Automatic classification of killer whale vocalizations using dynamic time warping

Judith C. Brown[a)]
Physics Department, Wellesley College, Wellesley, Massachusetts 02481 and Media Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Patrick J. O. Miller[b)]
Sea Mammal Research Unit, University of St. Andrews, St. Andrews, Fife KY16 9QQ, United Kingdom

(Received 15 November 2006; revised 30 April 2007; accepted 14 May 2007)

A set of killer whale sounds from Marineland were recently classified automatically [Brown *et al.*, J. Acoust. Soc. Am. **119**, EL34–EL40 (2006)] into call types using dynamic time warping (DTW), multidimensional scaling, and kmeans clustering to give near-perfect agreement with a perceptual classification. Here the effectiveness of four DTW algorithms on a larger and much more challenging set of calls by Northern Resident whales will be examined, with each call consisting of two independently modulated pitch contours and having considerable overlap in contours for several of the perceptual call types. Classification results are given for each of the four algorithms for the low frequency contour (LFC), the high frequency contour (HFC), their derivatives, and weighted sums of the distances corresponding to LFC with HFC, LFC with its derivative, and HFC with its derivative. The best agreement with the perceptual classification was 90% attained by the Sakoe-Chiba algorithm for the low frequency contours alone. © *2007 Acoustical Society of America.* [DOI: 10.1121/1.2747198]

PACS number(s): 43.80.Ka [WWA] Pages: 1201–1207

I. INTRODUCTION

Marine mammals produce a wide range of vocalizations, and an improved ability to classify recorded sounds could aid in species identification as well as in tracking movements of animal groups. For the most part, the sounds produced by killer whales have been classified by humans into groups called "call types" from listening to the calls and observing their spectra. For killer whale sounds classification by eye and ear is consistent, and this type of classification has been useful to reveal group-specific acoustic repertoires and matching vocal exchanges (Yurk *et al.* 2002). 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 a previous study we examined a group of captive killer whale sounds recorded in Marineland in the French Antilles and consisting of nine call types with at least five examples in each (Brown *et al.* 2006). We found that dynamic time warping (DTW) gives an accurate measure of the dissimilarity of calls and were able to classify this set automatically with near-perfect accuracy. Here we extend this work with a larger group of whale sounds recorded on the open sea and examine the effectiveness of four different DTW algorithms. This set of sounds consists of biphonic (two simultaneous, independently modulated) calls of northern resident whales and contains over 100 calls previously classified perceptually into seven call types. This is the first automatic classification study using frequency contours of biphonic calls as well as the first full-length article on classification of marine mammal calls using DTW. Preliminary results were reported by Brown and Miller (2006a, b).

II. BACKGROUND

A. Killer whale vocalizations

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 with many harmonics. Among these pulsed calls are a number of highly stereotyped (repeated and recognizable) calls, which are thought to be learned within the pod or living group. Repertoires of these stereotyped calls are pod specific, and the time-frequency contours of shared stereotyped calls are also group specific from matrilineal lines (group with same mother) to larger pods (consisting of several matrilineal lines) to clans (larger groups sharing calls).

B. Fundamental frequency tracking and perceptual classification

One of the remarkable features of some northern resident killer whale pulsed calls is that they contain two overlapping but independently modulated contours or "voices" as shown in Fig. 1. Biphonation, as this is called, is common in birds but has been described for few marine mammal sounds (Tyson, 2006; Tyson *et al.*, 2006). One of the challenges of analyzing these complex sounds is to determine the fundamental frequency or to "pitch-track" these two components

[a)]Electronic mail: brown@media.mit.edu
[b)]Electronic mail: pm29@st-andrews.ac.uk

