Automatic Classification of Humpback Whale Social Calls

Irina Tolkova[1] and Lisa Bauer[2], Antonella Wilby[3], Ryan Kastner[4], Kerri D. Seger[5], Aaron M. Thode[6]

[1]Applied Math and Computer Science Departments, University of Washington, WA, USA

[2]Department of Computer Science, Johns Hopkins University, MD, USA

[3]Jacobs School of Engineering, University of California San Diego, CA, USA

[4] School of Marine Science and Ocean Engineering, University of New Hampshire, NH, USA

[5]Scripps Institution of Oceanography, University of California San Diego, CA, USA

Abstract

Acoustic methods are becoming increasingly common in the study of marine mammal populations and behavior. Automating the detection and classification of whale vocalizations has been a central aim of these methods. The focus has primarily been on intra-species detection and classification, however, humpback whale (Megaptera novaeangliae) social call detection and classification has largely remained a manual task in the bioacoustics community. To automate this process, we processed spectrograms of calls using PCA-based and connected-component-based methods, and derived features from relative power in the frequency bins of these spectrograms. We then used these features to train and test a supervised Hidden Markov Model (HMM) algorithm to investigate classification feasibility.

Introduction

There is a growing need for the automated detection and classification of vocalization classification to expedite research requiring analysis of massive quantities of recorded whale vocalizations. Some automated methods have explored detecting and classifying humpback whale song using HMMs (Pace, White, & Adam 2012; Rickwood & Taylor 2008), however whale song and social calls differ in the rhythms and patterning of song that are not present in social calls (Silber 1986). There is a lack of exploration concerning the different social calls humpback whales produce and while a few studies have focused on the detection of such social calls (Stimpert 2011), social call classification has not been extensively examined and remains mostly a manual task.

In recent years, HMMs have gained popularity in bioacoustics vocalization classification applications due to their flexible nature (Ren 2009). HMMs consider a temporal progression of the data and therefore characterize the spectral changes of a call over time. For many years, HMMs have been a well-established tool in Automated Speech Recognition and have been used for both supervised and unsupervised applications. Since they have gained popularity in marine mammal bioacoustics signal classification, they have been successfully used in many studies. Datta & Sturtivant (2002) utilized HMMs to classify three different groups of dolphin whistles. A HMM was learned for each whistle class and the resulting HMMs were used to classify new whistles by computing the likelihood that a new recorded whistle signal belonged to each class. HMMs have also been successfully employed to classify cetacean calls. Brown & Smaragdis (2009) classified seven different killer whale call types using HMMs, with the top results yielding about 95%. We examined a supervised HMM approach to classify Humpback whale social calls and verify the ability to classify social calls given ecological classification categories.

To convert the acoustic signals to a compact representation which can be analyzed with an HMM, we calculate spectrograms over the audio data and use a Principal-Component-Analysis-based method for subtracting the ambient background noise. Variants of PCA have long been used for dimensionality reduction in image processing problems including face recognition and video surveillance (Oliver et al, 2000). While most studies develop more sophisticated, robust variants of PCA, the basic singular value decomposition has been shown to be effective in cases where foreground objects are small relative to the image size, and don't appear stationary for multiple frames (Guyon et al, 2012). As these conditions fit our problem, and make for a simpler and less computationally intensive approach, we employed PCA for background subtraction. We followed this approach with a connected-components-based method for removal of leftover noise. Then, we derived features from relative power in the frequency bins of these spectrograms, and trained and tested a supervised HMM algorithm to investigate classification feasibility.

Materials and Methods

The data for this project was collected and studied by Dr. Kerri Seger and Dr. Aaron Thode of the Scripps Institute of Oceanography. Acousonde acoustic tags were deployed on humpback whales in Los Casbos, Mexico, a region which falls within a migration route and also serves as a breeding ground of the North Pacific humpback whale subpopulation (Seger, 2016). Acoustic data was collected over a total of 31 days in February and March of 2014 and 2015, resulting in just under 70 hours of recordings at a sampling rate of around 8000 samples per second. Each recording was manually screened for whale calls, and the time range and frequency range of each whale call was recorded, as well as shape metrics such as the slope and the number of harmonics, and SNR metrics such as the RMS of noise and of the signal. Additionally, each call was classified into a call type category, for about 50 distinct categories of calls.

