Strategies for orca call retrieval to support collaborative annotation of a large archive

Steven R. Ness [#1], Alex Lerch [*2], George Tzanetakis [#3]

[#] *Computer Science, University of Victoria*
Canada
[1] sness@sness.net, gtzan@cs.uvic.ca

[*] *zplane.development inc.*
Germany
[2] lerch@zplane.de

***Abstract*—The Orchive is a large audio archive of hydrophone recordings of Killer whale (*Orcinus orca*) vocalizations. Researchers and users from around the world can interact with the archive using a collaborative web-based annotation, visualization and retrieval interface. In addition a mobile client has been written in order to crowdsource Orca call annotation. In this paper we describe and compare different strategies for the retrieval of discrete Orca calls. In addition, the results of the automatic analysis are integrated in the user interface facilitating annotation as well as leveraging the existing annotations for supervised learning. The best strategy achieves a mean average precision of 0.77 with the first retrieved item being relevant 95% of the time in a dataset of 185 calls belonging to 4 types.**

I. Introduction

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 domains, such as sports video, news video, and natural images [1]. 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. Some of the important research trends in multimedia retrieval research have been the use of large collections for supervised learning, the integration of the user interface and the annotation/retrieval system, and the shift from single user system to collaborative web-based interfaces which enable client-cloud architectures.

Our focus in this paper is applying similar ideas to the Orchive [2], a large archive of audio recordings of killer whale (*Orcinus orca*) vocalizations from the Northern resident community of British Columbia. It has been shown that different killer whale communities use distinct vocal signals [3]. Pods are stable kin groups and have unique vocal repertoires consisting of 7-17 distinct calls. Related pods often use structurally distinct versions of the same class type.

Currently the Orchive contains approximately 10000 hours of audio data digitized from the original analog cassettes with a projected total size of 20000 hours (it would take approximately 6 years listening 8 hours every day to cover the entire archive). Traditionally researchers had to digitize the analog cassettes and process the resulting files individually on their computers. This tedious process has inhibited researchers from analyzing the amounts of data that are available. Harnessing the large amounts of data provided in the archive holds enormous potential in advancing our understanding of how these animals communicate. However, in order for the data to be effectively analyzed it needs to be annotated and automatic retrieval tools need to be developed. We have developed a collaborative web-based interface that is enhanced with retrieval and classification capabilities which are used to support the annotation process.

Unlike other areas of multimedia retrieval such as internet videos and images for which annotations are easily obtained either by text analysis or by manual entry from users, annotation is major challenge in our application domain. Correct identification of the different types of discrete Orca calls requires training and in the most difficult cases can only be performed by an expert. The main challenge that motivated the work described in this paper has been to obtain high quality annotations for this large specialized audio archive. We have followed a multi-pronged strategy in order to address this challenge: using a client-cloud web-based collaborative interfaces we can utilize volunteers from all around the world to perform the annotation, using automatic similarity retrieval we assist the annotation process especially for inexperienced users without affecting the quality of the resulting annotations, finally we use the annotated data to build machine learning algorithms to further annotate the data.

Most of existing work in the automatic analysis for Orca calls has focused either on the automatic detection of calls either in real-time [4] or offline [5] or on their classification [6]. In contrast our focus is on similarity-based retrieval. There are several reasons why retrieval is more important than classification in our situation. Retrieval provides more fine grained information than classification and supports the study of variation within a particular call type. In addition it can deal with calls of an unknown type or with classes that have a very small number of examples. Users of our interface fall into two categories: experts and volunteers. In many cases volunteers might not be able to classify a call

978-1-4577-1434-4/11/$26.00 ©2011 IEEE

type by listening to it but can easily identify to which call it is more similar from a limited set of examples. That way the call can be indirectly classified. Similarity retrieval can provide this limited set of examples. There have been several strategies and representations proposed in the literature for the classification and retrieval of Orca calls, but to the best of our knowledge they have mostly been evaluated on small amounts of data and have not been compared directly on the same data using retrieval effectiveness metrics.

The main contributions of this work are: 1) a collaborative web-based interface that integrates automatic similarity retrieval to enhance the annotation process 2) a description of different strategies for the retrieval of Orca calls 3) an experimental evaluation of these different strategies on a large dataset using established retrieval effectiveness metrics and 4) a post-processing step for denoising and exact boundary identification for presentation of the Orca calls.

