

©Copyright 2016

Nicole Nichols

# Marine mammal species detection and classification

Nicole Nichols

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Mari Ostendorf, Chair

Les Atlas

Ann Bowles

Program Authorized to Offer Degree:  
Electrical Engineering

University of Washington

**Abstract**

Marine mammal species detection and classification

Nicole Nichols

Chair of the Supervisory Committee:  
Professor Mari Ostendorf  
Department of Electrical Engineering

Transient source detection and classification is a particularly challenging problem. Marine mammal vocalizations are a well-known example of these non-stationary sources, including a variety of clicks, pulse bursts and frequency sweeps. There are both environmental and legal needs to improve remote marine mammal monitoring, which can be done efficiently using passive acoustic monitoring (PAM), and public data are available to test new methods. In this thesis, I propose the use of non-negative matrix factorization (NMF) based feature representation for detection and classification of marine mammals for the following reasons: NMF can learn non-stationary signals, training requires less detailed annotations than existing species classification techniques, it can capture species-specific information from non-stereotyped vocalizations, and some NMF-based methodologies incorporate noise removal and session effect compensation. In particular, co-occurrence constraints in NMF analysis were helpful in addressing session effects in species classification.

An additional direction of the research was to minimize the need for strictly labeled training data, which is arduous to create and thereby limits performance. I investigated weakly supervised learning techniques to leverage data with incomplete annotations. In these trials, recordings were made in the visual presence of a single species, but there were no annotations to indicate when vocalizations occurred. Automated detection algorithms identified potential vocalizations and then confidence-based selection methods filtered the best examples in an iterative training procedure. This method was particularly beneficial



for species classification from clicks, which is very sensitive to on- and off-axis variations. Changes in orientation make the signal more variable and interfere with establishing consistent features for species classification.

Weakly supervised species classification from clicks automatically identified the clicks that were most representative of species. This method improved species classification by 7-15% as compared to models built with all detected clicks. Weak supervision for updating noise bases also lead to a 30% reduction in cross species error for a mismatched data scenario in species classification based on whistles. Together these methods contributed to algorithm improvements for transient source detection and classification system. Further experiments in marine mammal identification using NMF would provide further understanding of methods for compensating for variability associated with recording conditions.



## TABLE OF CONTENTS

	Page
List of Figures . . . . .	iii
Chapter 1: Introduction . . . . .	1
1.1 Thesis Contributions . . . . .	5
1.2 Thesis Outline . . . . .	7
Chapter 2: Background . . . . .	9
2.1 Marine Mammal Detection . . . . .	9
2.2 Species Classification . . . . .	12
2.3 Non-Negative Matrix Factorization . . . . .	14
2.4 Learning from Data with Incomplete Labels . . . . .	19
2.5 Summary . . . . .	22
Chapter 3: Classification System Design . . . . .	23
3.1 Data Description . . . . .	23
3.2 Experiment Paradigm . . . . .	30
3.3 Performance Evaluation . . . . .	35
3.4 Summary . . . . .	39
Chapter 4: Variations of Non-Negative Matrix Factorization for Marine Mammal Whistles . . . . .	40
4.1 Noise Training . . . . .	41
4.2 Multi-pass Training . . . . .	42
4.3 Co-occurrence Constraints . . . . .	42
4.4 Shuffle Training . . . . .	43
4.5 Whistle Detection Experiments . . . . .	43
4.6 Whistle Detection Comparison with Silbido . . . . .	51
4.7 Summary . . . . .	52

Chapter 5: Species Classification from Whistles . . . . .	54
5.1 Species Classification Methods . . . . .	54
5.2 Experiments . . . . .	55
5.3 Summary . . . . .	63
Chapter 6: Weak Supervision . . . . .	65
6.1 General Methods . . . . .	65
6.2 Clicks . . . . .	67
6.3 Whistles . . . . .	71
6.4 Summary . . . . .	80
Chapter 7: Conclusions and Future Work . . . . .	83
7.1 Summary of Contributions . . . . .	83
7.2 Future Work . . . . .	85
Bibliography . . . . .	89
Appendix A: Appendix A . . . . .	99
A.1 File List for Silbido Annotated partitions 1,2,3 . . . . .	100
A.2 File List for Hand Annotated partitions A,B,C . . . . .	112
A.3 File List for Click partitions A,B,C,D,E . . . . .	114

## LIST OF FIGURES

Figure Number	Page
1.1 Example spectrograms of vocalizations:(a) Clicks (b) Whistles (c) Burst-pulse	3
1.2 Spectrogram of a killer whale vocalization, which exhibits biphonation. Source: Hubbs-SeaWorld Research Institute . . . . .	4
3.1 Example of whistle unions, single whistles are denoted in pink, with pink rectangles to outline start and end times. Yellow boxes denote the start and end times of whistle unions. For single whistles the union start-end times are unchanged. . . . .	24
3.2 Block diagram of system modules . . . . .	26
3.3 Noise Removal Process: (a) Original spectrogram (b) Detected clicks (c) Median smoothing estimate of background noise, and (d) Result of noise removal process. . . . .	29
3.4 Cross validation partitions. Purple is a partition with hand annotated whistles, green is a partition without hand annotated whistles. Each row corresponds to a cross validation fold. The A,B,C and 1,2,3 partitions are used in only one stage of the experiment to keep training and evaluation data independent, but the permutations are different among cross validation folds to provide variability. . . . .	32
3.5 Five-fold cross validation partitions used for weakly supervised species classification from click experiments. Common subsets between click and whistle partitions are not readily summarized, however exact file lists for all partitions are provided in Appendix A . . . . .	36
4.1 Comparison of sample whistle bases using standard and joint training . . . .	42
4.2 F-score comparison for one-pass and two-pass NMF bases training . . . . .	45
4.3 Median performance with varying GMM's for (a) 20N/30W (left) and (b) 20N/60W (right) . . . . .	46
4.4 Per fold performance for (a) 20 noise/60 whistle GMM10 and (b) 20 noise/60 whistle GMM5 configuration . . . . .	47
4.5 Comparison of Q1 and Q2 co-occurrence constraint relationships . . . . .	48
4.6 Comparison of median performance with NoQ, Q1, and Q2 co-occurrence constraints. . . . .	48

4.7	Comparison of 20N30W NoQ GMM5 median performance with and without shuffle training . . . . .	50
4.8	Median and per fold performance of Shuffle (left; 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint), and No Shuffle (right; 20N30W joint bases, GMM5, no shuffle and no co-occurrence constraint) . . . . .	50
4.9	Comparison of species-dependent and -independent whistle detection . . . . .	51
4.10	Median performance of best detection configuration which used 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint. This best performance is labeled (E20) in the figure. The configuration with the least variance across folds had 20N60W joint bases, GMM5, no shuffle and no co-occurrence constraint. This configuration is labeled (E8) in the figure. Silbido, including maximum and minimum Silbido performance are labeled accordingly. . . . .	53
4.11	Median and per fold performance of best detection (left; 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint), and detection with least variance (right; 20N60W joint bases, GMM5, no shuffle and no co-occurrence constraint) . . . . .	53
6.1	Comparison of MAP vs. Vote decision functions for best threshold configuration ( $\tau_c = 0.7, \tau_n = 0.6$ ): (a) Cross Species Error and (b) Macro Error . . . . .	68
6.2	Comparison of Threshold settings: (a) Cross Species Error and (b) Macro Error . . . . .	70
6.3	Comparison of GMM re-seeding for ( $\tau_c = 0.7, \tau_n = 0.6$ ): (a) Cross Species Error and (b) Macro Error . . . . .	71
6.4	Comparison of GMM re-seeding for ( $\tau_c = 0.7, \tau_n = 0.7$ ): (a) Cross Species Error and (b) Macro Error . . . . .	72
6.5	Comparison of GMM re-seeding for ( $\tau_c = 0.6, \tau_n = 0.6$ ): (a) Cross Species Error and (b) Macro Error . . . . .	72
6.6	Comparison of GMM re-seeding for ( $\tau_c = 0.6, \tau_n = 0.7$ ): (a) Cross Species Error and (b) Macro Error . . . . .	73

## ACKNOWLEDGMENTS

I would like to express sincere appreciation, to all who have contributed to this effort. I am particularly grateful for the feedback and guidance of my committee, Mari Ostendorf, Ann Bowles, Les Atlas, and Peter Dahl. The support of friends, family, labmates, colleagues, and coworkers has been tremendous. Thank you!



## Chapter 1

### INTRODUCTION

Passive acoustic monitoring (PAM) is a cost effective new tool to help monitor marine mammals, providing data on presence, population status, habitat usage patterns, etc. Buoys can be deployed in locations that are remote or where conditions are likely to make surface-based observations impractical. Such PAM data has significantly improved temporal resolution of the baseline movement and behavior information crucial for policy. However, there are still a lot of challenges to processing and interpreting the data [41]. Initially, post-processing was exclusively conducted via trained human listeners, but it is a very time consuming and subjective process. Computer automation of some tasks, such as detection, have achieved limited success [8, 101, 78]. Other tasks such as species identification are in research development and have not been adopted for routine use due to the difficulty generalizing for multiple species [66, 9, 72].

There are numerous factors that contribute to the difficulty of automated acoustic detection and classification of marine mammals. Improvement of automated detection and classification algorithms, is the aim of this thesis. Foremost among the difficulties in PAM data processing is the assortment of population and regional differences in vocalizations within species [102]. Marine mammals are a heterogeneous group of taxa that evolved adaptations allowing them to spend a large proportion of their time in marine waters, including polar bears; pinnipeds (seals, sea lions, and walruses); sirenians (manatees and dugongs); and the Cetartodactyla, which include whales, dolphins, and porpoises. The Cetartodactyla are divided into two suborders, the odontocetes (toothed whales, including dolphins and porpoises) and mysticetes (baleen whales), which have very different behavior and acoustics. Broadly speaking, odontocetes are smaller, faster, eat fish, squid or other mammals, have sophisticated echolocation systems and produce social vocalizations with dominant energy in the range of 400Hz-50,000Hz. Baleen whales are typically much larger, slower, filter feed

on krill, plankton or small fish, are not known to echolocate, and produce low frequency vocalizations in the range of 6Hz-4,000Hz [40]. The methods I describe in this thesis were developed with data representative of small to mid-sized odontocete vocalizations.

The descriptions of marine mammal vocalizations that occur in the literature include grunts, groans, moans, chirps, boings, bursts, etc. These terms indicate the diversity of the sounds, but lack quantifiable characteristics. For this work, I have followed the classification established by [69] and used by [73], which group marine mammal vocalizations into three broad categories: clicks, whistles, and burst-pulses. Clicks, as shown in Figure 1.1a, can be very short (less than 1 ms for echolocation clicks of some species) [104, 3, 53], and relatively broad in frequency range. Trains of clicks and pulses that do not function in echolocation are also included in this category, so there can be overlap with the category of burst-pulses. When echolocation click trains are isolated based on heuristic criteria, species differences in spectral characteristics can be identified [25]. Whistles, as shown in Figure 1.1b, have relatively narrowband frequency components, characterized as tonals, that sometimes exhibit harmonic structure and are generally about 0.5-3s in duration. Burst-pulses, as shown in Figure 1.1c include a heterogeneous assortment of long duration broadband sounds that may be difficult to distinguish from click trains if the vocal behavior of a species is poorly understood. The particular example in Figure 1.1c shows trains of clicks so closely spaced that harmonic structures begin to emerge, characteristic of FFT-based spectrograms with sampling rates and windowing typical of vocalization analyses.

The narrowband components in whistles are not pure tones, but the result of tight interpulse interval spacing and choice of sampling window [97]. The example in Figure 1.1c shows overlapping burst-pulse sounds and apparently pure-tone whistles, a frequent finding when recordings of large schools are made. However, although supporting literature on production mechanisms is incomplete, the best evidence [50] suggests that the whistles and burst-pulse sounds are produced using similar mechanisms. In Figure 1.1c, the segment from 1.2 seconds to the end is free of background whistles, illustrating burst-pulse sounds.

Given the wide range of acoustic characteristics among the three vocalization classes and the possibility of intergradations, feature detectors and classifiers must be tuned to particular time-frequency patterns to perform with sufficient accuracy.

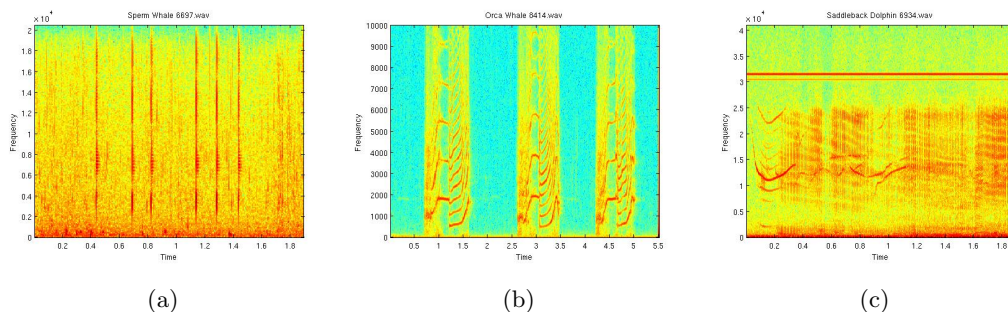


Figure 1.1: Example spectrograms of vocalizations:(a) Clicks (b) Whistles (c) Burst-pulse

Ambient noise is pervasive in the ocean and another major challenge for PAM data processing. Natural phenomena such as wind, rain and waves are dominant sources. Human activities such as shipping traffic and offshore energy development also contribute significantly to the acoustic environment, particularly in the northern hemisphere. Quantitative contributions of sound sources in the ocean were classically discussed in Wentz et al. [98] and the topic continues to be an active area of research [19]. Marine mammals themselves are often significant contributors [4], even in areas with intensive human activity.

Some species, such as common dolphins (*Delphinus delphis*), are quite social, traveling in large groups of tens or hundreds of individuals [72, 6] that can be composed of more than one species. Detecting the presence of one or more secondary species amid the background of constant vocalizations from a large school is a formidable challenge.

The transmission of sound waves through air and sea water are quite different processes. For the purpose of modeling sound transmission, air is relatively homogeneous at short to moderate ranges. Sea water, however, is often stratified even at short ranges because sound speed is strongly affected by temperature, pressure, salinity, and local parameters such as the inflow of terrigenous fresh water. In addition, the substrate and bottom topography affect propagation. The frequency patterns of a source signal become distorted as sound travels through the ocean to the point of reception, with higher frequencies also affected by adsorption. Frequently, the end result is frequency spreading as well as temporal distortion introduced by reverberation [37]. Recordings of odontocetes are almost always recorded at distances where such phenomena become important, making it more difficult to recognize

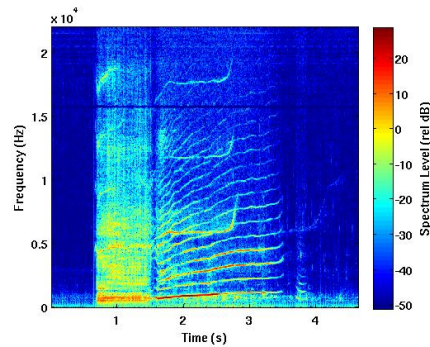


Figure 1.2: Spectrogram of a killer whale vocalization, which exhibits biphonation. Source: Hubbs-SeaWorld Research Institute

particular time-frequency patterns reliably.

Compensation techniques include using a variety of feature sets such that each provides some invariance to the distorting effects. For example it has been shown that modulation features are more invariant to distortion in highly reverberant environments as demonstrated in a speech recognition task [44]. The vocal systems of small to mid-sized social delphinids universally include whistled vocalizations with prominent modulation features. Odontocetes have excellent control over the sounds they produce, although production mechanisms are still not perfectly understood. The most recent research [18] indicates that pressure differences across vibrating phonic lips, which are located immediately behind the melon, are responsible for producing most sounds. Given the two vibrating structures, it is not surprising that a number of species are capable of biphonation [29], production using two independent and simultaneous vocal sources. An example can be seen in Figure 1.2, which clearly shows two independently modulated harmonics as recorded from an individual killer whale (*Orcinus orca*). From a classification standpoint, methods developed to extract harmonic contours from a single source, already susceptible to noise interference, could be further disrupted by biphonation.

An additional data processing complication of odontocete sound production arises from the capacity to control the direction of sound being produced. The melon, a structure composed of fatty tissue that lies in the front of the head, acts as a lens to focus outgoing

sounds, at least at echolocation frequencies. Echolocation thus produces a highly directional signal with significantly less signal strength in the off-axis direction and a different pattern of relative intensity across frequency. Significant differences occur when clicks are recorded as little as  $15^\circ$  off axis. For an illustration of these differences, refer to Madsen [52], Figure 2, which shows both the time domain waveform and frequency domain representation of individual clicks recorded at a variety of angles to the vocalizing animal.

Signal processing of odontocete vocalizations has historically focused on echolocation. More recently, work has focused on ensemble click trains, without reference to function or directionality [7, 82] and the whistles of delphinid cetaceans (including the tonal components of complex killer whale vocalizations) [5, 67, 13]. Computer automated extraction of a marine mammal’s whistle contour can be error prone due to many sources of variability, particularly when many vocalizations overlap [71, 38]. Unfortunately, this is a common first step for many marine mammal species classification algorithms, which first extract the whistle contour, then parameterize the shape and assign a species prediction based on the resulting features. Particularly in a noisy environment, it is easy for these algorithms to extract only partial whistle contours or misidentify other sounds, even noise, as whistles, causing a problem of compounding errors. Accurate contour extraction of marine mammal vocalizations is an active topic of research [79]. In a state-of-the-art algorithm [71], it was reported that average deviation from hand annotated frequency contours was 161 Hz,  $\sigma = 51$  for a particle filter based approach and 70 Hz,  $\sigma = 76$  for a graph search approach.

### **1.1 Thesis Contributions**

This thesis makes technical contributions to the passive acoustic tools used for marine mammal detection and classification for use in species surveys. Three contributions are presented.

First is a demonstration that non-negative matrix factorization (NMF) is a valuable method for transient whistle detection and species classification that has several advantages over prior methods. I addressed the challenge of automated whistle contour extraction by looking for a classification feature that could bypass or complement explicit contour extraction. NMF is a method of decomposing a matrix into two matrices (the bases and

weights) which when multiplied, reconstruct the original matrix. In the training stage, NMF bases, a learned dictionary, were designed to represent physically meaningful frequency patterns in training data representing species-specific or species-independent whistles, or associated noise. The activation weights for these bases then served as a classification feature. This approach made no assumptions about the structure of the vocalizations, but simply learned from the data. Thus this method could be used with non-stereotyped and complex vocalizations which, depending on behavioral context can account for as much as 4.8-34.5% of the repertoire of killer whales [31], a species often studied because of its well cataloged repertoire of stereotyped calls [20, 86, 30]. For other species, the proportion can be substantially higher. Historically, non-stereotyped and particularly burst-pulse vocalizations have been excluded from species classification because of incompatibility with template matching methods. NMF has the potential to detect and classify these sounds as well.

A second contribution of this work was the use of co-occurrence constraints which impose penalties to prefer defined patterns of weights [87]. The original motivation for using constraints between species or noise-specific subsets of the collective set of bases was to minimize the impact of potentially redundant bases learned for each species, since species bases are learned separately then combined to create the species classification models. Experimental results instead showed that co-occurrence constraints reduced the impact of session effects.

To address the challenge of having only small or incomplete annotations of the true marine mammal vocalizations in the data, I developed a weakly supervised training procedure to make use of partially labeled data, i.e. data for which species was known, but individual vocalizations were not marked and classified. Hypothesized vocalizations are identified by the click detection method described in this thesis, or by *Silbido*, a state of the art whistle detection algorithm.<sup>1</sup> I sorted hypothesized vocalizations by probabilistic confidence, as determined by the Gaussian mixture model, and included the most confident examples in model training. This third research contribution allowed model training to span a much greater volume of training data and consequently made the resulting system more robust.

---

<sup>1</sup>[http://roch.sdsu.edu/software/silbido\\_JASA2011baseline.zip](http://roch.sdsu.edu/software/silbido_JASA2011baseline.zip)

Applied to both clicks and whistles, it was particularly successful for clicks, showing 7-15% improvement and potentially allowing species classification from clicks to automatically select the on- or off-axis clicks that best separated the species classes. Whistle based weak supervision was more susceptible to session effects and further work will be needed before weak supervision can realize its full potential for the whistle domain. Weakly supervised training did not benefit the whistle performance in the same way as clicks, but it did help with session effects. Updating just the noise bases reduced the cross species error by 30% and macro-error by 15% for mismatched evaluation data.

These contributions are assessed experimentally using data from the 5th International Workshop on Detection, Classification, Localization and Density Estimation of Marine Mammals using Passive Acoustics.<sup>2</sup> Files are provided for three species bottlenose dolphins (*Tursiops truncatus*), melon-headed whales (*Peponcephala electra*), short- and long-beaked common dolphins (*Delphinus delphis* and *D. capensis*), and contain both clicks and whistles. However, only a small portion of files have whistle annotations and no files have click annotations. This work serves to advance the overall capabilities of marine mammal monitoring systems. Though the work was developed for odontocete species, the methods are applicable to other transient acoustic sources.

## 1.2 Thesis Outline

The thesis is structured as follows:

- Chapter 2 describes related prior work on marine mammal detection and species classification of clicks and whistles. Note that burst pulse sounds have been pooled with clicks to some degree in these studies. I provide background information on several variations of non-negative matrix factorization (NMF), the mathematical underpinning of the first contribution. I additionally review techniques for learning classifiers from partially labeled data, which informed the methodologies of the third, weak supervision contribution.

---

<sup>2</sup><http://www.bioacoustics.us/dcl.html>

- Chapter 3 discusses all aspects of the classification system architecture and experimental framework for the thesis, including data, data variability, data partitioning strategy, noise removal techniques, feature extraction, classification models, decision functions and methods used to evaluate performance.
- Chapter 4 details the design of the NMF training algorithm, refinements made to improve whistle detection results, and experiments on whistle detection. The whistle detection shows regimes where NMF-based detection was more effective than the state-of-the-art Silbido method, specifically when less fragmentation is desired. Initial analysis of co-occurrence and basis training strategies are also discussed.
- Chapter 5 describes the system for species classification and experimental results using different configurations for NMF. These results discuss the influence of general parameter settings that contribute to the effectiveness of the first contribution and further analysis of the co-occurrence constraints related to session effects in the second contribution.
- Chapter 6 discusses the strategy for implementing weak supervision for click and whistle based species classification, which is the third contribution of this thesis, and describes experiments demonstrating the impact on classification performance. Analysis of session effects provides insights into the impact of session effects on all contributions.
- Chapter 7 summarizes the impact and conclusions of the research and additionally outlines potential directions for future work.

