ESTIMATING THE NUMBER OF MARINE MAMMALS USING RECORDINGS OF CLICKS FROM ONE MICROPHONE

Xanadu C. Halkias and Daniel P. W. Ellis

LabROSA, Columbia University Department of Electrical Engineering
{xanadu, dpwe}@ee.columbia.edu

ABSTRACT

An important but challenging task is to extract information on the location and/or number of marine mammals present given recordings from an array of hydrophones. Systems such as the Marine Mammal Monitoring on Navy Ranges (M3R) attempt to localize marine mammals as well as to get an estimate of their number using cross-correlation techniques on all available hydrophones. Our methodology offers the possibility to extract an estimate of the number of marine mammals given recordings from a single hydrophone, thus providing information to a researcher who does not have access to a larger array. The algorithm is based on three steps: detection of the clicks in the spectrogram using their energy, extraction of meaningful features, such as cepstral coefficients that are descriptive of the detected calls, and, lastly, choosing the appropriate number of clusters when using spectral clustering through the maximization of a given metric. The chosen number of clusters that best represents the data is an estimate of marine mammals present in the area. Informal analysis of the clustered clicks from example recordings shows that they are a good fit of the data, although a formal evaluation would require additional ground-truth. The algorithm was performed on several hydrophones in order to obtain some cross-validation of our results. Finally, the clusters were tracked in time using KL divergence. This algorithm could provide a first approximation on the number of vocalizing marine mammals using only one hydrophone.

1. INTRODUCTION

One of the most challenging tasks in marine mammal research is the localization of pods in a region were hydrophones have been employed. Several methodologies have been proposed that track and localize marine mammal pods, such as the Marine Mammal Monitoring on Navy Ranges [5].

These methodologies provide a fairly accurate estimate of the presence of marine mammals as has been confirmed by visual sightings. However, the majority of these techniques requires and depends on the use of multiple hydrophones.

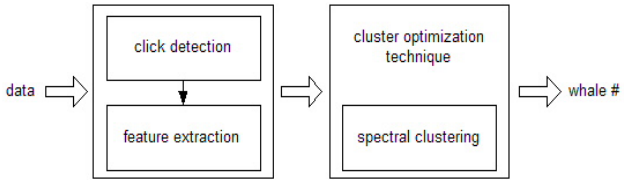

Figure 1: System Organization

In this work we approach the problem of localization by trying to provide an estimate of marine mammals in a region given recordings of clicks from a single microphone. The idea is to assist a field researcher who lacks the means to employ a larger array of hydrophones.

The paper is organized as follows: Section 2 provides the systems organization with detailed explanation of the implemented algorithm. Experimental results are shown in Section 3 and finally, Section 4 discusses the conclusions drawn and future work.

2. SYSTEM ORGANIZATION

Figure 1 provides a schematic description of the system. The basic algorithm is comprised of four steps:

1. Pre-processing of the data
2. Click detection
3. Extraction of meaningful features for the detected clicks
4. Extracting the optimal number of clusters using spectral clustering

Pre-processing of the data is obtained by high-pass filtering of the data in order to eliminate possible undesired low frequency noise such as waves or engine noise. Moreover, there is some additional denoising performed by standardizing the spectrogram along the time axis.

In order to detect the possible clicks in the recording we perform a simple peak detector based on the variance of the sum of the energy in all frequency channels across every time slice of the spectrogram as seen in Eq. 1. This procedure will yield a good approximation on the number of clicks present in the data as seen in Figure 2.

$$p[n] = \sum_{f} \left| X[f,n] \right| \Rightarrow clicks = p[n] \geq th \quad (1)$$

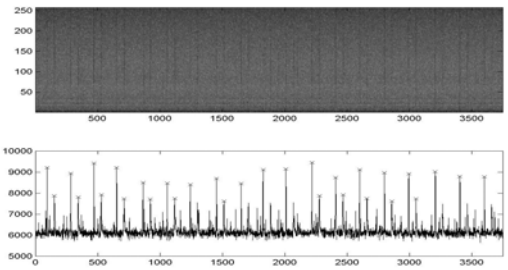

Figure 2: Click Detection

Where *X[f,n]* is the spectrogram of the data and *th* is an empirical chosen threshold that will yield the desired clicks.