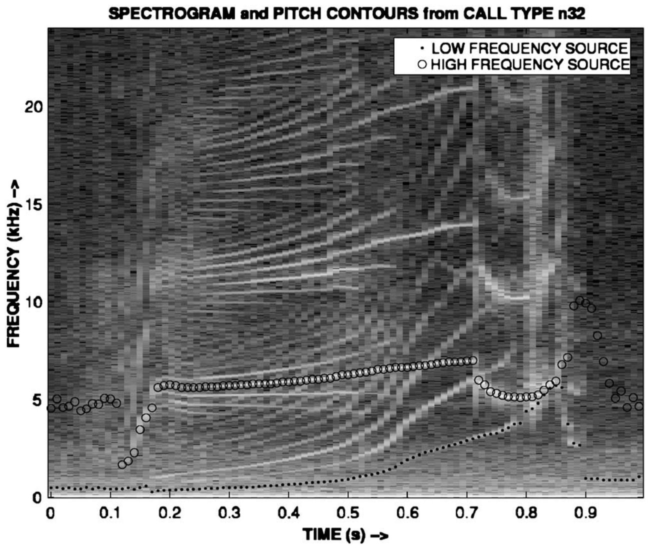

FIG. 1. Spectrogram showing pitch contours of the low frequency and high frequency sources in a killer whale pulsed call. Note there is noise before and after the onset of the calls.

from the same sound. See the example in Fig. 1 where the upper and lower frequency components are superposed on the spectrum.

Pitch tracking has had a long and abundant history in the speech literature (Hess, 1983). Some of these methods have proven successful for determining the repetition rate, or fundamental frequency, for pulsed killer whale sounds and have been described in Brown (1992), Wang and Seneff (2000), and Brown *et al.* (2004). The pitch contours of our northern resident group are arranged by perceptually determined call types in Fig. 2 for the low frequency contours and in Fig. 3 for the high frequency contours. As can be seen in these figures, the shapes of the contours within each group are similar though the lengths of the calls differ. The call types are graphed separately because of the considerable overlap in frequency range of several of the groups; this foreshadows difficulty for automatic classification.

C. Dynamic time warping

For automatic classification, a technique for quantitatively comparing curves of similar shape but different lengths is required. Dynamic time warping (DTW) is ideally suited to this task. It was used widely in the early days of speech recognition, and the different algorithms used by the speech community are described and evaluated in an excellent paper by Myers *et al.* (1980). See also Rabiner and Juang (1993). More recently DTW has been used for "query by humming" searches in musical information retrieval (Hu *et al.*, 2003).

For marine mammal sounds DTW was first used for the classification of 15 dolphin signature whistles into five groups by Buck and Tyack (1993). In the past year it has been applied to pulsed killer whale sounds by Deecke and Janik (2006) on a set of 20 calls in six categories, as well as our (Brown *et al.*, 2006) Marineland classification of 57 calls into nine call types. In the smaller sets of 15 and 20 calls, the contours within call types were virtually identical. While there was much more contour variation in the 57-call data set, these calls still separated sufficiently in absolute frequency to be identifiable on the same graph; this is not the case for our current, much larger set of calls.

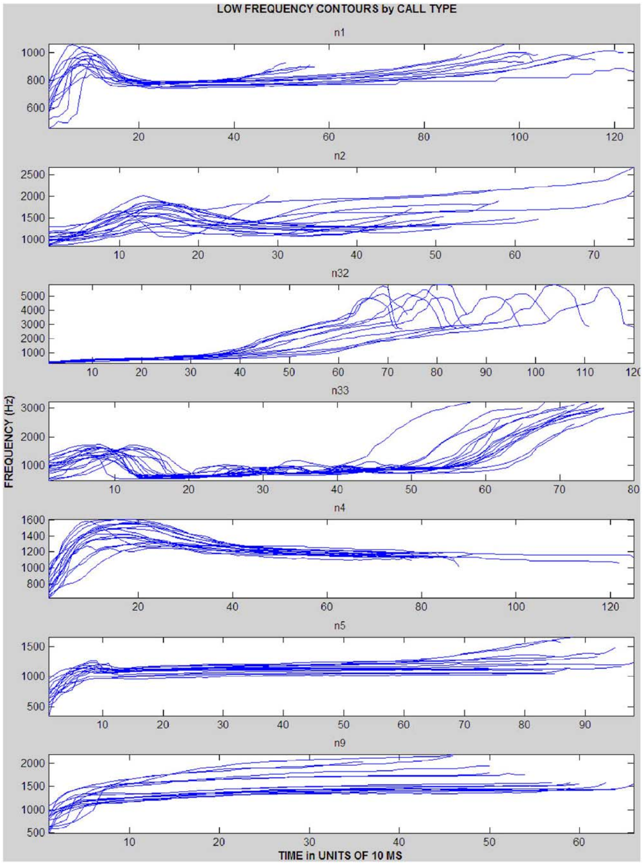

FIG. 2. (Color online) Pitch contours of the low frequency calls of the northern resident group of killer whales plotted in the perceptually designated call types. The calls are plotted separately since there is too much overlap to distinguish them if plotted on the same graph.

