Available online at www.sciencedirect.com
ScienceDirect
Applied Acoustics 67 (2006) 1164–1174
applied acoustics
www.elsevier.com/locate/apacoust

# Call detection and extraction using Bayesian inference

Xanadu C. Halkias *, Daniel P.W. Ellis

*LabROSA, Columbia University, Department of Electrical Engineering, 500 West 120th Street, 1300 S.W. Mudd, New York, NY 10027, United States*

Received 26 January 2006; received in revised form 2 April 2006; accepted 21 May 2006
Available online 28 July 2006

**Abstract**

Marine mammal vocalizations have always presented an intriguing topic for researchers not only because they provide an insight on their interaction, but also because they are a way for scientists to extract information on their location, number and various other parameters needed for their monitoring and tracking. In the past years field researchers have used submersible microphones to record underwater sounds in the hopes of being able to understand and label marine life. One of the emerging problems for both on site and off site researchers is the ability to detect and extract marine mammal vocalizations automatically and in real time given the copious amounts of existing recordings. In this paper, we focus on signal types that have a well-defined single frequency maxima and offer a method based on Sine wave modeling and Bayesian inference that will automatically detect and extract such possible vocalizations belonging to marine mammals while minimizing human interference. The procedure presented in this paper is based on global characteristics of these calls thus rendering it a species independent call detector/extractor.
© 2006 Elsevier Ltd. All rights reserved.

*Keywords:* Whistles; Call detection; Extraction; Bayesian inference; Sinewave modeling

## 1. Introduction

In recent years, new ideas and new questions have emerged as scientists equipped with innovative technological tools try to acquire a deeper understanding of inter- and

* Corresponding author. Tel.: +1 212 877 3467.
*E-mail addresses:* xanadu@ee.columbia.edu (X.C. Halkias), dpwe@ee.columbia.edu (D.P.W. Ellis).

0003-682X/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apacoust.2006.05.006

intra-species interactions. Marine mammals have become an important link in the history and comprehension of life.

In 1958, Watkins and Schevill made the first marine mammal audio recordings we know of [4]. Thus, started a long journey of collecting observations and information on marine mammals. For various reasons, certain species such as whales and dolphins became commercially popular and attracted enormous public interest.

In the mid-1950s, Kenneth S. Norris [3], with his first experiments on echolocation, together with the US navy acknowledged the possibility of intelligence within marine mammals. Further scientific exploration as well as the need to protect sea life led to more organized and joint collaborations of researchers from different fields.

In recent years, marine biologists as well as engineers have come together as teams and work on or off site collecting multimodal data such as audio, video, etc. The technological advances have allowed researchers to track, monitor and analyze marine mammals in loco, and in a simulated environment through the use of computers.

In terms of engineering, three major categories of underwater acoustic research have emerged: (i) call detection/extraction within recordings; (ii) localization; and (iii) echolocation. This paper proposes a method that belongs in the first category. Call detection/extraction is the first step needed in order to analyze recordings that include marine mammal vocalizations. However, the amount of collected data as well as the diversity of the species requires a robust technique that will alleviate manual handling and labeling.

Moreover, an automatic system that will detect and extract marine mammal vocalizations is needed for both localization – the ability to track groups of marine mammals – as well as echolocation (e.g. sonar).

This paper is organized as follows: Section 2 provides the overview of the system, which is comprised of two parts. Section 2.1 explains the front end of the system. Section 2.2 describes the back end of the system. In Section 3 experimental results are provided and finally Section 4 consists of concluding remarks as well as future work.

## 2. Overview of the automatic call detection/extraction system

Fig. 1 presents the schematic overview of the proposed system for automatic call detection and extraction. As seen in the figure, the system is composed of two subsystems: the front-end and the back-end.

