J. Mar. Biol. Ass. U.K. (2007), **87**, 35–38
Printed in the United Kingdom

doi: 10.1017/S0025315407054756

Neural network-based sperm whale click classification

M. van der Schaar*, E. Delory*, A. Català[†] and M. André*[‡]

*Laboratori d'Aplicacions Bioacústiques, Universitat Politècnica de Catalunya, Spain. [†]Departament ESAII, Universitat Politècnica de Catalunya, Spain. [‡]Corresponding author, e-mail: michel.andre@upc.edu

Recordings of a group of foraging sperm whales usually result in a mixture of clicks from different animals. To analyse the click sequences of individual whales these clicks need to be separated, and for this an automatic classifier would be preferred. Here we study the use of a radial basis function network to perform the separation. The neural network's ability to discriminate between different whales was tested with six data sets of individually diving males. The data consisted of five shorter click trains and one complete dive which was especially important to evaluate the capacity of the network to generalize. The network was trained with characteristics extracted from the six click series with the help of a wavelet packet-based local discriminant basis. The selected features were separated in a training set containing 50 clicks of each data set and a validation set with the remaining clicks. After the network was trained it could correctly classify around 90% of the short click series, while for the entire dive this percentage was around 78%.

INTRODUCTION

There has been a continuous interest in the tracking and monitoring of sperm whales (*Physeter macrocephalus*) through passive means (Leaper et al., 1992; Mellinger et al., 2004; Thode, 2004). Sperm whales especially make excellent targets for passive acoustic observation, because, while foraging, they produce continuous series of clicks that can be detected at large distances. More recently, the monitoring effort has shifted from a mostly scientific interest to more practical interests, as for example in the avoidance of collisions with shipping traffic (André et al., 1997; André & Potter, 2000; Delory et al., 2006).

Acoustic recordings of group dives of sperm whales often result in a mixture of signals, making analysis by automated tools difficult. In order to track an individual animal it is necessary to find unique characteristics in its signals. This is complicated by the fact that the frequency content of the clicks is influenced by various factors, such as the animal's depth, orientation, and distance to the hydrophone (Thode et al., 2002; Møhl et al., 2003). However, for this study we were not interested in the unique identification of a whale, which would require stable and unique characteristics, but in distinguishing between individuals in a small group for the duration of a single dive, which merely requires features not to overlap.

Some authors have previously looked at the processing of sperm whale vocalizations, like click detection and feature extraction (Adam et al., 2005; Lopatka et al., 2005; van der Schaar et al., 2007). Wavelet-based identification of sperm whales was also attempted in Dougherty (1999), where the author noted difficulties due to the variability of click characteristics during a dive. A neural network solution to identify marine mammal sounds was proposed in Huynh et al. (1998), where the authors were successful in identifying different species. In Murray et al. (1998) the authors used a neural network approach to classify killer whale vocalizations. Based on the specific distribution of the characteristics found in our data sets we chose a different type of neural network known as a radial basis function network (Bishop, 1995). This type of network acts as a non-linear classifier with the advantage of a simple structure, which allows fast training through the combination of unsupervised and (linear) supervised techniques.

MATERIALS AND METHODS

Data collection

The sperm whale data were collected from an inflatable boat during four field seasons spanning four to ten weeks each (from 1997 to 1999) at Kaikoura, New Zealand (Jaquet et al., 2001). Recordings were made of solitary diving male sperm whales using an omni-directional hydrophone (Sonatech 8178; minimum frequency response 100 Hz to 30 kHz ±5 dB) lowered to 20 m. Visual identification of the whales when they were surfacing allowed the recordings to be identified. The hydrophone was first connected to a fixed gain amplifier (at response from 0 to 45 kHz), and then to one channel of a Sony TCD-D10PROII digital audio tape recorder (frequency response 20 Hz to 22 kHz ±1 dB with an anti-aliasing filter at 22 kHz). Recordings were digitized at 48 kHz and 16 bits.