## Chapter 2

### BACKGROUND

#### **2.1 Marine Mammal Detection**

##### *2.1.1 Clicks*

In an end-to-end marine mammal species classification system, the first task is detection of marine mammal vocalizations. Because clicks and whistles have opposite time frequency characteristics, detectors are usually optimized for one type of vocalization or the other. Much of the existing research on click detection has emphasized sperm whales, delphinids, harbor porpoises, and beaked whale species because clicking is a defining acoustic behavior of these species and many of these species can be detected more efficiently by PAM. Several methods exist, such as binary thresholded FFT [58] and the Teager-Kaiser energy operator [43]. A quantitative comparison of six additional methods for beaked whales was presented in [101]. The six methods included several publicly available tools, (Ishmael, XBAT and PAMGUARD), and three additionally published methods, (ERMA (an energy band ratio test), a GMM, and a frequency modulated echolocation click detector (FMCD)). Accuracy ranged from 67-89% correct for a beaked whale click detection task, and the highest accuracy was achieved by a Gaussian mixture model (GMM). The approach I chose to use for click detection is most similar to that of Roch et al. [72], the GMM approach that showed the best performance. The procedure is detailed in Section 3.1.2.

##### *2.1.2 Whistles and Burst-pulse Sounds*

Detection methods for whistle vocalizations are more complex because of the non-stationary, transitory, and harmonic properties that make it more difficult to describe the target signal. The tonal vocal repertoires for species such as killer whales also have regional variations which often force detection methods to be tuned to a specific species or sub group. This specialization makes methods difficult to generalize and compare. A review of the literature

showed that whistle detection methods can be grouped into 1) kernel based classifiers which simultaneously detect and classify species or call type based on one set of features [83, 54, 1, 56] or 2) separate detection and classification stages that allow different features and classifiers to be used for the two tasks [35, 8, 71, 67, 65, 73]. A detection method based on information entropy [24] that was designed to generically detect marine mammal tonal vocalizations from any species has also shown some promise. This method used normalized power spectral density (PSD) to compute an entropy metric ( $PSD * \log(PSD)$ ). Entropy measures the peakedness in the PSD that was observed to occur for a variety of marine mammal species vocalizations. Detection was recorded when the instantaneous PSD entropy exceeded the median by a threshold times the standard deviation.

The kernel based methods have predominantly been applied to baleen whales, which were not the target of this research. However, the detection method is worth contrasting. Kernel based detection functions by using a set of templates that are representative of the audio signals of interest. By taking a cross correlation function with each template and identifying a threshold that yields satisfactory detection results with respect to false alarms, signals of interest can be identified. The advantage of this method is that species classification is inherent in the process as the templates correspond to species specific vocalizations. Additionally, it is simple enough to operate in a real-time environment. The drawback is that if there is high variability in the call type or the vocal repertoire is unknown, it can be difficult to build an accurate library of templates. As an example of performance, for a humpback whale detection task, kernel based detection showed 74% correct detection [1]. In this study data were collected from Hawaii, Alaska and Stellwagen Bank, in one year and tested from an independent set recorded at one of the three locations. Humpback whales produce complex, stereotyped song elements with components ranging from tonals to short transients.

Because of the high variability in dolphin whistles, traditional kernel-based methods are inadequate. Most of the other approaches to whistle detection focus on using some method of extracting the whistle contour, then parameterizing the minimum frequency, maximum frequency, duration, derivatives and other similar metrics as input features to a classifier. Early acoustic analysis relied on human annotation of these frequency contours [84, 70], a

very time consuming task. More recent studies have moved to semi-automated methods [14, 32], where computational algorithms estimate an initial frequency track which is then manually corrected or users define starting points for algorithmic extraction of frequency contours. A variety of creative methods have been demonstrated in the literature. For example, Baumgartner et al. [8] describe a pitch tracking algorithm which minimizes a cost function that estimates the most probable whistle path given a starting coordinate. Roch et al. [71] perform contour extraction via a particle filter and Parada et al. [67] compare two audio coding algorithms, called the unpredictability measure [85] and the MUSIC algorithm [77]. The unpredictability measure assumes the amplitude and phase of tonal signals is more predictable than for noise or other impulsive sounds. The difference between predicted and actual amplitude and phase are combined to create the unpredictability measure which then extracts an estimated frequency contour for the whistles. The MUSIC algorithm was used to estimate broad class labels (noise, pulses, whistles, or pulses and whistles) by using the variance of an estimate of the dominant frequency. These pitch tracking algorithms work well in clean audio and some have shown some robustness to noise but quantitative comparison has not been published. However, many research groups have made their code available including PAMGUARD [34], Silbido, Triton, and Ishmael. I chose to compare the NMF approach with the Silbido package as they have published results [72] on a superset of the data used herein. The Silbido method is based on pitch tracking with a graph search algorithm.

As discussed in Section 1, pitch tracking algorithms tend to break down in the presence of noise, resulting in compounding errors when species classification features are extracted from the error prone tracks. A new technique for detection and manipulation of speech that does not rely on pitch tracks is non-negative matrix factorization (NMF). This technique, NMF, has successfully been applied to a variety of challenging tasks including separation of overlapping speakers from a single microphone [80, 92] and noise removal [100]. The generic process of using NMF for classification operates in two stages, one to learn a set of bases which are additive and can be representative of physically meaningful components in the data. The second stage is used to decompose arbitrary data by assigning weights to the learned bases. The weights are then used as classification features because they represent

the amount of each basis type in the observed data at each time frame. I compare the impact of a variety of NMF bases training methods, and calculation details are discussed in section 2.3. I demonstrate the effectiveness of NMF to perform whistle detection without explicit extraction of the frequency contours.

## **2.2 Species Classification**

### *2.2.1 Clicks*

There has been significant prior effort to classify marine mammal species from their vocalizations [9]. There are several ways to approach the species identification task. For species with a limited repertoire of vocalizations, such as the North Atlantic right whale, autocorrelation or kernel based methods can simultaneously detect and classify species [55, 54, 88, 59]. However, because this method was based on a threshold parameter and no standardized evaluation data was available, comparisons of method accuracies are not meaningful. For species with a more diverse vocal repertoire, additional techniques are used but approaches typically focus on either clicks or whistles. Burst-pulse sounds are often overlooked even though they can represent a significant portion of the repertoire for some species. Lu et al. have put some effort to combine click and whistle information [48]. They used cepstral features were used for clicks and several whistle classification features were extracted from contours including maximum, minimum and median frequency, duration, slope, and number of inflection points. Features were used individually when only whistles or clicks were present, but merged when whistles and clicks overlapped. A random forest classifier was built using random subsets of the data. Results were compared to using clicks or whistles alone, and the joint classification greatly improved whistle classification but had little impact on click classification.

With regard to clicks, the temporal and spectral characteristics of sperm whale, beaked whale and porpoise clicks are quite distinct from those of schooling delphinid species and thus much of the click classification effort has focused on distinguishing among dolphin species, [7]. Numerous publications have demonstrated the potential effectiveness of species classification from clicks [36, 33, 73, 72, 82, 74].

In an analysis of click spectral features such as peak frequency, center frequency, duration, and inter-click interval, discriminant function analysis resulted in 54-93% accuracy for the three species included in the particular classification task [7]. In later work, clicks were parameterized by 14 cepstral coefficients for bins between 10-92kHz. These features were classified with 16 mixture GMMs and showed a 78% accuracy over 6 dolphin species [72]. Based on the comparative success of this approach [57], I chose to use Mel-frequency cepstral coefficients (MFCC) and GMMs for the click classification baseline.

There have been numerous studies documenting significant differences in the on- and off-axis Spectral characteristics of clicks from bottlenose dolphins [93, 2] and other odontocete species [51, 49]. It is commonly understood that this and other variability contributes to the limits of successful species classification. Two recent studies noted the need to average large groups of clicks to compensate for on- vs. off-axis variability when computing species classification [48, 72].

The contribution here directly addresses this limitation by using weakly supervised learning to identify which clicks are most discriminative of species. Background information on this learning technique is discussed in section 2.4, and the results of these experiments are discussed in Chapter 6.

### 2.2.2 Whistles

Methods for species identification from whistles have similarly focused on dolphin species. Early work by Steiner [84] used linear discriminant analysis with 6 variables to describe frequency and inflection of the whistles. Correct classification ranged from 57-80% for the 5 dolphin species. In a study by Oswald et al. [65], 12 features were extracted from the whistle contours of 9 dolphin species. These features measured beginning, ending, minimum and maximum frequency, duration, and slope. Using a CART classifier accuracy for individual species ranged from 24%-88% with a global correct classification rate of 51%. This work was extended by Gannier et al. [32] to include more feature parameters and accuracies ranged between 46-79% correct classification for each of the individual species and 62% global accuracy. The most comprehensive system is based upon this method but expanded

to use 54 feature parameters and a random forest classifier [5]. In Parada et al., [67], GMM classification was performed based on features derived from their extracted contours and unpredictability metric. In a task classifying four species (three delphinids and pilot whales), correct species classification ranged from 30.0%73.3%. The wide range of accuracy in these whistle-based classifications likely reflects a combination of factors, including variability in the behavioral state of callers, difficulty in identifying the species present (particularly for the common dolphin), and session effects. The highest accuracies were obtained for the pilot whale species, but the authors analyzed only 6 seconds of data.

The authors attempted to control for the data imbalance by conducting an experiment with all species limited to 6 seconds of data. This test also used a leave-one-out strategy where all but one whistle was used to train species models, which were then used to evaluate the held out whistle. This resulted in a similar performance spread, 43%-73%)

My approach to species classification combined the benefits of NMF, which bypasses the need for frequency tracking, and weakly supervised learning to maximize the training benefits that can be achieved from partially labeled data.

### **2.3 Non-Negative Matrix Factorization**

Non-Negative matrix factorization (NMF) is one of many ways to decompose a matrix into bases and weights that can estimate and reconstruct the original data [68]. Each method has its own properties and advantages. For NMF, the principal advantage is the additive and intuitive interpretation of the bases [45]. In spectrogram reconstruction, the 1-D bases represent the frequency pattern of a single time frame. Here, I used the 2-D convolutive bases described in Section 2.3.1. Bases are representative of the real frequency patterns observed in the labeled training data. Reconstruction is intuitive because it is a weighted sum of these learned, interpretable patterns. The simplest formulation of NMF represents a length- $n$  sequence of  $c$ -dimensional vectors  $V_i$ ,  $V = [V_1 V_2 \cdots V_n]$ , with an approximate reconstruction  $R = [R_1 R_2 \cdots R_n]$  where  $R_i = \sum_{j=1}^r H_{ij} W_j$ . The weight vectors  $H_i$  and  $r$  basis vectors  $W_j$  are chosen to minimize mean-squared error of the reconstruction. The bases are learned in task-specific training, then fixed when learning weights (the features used in classification). The elements of the vectors  $V_i$  are assumed to be non-negative, and

so all the elements of the  $r$  basis vectors  $W_j$  and weights  $H_{ij}$  are required to be non-negative. The matrix formulation of the problem is as follows:  $V \approx R = WH$  where

$$\begin{aligned} V &= c \times n, \text{ original matrix} \\ R &= c \times n, \text{ approximation of the original matrix} \\ W &= c \times r, \text{ set of } r \text{ NMF basis vectors, each of dimension } c \\ H &= r \times n, \text{ sequence of } r\text{-dimensional NMF weight vectors} \end{aligned}$$

NMF gained attention in the classic 1999 paper by Lee and Seung [45]. In this and a subsequent paper detailing the computation [46], Lee and Seung demonstrate how bases are learned via an iterative update procedure which minimizes a cost function. Any distance metric or divergence function can be used to construct the cost function. The most common metric is the Euclidian distance, in which case the cost function is:

$$f(W, H) = ||V - WH||^2 \tag{2.1}$$

One of the impacts of the Lee and Seung paper was to popularize the multiplicative update approach to solving these cost functions,

$$\begin{aligned} W &\leftarrow W \frac{-VH^T}{WHH^T}, \\ H &\leftarrow H \frac{-W^T V}{W^T W H}, \end{aligned}$$

where the T superscript indicates transpose. In their paper [46], they show a proof of convergence and contrast this method with traditional gradient descent and additive update rules. Some additional key insights regarding the computation of the update rules can be found in Fevotte et al. [28]. There are numerous extensions of the basic NMF formulation, such as bases normalization, sparsity constraints, online computation, and convolutive bases that include an additional temporal component. My system combines several of these extensions, as described in the next subsections.

### 2.3.1 Convolutional Non-Negative Matrix Factorization

With a traditional NMF formulation, each columnwise basis vector is of size  $c \times 1$  and the collection of  $r$  bases is  $c \times r$ . It is important to emphasize that each  $c \times 1$  basis vector is a separate unit from every other basis vector, and no temporal structure is captured. Because temporal structure is important for capturing important information such as smoothness, a method for incorporating some temporal features has been added for many of the successful speech processing applications. To incorporate this type of information into an NMF framework, Smaragdis [80] introduced a convolutional extension of the traditional NMF computation. In this formulation, the basis matrix  $\mathbf{W}$  has dimensions  $c \times r \times t$ , where  $t$  is the temporal length in time frames for each of the  $r$  bases. In this case, the reconstruction of the original matrix is achieved by a shifted sum of products:

$$\mathbf{R} = \sum_{p=0}^{T-1} W_p \vec{H}^p \quad (2.2)$$

The boldface  $\mathbf{R}$  notation is used to distinguish the convolutional reconstruction from the simpler vector approach. The notation  $\vec{H}^p$  implies that the weight matrix  $H$  is shifted by  $p$ , with zeros filling in behind. For example:

$$\begin{array}{r} \vec{H}^0 = \\ \begin{array}{cccc} 1 & 5 & 9 & 13 \\ 2 & 6 & 10 & 14 \\ 3 & 7 & 11 & 15 \\ 4 & 8 & 12 & 16 \end{array} \end{array} \quad \begin{array}{r} \vec{H}^1 = \\ \begin{array}{cccc} 0 & 1 & 5 & 9 \\ 0 & 2 & 6 & 10 \\ 0 & 3 & 7 & 11 \\ 0 & 4 & 8 & 12 \end{array} \end{array} \quad \begin{array}{r} \vec{H}^2 = \\ \begin{array}{cccc} 0 & 0 & 1 & 5 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 3 & 7 \\ 0 & 0 & 4 & 8 \end{array} \end{array}$$

Substituting this revised computation of the reconstruction into the cost function, new multiplicative update rules can be derived [80] and are defined:

$$W_p \leftarrow W_p \frac{-V(\vec{H}^p)^T}{W_p \vec{H}^p (\vec{H}^p)^T} \quad (2.3)$$

$$H \leftarrow H \frac{-W_p^T V}{W_p^T W_p (\vec{H}^p)^T} \quad (2.4)$$

This formulation of NMF has been applied to some very challenging speech processing tasks, including separation of overlapping talkers from a single microphone [80, 60] and noise robust speech recognition [91].

### 2.3.2 Weight Sparsity and Constraints

NMF can be used as a means to perform data compression, particularly when there is sparsity in the weight parameters such that any given time point can be represented by weights of just a few primary bases. In order to achieve this sparsity, a regularization term  $g(H)$  is added to the cost function:

$$f(W, H) = \|V - \bar{W}H\|^2 + \lambda g(H). \quad (2.5)$$

The  $g(H)$  term is often defined to be an L1 norm <sup>1</sup> when sparsity is a goal. In this thesis I was interested in using NMF for classification, not compression, therefore drew on co-occurrence constraints, an alternative type of sparsity, as introduced by Tjoa et al. 2010 [87]. In this formulation, rather than having a sparsity constraint for each individual basis, weights are chosen selectively for subgroups of bases. The application is highly relevant to classification tasks. In this scenario, a set of bases for the whistles of each species are learned, but there is potential redundancy between these sets of bases. Left unconstrained, the system could use weights from any of the species across the different sets of bases. However, because the weights of these bases are the features used in species prediction, if the weights from similar bases are distributed, the detection signal becomes weaker. By adding a co-constraint penalty, the reconstruction weights are encouraged or required to be selected from specified sub groups. To implement this in an NMF framework, I followed the derivation detailed in Tjoa et al. [87]. They introduce a  $Q$  matrix to define the co-occurrence relationships among bases. For  $R$  bases,  $Q$  is a square matrix,  $R \times R$  and I minimize the divergence:

$$g(H) = d_{EUC}(Q, HH^T) = \|Q - HH^T\|_F^2 \quad (2.6)$$

To prevent division by zero, and to guarantee a decrease in the divergence  $\epsilon$  is added to the numerator and denominator and is set to 0.2, a small positive number. This divergence leads to the following co-occurrence constrained update rules for  $H$ :

$$H \leftarrow H \frac{W^T V + \lambda Q H + \epsilon}{W^T W H + \lambda H H^T H + \epsilon}$$

---

<sup>1</sup>The L1 norm is the sum of absolute values  $\|x\|_1 = \sum_{i=1}^n |x_i|$

Because bases are learned per species, then concatenated, no co-occurrence is used in training the bases. Additional discussion of my specific implementation is in Section 4.5.4.

### 2.3.3 Online computation

Another challenge of NMF is the computational and memory costs involved for estimating reconstruction of large matrices. For small data sets, batch algorithms are straight forward, but when processing hours of data, online or on-the-fly computation techniques need to be considered. A few key papers [96, 94, 95] address this challenge for NMF, by breaking the bases learning task into a piecewise process. Consider the original matrix  $V$  to be the collection of  $u$  pieces, which could be considered a vector of matrices. Each piece is a temporal subsection of the larger data, with no overlap between pieces:

$$V = \{V(1), V(2), \dots, V(u)\}; u \text{ is the piece index.}$$

With this framework, the reconstruction of convolutive NMF is defined to be:

$$R(u) = \sum_{p=0}^{T-1} W(p) H^p(u) \quad (2.7)$$

Once bases are learned it is a straightforward process to learn weights for any set of pieces, as each piece is independent of prior pieces. However, in basis training, the patterns in all pieces need to be considered in order for the bases to be representative of all the data. The paper by Wang et al. [94] illustrates the process of using cumulative statistics to compute convolutive NMF bases in a piecewise fashion. The basis update rule for this method is:

$$W_p \leftarrow W_p \frac{-\sum_u V(u) (H^p(u))^T}{\sum_q W_q H^q(u) (H^p(u))^T}$$

This implementation of NMF explores all of these variations.

## 2.4 Learning from Data with Incomplete Labels

When annotating marine mammal recordings, different levels of granularity may be needed or provided based on the task. Three types of annotations are considered: 1) species labels, 2) times when vocalizations occur, and 3) the specific frequency contour of vocalizations.

Annotations can often make a task easier by highlighting the exact signal to be modeled. However, for many tasks the process of creating comprehensive and accurate annotations is time consuming, costly and qualitative. When training data are restricted to data with fine-grained hand annotation, only a limited set of recording conditions can be represented.

The 5th DCL dataset (used herein and described in Section 3.1) does provide specific frequency contours for each vocalization for a subset of the data. The time when a vocalization occurs can be inferred from the frequency tracks and it is used for this purpose. I did not use the additional frequency information in order to validate the possibility of species classification without the intermediate step of frequency tracking. To distinguish between annotation types, I will use the term 'annotation' to denote when a vocalization occurs, and 'label' to denote the species attributed to the vocalization.

The dataset includes species labels for both clicks and the unannotated portion of whistle data. In this thesis, 'partial label' will refer to species labels applied to a block of data, but without further annotation of the recording. Because time annotations are not available for vocalizations in this data, I detected them using an automated procedure described in Sections 3.1.2 and 3.1.4. I could infer that an automatically detected vocalization was from the particular species visually observed during recording, but would also expect error in the label because some detections could potentially have been in error, e.g., noise. Partially labeled data is characteristic of weakly supervised learning scenarios.<sup>2</sup> Partial labels allow us to balance the trade-off of annotation costs and the need to have a large and sufficiently diverse training set to compensate for recording session effects.

In the rest of the thesis, I will contrast five learning types: supervised, unsupervised, semi-supervised, weakly supervised, and active learning.

---

<sup>2</sup>In this thesis, the term "partially labeled" is used to mean that the labels are incomplete, not to be confused with the case where a subset of the data is labeled.

- A supervised system has annotations and class labels for all training data.
- In contrast, unsupervised learning has neither annotations nor class labels for any of the data.
- In semi-supervised learning, some portion of the data has annotations and class labels, but there is additionally a large, unlabeled set of data with neither annotations or class labels that can be used to enhance the learning system.
- In weakly supervised learning, some or all of the data has partial or noisy labels. Weakly supervised learning is typically built on semi-supervised methods. This is the scenario for the data in this thesis.
- Active learning [18, 19] starts with a small set of labeled data and iteratively determines which unlabeled samples would most benefit the system with human annotations. Active learning can be leveraged in combination with other learning types as discussed in Chapter 7.

There are a variety of semi-supervised approaches to combining labeled and unlabeled data. A classic summary of these methods is provided by Zhu [103]. An important subclass of methods involves iteratively labeling the unlabeled data and updating the model by techniques including self-training [76] and co-training [11]. Self-training is a simple approach that uses all of the features from the labeled portion of data to construct one model. Unlabeled data is then processed with that model to score class posteriors for each example, and the highest ranked examples are added to the training set. Co-training has three basic assumptions. First, the feature space can be divided into two portions. Second, each partition of the feature set can construct a classifier, and third, the two classifiers are conditionally independent. The two learned classifiers are then used to “teach” each other new training examples from the unlabeled data. The work of Nigam and Ghani [64] demonstrated this third condition, conditional independence is fundamental to the success of this method.

Weakly supervised learning is a variation on semi-supervised learning and is defined by the use of partially labeled data. There are several ways in which data can be partially

labeled, including: having presence vs. absence labels when there are multiple possible classes that could be present, having multiple labels (only one of which is true), or having a class label without time information. In the DCL dataset, it is known that the unannotated data were collected in the visual presence of only one species. There are no annotations of when vocalizations occur, but species labels can be used to filter out incorrect hypotheses from self training when these hypotheses are classified as an incorrect species.

The weakly supervised approach has been used successfully in other applications particularly object recognition in images [10, 27, 22]. Additionally, it has been applied to object recognition in video, introducing a time-series component [61]. Similar to the marine mammal task, visual object recognition currently depends on having annotations of both object location and class. Because of the variations in scale and pose of objects and the near infinite set of possible object classes it is difficult to obtain sufficient size and quality of hand annotated data. Fergus et al. [27] demonstrate the ability to use Google’s image search to generate weakly supervised labels for an arbitrary class and build accurate object classification models from the weakly supervised data. The data returned from an image search is considered weakly supervised because it does not contain information regarding the location or number of objects in the image and search results often include images of unrelated classes. They compare the results of three classification models (pLSA, ABS-pLAS, and TSI-pLAS) as trained on traditional hand annotated data as compared to the weakly supervised data and show competitive performance. They extend this result by using the learned topic models to re-rank Google’s image search. Deselaers et al. [22] take a slightly different approach to training with weakly supervised data. In their work, they propose a two stage system. First they use data with only location annotations, to create a generic detection algorithm. In the second stage, they use a conditional random field (CRF) to simultaneously localize and learn a class model, iterating between the two tasks. The generic detection model created in the first stage is used to initialize a localization for the CRF training. Though our techniques are different, this research provides context for other domains where weak supervision has been used to overcome challenges of simultaneous object detection and classification.