Marine mammal vocalizations e.g. from dolphins and whales can be roughly classified into two main categories: (i) whistles and (ii) clicks. Whistles can be considered as calls that promote interaction and "communication" between marine mammals and other species, and have a sinusoidal or harmonic structure. Clicks on the other hand are used mostly

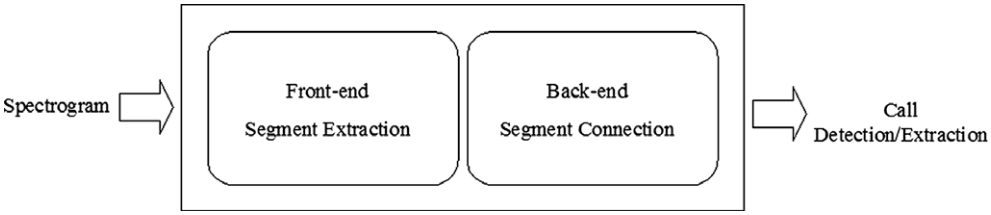


Fig. 1. System overview.

for underwater navigation. Our system can be utilized for the detection and extraction of whistle calls found in various marine mammal recordings. Our methodology utilizes sine-wave modeling and a probabilistic framework that is based on Bayesian inference [5].

In the general case, whistles can be viewed as frequency modulated signals that could be periodic in time and could also have a clear harmonic structure. A typical whistle signal could be described as:

$$y(t) = A(t) \cdot \sin(2\pi f(t)t) \tag{1}$$

where $A(t)$ is some kind of amplitude function and $f(t)$ is the frequency modulator.

The existing literature, such as Mellinger and Clark [1,2], deals with species specific calls of high signal-to-noise ratio (SNR) and with no overlaps between calls. Their approach is based on kernel matching/correlation that is sensitive to specified calls. This approach has difficulties with generalizing and capturing the diversity of calls that an in-the-field researcher would encounter.

Our method is based on a Bayesian probabilistic framework, which can be described as follows.

We are trying to find the most probable call structure, $C$ given our observations, recordings, $o$ as seen in Eq. (2) and as obtained from Bayes rule [5]

$$p(C|o) = p(o|C)p(C) \tag{2}$$

where $p(o|C)$ is the probability of the actual observations given the hypothesized call parameters that define the call structure, and $p(C)$ is the prior of that call.

The recordings used can be defined as the observation signal given by:

$$o(t) = y(t) + n(t) \tag{3}$$

$$p(o|y) = \aleph(o; y, \sigma_n) \tag{4}$$

where $o(t)$ is the observed signal, $y(t)$ is the ideal track waveform and $n(t)$ is the noise. This allows us to define the likelihood presented in Eq. (4), which indicates that the observation is normally distributed with mean equal to the underlying call, and variance resulting from the noise.

As mentioned in Eq. (2), $C$ is the call structure as defined by a set of global characteristics e.g. smoothness in energy, etc. and is further explored in Section 2.2. We can define $p(C)$ as the prior for these parameters described by a $D$-dimensional normal distribution given by:

$$p(C) = \aleph(\mu_D, \sigma_D) \tag{5}$$

We can assume that our ideal waveform $y(t)$ is randomly related to the call parameters $C$, since it is described by the two random variables $A(t)$, $f(t)$ as seen in Eq. (1), thus introducing $p(y|C)$ in our computations.

Combining these equations, we obtain the desired result as seen:

$$p(C|o) = \frac{p(o|y)p(y|C)p(C)}{p(o)} \tag{6}$$

where the maximum – likelihood inference of the call parameters $C$ comes from maximizing the numerator since the denominator $p(o)$ does not depend on $C$.

As mentioned, the system is comprised of two subsystems. Eqs. (3) and (4) describes the front-end of the system, where sinusoidal segments that are part of calls are extracted.

Eqs. (2)–(5) describe the back-end of the system where the segments extracted from the front-end need to be connected in order to extract the final calls. The decision concerning the possible connections is made according to some chosen distribution e.g. normal that describes certain characteristics of the calls e.g. smoothness in frequency and energy and in accordance to Eq. (6).

