Chants and Orcas: Semi-automatic Tools for Audio Annotation and Analysis in Niche Domains

Steven R. Ness
Department of Computer Science
University of Victoria
sness@sness.net

Matthew Wright
Department of Computer Science
University of Victoria
mattwrig@uvic.ca

L. Gustavo Martins
Telecommunications and Multimedia Unit
INESC Porto
Porto, Portugal
lmartins@inescporto.pt

George Tzanetakis
Department of Computer Science
University of Victoria
gtzan@cs.uvic.ca

ABSTRACT

The recent explosion of web-based collaborative applications in business and social media sites demonstrated the power of collaborative internet scale software. This includes the ability to access huge datasets, the ability to quickly update software, and the ability to let people around the world collaborate seamlessly. Multimedia learning techniques have the potential to make unstructured multimedia data accessible, reusable, searchable, and manageable. We present two different web-based collaborative projects: *Cantillion*, and the *Orchive*. *Cantillion* enables ethnomusicology scholars to listen and view data relating to chants from a variety of traditions, letting them view and interact with various pitch contour representations of the chant. The *Orchive* is a project to digitize over 20,000 hours of *Orcinus orca* (killer whale) vocalizations, recorded over a period of approximately 35 years, and provide tools to assist their study. The developed tools utilize ideas and techniques that are similar to the ones used in general multimedia domains such as sports video or news. However, their niche nature has presented us with special challenges as well as opportunities. Unlike more traditional domains where there are clearly defined objectives one of the biggest challenges has been the desire to support researchers to formulate questions and problems related to the data even when there is no clearly defined objective.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Human Factors

MS'08, October 31, 2008, Vancouver, British Columbia, Canada.

Keywords

multimedia annotation, multimedia analysis, audio feature extraction, semi-automatic annotation, machine learning

1. INTRODUCTION

Web-based software has been helping connect communities of researchers since its inception. Recently, advances in software and in computer power have dramatically widened its possible applications to include a wide variety of multimedia content. These advances have been primarily in the business community, and the tools developed are just starting to be used by academics.

In recent years there has been increasing research activity in the areas of multimedia learning and information retrieval. Most of it has been in traditional specific domains, such as sports video [10], news video [9] and natural images. There is broad interest in these domains and in most cases there are clearly defined objectives such as highlights in sports videos, explosions in news video or sunsets in natural images. Our focus in this paper is two rather niche domains that share the challenge of effectively accessing large amounts of data but have their own specific characteristics and challenges [5]. Interest in these domains is much more focused and specific. Unlike traditional multimedia data where most users, including the developers of tools, can be used as annotators, in these niche domains any type of annotation requires highly trained experts. These are also problem seeking domains where there are no clearly defined objectives and formulating problems is as important as solving them. We believe that despite these challenges it is possible to develop semi-automatic tools that can assist researchers in significantly improving access and understanding of large data collections.

We have been working on applying these technologies to ongoing collaborative projects that we are involved in. By leveraging several new technologies including *Flash*, *haXe*, *AJAX* and *Ruby on Rails*, we have been able to rapidly develop web-based tools that have utility for our scientific partners. Rapid prototyping and iterative development have been key elements of our collaborative strategy. This agile development strategy has proven its effectiveness in these

projects. Although our number of users is limited compared to other areas of multimedia analysis and retrieval, this is to some degree compensated by their passion and willingness to work closely with us in developing these tools.

The first of these collaborations is a project to develop tools to study chants from various traditions around the world including Hungarian *siratok* (laments)[12], Torah cantillation[23], tenth century St. Gallen plainchant[11] [14], and Koran recitation[16] . These diverse traditions share the common theme of having an origin in primarily non-notated melodies which then later became codified. The evolution and spread of differences in the oral traditions of these different chants are a current topic of research in Ethnomusicology.

It has proved difficult to study these changes using traditional methods and it was decided that a combined approach, using field recordings marked up by experts, mathematical models for analyzing the fundamental frequency content of the audio, and a flexible graphic user interface, would help figure out what questions needed to be asked.

The second project involves the analysis of a large archive of recordings of *Orcinus orca* (killer whale) vocalizations [6] recorded at OrcaLab, a research station on the west coast of Canada. There are stable resident populations [7] of *Orcinus orca* in the northwest Pacific Ocean, and some of these populations [8] are found near Hanson Island, off the north tip of Vancouver Island in Canada. Orcalab is a research station that has been recording audio of these Orca populations since 1972 [3, 22]. They have amassed a huge archive of more than 20,000 hours of audio recordings collected via a permanent installation of underwater hydrophones. The archive was recorded onto cassette and DAT tapes. In a previous work [21] a system for digitizing the audio was presented as well as some preliminary results in denoising orca vocalizations.

