Classification of vocalizations of killer whales using dynamic time warping

Judith C. Brown
Physics Department, Wellesley College, Wellesley, Massachusetts 02481 and Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139
brown@media.mit.edu

Andrea Hodgins-Davis
Biology Department, Wellesley College, Wellesley, Massachusetts 02481
ahodgins-davis@mbl.edu

Patrick J. O. Miller
Sea Mammal Research Unit, University of St. Andrews, St. Andrews, Fife KY16 9QQ, United Kingdom
pm29@st-andrews.ac.uk

Abstract: A large number of killer whale sounds have recently been classified perceptually into Call Types. [A. Hodgins-Davis, thesis, Wellesley College (2004)]. The repetition rate of the pulsed component of five or more examples of each call type has been calculated using a modified form of the FFT based comb-filter method. A dissimilarity or distance matrix for these sounds was calculated using dynamic time warping to compare their melodic contours. These distances were transformed into a component space using multidimensional scaling and the resulting points were clustered with a kmeans algorithm. In grouping 57 sounds into 9 call types, a single discrepancy between the perceptual and the automated methods occurred.

PACS numbers: 43.80.Ka [CFM]
Date Received: October 10, 2005 **Date Accepted:** December 19, 2005

1. Introduction

Marine mammals produce a wide range of vocalizations, and an improved ability to classify recorded sounds could aid in species identification or in tracking movements of animal groups. In the case of killer whales, time-frequency contours of stereotyped pulsed calls are group-specific from matrilineal lines (group with same mother) to pods (living group consisting of a number of matrilineal lines) to clans (larger groups sharing calls) (Ford, 1991; Miller and Bain, 2000). There are a number of call types within these groups, which are thought to be learned in the pod, and for the most part, these types have been classified by humans from listening and looking at their spectra. For killer whale sounds classification by eye and ear is quite consistent (Yurk *et al.*, 2002), and this type of classification has been useful to reveal group-specific acoustic repertoires and matching vocal exchanges (Ford, 1991; Miller *et al.*, 2004).

It would, nonetheless, be useful to replace human classification with an automatic technique because of the large amounts of data to be classified, and the fact that automatic methods can be fully replicated in subsequent studies. In the last decade, there have been attempts at automatic classification, including a single dynamic time warping study (DTW). Buck and Tyack (1993) took three signature (specific to the individual) whistles from each of five dolphins and used DTW to compare to five reference whistles, one from each of the five dolphins. Correct identification was achieved for all 15 of the signature whistles.

Killer whales produce three forms of vocalizations: clicks, whistles, and pulsed calls. Clicks consist of an impulse train (series of broadband pulses); whistles consist, for the most part, of a single sinusoid with varying frequency; and pulsed calls are more complex sounds

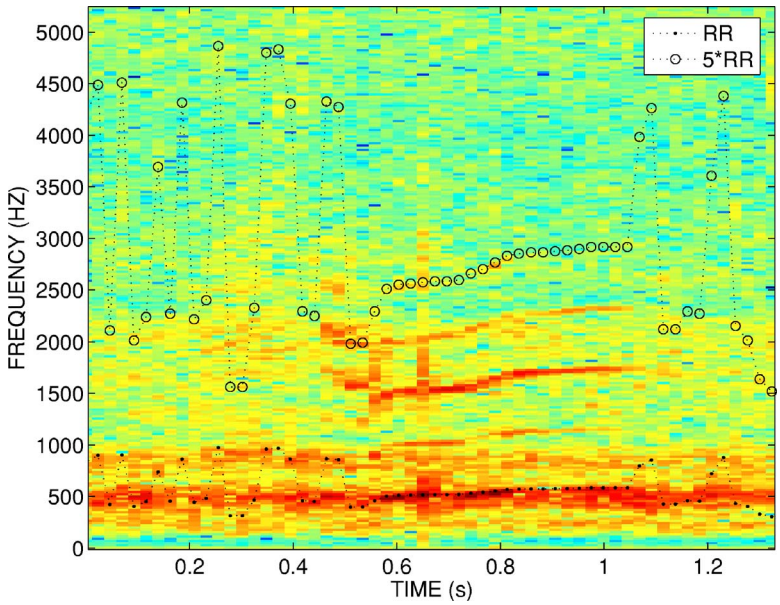

FIG. 1. (Color online) Spectrogram of a sound with fundamental frequency (RR for repetition rate) and fifth harmonic (5*RR) displayed. This sound has a fundamental frequency which falls in a strong band of noise. Noise alone can be seen before and after the call.

with many harmonics. The repetition rate is a measure of the periodicity of the signal, and its measurement is called "pitch tracking" or "fundamental frequency tracking" in the speech literature.

