Proceedings of Meetings on Acoustics

Volume 19, 2013 http://acousticalsociety.org/

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Animal Bioacoustics
Session 3aAB: Perceiving Objects I

3aAB3. Exploring the capacity of neural networks to recognize objects from dolphin echoes across multiple orientations

Matthew G. Wisniewski*, Caroline M. DeLong, Amanda L. Heberle and Eduardo Mercado III

***Corresponding author's address: Psychology, University at Buffalo, The State University of New York, Buffalo, NY 14260, mgw@buffalo.edu**

Dolphins naturally recognize objects from multiple angles using echolocation. With training, humans can also learn to accurately classify objects based on their echoic features. In this study, we used neural networks to identify acoustic cues that enable objects to be recognized from varying aspects. In Simulation 1, a self-organizing map was able to differentiate a subset of objects using only amplitude and frequency cues, but it classified some echoes from different objects as being from the same object. In Simulation 2, a multilayer perceptron was trained through error correction to identify objects based on echoes from a single aspect, and then tested on its ability to recognize those objects using echoes from different orientations. Overall, perceptrons performed similarly to trained undergraduates. Analysis of network connection weights revealed that both the amplitude and frequency of echoes, as well as the temporal dynamics of these features over the course of an echo train, enabled perceptrons to accurately identify objects when presented with novel orientations. These findings suggest that learning may strongly impact an organism's ability to echoically recognize an object from any viewpoint.

Published by the Acoustical Society of America through the American Institute of Physics

 [DOI: 10.1121/1.4799305]
Received 22 Jan 2013; published 2 Jun 2013

INTRODUCTION

Dolphins use echolocation (i.e., biological sonar) by emitting a series of ultrasonic clicks and listening to the returning echoes (Au, 1993). Echolocation is used to navigate, avoid predators, and track moving prey. Most of the objects dolphins encounter are aspect-dependent; meaning that the size and shape of the surfaces of the object will change as they are viewed from different orientations. The echoes from objects can vary considerably depending on the angle from which they are inspected by the dolphin (Helweg, Au, Roitblat, & Nachtigall, 1996; Helweg, Roitblat, Nachtigall, & Hautus, 1996), and echoes from different aspects of a single object can vary more from each other than echoes from different objects (DeLong, Au, Lemonds, Harley, & Roitblat, 2006). This makes it difficult to determine how organisms identify an object with such large differences between echoes from different object orientations.

One strategy for finding object recognition cues is to use human listening studies (Au & Martin, 1989; DeLong, Au, Harley, Roitblat, Pytka, 2007; DeLong, Au, Stamper, 2007; Gorman & Sawatari, 1985; Gorman & Sejnowski, 1988; Helweg et al., 1995). In tasks presented to humans, echo stimuli altered to fall within the range of human hearing are played via headphones and participants are trained to identify the object corresponding to a particular train of echoes. The advantage of these studies is that participants can report the acoustic features of echoes they used to identify objects. DeLong, Heberle, Mata, Harley, and Au (2013) trained human participants to identify wood, copper, and ceramic objects from echo trains. DeLong et al.'s (2013) participants learned to correctly identify objects at trained and untrained aspects (i.e., participants generalized learning to novel aspects). Most of their participants reported using frequency and timbre to identify objects, and several also reported using amplitude features to recognize objects at untrained aspects. However, without an analysis of the actual echo stimuli, the usefulness of the reported cues cannot be confirmed. It could also be the case that there exist cues that cannot be described adequately in verbal reports due to the limits of vocabulary (e.g., some combination of frequency and amplitude or the tracking of frequency and amplitude over time).

A complementary strategy for cue identification is to train artificial neural networks to identify echo trains. Major advantages of neural networks are that they can be used to rapidly test various hypotheses about which features are critical for object recognition and generalization, and can be used to explore how different networks transform stimulus features in ways that facilitate recognition of novel exemplars (Guillette et al., 2010; Wisniewski, Radell, Guillette, Sturdy, & Mercado, 2012), as well as recognition of familiar exemplars presented in novel orientations (Gorman & Sejnowski, 1988). Early attempts to classify sonar targets with artificial neural networks revealed that relatively simple networks with multiple layers of processing discovered features within sonar returns that were similar to the features that human listeners reported using (Gorman & Sejnowski, 1988). Later studies of neural networks trained to recognize objects using either single echoes or multiple successive echoes generated from a dolphin's sonar signal showed that networks that integrate information from successive echoes are better able to mimic the recognition performance of dolphins (Moore, Roitblat, Penner, & Nachtigall, 1991), and that such networks can recognize objects at above chance levels even when the echoes were generated by objects that varied freely in orientation (Helweg, Roibtblat, & Nachtigall, 1993). Neural networks almost certainly do not replicate the perceptual processes employed by echolocating dolphins or humans. Nevertheless, the ability of such networks to recognize objects at levels comparable to dolphins and humans indicates that they may be extracting similar acoustical information.

