the similar "A" calls and the difference in the "B" and "C" calls as shown in the topic probability plots in Figure 8. The "A" calls were labeled as Topic 2 (green), and the "B" call became Topic 0 (blue) and the "C" call was labeled as Topic 3 (red). Topic 1 (orange) represents the background noise.

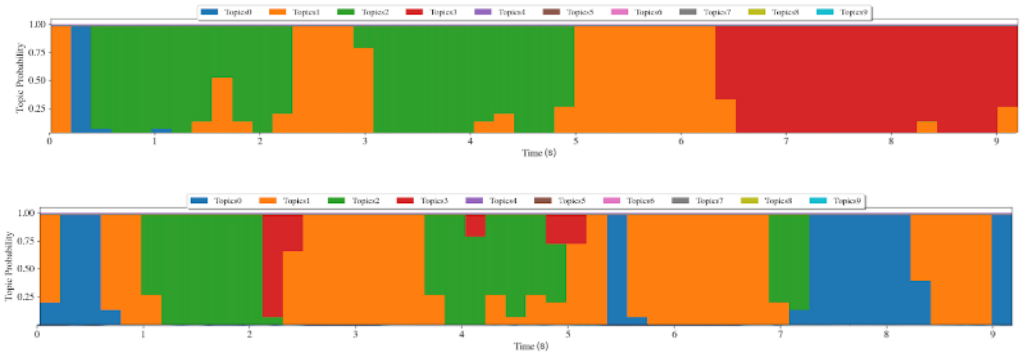


Figure 8: The topic probability output of the two humpback clips demonstrates the "AAB" and "AAC" phrases. The call "A" is represented by Topic 2 (green), the call "B" is Topic 0 (blue), and the call "C" is represented by Topic 3 (red). Topic 1 (orange) is background noise.

Using both topic probability outputs, coherence was calculated using a window of 4 and an overlap of 3. As highlighted in Figure 9, the first and second set of "A" calls resulted in a high coherence of nearly 1. As for the "B" and "C" comparison, the coherence score was practically 0 as expected because they are different calls and different topics.

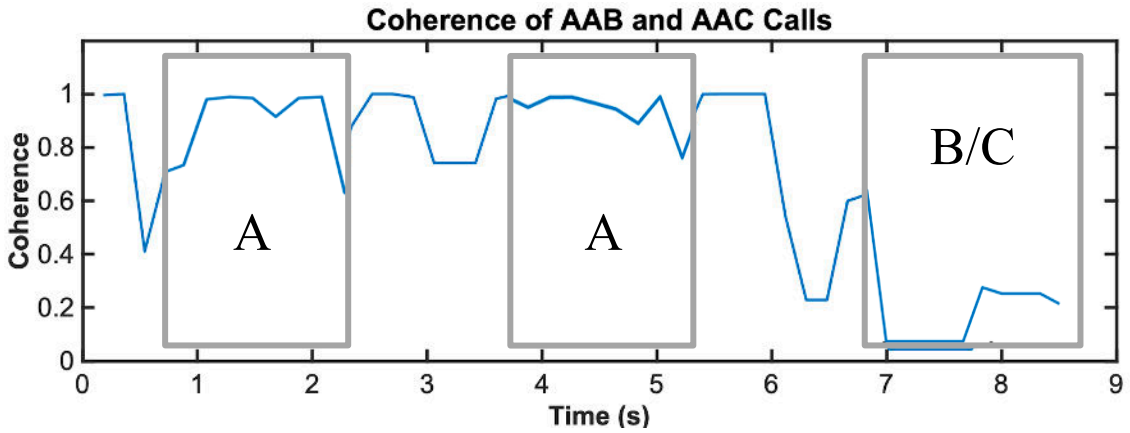


Figure 9: Plot of coherence between the two humpback clips. The sets of similar "A" calls (green Topic 2) scored high coherence of nearly 1 proving a high similarity. The set of different "B" and "C" calls (blue Topic 0 and red Topic 3) scored a low coherence of nearly 0 demonstrating the lack of similarity in the calls.

COHERENCE OF EMBEDDING FEATURES

Instead of using the topic probability output from the topic model to calculate coherence, the .wav file itself was used to calculate coherence. Audio files consist of layers of embedding features that represent or describe the file. The code utilized was developed by Google and it extracts the embedding features by using a VGGish convolutional neural network. The feature representation output is a matrix in which the columns display topics and rows portray the layers of the representation at each time frame. Figure 10 displays the output from calculating coherence using the embedding features rather than the topic probability distribution.

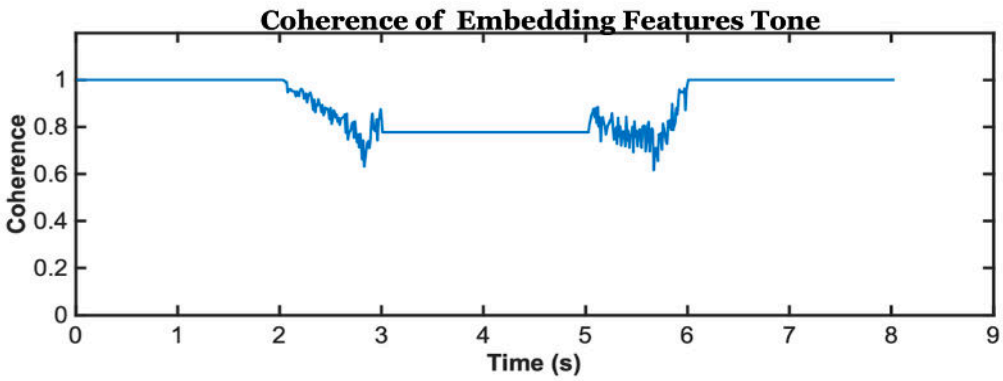


Figure 10: Plot of coherence using the embedding features from the original practice tone files displayed in Figure 5. The coherence calculated from the embedding features is very similar to the coherence calculated from the tone files' topic probability output in Figure 5c. There is a high coherence of 1 when the tones are the same for the first and third tones, and there is a low coherence for the middle tone when the tones were different.