Our classification of pulsed whale sounds is a much more difficult problem than the previous work using whistles. There are many more calls here, and these pulsed calls are far more complex than whistles which generally consist of a single harmonic eliminating the difficult step of obtaining the repetition rate. No reference calls are used for our study; all sounds are clustered with no assumptions other than the number of types previously classified perceptually.

2. Repetition rate measurement

A large number of sounds from the Captive Killer Whale Population at Marineland of Antibes, France were recorded using an HTI hydrophone directly to a computer hard-drive. These were recently classified into call types by listening to the calls and examining their spectrograms (Hodgins-Davis, 2004). Five or more examples of each of nine perceptually identified call types of these killer whale vocalizations were chosen for this study. These sounds were among those with the highest signal to noise ratio estimated visually from their spectrograms.

Repetition rates based on calculation of the autocorrelation function (Brown and Zhang, 1991) and the constant Q transform (Brown, 1992) were measured initially since these have proven accurate for fundamental frequency tracking of musical sounds. A third method (Brown *et al.*, 2004), called the comb filter method, proved more effective in dealing with the principal difficulty, which is the noise of the water circulation pump in some frequency bands. The comb-filter method is FFT-based and involves adding up a variable number of Fourier components with a fixed spacing (Hess, 1983); it can be adapted to accommodate several input parameters:

(1) The upper and lower range of frequencies which are searched as possible repetition rate candidates.

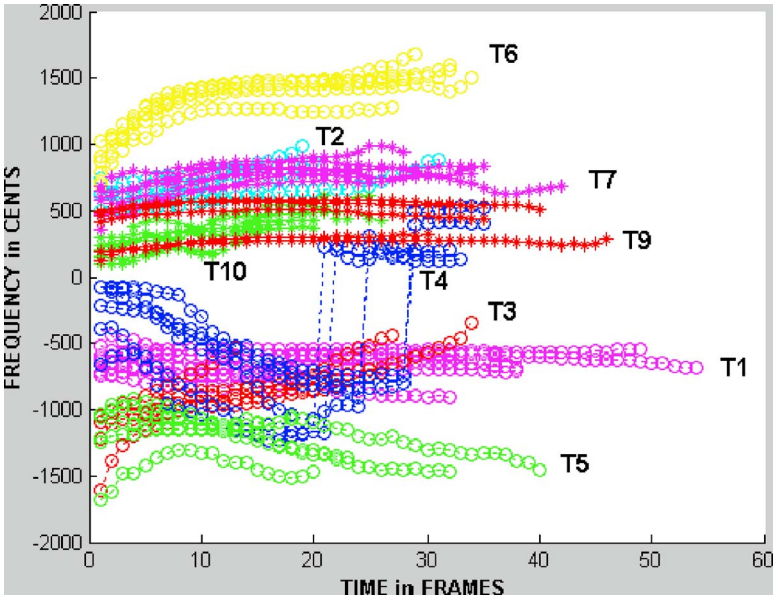

FIG. 2. (Color online) Fundamental frequency contours in cents for each of the sounds with perceptual grouping indicated. The total time for 60 frames is 1.4 s. Frequencies in Hz range roughly from 185 Hz (−1500 cents) to 1050 Hz (1500 cents). This figure is much clearer in color in the online publication.

(2) The lowest frequency at which the comb is fit. This made it possible to jump over the pump noise which in general was below around 500 Hz.
(3) The number of "teeth" (harmonics) in the comb.

The FFT was calculated with a frame size of 46.4 ms (window size of 1024 at sample rate 22050) and hop size of 23.2 ms giving 50% overlap and a frequency resolution of 21.6 Hz. The FFT was upsampled by a factor of 10 prior to applying the "comb" allowing identification of peak position to within 2.2 Hz. Parameters were adjusted for each of the call types, and the calls were extracted manually.