We have chosen four very different DTW algorithms, including the three used previously in the marine mammal studies mentioned above, for our current classification to determine their relative performance on this extremely challenging set of calls.

III. CALCULATIONS

A. Dynamic time warping (DTW) and contour dissimilarity

As an example of a DTW calculation, we consider two calls of different lengths, both from call type n32. By convention the shorter call is referred to as the query $\mathbf{Q}[i]$ and is aligned along a vertical axis, and the longer call is the target $\mathbf{T}[j]$ aligned horizontally as shown in Fig. 4. For all algorithms the first step is to construct a difference matrix where each element $\mathbf{D}[i,j]$ is equal to the difference in corresponding elements,

$$\mathbf{D}[i,j] = |\mathbf{Q}[i] - \mathbf{T}[j]|. \tag{1}$$

From this difference matrix, a cost matrix $\mathbf{M}$ is constructed that keeps a running tab on the dissimilarities of the elements making up the curves while adding up these costs

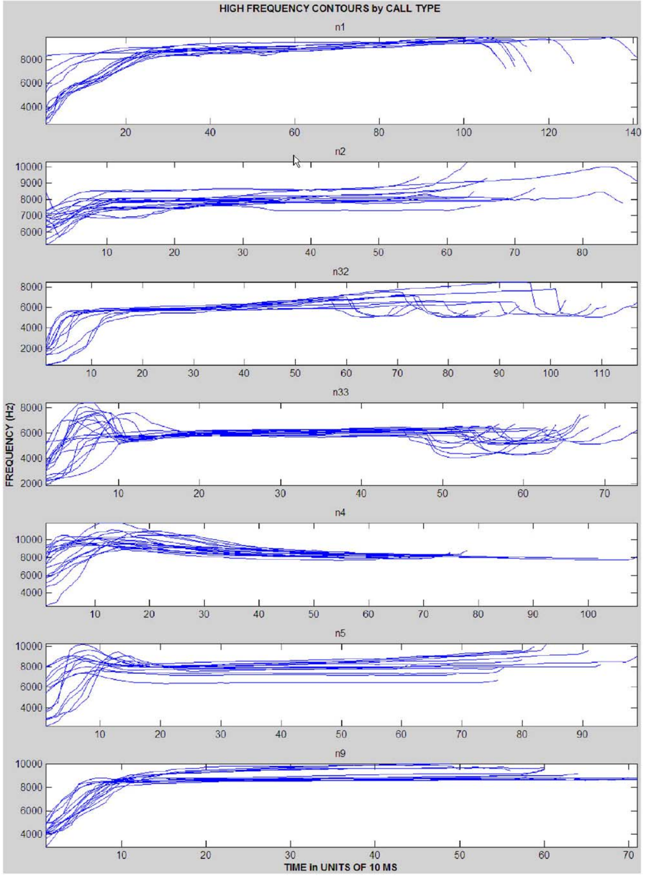

FIG. 3. (Color online) Pitch contours of the high frequency calls of the northern resident group of killer whales plotted in perceptually designated call types. They are plotted separately due to overlap as in Fig. 2.

to give a final number called the "dissimilarity" or distance between the query and target. We examine the cost matrices of our four algorithms below.