In the current study, we addressed the acoustic cues that may have been utilized for object identification by humans in DeLong et al.'s (2013) experiment with two simulations. In simulation 1, echo stimuli used by DeLong et al. (2013) were used to train a self-organizing map (SOM) neural network. SOMs employ unsupervised learning algorithms that learn only with information about their input (Kohonen, 2001), making them useful for testing the hypothesis that acoustic similarity is enough to group echoes coming from the same object together without information about object identity. In Simulation 2, multilayer perceptron neural networks were trained with feedback about object identity to classify objects using the echo sequences learned by DeLong et al.'s (2013) participants. Networks were then tested on their ability to accurately classify novel echo sequences corresponding to different aspects. Neural network performance was compared to human performance and network structure was analyzed to determine the features networks used to identify objects at trained and novel aspects. We hypothesized that networks would confirm the usefulness of some of the cues reported by humans and that they would reveal other cues that humans did not report using.

SIMULATION 1

SOMs learn to spatially organize inputs in a map based on similarity, making it possible to visualize differences between stimuli varying along several dimensions. Units in SOMs learn with a competitive algorithm by which the unit with weights most similar to the input (the winner) and units close to it in the map adjust their weights to be more similar to the input. The result of this process is that similar inputs activate similar parts of the map and that each unit has a prototypical input (the unit's weight vector). If acoustic similarity alone is enough to separate echoes into different objects, then units in the map should not be activated by more than one object after training. Furthermore, if participants in DeLong et al.'s (2013) experiment were correct in their reports of using frequency mostly to identify objects, then the SOM should be able to separate objects in the map based on frequency information.

Methods

Echo Representations

The waveform of one example echo train from DeLong et al.'s (2013) study is shown in figure 1. DeLong et al. (2013) presented humans with echo trains (18 echoes per train) recorded from wood, copper, and ceramic objects at five different aspects covering -30 to +30 degrees (where 0 degrees, aspect 3, is the front face of the object). Each of these echo trains was automatically analyzed using a customized Matlab script to extract measurements of amplitude and frequency from echoes. Amplitude and peak frequency values were measured from each echo in a train (i.e., there was a peak frequency and amplitude measure for each echo). These measures were also used to compute means and standard deviations of amplitude and frequency across all echoes within a train. Measures from each individual train were combined into a vector and used as inputs for the SOM.

FIGURE 1. An example of an echo train used in DeLong et al.'s (2013) human listening experiment.

Network Architecture

A 3x3 SOM was constructed with neural network toolbox running in Matlab R2010a (MathWorks, 2010), allowing echo trains to be sorted in terms of their similarity to nine prototypes determined after learning with all echo trains. The SOM was trained for 2000 trials at a learning rate of .05.

Results & Discussion

Figure 2 shows units in the SOM that were designated winners after training. Each unit within the SOM corresponds to one of the nine prototypical sets of input values identified by the SOM. Pie charts associated with each unit show how many inputs of each object type won (i.e., fit the prototype of that unit). Some units within the SOM were selective to echo trains from only one type of object, showing that the features given to the SOM were sufficient for identifying a subset of objects. For instance, nine echo trains from the wooden object activated one unit at the bottom of the map (unit 3,2), and nine echo trains from the ceramic object activated an adjacent unit (unit 3,1). Thus, activation of either of these two units identifies the object that was ensonified. Echo trains from the copper object activated unique units on both the left and right sides of the SOM, effectively identifying that object for at least a subset of echo trains. The three SOM units along the top, however, each responded to echo trains from multiple objects. The overlap in these units suggests that several echo trains have amplitude and frequency properties that are not systematically associated with particular objects.