An example of these results is found in Fig. 1 with audio in Mm. 1. The noise in this figure before the call starts can be easily identified. The pitch track is graphed for both the fundamental and a higher harmonic, in this case the fifth. This was necessary for some of the calls with a fundamental frequency buried in the noise, such as this one, where only the higher harmonics could be seen clearly. It is also a much better visual indicator of errors for a low frequency fundamental where a small error in the fundamental is difficult to see, but is multiplied by the harmonic number for the higher harmonic and becomes more evident.

Mm. 1 Audio file in wav format (60 kb) for the example shown in Fig. 1.

A summary figure including all fundamental frequency contours for all types is given in Fig. 2. Frequencies are measured in cents defined as

$$f_{\text{cents}} = 1200 \cdot \log_2(f/f_{\text{ref}}), \tag{1}$$

where we have chosen the reference frequency f_{ref} as 440 Hz. This unit is commonly used for musical frequencies, and gives a 100 cent increase for each semitone of the equal tempered scale which is roughly 6% of the frequency. This is a geometric measure in that equal percentage changes are transformed into equal additive differences. The advantage is that a given melody has the same shape independent of its initial frequency. For example, consider doubling the frequency from 100 Hz to 200 Hz compared to 400 Hz to 800 Hz. In each of these cases the change measured in cents is 1200.

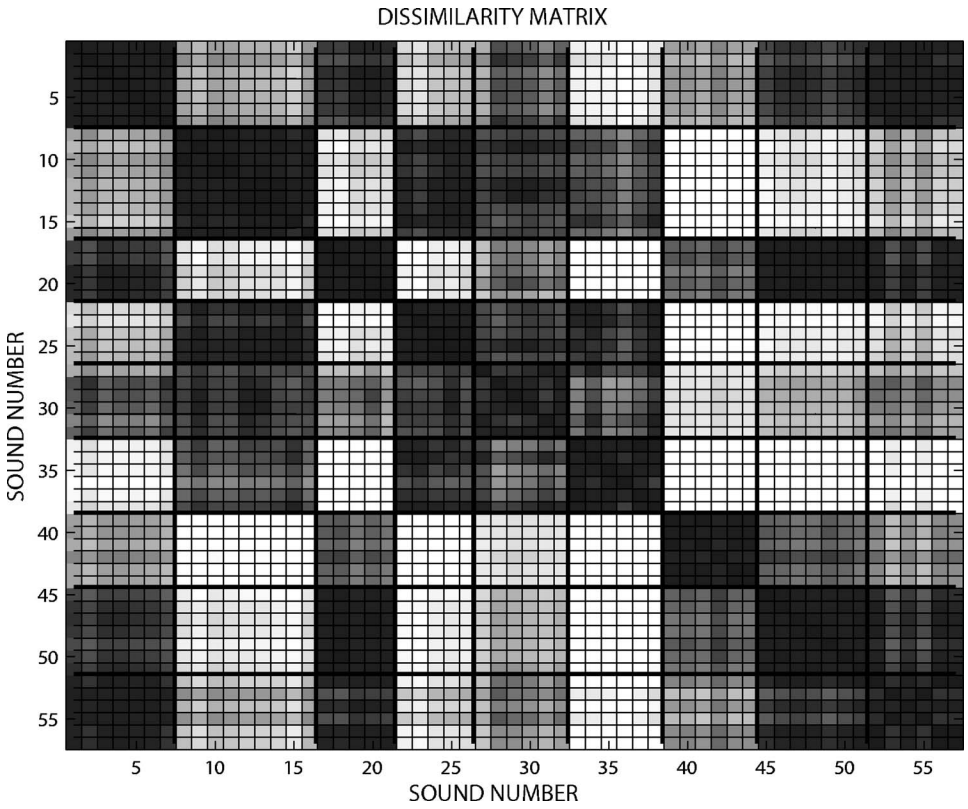

FIG. 3. Dissimilarity matrix calculated with dynamic time warping for fundamental frequency contours of Fig. 2. The sound blocks are in the order T10, T1, T2, T3, T4, T5, T6, T7, T9 as labeled in Fig. 2.

In Fig. 2 some call types are well-separated in frequency, while others overlap considerably. This figure distinguishes between the types much more clearly in color in the online publication than in black and white.

3. Dynamic time warping

The pitch contours were compared pairwise using the dynamic time warping (DTW) method described in Chai and Vercoe (2003). See also Rabiner and Juang (1993) for a thorough discussion of methods of dynamic time warping. The frequency values in cents from Eq. (1) were divided by 100 to retain the logarithmic measure of frequency but to have one or two digit differences. This facilitated a quick grasp of the numerical values involved for choosing parameter values.