## **2.5 Summary**

This chapter outlines the prior work for each area of this thesis. Here I summarize the specific pieces used to conduct the thesis experiments. I used NMF as a means of learning features for non-stationary spectral features. This was the first application of this method to marine mammal vocalizations for detection and classification. It provided several advantages over previous methods of marine mammal species classification, including the ability to incorporate both stereotyped and non-stereotyped vocalizations of all types (clicks, whistles, and burst-pulse sounds) in species models and an opportunity to use annotations that only mark start and end times of vocalizations, rather than precise frequency contours. With regard to methods for clicks, I use detection and classification techniques as used by Roch et al. [72], but build upon that work by introducing the iterative training process which automatically selects clicks that are most predictive of species.

## Chapter 3

## CLASSIFICATION SYSTEM DESIGN

**3.1 Data Description**

The data used to evaluate our whistle detection and species classification model is from the 5th International Workshop on Marine Mammal Detection, Classification Localization and Density Estimation. It is publicly available at Mobysound.org. In this analysis I constrained experiments to the data from the three primary species, bottlenose dolphins (*Tursiops truncatus*), common dolphins (*Delphinus delphis* and *Delphinus capensis*), and melon-headed whale (*Peponocephala electra*). A detailed description of the recording equipment and environment in which the data was recorded are provided in Baumann-Pickering et al. [7] and Soldevilla et al. [82]. The recordings as provided were sampled at 192kHz (96kHz bandwidth). In order to evaluate potential performance with data collected using sonobuoys, I downsampled the data to 40kHz and low-pass filtered at 20kHz bandwidth.

For evaluation, the provided files were divided into sub-files of 30 second length. This subdivision is performed to make the number of files correct is a more meaningful evaluation metric. When files are of non-uniform length, the longer files are more likely yield correct predictions because there are more vocalizations on which to base the prediction. Additionally, because short and long files, taken as a whole, would count equally in the analysis, there would be artificially greater weight on short files. Ideally, there should be enough data that segments discarded after sub-file creation would be unimportant. However, some of the segments of recorded data were short, and I was concerned about further limiting an already small data set. I chose to keep the remainder segments in part because the time difference between, for example, 15 and 30 sec is significantly less than the original 30s and 10 min differences. Additionally, because many of the files were quite long ( 10 minutes), there were only a few remainders. There was no minimum length constraint. For the annotated portion of the data there were 17 fragments that were less than 29 seconds

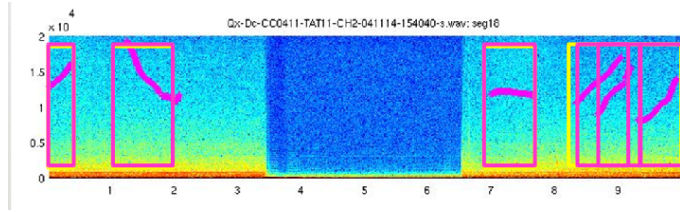


Figure 3.1: Example of whistle unions, single whistles are denoted in pink, with pink rectangles to outline start and end times. Yellow boxes denote the start and end times of whistle unions. For single whistles the union start-end times are unchanged.

out of 219 total segments.

In many applications, the terms labels and annotations are used interchangeably. In this thesis, I attribute different meaning to these words in order to describe known information associated with two desired tasks. Labels referred to the species identification, and a single species label was associated with each 30s segment. For the portion of data where hand annotations of whistles were available, if no annotations are present in the 30s segment, it was assigned as noise. Annotation referred to the time (and frequency) of a whistle occurrence. Annotations were provided in a format that labeled the time-frequency track of each hand identified whistle for a subset of files. In this work I discarded the frequency information and used only the start and end times. For all whistles that overlapped in time, the collective start and end times were used to denote when whistles were occurring, and is referred to as the whistle union. For an example, see Figure 3.1. No annotations were provided for clicks, though each file was recorded in the visual presence of only one species such that it could be assumed that any detected click or whistle could be attributed to either that species, or noise. The frequency contours were provided by hand or Silbido annotations and the whistle unions are derived from this information. The sudden shift in background noise was noted as a recording artifact.

For the three species, different volumes of training data were available, both in units of recording time and in counts of clicks or whistles. Tables 3.1 and 3.2, show some statistics regarding the provided files. I performed a downselection of the unannotated files such that only 30 second segments with at least 8 Silbido-detected whistle contours were included

Table 3.1: Number of whistles in annotated data

	# files	# 30s files	# individual whistles
bottlenose	11	101	3208
common	17	165	8931
melon	8	64	5423

Table 3.2: Number of files in unannotated data, restricted to exclude files with less than 8 detected whistles.

	# files	# 30s files
bottlenose	136	1200
common	31	359
melon	133	1620

in the cross validation partitions. This was done to have a high confidence that a genuine whistle was contained in the file, which was particularly important for maintaining accuracy when evaluating performance on unannotated data. The published performance of the Silbido graph search whistle detection algorithm was 80.0% recall and 76.9% precision [71]. When automated Silbido annotations are compared to human annotators, coverage (percent of tonal detected) and fragmentation rate (number of fragments detected per tonal) are used to evaluate quality. Reported results were a coverage of 80 to 85% and a fragmentation rate of 1.2 [71].

Common terms are defined below for reference:

1. File - a .wav format audio file of varying length, between 30 seconds to 10 minutes
2. Segment - same audio data as in the files, but broken into 30 second segments for faster computation and uniform analysis.

3. Contour - the specific time-frequency points that define an individual whistle contour.
4. Union - the annotation of start and end times of the union of any temporally overlapping whistle contours; also called whistle union.
5. Fold - groups of file segments (from the same parent file) used to conduct multiple independent evaluations of an experiment configuration; also called cross validation fold.
6. Annotation - In much of the literature it refers to both time and frequency of the whistle contour. In this work it refers only to the start and end times of individual whistles or whistle unions.
7. Label - classification category, e.g. bottlenose, common, melon or noise.

### 3.1.1 System Overview

The overall species classification system is composed of modules to perform subtasks, each of which is shown in Figure 3.2 and described in this chapter.

As an overview, the subtasks were as follows. The first step was detection and isolation of click-like sounds. Clicks were processed separately from the remaining audio by extracting Mel-Frequency Cepstral Coefficient (MFCC) features and making a species prediction based on species-specific GMM models. Once click events were removed, the remaining data were

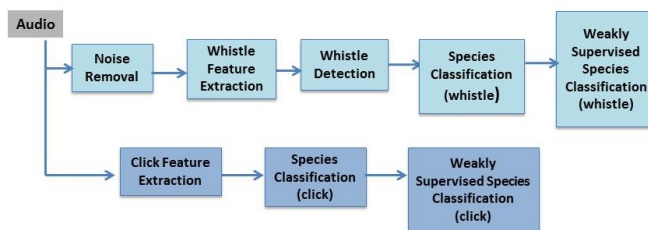


Figure 3.2: Block diagram of system modules

passed through an additional noise removal process before identifying segments that were likely to contain whistles. NMF features were extracted from whistle events and a species prediction generated.

### 3.1.2 Click Detection and MFCC Feature Extraction

Click-like slices (all referred to here as clicks) are automatically detected and extracted via an iterative thresholding process that identifies regions of atypically high broadband energy. These extracted clicks include echolocation clicks and other short duration broad band sounds. For an example, see Figure 3.3a and 3.3b, which show the original spectrogram and extracted clicks respectively. Click detection and removal was the first step in the audio processing procedure because it helps the median smoothing achieve a better estimate of any slowly varying background for noise estimation. Also, others [72] have experimentally observed that noise removal can degrade the species classification performance from clicks.

A time slice<sup>1</sup> of the spectrogram is identified as a click when the mean energy in the upper half of the frequency range is greater than a unitless constant ( $T$ ) times the mean of the energy in the  $\pm 2$  neighboring time slices. When a slice detected as a click is removed, it is replaced with a vector that is the average of the  $\pm 2$  time frames around but not including the click time. We iterate over the data four times, increasing  $T$ , such that we first extract the strongest clicks, and as they are removed we can then extract any additional neighboring clicks that may have been obscured. In my experiments,  $T$  was set to perform two passes at 1.8 and an additional two passes at 2.0. The spectrogram settings were 40kHz sampling, 2048 point FFT, Hamming window of length 1024 and 512 point overlap. This implies the shortest inter-pulse interval that could be fully resolved as a click would conservatively be 0.0396 seconds. Because of the iterative extraction based on the energy in each click, clicks in neighboring time slices can still be extracted, depending on the specific circumstances. Click detection performance could not be evaluated because annotated clicks were not available in the 5th DCL data. Thus, I had to assume that the automatically detected clicks included both clicks that were noise and clicks from the

---

<sup>1</sup>a time slice is single column of frequency information from the spectrogram

target animals. Noise clicks were defined to encompass clicks from non-marine mammal sources and clicks that were a poor representation of the species characteristics, e.g., due to reverberation. The noise clicks were identified in a weakly supervised learning process described in Chapter 6.

Prior work has shown MFCCs [16], a common and highly successful feature type for human speech processing [15], are also reliable features for species classification of odontocete clicks [57] [72]. Here, 18 MFCCs are computed by finding the energy in log-spaced frequency subbands using a triangular window on a 2048-point FFT and then taking the log magnitude. MFCC coefficients are calculated for each click and are the features used in scoring and click-based species classification. As a note, the same MFCC procedure is used to establish a baseline for species classification for whistles, and is compared against the proposed NMF features. For whistle features, 18 MFCC coefficients are computed in addition to the first and second deltas<sup>2</sup>.

### 3.1.3 Noise Estimation and Removal

After click-like sounds were removed from the data, a noise removal stage was performed on all data, including segments with whistles. This served to identify and enhance the remaining data for whistle detection. I used a 2D median smoothing algorithm to estimate the background noise [63], then performed spectral subtraction to remove this contribution from the data. Median smoothing is an effective procedure for removing slowly varying noise [63]. The process defines a rectangular spatial mask of width  $X$  and height  $Y$ , which we specified as 15 and 3 respectively. The mask is shifted across the image and the centroid of the mask is replaced by the median of the values contained in the mask. This process decreases the sensitivity to outlier values. The selected shape of the mask,  $1 \times X$  or  $Y \times 1$ , can help preserve vertical or horizontal structure, respectively. This shape was tuned to preserve whistle structure, but no knowledge of actual whistle locations is used. This process captures the slowly varying background noise, which can then be subtracted by spectral

---

<sup>2</sup>The MFCC deltas are a features derived from the MFCC coefficients by taking the difference between consecutive coefficients. The second deltas are computed by taking the difference between consecutive coefficients of the first deltas.

mean subtraction [12]. An additive noise model is assumed,  $y = x + n$  where  $y$  is the time domain noise corrupted signal, and  $x$  and  $n$  are the signal and noise respectively.

Generalized spectral mean subtraction is computed by  $x = IFFT(|Y|^g - |N|^g)$ . Standard magnitude subtraction is performed if  $g = 1$  and power spectral subtraction when  $g=2$ , though other values can also be used. Figure 3.3 is an example of the noise removal process for a section of the 5th DCL data.

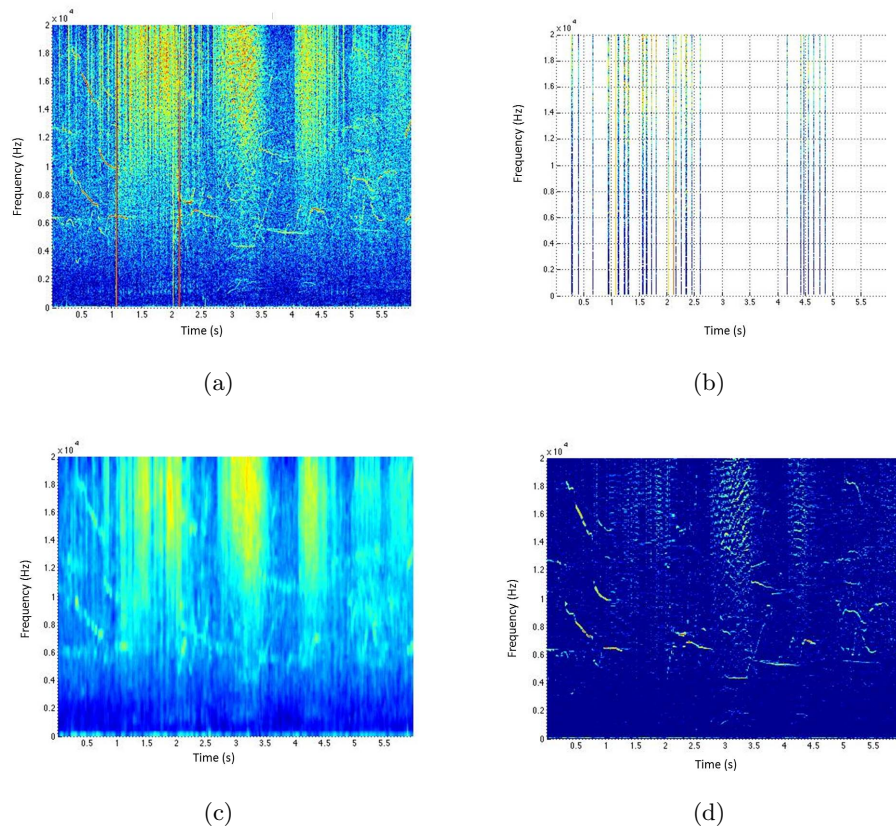


Figure 3.3: Noise Removal Process: (a) Original spectrogram (b) Detected clicks (c) Median smoothing estimate of background noise, and (d) Result of noise removal process.

### 3.1.4 Whistle Feature Extraction and Detection

Hand annotations are used when available. For data without hand annotations, the Silbido whistle detection tool is used to identify regions where whistles likely occur. The annotation

quality between hand and automated methods are not equivalent, and this influence is incorporated into the experiment design (see Section 3.1). Once whistle regions were isolated, features were extracted via NMF from the identified time segments. I discuss the ability of NMF to perform whistle detection and present those preliminary results in Chapter 4 .

The whistle features were extracted from detected segments after spectral subtraction of noise. The fixed NMF bases were learned in the system training phase. Algorithms to perform this training are described in Section 2.3. An NMF decomposition was performed to determine the weights corresponding to the fixed NMF bases. These weights were used as features in whistle detection and species classification. To quantify the impact of using automated detections, I also contrasted species classification performance using hand or Silbido automated detections, for the portion of data where hand annotations were available.

### *3.1.5 Classification Model*

The Gaussian mixture model (GMM) is a common statistical model for classification. A GMM is fully defined by the vector mean and covariance matrix parameters. These parameters are first assigned via random initialization then estimated from training data by applying the expectation-maximization (EM) algorithm. Details of the EM algorithm are described in a classic paper on the method [21]. I used GMMs to learn models for click based species classification, whistle detection and whistle based species classification. Note that the NMF bases used to train the GMM must match the bases used in testing.

## **3.2 Experiment Paradigm**

### *3.2.1 Cross Validation Data Partitioning*

It was observed in [62] that session effects strongly influenced the performance of the classifier in tasks involving field recordings. Session effects occur when data from different species are collected under significantly different recording conditions, so that the classifier can inadvertently use acoustic conditions of the recording in addition to vocalizations in the classification of species. Session effects can artificially inflate the estimate of classifier performance, a source of bias. A detailed discussion quantifying the impact of session effects

is given by Roch et al. [75].

In the 5th DCL data, there were several influences to consider. First, the data were collected from two disparate geographic sites, Palmyra Atoll and the Flip Observation Platform [an artificial platform] stationed in the Gulf of California. Hydrophones were either towed or dipped at a depth between 10-80m. These two locations had different acoustic environments and propagation characteristics. Thus, to train on data from one and evaluate on data from the other would produce poor results unless session effects were removed as a source of bias. An alternate approach, training and testing on independent data from the same geographic location, could produce falsely high results. Likewise, independent segments of data recorded on the same day, in the same location could produce overly optimistic performance estimates.

To negate these influences, I distributed the data into cross validation folds.<sup>3</sup> Based on the amount of data available, I chose to construct three folds for the annotated and unannotated data. For each species-annotated data I first sorted the data by location. Using the provided filenames, which included a date and time stamp syntax, I also applied a time order sort. Files were then assigned by counting off by three through the time sorted lists. This provided an even geographic and time distribution of the data into each of the folds. Additionally, when assignments were made, any data from the same day were assigned to the same cross-validation fold. Once the folds were constructed, the complimentary folds for the two geographic regions were combined to define the final fold, for example Flip A and Palmyra A. This procedure was repeated for the unannotated data files, but with one other selection criterion.

The annotated partitions were identified by the letters A, B, and C, and the unannotated partitions by 1, 2 and 3. I systematically created permutations of these partitions to construct training, tuning, and evaluation scenarios for all stages of the classification system. The result is diagrammed in Figure 3.4. Specific file-segment names included in each of the A, B, C, 1, 2 and 3 partitions are listed in Appendix A.

---

<sup>3</sup>Cross validation is a method of model evaluation that helps assess performance on new and independent data. An experiment configuration is repeated over independent partitions (sometimes called folds) of the data. The reported score is the average performance across the each of the individual partitions.

Train NMF	Train GMM	Eval $\alpha$	Eval $\beta$

Figure 3.4: Cross validation partitions. Purple is a partition with hand annotated whistles, green is a partition without hand annotated whistles. Each row corresponds to a cross validation fold. The A,B,C and 1,2,3 partitions are used in only one stage of the experiment to keep training and evaluation data independent, but the permutations are different among cross validation folds to provide variability.

Table 3.3: Whistle union statistics for hand annotated partitions A, B, C. (Standard deviation noted in parentheses; date:YY-MM-DD)

	Mean # unions	Mean union duration	Mean cumulative duration	Min date	Max date
partition A	14.2 (11.8)	0.39 (0.50)	5.65 (5.92)	04-11-14	07-10-04
partition B	16.7 (16.2)	0.42 (0.52)	7.10 (6.96)	06-04-06	07-10-12
partition C	18.9 (20.0)	0.41 (0.76)	7.80 (8.69)	06-05-16	07-09-25

Table 3.4: Whistle union statistics for Silbido annotated partitions 1, 2, 3 (Standard deviation noted in parentheses; date:YY-MM-DD)

	Mean # unions	Mean union duration	mean cumulative duration	min date	max date
partition 1	24.13 (9.28)	0.56 (0.61)	13.58 (7.52)	04-11-14	07-10-01
partition 2	24.18 (7.54)	0.56 (0.62)	13.68 (6.51)	06-04-06	07-10-03
partition 3	27.49 (8.25)	0.61 (0.64)	16.71 (6.75)	06-05-13	07-10-04

For the A, B, C and 1, 2, 3 partitions, I computed some basic statistics for the number and duration of whistle unions. Comparing Table 3.4 and Table 3.3, the difference in mean unions per file is significantly larger among the A, B, C partitions which range from 14-18, as compared to the 1, 2, 3 partitions, which range from 24-27. The mean union duration was approximately the same regardless of partition. The largest difference between partitions was the cumulative duration, which was 5-7 seconds in the A, B, C partitions and 13-16 seconds in the 1, 2, 3 partitions, or almost double that in the annotated data.

Taking into account that the 1, 2, 3 partitions were restricted to segments with whistles - a minimum 8 automatically detected whistles were required in the segments annotated using Silbido - I recomputed the mean unions per file to exclude annotated segments without

Table 3.5: Whistle union statistics for hand annotated partitions A, B, C. **Empty files excluded** (Standard deviation noted in parentheses)

	Mean # unions	mean cumulative duration
partition A	19.5 (9.3)	7.77 (5.62)
partition B	19.8 (15.7)	8.44 (6.79)
partition C	30.5 (17.1)	12.54 (7.85)

Table 3.6: Whistle union statistics for partitions A, B, C using Silbido annotations. **Empty files excluded** (Standard deviation noted in parentheses)

	Mean # unions	mean cumulative duration
partition A	10.9 (10.5)	4.82 (5.06)
partition B	14.7 (10.7)	6.42 (5.86)
partition C	11.5 (12.2)	5.68 (7.20)

whistles in the A, B, C partitions. The mean cumulative duration per file for partitions A, B, C is shown in Table 3.5 The difference in mean unions per file was slightly smaller but the mean cumulative duration continued to be significantly different.

Tables 3.7 and 3.8 describe the geographic distribution of data among species, per partition for Silbido and hand annotated data. For both Silbido and hand annotations, the data for common dolphins is 100% recorded at Flip, and Melon is 100% recorded at Palmyra. The only variability between partitions is due to data in the bottlenose dolphin portion. For hand annotated data, partition B has the most data from Flip. For Silbido annotated data, the majority of the bottlenose data is from Palmyra, for all partitions.

Experiments on weakly supervised click based species classification were conducted before the full system design had been implemented. Cross-validation was used in these experiments; however, the partitions were slightly different from Figure 3.4. This configuration used a five-fold configuration and had more unannotated data. The total set of unannotated

Table 3.7: Geographic distribution of whistle unions per species for hand annotated partitions A, B, C. (%Palmyra:%Flip)

	Bottlenose	Common	Melon
partition A	50:50	0:100	100:0
partition B	11:89	0:100	100:0
partition C	93:7	0:100	100:0

Table 3.8: Geographic distribution of whistle unions per species for Silbido annotated partitions 1, 2, 3. (%Palmyra:%Flip)

	Bottlenose	Common	Melon
partition 1	99:1	0:100	100:0
partition 2	82:18	0:100	100:0
partition 3	90:10	0:100	100:0

data used in the click partitioning is a strict superset of the total unannotated data used in the whistle partitions (A, B, C). The procedure for performing the split was identical. A diagram of the five-fold partition is shown in Figure 3.5, and a file list is included in Appendix A.

### 3.3 Performance Evaluation

To quantify the system performance, there were two primary tasks to be evaluated: whistle detection and species classification. The latter can be achieved by either classification from clicks or classification from whistles. These different tasks require different evaluation methods in order to effectively analyze performance.