For the detected clicks we extract three meaningful features: the first two cepstral coefficients, C_0 and C_1 and the slope of the onset of a click in a 20msec window [1]. The above can be seen in Eq. 2-3.

$$c_n = idft(log|dft(x[n])|) \tag{2}$$

$$s = \frac{1}{\sum_{n=t_0}^{t_0+t_w}(n-\bar{n})^2}\sum_{n=t_0}^{t_0+t_w} x^2[n](n-\bar{n}) \tag{3}$$

Where $\bar{n} = t_0 + t_w/2$ is the time average. The cepstral coefficients will give us the average energy of a click and the "skewness" respectively. The slope within the 20msec window will hopefully discriminate between reverberated clicks or not. Each detected click will be described by a triplet formed from the features mentioned above. The above can be seen in Fig. 3 validating the discriminative nature of the chosen features.

2.1 Spectral Clustering

We proceed by implementing spectral clustering as seen in [4]. Spectral clustering is an appealing and simple algorithm. It is based on building an affinity matrix (kernel), *A* using some known similarity metric. In this paper we formed A_{ij}, $i \neq j$ as the reciprocal of the Euclidean distance between the feature vectors and $A_{ii} = 0$.

We then define the diagonal matrix *D* whose *(i,i)*-element is formed by summing *A*'s *i*-th row and form the new matrix $L = D^{-1/2}AD^{-1/2}$.

We choose the dominant eigenvectors of matrix *L* and form a new matrix *X* by stacking those eigenvectors in columns. We normalize *X's* rows to have unit length.

Finally, we treat *X*'s rows as feature vectors and cluster them using a simple K-means [6].

2.2 Optimal K for K-means

Also, an algorithm based on cluster distortion is implemented as seen in [3] in order to extract the optimal number of clusters given the features. This number serves as

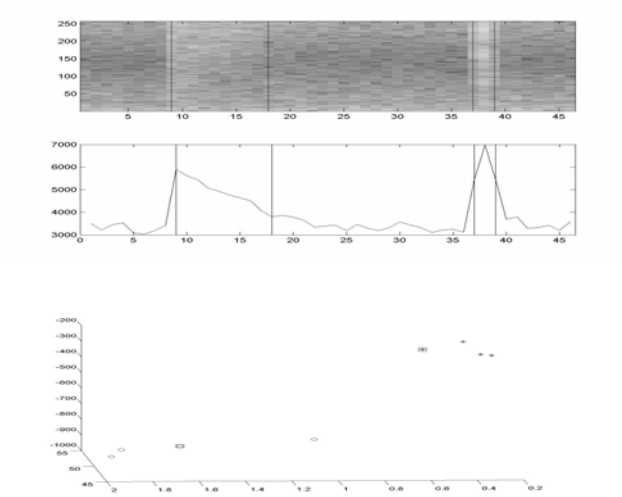

Figure 3: Sample clicks, with and without reverb, and their values in feature space (inside square) along with similar examples

a relative approximation on the number of marine mammals present in the area.

We define the distortion of the distortion I_j of a cluster *j* to be, Eq. 4a.

$$I_j = \sum_{t=1}^{N_j}\left[d\left(x_{jt}, w_j\right)\right]^2 \tag{4a}$$

Where w_j is the center of the cluster, N_j is the number of objects belonging to cluster *j*, x_{jt} is the *t*th object belonging to cluster *j*, and $d(x_{jt},w_j)$ is the distance between the object and the center of the cluster.

Each cluster is represented by its distortion and the impact it has on the entire data set is measured by its contribution to the sum of all distortions, S_K, Eq. 4b.

$$S_K = \sum_{j=1}^{K} I_K \tag{4b}$$

The optimal number of clusters is given by the function *f(K)* evaluated for different cluster numbers, *K*. The function is derived through Eqs. 5, 6.

$$f(K) = \begin{cases} 1 & if\ K = 1 \\ \dfrac{S_K}{\alpha_K S_{K-1}} & if\ S_{K-1} \neq 0, \forall K > 1 \\ 1 & if\ S_{K-1} = 0, \forall K > 1 \end{cases} \tag{5}$$

$$\alpha_K = \begin{cases} 1 - \dfrac{3}{4N_d} & if\ K = 2\ and\ N_d > 1 \\ \alpha_{K-1} + \dfrac{1-\alpha_{K-1}}{6} & if\ K > 2\ and\ N_d > 1 \end{cases} \tag{6}$$

Where S_K is the sum of cluster distortions for *K* clusters, N_d is the number of dimensions of the data and α_K is the weight factor.