1. Ellis method

This is the simplest and most straightforward algorithm.[1] Each element of the cost matrix is obtained by adding the difference element for that position [obtained from Eq. (1)] to the minimum of the three previously determined elements of the cost matrix, which are (1) diagonal, (2) above, and (3) to the left:

$$\mathbf{M}[i,j] = \min\begin{pmatrix} \mathbf{M}[i-1,j-1] \\ \mathbf{M}[i-1,j] \\ \mathbf{M}[i,j-1] \end{pmatrix} + \mathbf{D}[i,j]. \tag{2}$$

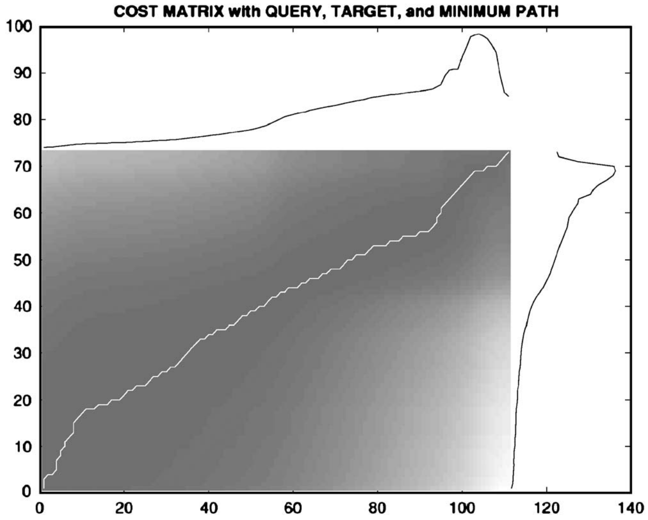

FIG. 4. Cost matrix with minimum cost path in white and input contour shapes above and to the right. The shorter sound is called the query and the longer sound the target.

2. Sakoe-Chiba method

The method of Sakoe and Chiba (1978) in an altered form was used by Deecke and Janik (2006). It is more complex and compares the weighted sum of difference elements from two columns and two rows distant with the weighted diagonal as shown in the equation below. We have chosen the form indicated by Sakoe and Chiba to give the best results:

$$\mathbf{M}[i,j] = \min\begin{pmatrix} \mathbf{M}[i-1,j-1] + 2\cdot\mathbf{D}[i,j] \\ \mathbf{M}[i-2,j-1] + 2\cdot\mathbf{D}[i-1,j] + \mathbf{D}[i,j] \\ \mathbf{M}[i-1,j-2] + 2\cdot\mathbf{D}[i,j-1] + \mathbf{D}[i,j] \end{pmatrix}. \tag{3}$$

3. Itakura method

This method (Itakura, 1975) was used by Buck and Tyack (1993):

$$\mathbf{M}[i,j] = \min\begin{pmatrix} \mathbf{M}[i-2,j-1] \\ \mathbf{M}[i-1,j-1] \\ \mathbf{M}[i,j-1] \end{pmatrix} + \mathbf{D}[i,j]. \tag{4}$$

It differs from other algorithms in that there is a constraint that two elements cannot be chosen sequentially from the same row, i.e., if $\mathbf{M}[i,j-1]$ is the minimum element, then it is not an option for the next element of the cost matrix in that row.

4. Chai-Vercoe method

This is the method often used in music information retrieval (Chai and Vercoe, 2003; Foote, 2000; Kruskal and Sankoff, 1983) and was extremely successful in classifying our killer whale calls from Marineland. The cost matrix is generated with

$$\mathbf{M}[i,j] = \min\begin{pmatrix} \mathbf{M}[i-1,j] + a, \\ \mathbf{M}[i,j-1] + b, \\ \mathbf{M}[i-1,j-1] + \mathbf{D}[i,j] \end{pmatrix}. \tag{5}$$

Here each element of the cost matrix can come from (1) the cost element directly above and adding a, the cost of an insertion; (2) the cost element to the left and adding b, the cost of an deletion; or (3) the previous element along the diagonal with the addition of the difference in corresponding elements. Since a deletion means a difference in lengths,