These two parts of the system are combined in Eq. (6) where $p(o|y)$ stands for sinewave modeling and $p(y|C)p(C)$ defines the formation of the call based on global characteristics.

### 2.1. The front end of the system: extracting sinusoidal fragments

Fig. 2 shows a detailed schematic of the front end of the system. The goal is to extract as many fragments of calls as possible while minimizing false positives – i.e. segments that are not part of the desired calls.

In order to achieve that, we perform a column wise scan of the spectrogram and keeping all the regional maxima. Regional maxima are defined as a set of connected points of constant value from which it is impossible to reach a point with a higher value without first descending.

However, this definition will yield a large number of false positive segments due to the appearance of non-uniform colored noise within the recordings that can be attributed to bad recording conditions e.g. vehicle engine, other marine life, weather conditions, hydrophone responses, etc.

In order to alleviate this problem a new hybrid regional maxima technique was employed. It is an iterative algorithm based on the variance of the spectrum at each time frame. Eq. (7) describes the methodology in a succinct way. For each time slice, $t$ of the spectrogram we perform the following as described in Eq. (7).

$$\left.\begin{aligned} &\text{init}: i = 1 \text{peak}_{t,i=1} \max(\text{Regmax}_{t,i=1}) = M_t \\ &\text{peak}_{t,i+1} = \max(\text{Regmax}_{t,i+1}), \quad \text{while } \vartheta > .35 \\ &\vartheta = 1 - \frac{\text{var}\left(\text{Regmax}_{t,i+1}\right)}{\text{var}\left(\text{Regmax}_{t,i}\right)}, \quad t = 1,2,\ldots, \quad i = 1,2,\ldots \\ &\text{Regmax}_{t,i+1} = \text{Regmax}_{t,i} \neq \text{peak}_{t,i} \end{aligned}\right\} \tag{7}$$

where $\text{peak}_{t,i}$ defines the extracted peaks from each time slice, $t$ at every iteration $i$. $\text{Regmax}_{t,i}$ defines the regional maxima of that time slice $t$ at each iteration $i$. $\text{Regmax}_{t,i+1}$ is the regional maxima of time slice $t$ and iteration $i$ excluding the extracted peaks at time slice $t$ and iteration $i$. $\vartheta$ is the threshold that determines the number of peaks that are extracted from each time slice. It can be described as the drop of the variance at each iteration $I$ and changes for every iteration $i$ and time slice $t$. Finally, for the initialization of the algorithm

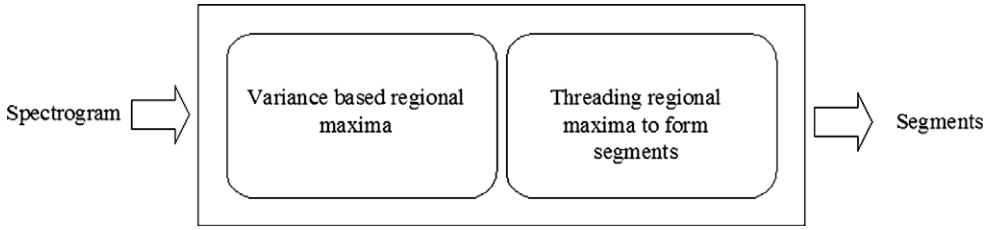


Fig. 2. Front end.

for $i = 1$ we extract the regional maxima of a time slice $t$ and the first peak extracted is the maximum $M_t$ of that regional maxima belonging to time slice $t$.

Empirical results have shown that the algorithm can extract the correct regional maxima for most recordings when $\vartheta > 0.35$. Basically, this method will separate the signal from the noisy background using the fact that the signal will appear as accentuated peaks in the spectrum of larger amplitude than that of the existing ambient noise. However, this assumption could lead to false positives given that other tonal signals can be considered noise e.g. boat like signals.

We then proceed by implementing a quadratic interpolation of the frequencies and amplitudes that correspond to the extracted hybrid regional maxima. This will provide a smoother effect to the extracted segments to capture the calls more accurately.

