ORCHIVE: Digitizing and Analyzing Orca Vocalizations

George Tzanetakis & Mathieu Lagrange
Department of Computer Science
University of Victoria, Canada
{gtzan, lagrange}@uvic.ca

Paul Spong & Helena Symonds
Orcalab, Hanson Island, BC, Canada

Abstract

This paper describes the process of creating a large digital archive of killer whale or orca vocalizations. The goal of the project is to digitize approximately 20000 hours of existing analog recordings of these vocalizations in order to facilitate access to researchers internationally. We are also developing tools to assist content-based access and retrieval over this large digital audio archive. After describing the logistics of the digitization process we describe algorithms for denoising the vocalizations and for segmenting the recordings into regions of interest. It is our hope that the creation of this archive and the associated tools will lead to better understanding of the acoustic communications of Orca communities worldwide.

Introduction

The fish eating killer whales or orcas, Orcinus Orca, that are the focus of this project live in the coastal waters of the northeastern Pacific Ocean. They are termed *residents* and live in the most stable groups documented among mammals. Resident orcas emit a variety of vocalizations including echolocation clicks, tonal whistles, and pulsed calls (Ford et al., 2000). The Northern Resident Community consists of more than 200 individually known orcas in three acoustic clans. It is regularly found in the study area of the Johnstone Strait and the adjacent waters off Vancouver Island, British Columbia from July to October.

The goal of the Orchive project is to digitize acoustic data that have been collected over a period of 36 years using a variety of analog media at the research station OrcaLab (http://www.orcalab.org) on Hanson Island which is located centrally in the study area. Currently we have approximately 20000 hours of analog recordings mostly in high quality audio cassettes. In addition to the digitization effort which is underway we are developing algorithms and software tools to facilitate access and retrieval to this large audio collection. The size of this collection makes access and retrieval especially challenging (for example it would take approximately 2.2 years of continuous listening to cover the entire archive). Therefore the developed algorithms and tools are essential for effective long term studies employing acoustic techniques. This project is just beginning but we believe it provides many opportunities and challenges related to large scale semantic access in a non-typical application scenario. We are looking forward to receiving feedback from other researchers in information access and retrieval regarding this project. The people involved in the project are all volunteers and the software developed is open source. We welcome help and contributions from any interested parties. Finally we hope that in the future researchers from around the world will be able to access this repository and use the developed tools to improve understanding of acoustic communication of orcas.

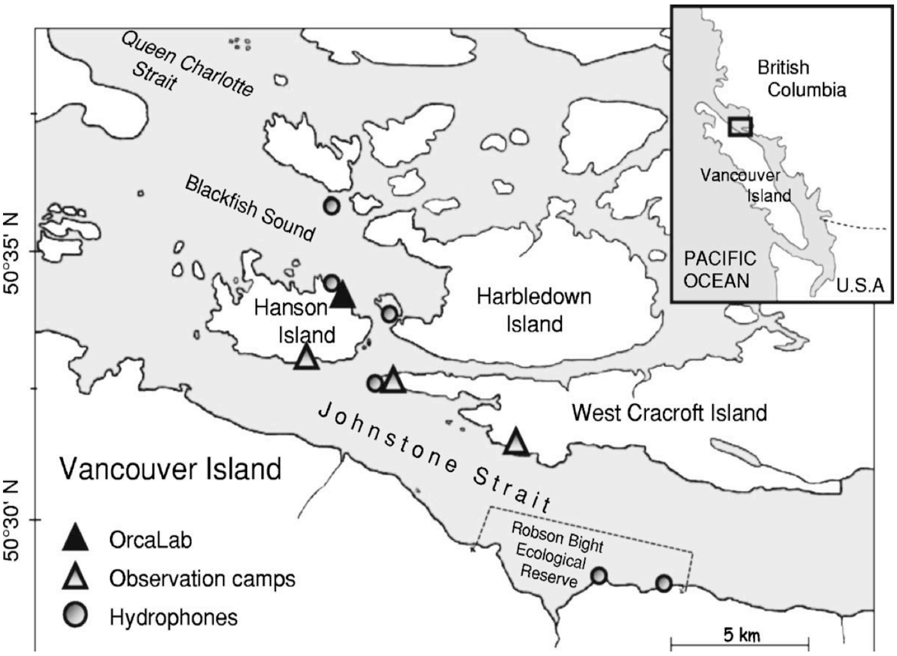

Figure 1: Summer core area of the Northern Resident community with land-based observation sites and the OrcaLab hydrophone network.