Although these recordings contain large amounts of Orca vocalizations, the recordings also contain other sources of audio, including voice-overs describing the current observing conditions, boat and cruise-ship noise, and large sections of silence. Finding the Orca vocalizations on these tapes is a labor-intensive and time-consuming task.

In the current work, we present a web-based collaborative system to assist with the task of identifying and annotating the sections of these audio recordings that contain Orca vocalizations. This system consists of a dynamic and user-informed front end written in *XHTML/CSS* and *Flash* which lets a researcher identify and label sections of audio as Orca vocalization, voice-over or background noise. By using annotation boot-strapping [19], an approach inspired by semi-supervised learning, we show that it is possible to obtain good classification results while annotating only a small subset of the data. This is critical as it would take several human years to fully annotate the entire archive. Once the data is annotated it is trivial to focus on data of interest such as all the orca vocalizations for a particular year without having to manually search through the audio file to find the corresponding relevant sections.

2. DOMAINS:

2.1 Chants

Our work in developing tools to assist with chant research is a collaboration with Dr. Daniel Biro, a professor in the School of Music at the University of Victoria. He has been collecting and studying recordings of chant with specific focus on how music transmission based on oral transmission and ritual was gradually changed to one based on writing and music notation. The examples studied come from improvised, partially notated, and gesture-based [13] notational chant traditions: Hungarian siratok (laments) [1], Torah cantillation [24] [2], tenth century St. Gallen plainchant [18] [3], and Koran recitation [4].

Although Dr. Biro has been studying these recordings for some time and has considerable computer expertise for a professor in music, the design and development of our tools has been challenging. This is partly due to difficulties in communication and terminology as well as the fact that the work is exploratory in nature and there are no easily defined objectives. The tool has been developed through extensive interactions with Dr. Biro with frequent frustration on both sides. At the same time, a wonderful thing about expert users like Dr. Biro is that they are willing to spend considerable time preparing and annotating data as well as testing the system and user interface which is not the case in more traditional broad application domains.

2.2 Orca vocalizations

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 on the west coast of Vancouver Island in Canada. 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 for 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. Currently such studies require enormous effort as the relevant acoustic tapes need to be recovered and the relevant segments need to be tediously digitized for analysis.

The majority of the audio recordings consist of three broad classes of audio signals: background noise caused mainly by the hydrophones, boats, 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

[1]Archived Examples from Hungarian Academy of Science (1968-1973)

[2]Archived Examples from Hungary and Morocco from the Feher Music Center at the Bet Hatfatsut, Tel Aviv, Israel

[3]Godehard Joppich and Singphoniker: Gregorian Chant from St. Gallen (GorgmarienhÃijtte: CPO 999267-2, 1994)

[4]Examples from Indonesia and Egypt: in Approaching the Koran (Ashland: White Cloud, 1999)

allow expert researchers to identify their social group (matriline and pod) and in some cases even allow identification of individuals.

Even when the data is digitized, locating a particular segment of interest in a long monolithic audio recording can be very tedious as users have to listen to many irrelevant sections until they can locate what they are looking for. Even though visualizations such as spectrograms can provide some assistance this is still a task that requires much manual effort. In this paper we describe experiments for the automatic classification and segmentation of the orca recordings for the purposes of locating segments of interest and facilitating interaction with this large audio archive.

3. ANALYSIS AND BROWSING

3.1 Melodic Contour Analysis

Our tool takes in a (digitized) monophonic or heterophonic recording and produces a series of succesively more refined and abstract representations of the segments it contains as well as the corresponding melodic contours . More specifically the following analysis stages are performed:

- Hand Labeling of Audio Segments
- First Order Markov Model of Sign Sequences
- F0 Estimation
- F0 Pruning
- Scale Derivation: Kernel Density Estimation
- Quantization in Pitch
- Scale-Degree Histogram
- Histogram-Based Contour Abstraction
- Plotting and Recombining the Segments

The recordings are manually segmented and annotated by the expert. Even though we considered the possibility of creating an automatic segmentation tool, it was decided that the task was too subjective and critical to automate. Each segment is annotated with a word/symbol that is related to the corresponding text or performance symbols used during the recitation.

In order to study the transitions between signs/symbols we calculate a first order Markov model of the sign sequence for each recording. We were asked to perform this type of syntagmatic analysis by Dr. Biro. Although it is completely straightforward to perform automatically using the annotation, it would be hard, if not impossible, to calculate manually. Figure 3.1 shows an example transition matrix. For a given trope sign (a row), how many total times does it appear in the example (numeral after row label), and in what fraction of those appearances is it followed by each of the other trope signs? The darkness of each cell corresponds to the fraction of times that the trope sign in the given row is followed by the trope sign in the given column. (NB: Cell shading is relative to the total number of occurrences of the trope sign in the row, so, e.g., the black square saying that "darga" always precedes "revia" represents 1/1, while the black square saying that "zakef" always precedes "katon" represents 9/9.)