The most prominent source of noise in the data were high-amplitude transmitter pings, about 0.06 seconds in length, that occurred every half-second. To reduce the interference of these pings with the signal, we broke the data into 3550-sample windows in between the pings. First, we extracted the first minute of five recordings, separated this range into windows, computed the spectrograms of each window, and saved these values in a matrix. The frequency range of the spectrograms stretched from 0 to 4 kHz. Then, we calculated the singular value decomposition of this matrix, and observed the structure of the resulting principal components. The first two components were chosen to represent the time-frequency structure of the ambient acoustic background of the recordings.

To analyze a new dataset, we similarly restructured the data into a series of windows. For each window, the spectrogram was computed. Then, we subtracted the contribution of the two "background components" from the spectrogram by calculating the projection of the spectrogram onto each component. All pixels with a value of less than 0.75 were set to zero. As this yielded a sparse matrix, we used a connected-components-based approach to further remove noise from the spectrogram. First, we set the elements in the 0hz-200hz frequency range to zero, as there were few traces of calls in this frequency band, and a larger degree of residual noise. Then, we calculated the connected components within each spectrogram with a built-in Matlab function, and applied a threshold in both the minimum pixel count and the minimum power of the component, where power was taken to be the sum of the amplitudes of the pixels in that component. Every component that consisted of fewer than 20 pixels and had a power value of less than 20 was removed.

After processing the spectrograms with background subtraction and connected-component analysis, we separated these spectrograms into smaller windows in the time domain. If a window fell within a labeled call, it was given a label corresponding to that call; otherwise, it was given a “no call” label. A feature vector was calculated for each window by dividing the 200hz to 3800hz frequency range into equal-size bins and assigning a value to represent the relative amplitude within each bin. These time-varying features were then loaded into a classification algorithm that used a supervised HMM approach.

For classification purposes, each call category was represented by a particular HMM that maximized the likelihood of its respective call category. During the training phase, labeled data was segmented by category into k different subsets, where k was the number of different categories present in that data set. Each data point consisted of a call that comprised several time varying feature vectors. Each of the calls in the k subsets were used to train a specific HMM that corresponded to that subset, thus creating k models where each model corresponded to a particular call category. The training of a particular HMM was accomplished by using the Baum-Welch algorithm to estimate model parameters.

Once the training stage completed, the forward-backward algorithm was used to assign test data to the model that maximized the likelihood of the data. This model determined the classification of the test data. The algorithm’s classification was then compared to the manual classification to compute the overall accuracy, and the precision and recall for each class. We used k-fold cross validation to evaluate estimation performance and obtain averages for all of our evaluations.

We tested 4 different data sets and a combination of all 4 data sets on the classifier. We tested 3 categories found in these data sets (data points containing no whale calls, squeak, and low yap). There were a total of 21049 no whale calls, 119 squeaks, and 62 low yaps present in the data. We experimented with feature vectors that consisted of 9 and 30 features and examined the differences in performance across data sets.

The processing work in this study was done in Matlab and the HMM classification algorithm was implemented using python.

Results and discussion

The principal components resulting from applying principal components analysis to a collection of unprocessed spectrograms, and the corresponding singular values, are shown in Figure 1. It is visible that some of the components, such as components 1, 2, 3, 5, and 7, are time-invariant, and so can be considered to be representative of the ambient background structure. We used the first two components for background subtraction, and these accounted for 39% of the energy present in those spectrograms. This approach has the advantage of being able to represent backgrounds of different intensities, with varying structure on the frequency axis, without removing irregularities (such as calls) in the foreground. One downside of PCA is the possibility of negative components or negative coefficients, which has a potential consequence of cancellation between two components. Since there is no physical cancellation of acoustic signals – sounds are additive –an alternative, possibly more intuitive approach to this analysis would use a nonnegative matrix factorization instead of PCA.