After extracting the regional maxima for a time frame of the spectrogram we need to decide whether they are connected to the ones extracted from the subsequent time frames. To do so we utilize two parameters: their frequency and their amplitude.

Each regional max lies within a frequency bin and in order to connect it with a subsequent regional max they cannot be separated by more than a predetermined frequency "distance", which we have set as five frequency bins. We perform an exhaustive search on the regional maxima and prefer the ones that have the smallest distance.

A second level of decision is also employed through the use of the ratio of the amplitudes, which cannot exceed a predefined amount. This level of decision will deal with possible frequency ties. In the case where we have a tie in this stage as well, the choice for continuation is made randomly.

The whole procedure is continued and a segment is considered extracted if there is no continuation found after a number of time steps, which we call dead steps. In our experiments we consider that a segment is over when there has been more than three dead steps.

Finally, we only consider as valid segments the ones that satisfy a minimum length parameter, thus keeping only the segments that have more than two points.

Fig. 3 presents an example of the final resulting segments of a sample recording. The lower left corner of the bottom left image shows two small segments that were extracted from a call that is present there.

### 2.2. The back-end: forming calls from segments

Fig. 4 provides a detailed description of the back-end of the system. The theory is given in Eq. (3) where the main idea is to connect the extracted segments from the front-end of the system according to some decision made through maximum likelihood.

After having extracted the segments using the methodology described in Section 2.1 we proceed into defining the problem as one that can be divided into two categories: intra-call discontinuities and inter-call discontinuities.

For the first category, we basically want to connect gaps that might appear within a segment. In order to accomplish that we define a time gap that signifies the maximum gap allowed within the segment. When such gaps are found we perform linear interpolation of the edges in both frequency and amplitude.

This procedure will ensure the extraction of the calls even when there is a significant difference of amplitude within a call, but it is clear that the segments are part of some larger call. The second category is the core of the whole system since it tries to connect a segment with the best choice from a set of fragments.