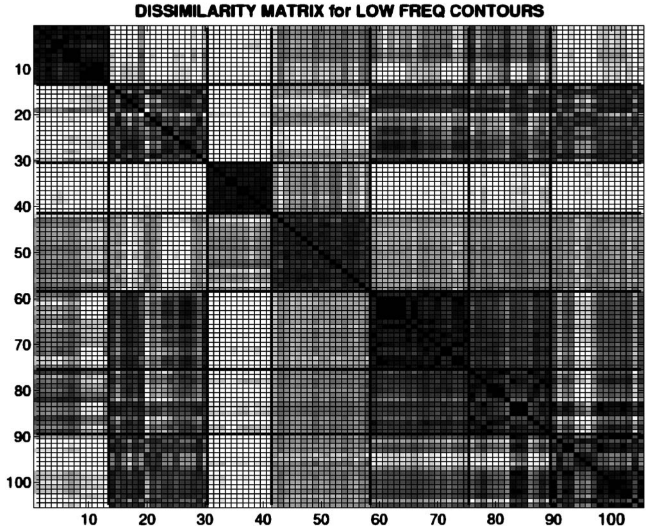

FIG. 5. Dissimilarity matrix of the LFC using the method of Sakoe and Chiba. Each point represents the dissimilarity of (or distance between) a pair of sounds.

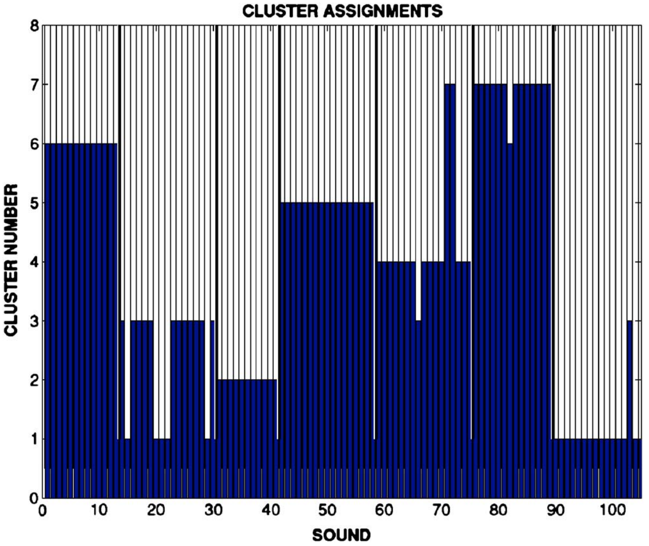

FIG. 6. (Color online) Clustering results for the difference matrix of Fig. 5.

which we do not want to penalize, b was chosen to be 0. The principal disadvantage of this method is that it contains the adjustable parameter "a."

For each of these four methods a running tab is kept of which choice is made for each element. Thus the minimum path can be retraced, and an example is shown in Fig. 4. The final "dissimilarity" is the number $\mathbf{M}[i_{\max},j_{\max}]$ normalized by dividing by the length of the query; this is a measure of the difference in the two contours. Identical signals will have a diagonal best path and a cost of zero, while larger differences will increase the matching cost. For classification these costs are a means of grouping (clustering) the calls with the smallest dissimilarities.

5. Dissimilarity matrices

Dissimilarity matrices were obtained by calculating a cost matrix for each pair of the low frequency calls shown in Fig. 2 to give a matrix with elements equal to these dissimilarities. The frequencies in Hz, which are graphed, were transformed for the cost matrix calculation using

$$f_{\text{cents}} = 12 \log_2(f/f_{\text{ref}}), \tag{6}$$

where f_{ref}=440 Hz as described in Brown *et al.* (2006). This unit means that we are comparing ratios of frequencies rather than absolute values, and, for example, a difference of 100 Hz and 200 Hz will be weighted the same as a difference of 400 Hz and 800 Hz. An identical procedure was carried out for obtaining a dissimilarity matrix for the high frequency calls shown in Fig. 3 as well as the derivatives (point to point differences of each curve in Figs. 2 and 3) for both groups. The derivatives are a measure of the shape alone of the curves and are independent of absolute frequency.

An example of a dissimilarity matrix is given in Fig. 5. Each element of this matrix represents the result of calculating a cost matrix for a particular pair of calls. Since there were a total of 105 calls, each dissimilarity matrix represents the results of calculating a cost matrix for 105(104)/2 or 5460 pairs of calls. The matrix is not truly symmetric but comparison of the shorter (query) to the longer (target) sound has been found to be a more accurate measure of the difference; therefore the elements below the diagonal were obtained by transposition.