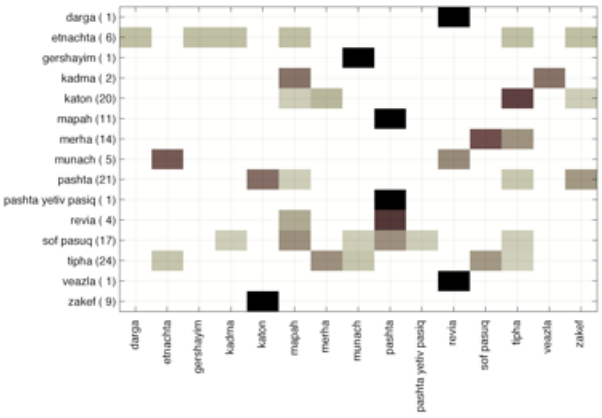

Figure 1: Syntagmatic analysis with a first-order Markov model of the sequence of Torah trope signs for the text Shir Ha Shirim ("Song of Songs").

After the segments have been identified, the fundamental frequency ("F0" in this case equivalent to pitch) and signal energy (related to loudness) are calculated for each segment as functions of time. We use the SWIPEP fundamental frequency estimator [1] with all default parameters except for upper and lower frequency bounds that are hand-tuned for each example. For signal energy we simply take the sum of squares of signal values in each non-overlapping 10-ms rectangular window.

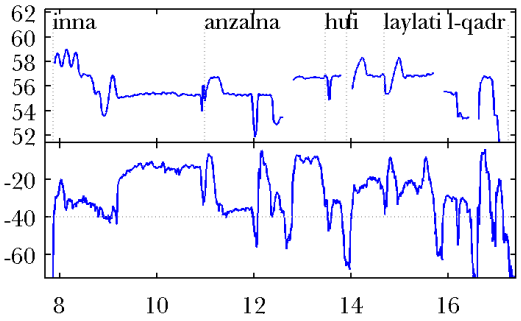

Figure 2: F0 contour

The next step is to identify pauses between phrases, so as to eliminate the meaningless and wildly varying F0 estimates during these noisy regions. We define an energy threshold, generally 40 decibels below each recording's maximum. If the signal energy stays below this threshold for at least 100 ms then the quiet region is treated as silence and its F0 estimates are ignored. Figure 3.1 shows an excerpt of the F0 and energy curves for an excerpt from the Koran sura ("section") Al-Qadr ("destiny") recited by the renowned Sheikh Mahmud Khalil al-Husari from Egypt.

Following the pitch contour extraction is pitch quantization, which is the discretization of the continuous pitch contour into discrete notes of a scale. Rather than externally imposing a particular set of pitches, such as an equal-tempered chromatic (the piano keys) or diatonic scale, we have developed a novel method for extracting a scale from an F0 envelope that is continuous (or at least very densely sampled) in both time and pitch. Our method is inspired

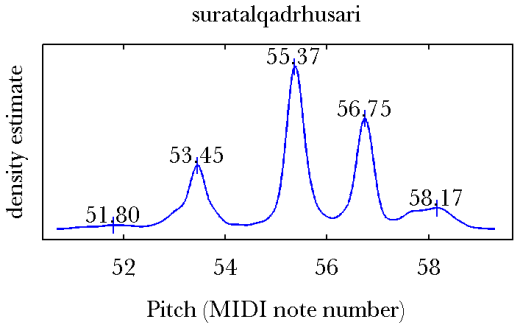

Figure 3: Recording-specific scale derivation

by Krumhansl's time-on-pitch histograms adding up the total amount of time spent on each pitch [13]. We demand a pitch resolution of one cent [5], so we cannot use a simple histogram. Instead we use a statistical technique known as nonparametric kernel density estimation, with a Gaussian kernel [6]. More specifically a Gaussian (with standard deviation of 33 cents) is centered on each sample of the frequency estimate and the Gaussians of all the samples are added to form the kernel density estimate. The resulting curve is our density estimate; like a histogram, it can be interpreted as the relative probability of each pitch appearing at any given point in time. Figure 2 shows this method's density estimate given the F0 curve from Figure 1.

We interpret each peak in the density estimate as a note of the scale. We restrict the minimum interval between scale pitches (currently 80 cents by default) by choosing only the higher peak when there are two or more very close peaks. This method's free parameter is the standard deviation of the Gaussian kernel, which provides an adjustable level of smoothness to our density estimate; we have obtained good results with a standard deviation of 30 cents. Note that this method has no knowledge of octaves.