Because of the difficult, expensive, and opportunistic nature of obtaining marine mammal acoustic data, the resulting datasets are often small and have uneven volumes of data for each class. Whistle detection can be assessed at a frame-level or a whistle-level. In a

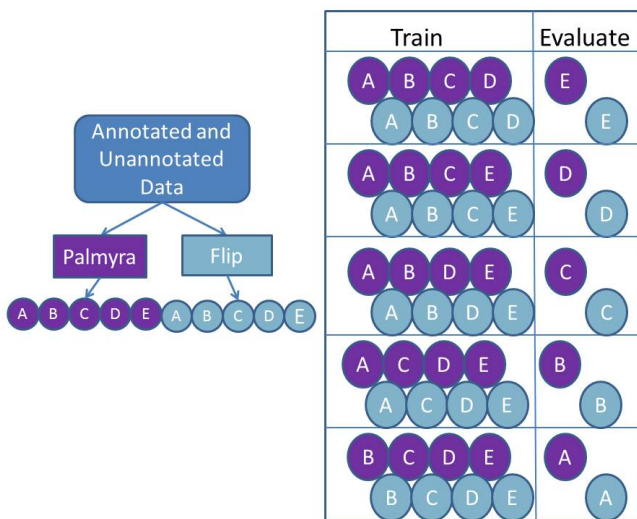


Figure 3.5: Five-fold cross validation partitions used for weakly supervised species classification from click experiments. Common subsets between click and whistle partitions are not readily summarized, however exact file lists for all partitions are provided in Appendix A

frame-level detection, a prediction (whistle or no whistle) is made per time-slice of the spectrogram. Performance is evaluated with a Receiver Operator Curve (ROC), an efficient way to assess the trade off between correct detections and false alarms. Computing the area under the curve (AUC) provides a statistic that can summarize the ROC performance. Early experiments computed these metrics but they are not reported here because a drawback to frame-level analysis is that it does not penalize the fragmented detection of whistles. The second approach, for which I present results, is a whistle-level statistic. In this approach, consecutive frames need to correctly predict the whistle class, for a minimum percent of overlap with the known labeled whistle to be considered correct. The whistle-level task is more challenging, and performance varies depending on the required percent overlap. In the presented results, I plot the F-score (the geometric mean of precision and recall) with respect to percent overlap, ranging from 10% to 80%.

In a data-driven modeling approach, it is essential to have as much training data as possible in order to capture variability. It is not practical to discard samples to achieve ideal conditions such as an equal number of samples per species because too much would

be thrown away. However, imbalance in the data by species has direct impact on choosing evaluation metrics because typical metrics such as cross species error, F-statistics and overall percent error are calculated assuming that samples are independent and weighted equally. They will be dominated by performance of the majority class <sup>4</sup>. I report these standard statistics, but also provide macro statistics, which average performance across species. Macro statistics have been shown to be more robust in situations of unbalanced data [81]. I report macro-error in the experimental results. MacroEN is a variation of the macro error, where noise is also allowed as a class.

For a standard confusion matrix  $M$ , the cross-species error (CSE) is defined in equation 3.1, where  $M_{ij}$  is the number of cases where the true species  $i$  is recognized as species  $j$ .

$$CSE = \frac{\sum_i \sum_{j \neq i} M_{ij}}{\sum_i \sum_j M_{ij}} \quad (3.1)$$

The false detection rate for species  $k$ ,  $P_f(k)$ , is the relative number of 30s file segments predicted as class  $k$  that were in fact from a different class. Likewise, the missed detection rate,  $P_m(k)$ , is the relative number of true file segments of class  $k$  that were mistakenly predicted as a different class. The macro-false and macro-miss combine the false and miss scores from each class using macro statistics. As discussed in [81] the macro scores represent each class evenly, unlike the micro score computation that favors larger classes.

Let  $N$  equal the number of species modeled. To compute the macro statistics, we define  $M_k = \sum_{j=1}^N M_{kj}$ , the number of true species  $k$ , where  $M_{kj}$  is the number of cases where species  $k$  is classified as species  $j$ . This is used to define the variables  $P_f(k)$  and  $P_m(k)$  as the per

---

<sup>4</sup>As an intuitive example, if there are 10,000 examples of class A but only 10 examples of class B. If the class A model is good, overall results will be good, regardless of how good or bad the class B model performs. Macro statistics average the per class performance to offset this bias.

species false and miss rates in equations 3.2 and 3.3 respectively.

$$P_f(k) = \frac{\sum_{l \neq k}^N M_{lk}}{\sum_{l \neq k}^N \sum_{j=1}^N M_{lj}} \quad (3.2)$$

$$P_m(k) = \frac{\sum_{i \neq k} M_{ki}}{\sum_{j=1}^N M_{kj}} \quad (3.3)$$

$$(3.4)$$

The macro-false and macro-miss rates are defined in equations 3.5 and 3.6, where  $N$  is the number of species classified ( $N=3$  for experiments here).

$$\overline{P}_f = \frac{1}{N} \sum_{k=1}^N P_f(k) \quad (3.5)$$

$$\overline{P}_m = \frac{1}{N} \sum_{k=1}^N P_m(k) \quad (3.6)$$

The macro-error is the geometric mean of the macro-false and macro-miss:

$$MacroE = 2 * \frac{\overline{P}_f * \overline{P}_m}{\overline{P}_f + \overline{P}_m} \quad (3.7)$$

In the species classification experiments using whistle detection, I made predictions only from the time regions identified to have whistles (by hand or Silbido annotations). I assumed this was an accurate detection and used a decision function that did not allow a noise class. In addition, for a subset of the conditions, I ran experiments where the model was allowed to assign segments the class label of noise, which is potentially useful with automatic annotations. In this case, I refer to the error statistics as macroEN. The macroEN is the same equation as macro-error, except it is computed from  $P_{fN}(k)$  and  $P_{mN}(k)$ , and the noise class is allowed as a decision.

To compute the macroEN statistics, define the variables  $P_f(k)$  and  $P_m(k)$  as the per species false and miss rates and the  $N + 1$  class is the noise class. Formulas are defined

equations 3.8 and 3.9 respectively.

$$P_{fN}(k) = \frac{\sum_{l \neq k}^{N+1} M_{lk}}{\sum_{l \neq k, l=1}^{N+1} \sum_{j=1}^{N+1} M_{lj}} \quad (3.8)$$

$$P_{mN}(k) = \frac{\sum_{i \neq k} M_{ki}}{\sum_{j=1}^{N+1} M_{kj}} \quad (3.9)$$

$$(3.10)$$

### 3.4 Summary

- The 5th DCL dataset is cut into file segments of uniform 30 second length and annotations are converted whistle union format.
- Clicks are iteratively extracted from the audio and MFCC features are computed for GMM species classification.
- Median smoothing based noise removal is applied before NMF features are extracted from background noise and whistle union segments.
- Cross validation is used to balance data based on available temporal and geographic information.
- Detection is evaluated with F-score and species classification is evaluated with cross species error, MacroE and MacroEN statistics.

## Chapter 4

**VARIATIONS OF NON-NEGATIVE MATRIX FACTORIZATION FOR MARINE MAMMAL WHISTLES**

For the task of marine mammal detection and classification, the intuitive interpretation of NMF allows us to train NMF bases representative of acoustic components: noise, whistles, and species-specific whistles. Because of the important temporal information contained in the whistle structures, convolutive NMF bases are a logical choice over standard NMF. Additionally, due to the size of the 5th DCL dataset it was necessary to use an online formulation for learning the bases, because it was computationally more efficient. This dataset is relatively large for public marine mammal datasets, but it is unclear it would be large enough to fully characterize the species vocal behavior due to the natural variability and dependence on behavioral state [39]. The online algorithm allowed the dataset to be processed as 30 second file segments and output bases that would be representative of a large sample of observations. Online algorithms use intermediate statistics such that each training example can be processed in sequence rather than having all the data in memory at the same time. The formulation by Wang et al. [94], summarized in Section 2.3.3, was adopted as the primary method of NMF whistle computation. This 2D approach which is inappropriate for clicks due to their very short duration. Several changes to the Wang et al. methods were tested for the potential to improve detection and classification performance, including multi-pass and shuffle training, noise training strategies, and co-occurrence constraints. These additions are the topic of this chapter. The different variants were assessed in whistle detection experiments. The task of whistle detection was not the primary objective of this research, however it was a necessary step in system development. Detection was needed to identify segments of recordings to be used in species classification. It could also play a role in identifying potential vocalization examples to include in the weakly supervised learning process, as will be discussed in Chapter 6. Whistle detection experiments also provide a

fast turn around platform for exploring new features of the classification system.

#### **4.1 Noise Training**

Because the end goal of the system was species classification, NMF bases were initially learned separately for each species. As a realistic dataset, the training data also contains periods of background noise that must also be modeled with NMF bases. Using the provided whistle contour annotations, whistle union annotations were generated and used learn species-specific NMF bases. All audio between whistle unions, was considered background noise and used to train the noise bases. The number of noise and whistle bases was a parameter that could be changed via an experiment configuration file. Once species-specific training of all bases was complete, all noise and species bases were concatenated with each other to define a species-independent feature space used for learning the species-dependent GMMs. I refer to this approach as standard basis training.

Standard basis training was simple, but it had some flaws. First, having three sets of noise bases, one from each species training data, could contribute to problems with over-training to particular recording conditions (session effects). While it was possible that a given noise phenomenon could be present in some recordings and not others, and perhaps this concatenation maintained some additional diversity that might otherwise be averaged away, noise was ultimately a single class and the noise bases had to be trained as such. The second flaw was that even though the noise removal process implemented here was widely used for acoustic applications, some noise remained in the recordings even after noise removal, and would thus be present in the whistle training segments, adding unintended bias. As an alternative, which I will refer to as joint basis training, I designed the system to first learn the noise bases using the identified noise segments from all species in the dataset. Next, the species-specific bases are learned, but with the noise bases present and fixed. In this approach, the training algorithm could assign weights to noise bases, such that background noise could be represented by those bases, leaving the whistle bases with the flexibility to learn just the whistle-like sounds. Once the different bases were learned in this manner, the resulting noise and whistles bases were concatenated together for the final feature space to be used in GMM training and evaluation.

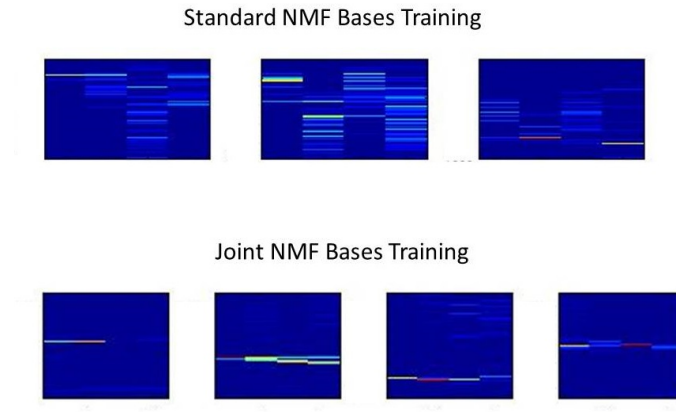


Figure 4.1: Comparison of sample whistle bases using standard and joint training

## 4.2 *Multi-pass Training*

An online learning strategy for training NMF bases was used to sequentially process training files. For the first file, bases were initialized randomly. As each file was processed, bases and cumulative statistics were brought forward to the next training file. Thirty iterations of the NMF update rules were performed for each file. The bases output after processing the last training file were the final fixed bases, which were then used for learning the GMMs for species and evaluating species classification performance.

## 4.3 *Co-occurrence Constraints*

The standard procedure for learning NMF weights considers all bases equal and independent, with no regard to possible groupings. Redundancy between bases have less impact for binary decisions, however, the classification task of this research had three species classes plus a noise class. With a traditional NMF algorithm for learning weights, if there are redundancies among bases from different species, the corresponding basis weights would be spread across species and would result in a weaker species classification performance. To prevent this behavior, I employed a method [87] that encouraged weights to non-zero values for only one species. I used an intermediate  $Q$  matrix to define co-occurrence constraints between NMF bases. This had the effect of forcing the system to choose weights belonging to a defined

subgroup of bases, which should have resulted in better classification performance. The  $Q$  matrix is constructed by defining elements that must co-occur, can co-occur or cannot co-occur. Assuming  $Q$  is indexed by  $i$  rows and  $j$  columns, Tjoa et al. [87] advise values of  $Q_{ij} = 1$  for elements that must co-occur,  $Q_{ij} = 10e - 8$  for elements that cannot co-occur, and  $Q_{ij} = H_i^T H_j$  (or  $HH^T$ ) for elements that can co-occur, where  $H$  is the matrix of NMF weights.

#### 4.4 *Shuffle Training*

An additional consideration is that the model resulting from an online algorithm depends on the order in which the data were processed. With a linear presentation of data, the algorithm could get stuck in a local optimum and produce overfitting in the model. Several articles [47, 99], note the benefit of randomizing the training data to minimize this impact. In this work, the data are shuffled just before NMF bases training. The bases training algorithm is given the list of 30 second file segments to process for the particular cross validation fold. An intermediate module shuffles this list of file segment names such that each file is still processed once, but in a new and randomized order. The remainder of the system remains the same. This occurs for each cross validation fold and preserves the independence of data between cross validation folds.

#### 4.5 *Whistle Detection Experiments*

The system module that performed whistle detection was computed from NMF features. Using the first row of the cross validation data table (Figure 3.4) as an illustrative example, the overall steps to train the whistle detection model were: i) train a set of NMF bases from partition A, ii) Using the bases learned in step i, extract weights from partition B and, iii) train the GMM detection model from these weights. Evaluation would be measured with  $\text{Eval } \alpha$  which correspond to partitions C for this example. This formula is repeated for each cross validation partition. Both species-dependent and species-independent configurations of the detection system were tested.

There were four aspects of the system configuration that were controlled experimentally when comparing results: standard vs. joint basis training, co-occurrence, the number of

noise and whistle bases, and the number of GMM mixtures. In the remaining sections of this chapter, I present the results of experiments to assess the impact of the different NMF basis training approaches in the context of whistle detection. All experiments used the CV training and testing strategy described in Section 3.2.1. Results were reported only on the Eval  $\alpha$  data since hand annotations were required for assessing whistle detection. The median performance of the cross validation set was reported for all metrics and experiments.

#### 4.5.1 *Joint Basis Training*

The first question I addressed was how the joint basis training impacted whistle detection. This could be answered by comparing detection experiments for standard and joint basis training. I first computed two configurations of joint bases (20 noise, 60 whistle and 20 noise,30 whistle). These configurations were chosen because they were similar to the configurations of Smaragdis et al. [80]. I used fewer bases per class, 20 instead of 40, because I had 4 classes instead of 2, and less training data to effectively model a larger number of bases per class. The number of whistle bases was reduced to 10 per class when it was observed there may be redundancies between species. These configurations were compared against the nearest comparable standard bases configurations (21 noise 60 whistle and 21 noise 30 whistle). The standard bases experiment (21 noise 30 whistle) and other unreported experiments with standard bases did not achieve GMM model convergence for all species and cross validation folds.

I was able to achieve a reasonably close match in one experiment pair. The parameters are summarized in Table 4.1. The results showed that the median F-statistic was nearly identical across all values of percent overlap. However, considering both the convergence problems and the identical detection performance, we chose to use joint basis training.

#### 4.5.2 *Multi-pass Training*

In many sequential learning applications, training involves multiple passes over the data. I conducted experiments to determine whether whistle detection improved if the training algorithm was run a second and third time over the same training data. In the second and

Bases	Co-occurrence	N/W Bases	# GMM
Standard	No	21/60	5
Joint	No	20/60	5

Table 4.1: Table of standard vs joint basis training experiment configurations

third iterations, I reduced the number of NMF iterations performed for each file from thirty to five. Based on pilot experiments, I concluded that performing a second pass over the training files yielded notable improvements but the third and further iterations were not necessary.

These pilot experiments were conducted early in the research and before cross validation was integrated into the system. To confirm the initial conclusions of the multi-pass training experiments, I compared the F-scores of one-pass and two-pass whistle detection for an experiment with 5 GMM mixtures, joint basis training (20 noise/30 whistle bases) and no co-occurrence. The results were nearly identical for the one-pass and two-pass configurations (Figure 4.2). Because there was no performance loss, and it had been integrated into numerous other experiments, I continued to use the multi-pass training. Future evaluation with an alternate dataset could provide further insight to potential benefits.

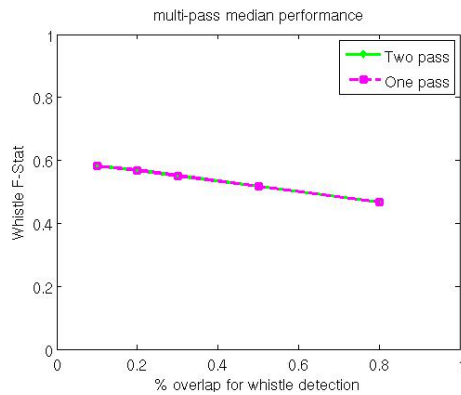


Figure 4.2: F-score comparison for one-pass and two-pass NMF bases training

### 4.5.3 NMF Bases and GMM Mixtures

I explored using different numbers of NMF bases and GMM mixtures to try to improve detection performance. I used 5 mixture components for the majority of classification experiments. Results showed that increasing the model complexity (to 10 mixtures) was only useful when the number of whistle bases increased. However, gains were marginal compared to the performance when using 5 mixtures and fewer bases. The specific experiments compared the use of 20 noise / 30 whistle bases to the use of 20noise / 60 whistle bases for 5, 10 and 15 GMM mixtures. The median performance of each these experiments is shown in Figure 4.3. The parameters are summarized in Table 4.2.

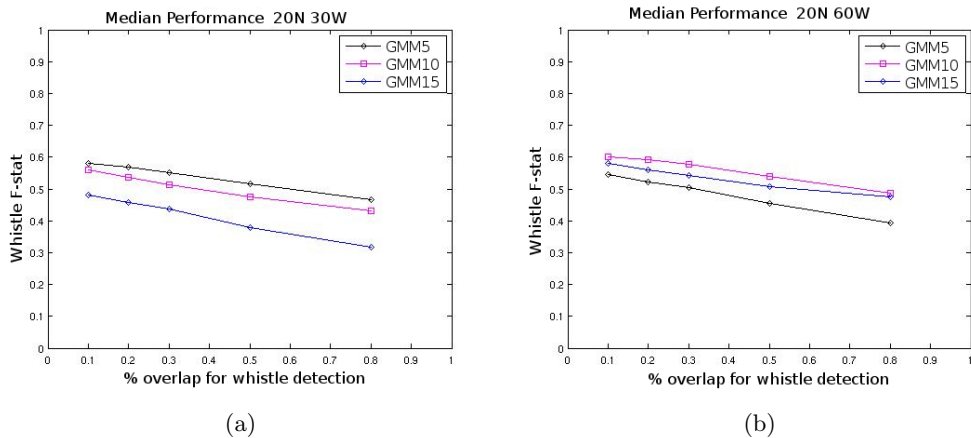


Figure 4.3: Median performance with varying GMM's for (a) 20N/30W (left) and (b) 20N/60W (right)

Based on the two graphs in Figure 4.3, the 20 noise/60 whistle GMM10 configuration had the best performance in this experiment group. However, there is a wide variance in the fold specific performance. The 20 noise/60 whistle GMM5 configuration had significantly lower variance compared to other experiments. Figure 4.4 shows a side by side comparison of the GMM5 and GMM10 experiments with 20 noise and 60 whistle bases.

Joint-Bases	Co-occurrence	N/W Bases	# GMM
Yes	No	20/60	5
Yes	No	20/60	10
Yes	No	20/60	15
Yes	No	20/30	5
Yes	No	20/30	10
Yes	No	20/30	15

Table 4.2: Table of GMM mixture experiment configurations

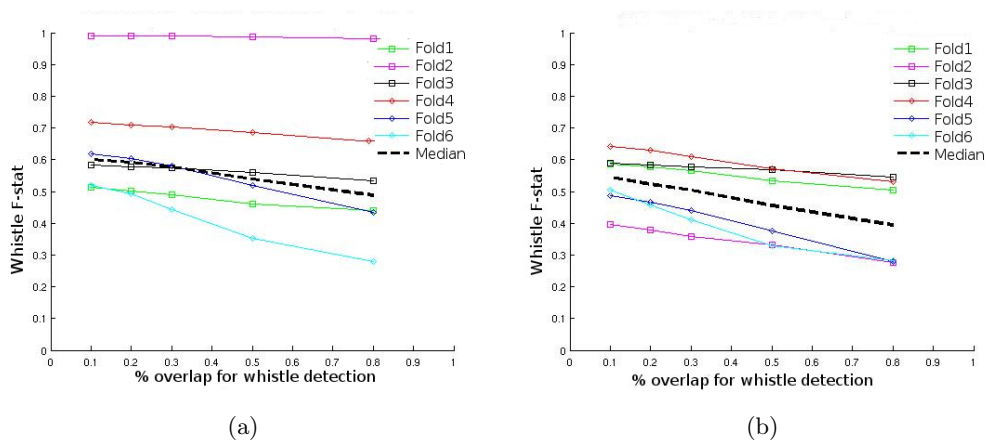


Figure 4.4: Per fold performance for (a) 20 noise/60 whistle GMM10 and (b) 20 noise/60 whistle GMM5 configuration

#### 4.5.4 Co-occurrence Constraints

The next variant explored was the use of different co-occurrence constraints. The two co-occurrence matrices used for evaluation are shown in Figure 4.5. With Q1, I did not allow noise bases to co-occur with whistle bases. This forced the system to choose among the three species or noise, even though noise often co-occurred in the background of a segment with vocalizations. The Q2 matrix allowed noise to co-occur with species, but different species could not co-occur. A summary of the co-occurrence constraint experiment configurations

is shown in Table 4.3.

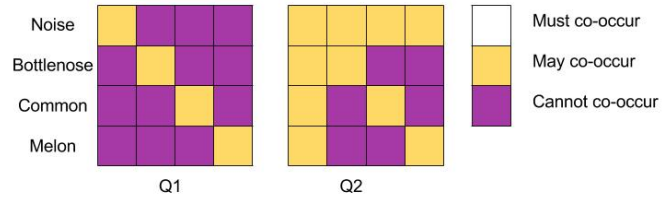


Figure 4.5: Comparison of Q1 and Q2 co-occurrence constraint relationships

Figure 4.6 shows that co-occurrence constraints led to significantly worse detection performance. Given that the task was species-independent whistle detection and not species classification, the additional species specific constraint was an unnecessary penalty in the system and I did not retain it for whistle detection.

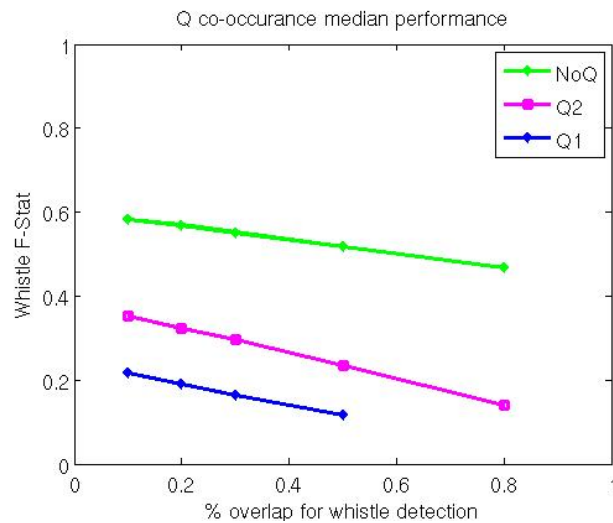


Figure 4.6: Comparison of median performance with NoQ, Q1, and Q2 co-occurrence constraints.