The fundamental frequencies of all possible pairs of sounds were compared number by number. We will refer to a typical pair as sound 1 and sound 2. A scoring matrix was constructed using the algorithm (Chai and Vercoe, 2003):

$$M[1,1] = 0,$$

$$M[1,j] = M[1,j-1] + b,$$

$$M[i,1] = M[i-1,1] + a,$$

and

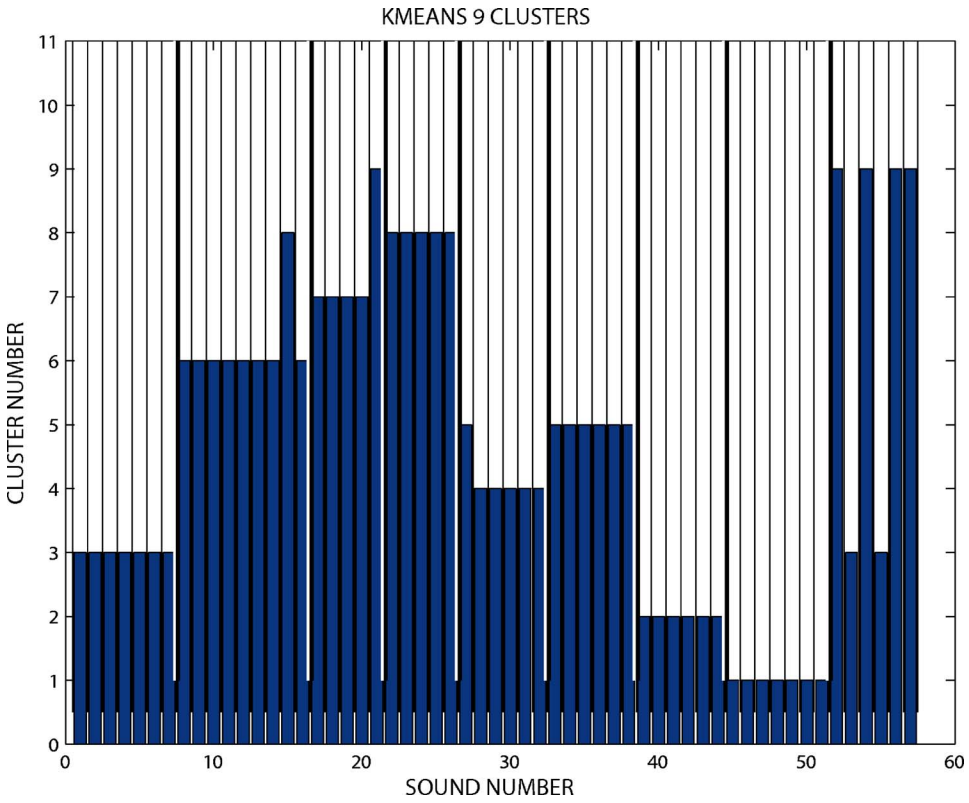

FIG. 4. (Color online) Clustering results for dissimilarity matrix for differences in fundamental frequency contours alone.

$$M[i,j] = \min(M[i-1,j] + a, M[i,j-1] + b, M[i-1,j-1] + D), \tag{2}$$

where $i,j \geqslant 2$ and D is the difference in the jth frequency of sound 1 and the ith frequency of sound 2. Here a is the cost of an insertion and b is the cost of deletion.

The resulting score is the lowest value in the last row normalized by dividing by the length of the shorter sound. The matrix of scores with one score tabulated for each pair of sounds is shown in the dissimilarity (or distance) matrix of Fig. 3. For identification purposes, blocks which were grouped in the previous perceptual classification are outlined by bold lines. Perfect agreement with the perceptual classification, would give dark (low number for small distance) in the blocks along the diagonal (comparisons within the same perceptual call type), and white elsewhere, corresponding to large differences between dissimilar call types. Dark blocks off the diagonal indicate call types which are similar and difficult to separate in classification.

4. Results

4.1 Experiment 1

Adjustable parameters in Eq. (2) for the calculation of the dissimilarity matrix of Fig. 3 are the costs a and b of insertions and deletions. The deletion cost corresponds to a difference in the duration of a call rather than the frequency contour so b was chosen to be zero. An initial guess of 15 was made for the insertion cost, and this turned out to be very near optimum for matching the perceptual results. This value was twice the average deviation from the mean of the numbers for the melodic contours.