II. Related Work

The main motivation behind our work has been creating better interfaces for interacting with the Orchive [2] a large archive of audio recordings of Orca vocalizations. There are stable resident populations of *Orcinus orca* in the northwest Pacific Ocean, and some of these populations [3]] 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 [7]. They have amassed a huge archive of more than 20,000 hours of audio recordings collected via a permanent installation of underwater hydrophones.

Most of existing work in the automatic analysis of Orca calls has focused on detection and classification rather than retrieval. A real-time system with low computational requirements for the detection of Orca vocalizations is described in [4]. Annotation bootstrapping is a technique used to classify/segment hydrophone recordings into three broad categories: voiceover, background, and vocalizations [5].

Our work was influenced by two publications that described different representations and methodologies for the classification of Orca calls. Orca vocalizations consists of well-defined, discrete calls with tonal signal components. They can be characterized by the pulse rate contour of the goal which can be viewed as analogous to the pitch contour of a speech or monophonic music signal. A method for computing acoustic similarity between pulse rate contours (normalized so that they all have the same duration) using the discriminative error of a Artificial Neural Network is described in [7].

Dynamic time warping (DTW) is a technique for measuring the similarity of two sequence that many vary in time. It is mostly known in the context of speech recognition [8] but it has found applications in many areas including video, motion and DNA sequence analysis. The use of DTW to compute the similarity between two pulse rate contours in the context of Orca calls has been explored in [6]. In that work, a similarity matrix is calculated containing all the DTW alignment costs between pairs of pulse rate contours. This similarity matrix is then subsequently used to calculate clusters which are then compared the ground truth call labeling to assess the feasibility of call classification using this approach. Figure II shows two similarity matrices and the corresponding alignments betwen two instances of the same call and two different calls. As can be seen the alignment of the two calls of the same type is much closer to a diagonal and the total score is lower.

In this paper we focus on retrieval rather than classification and only use the ground truth labels as a way to measure retrieval effectiveness. We compare different strategies over a large dataset (185 calls, 4 classes) using well established retrieval effectiveness measures. To the best of our knowledge this is the first systematic evaluation of these different design choices over a data set that is significantly larger than the ones used in the existing literature. It also the first time a full collaborative web-based client-cloud system has been developed to support research in bioacoustics which is typically performed individually using desktop applications.

III. System Description

A. Contour Extraction and Retrieval Strategies

The discrete calls of killer whales are pulsed signals in which a tone (of a certain tonal frequency) is not emitted continuously but in pulses given by the pulse-repetition rate. Unlike the tonal signals of many birds and other delphinids, the highest energy is not always contained in the first or second harmonic [7]. The high levels of background noise and variety of recording conditions compound the difficulty of obtaining pulse rate contours. The pulse rate contour is used as the primary representation for Orca calls because it is more robust as compared to spectral features to levels of background noise typical in field recordings.

We have compared three pitch extraction methods for obtaining the pulse rate contour. The first method (**PRAAT**) is based on time-domain autocorrelation and is similar to the pitch extraction algorithm implemented in Praat [9]. It is based on calculating the time-domain autocorrelation of the signal:

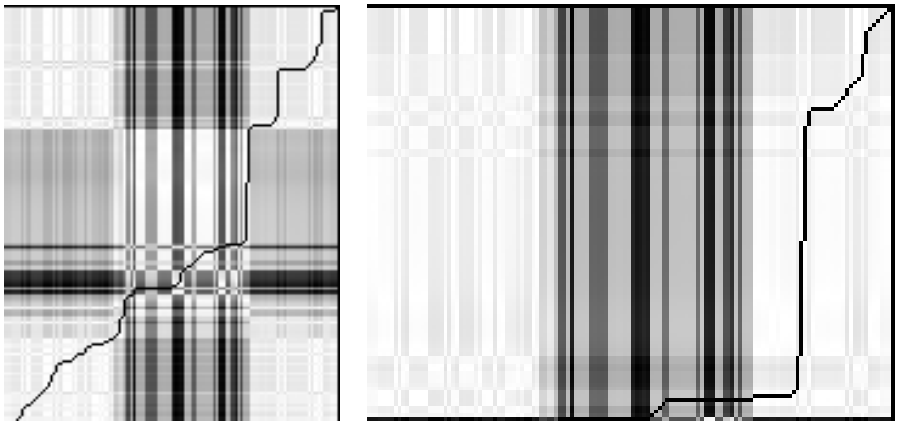
$$R(\tau) = \frac{1}{N} \sum_{n=0}^{N-1-m} x[n]x[n+m] \quad 0 \leq m < M \qquad (1)$$

