The Orchive : Data mining a massive bioacoustic archive

Steven Ness SNESS@UVIC.CA
Department of Computer Science, University of Victoria, Canada

Helena Symonds INFO@ORCALAB.ORG
OrcaLab, P.O. Box 510 Alert Bay, BC, Canada

Paul Spong INFO@ORCALAB.ORG
OrcaLab, P.O. Box 510 Alert Bay, BC, Canada

George Tzanetakis GTZAN@CS.UVIC.CA
Department of Computer Science, University of Victoria, Canada

Abstract

The Orchive is a large collection of over 20,000 hours of audio recordings from the OrcaLab research facility located off the northern tip of Vancouver Island. It contains recorded orca vocalizations from the 1980 to the present time and is one of the largest resources of bioacoustic data in the world. We have developed a web-based interface that allows researchers to listen to these recordings, view waveform and spectral representations of the audio, label clips with annotations, and view the results of machine learning classifiers based on automatic audio features extraction. In this paper we describe such classifiers that discriminate between background noise, orca calls, and the voice notes that are present in most of the tapes. Furthermore we show classification results for individual calls based on a previously existing orca call catalog. We have also experimentally investigated the scalability of classifiers over the entire Orchive.

1. Introduction

The Orchive is a large archive containing over 20,000 hours of recordings from the Orcalab research station. These recordings were made using a network of hydrophones and originally stored on analog cassette tapes. OrcaLab is a research station on Hanson Island which is located at the north part of Vancouver Island on the west coast of Canada. It has been in continuous operation since 1980. It was designed as a land based station in order to reduce the impact on the orcas under study, as the noise and disturbance from boats affects the orcas in observable but currently unquantified ways. In collaboration with OrcaLab, we have digitized the tapes and have made these recordings available to the scientific community through the Orchive website (http://orchive.cs.uvic.ca).

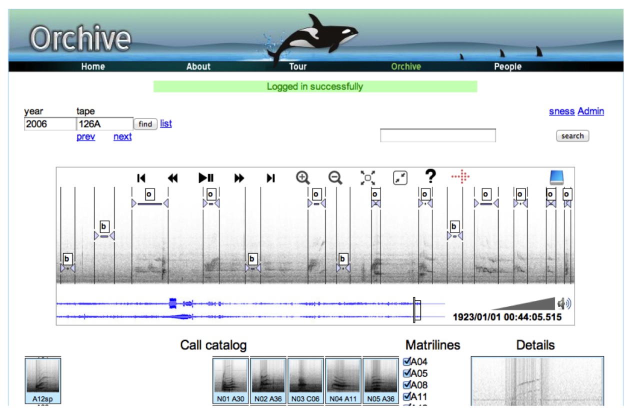

Figure 1. Annotated audio from from the Orchive

Over the past 5 years, a number of orca researchers using our website have added over 18,000 clip annotations to our database. A small section of annotated audio from the Orchive is shown in Figure 1. These clip annotation are of two main types: The first is clips that differentiate background noise from orca calls and from the voice notes of the researchers that collected

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

the data. The second type of clip annotations classify orca vocalizations into different calls. Orcas make three types of vocalizations, echolocation clicks, whistles and pulsed calls. The pulsed calls are highly conserved stereotyped vocalizations which have been classified into a catalog of over 52 different calls by John Ford (?). Of the 18,000 annotations currently in the Orchive, 3000 are of these individually classified calls. In addition, we have a curated call catalog containing 384 different recordings of different calls vocalized by a variety of different pods and matrilines. This catalog is used for training the annotators.

Many parts of the recordings contain boat noise which makes identifying orca calls both difficult and tiring. In addition, the size of the Orchive makes full human annotation practically impossible. Therefore we have explored machine learning approaches to the task. One data mining task is to segment and label the recordings with the labels background, orca, voice. Another is to subsequently classify the orca calls into the classes specified in the call catalog.

2. Related Work

Audio feature extraction is the first step in classifying audio using machine learning algorithms. Mel-Frequency Cepstral Coefficients (?) (MFCC) have been widely used for this purpose. MFCCs have also been used in bioacoustics, and have been used to classify bird songs (?) and orca calls (?). In this work we also use MFCCs, but supplement them with other audio features including Centroid frequency, Rolloff frequency, Flux, and Zero Crossings.

Our system uses two types of web based interfaces. The first are tools aimed at expert users, and the second are simpler interfaces designed for crowdsourcing the annotation. There are a number of tools that experts use to segment and analyze audio and specifically bioacoustic data. One of the most popular is Raven (http://www.birds.cornell.edu/raven), a toolkit developed at the Cornell Lab of Ornithology. The biggest difference our system compared to systems such as Raven is that our system web-based, can more easily view and analyze large amounts of data.