Once we have determined the scale, pitch quantization is the trivial task of converting each F0 estimate to the nearest note of the scale. In our opinion these derived scales are more true to the actual nature of pitch-contour relationships within oral/aural and semi-notated musical traditions. Instead of viewing these pitches to be deviations of pre-existing "normalized" scales our method defines a more differentiated scale from the outset. With our approach the scale tones do not require "normalization" and thereby exist in an autonomous microtonal environment defined solely on statistical occurrence of pitch within a temporal unfolding of the given melodic context.

Once the pitch contour is quantized into the recording-specific scale calculated using Kernel density estimation, we can calculate how many times a particular scale degree appears during an excerpt. The resulting data is a scale-degree histogram which is used create simplify abstract visual representations of the melodic contours.

[5]One cent is 1/100 of a semitone, corresponding to a frequency difference of about 0.06%

[6]Thinking statistically, our scale is related to a distribution given the relative probability of each possible pitch. We can think of each F0 estimate (i.e each sampled value of the F0 envelope) as a sample drawn from this unknown distribution so our problem becomes one of estimation the unknown distribution given the samples

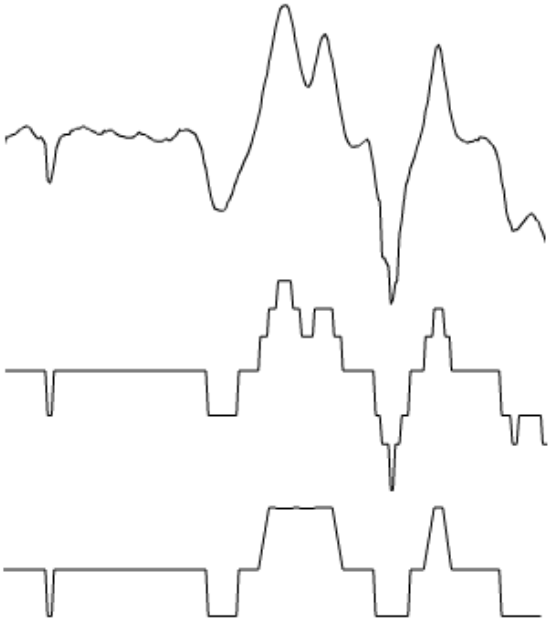

Figure 4: Melodic contours at different levels of abstraction (top: original, middle: quantized, bottom: simplified using 3 most prominent scale degrees

The basic idea is to only use the most salient discrete scale degrees (the histogram bins with the highest magnitude) as significant points to simplify the representation of the contour. By adjusting the number of prominent scale degrees used to represent the simplified representation the researchers can view/listen to the melodic contour at different levels of abstraction and detail. Figure 3.1 shows an original continuous contour, the quantized representation using the recording-specific derived scale and the abstracted representation using only the 3 most prominent scale degrees.

3.2 Cantillion interface

We have developed a browsing interface that allows researchers to organize and analyze chant segments in a variety of ways (`http://cantillation.sness.net`). Each recording is manually segmented into the appropriate units for each chant type (such as trope sign, neumes, semantic units, or words). The pitch contours of these segments can be viewed at different levels of detail and smoothness using a histogram-based method. The segments can also be rearranged in a variety of ways both manually and automatically. That way one can compare the beginning and ending pitches of any trope sign, neume or word.

The interface 3.2 has four main sections: a sound player, a main window to display the pitch contours, a control window, and a histogram window. The sound player window displays a spectrogram representation of the sound file with shuttle controls to let the user choose the current playback position in the sound file. The main window shows all the pitch contours for the song as icons that can be repositioned automatically based on a variety of sorting criteria, or alternatively can be manually positioned by the user. The name of each segment (from the initial segmentation step) appears