The recordings

The acoustic data has been collected with a network of up to six radio-transmitting custom-made hydrophone stations (overall system frequency response 10 Hz to 15 kHz) monitoring the underwater acoustic environment of the area continuously, 24 hours a day and year around. Figure 1 shows the geographical layout of the land observation sites and hydrophone network used for collecting the data. Whenever whales are vocal the mixed output of radio receivers tuned to the specific frequencies of the remote transmitters is recorded on a two-channel audio cassette recorder. Use of mixer controls allows distinction of hydrophone stations and thus basic tracking of group movements. In addition to the acoustic data corresponding information is recorded in logs that are associated with specific tapes. This information is based on acoustic and visual data collected by volunteers, other independent researchers and whale watch operators. It includes various types of information such as number and identity of individuals, group composition, group cohesion, direction of movement and behavioral state (travel, motionless, forage and socialize). The logs also record the mixer settings.

The majority of the audio recordings consist of three broad classes of audio signals: background noise caused mainly by the hydrophones, boats etc, background noise containing orca vocalizations and voice over sections where the observer that started the recording is talking about the details of the particular recording. In some cases there is also significant overlap between multiple orca vocalizations. The orca vocalizations frequently can be categorized into discrete calls that allow expert researchers to identify their social group (pod and matriline) and in some cases even allow identification of individuals.

Digitization

The logistics of digitizing all these analog audio tapes are challenging. We plan to record the 20000 hours of audio to uncompressed digital audio in stereo with a sampling rate of 44100 Hz, and 16-bit dynamic range. This results in approximately 1 GB / hour of audio so we will eventually require storage for 20 Terrabytes (TB) of data.

After taking into consideration various tradeoffs between time, cost, robustness, power consumption (important for Orcalab which is not connected to a power grid) and other factors we decided to start the pilot phase of the project using two digitization stations. One of the stations is located at the University of Victoria and the other is located at Orcalab. Once the pilot phase is completed and any potential problems are worked out we will be able to scale the digitization effort by purchasing more digitizing stations. Each digitizing station consists of the following components: 1 Apple Mac Mini small-factor desktop computer, 2 dual tape Tascam 322 cassette players and 1 Tascam FW1804 multichannel firewire audio interface.

Each digitization station is capable of recording 4 stereo analog audio cassettes (8 channels) simultaneously at 44100 Hz. Although it is possible with more specialized hardware to record more channels simultaneously into a single computer we have determined that 8 channels is best in terms of cost effectiveness and robustness. Custom software has been written to simplify the process of digitization requiring minimum human involvement (just manual loading of the cassettes and pressing the play buttons). Volunteer students and researchers conduct the digitizing process. Assuming 6 hours of digitization per day one station is capable of processing 16 tapes per day and it would take approximately 3.5 years to digitize the entire archive. More digitization speeds up the process linearly. Currently we have digitized approximately 200 tapes corresponding to 1% of the total archive. Digitizing these data has enabled us to iron out problems with the throughput and software as well as provide a test-bed for the tools described in the following sections. For storage we are currently using a combination of three storage devices: high-quality DVD+R are used for long term archiving as they have a much longer life than regular DVDs, 360 Gigabyte external hard drives are used for communication of data between the two locations and temporary storage, and a 10 Terabyte Apple XServe-Raid, that we eventually plan to scale up to 20 TB is used for storing the overall archive and for data processing.

Although tedious the process of creating the digital archive is straightforward. However, as is increasingly the case with large multimedia archives, the central challenge is not the creation and storage of the archive but effective and efficient access. In order to address this challenge we have been developing various algorithms and software tools designed for audio analysis and adapted for the specific constraints of our archive. In the following sections we describe tools for denoising, segmentation and classification.