The peaks of the autocorrelation function correspond to the lags in which the signal is self-similar. The signal is processed in windows and the autocorrelation of the windowed signal R_{xw} is divided by the autocorrelation of the window R_w providing better robustness to noise and better accuracy.

$$R_x(\tau) = R_{xw}(\tau)/R_w(\tau) \qquad (2)$$

The second method is based on the **YIN** pitch extraction method. The YIN method is based on the difference function which is similar to the autocorrelation:

$$dt = \sum_{n=0}^{N-1} (x[n] - x[n+\tau])^2 \qquad (3)$$



(a) N1 versus N1 (b) N1 versus N4

Fig. 1. A representation of the output of the Dynamic Time Warping (DTW) algorithm which is used to calculate the path of best similarity between two sequences of pitch values. Shown here are similarity matricies, one is between two different instances of the N1 call, and the other is the N1 call compared to an N4 call. The black line shows the path of best match as determined by the DTW algorithm. The N1 call compared to itself had a DTW score of 0.60333, while the N1 call compared to the N4 call had a worse DTW score of 3.35097.

The dips in the difference function correspond to periodicities. In order to reduce the occurrence of subharmonic errors, YIN employs a cumulative mean function which deemphasizes higher period dips in the difference function.

The third method (**SACF**) is based on the multipitch detection algorithm described by Tolonen and Karjalainen [10]. In this algorithm, the signal is decomposed into two frequency bands (below and above 1000 Hz) and amplitude envelopes are extracted for each frequency band. The envelope extraction is performed by applying half-wave rectification and low-pass filtering. The envelopes are summed and an enhanced autocorrelation function is computed so that the effect of integer multiples of the peak frequencies to multiple pitch detection is reduced.

We explore three retrieval strategies/representations. Statistical features characterizing the entire pulse rate contour are computed and each call is characterized by a single vector of features. The features are normalized by max-min normalization so that they range from 0 to 1 over the entire dataset. Similarities are then computed by taking the Euclidean distance in the normalized space. This strategy is used as reasonable baseline. The features used in this work are the mean, median, standard deviation, min and max of the pulse rate contour. The second strategy consists of resampling the pulse rate contour using linear interpolation to a fixed number of points. This strategy is similar to the one used in Deecke [7]. Essentially it assumes that the duration of the call does not play a major role in its characterization and temporal scaling is applied uniformly across the contour. The third strategy utilizes dynamic time warping to align the pulse rate contours. The alignment cost is used to measure the similarity between calls. The two sequences to be matched are arranged on the sides of a grid. To find the best match between the sequences we can find a path through the grid that minimizes the total distance between them. More details can be found in [8].

B. Collaborative web-interface

Collaborative web-based interfaces have significant advantages compared to desktop applications especially in the context of research in bioacoustics. In the most common traditional setup each researcher has access to their own data collection and performs annotations using desktop applications such as audio editors and general purpose software applications such as spreadsheets. In such a context it is hard to leverage annotations from multiple users or share large datasets such as the Orchive. By moving into a client-cloud mode of interaction we are able to have several users from around the world access the Orchive simultaneously and can store and analyze their annotations centrally. The entire archive is also regularly backed up and replicated as more files are digitized and more annotations are added. The flexibility of the architecture enables the easy creation of new interfaces such as the mobile game client described later in this paper.

Figure III-A shows a screen-shot of the collaborative web-interface. The user has selected a region corresponding to call that needs to be annotated. Under the spectrogram a call catalog is provided to help the user with annotation. The interface uses the similarity retrieval described in this paper to suggest the most likely call type to the user by showing it zoomed in the window to the right. The call is then correctly annotated as of N4 type. The interface is written in the Django web development framework [1] and interfaces directly with the Marsyas open source audio analysis software [2] through Python bindings. The entire interface runs inside a web browser and can be accessed at http://orchive.cs.uvic.ca. Viewing is public but annotation is restricted to registered users.