Joint-Bases	Co-occurrence	N/W Bases	# GMM
Yes	No	20/30	5
Yes	Yes-Q2	20/30	5
Yes	Yes-Q1	20/30	5

Table 4.3: Table of co-occurrence constraint detection experiment configurations

#### 4.5.5 *Shuffle Training*

I compared linear and shuffle training strategies for the whistle detection task. Both experiments used 20 noise 30 whistle bases, no Q co-occurrence constraints, and 5GMM mixtures. The results are illustrated in Figure 4.7. There was a slight improvement as compared to the equivalent experiment and ultimately this configuration shows the best performance of all configurations. Per fold variations of these two experiments are shown in Figure 4.8. The per fold performance differences were quite minimal and the relative performance of each fold was the same between both experiments. The largest difference was that fold 3, the worst performing fold, was slightly worse in the non-shuffle experiment. Another subtle difference was that fold 6 did worse at 80% overlap in the shuffle experiment as compared to the non-shuffle experiment. The non-shuffle experiment with 20 noise 60 whistle bases had better median performance than the 20 noise 30 whistle configuration discussed here. A comparable shuffle experiment with 20 noise 60 whistle bases was performed but a convergence problem with the GMM model for one fold prevented full comparison. These results suggest that across all cross validation folds, shuffling provides a consistent but very small benefit averaged over the folds and a small reduction in variation across folds.

#### 4.5.6 *Species Dependent Detection*

It is possible that for the detection task, a species-dependent model would outperform a species-independent model. To test this possibility I used the species-independent set of bases, and, using only species specific training data, learned a species-specific whistle

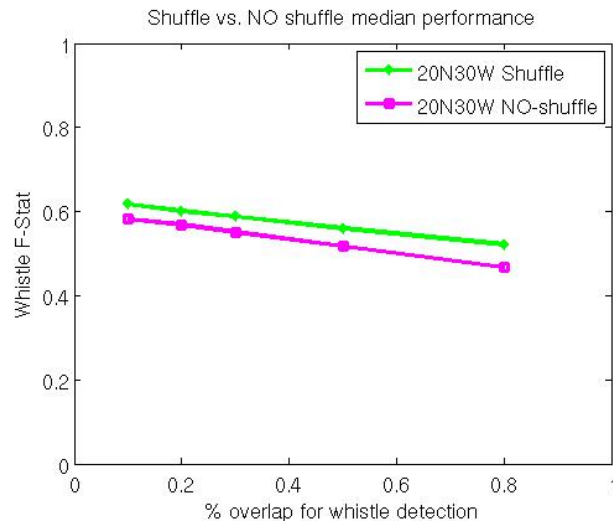


Figure 4.7: Comparison of 20N30W NoQ GMM5 median performance with and without shuffle training

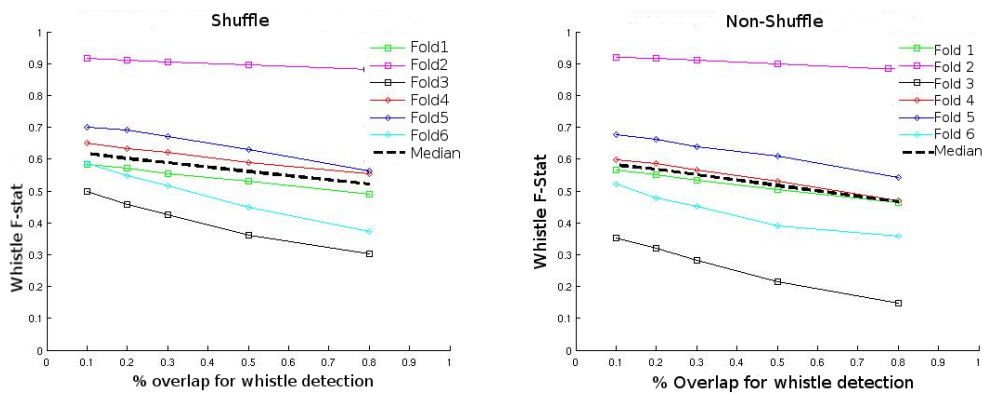


Figure 4.8: Median and per fold performance of Shuffle (left; 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint), and No Shuffle (right; 20N30W joint bases, GMM5, no shuffle and no co-occurrence constraint)

model. All models used joint bases, no Q constraints, 20 noise bases and 30 whistle bases, and GMM5. I evaluated each species model with the portion of the Eval  $\alpha$  dataset associated with that species and compared the result to the evaluation of the multispecies model with pooled data (Figure 4.9). It is clear that this configuration of species-dependent models did not perform as well as the species-independent model. I did note that several folds of

the bottlenose dolphin model did not converge, so the results for bottlenose data do not contain all folds. However, based on the trends from the other species, it is unlikely that the missing data would have changed the conclusions.

The most obvious explanation for the poor performance of species-dependent models was the reduced volume of training data. It is also possible that session effects could have been a factor.

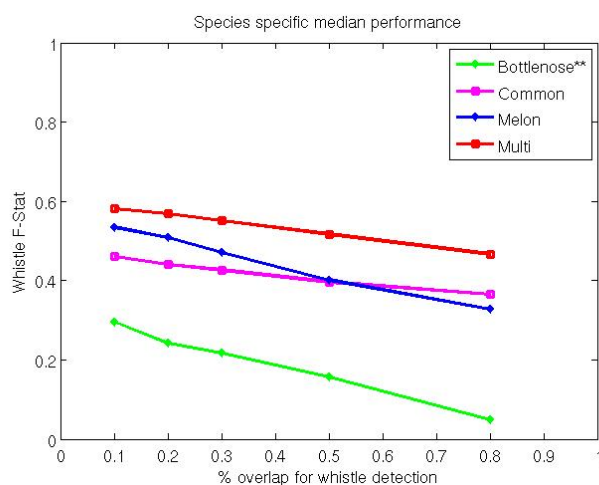


Figure 4.9: Comparison of species-dependent and -independent whistle detection

#### 4.6 Whistle Detection Comparison with Silbido

Silbido is a state-of-the-art whistle detection program that is publicly available from the authors.<sup>1</sup> Their system is designed to extract precise contours of whistles in complex streams of vocalizations, allowing classification features to be extracted from individual contours. In a comparison with hand annotations, Silbido annotations will cover an average of 80% of the complete whistle and will fragment into multiple whistles [71]. It should also be noted that hand annotations are not perfect. People can misidentify the pitch track or not annotate the full whistle contour because the track is confusing or faint. To examine the performance of whistle detections made with Silbido, I ran the algorithm across the hand annotated

<sup>1</sup>[http://roch.sdsu.edu/software/silbido\\_JASA2011baseline.zip](http://roch.sdsu.edu/software/silbido_JASA2011baseline.zip)

test data, identifying individual whistles automatically for comparison. To create a test dataset, I ran a script that performed a time union of individual whistles to identify whistle regions containing no noise. The results of the best detection system (measured by median performance) and a second system with good median performance but lower variance were compared to the results from Silbido-annotated data Figure 4.10. The system with the best performance had the following configuration: 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint. The other configuration with good performance and lower variance had configuration: 20N60W joint bases, GMM5, no shuffle and no co-occurrence constraint. A per fold performance of these configurations is shown in Figure 4.11. When the required percent overlap with the true detection is low, (10, 20 and 30%) the median NMF performance roughly equalled the median Silbido performance. However, at higher thresholds (50 and 80% overlap), the median NMF detection performance was better. This could be due to Silbido having a higher fragmentation rate when detecting whistles, as noted in their paper [71].

Based on this series of tests, the challenge with an NMF based detection system was high variance observed between individual cross validation folds. Even in the case with minimum variance, the difference in F-score between the best and worst folds was on the order of 20% difference. This suggests the cross validation folds may be inhomogeneous despite the best effort to divide the overall data as evenly as possible for temporal and geographic variations.

#### 4.7 Summary

- Co-occurrence constraints and species-specific NMF bases decrease whistle detection performance
- Shuffle training showed small improvements for whistle detection in terms of average performance and across fold variation
- Higher order models showed better performance but additionally considering results of Chapter 5 this could be due to overfitting.

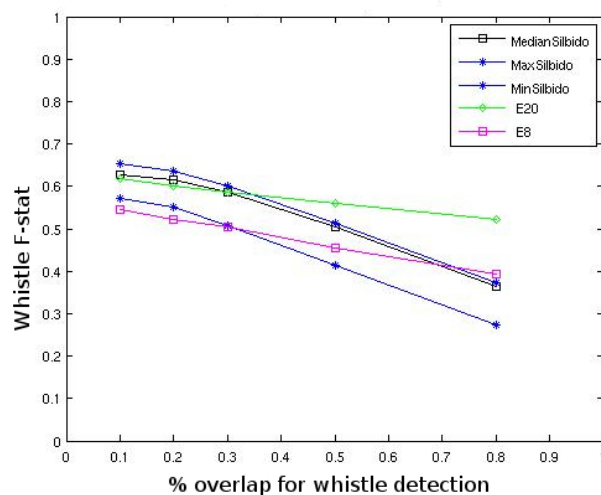


Figure 4.10: Median performance of best detection configuration which used 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint. This best performance is labeled (E20) in the figure. The configuration with the least variance across folds had 20N60W joint bases, GMM5, no shuffle and no co-occurrence constraint. This configuration is labeled (E8) in the figure. Silbido, including maximum and minimum Silbido performance are labeled accordingly.

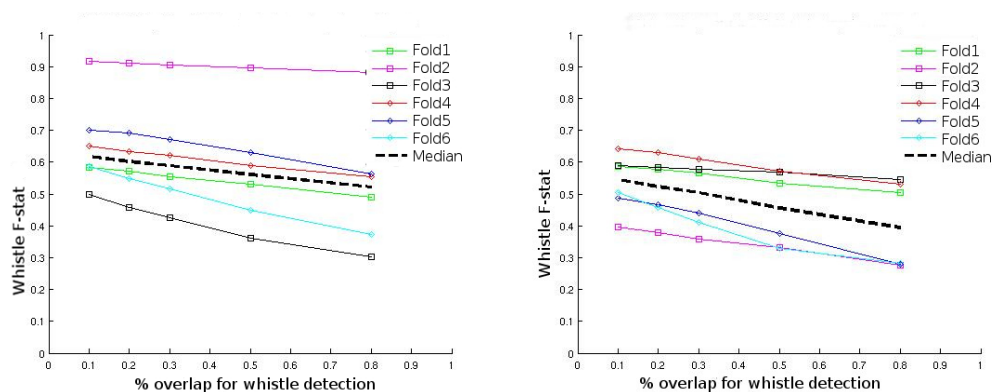


Figure 4.11: Median and per fold performance of best detection (left; 20N30W joint bases, GMM5, shuffle training and no co-occurrence constraint), and detection with least variance (right; 20N60W joint bases, GMM5, no shuffle and no co-occurrence constraint)

- NMF detection showed better performance than Silbido when more than 30% overlap was required.

## Chapter 5

**SPECIES CLASSIFICATION FROM WHISTLES**

The task of species classification is particularly important for a marine policy planning because regulations are species specific. In order to generate the science to validate if for example offshore energy permits can be granted, baseline usage statistics need to be taken into consideration, particularly for threatened or endangered species. In this chapter we aim to evaluate method configurations for species classification from whistles. There are several considerations we evaluate in the NMF system, including co-occurrence constraints that force the model to choose bases from within species specific subgroups. In a special experiment we compare the impact of using Silbido generated annotations by exploring the impact of including noise-only segments, as determined by hand annotated ground truth. System optimization is performed to determine the number of NMF bases and GMM mixtures. Training the NMF bases from a linear or randomly shuffled ordering of the data is also evaluated. For most configurations, including MFCC baseline configurations, hand annotations (when available) or Silbido based annotations are used to identify the time regions when whistles are occurring and make species predictions using features extracted only from those time regions. We establish a baseline with MFCC's, and compare it with the best case NMF based system. All species classification experiments were evaluated in a cross fold construction. Eval  $\alpha$  has hand annotations available which are used for development purposes and as a ground truth. Eval  $\beta$  has only automated annotations. Comparing performance on Eval  $\alpha$ , using automated Silbido annotations can quantify the degradation caused from using imperfect annotations facilitating interpretation of Eval  $\beta$ .

**5.1 Species Classification Methods**

The method we employ to perform species classification from whistles is similar to whistle detection. An important difference is the training data used to create the GMM classification

models is now species specific. Another difference is that species classification is operating on the output of whistle detection (see Section 3.1.4).

Based on the whistle detection experiments of Chapter 4, the species classification experiments are focused on a subset of configurations, resulting in three experimental comparisons. First we contrast the co-occurrence constraint settings for two configurations: 20 noise/60 whistles bases, and 20 noise/30 whistle bases, both with 5 GMM mixture components. In the second experiment we compare the impact of increasing the number of GMM mixtures for 20 noise/30 whistle and 20 noise/60 whistle bases, all with no co-occurrence constraint. Lastly, shuffle training is analyzed in contrast to an identical configuration without shuffle training.

Throughout the chapter, best performance per condition of each results table is highlighted in boldface.

## **5.2 Experiments**

### *5.2.1 Influence of Noise Class and Automated Annotations*

The species classification methodology relies on having some method of identifying the time regions containing whistles; this could be hand annotations or an automated detection algorithm. This information allows us to disregard the ambient background sections which would only provide additional recording condition bias. In this section we compare the difference in classification performance:

1. when we allow the decision function to choose a noise class in addition to the three possible marine mammal species.
2. when noise-only segments are included.
3. when Silbido based automatic annotations (SA) are used in contrast to the available hand annotations (HA).

Several definitions are required to understand this comparison.

	MacroE	Pf/Pm	MacroEN	Pf/Pm
Eval $\alpha$ (HA)	.303	.235/.427	.246	.149/.708
Eval $\alpha+$ (HA)	.218	.147/.427	.161	.091/.708
Eval $\alpha$ (SA)	.310	.240/.438	.243	.147/.690
Eval $\alpha+$ (SA)	.343	.281/.438	.219	.130/.690

Table 5.1

	Eval $\alpha+$ (HA)	Eval $\alpha+$ (SA)
No Q	<b>.161</b>	<b>.219</b>
Q1	.176	.232
Q2	.201	.305

Table 5.2: Comparison of hand and Silbido annotation types for the Eval  $\alpha+$  dataset as reported by MacroEN scores for three co-occurrence constraint conditions.

1. Eval  $\alpha$  = subset of hand annotated segments where each segment has at least one whistle
2. Eval  $\alpha+$  = full set of hand annotated segments, including noise-only segments.
3. MacroE - Macro error for the case when the classifier is forced to decide between one of the three species
4. MacroEN - Macro error for the case where classifier is allowed to choose a species or noise class.

Analyzing the results in Table 5.1, there are three key observations. First, comparing the scores for MacroE and MacroEN, it is the MacroEN scores that are lower. This is surprising because the addition of a fourth class would typically increase the difficulty of correct classification. In this data, having the noise class is practical because some segments

will be noise. Difficult species predictions can also be assigned to noise. Looking at the ratio of Pf (false detection) to Pm (missed detection), it confirms the false positive rate decreases when using MacroEN, and the missed detection rate increases.

The second key observation is obtained from the comparison of Eval  $\alpha$  to Eval  $\alpha+$ . The error rates are lower for Eval  $\alpha+$ , because many noise segments are correctly classified as noise. The hand annotated data is 100% correct because of the oracle condition and for Silbido annotations 69% of noise only segments are correctly predicted as noise.

The third key observation is obtained from comparing hand annotations to Silbido annotations. So me performance degradation from using automatic whistle detection was expected, but the difference for Eval  $\alpha$  (where all segments included whistles) was not large. The main effect was in the noise segments. This suggests that the NMF noise model is helpful for identifying unreliable Silbido whistles. This is important when no species is present but not as important when there is a mix of good and bad whistles.

Table 5.2, uses the Eval  $\alpha+$  data, which includes the noise only segments. Here it is shown that using automated (Silbido) detections degrades species classification by 5.6% to 10.4%. This is primarily due to false detections from the noise-only segments, even though noise is an allowed class.

### 5.2.2 *Co-occurrence Constraints*

Co-occurrence constraints provide a means of specifying NMF bases that must, may, or cannot co-occur with each other. This concept is introduced in Section 2.3.2 and is achieved by adding a Q matrix to define the co-occurrence relations that constrain the computation of NMF weights. The two test configurations are illustrated in Figure 4.5. Specifically, Q1 forces the weights to be associated with either noise or a single species, and Q2 allows noise weights together with a single species. The co-occurrence based formulation for learning weights is used both when training the GMM species models, as well as in evaluation. Although the co-occurrence constraints did not help in whistle detection, these experiments were species-independent. The constraints are more relevant to the problem of species classification.

	Cross Species Error		Macro Error		MacroEN	
	20N/30W	20N/60W	20N/30W	20N/60W	20N/30W	20N/60W
No Q	.465	<b>.450</b>	.309	<b>.303</b>	<b>.238</b>	.246
Q1	.513	.522	.328	.333	.285	.253
Q2	.525	.547	.340	.350	.336	.343

Table 5.3: Species classification performance on Eval  $\alpha$  (Hand annotations), for different Q conditions and different numbers of whistle bases. The best results are in bold.

In this series of experiments, the impact of three Q configurations is assessed, (Q1, Q2 and No Q) under two evaluation sets; Eval  $\alpha$ (HA) and Eval  $\beta$  (SA). Because Silbido or hand annotated whistle detection has been applied, it is assumed a marine mammal is present and force the decision function to choose between the three species, thereby excluding the noise class. However, for the MacroEN metric the noise class is allowed. All experiments in this set used 5 GMM mixtures and linear training. The performance for these co-occurrence experiments with cross-species error and macro-error statistics are reported.

When considering evaluation set  $\alpha$  with hand annotations (Table 5.3), best performance occurs for the No Q condition and worst performance for Q2 condition. There are performance differences between using 20 Noise/30 Whistle NMF bases as compared to 20 Noise/60 Whistle bases; however, the impact is smaller than the differences based on Q type.

For evaluation set  $\beta$ , no hand annotations are available, however automated Silbido annotations are used to identify whistle regions. To ensure there are at least some correct whistle detections in the file, files with less than ten detected whistles were discarded. Results are summarized in Table 5.4 and I observed different conclusions as compared to evaluation set  $\alpha$  with hand annotations: best performance with Q1 and worst performance with no Q. This was the intent when co-occurrence was introduced. However, for the MacroEN score, best performance is again for the No Q configuration when the noise class is allowed. This is because the miss rate increases, particularly for Q2 because it always

	Cross Species Error		Macro Error		MacroEN	
	20N/30W	20N/60W	20N/30W	20N/60W	20N/30W	20N/60W
No Q	.675	.679	.405	.402	<b>.237</b>	<b>.238</b>
Q1	.492	<b>.400</b>	.349	<b>.319</b>	.293	.290
Q2	.500	.469	.342	.340	.309	.331

Table 5.4: Species classification performance on Eval  $\beta$  (Silbido annotations), for different Q conditions and different numbers of whistle bases. The best results are in bold.

	Eval $\alpha$ (HA)		Eval $\alpha$ (SA)	
	MacroE	MacroEN	MacroE	MacroEN
NoQ	<b>.303</b>	<b>.246</b>	<b>.310</b>	<b>.243</b>
Q1	.333	.253	.341	.259
Q2	.350	.343	.339	.339

Table 5.5: Comparison of MacroE and Macro EN for hand and Silbido Annotations with no noise-only segments.

allows weight on the noise bases.

To better understand the result, performance was assessed with different Q constraints using hand annotations vs. Silbido annotations for Eval  $\alpha$  and Eval  $\alpha+$ .

Because no hand annotations exist for Eval  $\beta$ , the impact of these test conditions is observed with Eval  $\alpha$  data. The experiment configuration that showed the best performance on Eval  $\alpha$  (20N/60W, GMM5, NoQ, linear training) as well as the corresponding variations with Q1 and Q2 co-occurrence constraints, were selected for this test. Results are summarized in Table 5.5 and 5.2.

There are several observation to be made from Table 5.5 that are suggestive of the importance of noise and whistle bases as controlled by the co-occurrence constraints. Looking at the MacroEN scores, Q1 and Q2 both do worse than NoQ, with Q2 being worst. The

Q1 construction does not allow weight on the noise bases, unless noise is considered the primary class. The Q2 construction allows noise to co-occur with the other classes. Because macroEN includes noise as a class, it is logical that Q2 would do worst because it would be more prone to having high weights on the noise bases. Q1 would only allow noise weights if noise was truly dominant. With that interpretation in mind, it must be addressed that NoQ performs better than Q1. The dataset in this study was split as evenly as possible based on provide temporal and geographic information, but likely still remains inhomogeneous. One explanation for the NoQ performance is that the whistle bases of one species are useful for prediction of other species, and the limits imposed by the Q1 co-occurrence constraints negates that advantage, resulting slightly worse performance. In Table 5.4, Q1 does perform better than NoQ, but this is maybe because the bases are also capturing channel effects that are not present in Eval  $\beta$ .

Another observed trend from Table 5.5 is consistently better performance for the MacroEN. In this comparison, noise-only segments were withheld, suggesting that for segments that do have whistles, allowing the noise decision helps by reducing false detections more than correct detections. Additionally, the Silbido based species prediction errors are predominantly going to noise, suggesting that an annotation strategy that misses a few whistles is not a problem since there are other whistles to make the decision.

### 5.2.3 Influence of GMM mixtures

The second question addressed in the analysis of species classification from whistles is the influence of the number of GMM mixtures. For this set of experiments we hold fixed the No-Q configuration and linear training. Consider two bases configurations, 20N/30W and 20N/60W and report results from each of the three evaluation scenarios, starting with Eval  $\alpha$ (HA) in Table 5.6.

As also observed in the Q configuration experiments, the number of NMF bases has a relatively small impact compared to our other test variable, which in this case is the number of GMM mixture components. Considering the 20N/30W case best performance was with 5 mixture components as measured by both CSE and MacroE. Overall, the best performance

	Cross Species Error		Macro Error		MacroEN	
	20N/30W	20N/60W	20N/30W	20N/60W	20N/30W	20N/60W
GMM5	<b>.465</b>	.450	<b>.309</b>	.303	<b>.238</b>	.246
GMM10	.469	.472	.314	.325	.255	.259
GMM15	.484	<b>.393</b>	.333	<b>.268</b>	.258	<b>.228</b>

Table 5.6: Species classification performance on Eval  $\alpha$  (Hand annotations), for different GMM mixtures and different numbers of whistle bases, with no Q co-occurrence constraints. The best results are in bold.

was with the model with the most degrees of freedom: 20N/60W NMF bases and 15 GMM mixture components. This could occur because with the lower numbers of GMM mixtures (5 and 10), there are insufficient degrees of freedom to capture the variability in 20N/60W base. In contrast, 20N/30W bases have less variability due to the smaller number of bases and there are not gains to be had by increasing the number of mixtures.