Denoising

In most of the recordings the background noise level is very high. This is caused by water movement, rain drops, passing boats and underwater acoustic transmission. Therefore denoising is more challenging than standard audio recordings using regular microphones. Standard denoising algorithms require the user to provide a recording of the background noise, calculate its statistics and subsequently use these statistics to filter out the corresponding frequencies, see Vaseghi et al (1992) for further references. In many of the orca vocalizations standard denoising approaches fail. For example boat noise frequently changes in frequency over time because of Doppler shifts as well as changes in engine speed.

On the other hand orca vocalizations are optimized for transmission in the noisy underwater environment and therefore exhibit strong harmonic peaks with smoothly varying amplitudes and frequencies. We take advantage of this property in our denoising algorithm.

In order to separate the orca vocalizations from the background noise we utilize a new data-driven algorithm for simultaneous partial tracking and sound source formation proposed in Lagrange and Tzanetakis (2006). The algorithm is inspired by ideas in Computational Auditory Scene Analysis (Bregman, 1990) and allows a variety of grouping criteria based on in frequency, amplitude, time and harmonic proximity.

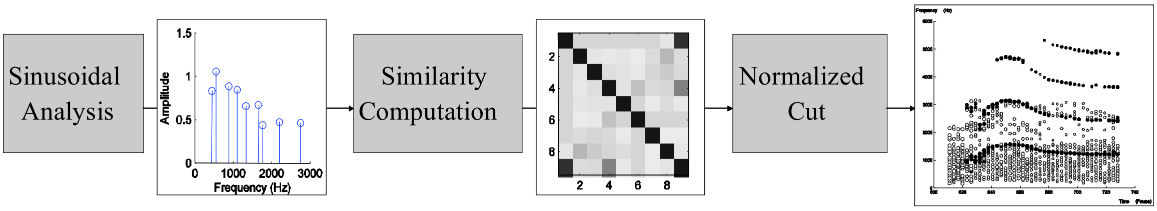

Figure 2: Block-Diagram of the audio analysis unit used for denoising orca vocalizations.

Computational Auditory Scene Analysis (CASA) systems aim at identifying perceived sound sources (e.g. notes in the case of music recordings) and grouping them into auditory streams using psycho-acoustical cues. However, as remarked in (Vincent, 2006) the precedence rules and the relevance of each of those cues with respect to a given practical task is hard to assess. Our goal is to provide a flexible framework where these perceptual cues can be expressed in terms of similarity between time/frequency components. The identification, and separation task is then carried out by clustering components that are close in the similarity space. The underlying representation we used as input to the clustering algorithm is sinusoidal analysis.

Sinusoidal modeling aims to represent a sound signal as a sum of sinusoids characterized by amplitudes, frequencies, and phases. A common approach is to segment the signal into successive frames of small duration and identify local maxima in the spectrum of those frames, usually called peaks. In order to determine whether a peak belongs to the background or to a potential orca call, we use a graph partition algorithm called the *normalized cut*, successfully applied to image and video segmentation (Shi et al, 2000).

In our approach, each partition is a set of peaks that are grouped together such that the similarity within the partition is minimized and the dissimilarity between different partitions is maximized over a texture window of several audio analysis frames. The edge weight connecting two peaks depends on the proximity of frequency, amplitude and harmonicity. Preliminary experiments show that this algorithm is able to optimize continuity and harmonicity constraints globally at larger time scale. As a result, components with coherent frequency evolutions like orca calls are more easily tracked over time, even at low Signal-to Noise ratios.

Figure 2 shows a block diagram of the denoising process. The audio signal is initially analyzed using a Short Time Fourier Transform (STFT) and the peaks of the magnitude spectrum are identified. We also apply amplitude and frequency correction based on phase information to compensate (to some extent) inaccuracies due to windowing and the limited frequency resolution of the FFT. Once the peaks of a "texture" window of approximately 1 second (10-20 audio analysis frames) have been detected a similarity matrix that contains all the pairwise similarities between peaks is calculated. This similarity matrix is used as input to the clustering algorithm that utilizes the Normalized Cut criterion. The cluster with the largest within self-similarity is selected as the separated (denoised) signal.