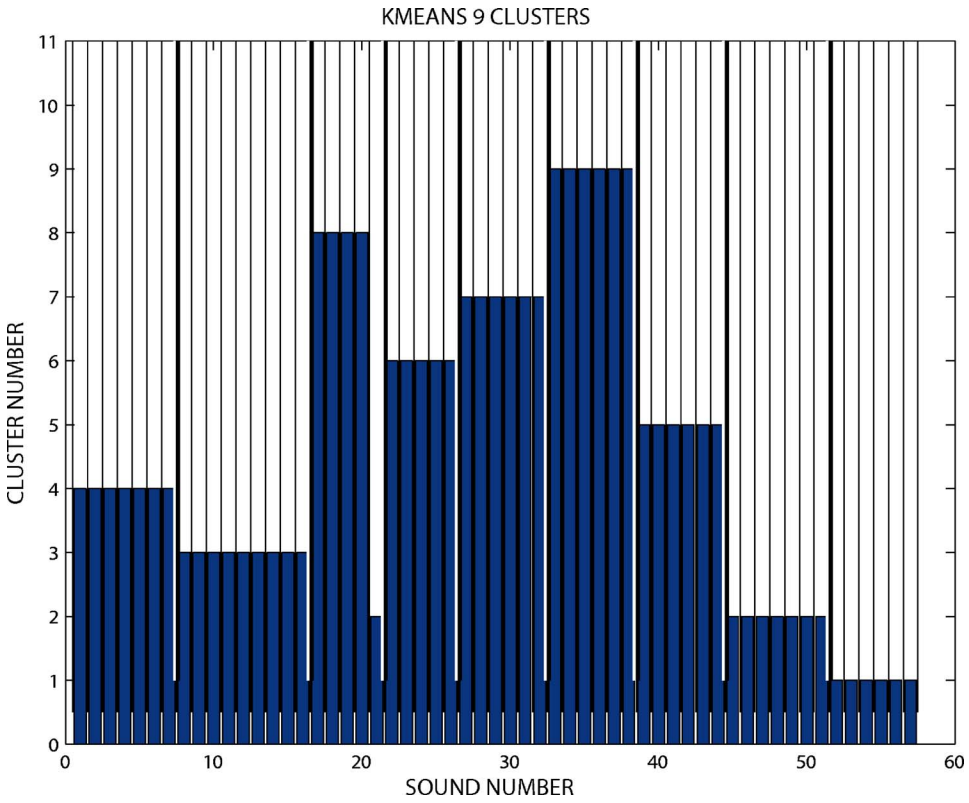

FIG. 5. (Color online) Clustering results for sum of dissimilarity matrices for differences in fundamental frequency contours and for derivatives (point to point differences) of the frequency contours shown in Fig. 3.

To obtain clusters from the distance values in the dissimilarity matrix, we first used multidimensional scaling (MDS) with matlab function *cmdscale*. MDS is a tool usually used for dimensionality reduction and/or data visualization (Kruskal and Wish, 1978). For example one can take perceptual differences in sounds such as those generated by musical instruments, transform them with MDS into points in a space with as few dimensions as possible, and then attempt to associate those dimensions with acoustical correlates. The mathematics of this transformation from distances to a coordinate representation is presented clearly in Borg and Groenen (1997).

For our purposes MDS was used to transform the dissimilarity values into positions in a Euclidian-type space; they could then be grouped using a clustering algorithm, implemented in this case with matlab function *kmeans*.

The number of output clusters was chosen to be 9 corresponding to the number of perceptual clusters. Clustering algorithms are nondeterministic as they depend on random initial conditions, and risk finding local rather than global minima. To insure that the global (true) minimum is found, a large number of runs must be made. Out of 100 total runs, in over 90 of these, the number of errors (compared to perceptual results) was 5 or 7 out of a total of 57 sounds. A typical classification can be found in Fig. 4, which is a bar graph of the cluster assignment vs sound numbers. Any of the results with 5 errors would be equivalent to this one but would have different cluster assignments on the vertical axis. These numbers are not significant except in showing that different call types are separated into different clusters.