As a second test, the same models with Eval  $\beta$ , using Silbido annotations, Table 5.7 are evaluated. As before performance is overall worse than as compared with Eval  $\alpha$ , but again the findings from Eval  $\alpha$  do not hold with Eval  $\beta$ . Notably, for both CSE and MacroE performances measure of the 20N/60W NoQ case, best performance is with 5 GMM mixture components. This observation suggests for mismatched conditions, fewer degrees of freedom provided better results.

#### 5.2.4 Influence of Shuffled Training Data

It has been noted in prior research [47, 99] that if time series data such as audio is processed in a consecutive manner, the algorithms are more prone to over-fit local phenomena. To mitigate that problem, linear and a random ordering of the NMF bases training data are contrasted and the influence on species classification was observed. In the shuffle training configuration, the exact same data was presented, but in a randomized order. Other parameters were held constant: 5 GMM mixture components, 20N/30W NMF bases and no

	Cross Species Error		Macro Error		MacroEN	
	20N/30W	20N/60W	20N/30W	20N/60W	20N/30W	20N/60W
GMM5	<b>.675</b>	<b>.679</b>	.405	<b>.402</b>	.237	.238
GMM10	<b>.679</b>	.747	<b>.400</b>	.407	.268	<b>.212</b>
GMM15	.688	.703	.411	.410	.282	.290

Table 5.7: Species classification performance on Eval  $\beta$  (Silbido annotations), for different GMM mixtures and different numbers of whistle bases. The best results are in bold.

	Cross Species Error		Macro Error		MacroEN	
	Eval $\alpha$ (HA)	Eval $\beta$ (SA)	Eval $\alpha$ (HA)	Eval $\beta$ (SA)	Eval $\alpha$ (HA)	Eval $\beta$ (SA)
Linear	<b>.465</b>	.675	<b>.309</b>	.405	<b>.238</b>	.237
Shuffle	.475	<b>.648</b>	.314	<b>.403</b>	.252	<b>.218</b>

Table 5.8: Cross species error and Macro-error comparison for linear and shuffle training.

Q co-occurrence constraints.

The cross-species error and macro-error were compared for both linear and shuffle conditions. Results are reported in Table 5.8. It was found that shuffle training showed no improvements for Eval  $\alpha$ , but nearly 3% improvement in CSE for Eval  $\beta$ . Differences in MacroE were minimal for both evaluation sets. This was consistent with the possibility of overfitting to the channel conditions of the hand annotated data. Based on the GMM mixture and co-occurrence experiments, it is likely the opposite conclusions seen in shuffle training with Eval  $\beta$  are likely due to the data itself and not due to differences in annotation type because this trend also holds for the MacroEN score as well.

	Eval $\beta$ (SA)		Eval $\alpha$ (HA)	
	MFCC	NMF-Q1	MFCC	NMF-NoQ
CSE	.619	<b>.400</b>	.519	<b>.450</b>
MacroE	.395	<b>.319</b>	.359	<b>.303</b>
MacroEN	.333	<b>.290</b>	.257	<b>.256</b>

Table 5.9: MFCC baseline results for Eval  $\alpha$  and Eval  $\beta$  as compared to best performance NMF

### 5.2.5 Baseline Comparison

To establish a baseline for this particular dataset Mel-Frequency Cepstral Coefficients with GMM models are used. Specifically, 18 MFCC's (with first and second deltas) are used as features for a GMM with 5 mixtures. Noise subtraction was applied via median smoothing, as was used in all other experiments. Results are summarized in Table 5.9. I compare the resulting classification model performance with MacroE and CSE statistics for two evaluation scenarios. For both Eval  $\beta$  and Eval  $\alpha$ , the best case NMF configurations outperformed the MFCC baseline as measured by CSE, MacroE and MacroEN. By all error metrics, the amount by which NMF outperformed MFCC's is greater for Eval  $\beta$ , where the channel mismatch plays a bigger role.

## 5.3 Summary

This chapter compares species classification results for a variety of system parameters. In the course of these experiments a few key points can be concluded:

1. Comparing hand annotations with Silbido annotations, when noise only segments are withheld, the Silbido annotations resulted in very little degradation to performance. For MacroEN, Silbido annotations performed slightly better. When noise only segments are included, the more realistic scenario, hand annotations perform better.
2. Less complex models (fewer bases and fewer GMM mixtures) are more useful in a

scenario with channel mismatch.

3. Co-occurrence comparisons suggest that the species bases may contain channel effects, and/or that the bases of other species can be helpful in producing better species predictions.
4. The use of shuffle training only showed benefit for Eval  $\beta$ , likely because it is less well matched to the training set than Eval  $\alpha$
5. NMF has been demonstrated to substantially beat the MFCC baseline in a species classification task.

## Chapter 6

**WEAK SUPERVISION**

A common problem for PAM applications is the time consuming and expensive nature of annotating large volumes of data, particularly by hand. Weak supervision aims to address this by providing a means of adding training examples from partially labeled data. Weak supervision is often accomplished in several stages. The first stage builds initial models from a small dataset with annotations and labels, if available. The second stage, which may be repeated multiple times, re-trains the model with additional examples from the partially labeled data. In the DCL dataset, partial labeling in the data consisted of species identification, detected visually when the data were recorded. In these data files, there were no hand annotations of time frequency tracks for whistles. For the experiments conducted in this chapter, the technique was applied to both click and whistle data, although the details of implementation differed.

**6.1 General Methods**

A naive approach to weak supervision uses an automated detection method to extract vocal events from partially labeled data and incorporates all detected examples into training. A self-training approach will downselect these additional data based confidence or other quantifiable metric. An example of the self-training approach is outlined in the following steps:

1. Detect vocalizations in unlabeled data using a species-independent detector
2. Build initial GMM models per species.  $\{\Theta_j^0\}, j \in \{b, c, m, noise\}$ <sup>1</sup>
3. Iterate  $p = 0, 1, \dots, P$ , where  $p$  counts the number of iterations.

---

<sup>1</sup>As a reminder, b=Bottlenose, c=Common, and m=Melon.

- i) Use species classifier to determine species posteriors,  $p(j|\Theta_j^p)$  for all detected vocalizations  $x_i$  in the unlabeled dataset
- ii) Select vocalizations for inclusion in the next pass of training based on a match of the maximum species posterior probability to the true species file label.
- iii) Build new model  $\{\Theta^{p+1}\}$

4. Iteration stops when  $p = P$ .

For the click classification task, I had only partial labels for all data. Click-like sounds could be produced by either the visually observed species or a variety of noise sources, including self-noise from the recording system. I applied the self-training approach to this task, experimenting with different numbers of iterations, assessing the models in file-level species classification test for each iteration. Species decisions were being made per click, but the model performance was judged based on the number of 30s file segments having correct species prediction. The details of this decision function are discussed in Section 6.2.1. Because I was initially unable to discriminate between species and noise clicks, all were included in the initial model  $\{\Theta_j^0\}$ . The self-training process computes a new model  $\Theta_j^{p+1}$  by selectively adjusting which examples were used from the partially labeled and unannotated data. The pool of possible clicks added in each iteration of the self-training process remained fixed because the click detection model is unchanged. However, membership in training data of the  $p^{th}$  model is determined by confidence of correct classification from the  $p - 1$  species classification model.

With clicks, the weak supervision process was used to automatically learn which clicks were potentially noise, such as from a man-made source or fish, rather than a marine mammal, or any click with weak predictive power, such as off-axis clicks. The work of [52] clearly illustrates the significant variation of characteristic peak frequencies that occur with as little as +/-15 Deg off-axis. This problem has been a significant weakness to prior attempts at classifying marine mammal species from their clicks, because data recorded in an open environment has no inherent means of determining the orientation of the animal to the hydrophone. Dolphins swim in a dynamic manner. Individuals may engage in changes of

direction and relative position even when the movement of the school as a whole is generally directional. Attempts have been made to characterize this variability [2, 90, 93], but the approach here removes the intermediate attempt to model or make assumptions about the orientation and automatically learns the click types that are most uniquely distinguishing for the particular species.

The weakly supervised method can also be applied to whistles. As discussed in Section 3.1 some of files with whistles had hand annotations but most of the whistle data were only partially labeled. In these cases, I knew that a hypothesized whistle was either noise or the identified species, but I had no annotations to indicate when the whistles occurred. I compared the naive and self-training approaches applied to the whistle data in the same manner as for the clicks. Having an accurate species classification model was important and the experiments of Chapter 5, informed that choice.

## 6.2 Clicks

Initial investigation of weak supervision with click based species classification was presented in Nichols et al. [62], where 3 iterations of weak supervision were tested and a 7-12% improvement in species classification was observed. This prior work and the current experiments used the 5th DCL data set. The method for detecting clicks and extracting MFCC features are both described in Section 3.1.2, and the MFCC features are used to build a GMM to perform species classification. The work described in this chapter includes more extensive analysis, investigated GMM re-seeding, performed further tests on the optimal number of weak iterations, and used an improved data partition to reach more definitive conclusions

Session effects are a common problem in datasets recorded at different locations and are described at length in Roch et al. [75]. In prior work with species classification from clicks, I observed what were likely significant session effects, attributable to data partitions. This suggested the cross validation construction for further tests. For the current set of experiments, I used the 5-fold cross validation configuration, which had an even distribution of temporal and geographic variability in each fold. This division is discussed in detail in Section 3.2.1. The following experiments tested key aspects of the system configuration:

decision functions, threshold selection, and GMM re-seeding.

### 6.2.1 Decision functions

When assigning an individual click to a species category, posterior values are the obvious choice. For the file level species assignment, two strategies were tested, MAP and vote. The vote method makes its choice based on the counts of individual clicks classified to each possible species, and the winning species is the species with the most votes. The MAP strategy uses the total log posterior probability of all detected clicks and chooses the species class with the highest probability. Initial work [62] suggested the MAP decision function performed slightly better, but session effects masked the results, making strong conclusions difficult. This finding was re-evaluated with the 5-fold cross validation construction and up to 6 iterations of self-training.

Results of the experiment are shown in Figure 6.1. As before, the MAP decision strategy gave slightly better performance for both cross species error and macro error.

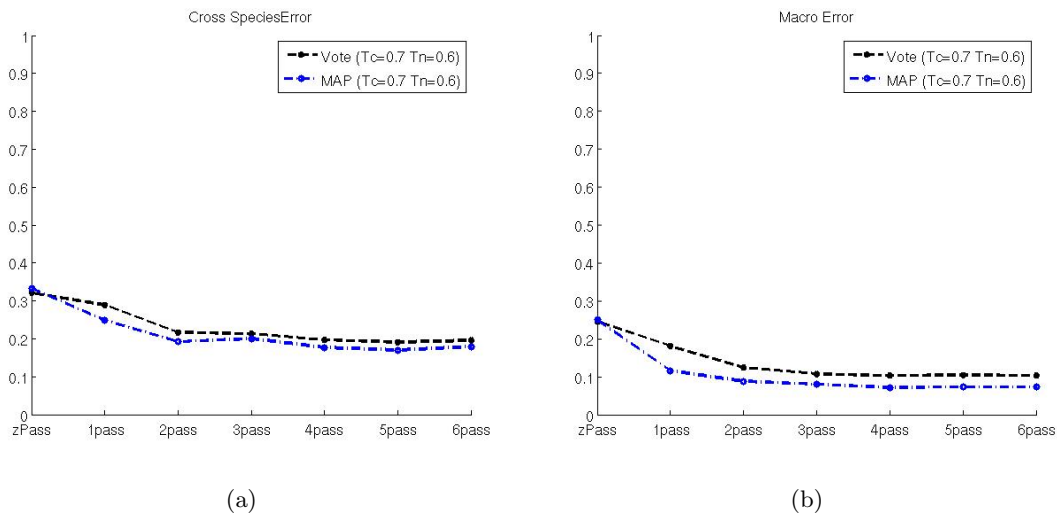


Figure 6.1: Comparison of MAP vs. Vote decision functions for best threshold configuration ( $\tau_c = 0.7, \tau_n = 0.6$ ): (a) Cross Species Error and (b) Macro Error

$\tau_c$	$\tau_n$
0.6	0.6
0.7	0.6
0.7	0.7
0.6	0.7

Table 6.1

### 6.2.2 Threshold selection

Perhaps the most important aspect of the weak supervision processes for clicks was selection of events to be included in future model iterations. Adding or keeping potentially poor examples would obviously reduce model accuracy. To identify which clicks should be rejected as noise, I examined the posterior values of individual click events, assigning each to one of three categories, confident, weak, or noise. Confident clicks were classified when the species was known to be present in the segment and had a high posterior probability. Weak clicks were classified as the target species but with a lower posterior probability. A click was assigned to the noise category if it received an incorrect species label with a high probability. A click classified with an incorrect species label with a lower probability was not used in training noise or species models.

It was generally observed that the weak supervision process benefited from including correctly labeled clicks in early iterations, regardless of the confidence. Initially this was controlled by building models including confident and weak clicks for the first one or two iterations. Subsequent iterations used only confident clicks. Further investigation showed that the benefit could be controlled by adjusting posterior thresholds, specifically determining the value above which a posterior was considered high probability and below which it was considered low probability. I controlled the two thresholds (noise confidence and click confidence) independently.

By looking at the posterior probabilities for a subsample of clicks, candidate values of 0.6

and 0.7 were chosen for  $\tau_c$ , threshold of confidence for clicks with the correct species label and  $\tau_n$ , threshold of clicks incorrectly labeled and to be assigned as noise. These parameters  $\tau_c$  and  $\tau_n$  were adjusted to a range of possibilities but the most insight was gained from comparing the four combinations noted in Table 6.1. Figure 6.2 shows the experimental results. Of these options, the best strategy for improving model classification via weak learning was to be conservative in assigning clicks to the click category, via  $\tau_c = 0.7$ , but liberal in assigning clicks to noise,  $\tau_n = 0.6$ . With this best configuration, cross-species error was reduced 15% between the zero pass (33% CSE) and the 6th pass (18% CSE).

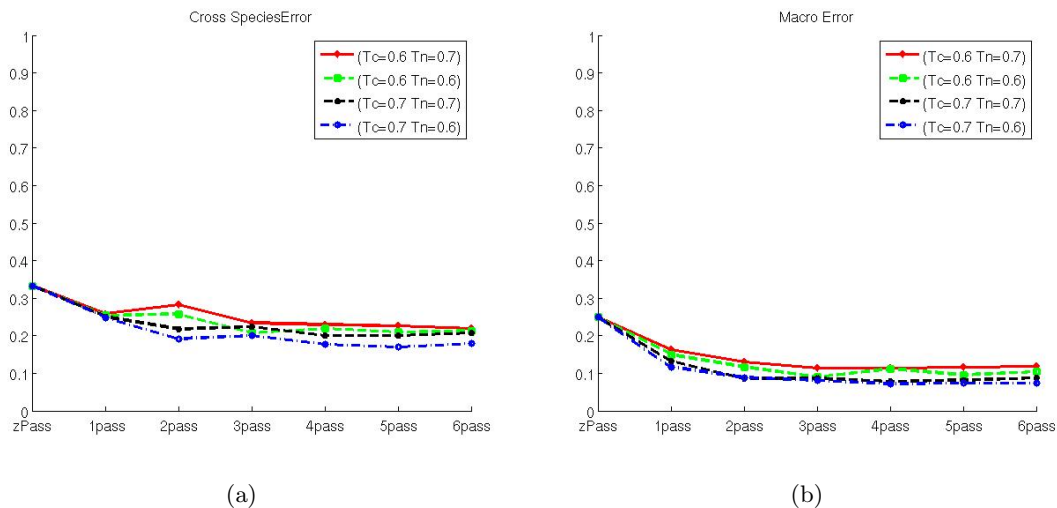


Figure 6.2: Comparison of Threshold settings: (a) Cross Species Error and (b) Macro Error

### 6.2.3 GMM re-seeding

When training GMMs in a weakly supervised training scenario, a new GMM model is calculated at each weak iteration step. The initial species models are constructed using all detected clicks or whistles in the files for each species. The  $p^{th}$  GMM is constructed from a subset of the species clicks or whistles, which are chosen to most confidently represent the species, as measured by the posterior probability with respect to the last GMM ( $p - 1$ ). The mean and covariance parameters for the  $p^{th}$  model can be initialized for GMM training

in two ways. The mean and covariance parameters for the  $p$ th model can be initialized for GMM training in two ways. Either they could be randomly reinitialized, as for the very first model, or the  $p$ th model could be seeded with the parameter values from the  $p - 1$  model. Both methods of initialization are evaluated. For completeness I compared the two approaches with all four threshold comparisons. In two cases GMM re-seeding performed nearly identically to the comparable experiment configuration with random initialization, Figure 6.6 and 6.5. In the other two cases, Figure 6.4 and Figure 6.3, GMM re-seeding performs slightly worse.

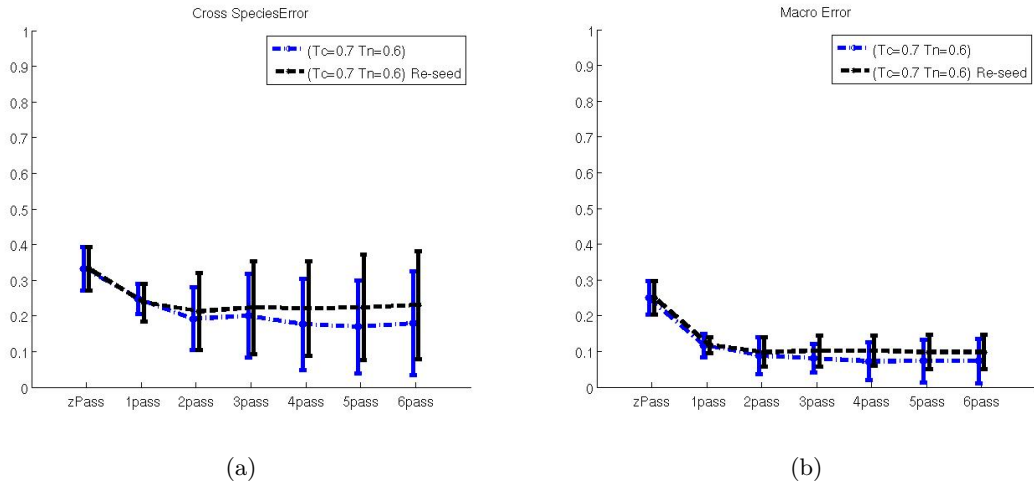


Figure 6.3: Comparison of GMM re-seeding for  $(\tau_c = 0.7, \tau_n = 0.6)$ : (a) Cross Species Error and (b) Macro Error

### 6.3 Whistles

When applying weak supervision to the whistle portion of the data, I had the advantage of some hand labeled annotations from which an initial model could be trained. This gave a higher confidence that all training examples were correct examples. However, because the annotated set was small, the full variability that naturally occurred may not have been incorporated into the model. Ideally, with accurate initial models and correct threshold settings, the sequence of models could be iteratively improved as partially labeled data that

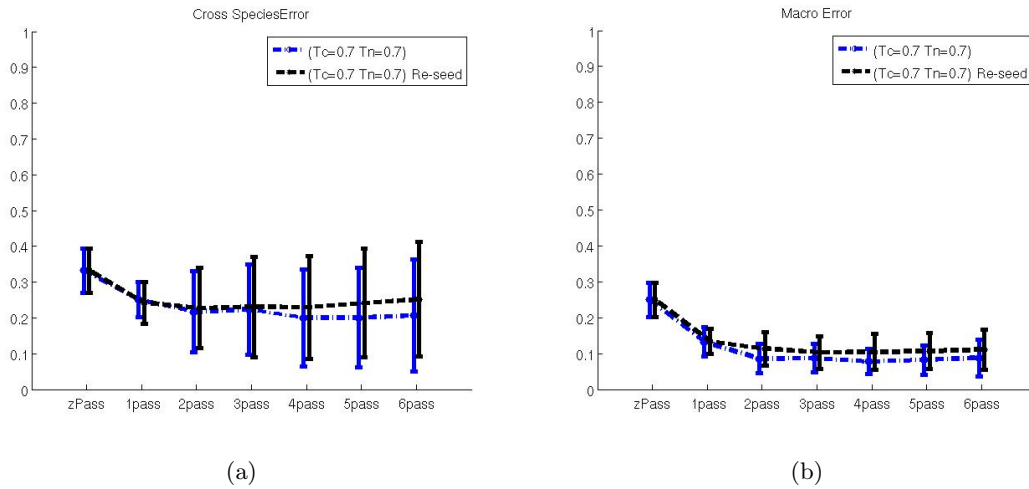


Figure 6.4: Comparison of GMM re-seeding for  $(\tau_c = 0.7, \tau_n = 0.7)$ : (a) Cross Species Error and (b) Macro Error

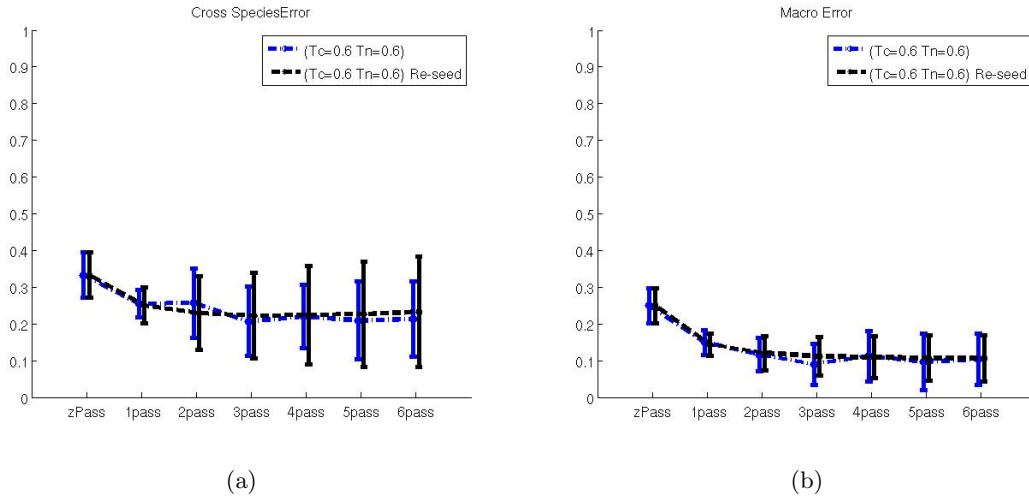


Figure 6.5: Comparison of GMM re-seeding for  $(\tau_c = 0.6, \tau_n = 0.6)$ : (a) Cross Species Error and (b) Macro Error

did span the larger set of variability were incorporated.