[1]http://www.djangoproject.com/
[2]http://marsyas.info

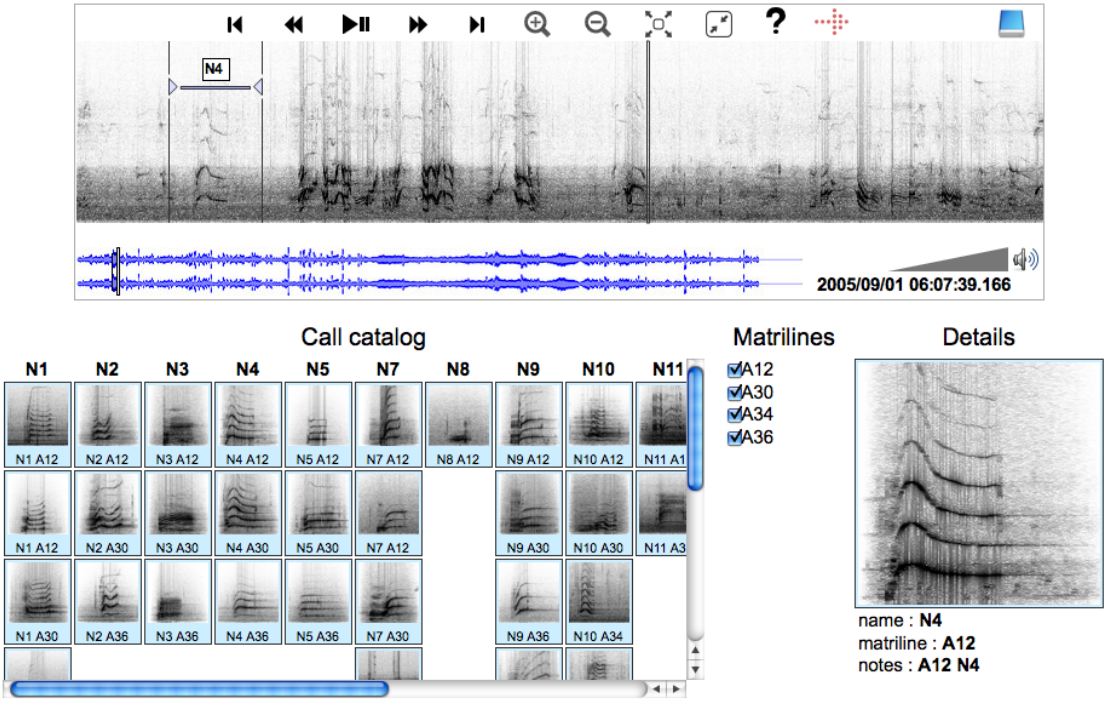


Fig. 2. Orchive collaborative web-interface showing how annotation can support retrieval. The automatically suggested call is shown on the right and can be used for immediate annotation if the user agrees that the selected call is of the same type after listening to it.

Fig. 3. Mobile client interface for Orca call annotation.

C. Mobile client

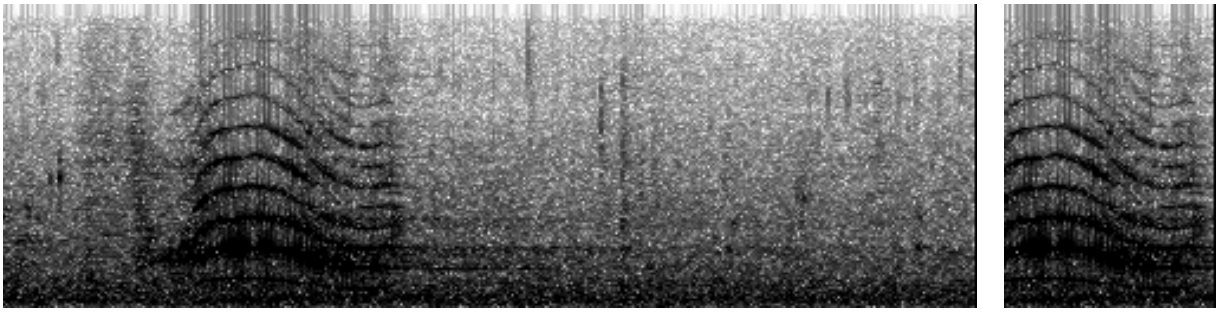
A game with a purpose (GWAP) [11] is a type of computer game that can be used to collect interesting data about a task while at the same time being entertaining to the players. The most well known example is the ESP game [12] in which two players attempt to assign labels to images and in the process provide a wealth of annotation data. We have developed a similar game for classifying Orca vocalizations. The player is presented with a call that needs to be annotated and the two closest examples from a labeled dataset based on content-based similarity retrieval with the constraint that the two examples come from different classes. The player then selects the call that is more similar and the label is sent to the database for storage. By probing the user with labeled samples we can assess their skill in annotation and decide whether to use their labels or not. The game and client is written for the Android mobile operating system.