The example of Fig. 5 was calculated for the low frequency calls using the Sakoe-Chiba method. Here the dark elements indicate a small distance and the lighter ones a larger distance. The perceptual groupings of Fig. 2 are indicated with bold horizontal and vertical lines. Perfect agreement with the perceptual results would give black blocks along the diagonal corresponding to small distances for the perceptual groupings and white elsewhere indicating large distances. The third group (n32) is closest to this ideal with white for all other groups except the fourth (n33). The last three groups are mixed with each other as well as with the second group, and this was typical of all calculations.

B. Classification

For each method the distances given by the dissimilarity matrices were transformed into a Euclidean-like space using multi-dimensional scaling. They were then clustered using a kmeans algorithm (Brown *et al.*, 2006) from Matlab into seven call types to compare to the perceptual classification. An example classification corresponding to the dissimilarity matrix of Fig. 5 for the Sakoe-Chiba method is given in Fig. 6. There are ten errors in this example all involving clusters 2, 5, 6, and 7, as could be predicted from the dissimilarity matrix.

IV. RESULTS AND DISCUSSION

Classification results for each of the four algorithms used to classify the low frequency component (LFC), the high frequency component (HFC), and their derivatives are given numerically in Table I as well as in the corresponding bar graphs of Figs. 7–9, where they are more easily visualized. In Table I the column labeled "Double group" indicates

TABLE I. Summary of results. The upper third of the table gives the percent agreement with the perceptual classification of the LFC and HFC contours alone in columns 1 and 3 and for their sum in column 6. The middle third of the table does the same for the LFC and its derivative. The lower third does the same for the HFC and its derivative.

	Summary of results						
	Low frequency and high frequency components						
	Low freq	Double group	High freq	Double group	Ratio	Sum	Double group
Ellis	77		70		1.6	80	
Sakoe-Chiba	90		69		1.6	90	
Itakura	86		68	1	1.6	81	
Chai-Vercoe	83	1	79	1	1.8	90	1
	Low frequency component and its derivative						
	Low freq	Double group	Low freq derivative	Double group	Ratio	Sum	Double group
Ellis	77		81	1	12	77	
Sakoe-Chiba	90		82	1	12	88	
Itakura	86		70	1	11	73	
Chai-Vercoe	83	1	86	1	20	86	1
	High frequency component and its derivative						
	High freq		High freq derivative	Double group	Ratio	Sum	Double group
Ellis	70		76		14	77	
Sakoe-Chiba	69		76		16	86	
Itakura	68	1	57	1	14	73	
Chai-Vercoe	79	1	77	1	26	80	1

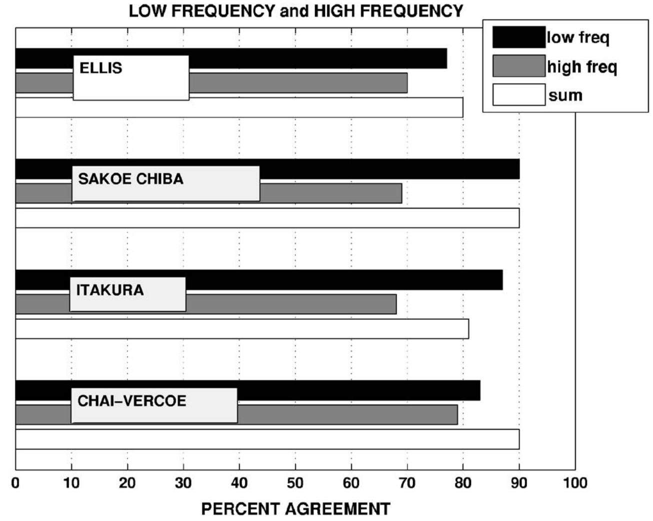

FIG. 7. Percent agreement with perceptual results for each method of calculating for the LFC, HFC, and their sum.