Since one can rarely hear more than five animals simultaneously in the recordings of a group dive (Whitehead & Weilgart, 1989), we restricted the number of animals that should be recognized to six. The data sets were therefore created from the clicks of six different whales. Five sets contained a single click train (a consecutive series of clicks) each from a specific whale, the smallest set containing 83

https://doi.org/10.1017/S0025315407054756 Published online by Cambridge University Press

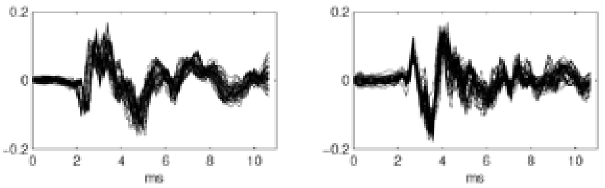

Figure 1. Synchronization of 50 clicks from Set 3 (left) and Set 4 (right). The clicks have been low-pass filtered, downsampled and normalized by their energy.

clicks and the largest 248. The sixth set contained the clicks from an entire dive of another whale, leaving out creak sequences, and included well over a thousand clicks. This set was especially important to assess the capacity of the algorithm to generalize.

Feature selection

In order to classify the clicks, suitable characteristic features of the signals had to be found, and we searched for these in the lower frequency components. Although on-axis clicks can present a peak frequency around 12 kHz (Møhl et al., 2003), this type of click is rarely found in recordings. More importantly, higher frequency components are not expected to be suitable since they suffer more from transmission loss and directional effects. Research has shown (Goold & Jones, 1995) that clicks recorded off-axis contain dominant low frequencies; specifically for male sperm whales a dominant frequency can be found below 2000 Hz. Since lower frequency components are more stable, we focused our attention on these. In preparation for the classification algorithm, the clicks were manually detected and filtered for echoes and acceptable noise levels. The clicks were then denoised using a standard soft-thresholding algorithm, available from Wavelab (Donoho et al., 1999). In soft-thresholding, coefficients below a certain threshold are set to zero, while the magnitude of the other coefficients is reduced by the threshold to retain a smooth signal. After denoising, the clicks were synchronized at low frequencies using a typical low-pass filtered click as matched filter. After synchronization, all signals were low-pass filtered at 2000 Hz, and subsequently normalized by their energy and downsampled to remove redundant information. The result of these operations can be seen in Figure 1. A local discriminant basis based on a wavelet packet decomposition of the signals (Saito & Coifman, 1994) was created for the six classes, and the 15 strongest wavelet coefficients (according to Fisher's discriminant) were selected and used for the classification.

Radial basis function networks

A radial basis function network (Bishop, 1995) is a two layer neural network (see Figure 2). The activation functions in the first (hidden) layer consist of non-linear radial functions, which are characterized by having a single global maximum (or minimum) and a monotonically decreasing (increasing) value away from this point. The most commonly-used activation function is the Gaussian given by:

$$\phi_{\mu,\sigma}(x) = \exp\left(-\|x-\mu\|^2/2\sigma^2\right)$$

The Gaussian function has a maximum response of 1 when $x = \mu$ and its value lowers when the distance to μ increases. Its width, or locality, can be controlled through σ. The second (output) layer of the network may contain non-linear activation functions as well, but it can be particularly advantageous to make it linear, i.e. summing the weighed inputs from the hidden layer. This will allow the network to be trained without the use of time consuming non-linear optimization algorithms. The choice for this type of network can be motivated with Figure 3. The left image in this Figure shows the two strongest discriminating features for four data sets, plotted against each other. A natural way of characterizing the classes would be by placing centres in the clusters and assigning a point to a class depending on the distance to each of the centres. This approach naturally leads to the use of a distance based network, like a RBF network.