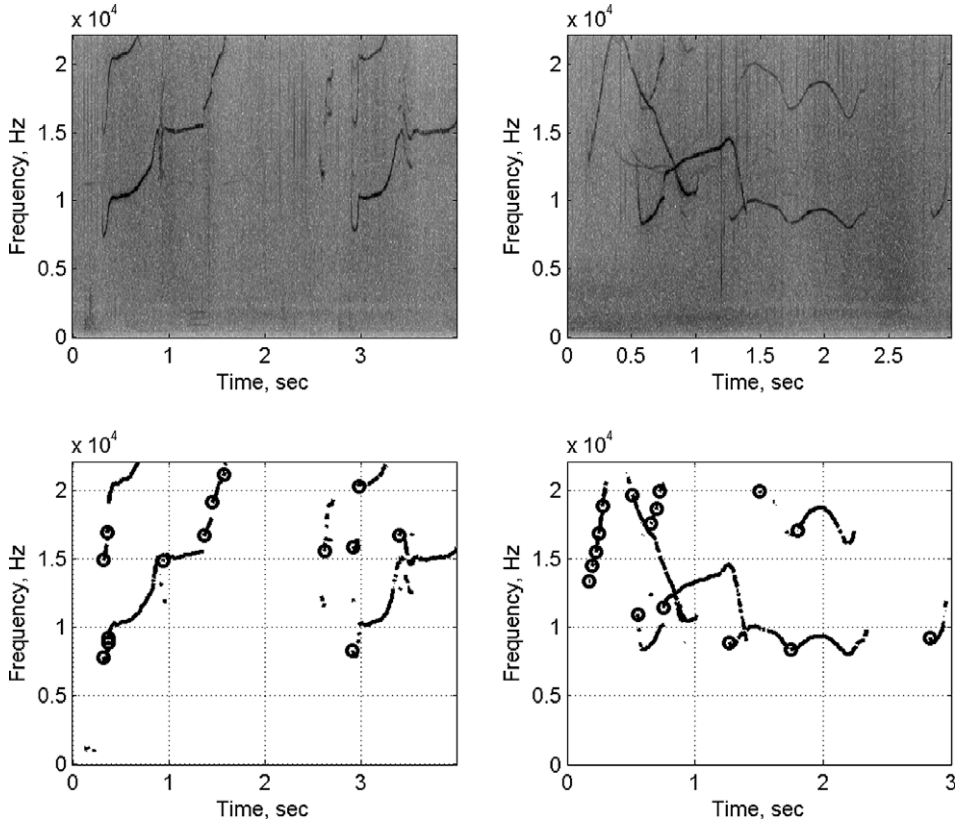


Fig. 3. Segment extraction through the front-end of the system. Top row: original spectrograms depicting multiple calls. Bottom row: extracted segments. Circles indicate sample segment edges for two different recordings.

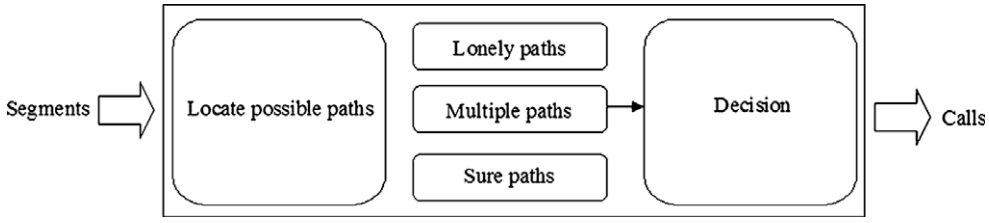


Fig. 4. Back-end.

We begin by defining a suitable search neighborhood for each segment. This neighborhood is adaptive ensuring that each segment has a neighbor. We utilize two parameters, frequency and time that indicate the final boundaries of the search area e.g. $\pm 50$ frequency bins, $\pm 10$ time steps. Note that each segment has two search areas, one for each tip. We proceed by placing the restriction that the segments cannot be overlapping in time, which implies temporal consistency.

In order to decrease the size of the search neighborhood even more we add a tip directionality criterion. For each tip of each segment within the search neighborhood we measure

its slope. From training data we have extracted normal distributions from instances of pairs of segments with an upward or a downward directionality that belong to the same track.

The slope of the tips is used to figure out the directionality likelihood for a pair of segments. We keep the segments that have the highest likelihood. This procedure can be seen more clearly in:

$$p(sl_{\text{up}}|\Theta_{\text{up}}) = N\left(\mu_{x,y}, \sum_{2\times 2}\right) \tag{8}$$

$$p(sl_{\text{down}}|\Theta_{\text{down}}) = N\left(\mu_{x,y}, \sum_{2\times 2}\right) \tag{9}$$

$$\text{if } p(sl|\Theta_{\text{up}}) > p(sl|\Theta_{\text{down}}) \Rightarrow \text{upward} \tag{10}$$

where $p(sl_{\text{up}}|\Theta_{\text{up}})$ describes the likelihood of a pair of segments belonging to the same call and have an upward directionality. The distribution is approximated through a 2D Gaussian whose parameters are the slopes of pairs of tips that belong to the same track. These values are obtained through training data. The same is applied for the downward directionality thus providing us with a smaller set of possible neighbors to a segment.

After having extracted all the possible paths using the methodology described above we can assume that there exist three types of paths:

1. Sure path: a segment that has one connection only.
2. Lonely path: a segment that has no connection.
3. Multiple paths: a segment that has multiple connections.