The vertical boldface black lines show the breaks in perceptually defined types. For example the first perceptual group consists of sounds 1–7 and these are assigned to the same cluster with no errors; this means that all sounds judged by a human to belong to the same call

type are classified by DTW differences into the same cluster. There is one error each in the second, third, and fifth perceptual groups. The last perceptual group has two of its members assigned to the first group for two errors; so in this example there are five errors overall for an 88% agreement with perceptual results.

The DTW parameter a was varied to determine its effect on the results. If it is roughly halved from 15 to 8, the errors increase slightly to 8, but also there is mixing of types 2 and 7. From Fig. 2 it can be seen that these curves are very similar.

4.2 Experiment 2

The DTW distances as calculated for Fig. 3 weigh the similarity of absolute frequencies more than the shape of the contour. The derivative of these curves, calculated as point to point differences of the contour, is a measure of the shape alone as the absolute frequency is subtracted out. To combine these effects a dissimilarity matrix for the derivatives was calculated using the same procedure as described for the contours. Since these numbers for the differences are roughly a factor of 10 (ratio of their standard deviations is 9.8) smaller than those for the absolute frequencies, this matrix was multiplied by 10 and added to the matrix of Fig. 3. This sum takes both absolute frequencies and contours into account. Multidimensional scaling and kmeans calculations were performed as before. In over 80% of the runs, the results shown in Fig. 5 give a single sound differing from the perceptual classification for essentially perfect agreement.

5. Conclusions

Dynamic time warping has proven very successful for the automatic classification of killer whale vocalizations with the limitation that the determination of the melodic contours of the input sounds can be time-consuming. Further testing with diverse call repertoires recorded under natural conditions is planned in order to determine the full potential of this technique.

Acknowledgements

J.C.B. is very grateful to Wei Chai, who provided her matlab code for DTW, plus much helpful advice on its use and on dynamic programming in general. Thanks to Marineland, Antibes for support installing and using the pool hydrophones. Funding was provided by WHOI's Ocean Life Institute and a Royal Society fellowship to P.J.O.M.

References and links

Borg, I., and Groenen, P. (**1997**). *Modern Multidimensional Scaling. Theory and Applications*, Springer, New York.

Brown, J. C., and Zhang, B. (**1991**). "Musical frequency tracking using the methods of conventional and 'narrowed' autocorrelation," J. Acoust. Soc. Am. **89**, 2346–2354.

Brown, J. C. (**1992**). "Musical fundamental frequency tracking using a pattern recognition method," J. Acoust. Soc. Am. **92**, 1394–1402.

Brown, J. C., Hodgins-Davis, A., and Miller, P. J. O. (**2004**). "Calculation of repetition rates of the vocalizations of killer whales," J. Acoust. Soc. Am. **116**, 2615.

Buck, J. R., and Tyack, P. L. (**1993**). "A quantitative measure of similarity for Tursiops truncatus signature whistles," J. Acoust. Soc. Am. **94**, 2497–2506.

Chai, Wei, and Vercoe, Barry (**2003**). "Structural analysis of musical signals for indexing and thumbnailing," in Proceedings of ACM/IEEE Joint Conference on Digital Libraries.

Ford, J. K. B. (**1991**). "Vocal traditions among resident killer whales Orcinus orca in coastal waters of British Columbia," Can. J. Zool. **69**, 1454–1483.

Hess, W. (**1983**). *Pitch Determination of Speech Signals: Algorithms and Devices*, Springer-Verlag, Berlin.

Hodgins-Davis, A. (**2004**). "An analysis of the vocal repertoire of the captive killer whale population at Marineland of Antibes, France," thesis, Wellesley College.

Kruskal, J. B., and Wish, M. (**1978**). *Multidimensional Scaling*, Sage Publications, Beverly Hills, CA.

Miller, P. J. O. and Bain, D. E. (**2000**). "Within-pod variation in the sound production of a pod of killer whales, Orcinus orca," Anim. Behav. **60**, 617–628.

Miller, P. J. O., Shapiro, A. D., Tyack, P. L., and Solow, A. R. (**2004**). "Call-type matching in vocal exchanges of free-ranging resident killer whales, Orcinus orca," Anim. Behav. **67**, 1099–1107.

Rabiner, L., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition*, Prentice-Hall, New Jersey.

Yurk, H., Barrett-Lennard, L., Ford, J. K. B., and Matkin, C. O. (**2002**). "Cultural transmission within maternal lineages: vocal clans in resident killer whales in southern Alaska," Anim. Behav. **63**, 1103–1119.