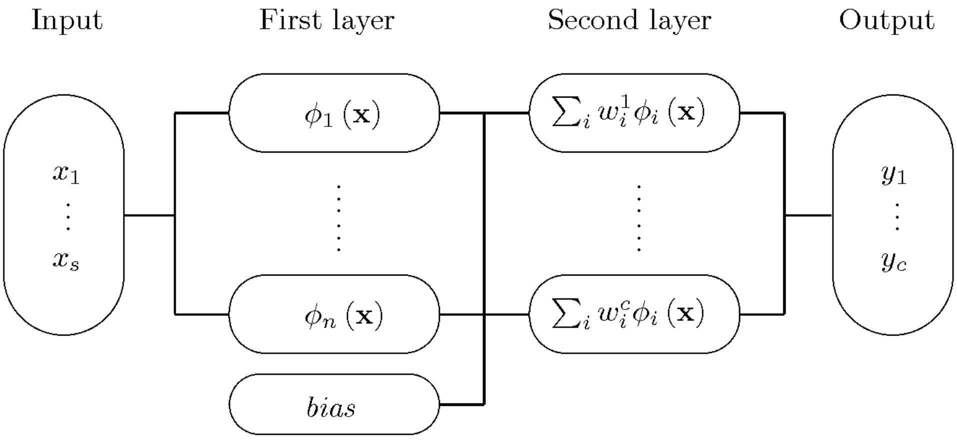

Figure 2. Schematic of a RBF-network. A s-dimensional sample **x** enters on the left. It is first run through the n hidden layer nodes, where the distances between **x** and centres μ are evaluated through Gaussian functions ϕ_n(x). The outputs of the n Gaussian functions are then weighed with weights w_i^j and linearly combined in the second layer nodes (containing one node per class). An additional weight (w_0^j) is usually added to each second layer node to account for the bias factor. This bias is then represented by an additional constant activation function in the first layer, $\phi_0 \equiv 1$. Taking the output vector **y**, the class of the sample **x** is computed by $\arg\max_i y_i$.

https://doi.org/10.1017/S0025315407054756 Published online by Cambridge University Press

Table 1. *Percentages of correctly classified clicks from the validation sets using different numbers of centres per class. In the first column the value gives the number of clusters used per class within parentheses the total number of clusters used to model the data. The eighth column contains the data from the complete dive, with two extraordinary noisy segments removed. The last column gives the average of the correct classification over the six data sets (using Set 6).*

No. centres per class (total)	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 6*	Avg
1(6)	80	94	92	70	56	73	86	78
2(12)	89	96	93	85	50	78	89	82
3(18)	90	94	92	91	91	78	90	89
4(24)	94	96	86	85	100	80	92	90
5(30)	93	96	85	85	100	80	92	90
6(36)	93	96	86	85	100	81	92	90
7(42)	91	96	85	82	100	81	92	90
No. samples	198	69	148	33	54	1207	1013	

The two layers in the RBF network perform the respective roles of clustering and classifying samples. The hidden layer covers the data with centres and evaluates the distances of an input to the centres through the Gaussian functions. The output layer then linearly combines the information from all the Gaussian functions to assign a class to the input.

The different roles of the two layers also allow them to be trained in two separate steps. The parameters for the hidden layer that need to be learned are the number of Gaussian functions needed to cover the data, and for every function its centre and width. One way to cluster the data, and to learn these parameters, is to use *k*-means clustering. This does not directly give the optimal number of clusters to use, but since the training of the network goes fast, in the order of minutes, different network configurations can be tested easily. When the network is trained in two steps the learning of the hidden layer can be done unsupervised, clustering all training data together ignoring the different classes, or the class information can be used by clustering the individual classes. The clusters then provide the centres for the Gaussian functions, and the cluster variances could also be used to define the widths of the Gaussians. However, we found better performance when the widths were set to a fixed value, ensuring that the functions have sufficient overlap.

The number of nodes in the output layer was set equal to the number of classes, using binary encoding for the targets (in the case of four classes, a sample from class 1 has target $[1\ 0\ 0\ 0]^t$, while a class 3 sample has target $[0\ 0\ 1\ 0]^t$). After the hidden layer has been trained the weights of the nodes in the output layer can be calculated directly and efficiently due to its linearity (Bishop, 1995).

In the example of Figure 3, *k*-means clustering was used to place eight centres in the data in the second image, without taking into account the class labels. For this example the number of centres was set to twice the number of classes based on visual inspection of the first image. The third image shows the decision regions around these centres as defined by the neural network. It should be noted that further away from the data samples the classification is less reliable due to a lack of information.

RESULTS

The training data were created by taking the 15 most discriminating coefficients from the local basis representations of 50 clicks from the start of the data sets; the remaining clicks were used for the validation sets. The hidden layer of the RBF network was initialized by clustering the classes individually with *k*-means clustering. Since *k*-means cannot calculate an optimal number of clusters by itself we performed the classification using 1 to 7 clusters per class. The *k*-means runs were repeated 30 times in an attempt to find an optimal solution. Individual clustering of the classes gave better results than clustering all data together with the same total number of clusters. This may partly be caused by *k*-means, which had difficulties finding an optimal stable solution when the number of clusters was high. Classification of the training data was not good when only 1 or 2 clusters per class were used, but with more than two clusters the results were always 100% correct, with the exception of a single sample in Set 3, which was always wrongly classified. This single outlier should not significantly affect the results. The results of classifying the remaining clicks (validation set) in the data sets are shown in Table 1; Sets 1 to 5 contained the single click trains, Set 6 contained data from a complete