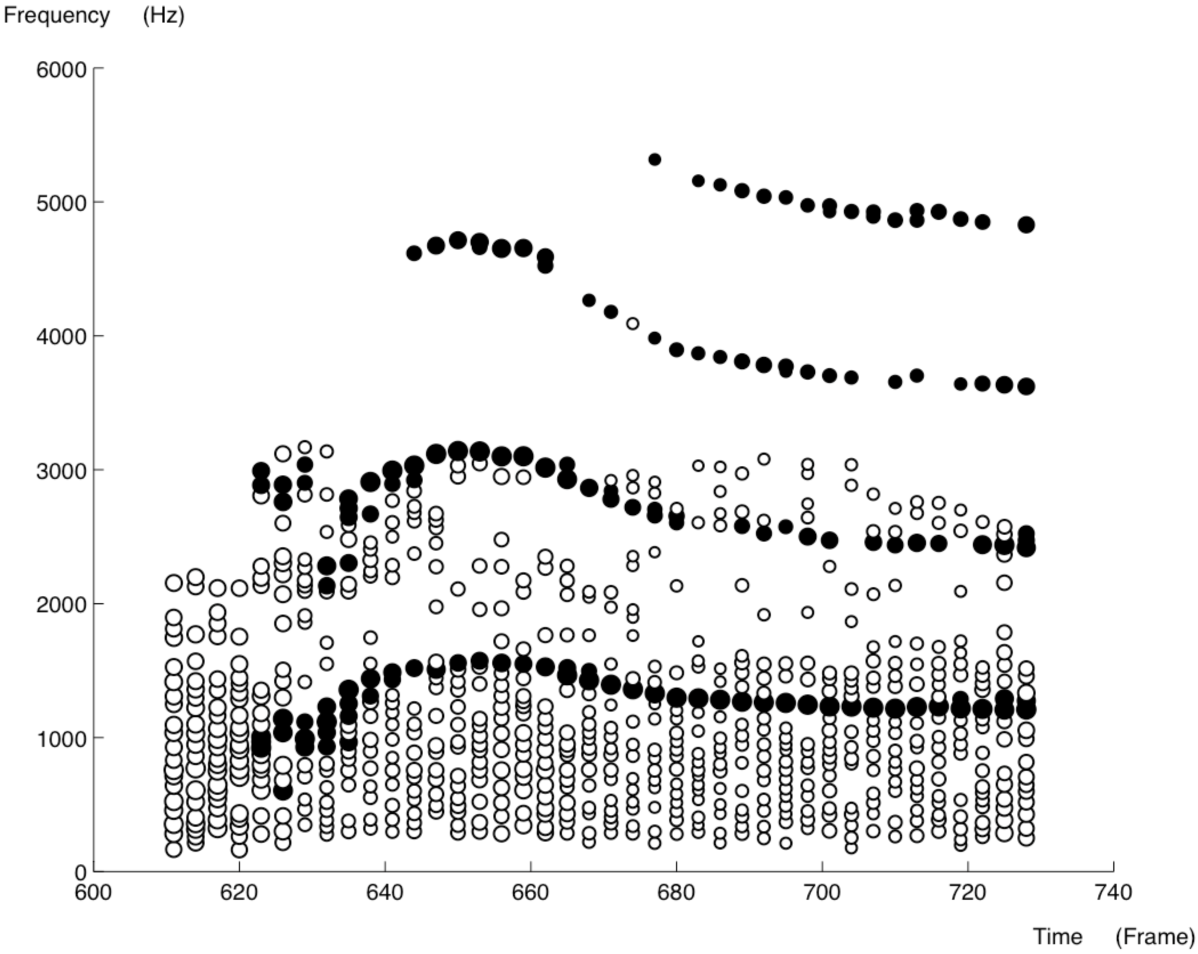

Figure 3: Orca vocalization (black circles) and background noise (transparent circles)

Figure 3 shows how an orca vocalization is separated from the background noise. Each circle corresponds to a sinusoidal peak and their radius is proportional to their amplitude. The circles (peaks) corresponds to the orca vocalization are shown in black. As can be seen from the figure the amplitude of the background noise is significant and broadband. It is important to note that traditional partial tracking algorithms such as (McAuley & Quatieri, 1986) have difficulty tracking partials in such noise as they only consider two successive frames (columns of circles in the figure).