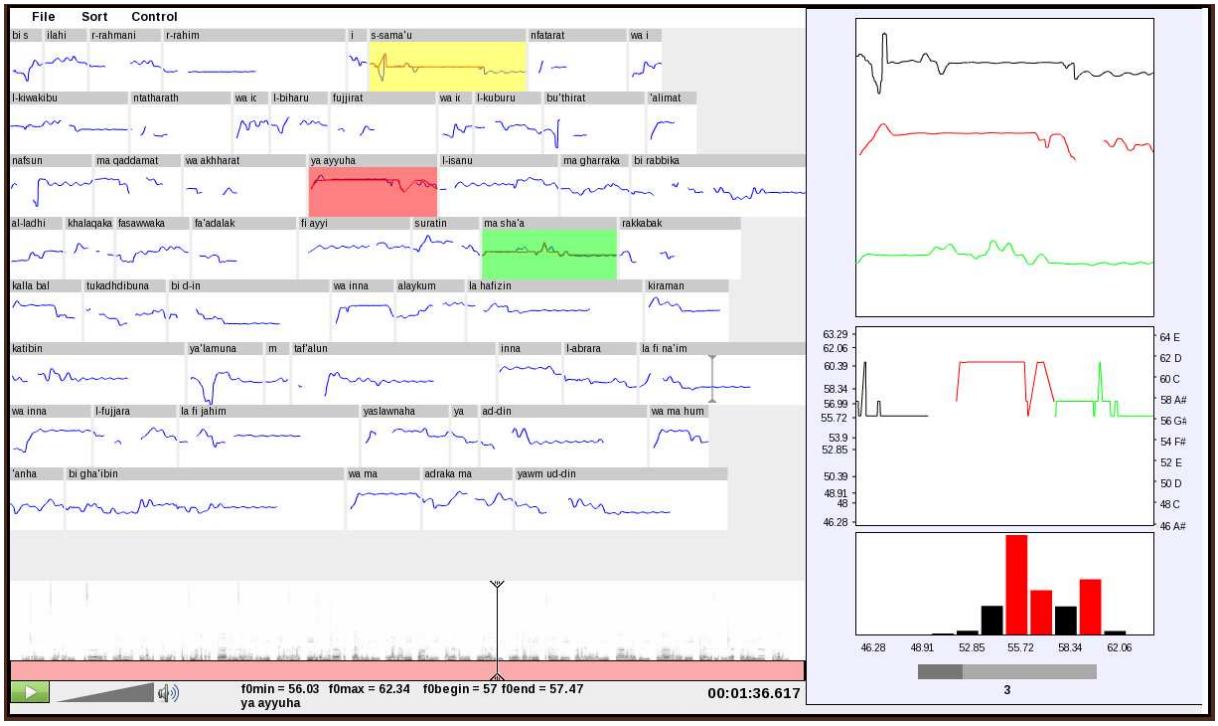

Figure 5: Web-based *Flash* interface to allow users to listen to audio, and to enable interactive querying of gesture contour diagrams.

above its F0 contour. The shuttle control of the main sound player is linked to the shuttle controls in each of these icons, allowing the user to set the current playback state either by clicking on the sound player window, or directly in the icon of interest.

When an icon in the main F0 display window is clicked, the histogram window shows a histogram of the distribution of quantized pitches in the selected sign. Below this histogram is a slider to choose how many of the largest histogram bins will be used to generate a simplified contour representation of the F0 curve. In the limiting case of selecting all histogram bins, the reduced curve is exactly the quantized F0 curve. At lower values, only the histogram bins with the most items are used to draw the reduced curve, which has the effect of reducing the impact of outlier values and providing a smoother "abstract" contour. Shift-clicking selects multiple signs; in this case the histogram window includes the data from all the selected signs. We often select all segments with the same word, trope sign, or neume; this causes the simplified contour representation to be calculated using the sum of all the pitches found in that particular sign, enhancing the quality of the simplified contour representation. Figure 3.2 shows a screenshot of the browsing interface.

The identity of chant formulae in oral/aural chant traditions is to a large extent determined by gesture/contour rather than by discrete pitches. Computational approaches assist with the analysis of these gestures/contours and enables the juxtaposition of multiple views at different levels of detail in a variety of analytical (paradigmatic and syntagmatic) contexts. The possibilities for such complex analysis methods would be difficult if not impossible without such computer-assisted analysis. Employing these tools we hope to better understand the role of and interchange between melodic formulae in oral/aural and written chant cultures. While our present analysis investigates melodic formulae primarily in terms of their gestural content and semantic functionality, we hope that these methods might allow scholars to reach a better understanding of the historical development of melodic formulae within various chant traditions.

3.3 Orchive

The *Orchive* (`http://orchive.cs.uvic.ca`) is a web-based collaborative system designed to assist with the task of identifying and annotating sections of audio recordings that contain orca vocalizations. This system consists of a dynamic front end written in *XHTML/CSS* and *Flash*. The interface allows the user to annotate regions of the recording as "orca" and "voiceover", and automatically assigns the "background" label to unlabeled regions of the audio. In voiceover sections the observer that started the tape recording talks about the details of the particular recording such as the geographic location of the Orcas, the time of the day, the weather conditions and other items of note. A sample section of audio with voiceover, orca vocalizations and background is shown in Figure 3.3. Although we eventually want to provide more detailed classification, such as the type of orca calls, in practical terms this basic classification to three categories is very important to the researchers involved.

This web server then runs audio feature extraction and performs supervised and semi-supervised learning using the *Marsyas* [20] (`http://marsyas.sness.net`) open source software framework for audio analysis.