DISCUSSION

The coherence calculations of the embedding features are similar to the coherence output of the topic probabilities. However, the embedding features have more points, so the coherence plot using the embedding features is noisier. The embedding features directly describe the .wav file itself and have more data which means the coherence calculation is more precise. Nonetheless, the embedding features coherence calculation skips the topic model step; therefore, it is not contributing to the optimization of the topic model. To continue to evaluate the topic model process, coherence needs to be calculated using the topic probability output. However, the embedding features can be a way to verify the calculations of coherence and check the topic probability output as well.

CONCLUSIONS

The findings presented in this paper demonstrate how coherence can be utilized to evaluate and optimize topic modeling for humpback whale song. By calculating the similarity between topics, coherence can be a valuable metric to filter the good and

bad topics or the consistency between frequent humpback whale songs. Overall, the use of coherence can improve the optimization of topic modeling. However, further exploration is necessary to better optimize topic modeling for humpback whale song.

FUTURE WORK

The most immediate next step is to calculate coherence of a humpback whale song using the embedding features to ensure a resemblance between both coherence calculations. More research needs to be conducted on how to utilize both methods – using topic probability distribution output or using the embedding features – to calculate coherence and evaluate the topic model. In addition, coherence needs to be calculated on longer songs that might not show distinguishable topics in the topic probability output. This would test how `mscohere` as a function can handle larger sets of data. Another future step is to calculate the coherence of each of the labeled calls. For example, calculating the coherence between all of the labeled "A" calls could lead to the discovery of consistency between labels. Lastly, it would be interesting to cross correlate a single call across the rest of the song to find similar calls and their coherence. Ideally, coherence would be integrated into our current topic model such that the model can optimize itself based on the coherence outputs.

ACKNOWLEDGEMENTS

I would like to thank my mentors Danelle Cline and John Ryan for their knowledge, support, and guidance. Thank you to Ben Raanan for sharing his expertise and to Thomas Bergamaschi, a previous intern, for his work on topic modeling. Lastly, a huge thank you to Monterey Bay Aquarium Research Institute (MBARI) and George Matsumoto for a fantastic summer internship and to the David Lucile Packard Foundation, the Dean and Helen Witter Family Fund, and the Rentshler Family Fund for funding this internship program.

References:

[1] M. E. Fournet, A. Szabo, and D. K. Mellinger, "Repertoire and classification of non-song calls in Southeast Alaskan humpback whales (Megaptera novaeangliae)," *The Journal of the Acoustical Society of America*, vol. 137, no. 1, 2015.

[2] D. M. Choweliak, S. Cerchio, J. K. Jacobsen, J. Urbán-R, and C. W. Clark, "Songbird dynamics under the sea: acoustic interactions between humpback whales suggest song mediates male interactions." *Royal Society Open Science*, vol. 5, 2018.

[3] R. A. Dunlop, M. J. Noad, D. H. Cato, and D. Stokes, "The social vocalization repertoire of east Australian migrating humpback whales (Megaptera novaeangliae)," *The Journal of the Acoustical Society of America*, vol. 122, no. 5, 2007.

[4] D. M. Choweliak, R. S. Sousa-Lima, and S. Cerchio, "Humpback whale song hierarchical structure: Historical context and discussion of current classification issues," *Marine Mammal Science*, vol. 29, no. 3, 2013.

[5] H. Pines, "Mapping the phonetic structure of humpback whale song units: extraction, classification, and Shannon-Zipf confirmation of sixty sub-units," *The Journal of the Acoustical Society of America*, vol. 35, 2018.

[6] M. Harvey, "Acoustic Detection of Humpback Whales Using a Convolutional Neural Network," *Google AI Blog,* 2018.

[7] T. Bergamaschi, "A Topic Modeling Framework for Humpback Whale Song," *MBARI Summer Internship*, 2018.

[8] D. M. Blei, "Probabilistic Topic Models," *Communications of the ACM*, vol. 55, no. 4, pp. 77-84, 2012.

[9] T. L. Griffiths, and M. Steyvers, "Finding scientific topics," *PNAS*, vol. 101, suppl. 1, 2004.

[10] S. Kapadia, "Evaluate Topic Models: Latent Dirichlet Allocation (LDA)," *Towards Data Science*, 2019, https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[11] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing Semantic Coherence in Topic Models," *Association for Computational Linguistics*, 2011.

[12] "mscohere," *Magnitude-Squared Coherence – MATLAB*, The MathWorks, Inc. https://www.mathworks.com/help/signal/ref/mscohere.html#bvi4lxm-window