D. Usage statistics

Since launch in July 2008, there have been just over 5000 visitors to the site from 74 countries. These users have viewed just under 30,000 pages, with an average of 5.73 pages per visit. Each visitor spent an average of 3 minutes 44 seconds on the website. 24 researchers have been granted annotation access to the Orchive, and amongst them they have entered 5708 annotations of orca calls using our web based interface. These annotations are on only about 1000 recordings, which means the vast majority of the 14862 recordings in the database have yet to be annotated. These recordings are 45 minutes long, which means a total of almost 700,000 minutes of data currently is in our archive, with more data being digitized constantly.

E. Post-processing

The annotation process by users is not very accurate and frequently includes noisy background parts of the signal. A spectral peak picking approach is used to identify the tonal



(a) Before (b) After

Fig. 4. Post-processing of Orca calls for presentation. The selected call on the left is processed in order to remove the noisy hydrophone parts preceding and following it as well as enhancing the tonal components.

components of the input signal. These components, separated from the noise-like signal parts are resynthesized to provide a more clear sounding call that still retains its identifying characteristics. A result of this denoising compared to traditional denoising based on spectral profiling can be found and heard at: http://orchive.cs.uvic.ca/main/tour.

In a first processing step, peaks are pre-selected as candidates for tonal components by ensuring that a) their amplitude is the largest local maximum within small blocks of neighboring frequency bins, b) their frequency is in a pre-defined frequency range and c) their amplitude is above a threshold that is computed relative to the overall spectral maximum, the overall spectral energy and a smoothed version of the magnitude spectrum. Then, a "tonal probability" is computed for every peak candidate and the candidates with the highest probability are selected as tonal components. The probability is computed as the arithmetic mean of a) the Gaussian of the distance of the peak's bin frequency and its instantaneous frequency, b) the normalized logarithmic distance between the peak's amplitude and a simultaneous masking threshold computed according to the psycho-acoustic model in ITU recommendation BS.1387, and c) the amplitude of a peak spectrogram smoothed in both time and frequency domain so that sporadic peaks have low amplitude and "steady" peaks have higher amplitude. This tonal probability is also used to remove the parts of the annotation that contain background noise as shown in Figure III-B. It is hard to evaluate quantitatively the effectiveness of the post-processing step as there is no ground-truth for what the clean Orca calls would sound like. In all the cases that we tried the boundaries of the call were correctly identified and the perceived quality of the call is much better.

IV. Experimental Evaluation

In order to systematically explore the different strategies for Orca call retrieval we utilized a dataset consisting of 185 recordings of vocalizations. They have been annotated using the Orchive collaborative user interface and classified into 4 discrete call types by volunteers. The ground truth labels have been verified by experts. Table I shows the composition of the dataset used for evaluation. We use two established evaluation metrics that measure the retrieval effectiveness. Precision at 1 is simply the number of queries for which the first retrieved call has the same class as the query. The mean average precision (MAP) is the most frequently used summary measure of a ranked retrieval run. Average precision of a single query is the mean of the precision scores after each relevant document has been retrieved. The value for the run (a set of queries) is the mean of the individual average precision scores. MAP combines aspects of both precision and recall and rewards returned relevant items higher in the list.