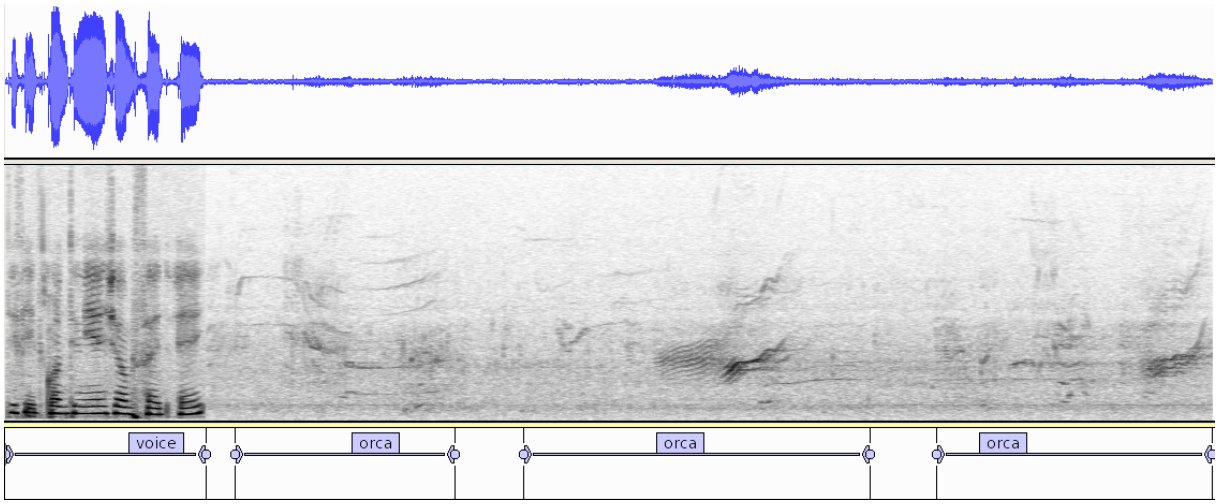

Figure 6: An annotated region of audio from the *Orchive*, with regions of voice and orca vocalization shown. Unlabeled regions are automatically assigned a label of background noise.

OrcaAnnotator is a Model-View-Controller system containing well-defined and well-separated sections, each of which presents a uniform interface to the other sections of the system. Each part is made to be a simple and well-defined unit, making them easier to test and maintain.

The primary mode of communication with the user is via an *XHTML/CSS* and *Flash* based interface. The user is presented with a simple and attractive *XHTML/CSS* web page that has been designed to be standards compliant which will facilitate accessibility by the research community on a wide variety of different web browsers and computer platforms. The *Flash* based interface is written in the haXe [15] programming language, which compiles the ECMAScript language *haXe* down to *Flash* bytecodes. The *Flash* interface presents a simple interface to the user with a spectrogram of the sound file, shuttle and volume controls, a time display, and an interface for labeling the audio file. We used the labeling functionality in *Audacity* as a model for our user-interaction paradigm. To add a label, the user simply clicks and drags the mouse on the label region. This creates a label with left and right extents, and a text region where the user can enter a text description of the audio. In addition, a pull-down menu with labels can be used for quick annotation.

Labels are saved to the database with the user that created them and the time that they were created. This user can be an actual user on the system, or can be labeled with *Marsyas* and the name and parameters of the classifier that was used for labeling. *Marsyas* contains a number of machine-learning classifiers, including Gaussian (MAP), Gaussian Mixture-Model (GMM), and Support Vector Machines (SVM). We used the "bextract" program which is part of *Marsyas*, which now includes a new Timeline module that allows the import of human-annotated sections of audio into *Marsyas* as a start for a bootstrapping approach. A variety of standard audio feature extraction algorithms such as Mel-Frequency Cepstral Coefficients (MFCC) as well as various types of spectral features are also provided. The integration of machine learning and audio signal processing is essential in creating a semi-automatic annotation interface.

To provide communication between the *Flash* user-interface and the *Marsyas* classifier algorithms, we have employed the *Ruby on Rails* web framework[17]. *Ruby on Rails* allows for quick and easy development and deployment of websites, and it provides a tight interface layer to an underlying database like *MySQL*.

Ruby on Rails also has the advantage that it makes it simple to build REST based applications[4]. REST is the model on which the internet is built and has the ability to minimize latency and network communication, while simultaneously maximizing the independence and scalability of network services. *Ruby on Rails* queries the database for user data, label data and locations of audio files. It then generates all the *XHTML/CSS* files displayed to the user and sends the required XML data to the *Flash* application. Once the user submits their annotated data back to the web server, it first stores this data in the database and then queues this data for *Marsyas* to run in a separate background process, perhaps on another machine, or network of machines. Once *Marsyas* completes processing the audio, the results are automatically sent back to the web server using REST web services.

Being able to segment and label the audio recordings into the three main categories (voiceover, orca vocalizations and background noise) is immensly useful to researchers working with this vast amount of data. For example background noise comprises approximately 64% of the recordings, and is much higher in some individual recordings. Fully annotating the data even using a well-designed user interface is out of the question given the size of the archive. To address this problem we have designed a semi-supervised learning system that only requires manual annotation of a small percentage of the data and utilizes machine learning techniques to annotate the other part. This recording-specific annotation bootstrapping can potentially be used with other types of time-based multimedia data.