Experiments demonstrated the results of two alternative methods for adding partially labeled data to whistle based species classification: naive training and self-training. In

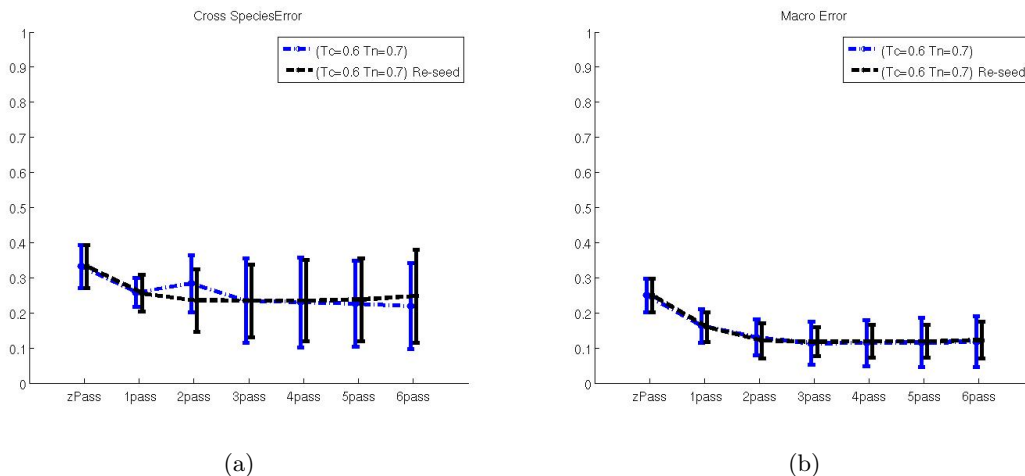


Figure 6.6: Comparison of GMM re-seeding for  $(\tau_c = 0.6, \tau_n = 0.7)$ : (a) Cross Species Error and (b) Macro Error

both methods Silbido was used to detect hypothesized whistles in the partially labeled data. Naive and self training methods then diverged, each using different strategies to incorporate detections.

There are several phases of self training

1. pick a zero pass model
2. identify whistle unions to be added to NMF bases training from partially labeled data. (See Section 3.1 for a reminder of how whistle unions are obtained.)

for naive training, this was all whistle unions that were detected

for self training, this was the subset of detected whistle unions that are classified to the correct species via the zero pass model.

3. update NMF bases
4. with new bases, extract new weights for learning new GMMs
  - i) extract all annotations from the hand annotated data
  - ii) downselect only correct species predictions from partially labeled data

5. evaluate performance of new naive or self-trained models using Eval  $\alpha$  and Eval  $\beta$

### 6.3.1 Naive Weak Supervision

Naive experiments took the simple approach of incorporating all hypothesized whistles that Silbido reported; 50% were added to NMF bases training and the remainder were used in GMM training. Within this context, I compared four variations of naive training with two baseline configurations that used only fully annotated data. The first three experiments were designed to determine the importance of which bases (noise, whistle, or both) would most benefit from being augmented with partially labeled data. The fourth experiment evaluated the influence of Q1 co-occurrence constraint (see Section 4.3 and Figure 4.5 for details of this configuration). To set the parameters for the naive experiment, I choose one of the best whistle detection configurations (Split Bases training, No Q co-occurrence, 20N60W, 5 GMM), used the outcome of this experiment as a baseline for performance (Base-NoQ) and used identical parameter configurations for the three naive experiments. In the first naive whistle experiment (NaiveNW-NoQ), I updated the noise and whistle NMF bases, and used these new bases to re-learn GMM species models, using 50% of the partially labeled data for each task. The second and third naive experiments repeated this parameter configuration but modified which bases were updated. In an effort to understand whether it was more important to account for session effects (noise) or whistle variability, the second experiment (NaiveN-NoQ) updated only noise bases, and in the third experiment (NaiveW-NoQ) updated only whistle bases. To perform the partial updates, the designated bases from the baseline and first naive experiment were concatenated (e.g. baseline whistle bases with updated noise bases). With the appropriate sets of bases constructed, GMM training proceeded using the fully annotated data plus 50% of the remaining partially labeled data, as in the NaiveNW-NoQ experiment. The results of these experiments are shown in Table 6.2.

For both error measures, one trend stood out. All three naive experiments performed worse on the Eval  $\alpha$  data but performed significantly better on the Eval  $\beta$  data as compared to the baseline. The Eval  $\beta$  performance was nearly identical for all naive configurations.

	Cross Species Error		Macro Error	
	Eval $\alpha$	Eval $\beta$	Eval $\alpha$	Eval $\beta$
Base-NoQ	0.45	0.68	0.30	0.40
NaiveNW-NoQ	0.70	0.47	0.48	0.32
NaiveN-NoQ	0.60	0.46	0.43	0.34
NaiveW-NoQ	0.63	0.47	0.43	0.32

Table 6.2: File level species classification performance for naive weakly supervised training of whistle and/or noise models.

These two observations implied that despite efforts to split the data evenly based on time and geographic location, a strong session effect could still have remained. There are many contributors to session effects that were not controlled, for example hydrophone depth in the water column, variation in the animal’s genetics, behavioral state, school social structure, etc. Another consideration is the Eval  $\beta$  dataset is much larger, so files that are potential outliers will have less impact on the results.

Based on results in Chapter 5, I theorized that the Q co-occurrence constraint could reduce the impact of session effects. I tested this theory by performing one more comparison. First I selected a new baseline (Base-Q1) trained identically to Base-NoQ, but using the Q1 co-occurrence constraint. This new baseline was compared to a naive experiment with naive updates of both noise and whistle bases and the Q1 constraint (NaiveNW-Q1). The results of this experiment are shown in Table 6.3. I observed that with a Q1 co-occurrence, the performance of NaiveNW-Q1 was worse than Base-Q1 on the Eval  $\alpha$  data. This trend was also observed in Table 6.2, when no Q co-occurrence is used, but the performance degradation is slightly higher for the no Q. While the relative benefit of the naive data is small for Eval  $\beta$  when Q1 is used, the result is much better for cross species error than without Q1. Comparing Base-NoQ to NaiveNW-Q1 performance on Eval  $\beta$ , the cross-species error decreased 31% and the macro-error is decreased 20%.

Both of these Base experiment models (with and without Q) were trained exclusively

	Cross Species Error		Macro Error	
	Eval $\alpha$	Eval $\beta$	Eval $\alpha$	Eval $\beta$
Base-Q1	0.45	0.68	0.30	0.40
Base-Q1	0.52	0.40	0.33	0.32
NaiveNW-NoQ	0.70	0.47	0.48	0.32
NaiveNW-Q1	0.60	0.39	0.38	0.32

Table 6.3: File level species classification performance for naive whistle models with and without Q1 co-occurrence

from hand annotated data (of the type used in Eval  $\alpha$ ). Assuming a session effect between the annotated and unannotated data partitions, this would result in a mismatch for the partially labeled data of Eval  $\beta$ .

### 6.3.2 Self-training experiments

The whistle-based self-training experiments had a more complex setup as compared to the naive experiments, which incorporated all detected whistle unions. The prior whistle based species classification reported in Chapter 5 used a MAP decision based on all detected whistle unions in a 30 second file. At no point were species predictions calculated per individual whistle union within a file. In the weakly supervised click classification, the decision to add or withhold a click was made for each individual click. For the whistle based weakly supervised experiments to have comparable methodology to the click based weakly supervised experiments, I changed the resolution at which species predictions were made and computed species probabilities for each whistle union. The decision to add or withhold a new Silbido-detected whistle was independent for each whistle union, but when evaluating model performance I continued to use the MAP decision based on all whistle unions in the file.

It was difficult to find a zero pass model that could correctly predict species per individual whistle union, for two likely reasons. First, there were fewer samples from which to make a

prediction. Based on data in Table 3.4, I was making the prediction from approximately .5 seconds of data and generally one whistle, as compared to 14+ seconds of data and multiple whistles. The second reason was session effects noted in the naive training experiments. Using a zero pass model, which was only trained on the hand annotated dataset (Base-NoQ), to find new whistle unions in the partially labeled data was completely inadequate. Species classification of the individual whistle unions was worse than random chance. Using co-occurrence constraints with hand annotated data (Base-Q1) was not sufficient but some reasonable results were obtained by combining co-occurrence constraints with naive updates (NaiveNW-Q1) for the zero pass model.

See Table 6.4 for a description of the number of whistle unions that were classified as the known species for use in the supplemental NMF bases training phase of self-training. In this table the number of total unions was the number of unions detected by Silbido. I did not expect all of these unions to be classified as the known species since some might not have been true whistles. In addition, Silbido might have detected whistles differently, e.g., as fragments of whistles.

I examined the data to determine relations that might explain the results. I observed that melon-headed whale whistles were correctly classified as the known species at a much higher rate. This could have been due to the fact that all data for this species was recorded in the same geographic location (Palmyra) and generally had a larger volume of available data. For the bottlenose dolphin data, folds 1 and 5 were particularly problematic for Base-Q1. These both used partition B, which had very little Palmyra data, to train the GMM. For common dolphins, folds 5 and 6 performed better than all other folds and both of these configurations did not use partition A in training NMF bases or the GMM model. It is possible this was because the common dolphin A partition had no Palmyra data.

With the zero pass model chosen (NaiveNW-Q1), I now proceeded with one pass of self training. Using NaiveNW-Q1 to perform species classification, I selected the Silbido-detected whistle unions that had correct species prediction and updated the NMF whistle bases with this new training. Using the newly updated bases, I extracted new weights and learned a new GMM using only Silbido detected whistle unions that had correct species prediction. I then evaluated performance of the new model with the Eval  $\alpha$  and Eval  $\beta$

Fold/Species	Base-Q1 Correct Unions	NaiveNW-Q1 Correct Unions	Total Unions
Fold1 - Bottlenose	1	314	457
Fold1 - Common	4	6	483
Fold1 - Melon	3886	4040	4139
Fold2 - Bottlenose	336	1453	1683
Fold2 - Common	19	77	232
Fold2 - Melon	2549	1582	4285
Fold3 - Bottlenose	121	201	535
Fold3 - Common	3	19	376
Fold3 - Melon	2346	3880	4818
Fold4 - Bottlenose	210	202	221
Fold4 - Common	0	9	254
Fold4 - Melon	5273	5522	6417
Fold5 - Bottlenose	0	353	960
Fold5 - Common	141	336	379
Fold5 - Melon	1706	1557	1707
Fold6 - Bottlenose	606	508	1854
Fold6 - Common	148	142	148
Fold6 - Melon	138	587	1372

Table 6.4: Table of the number of whistle unions added to NMF bases training by using the BaseQ1 and NaiveNW-Q1 models as the zero pass model. In this case, correct prediction means identified as the known species.

	Cross Species Error		Macro Error	
	Eval $\alpha$	Eval $\beta$	Eval $\alpha$	Eval $\beta$
Base-Q1	0.52	0.40	0.33	0.32
NaiveNW-Q1	0.60	0.39	0.38	0.32
SelfTrain-Q1	0.65	0.38	0.40	0.35

Table 6.5: File level species classification for Self-training results (SelfTrain-Q1) as compared to NaiveNW-Q1 (zero pass) model and Base-Q1 baseline.

datasets. The results are presented in Table 6.5. As also observed in the naive experiments, performance was degraded on the Eval $\alpha$  data; there was a marginal gain for the cross-species error for Eval  $\beta$ ; and a decrease in performance for Macro Error. This is due to the fact that most of the test samples are for the melon-headed whale, which improves at the expense of the other species.

As demonstrated in the click data, self training holds significant potential for learning from partially labeled data. However, for the whistle domain, I believe more advanced techniques to compensate for session effects will be needed for iterative self-training before it can show its full potential. Session effects were a key obstacle in whistle based self training. To analyze this performance I looked at the per fold confusion matrix for Eval  $\beta$  (Table 6.6). Note the evaluation data were identical in folded pairs 1&2, 3&4, and 5&6. For example, if confusion matrices are compared for fold 1 and 2, observations can be made about two models evaluated on the same data.

The results in a fold pair test on the same data but switch which fold use used for training NMF and GMMs. For example Fold 1 and 5 use partition B for GMM training. The B partition has the least Palmyra data for the bottlenose dolphin, but most of the Eval  $\beta$  test data is from Palmyra and bottlenose dolphin detection is worst on those folds.

Looking at Table 6.6 the best results for common dolphin is fold 6 which trains NMF on partitions B and 2, GMM on C and 2, and evaluates  $\beta$  performance with partition 3. From Table 3.8 and 3.7, the B and 2 partitions have the most data recorded at Flip and

all data for the common dolphin is recorded at Flip. This gives the model the best chance of learning noise bases for that region. fold 3 is the only other set with correct detections for the common dolphin. In this case the NMF bases are trained on the A partition, which also has a substantial amount of data from Flip.

There are numerous variants to the self-training method that could improve performance, such as updating the noise and whistle NMF bases instead of just the whistle bases. Alternatively, only whistle unions greater than a minimum length could be added to improve the quality of data being added to the NMF and GMM updates. The experiments here tested only one iteration of self-training, but it is possible that additional iterations could provide improvements in classification, particularly with a better starting model. We could try a more geographically balanced subset of the partially labeled data. It is also possible this technique would be more feasible if models can be successfully adapted to specific locations.

#### **6.4 Summary**

This chapter explored weakly supervised training for clicks and whistles with mixed results:

- Weakly supervised species classification from clicks was successful at reducing cross-species error by up to 15%. The largest improvements were gained in the first pass of training, but additional small gains were achieved with further iterations.
- The MAP decision criterion had consistently better performance than vote criterion.
- When determining the strategy for assigning thresholds, the best results were obtained by being liberal in assigning clicks to noise and conservative in assigning clicks to the click category.
- Weakly supervised species classification from whistles was successful at improving classification performance for the mismatched Eval  $\beta$  condition, but the benefit over using the Q1 constraint was small. Because of the recording mismatches, the weakly supervised training hurt performances on Eval  $\alpha$ , but less so with the Q1 constraint.

Fold/Species	Bottlenose	Common	Melon
Fold1 - Bottlenose	0	0	93
Fold1 - Common	18	8	23
Fold1 - Melon	0	0	223
Fold2 - Bottlenose	72	0	21
Fold2 - Common	2	4	43
Fold2 - Melon	82	0	141
Fold3 - Bottlenose	30	69	3
Fold3 - Common	7	13	13
Fold3 - Melon	10	58	148
Fold4 - Bottlenose	67	14	21
Fold4 - Common	13	0	20
Fold4 - Melon	56	0	160
Fold5 - Bottlenose	0	0	96
Fold5 - Common	0	0	26
Fold5 - Melon	5	0	351
Fold6 - Bottlenose	18	29	49
Fold6 - Common	0	22	4
Fold6 - Melon	108	33	215

Table 6.6: Table of confusion matrices for self-training results on the Eval  $\beta$  dataset.

- Analyses showed that the specific subsets of data used in training the NMF bases had a substantial impact on performance. There were several likely causes that limited the effectiveness of this method including session effects that were not controlled. (e.g. behavioral state of the animal, hydrophone depth in the water column, etc)

## Chapter 7

# CONCLUSIONS AND FUTURE WORK

### *7.1 Summary of Contributions*

To summarize, there were three main contributions of the thesis.

First I demonstrate the potential for NMF to bypass or complement the need for explicit contour extraction in marine mammal whistle detection and species classification tasks. Though we use Silbido to perform whistle detection to identify NMF training segments, an alternate method that only provided start and end times of whistles would be equally sufficient to provide input training for our methods. Prior species classification methods rely on precise extraction of the whistle contour which is used to compute the classification features but can add a compounding error. This contour extraction is easily hindered by background noise and the noise removal itself has potential to distort the frequency contour. In the NMF scenario, the NMF bases are trained to represent time-frequency patterns of whistles and the activation weights serve as a classification feature. The NMF bases and weights are learned without contour extraction and reducing the potential for compounding error. For tasks with complex audio structure, such as marine mammal vocalizations, the convolutional approach is more effective at capturing temporal variability. Experiments comparing NMF-based species classification with hand annotated vs automatically detected whistle regions shows that performance degradation is not large for automatically detected whistles except in long stretches of noise, showing that the NMF system is not sensitive to precise whistle tracks. The NMF whistle features are also designed to be jointly learned with noise features, which can be tuned to different noise environments if matched training data is available. I demonstrated that for a whistle detection task NMF can perform as well as Silbido and with less fragmentation of the detected whistles. In experiments, NMF-based species classification significantly outperformed an MFCC baseline. NMF has the additional advantage of being able to learn non-stereotyped vocalizations, which are typically omitted

from species classification analysis.

A second contribution is demonstration of the co-occurrence matrix providing compensation for session effects when there is a mismatch present between the training and evaluation datasets. In our data sets there was a mismatch between the hand annotated data and the data which only had partial labels. Using otherwise identical parameters in species classification, co-occurrence constraints decreased the cross-species error 28% and decreased the macro-error 8%, on the mismatched evaluation data.

The third contribution involved an exploration of methods of integrating partially labeled data into click and whistle based species classification. Adequate training data is another major challenge of marine mammal passive acoustics. There are several problems wrapped up in this subject including the size of training data, availability of annotations, session effects in the data and variability inherent to the diversity of the marine mammals themselves. Off-axis click variability has often been cited as a major challenge for species classification from clicks. The weakly supervised training approach is a way to circumvent this limitation.

In this procedure we can learn, from the data, which clicks are most discriminative for species prediction. Experimental results showed up to 15% improvement in species classification (as measured by cross-species error) when using this procedure. Weakly supervised training did not benefit whistle performance in the same way but it was observed to improve session effects. A naive update of only the noise bases reduced cross species error by 30% and macro-error by 15%.

Over the course of this work, other minor contributions were observed and are noted here to highlight the additional benefits beyond the three major contributions.

It is very rare to have ideal noise removal in a passive acoustic marine mammal detection/classification task. As a consequence it is difficult to have clean recordings of the target vocalizations. We explored a method of learning NMF bases for the whistle patterns, even in the presence of residual noise. The first step was learning a set of noise bases from the noise labeled samples. When learning the whistle bases, we co-learned them while keeping the noise bases fixed. When extracting weights, the training algorithm could use the noise bases for the noise-like components of the audio, leaving the whistle bases to learn just

the remaining whistle like sounds. Observationally, whistle bases visually appeared more whistle like and quantitatively performance was nearly identical. Algorithm convergence was more successful with joint bases, and they were used in the majority of experiments.

The bases constructed for whistle detection and species classification required different methodologies to be effective. For whistle detection, co-occurrence constraints and species-specific NMF bases reduced detection performance. For species classification, species specific bases were concatenated into a common feature set for classification. Co-occurrence constraints also showed species classification improvements when training and evaluation data were mismatched, but the main effect seemed to be related to constraints on the noise weights. Thus, it appears that some whistle bases are useful across species.

Experimental results in both whistle detection and species classification suggested that the higher order models, 10 and 15 GMM mixtures and 20/60 bases, though occasionally more successful, were potentially overfitting since lower order models were more effective in mismatched settings.

## **7.2 Future Work**

Many areas of this thesis research could be expanded and improved. Specifically, I highlighted NMF computation, system design, and weak supervision.

### *7.2.1 NMF Design*

Having an accurate and representative set of NMF bases is crucial for NMF based classification. Because NMF is a very data dependent technique, it is essential to use large volumes of training data.

Lastly, it would be useful to assess how the current system would perform with data from other dolphin species? Can we develop a method of combining bases trained from different datasets, to contribute to a parallelization algorithm to improve training speed and make training from larger datasets more practical? To compensate for a common lack of large volumes of annotated training data, would performance be improved by training general whistle-like bases from an agglomeration of species data then perform an adaptation of the bases from a smaller species specific dataset?

Other future work would perform additional experiments to evaluate the classification improvements gained from modifications to bases training, such as sparsity constraints as those in Eggert and Koerner [23].

### 7.2.2 *System Design*

The block diagram of the overall system, Figure 3.2, has several modules that were implemented with a basic algorithm in order to complete the overall system. Future iterations could improve system performance by replacing those with more advanced algorithms; specifically noise removal, click detection, whistle detection and system combination.

It would be a significant benefit to the academic community to create a large standardized dataset that controlled for differences in population, group size, behavior state, recording hardware, temporal variability, and balanced across species. This would be fundamental to drawing more definitive conclusions about method variations.

Current noise removal processing uses standard median smoothing techniques for estimation of the background. This approach is good for slowly varying noise, however transient sources of noise can be problematic for many marine acoustic situations. There is significant potential to use NMF for transient noise removal, particularly for a known noise source with a large amount of training examples, such as seismic airguns.

Click detection was based on a simple heuristic. While our pilot results suggest this was more effective than waveform methods such as Teager-Kaiser energy [42, 89, 43], more principled methods [26] could be tested. Moving towards a multi-band frequency decomposition may also be worth exploring.

Whistle detection is performed to identify hypothesized segments to be input to the species classifier, rather than performing species prediction on all incoming audio. Silbido is currently used to perform this detection and is configured to run in batches off-line, outputting time point annotations. The species classification module uses the identified whistles to make species predictions. Though Silbido is considered a state-of-the-art algorithm, it could be more efficient to use NMF for both stages, particularly if with additional system tuning, both detection and species classification models used the same bases. Current whis-

tle detection rates between NMF and Silbido are comparable when low overlap is required but NMF performs better for high overlap rates. On the other hand, the complementary methods used in NMF and Silbido may lead to a more robust solution.

Tests showed that species-dependent NMF based whistle detection performed worse than a species-independent model. It would be interesting to re evaluate that with equal volumes of data or simply larger volumes of data, could reverse that conclusion. NMF was also observed to have significantly higher variance across folds. If the proposed algorithmic improvements to NMF can stabilize that variability, it may outperform Silbido.

As designed the classification system presented has two parallel processing streams for the click and whistle based species classification. A future system could combine these two results into a unified prediction based on the features and predictions of the two models. There are a variety of approaches that could be used to learn this system combination. Click and whistle features could be combined and used to train a third species classification model based on both features. Another approach would be to compute the posterior values of the click and whistle features for their respective species models then use these in a classifier to make a combined prediction.

### 7.2.3 *Weak Supervision*

A major contribution of this thesis is the application of weak supervision to marine mammal detection and classification. We demonstrate significant classification improvement for species classification from clicks which we hypothesize could be mitigating the effects of off-axis click variability. Future work could directly assess this possibility. If click trains could be recorded with note of the animal's orientation to the hydrophone, processing that data with the weakly supervised algorithm would confirm if the rejected clicks are disproportionately from off-axis.

Another direction of future work related to weak supervision would be the method of inclusion of new examples. It is noted in Rosenberg et al. [76] that for self-training, the genre of method we use, the selection metric for incorporating unlabeled data as training examples is crucial. Rosenberg recommends that such a selection method should be based

on a distance measure defined independently of the detection model. Their research showed this to be more effective than using the posterior probability of a sample belonging to the classification model, as we used in this research. They theorize the reason for improvement is that an independent distance measure causes “orthogonal” failure modes will be chosen and will be less likely to reinforce incorrectly selected examples from the unlabeled data.