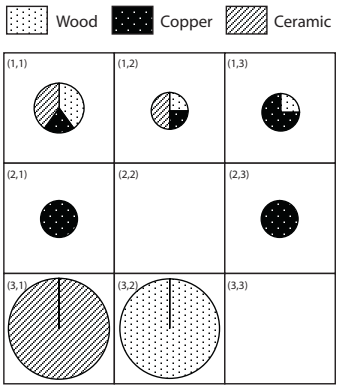

FIGURE 2. Results from a 3 x 3 SOM trained with all echo trains used in the human listening experiment. A depiction of units that represent prototypical echo trains from each object activated in the trained map. Pie chart size represents the number of echoes activating that unit and shading indicates the proportion of those echoes that were from each object. Numbers in parentheses indicate row and column numbers of units in the map.

Given SOM prototypes (the weight vectors of units in the SOM), participants in the human listening experiment may have picked up on the fact that the amplitudes of echoes coming from the wood and ceramic objects were higher than echoes coming from the copper object, or were varying less in peak frequency across echoes. Distinguishing wood examples from ceramic examples might then have been frequency-dependent, because the peak frequency for several echoes and mean peak frequency was higher for the weights of the SOM unit that ceramic examples activated. The weights of the (2,1) unit and (2,3) unit, where echo trains corresponding to copper objects won, had weight vectors containing both the lowest and highest values of mean frequency. If human participants in DeLong et al.'s (2013) experiment were using a decision rule such as "respond copper if the frequency is high" they would have not been accurate for many of the copper objects that had low frequencies. Participants performed greater than chance for all copper objects in the human listening experiment, suggesting that participants may have been using more sophisticated combinations of acoustic cues for identification (e.g., some combination of frequency and amplitude cues or differential weighting of individual echoes in the train).

SIMULATION 2

To more closely evaluate the acoustic information relevant for object identification within echo trains, we trained multilayer perceptrons to classify objects based on their echoes, and then examined how the networks accomplished this task. The advantages of using multilayer perceptrons compared to SOMs are that perceptrons are trained with feedback about object identity (like in the human listening experiment) and their accuracy can be examined and compared to human accuracy. If network accuracy matches human accuracy (i.e., identifies objects correctly above chance at all aspects) it could be the case that analysis of network structure reveals useful cues that were not reported by participants due to difficulty in verbalizing the cues.

Methods

Echo Representations

Representations used in Simulation 1 were also used in Simulation 2.

Network Architecture

Multilayer perceptrons had 40 input units, 30 hidden units, and 3 output units (Figure 3). Input units represented acoustic features of echo trains (described in methods section of Simulation 1), and output units corresponded to object categories. Each hidden unit's net input from connections with the input layer was converted to an activation level using a sigmoid activation function. In Figure 6, the dotted line in hidden units shows the point at which the sum of the weighted input is converted to an activity of 0. In this network architecture, weighted input exceeding .5 will be higher than 0 and input below .5 will be lower than 0. Weighted input from the hidden units is converted to

an activation level using a linear activation function in output units. Each hidden and output unit also had a bias (value added to the sum of weighted input to the unit). Twenty networks initialized with random weights were trained using a standard backpropagation learning algorithm (Rumelhart, Hinton, & Williams, 1986) with a learning rate of .05. All simulations were implemented in Neural Network Toolbox running in Matlab 2010a (MathWorks, 2010).

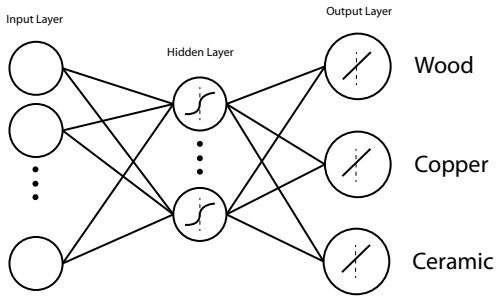

FIGURE 3. A depiction of the network architecture used in Simulation 2. Each input layer unit was associated with a feature extracted from echo trains. Input units had weighted connections to each hidden unit, and each hidden unit had a weighted connection with each output unit. Hidden units transformed the sum of the weighted input, plus a bias parameter, to an activity value with a sigmoid activation function. Output units operated similarly, except that they employed linear activation functions.