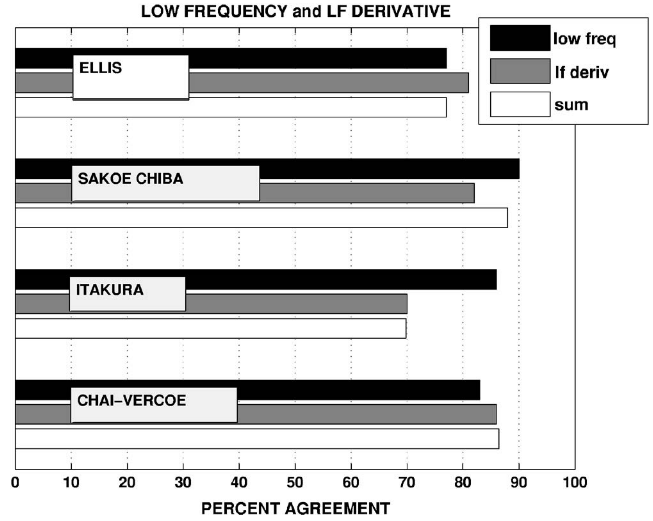

FIG. 8. Percent agreement with perceptual results for each method of calculating for the LFC, LFC derivative, and their sum.

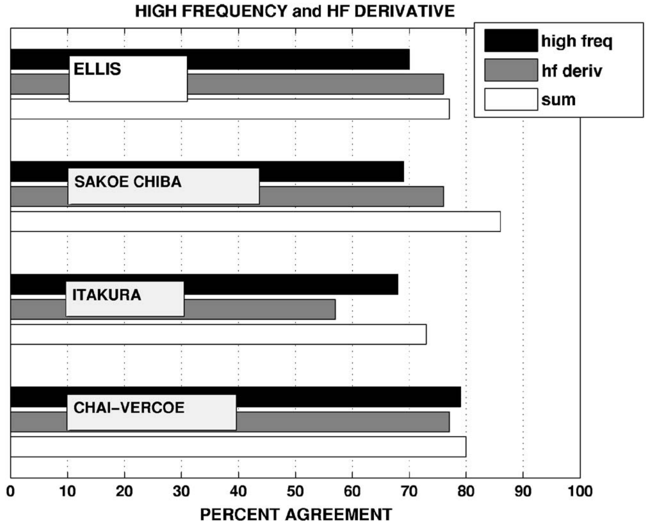

FIG. 9. Percent agreement with perceptual results for each method of calculating for the HFC, HFC derivative, and their sum.

that two of the perceptual call types were classified into the same cluster, so these results are not as good as the stated percentage would indicate.

A. Low frequency and high frequency contours

In the upper third of Table I (cf. Fig. 7), the agreement with perceptual results is greater for the low frequency component (LFC in column 1) than the high frequency component (HFC in column 3). Sakoe-Chiba does best on the LFC at 90% with Chai-Vercoe best for the HFC. To determine the effect of the LFC and HFC distances combined, their dissimilarity matrices were added with weighting corresponding to the ratio of the means (column 5), and a new classification calculation was carried out for each algorithm. The results are in column 6 with significant improvement for the Chai-Vercoe method.

B. Low and high frequency contour derivatives

Classification results for the LFC and HFC derivatives are found in column 3 of the remainder of Table I as well as in Figs. 8 and 9. With one exception these results are all over 70%, which is quite remarkable for these very irregular curves.

C. Sum of contours and derivatives

Results for the Marineland group (Brown *et al.*, 2006) were improved from 88% to 98% agreement with perceptual results by adding the dissimilarity matrix for the LFC derivative to the matrix for LFC alone weighted with the means of the two matrices. Results of the analogous calculation for this set of calls are given in column 6. Here they yield a marginal improvement for the Chai-Vercoe method. For the HFC summed with the HFC derivative calculation, Sakoe-Chiba was improved by 10% and 17%, respectively, over the HFC derivative and HFC alone results to 86%.

D. Summary

With the exception of the Itakura method on the HFC derivatives, the results of all algorithms were in agreement with the perceptual classification by near 70% or greater and could thus be considered successful given the difficulties of this data set. The shapes of the curves show variation within each perceptual call type, and there is considerable similarity among groups 2, 5, 6, and 7 (call types n2, n4, n5, and n9) in frequency range as well as shape. The best result was 90% using Sakoe-Chiba for the low frequency contours, which is truly outstanding.

It should be recalled that the perceptual classification was made by listening to the calls while observing their spectra, rather than by an examination of the contour alone. These perceptual decisions were probably influenced by spectral content (not present in the contours). Also DTW is most effective for curves differing in length by less than a factor of 2; in this set there was variation of lengths as great as a factor of 3. Thus, it is in fact remarkable that the computer classification reached 90% agreement.