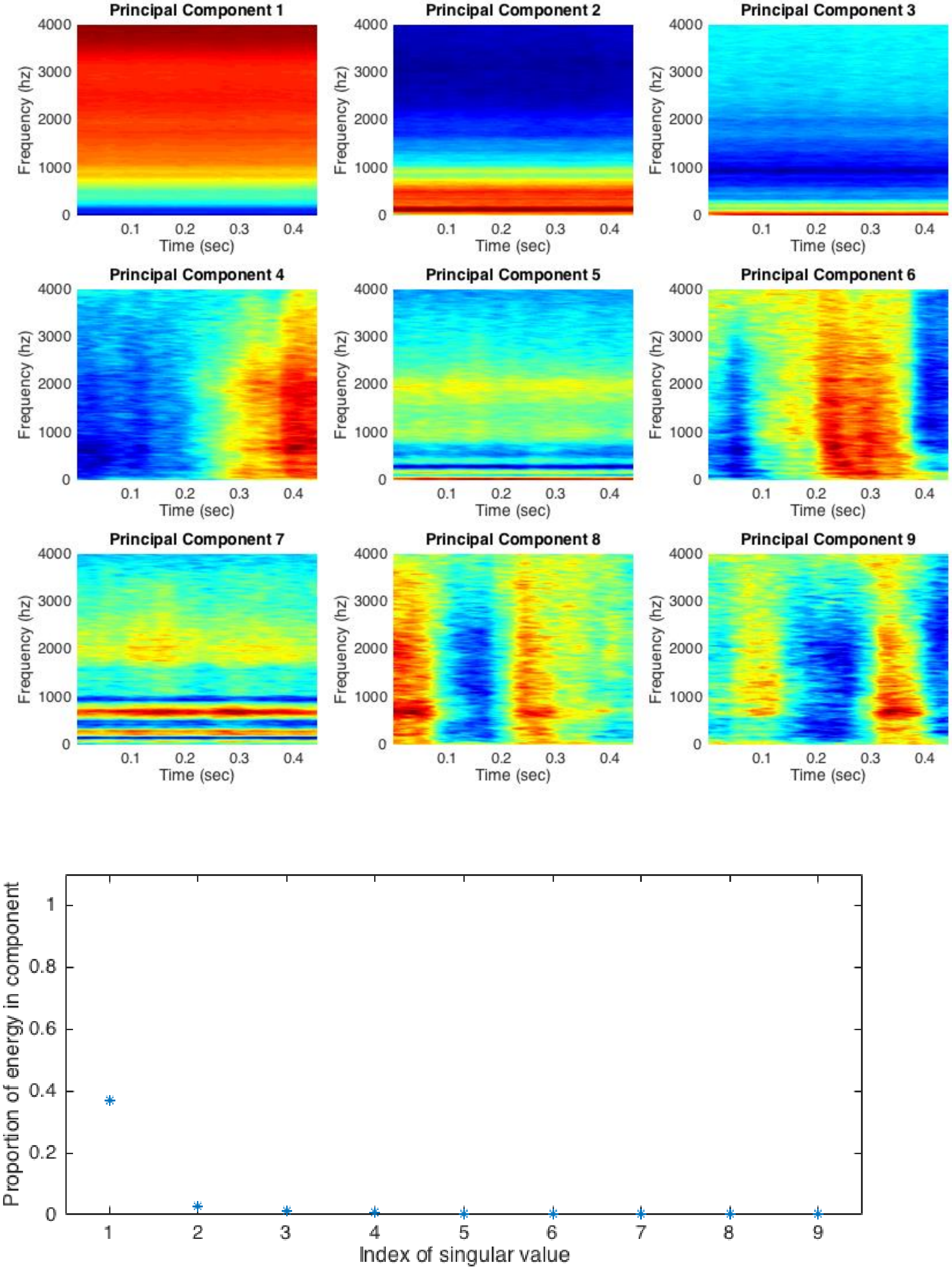

Figure 1. Top: first 9 principal components of PCA for background representation. Components 1 and 2 were used for subtracting the background from spectrograms. Bottom: first 9 singular values, corresponding to the principal components.

The result of applying this algorithm to a 10-second section of a recording is shown in Figure 2. The original spectrograms, spectrograms after background subtraction, and spectrograms after connected-component analysis are all shown. Visually, the segmentation of calls and irregularities from the background flow noise is reasonably successful. At the same time, the spectrograms show the difficulty of classification -- many of the irregularities in the acoustic data have similar structure, frequency range, and are often higher intensity than the calls.