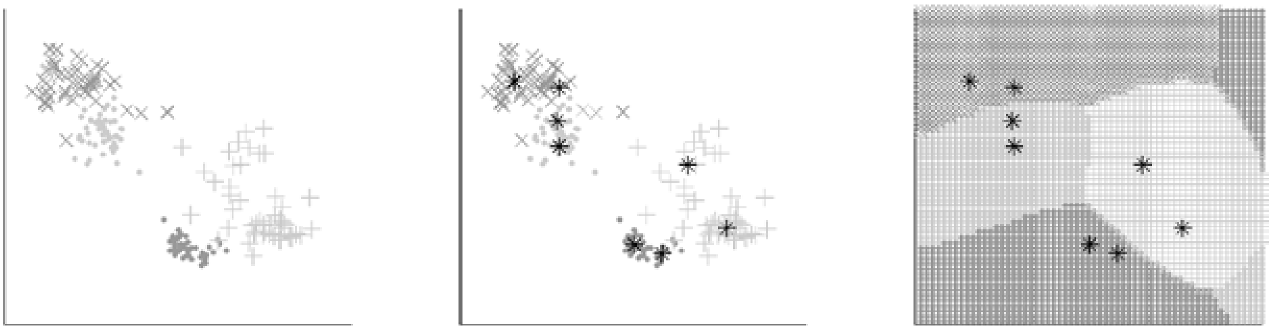

Figure 3. Example of classification with a RBF-network. The image on the left shows the two most discriminating characteristics of four data sets. In the second image eight centres are placed in the data using *k*-means (in this case without using class information). The third image shows the class regions as defined by the RBF-network. It should be noted that classification in areas further away from the centres (as for example in the upper right corner) is not reliable due to missing data.

https://doi.org/10.1017/S0025315407054756 Published online by Cambridge University Press

dive. Within this last data set there were two segments that classified particularly badly regardless of the number of centres used, likely due to a temporary increase in noise. Since these small segments had a disproportionate effect on the classification outcome, we also show the results with the segments removed in the eighth column in Table 1 (Set 6*). From the table, using three or four centres per class performs best. Adding more centres does not significantly improve the results, and may lead to poor generalization.

DISCUSSION

The radial basis function network was capable of distinguishing between the six individual animals. Using only 50 clicks at the start of the recordings the method was able to generalize to the entire dive of Set 6, although the limited amount of data available for the other classes does not allow making more general conclusions. It also needs to be noted that the network was tested with male sperm whales, while social groups mainly contain females and young calves. Goold & Jones (1995) showed that clicks from female sperm whales have somewhat different properties. Unfortunately, data of diving female sperm whales, where clicks can be assigned with certainty to a specific whale, were not available to us for analysis.

One important advantage of the radial basis network is that it allows fast training of the parameters. During the recording of a group dive some animals may return to the surface while other animals start foraging. A change in the diving group configuration might be detected automatically, either by a large number of clicks that cannot be classified because there is no significant maximum node in the output layer, or by a disproportionately large number of clicks falling in the same class within a short time interval because the clicks of the new animal are similar to those of a whale known to the network. The network can then be retrained quickly including the new animal, although creating the training set itself may still have to be done manually.

Occasionally, it may also be necessary to retrain the network with a set of recently classified clicks. It is possible that due to the variability of the features the click's characteristics are no longer modelled correctly by the cluster centres as they were defined by the training set. This can occur, for example, when the distance or the recording angle to the whales change significantly. When it is detected that the maximum network output for clicks is slowly lowering, indicating features moving away from the cluster centres, the network can be retrained automatically with recent clicks. More data from complete dives will be necessary to see how well the network will handle changing features and animals unknown to the network.

When recordings are made with more than one hydrophone the combination of RBF classification and the direction of arrival of clicks should make automatic separation of clicks from different animals in a small group possible.

The authors thank Nathalie Jaquet for providing the recordings, and the referees for their useful remarks on this paper. The study was funded by the BBVA (Banco Bilbao Vizcaya Argentaria) Foundation.