An alternate approach to preventing poor or incorrect examples from being incorporated is to add an explicit method of “unlearning” [103]. These techniques take several forms but the principal is that if error is observed to increase as compared to prior iterations, to remove some of the most recently added examples. In our current formulation of weakly supervised learning, we do not check the performance of the  $p$  to the  $p - 1$  model we simply use the  $p$  model to compute posterior probability and infer confidence. In our method, samples have the flexibility to be added and removed as the model performance evolves, however it does not include an explicit performance check as performed in “unlearning”. Future revisions could compare  $p$  with  $p - 1$  to explicitly confirm improvement as compared to a prior model. If a performance decrease was detected, the  $\tau_c$  and  $\tau_n$  could be adjusted to compensate. A key result from the species classification from clicks was that best performance was achieved when being generous in labeling an example as noise (low  $\tau_n$ ) but conservative in including a click in species training (high  $\tau_c$ ). Additional future variants could verify performance gains using a combination of prior models.

Though shuffle training is an effective technique, for situations where there is temporal drift in the data, active online learning [17] has been shown to be more effective in such data streams. This is likely to be a realistic need for long term passive acoustic recording environments.

## BIBLIOGRAPHY

- [1] Ted A Abbot, Vincent E Premus, and Philip A Abbot. A real-time method for autonomous passive acoustic detection-classification of humpback whales. *The Journal of the Acoustical Society of America*, 127(5):2894–903, may 2010.
- [2] Whitlow W. L. Au, Brian Branstetter, Patrick W. Moore, and James J. Finneran. The biosonar field around an Atlantic bottlenose dolphin (*Tursiops truncatus*). *The Journal of the Acoustical Society of America*, 131(January):569, 2012.
- [3] Whitlow W L Au, R W Floyd, and R H Penner. Measurement of echolocation signals of the Atlantic bottlenose dolphin, *Tursiops truncatus* Montagu, in open waters. *J. Acoust. Soc. Am.*, 56(4):1280–1290, 1974.
- [4] Whitlow WL Au, Michael Richlen, and Marc O Lammers. Soundscape of a nearshore coral reef near an urban center. In *The Effects of Noise on Aquatic Life*, pages 345–351. Springer, 2012.
- [5] Yvonne Barkley, Julie N. Oswald, James V. Carretta, Shannon Rankin, Alexis Rudd, and Marc O Lammers. Comparison of real-time and post-cruise acoustic species identification of dolphin whistles using ROCCA (Real-time odontocete call classification algorithm). Technical report, NOAA National Marine Fisheries, Southwest Fisheries Science Center, 2011.
- [6] Jay Barlow and Karin A Forney. Abundance and population density of cetaceans in the California Current ecosystem. *Fishery Bulletin*, 105:509–526, 2007.
- [7] Simone Baumann-Pickering, Sean M Wiggins, John a Hildebrand, Marie a Roch, and Hans-Ulrich Schnitzler. Discriminating features of echolocation clicks of melon-headed whales (*Peponocephala electra*), bottlenose dolphins (*Tursiops truncatus*), and Gray’s spinner dolphins (*Stenella longirostris longirostris*). *The Journal of the Acoustical Society of America*, 128(4):2212–24, oct 2010.
- [8] Mark F Baumgartner and Sarah E Mussoline. A generalized baleen whale call detection and classification system. *The Journal of the Acoustical Society of America*, 129(5):2889–902, may 2011.
- [9] Michael Bittle and Alec Duncan. A review of current marine mammal detection and classification algorithms for use in automated passive acoustic monitoring. *Proceedings of Acoustics: Science, Technology and Amenity*, (November):1–8, 2013.

- [10] Matthew Blaschko, Andrea Vedaldi, and Andrew Zisserman. Simultaneous Object Detection and Ranking with Weak Supervision. *Advances in Neural Information Processing Systems*, pages 1–9, 2010.
- [11] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory - COLT' 98*, pages 92–100, New York, New York, USA, 1998. ACM Press.
- [12] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2):113–120, 1979.
- [13] Judith C Brown, Paris Smaragdis, and Anna Nousek-McGregor. Automatic identification of individual killer whales. *The Journal of the Acoustical Society of America*, 128(3):EL93–8, sep 2010.
- [14] J R Buck and P L Tyack. A quantitative measure of similarity for tursiops truncatus signature whistles. *The Journal of the Acoustical Society of America*, 94(5):2497–2506, 1993.
- [15] J.P. Campbell Jr. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997.
- [16] D.G. Childers, D.P. Skinner, and R.C. Kemerait. The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, 1977.
- [17] Wei Chu, Martin Zinkevich, Lihong Li, Achint Thomas, and Belle Tseng. Unbiased online active learning in data streams. *17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 195–203, 2011.
- [18] Ted W. Cranford, Wesley R. Elsberry, William G. Van Bonn, Jennifer a. Jeffress, Monica S. Chaplin, Diane J. Blackwood, Donald a. Carder, Tricia Kamolnick, Mark a. Todd, and Sam H. Ridgway. Observation and analysis of sonar signal generation in the bottlenose dolphin (*Tursiops truncatus*): Evidence for two sonar sources. *Journal of Experimental Marine Biology and Ecology*, 407(1):81–96, oct 2011.
- [19] Peter H. Dahl, James H. Miller, Douglas H. Cato, and Rex K. Andrew. Underwater Ambient Noise. *Acoustics Today*, 3:23, 2007.
- [20] Volker B Deecke, Lance G Barrett-Lennard, Paul Spong, and John K B Ford. The structure of stereotyped calls reflects kinship and social affiliation in resident killer whales (*Orcinus orca*). *Die Naturwissenschaften*, 97(5):513–8, may 2010.
- [21] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 39(1):1–38, 1977.

- [22] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6314 LNCS(PART 4):452–466, 2010.
- [23] J Eggert and E Korner. Sparse coding and NMF. *Neural Networks, 2004. Proceedings. 2004 . . .*, 2(4):2529–2533, 2004.
- [24] Christine Erbe and Andrew R King. Automatic detection of marine mammals using information entropy. *Journal of the Acoustical Society of America*, 124(5):2833–2840, 2008.
- [25] Ida G Eskesen, Magnus Wahlberg, Malene Simon, and Ole Næsbye Larsen. Comparison of echolocation clicks from geographically sympatric killer whales and long-finned pilot whales (1). *The Journal of the Acoustical Society of America*, 130(1):9–12, 2011.
- [26] Jing Fang and Les E. Atlas. Quadratic Detectors for Energy Estimation. *IEEE Transactions on Signal Processing*, 43(11):2582–2594, 1995.
- [27] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from Google’s image search. *Tenth IEEE International Conference on Computer Vision ICCV05 Volume 1*, 2(13):1816–1823, 2005.
- [28] Cedric Fevotte. Majorization-Minimization Algorithm for Smooth Itakura-Saito Non-negative Matrix Factorization. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1980–1983, 2011.
- [29] OA Filatova, ID Fedutin, MM Nagaylik, AM Burdin, and E Hoyt. Usage of monophonic and biphonic calls by free-ranging resident killer whales (*orcinus orca*) in kamchatka, russian far east. *Acta ethologica*, 12(1):37–44, 2009.
- [30] John KB Ford and H Dean Fisher. Group-specific dialects of killer whales (*Orcinus orca*) in British Columbia. In *Communication and behavior of whales*, volume 76, pages 129–161. Westview Press Boulder, CO, 1983.
- [31] John Kenneth Baker Ford. *Call Traditions and Dialects of Killer Whales (Orcinus orca) in British Columbia*. PhD thesis, University of British Columbia, 1984.
- [32] Alexandre Gannier, Sandra Fuchs, Paméla Quèbre, and Julie N. Oswald. Performance of a contour-based classification method for whistles of Mediterranean delphinids. *Applied Acoustics*, 71(11):1063–1069, nov 2010.
- [33] O. Gerard, C. Carthel, S. Coraluppi, and P. Willett. Feature-Aided Tracking for Marine Mammal Detection and Classification. *Canadian Acoustics*, 36(1):13–19, 2008.

- [34] D Gillespie, D.K. Mellinger, J. Gordon, D. McLaren, P Redmond, R. McHugh, P Trinder, XY Deng, and A Thode. PAMGUARD: Semiautomated, open source software for real-time acoustic detection and localization of cetaceans. *Proceedings of the Institute of Acoustics*, 30(5), 2008.
- [35] Douglas Gillespie. Detection and Classification of Right Whale Calls Using an Edge Detector Operating on a Smoothed Spectrogram. *Canadian Acoustics*, 32(2):39–47, 2004.
- [36] Douglas Gillespie and Marjolaine Caillat. Statistical classification of odontocete clicks. *Canadian Acoustics*, 36(1):20–26, 2008.
- [37] Melania Guerra, Aaron M Thode, Susanna B Blackwell, and A Michael Macrander. Quantifying seismic survey reverberation off the alaskan north slope. *The Journal of the Acoustical Society of America*, 130(5):3046–3058, 2011.
- [38] X Halkias and D Ellis. Call detection and extraction using Bayesian inference. *Applied Acoustics*, 67(11-12):1164–1174, 2006.
- [39] EE Henderson, JA Hildebrand, MH Smith, and EA Falcone. The behavioral context of common dolphin (*delphinus* sp.) vocalizations. *Marine Mammal Science*, 28(3):439–460, 2012.
- [40] A Rus Hoelzel, editor. *Marine Mammal Biology: An evolutionary approach*. John Wiley & Sons, 2002.
- [41] Marla M Holt, Dawn P Noren, and Candice K Emmons. An investigation of sound use and behavior in a killer whale (*orcinus orca*) population to inform passive acoustic monitoring studies. *Marine Mammal Science*, 29(2):E193–E202, 2013.
- [42] James F Kaiser. On a Simple Algorithm to Calculate the Energy of a Signal. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 381–384. IEEE, 1990.
- [43] V Kandia and Y Stylianou. Detection of sperm whale clicks based on the TeagerKaiser energy operator. *Applied Acoustics*, 67(11-12):1144–1163, nov 2006.
- [44] Brian E.D. Kingsbury, Nelson Morgan, and Steven Greenberg. Robust speech recognition using the modulation spectrogram. *Speech Communication*, 25(1-3):117–132, aug 1998.
- [45] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

- [46] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13(1), 2001.
- [47] Jimmy Lin and Alek Kolcz. Large-scale machine learning at twitter. *Proceedings of the 2012 international conference on Management of Data SIGMOD 12*, pages 793–804, 2012.
- [48] Yang Lu, David Mellinger, and Holger Klinck. Joint classification of whistles and echolocation clicks from odontocetes. In *Proceedings of Meetings on Acoustics*, volume 19, Montreal, 2013.
- [49] P. T. Madsen. Echolocation clicks of two free-ranging, oceanic delphinids with different food preferences: false killer whales *Pseudorca crassidens* and Risso’s dolphins *Grampus griseus*. *Journal of Experimental Biology*, 207(11):1811–1823, 2004.
- [50] P T Madsen, F H Jensen, D Carder, and S Ridgway. Dolphin whistles: a functional misnomer revealed by heliox breathing. *Biology letters*, (July):7–9, sep 2011.
- [51] P T Madsen, I Kerr, and R Payne. Source parameter estimates of echolocation clicks from wild pygmy killer whales (*Feresa attenuata*). *The Journal of the Acoustical Society of America*, 116(4 Pt 1):1909–1912, 2004.
- [52] P. T. Madsen, M. Lammers, D. Wisniewska, and K. Beedholm. Nasal sound production in echolocating delphinids (*Tursiops truncatus* and *Pseudorca crassidens*) is dynamic, but unilateral: clicking on the right side and whistling on the left side. *Journal of Experimental Biology*, 216(21):4091–4102, oct 2013.
- [53] P T Madsen, R Payne, N U Kristiansen, M Wahlberg, I Kerr, and B Møhl. Sperm whale sound production studied with ultrasound time / depth-recording tags. *Journal of Experimental Marine Biology and Ecology*, 205:1899–1906, 2002.
- [54] David K Mellinger and Christopher W Clark. Recognizing transient low-frequency whale sounds by spectrogram correlation. *The Journal of the Acoustical Society of America*, 107(6):3518–3529, 2000.
- [55] D.K. Mellinger and C.W. Clark. A method for filtering bioacoustic transients by spectrogram image convolution. *Proceedings of OCEANS '93*, pages III122–III127, 1993.
- [56] Bashar Mohammad and Ronald Mchugh. Automatic Detection and Characterization of Dispersive North Atlantic Right Whale Upcalls Recorded in a Shallow-Water Environment Using a Region-Based Active Contour Model. *IEEE Journal of Oceanic Engineering*, 36(3):431–440, 2011.

- [57] David Moretti and N DiMarzio. Overview of the 3 RD International workshop on the detection and classification of Marine Mammals using passive acoustics. *Canadian Acoustics*, 36(1):7–11, 2008.
- [58] R Morrissey, J Ward, N Dimarzio, S Jarvis, and D Moretti. Passive acoustic detection and localization of sperm whales (*Physeter macrocephalus*) in the tongue of the ocean. *Applied Acoustics*, 67(11-12):1091–1105, nov 2006.
- [59] Xavier Mouy, Mohammed Bahoura, and Yvan Simard. Automatic recognition of fin and blue whale calls for real-time monitoring in the St. Lawrence. *The Journal of the Acoustical Society of America*, 126(6):2918–28, dec 2009.
- [60] Gautham J Mysore, Paris Smaragdis, and Bhiksha Raj. Non-negative Hidden Markov Modeling of Audio with Application to Source Separation. pages 1–8.
- [61] Minh Hoai Nguyen, Lorenzo Torresani, Fernando de la Torre, and Carsten Rother. Weakly supervised discriminative localization and classification: a joint learning process. *IEEE 12th International Conference on Computer Vision*, (Iccv):1925–1932, sep 2009.
- [62] Nicole Nichols and Mari Ostendorf. Weakly Supervised Click Models for Odontocete Species Classification. *OCEANS 2014 IEEE Taipei*, 2014.
- [63] Richard O Nielsen. *Sonar signal processing*. Artech House, Inc., 1991.
- [64] Kamal Nigam, AK McCallum, S Thrun, and T Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2-3):103–134, 2000.
- [65] Julie N Oswald, Jay Barlow, and Thomas F Norris. Acoustic Identification of Nine Delphinid Species in the Eastern Tropical Pacific Ocean. *Marine Mammal Science*, 19(January):20–37, 2003.
- [66] Julie N Oswald, Thomas F Norris, and Renata S Sousa-Lima. A review of computer-based methods for the automated detection, extraction, and classification of marine mammal sounds. *Journal of the Acoustical Society of America*, 128(4):2438, 2010.
- [67] Pablo Peso Parada and Antonio Cardenal-López. Using Gaussian mixture models to detect and classify dolphin whistles and pulses. *The Journal of the Acoustical Society of America*, 135(6):3371–80, jun 2014.
- [68] M D Plumbley, A Cichocki, and R Bro. Non-negative mixtures. In *Handbook of Blind Source Separation*, pages 515–547. 2010.

- [69] A. N. Popper. Sound emission and detection by delphinids. In L. M. Herman, editor, *Cetacean Behavior: Mechanisms and Functions*. Krieger, Malabar, FL, 1980.
- [70] L. E. Rendell, J. N. Matthews, A. Gill, J. C. D. Gordon, and D. W. Macdonald. Quantitative analysis of tonal calls from five odontocete species, examining interspecific and intraspecific variation. *Journal of Zoology*, 249(4):403–410, dec 1999.
- [71] Marie A Roch, T Scott Brandes, Bhavesh Patel, Yvonne Barkley, Simone Baumann-Pickering, and Melissa S Soldevilla. Automated extraction of odontocete whistle contours. *The Journal of the Acoustical Society of America*, 130(4):2212–23, oct 2011.
- [72] Marie A Roch, Holger Klinck, Simone Baumann-Pickering, David K Mellinger, Simon Qui, Melissa S Soldevilla, and John a Hildebrand. Classification of echolocation clicks from odontocetes in the Southern California Bight. *The Journal of the Acoustical Society of America*, 129(1):467–75, jan 2011.
- [73] Marie A Roch, Melissa S Soldevilla, Jessica C Burtenshaw, E Elizabeth Henderson, and John A Hildebrand. Gaussian mixture model classification of odontocetes in the Southern California Bight and the Gulf of California. *The Journal of the Acoustical Society of America*, 121(3):1737–1748, 2007.
- [74] Marie A Roch, Melissa S Soldevilla, Rhonda Hoenigman, Sean M Wiggins, and John A Hildebrand. Comparison of machine learning techniques for the classification of echolocation clicks from three species of odontocetes. *Canadian Acoustics*, 36(1):47–47, 2008.
- [75] Marie A Roch, Johanna Stinner-sloan, Simone Baumann-pickering, and Sean M Wiggins. Compensating for the effects of site and equipment variation on delphinid species identification from their echolocation clicks. *The Journal of the Acoustical Society of America*, 7720(2008), 2015.
- [76] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. In *Proceedings - Seventh IEEE Workshop on Applications of Computer Vision, WACV 2005*, pages 29–36, 2005.
- [77] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.
- [78] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M. von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin. Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls. *The Journal of the Acoustical Society of America*, 135(2):953–962, feb 2014.
- [79] Ari Daniel Shapiro and Chao Wang. A versatile pitch tracking algorithm: from human speech to killer whale vocalizations. *The Journal of the Acoustical Society of America*, 126(1):451–9, jul 2009.

- [80] Paris Smaragdis. Convolutional speech bases and their application to supervised speech separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):1–12, 2007.
- [81] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, jul 2009.
- [82] Melissa S Soldevilla, E Elizabeth Henderson, Gregory S Campbell, Sean M Wiggins, John a Hildebrand, and Marie a Roch. Classification of Risso’s and Pacific white-sided dolphins using spectral properties of echolocation clicks. *The Journal of the Acoustical Society of America*, 124(1):609–24, jul 2008.
- [83] K M Stafford, C G Fox, and D S Clark. Long-range acoustic detection and localization of blue whale calls in the northeast Pacific Ocean. *The Journal of the Acoustical Society of America*, 104(6):3616–25, dec 1998.
- [84] W.W. Steiner. Species-specific differences in pure tonal whistle vocalizations of five western North Atlantic dolphin species. *Behavioral Ecology and Sociobiology*, 9(4):241–246, 1981.
- [85] Jayaraman J. Thiagarajan and Andreas Spanias. *Analysis of the MPEG-1 Layer III (MP3) Algorithm Using MATLAB*, volume 3. 2011.
- [86] F Thomsen, D Franck, and J K Ford. Characteristics of whistles from the acoustic repertoire of resident killer whales (*Orcinus orca*) off Vancouver Island, British Columbia. *J Acoust Soc Am*, 109(3):1240–1246, 2001.
- [87] Stephen K. Tjoa and K.J. Ray Liu. Multiplicative Update Rules for Nonnegative Matrix Factorization with Co-occurrence Constraints. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 449–452, 2010.
- [88] Ildar R Urazghildiiev and Christopher W Clark. Acoustic detection of North Atlantic right whale contact calls using spectrogram-based statistics. *The Journal of the Acoustical Society of America*, 122(2):769–76, aug 2007.
- [89] David Vakman. On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *Signal Processing, IEEE Transactions on*, 44(4):791–797, 1996.
- [90] Anne Villadsgaard, Magnus Wahlberg, and Jakob Tougaard. Echolocation signals of wild harbour porpoises, *Phocoena phocoena*. *The Journal of experimental biology*, 210:56–64, 2007.

- [91] Ravichander Vippera, Simon Bozonnet, Dong Wang, and Nicholas Evans. Robust speech recognition in multi-source noise environments using convolutive non-negative matrix factorization. In *CHiME: Workshop on Machine Learning in Multisource Environments*, pages 74–79, 2011.
- [92] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.
- [93] Magnus Wahlberg, Frants H. Jensen, Natacha Aguilar Soto, Kristian Beedholm, Lars Bejder, Claudia Oliveira, Marianne Rasmussen, Malene Simon, Anne Villadsgaard, and Peter T. Madsen. Source parameters of echolocation clicks from wild bottlenose dolphins (*Tursiops aduncus* and *Tursiops truncatus*). *The Journal of the Acoustical Society of America*, 130(4):2263, 2011.
- [94] Dong Wang, Ravichander Vippera, and Nicholas Evans. Online Pattern Learning for Non-Negative Convolutive Sparse Coding. In *INTERSPEECH*, Florence, Italy, 2011.
- [95] Dong Wang, Ravichander Vippera, Nicholas Evans, and TF Zheng. Online non-negative convolutive pattern learning for speech signals. *IEEE Transactions on Signal Processing*, 61(1):44–56, 2013.
- [96] Wenwu Wang. Convolutive non-negative sparse coding. In *Proceedings of the International Joint Conference on Neural Networks*, number 3, pages 3681–3684. Ieee, jun 2008.
- [97] William A Watkins. The harmonic interval: fact or artifact in spectral analysis of pulse trains. Technical report, Woods Hole Oceanographic Institution, 1968.
- [98] Gordon M. Wenz. Acoustic Ambient Noise in the Ocean: Spectra and Sources. *The Journal of the Acoustical Society of America*, 34(12):1936, 1962.
- [99] Rob Wijnhoven, Kris van Rens, Egbert Jaspers, and Peter de With. Online Learning for Ship Detection in Maritime Surveillance. *Thirty-first Symposium on Information Theory in the Benelux*, pages 73 – 80, 2010.
- [100] KW Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. Speech denoising using nonnegative matrix factorization with priors. *ICASSP*, 2008.
- [101] Tina M Yack, Jay Barlow, Marie A Roch, Holger Klinck, Steve Martin, David K Mellinger, and Douglas Gillespie. Comparison of beaked whale detection algorithms. *Applied Acoustics*, 71(11):1043–1049, 2010.

- [102] H Yurk, L Barrett-Lennard, J.K.B Ford, and C.O Matkin. Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern Alaska. *Animal Behaviour*, 63(6):1103–1119, jun 2002.
- [103] Xiaojin Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005.
- [104] Walter M. X. Zimmer, Mark P. Johnson, Peter T. Madsen, and Peter L. Tyack. Echolocation clicks of free-ranging Cuviers beaked whales (*Ziphius cavirostris*). *The Journal of the Acoustical Society of America*, 117(6):3919, 2005.

Appendix A  
**APPENDIX A**







































































\*/melon/palmyra092007FS192-070928-031000seg240000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg340000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg440000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg540000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg640000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg740000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg840000hz.wav  
\*/melon/palmyra092007FS192-070928-031000seg940000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1040000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1240000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1440000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1540000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1640000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1740000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1840000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg1940000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg2040000hz.wav  
\*/melon/palmyra092007FS192-070928-032000seg640000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg140000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg1140000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg1240000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg1640000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg1740000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg1840000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg1940000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg240000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg2040000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg440000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg540000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg640000hz.wav  
\*/melon/palmyra092007FS192-070928-035000seg740000hz.wav  
\*/melon/palmyra102006-061025-191000\_4seg140000hz.wav  
\*/melon/palmyra102006-061025-191000\_4seg240000hz.wav  
\*/melon/palmyra102006-061025-191000\_4seg340000hz.wav  
\*/melon/palmyra102006-061025-191000\_4seg440000hz.wav  
\*/melon/palmyra102006-061025-191000\_4seg540000hz.wav  
\*/melon/palmyra102006-061025-191000\_4seg640000hz.wav