3. System Overview

We have developed a collaborative web interface that allows expert researchers to listen to, view and annotate large collections of audio data. The system also supports a variety of audio feature extraction and machine learning algorithms, and enables users to view the results of these algorithms. A diagram of this system is shown in Figure 2.

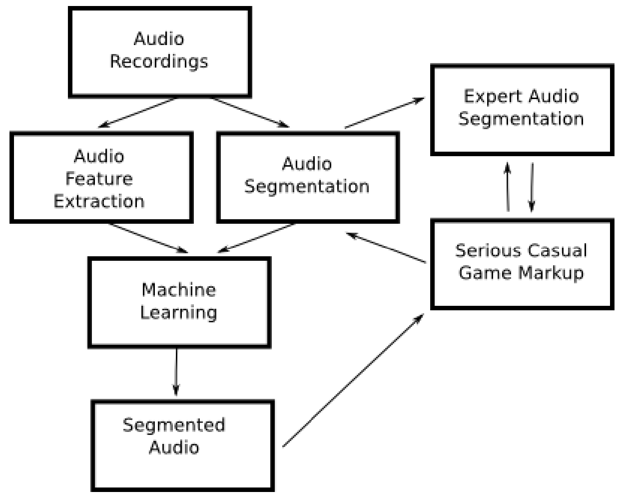

Figure 2. System Diagram

For audio features we use the Marsyas(?) Music Information Retrieval system. Marsyas allows us to perform both audio feature extraction and machine learning on audio data directly.

In order to efficiently analyze large audio archives we utilize distributed computing. There are many systems for distributing computation. We currently use the Portable Batch System (PBS) (?), a grid-computing system where similar data can be processed in parallel by a large number of computers.

4. Experimental Results

4.1. Audio Feature Extraction Parameters

The first set of parameters that needed to be optimized were the Window Size and Hop Size of the Digital Signal Processing (DSP) algorithms that take the input audio and calculate spectral information from them, the fundamental basis for which is the Fast Fourier Transform (FFT) algorithm. The length of time over which to calculate the statistical properties of the features, this is known in bextract as the “memory” and corresponds to the number of frames of features that are accumulated. We ran this on a 600 second audio dataset labeled as orca, background and voice with equal lengths of each label. In this dataset, the voice was trimmed by hand, the orca consisted of the middle 0.023 seconds of approximately 10,000 clips, and the background consisted of 0.15 seconds of approximately 1300 clips. The results for this are shown in Table 1. From this we can see that as we go to longer window sizes, the classification performance increases, and as we go to longer accumulation window sizes, the perfor-

winsize	hopsize	memory	# correct
20	512	256	70.16
20	1024	512	71.88
20	2048	1024	74.17
20	4096	2048	73.38
40	512	256	72.94
40	1024	512	75.67
40	2048	1024	78.29
40	4096	2048	80.58
80	512	256	76.53
80	1024	512	78.39
80	2048	1024	81.88
80	4096	2048	85.72

Table 1. In this table results of a systematic parameter search through different DSP parameters is shown. winsize is the window size of the FFT in samples, and hop size is the number of samples skipped between each successive application of the FFT. memsize refers to the number of FFT frames on which the mean and standard deviation are determined.

Training dataset	length (sec)	% corr. 10-fold	% corr. (calls)	% corr. (442A)
hand-10sec	30	99.4	93.5	93.1
hand-4min	720	99.9	96.1	95.0
ms 100	300	99.9	96.5	93.4

Table 2. Classification results with hand trimmed orca vocalizations using bextract using an SMO SVM classifier.

mance also increases. For the remaining experiments we use these optimal settings.

4.2. Orca/Background/Voice Classification

The first task we investigate is the classification of audio into three classes: orca, background, and human voice. In order to test the different distributed audio classification systems we first generated a set of training and testing data, one of these was a set of calls from the curated call catalog with silence removed, and the other was an entire 45 minute recording from the Orchive which had been annotated by an orca researcher. In a previous paper (?), we were able to obtain a classification performance of 82% when using a SVM classifier on hand labeled data. We looked in more detail at the training data, and found that there was a small amount of silence before and after the vocalization. The results can be found in the first line of table 2 and had 93.5% of the instances classified correctly. This large jump in performance was unexpected but easily understood, because if feature vectors of silence are labeled as orca, this will cause issues for the classifier. We then took a 4 minute region of orca calls and voice notes and removed all the silences from both of them, for this we obtained a classification accuracy of 96.1% when looking at the call catalog dataset, and 95.0% when looking at the annotated recording.

However, this process of hand trimming recordings would be unfeasible to do on the entire 18,000 current annotations. For this, we instead tested a procedure where we extracted a small section of audio from the middle of each clip where it was most probable that the orca call would be found.