Classification and Segmentation

Locating a particular segment of interest in a long monolithic audio recording can be very tedious as users have to listen to a lot of irrelevant parts until they can locate what they are looking for. Even though visualizations such as spectrograms can provide some assistance they still require a lot of manual effort. In this section we describe some initial proof-of-concept experiments for the automatic classification and segmentation of the orca recordings for the purposes of locating segments of interest.

We have built a classifier for the automatic detection of the three main types of audio encountered in the recordings namely: voice over, background noise and background noise with orca vocalizations, see middle of Figure 4. Initial evaluations using either artificial neural networks or support vector machines are very encouraging. By using a subset of the features proposed in Tzanetakis and Cook (2002) for musical genre classification we are able to correctly classify audio frames of 1 second with 95% accuracy. These initial experiments were done using 10 minutes of audio for training and 10 minutes for testing. The training data and testing data came from different analog cassettes. We are currently working on using larger datasets for training and testing. The developed segmentation and classification system is also going to be used in the Venus (http://www.venus.uvic.ca), and Neptune underwater observatory networks (http://www.neptunecanada.ca.uvic.ca) for analysis and processing hydrophone array data. Figure 4 shows time domain waveforms and spectrograms of an excerpt from a recording. The denoised versions are shown in the bottom of the figure and the automatic annotation to the main types of audio is also shown.

The most common vocalizations are “discrete calls” which are highly stereotyped pulsed calls that can be divided into distinct call types (Ford, 1989). Studies have showed that “pods” which are social units comprised for one or more closely related matrilines have unique vocal repertoires of 7 to 17 discrete call types. Even though we still haven’t developed algorithms to classify calls into types we utilize a variation of the segmentation methodology proposed in Tzanetakis and Cook (2000) to identify the onsets of these “discrete calls”. That way, researchers can directly skip forward and backward through the recording based on semantic units (the “discrete calls”) rather than arbitrary units selected by the zoom level of the audio editing software.

There has been work in quantifying patterns of variation in orca dialects using Artificial Neural Networks (ANN) (Deecke, et al., 1999). We plan to extend upon this research by using machine learning algorithms such as ANN that require large amounts of training data which we can provide using our archive. Another important characteristic of having a large archive is the possibility of studying the evolution of orca calls and their frequencies within different social groups across time. For example it has been observed that the frequency of certain calls increases in the days following the birth of a new calf returning prebirth values within 2 weeks. This may facilitate the learning process of this “acoustic family badge” and thereby help to recognize and maintain cohesion with family members (Weib et al., 2006). By using the automatic segmentation and classification tools we hope to be able to conduct such quantitative experiments over larger time scale and more data without extensive human annotation effort.