TABLE I
Dataset composition and MAP scores for best configuration (Hertz frequency scale, SACF pitch extractor and DTW matching)

Call Type	N1	N3	N4	N47
Instances	36	56	60	33
MAP	0.63	0.94	0.78	0.58

Table I shows the best MAP scores achieved for each type of call. As can be seen there is large variance in the MAP score for different types of calls. For example retrieval of N3 calls is very robust but retrieval of the N47 calls is not as much. These differences are also observed in the human classification of these calls.

Table II shows the MAP scores and average precision score at 1 over the entire dataset for combinations of different representations and pulse rate extraction strategies. As can be seen, the SACF pitch extractor is the best performing pitch extraction method independently of the retrieval strategy. The DTW matching is also the best performing retrieval strategy. It is hard to draw any conclusions with respect to SACF performing better than the other two pitch extractors. The better results obtained using the DTW retrieval strategy indicate there is important non-uniform timing variation in the structure of these calls.

We have also conducted experiments with different frequency scale representations such as the Bark-scale [13] and logarithmic frequency but in all configurations they performed worst than the default linear frequency representation in Hertz.

TABLE II
MEAN AVERAGE PRECISION SCORES FOR DIFFERENT PITCH EXTRACTION AND RETRIEVAL STRATEGIES

	Features	Contour	DTW
PRAAT	0.38	0.52	0.67
YIN	0.50	0.51	0.72
SACF	0.63	0.66	0.77

TABLE III
AVERAGE PRECISION AT 1 SCORES FOR DIFFERENT PITCH EXTRACTION AND RETRIEVAL STRATEGIES

	Features	Contour	DTW
PRAAT	0.38	0.40	0.4
YIN	0.77	0.72	0.95
SACF	0.79	0.82	0.95

V. CONCLUSIONS AND FUTURE WORK

We describe a web-based collaborative web interface for retrieval and annotation of a large archive of Orca vocalizations. A number of different strategies for pulse rate extraction and retrieval were experimentally compared. Excellent results were obtained by a combination of a computationally efficient pitch extractor based on the summary autocorrelation function and a matching strategy based on dynamic time warping. The best configuration achieves a mean average precision of 0.77 and the first retrieved call is relevant 95% of the time. In the future we plan to expand our dataset both in terms of instances and call types. In addition we plan to investigate more thoroughly different design choices especially in the pitch contour extraction possibly taking into account information about the hearing system of Orcas [14]. Finally we plan to expand our prototype annotation game with various types of challenges. Although the community of users interested in Orca vocalizations is small their are passionate and eager to help.

VI. ACKNOWLEDGMENTS

We would like to thank Paul Spong and Helena Symonds of Orcalab for providing the data and inspiration for this project. We would also like to thank the National Sciences and Engineering Research Council (NSERC) for their financial support. This work was partly supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

REFERENCES

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

[2] 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.

[3] 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*. Vancouver: UBC, 2000.

[4] J. Luke, J. Marichal-Hernandez, F.Rosa, and J.Almunia, "Real time automatic detection of orcinus orca vocalizations in a controlled environment," *Applied Acoustics*, vol. 71, no. 8, pp. 771 – 776, 2010.

[5] S. Ness, M. Wright, L. G. Martins, and G. Tzanetakis, "Chants and orcas: semi-automatic tools for audio annotation and analysis in niche domains," in *Proceeding of the 2nd ACM workshop on Multimedia semantics*, ser. MS '08, 2008, pp. 9–16.

[6] J. C. Brown and P. J. Miller, "Automatic classification of killer whale vocalizations using dynamic time warping," *J. Acoust. Soc. Am.*, vol. 122, pp. 1201–1207, Aug 2007.

[7] J. F. V.B. Deecke and P. Spong, "Quantifying complex patterns of bioacoustic variation: use of a neural network to compare killer whale (orcinus orca) dialects," *Journal of the Acoustical Society of America*, vol. 105, pp. 2499–2507, 1999.

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.

[9] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound." vol. 17, pp. 97–110, 1993.

[10] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 708 –716, Nov. 2000.

[11] L. von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92 –94, june 2006.

[12] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '04, 2004, pp. 319–326.

[13] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *J. Acoust. Soc. Am.*, vol. 33, 1961.

[14] S. Nummela, T. Reuter, S. Hemil?, P. Holmberg, and P. Paukku, "The anatomy of the killer whale middle ear (Orcinus orca)," *Hear. Res.*, vol. 133, pp. 61–70, Jul 1999.