V. CONCLUSIONS

These results with a maximum of 90% agreement with the perceptual data were not as impressive as the 98% reported previously on the Marineland set. However, this is easily understood in comparing Fig. 2 to the corresponding figure in Brown *et al.* (2006). The Marineland calls separated nicely in frequency and could be viewed on the same graph. In a similar graph (not included) for these northern resident calls, four of the call types were intermingled and unseparable visually. In other DTW studies on marine mammals (Buck and Tyack, 1993; Deeke and Janik, 2006) there were few contours, and they were virtually identical within groups. The current data set thus represented a severe test for DTW, and the 70%–90% agreement with perceptual classification is excellent.

Of the algorithms examined and combinations of dissimilarities, Sakoe-Chiba performed best on the LFC. While slightly more complicated than the other algorithms, it has the advantage of having no adjustable parameters. There is also a positive side to the fact that results were best for the low frequency component alone in that preprocessing reduces to pitch tracking a single component.

Dynamic time warping has proven to be an excellent technique for the automatic classification of killer whale call types. One of its most rewarding applications would be the ability to monitor the movements and habitat preferences of killer whales just by tracking sounds heard at remote monitoring stations. This will only be possible with systems developed to automatically process and identify calls heard at those locations so that the group producing them can be identified remotely.

ACKNOWLEDGMENTS

JCB is grateful to Dr. Eamonn Keogh for his MATLAB code, which was adapted to give Fig. 4. Funding was provided by WHOI's Ocean Life Institute and a Royal Society fellowship to PJOM.

[1] We are calling this the Ellis method as the code was obtained from Dan Ellis's website http://labrosa.ee.columbia.edu/matlab/dtw/.

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.

Brown, J. C., Hodgins-Davis, A., and Miller, P. J. O. (**2006**). "Classification of vocalizations of killer whales using dynamic time warping," J. Acoust. Soc. Am. **119**, EL34–EL40.

Brown, J. C., and Miller, P. J. O. (**2006a**). "Dynamic time warping for automatic classification of killer whale vocalizations," J. Acoust. Soc. Am. **119**, 3434.

Brown, J. C., and Miller, P. J. O. (**2006b**). "Classifying killer whale vocalization using time warping," Echoes **16**, 45–47.

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

Deecke, V. B., and Janik, V. M. (**2006**). "Automated categorization of bioacoustic signals: Avoiding perceptual pitfalls," J. Acoust. Soc. Am. **119**, 645–653.

Foote, J. (**2000**). "ARTHUR: Retrieving orchestral music by long-term structure," Proc. of the 1st Annual International Symposium on Music Information Retrieval (ISMIR 2000), pp. 1–6.

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

Hu, N., Dannenberg, R. B., and Tzanetakis, G. (**2003**). "Polyphonic audio matching and alignment for music retrieval," IEEE Workshop on Applications of Signal Processing to Audio, New Paltz, NY.

Itakura, F. (**1975**). "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-23**, 57–72.

Kruskal, J., and Sankoff, D. (**1983**). "An anthology of algorithms and concepts for sequence comparison," in *Time Warps, String Edits, and Macromolecules: The Theory and Practice of String Comparison*, edited by D. Sankoff and J. Kruskal (Addison-Wesley, Reading, MA).

Myers, C., Rabiner, L. R., and Rosenberg, A. E. (**1980**). "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-28**, 623–634.

Rabiner, L., and Juang, B. H. (**1993**). *Fundamentals of Speech Recognition* (Prentice Hall, Englewood Cliffs, NJ).

Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming optimization for spoken word recognition," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-26**, 43–49.

Tyson, R. (2006). "The presence and potential functions of nonlinear phenomena in cetacean vocalizations," thesis, Florida State University, Tallahassee, FL.

Tyson, R., Nowacek, D. P., and Miller, P. J. O. (**2006**). "Nonlinear phenomena in the vocalizations of North Atlantic right whales (*Eubalaena glacialis*) and killer whales (*Orcinus orca*)," accepted by J. Acoust. Soc. Am..

Wang, C., and Seneff, S. (**2000**). "Robust pitch tracking for prosodic modeling in telephone speech," Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 3, pp. 1343–1346.

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.