Table 1: Recording-specific classification performance

	Naive bayes % correct		SMO % correct	
	self	train with remaining	self	train with remaining
446A	89.42	93.10	95.00	73.39
446B	63.45	77.66	85.85	70.23
447B	75.46	57.32	82.02	68.17
448A	52.18	61.02	81.57	62.24
448B	84.63	67.62	83.64	67.87
449B	82.24	51.85	86.41	75.72
450A	94.66	90.91	96.12	91.58
450B	83.65	96.27	99.29	94.92
451A	70.92	89.58	97.04	78.72
451B	74.18	33.73	82.34	50.88

3.4 Annotation Bootstrapping

Annotation bootstrapping is inspired by semi-supervised learning [2]. It has been shown that unlabeled data, when used in conjuction with a small amount of labeled data, can produce considerable improvements in learning accuracy. The acquisition of labeled data for a learning problem often requires manual annotation which is a time consuming process so semi-supervised learning can significantly reduce annotation time for large multimedia archives.

We extend the idea of semi-supervised learning to take advantage of the strong correlation between feature vectors from the same audio recording. In the *Orchive* each audio recording has a duration of 45 minutes and corresponds to a particular date and time. There is considerable consistency within a recording as the same person is doing the voiceover sections, the mixing settings are the same and the orcas that are vocalizing typically come from the same group. A recording-specific bootstrap classifier is trained as follows: a small percentage of the specific audio recording is manually annotated and used to train a recording-specific classifier. This classifier is then used to label the remaining parts of the recording. Due to the consistency of the recording this classifier will be to some extent overfitted to the recording and will not generalize well to other recordings. However, that is not a problem in our case as we are mainly interested in obtained labels for the entire recording. This process is repeated for each recording. Once all the recordings have been semi-automatically fully labeled then feature extraction is performed for the entire archive and a generalizing classifier is trained using the full dataset.

In order to explore whether this idea would work for our data, we created a representative database consisting of 10 excerpts from our recordings with each excerpt lasting between 5 and 10 minutes. Table 1 shows classification results using 10-fold cross-validation for each particular recording using a recording specific classifier as well as using a classifier trained on the entire dataset. Two classifiers are used: a simple Naive Bayes classifier (NBS), as well as a Support Vector Machine (SVM). The results shown are based on the use of the standard Mel-Frequency Cepstral Coefficients (MFCC) as audio features. The "self" column shows the classification accuracy results of using a recording-specific classifier, whereas the "remaining" columns shows the classification accuracy results using the remaining nine recordings. As can be seen, recording-specific classifier can generate significantly better results than generalized classifiers, which is not surprising as they adapt to the specific data of the recording. This justifies the use of their annotation results to labeled the unlabeled parts of the audio recording.

Table 2: Classification performance using annotation-bootstrapping (SVM classifier)

% data used	%correct	F-measure
100	82.38	0.876
10	81.98	0.874
5	82.04	0.874
1	79.95	0.864
0.1	78.08	0.857
0.01	71.42	0.800

The goal of annotation bootstrapping is to only label a small part of each recording to train a recording-specific classifier which is then used to annotate the remainder of the recording. Table 2 shows the results in terms of classification accuracy and F-measure over the entire dataset for different amounts of labeled data. As one can see the classification accuracy remains quite good, even when only a small percentage of the data is labeled and annotation bootstrapping is used to label the rest. The first row shows the classification accuracy when all the data is used for training.

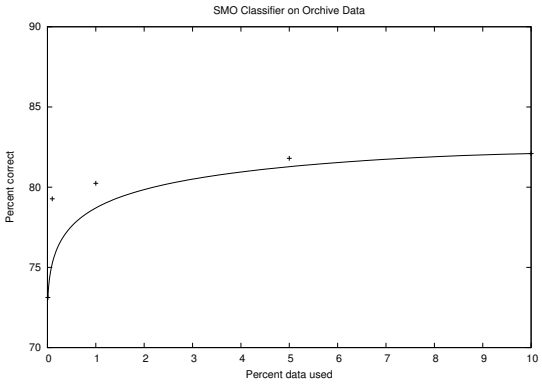

Figure 7: Graph of classification accuracy as percentage of labeling required. Data shown is for the performance of SMO classifier for different percentages of data used to train the classifier.

Figure 3.4 shows graphically how the classification accuracy increases with the amount of labeled data used for training. In both the table and the figure, the classifier used is a Support Vector Machine (SVM) and for evaluation a variation of 10-fold cross-validation where each of the 10 recordings is held out for testing, the remaining ones are used for training, and the process is iterated 10 times. We also experimented with different choices of window size for the feature calculation as well as different audio feature parametrization but there was no significant difference in the obtained results.