The optimal number of clusters, *K* is chosen to be the one

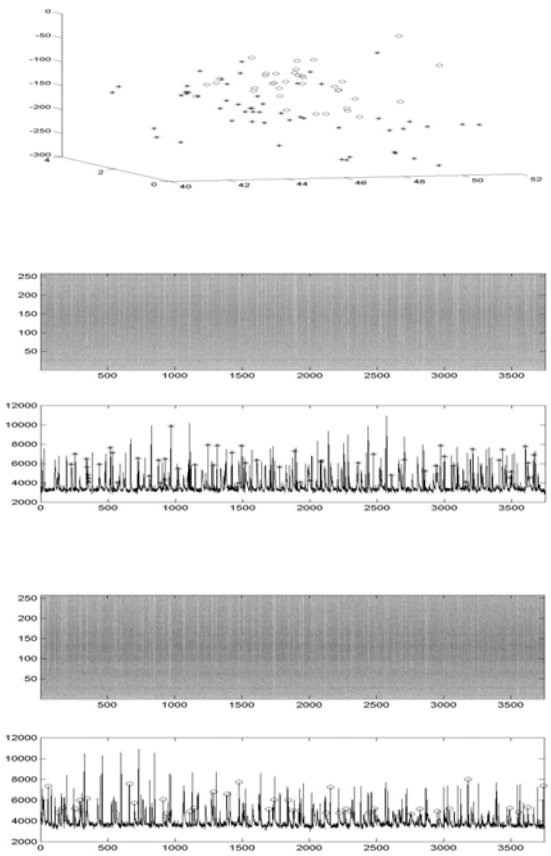

Figure 4: Sample results of KL divergence and clustering

that minimizes the function *f(K)*.

2.3 Tracking the clusters with KL-divergence

The clusters were tracked along time using the Kullback Leibler divergence (KL divergence). KL is a popular metric for comparing two distributions. In order to have a closed form we assume single gaussian mixtures for each cluster as seen in [2] and Eq. 4.

$$KL_N(\mu_p,\Sigma_p;\mu_q,\Sigma_q) = \log\frac{|\Sigma_q|}{|\Sigma_p|} + Tr\left(\Sigma_q^{-1}\Sigma_p\right) + \left(\mu_p-\mu_q\right)^{-1}\Sigma_q^{-1}\left(\mu_p-\mu_q\right) \quad (7)$$

Where $p(x)=\mathrm{N}(x;\mu_p,\Sigma_p)$ and $q(x)=\mathrm{N}(x;\mu_q,\Sigma_q)$

The number of clusters chosen was the average across each time frame as obtained from Tables 1-2 e.g. for the files 3M_ch4-6_35-40.wav we chose seven clusters and the tracking is performed on both sets of experiments. Figure 4 shows an example of how KL divergence provides a good measure for tracking the clusters along time. Two clusters of features belonging in subsequent time windows as well as their corresponding clicks in the spectrogram are shown (asterisk corresponds to the second row, open circle corresponds to the third row). These clusters were found to be the least divergent.

3. EXPERIMENTAL RESULTS

Two sets of experiments were performed. For each set, recordings from different hydrophones were also tested for cross-validation purposes.

The data set is comprised of clicks recorded with different hydrophones in a time frame of 20minutes. The file name as seen in Tables 1-2 is named according to some indicator followed by the channel, which corresponds to the different hydrophones and finally, it is followed by a time indicator of when the recording was made. Approximate ground truth is provided through visual observation as well as a comparison with the M3R system. Visual observations gave three sperm whales and an unknown number of pilot whales, while the M3R algorithm localized eleven whales. The M3R system provided the above results with the use of all hydrophones.

The idea behind the two sets is the inherent trade off between temporal resolution and cluster size. The first set of experiments is seen in Table 1. It is based on a small temporal resolution, where the system is implemented on 1min chunks of audio. For each 5min audio file we get 5 chunks. The average of those corresponds to the results in Table 1.

The second set is seen in Table 2. The algorithm is performed on 20sec windows of audio and the results are reported in the same way as mentioned above. This leads to empty clusters or small clusters that are dealt with by creating singleton clusters.