Network Training and Testing

Twenty ANNs were given the same number of training blocks, test blocks, and stimuli as participants in the human listening experiment (DeLong et al., 2013). Noise consisting of randomly generated values between ±0.2 was added to inputs before presentation to ANNs on each trial. This was necessary for ANNs to show variations in performance similar to those observed in humans. As with humans, networks were given feedback on all trials; weights were updated when the response generated in the output units differed from the target response. Target responses for each echo train were always set at 1 for the output unit associated with the object that generated the echoes and at 0 for the remaining two output units. The output unit with the highest activity was considered to be the object identity endorsed by a network on a given trial. The percent of correct identifications during each block was used as a measure of performance accuracy.

Analyzing Network Structure

Cue weighting. The absolute value of a weighted connection between an input unit and a hidden unit within the ANNs indicates how dependent hidden unit activity is on a particular input feature. Values close to zero indicate that hidden units are little affected by differences in the value of an input feature, whereas absolute values above zero indicate that the activation of hidden units is more strongly modulated by changes to that feature. The mean absolute values of weights between input features and hidden units were analyzed to determine which of the 40 acoustic features most strongly determined the identification of objects.

In the human listening experiment, participants reported using frequency and loudness cues, but we do not know from those interviews whether these cues varied in importance at different points in the echo trains. To get a sense of how different parts of the echo train and the features that described entire echo trains were weighted, mean absolute weight values were also calculated for the beginning, middle, and end of the train. To further evaluate the relative importance of frequency versus amplitude cues, as well as possible shifts in their usefulness across training sessions, we calculated the difference in frequency and amplitude weights post training and transfer testing.

Hidden unit activities. The ANNs created internal representations of echo trains via their hidden units in order to identify objects. It was the weights on these representations that ultimately determined the output values for any given echo train. Analyses of input weights only reveal the relative importance of input values. They do not reveal how ANNs use inputs to identify each object. To explore this issue, the most heavily positively and negatively weighted hidden units of each output unit for a single representative network were examined.

Results & Discussion

Training Sessions

Figure 4 shows the discrimination performance for the ANNs in the training sessions. Networks showed larger decreases in performance accuracy during the blocks where new echo exemplars were introduced than did humans (>80% in blocks 9 & 11). Nevertheless, on the last block of network training, performance accuracy was 80% or higher for all objects as it was with human participants (DeLong et al., 2013).

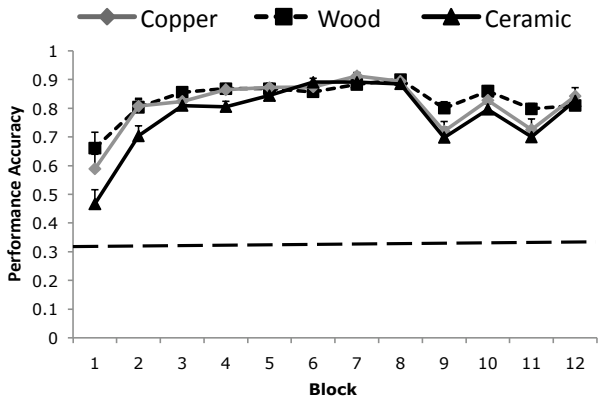

FIGURE 4. Discrimination performance in the training sessions for artificial neural networks (all training trials included only aspect 3). Error bars show standard error. The dotted line shows chance performance (33%).

Test Sessions

The ANNs successfully transferred object identification to the four novel object aspects. Figure 5 shows accuracy in the test sessions for the aspect ANNs were trained on (3) and the four transfer aspects (1, 2, 4, 5; see Simulation 1 methods for details of aspects). ANN performance matched human performance in that the networks correctly identified all three objects above chance at all five aspects.

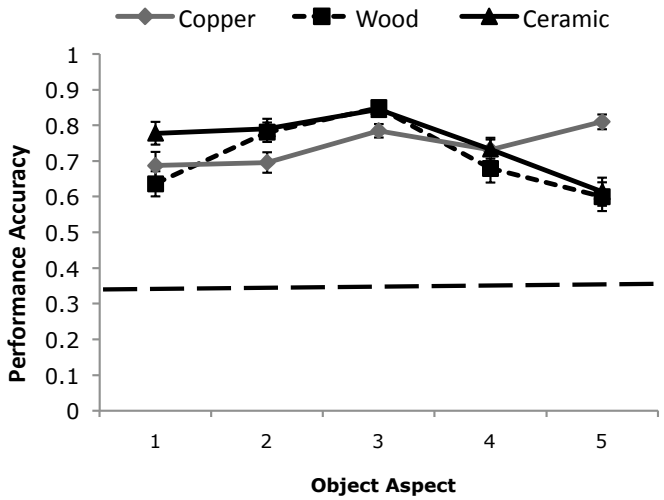

FIGURE 5. Accuracy in testing for artificial neural networks at all object aspects. The dotted line shows chance performance (33%). Error bars show standard error of the mean.