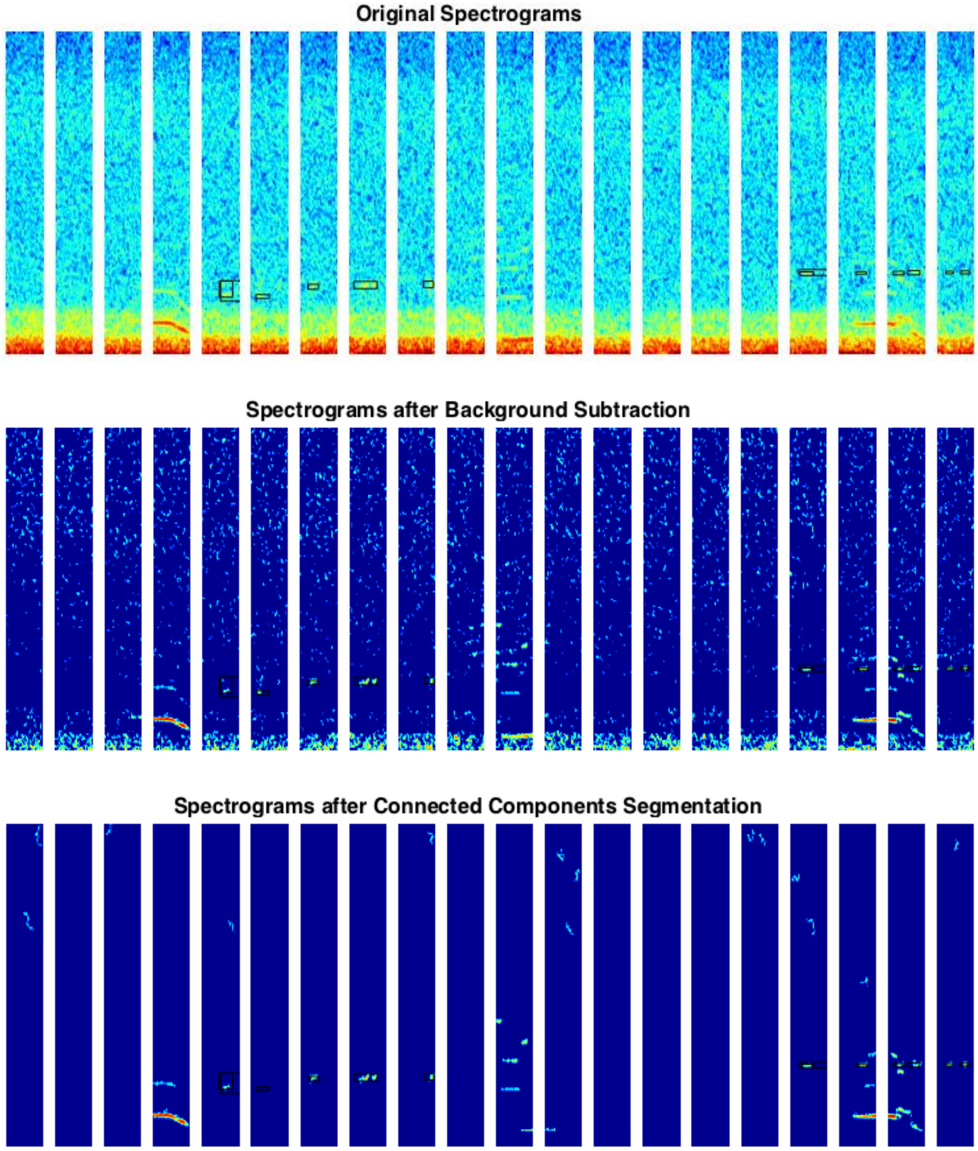

Figure 2. Top: original spectrograms of 20 windows. Labeled events are boxed. Middle: same 20 spectrograms after background subtraction and thresholding. Bottom: same 20 spectrograms after further connected-components-based filtering.

This analysis was also constrained by avoidance of periodic transmitter pings. These “pings” greatly surpassed the signals in amplitude, and had a frequency response spanning the 0-4kHz range, making it impractical to include them during spectrogram processing. After unsuccessfully attempting to detect and remove pings through cross-correlation of the ping waveform with the time series, we resorted to only considering portions of the time series in between the pings. In doing so, we were at risk of excluding potential calls. Future work may consider subtracting this noise by using a different method to characterize signals and noise present in the data, such as singular spectrum analysis.

Overall, feature vectors consisting of 9 features performed better on whale call classes than those consisting of 30. Table 1 shows how fewer features outperformed more features on the combination of all tested data sets. The success was measured by the generally higher precision and recall values for the whale call classes. While the overall accuracy for 30 features was higher, it is evident that this was a result of the high classification rates of the no call class and not of an overall better classification system. The better performance of 9 features could result from the tradeoff between resolution and accounting for variation that occurs with the number of features. Fewer features will put certain calls with slight variation in frequency position in the same bin whereas more features will separate it. While the results for the combination of all data sets are low, we examined the results for each individual data set.