From the above definitions it is clear that the existence of multiple paths requires a way of deciding which ones form the actual calls. In order to make that decision we form a linear combination of scalar normal distributions from features that we have extracted from training data. The features we have chosen are the mean across the call of the curvature smoothness in frequency and slope smoothness in energy as seen in:

$$F_{\text{sm}} = \left|\text{mean}(\mathrm{d}^2 f/\mathrm{d}t^2)\right| \tag{11}$$

$$E_{\text{sm}} = \left|\text{mean}(\mathrm{d}e/\mathrm{d}t)\right| \tag{12}$$

Fitting a normal distribution on the features extracted from training data allows us to use maximum likelihood and perform a greedy search amongst the multiple paths and pick the one with the highest likelihood i.e. highest $P(F_{\text{sm}})$ and $P(E_{\text{sm}})$. The algorithm proceeds and only keeps the paths whose likelihood does not drop more than a specific percentage e.g. 10%. Once the connections are established the two segments are then merged according to their linearly interpolated value.

In the case where the criterion is not met then the paths are set aside and the same algorithm is employed on them in order to extract possible new connections.

## 3. Experimental results

Fig. 5 shows the results of the system in two sample cases. Table 1 provides the success rates of the algorithm based on manual labeling. In total we used 5 min of audio recordings, which amounts to 400 calls of moderate difficulty as seen in Figs. 3 and 5. All the

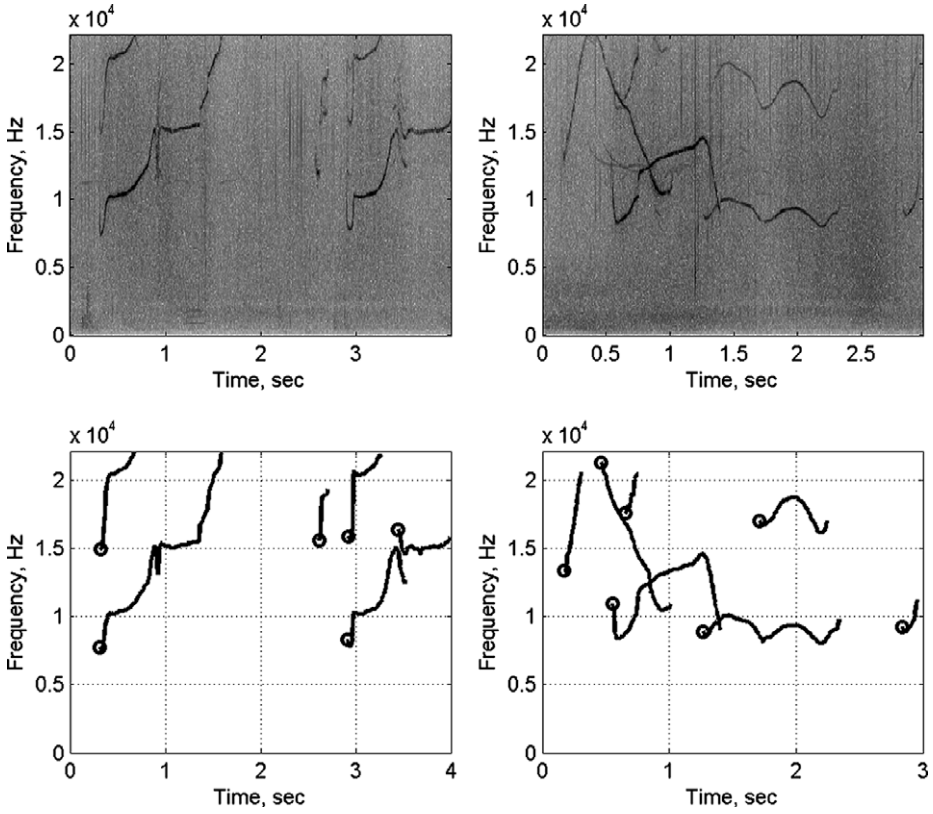


Fig. 5. Extracted calls. Top Row: original spectrograms depicting multiple calls. Bottom row: final extracted calls. Circles indicate beginning of calls.

Table 1
Results

| | |
|---|---|
| *Frame level* | |
| Extraction success rate | 82% |
| *Segment level* | |
| False positive rate | 5% |
| False negative rate | 3% |

recordings had a sampling frequency, $F_s$ of 44,100 Hz and were obtained from various recording hardware. The recordings were provided by Cornell University's Macaulay Library.