REFERENCES

Adam, O., Lopatka, M., Laplanche, C. & Motsch, J., 2005. Sperm whale signal analysis: comparison using the autoregressive model and the daubechies 15 wavelets transform. *Enformatika*, **4**, 188–195.

André, M. & Potter, J., 2000. Fast-ferry acoustic and direct physical impact on cetaceans: evidence, trends and potential mitigation. *Proceedings of the 5th European Conference on Underwater Acoustics (ECUA2000)*, (ed. M. Zakharia), vol. 1, 491–496.

André, M., Terada, M. & Watanabe, Y., 1997. Sperm Whale (*Physeter macrocephalus*) behavioural response after the playback of artificial sounds. *Report of the International Whaling Commission*, **47**, 499–504.

Bishop, C., 1995. *Neural networks for pattern recognition.* Oxford: Oxford University Press.

Delory, E., André, M., Navarro Mesa, J.-L. & Schaar, M. van der, 2006. On the possibility of detecting surfacing sperm whales using others' foraging clicks. *Journal of the Marine Biological Association of the United Kingdom*, **87**, 47–58.

Donoho, D., Duncan, M., Huo, X. & Levi, O., 1999. Version 8.02, http://wwwstat.stanford.edu/~wavelab/, visited 2002.

Dougherty, A., 1999. *Acoustic identification of individual sperm whales (*Physeter macrocephalus*).* Msc thesis, University of Washington.

Goold, J. & Jones, S., 1995. Time and frequency domain characteristics of sperm whale clicks. *Journal of the Acoustical Society of America*, **98**, 1279–1291.

Huynh, Q., Cooper, L., Intrator, N. & Shouval, H., 1998. Classification of underwater mammals using feature extraction based on time-frequency analysis and BCM theory. *IEEE Transactions on Signal Processing*, **46**, 1202–1207.

Jaquet, N., Dawson, S. & Douglas, L., 2001. Vocal behavior of male sperm whales: why do they click? *Journal of the Acoustical Society of America*, **109**, 2254–2259.

Leaper, R., Chappell, O. & Gordon, J., 1992. The development of practical techniques for surveying sperm whale populations acoustically. *Report of the International Whaling Commission*, **42**, 549–560.

Lopatka, M., Adam, O., Laplanche, C., Zarzycki, J. & Motsch, J., 2005. An attractive alternative for sperm whale click detection using the wavelet transform in comparison to the fourier spectrogram. *Aquatic Mammals*, **31**, 463–467.

Mellinger, D., Stafford, K. & Fox, C., 2004. Seasonal occurrence of sperm whale (*Physeter macrocephalus*) sounds in the Gulf of Alaska. 1999–2001. *Marine Mammal Science*, **20**, 48–62.

Møhl, B., Wahlberg, M., Madsen, P., Heerfordt, A. & Lund, A., 2003. The monopulsed nature of sperm whale clicks. *Journal of the Acoustical Society of America*, **114**, 1143–1154.

Murray, S., Mercado, E. & Roitblat, H., 1998. The neural network classification of false killer whale (*Pseudorca crassidens*) vocalizations. *Journal of the Acoustical Society of America*, **104**, 3626–3633.

Saito, N. & Coifman, R., 1994. Local discriminant bases. Mathematical imaging: wavelet applications in signal and image processing II, (ed. A. Laine & M. Unser) (Proc SPIE), vol. 2303.

Schaar, M. van der, Delory, E., Weide, J. van der, Kamminga, C., Goold, J., Jaquet, N. & André, M., 2007. A comparison of model and non-model based time-frequency transforms for sperm whale click classification. *Journal of the Marine Biological Association of the United Kingdom*, **87**, 27–34.

Thode, A., 2004. Tracking sperm whale (*Physeter macrocephalus*) dive profiles using a towed passive acoustic array. *Journal of the Acoustical Society of America*, **116**, 245–253.

Thode, A., Mellinger, D., Stienessen, S., Martinez, A. & Mullin, K., 2002. Depth-dependent acoustic features of diving sperm whales (*Physeter macrocephalus*) in the Gulf of Mexico. *Journal of the Acoustical Society of America*, **112**, 308–321.

Whitehead, H. & Weilgart, L., 1989. Click rates from sperm whales. *Journal of the Acoustical Society of America*, **87**, 1798–1806.

Submitted 13 June 2006. Accepted 24 September 2006.

https://doi.org/10.1017/S0025315407054756 Published online by Cambridge University Press