9 Features			**30 Features**		
Overall Accuracy	0.68		Overall Accuracy	0.84	
	Precision	Recall		Precision	Recall
No Call	0.9	0.73	*No Call*	0.89	0.91
Squeak	0.29	0.17	*Squeak*	0.2	0.19
Low Yap	0.14	0.07	*Low Yap*	0	0

Table 1. This table compares the overall accuracy of each data set given the number of features in a feature vector. It gives average precision and recall values for each class, derived from k-fold cross-validation where k=10. The HMMs involved used 2 states. The overall data distribution was 21049 no whale calls, 119 squeaks, and 62 low yaps.

We saw a similar trend in the individual data sets. In about half of the individual data sets, data with fewer features outperformed the data with more features. The first data set saw a 21% improvement in accuracy with 9 features, the second 4%, the third a drop of 5%, and the final data set retained a constant performance. In general, performance was much better in the smaller data sets. Since we already illustrated the risk of examining only accuracy, an individual data set was included to show the improved performances in precision and recall. Table 2 shows a similar comparison across a data set that contains 3692 non whale calls, 29 squeaks, and 19 low yaps. The precision and recall values of each class improved with 9 features, yielding an overall better accuracy. While there are presumably many factors influencing the better performance of an individual data set, we believe that much of the performance loss can be contributed to the variation in the noise that is encountered in each data set. Additionally, the data sets have an unbalanced representation of the number of calls, causing non whale calls to be much more likely than whale calls.

9 Features			**30 Features**		
Overall Accuracy	0.72		Overall Accuracy	0.51	
	Precision	Recall		Precision	Recall
No Call	0.6	0.77	*No Call*	0.49	0.48
Squeak	0.46	0.5	*Squeak*	0.41	0.48
Low Yap	0.5	0.3	*Low Yap*	0.2	0.1

Table 2. This table compares the overall accuracy of a particular data set a given the number of features in a feature vector. It gives average precision and recall values for each class, derived from k-fold cross-validation where k=10. The HMMs involved used 2 states. The data distribution for this data set was 3692 non whale calls, 29 squeaks, and 19 low yaps.

Conclusion

In all, we propose and evaluate a method for automatic detection and classification of humpback whale social calls in underwater acoustic recordings. Our classification accuracies are low, which is likely to be a result of the large variability of noise in the data, as well as an overwhelming number of points that do not contain whale calls. Further work could improve the feature extraction and learning procedures to better account for the presence of this acoustic noise and create a more balanced train data set.

Works Cited

Brown, Judith C., and Paris Smaragdis. "Hidden Markov and Gaussian mixture models for automatic call classification." *The Journal of the Acoustical Society of America* 125.6 (2009): EL221-EL224.

Datta, Sl, and C. Sturtivant. "Dolphin whistle classification for determining group identities." *Signal processing* 82.2 (2002): 251-258.

Guyon, C., Bouwmans, T., and El-hadi Zahzah. “Robust Principal Component Analysis for Background Subtraction: Systematic Evaluation and Comparative Analysis”. *Principal Component Analysis*, edited by Dr. Parinya Sanguansat, 2012.

Oliver, N., Rosario, B., and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831-843, Aug 2000.

Pace, Federica, Paul White, and Olivier Adam. "Hidden Markov Modeling for humpback whale (Megaptera Novaeanglie) call classification." *Proceedings of Meetings on Acoustics*. Vol. 17. No. 1. Acoustical Society of America, 2013.

Ren, Yao, et al. "A framework for bioacoustic vocalization analysis using hidden markov models." *Algorithms* 2.4 (2009): 1410-1428.

Rickwood, Peter, and Andrew Taylor. "Methods for automatically analyzing humpback song units." *The Journal of the Acoustical Society of America*123.3 (2008): 1763-1772.

Seger, Kerri D. *Ambient acoustic environments and cetacean signals: Baseline studies from humpback whale and gray whale breeding grounds*, University of California, San Diego, Ann Arbor, 2016.

Silber, Gregory K. "The relationship of social vocalizations to surface behavior and aggression in the Hawaiian humpback whale (Megaptera novaeangliae)." *Canadian Journal of Zoology* 64.10 (1986): 2075-2080.

Stimpert, Alison K., et al. "Common humpback whale (Megaptera novaeangliae) sound types for passive acoustic monitoring." *The Journal of the Acoustical Society of America* 129.1 (2011): 476-482.