We then extracted audio features from these sections of audio using Marsyas. Marsyas has a wide variety of audio features that it can calculate, including MFCCs, number of zero crossings per window and various high level descriptions of the spectrum including the centroid (center of mass of the spectrum), rolloff (the frequency for which the sum of magnitudes of its lower frequencies are equal to percentage of the sum of magnitudes of its higher frequencies) and the flux (the norm of the difference vector between two successive magnitue/power spectra). We tried different combinations of these, and found that using all of these features gave the best performance. All subsequent results in this paper use all of these features.

To classify these features, we used a Sequential Minimal Optimization implementation of a Support Vector Machine classifier (?), an algorithm which had shown its effectiveness in our previous work (?) in this problem domain.

The results for this procedure for a clip of 0.023 seconds from the middle of each orca call was 96.5% and for the recording from the Orchive, the accuracy was 93.4%.

4.3. Call classification

Using the Orchive interface we created a collection of 197 calls of 6 classes, these included the common calls "N1", "N3", "N4", "N7", "N9" and "N47". Audio features for each 20ms audio frame of these files were generated, these included the MFCC coefficients, Centroid, Rolloff, Flux and Zero crossings as described and justified in the previous section. The mean and standard deviation for each of these features were then calculated and were output as a .arff file. The SMO SVM classifier produced gave an accuracy of 98.5% accuracy on this set of calls, and the confusion matrix for this is shown in Table 3.

	N1	N4	N7	N9
N1	1726	0	0	0
N4	12	2858	0	0
N7	0	2	1297	59
N9	0	0	70	3231

Table 3. Confusion matrix for 10-fold crossvalidation with SVM classifier on labelled calls from Orchive.

Training data (sec)	% of Orchive	Run time (d:h:m:s)
30	1	00:00:05:18
30	5	00:00:25:20
30	10	00:00:50:58
30	100	00:09:01:05
240	1	00:06:16
240	5	00:00:31:21
240	10	00:04:47:12
240	100	02:04:18:32

Table 4. Performance results of timing on subsets of the entire Orchive dataset.

4.4. Performance

In order to investigate the performance of the classification of recordings into Orca, Background and Voice, we trained a SVM with a section of 30 and 240 seconds of hand trimmed data using the bextract program in Marsyas. We then used the sfplugin program in Marsyas to classify all the recordings in the Orchive on the Hermes/Nestor cluster, part of the Westgrid computational resource. For this we divided the data into sets of 1%, 5%, 10% and 100% of the Orchive. The timing results of these datasets run on 10 computers are shown in Table 4. From this we can see that the classifier that had more data took longer to classify, and that the speedup from taking samples of the data was almost linear.

5. Conclusion

In this paper we described a system that allows orca researchers to listen to, view and annotate the large amount of audio data in the Orchive. The system also allows researchers to run and view the results of audio feature extraction and machine learning algorithms on this data.

We investigated the performance of different parameters for the audio feature extraction process and showed that in general, large window sizes were beneficial, and that increasing the length of time that statistics were taken over the data was also beneficial. We showed that by carefully hand editing clips to remove silence was very useful, and boosted performance from around 90% to 96% on actual recordings. We then used these classifiers on a cluster to classify all the recordings in the Orchive into the classes, Orca, Background and Voice. The performance of call classification was also good, with a classification accuracy of 98.5% using a collection of 197 calls culled from the Orchive. The calls most often misclassified were the N7 and N9 calls, and these are also difficult for non-experts in orca vocalizations to differentiate.

References

Ford, J.K.B. A catalogue of underwater calls produced by killer whales (orcinus orca) in british columbia. Technical Report 633, Canadian Data Report of Fisheries and Aquatic Science, 1987.

Henderson, R. Job scheduling under the portable batch system. In Feitelson, Dror and Rudolph, Larry (eds.), *Job Scheduling Strategies for Parallel Processing*, volume 949 of *Lecture Notes in Computer Science*, pp. 279–294. Springer, 1995.

Lee, C.H., Lien, C.C., and Huang, R.Z. Automatic recognition of birdsongs using mel-frequency cepstral coefficients and vector quantization. In *IMECS*, pp. 331–335, 2006.

Logan, B. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

Ness, S.R., Wright, M, Martins, L.G., and Tzanetakis, G. Chants and orcas: semi-automatic tools for audio annotation and analysis in niche domains. In *Proc. of the 2nd ACM workshop on Multimedia semantics*, pp. 9–16, 2008.

Platt, John C. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Advances In Kernel Methods - Support Vector Learning, 1998.

Tzanetakis, G. *Marsyas-0.2: A case study in implementing music information retrieval systems*, chapter 2, pp. 31–49. Intelligent Music Information Systems: Tools and Methodologies. Information Science Reference, 2008. Shen, Shepherd, Cui, Liu (eds).