To make the importance of annotation bootstrapping concrete, fully annotating the archive would take approximately 2 and half years (assuming 24/7 manual annotation) whereas using one percent annotation bootstrapping would take 3 months (assuming 24/7 manual annotation) without significantly affecting the ability of the system to succesfully label all the recordings in the 3 classes of interest.

4. CONCLUSIONS

By combining the expert knowledge of our scientific collaborators with new multimedia web-based tools in an agile development strategy, we have been able to ask new questions that had previously been out of reach. The large and multi-dimensional datasets in both the chant community and in orca vocalization research provide challenging fields for study, and new web-based technologies provide the flexibility to allow true collaboration between scientific partners in widely disparate fields of study. We described an automatic technique for simplifying melodic contours based on kernel density estimation. Annotation of the archive of orca vocalizations is very time-consuming, we proposed annotation bootstrapping and show that it is an effective technique for automatically annotating recordings.

5. ACKNOWLEDGMENTS

We would like to thank Paul Spong and Helena Symonds of Orcalab and well as Daniel Biro for providing the data and inspiration for this project. We would also like to thank the National Sciences and Engineering Research Council (NSERC) and Social Sciences and Humanities Research Council (SSHRC) of Canada for their financial support.

6. REFERENCES

[1] A. Camacho. *A Sawtooth Waveform Inspired Pitch Estimator for Speech and Music*. PhD thesis, University of Florida, 2007.

[2] O. Chappelle, B. Scholkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[3] V. Deecke, J. Ford, and P. Spong. Dialect change in resident killer whales (orcinus orca): implications for vocal learning and cultural transmission. *Animal Behaviour*, 60(5):619–638, 2000.

[4] R. T. Fielding. *Architectural Styles and the Design of Network-based Software Architectures*. Phd dissertation, University Of California, Irvine.

[5] J. Foote. *Content-based Retrieval of Music and Audio*, pages 138–147. 1997.

[6] J. Ford. 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.

[7] J. Ford. Acoustic behaviour of resident killer whales (orcinus orca) off vancouver island, british columbia. *Canadian Journal of Zoology*, 64:727–745, 1989.

[8] J. Ford, E. G.M., and B. K.C. *Killer Whales : The natural history and genealogy of Orcinus orca in British Columbia and Washington, 2nd ed*. UBC, Vancouver, 2000.

[9] A. Hauptman and et al. Informedia at trec 2003 : Analyzing and searching broadcast news video. In *Proc. of (VIDEO) TREC 2003*, Gaithersburg, MD, 2003.

[10] A. Hauptman and M. Witbrock. *Informedia: News-on-demand Multimedia Information Acquisition and Retrieval*. MIT Press, Cambridge, Mass, 1997.

[11] T. Karp. *Aspects of Orality and Formularity in Gregorian Chant*. Northwestern University Press, Evanston, 1998.

[12] Z. Kodaly. *Folk Music of Hungary*. Corvina Press, Budapest, 1960.

[13] C. L. Krumhansl. *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford, 1990.

[14] K. Levy. *Gregorian Chant and the Carolingians*. Princeton University Press, Princeton, 1998.

[15] L. McColl-Sylvester and F. Ponticelli. *Professional haXe and Neko*. Wiley Publishing, Inc., Indianapolis, IN, 2008.

[16] K. Nelson. *The Art of Reciting the Koran*. University of Texas Press, Austin, 1985.

[17] D. Thomas, D. Hansson, L. Breedt, M. Clark, J. D. Davidson, J. Gehtland, and A. Schwarz. *Agile Web Development with Rails, 2nd Edition*. Pragmatic Bookshelf, Flower Mound, TX, 2006.

[18] L. Treitler. The early history of music writing in the west. *Journal of the American Musicological Society*, 35, 1982.

[19] G. Tzanetakis. Song specific bootstrapping of singing voice structure. In *Proc. Int. Conf. on Multimedia and Exposition ICME*, TaiPei, Taiwan, 2004. IEEE.

[20] G. Tzanetakis and P. Cook. Marsyas: A framework for audio analysis. *Organized Sound*, 4(3), 2000.

[21] G. Tzanetakis, M. Lagrange, P. Spong, and H. Symonds. Orchive: Digitizing and analyzing orca vocalizations. In *Proc. of the RIAO, Large-Scale Semantic Access to Content Conference*. RIAO, 2007.

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

[23] G. Wigoder and et al. *Masora, The Encyclopedia of Judaism*. MacMillan Publishing Company, New York, 1989.

[24] H. Zimmermann. *Untersuchungen zur Musikauffassung des rabbinischen Judentums*. Peter Lang, Bern, 2000.