We utilized the STFT (spectrogram) for visualization and computation purposes. The parameters for the STFT are 512 point FFT thus obtaining a frequency resolution of approximately 86.13 Hz. Moreover, we used a 512 point Hanning window with 50% overlap, which yielded a time resolution of 11.6 ms.

In order to obtain the rates that are presented in Table 1 we had to approach the system's performance in two levels. This need stems from the fact that we are implementing a

two-stage system whose individual parts, as explained in Section 2 contribute equally to the final results obtained from our system. We can then say that the overall success of the extraction algorithm depends not only on how well the front-end performs, since if a segment is not extracted then that call will not be represented, but also if the segments that are connected correspond to existing calls.

In order to incorporate the above in the performance measure of the system, we provide an overall extraction success rate that is obtained on the frame level. Thus, for every track that is extracted we measure the number of points that it includes and compare that with the ground truth that is obtained through manual labeling. As seen in Table 1, the total extraction rate for the data was 82%.

We also provide false positive and false negative rates on the segment level by obtaining the count of non-correct and correct connections respectively, thus extract dependencies in the sinewave modeling. The false positive rate that indicates a connection where there should not be a connection is 5%. Also, the false negative rate, which indicates connections that should have been made, is at 3%. The false positive rate appears to be higher due to the fact that the system is created under the assumption that the extracted segments are more likely to be close to each other when belonging to the same call and thus their connection is desired. Finally, it is worth noting that the audio recordings come from different

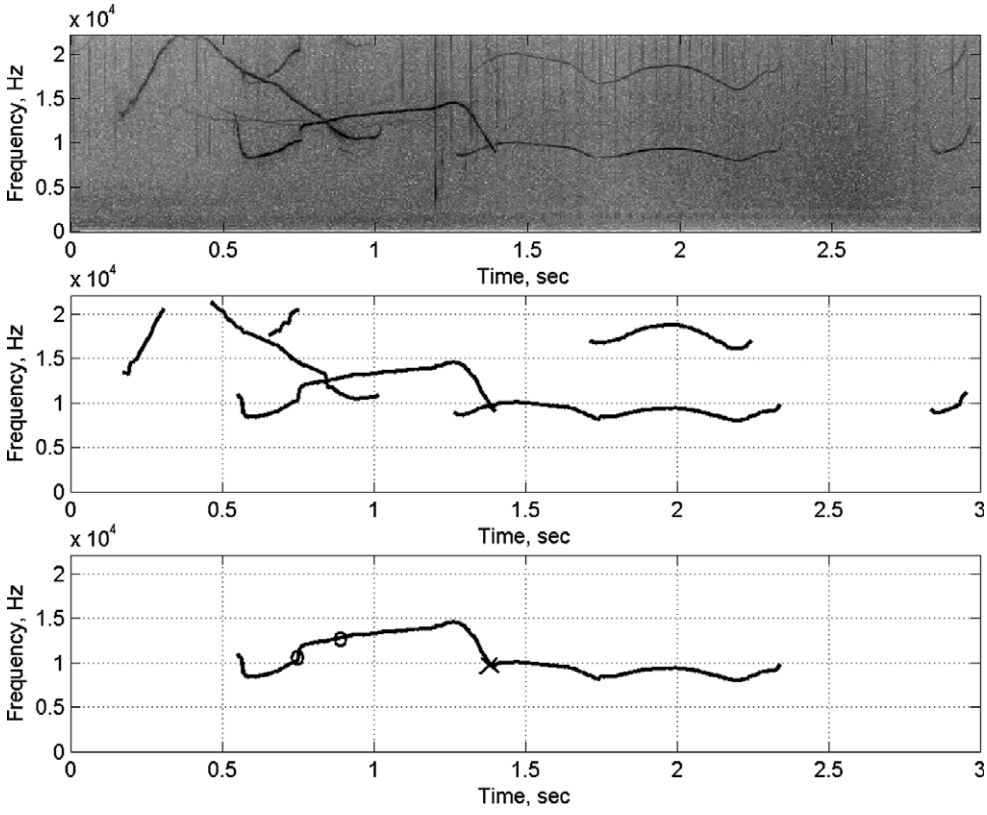


Fig. 6. Overlap example. Top row: original spectrogram depicting multiple calls. Middle row: extraction of calls. Bottom row: example calls. Circle indicates correct connection; × indicates false connection.