Audio file (wav)	# of clusters	Size per chunk (min)	Hydrophone
3M_ch4_35-40	5	1	A
3M_ch5_35-40	7	1	B
3M_ch6_35-40	9	1	C
3M_ch4_40-45	5	1	A
3M_ch5_40-45	7	1	B
3M_ch6_40-45	7	1	C
3M_ch4_45-50	5	1	A
3M_ch5_45-50	4	1	B
3M_ch6_45-50	8	1	C
3M_ch4_50-55	5	1	A
3M_ch5_50-55	4	1	B
3M_ch6_50-55	4	1	C

Table 1: Average results on 1 min chunks

Audio file (wav)	# of clusters	Size per chunk (sec)	Hydrophone
3M_ch4_35-40	4	20	A
3M_ch5_35-40	5	20	B
3M_ch6_35-40	4	20	C
3M_ch4_40-45	5	20	A
3M_ch5_40-45	5	20	B
3M_ch6_40-45	6	20	C
3M_ch4_45-50	4	20	A
3M_ch5_45-50	4	20	B
3M_ch6_45-50	4	20	C
3M_ch4_50-55	3	20	A
3M_ch5_50-55	4	20	B
3M_ch6_50-55	4	20	C

Table 2: Average results on 20 sec chunks

Attempts were made to label the data by the authors. However, without the expertise of an experienced marine biologist to label each individual click as belonging to a specific whale, the accurate ground truth is subjective rather than absolute. In order to compensate for the lack of absolute ground truth acoustic and visual analysis of the clustered clicks was employed. This analysis was performed on all sets. The analysis yielded that there is an approximate 30% overlap of the chosen clicks along time. This could be an indicator that the method is actually trying to decipher between individual marine mammals present.

3. CONCLUSIONS

As seen from Section 2 there are a few conclusions we can derive.

Firstly, if we consider the ground truth of the data, there are approximately eleven whales present in the area at the time of the recordings. Given that, it appears that the larger time windows, Table 1, capture a better approximation of the marine mammals present, since it gives us a number greater than five whales and an average across all hydrophones of seven whales. The differences between the hydrophones can be explained by the possible differences in the locations of the whales, which translate in weak features for clustering.

However, it is reassuring and a good indicator that the average numbers remain somewhat stable between the different times of the recordings e.g. Table 1 yields seven possible whales for recordings obtained at 35-40 min and six possible whales for recordings obtained at 40-45min.

Tracking of the clusters could provide some future information on the variability and trajectory of the whale movements since it provides an identifier of the clicks within a recording.

Finally, it would be interesting to proceed by trying to match the click sequences assigned to the extracted clusters with others within the recording thus extracting each whale's vocalization. This can be achieved with simple cross-correlation techniques. However, the lack of absolute ground truth does not allow us to proceed in such a direction since any results obtained would not be able to be classified as dismissive or accepted.

In this work we have provided a first step in extracting a gross approximation of the number of marine mammals based on recordings obtained from one microphone. The algorithm is comprised of small simple procedures and its usefulness is based on its ability to perform quick assessments for researchers who lack a larger on field system.

11. REFERENCES

[1] N. Lesser, D. Ellis (2005). "Clap Detection and Discrimination for Rhythm Therapy," *Proc. ICASSP-05*, Philadelphia, March 2005, pp. III-37-40. (4pp)

[2] M. Mandel, D. Ellis (2005). "Song-Level Features and Support Vector Machines for Music Classification," *Proc. Int. Conf. on Music Info. Retrieval ISMIR-05*, London, September 2005. (6pp)

[3] D. T Pham, S. S. Dimov and C. D. Nguyen, "Selection of K in K-means Clustering," *Proc. ImechE Vol. 219 Part C: J. Mechanical Engineering science*, pp. 103-119, September 2005.

[4] A. Y. Ng, M. I. Jordan, and Y. Weiss. In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), " On spectral clustering: Analysis and an algorithm," *Advances in Neural Information Processing Systems (NIPS) 14*, 2002.

[5] Moretti, D., et al., "Marine Mammal Monitoring on Navy Ranges (M3R)", *Proceedings of the Undersea Defense Technology* Hawaii 2001 Conference, October 2001.

[6] R. O. Duda, P. E. Hart and D. G. Stork, "Pattern Classification," *Wiley-Interscience,* pp. 526-528, 2001