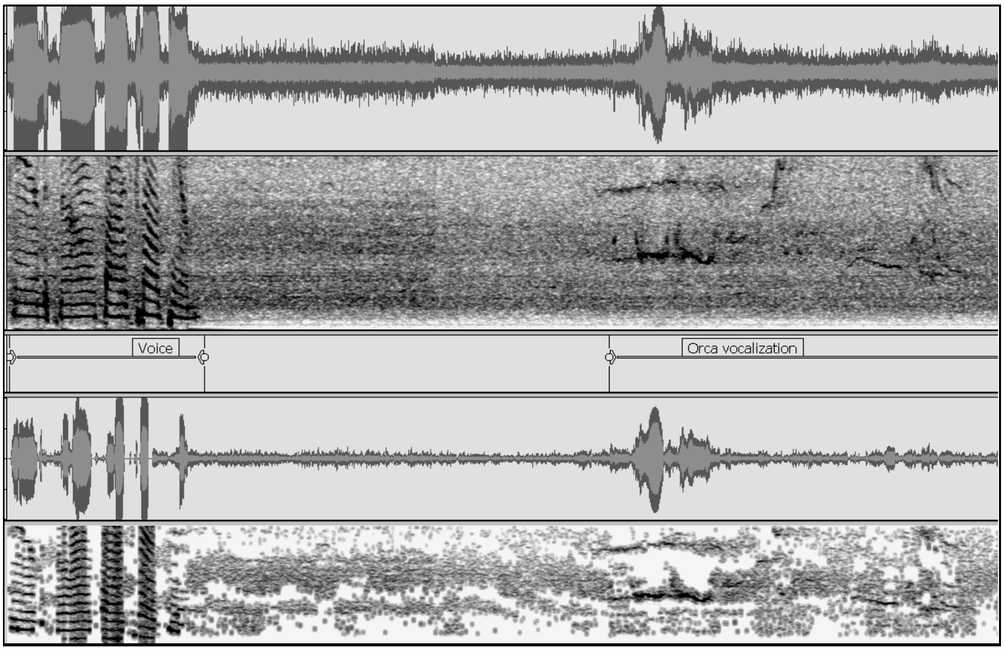

Figure 4: Automatic labeling and denoising of Orca recordings using Marsyas tools. The bottom two time domain and spectral plots correspond to the denoised signals.

Discussion and Future Work

All the software and tools used in the Orchive project are developed using Marsyas (http://marsyas.sourceforge.net) which is a free software framework for audio analysis, retrieval and synthesis. It is important to emphasize that our goal is to provide tools and support to assist researchers in understanding orca vocalizations rather than replacing the human element in the process. The digitization effort is under way and we believe we are ready to scale to more stations and larger archives. The Orchive project is just starting so there is a lot of room for future work. Future directions we plan to explore include: automatic identification of individual calls using supervised learning methods, classification to pod and matriline, and similarity detection to identify calls across time. Another goal is the development of a continuous monitoring system that automatically detects when orcas vocalize and starts direct digital recording. Finally we plan to provide many of our access tools as web services so that researchers can access directly the parts of the signal they are interested in without having to download the entire recording. As an example scenario a researcher might request all instances of N2 (a particular type of call) over the period of 1995-1997 with the background noise removed. We encourage anyone who is either interested in the acoustic data or the tools we are developing to contact us for further information and/or possible collaborations.

References

Bregman A. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, Massachusetts: MIT Press

Ford, J.K.B. (1989) Acoustic Behavior of Resident Killer Whales (Orcinus Orca) off Vancouver Island, British Columbia. *Canadian Journal of Zoology* 64, 727-745.

Ford, J.K. B., Ellis, G. M., and Balcomb, K.C. (2000). Killer Whales: The natural history and genealogy of Orcinus Orca in British Columbia and Washington, 2nd ed (UBC, Vancouver)

Lagrange, M., and Tzanetakis, G. (2006) Sound Sound Formation and Tracking using the Normalized Cut. in "Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)

McAuley, R. J., and Quatieri, T.F. (1986) Speech Analysis/Synthesis Based on Sinusoidal Representation. IEEE Transactions on Acoustics, Speech, and Signal Processing 34(4), 744-754.

Shi J. and Malik j. (2000) Normalized cuts and image Segmentation in IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 22(8), pp. 888-905.

Vaseghi S. V. et al (1992) Restoration of old gramophone recordings. *Journal of the Audio Engineering Society*, 40(10), Oct. 1992.

Tzanetakis, G. and P. Cook (2002) Musical Genre Classification of Audio Signals, *IEEE Transactions on Speech and Audio Processing,* 10(5)

Tzanetakis, G. and P. Cook (2000) Multi-Feature Audio Segmentation for Browsing and Annotation in Proc. IEEE Workshop on Applications of Signals Processing to Audio and Acoustics (WASPAA).

Vincent, E. (2006) Musical Source Separation using Time-Frequency Priors. *IEEE Transactions on Audio, Speech and Language Processing*. 14(1), 91-98.

Weib, B.M., Ladich, F., Spong, P. and Symonds, H. (2006) Vocal behavior of resident killer whale matrilines with newborn calves: The role of family signatures. *Journal of Acoustical Society of America*. 119(1), 627-635.