species of marine mammals and also that more than one species was present during the analysis. This is an important factor given that most existing experiments [1] are done on carefully selected recordings of one specific species with no overlaps of calls are present.

One of the innovations of our system is that it deals with one of the hardest tasks in call detection and extraction, which is the ability to handle overlapping calls.

Fig. 6 is an example of the system disambiguating one case of overlapping calls while failing to disambiguate another. The success can be attributed to the parameters used for deciding amongst multiple paths, while the failure of the system on this particular example is due to click interference on the intersection of the calls thus confusing the system into forming a connection between them. The aim of this work is to provide a general method of call extraction without altering the spectrogram, however future implementations will investigate automatic removal of click calls.

Finally, it is worth noting that the execution time on 4 s segments of moderate difficulty as shown in Figs. 3, 5, 6 is approximately 5 s on a 2.0 GHz Pentium 4M laptop PC with 512M of RAM. For recordings of higher difficulty and complexity the execution time increases and the overall success rate decreases.

## 4. Conclusions

In this paper, we presented a robust algorithm for detecting and extracting tonal calls e.g. dolphin whistles from marine mammals independent of species or number present in the recordings. Tonal calls have defined single frequency maxima and are of great importance since scientists have hypothesized that they are utilized for ''communication'' purposes.

The results are very promising and they indicate that a frame-by-frame extraction of regional maxima using a greedy algorithm is adequate for an in-the-field researcher in order to obtain a first analysis. An exhaustive search of the possible paths would yield much better results, but would lead to a computationally more expensive system. Furthermore, the ability of this method to provide a first step in disambiguating overlaps between calls is very promising. However, there are still a number of cases e.g. diverging paths from a single frequency point where unless there are significant excluding differences as obtained from the training data, the algorithm will choose randomly, thus increasing the possibility of a false positive or false negative.

The innovation of the algorithm lies in the probabilistic framework where a connection is decided according to the parameters expressed in Section 2.2.

Future developments could be to add more parameters that would describe the calls as a whole, thus rendering this methodology even more stable and minimizing the rates of false positives and/or negatives. We would also like to provide the option of automatically repairing interfering click calls. It would be interesting to create a one stage system where the decision is made at the frame level and compare the overall extraction rate with the system presented in this work.

Finally, an alternate approach would be to tune the algorithm by using species dependent parameters that could be extracted from training data e.g. bottlenose dolphins might have a distinguishable curvature in frequency, thus allowing us to model it through a specific distribution, which would yield a species-specific call detector/extractor. In a similar manner, this algorithm might enable the extraction of boat like sounds that are comprised of similar tonal signals.

In this work we managed to formalize a generalized call detector/extractor for marine mammals that requires minimal manual interaction and is diverse enough to be utilized as a species detector.

## References

[1] Mellinger DK, Clark CW. Methods for automatic detection of mysticete sounds. Mar Fresh Behav Physiol 1997;29:163–81.
[2] Mellinger DK, Clark CW. Recognizing transient low-frequency whale sounds by spectrogram correlation. JASA 2000;107(6).
[3] Norris KS. The echolocation of marine mammals. In: Andersen HT, editor. The biology of marine mammals. Academic Press; 1969. p. 425–75.
[4] Watkins WA, Schevill WE. Underwater paint marking of porpoises. Fish Bull 1976;74:687–9.
[5] Duda RO, Hart PE, Stork DG. Pattern classification. Wiley Interscience; 2001.