Acoustic Cues

Figure 6 shows the relative importance (mean absolute weight) given to input used by the ANNs to identify objects. It was not the case that all frequency information was weighted more heavily than amplitude information or vice versa. Rather, the activity of hidden units in ANNs was dependent on a combination of frequency and amplitude features.

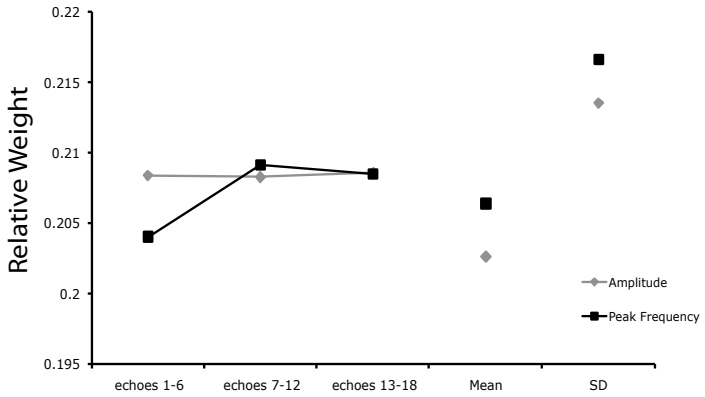

FIGURE 6. Analysis of the relative value of weights between input units and hidden layer units. Relative weights are shown for each third of the echo train along with mean and standard deviation features derived from the entire echo train.

Amplitude features were weighted similarly throughout echo trains (Figure 6). Peak frequency information from individual echoes, however, was weighted higher for the later portions of echo trains. Figure 6 shows that weights for the frequency of echoes 7-18 are higher than echoes 1-6. Figure 6 also shows that mean peak frequency was weighted more heavily than mean amplitude and that variation in frequency was weighted more heavily than variation in amplitude. The weighting of cues by networks is thus consistent with the weighting of cues reported by humans (i.e., more humans reported using frequency than amplitude). Also, comparisons of cue weighting before and after testing show that the ANNs learned to weight amplitude features more heavily during testing than was the case in training, as reported by humans (DeLong et al., 2013).

Hidden Unit Patterns

Table 1 lists the most positively and most negatively weighted connections for each output unit. The connections that maximally activated or maximally inhibited output units overlapped across objects. For example, the hidden unit that most strongly activated the output unit corresponding to the wooden object, simultaneously strongly inhibited the output unit corresponding to the copper object (Wood+/Copper-). This particular hidden unit responded to inputs with high amplitudes, high peak frequencies, and large variations in amplitude. Consequently, the ANN used these features as cues to both reject a train as being from a copper object and to accept the train as coming from a wooden object when the inputs are high. When the inputs are low, the reverse conclusion is reached. The hidden unit that most strongly inhibited activation of the wooden object unit was sensitive to these same cues, inhibiting this output unit most strongly when an echo train included low amplitude echoes, low peak frequencies, and little variation in amplitude. Other hidden units had similarly symmetric effects on the activation of output units indicating a copper or ceramic object (e.g., strong activation of the ceramic unit coupled with strong inhibition of the copper unit). The copper object output unit was triggered by echo trains with high peak frequencies and highly variable peak frequencies, whereas the ceramic output unit was activated by echo trains with low peak frequencies that varied little across echoes.

TABLE 1. The most positively and negatively weighted hidden units for each output unit

	Hidden Unit				
Input Feature	**Wood+/Copper-**	**Wood-**	**Copper+**	**Ceramic+**	**Ceramic-**
Mean amplitude	0.3838	-0.067	-0.1113	0.0391	-0.2189
Mean peak frequency	0.4182	-0.427	0.3296	-0.2885	0.0419
SD amplitude	0.3579	-0.3635	-0.0594	-0.0752	-0.2555
SD peak frequency	0.0706	0.1192	0.4054	-0.2793	0.3501

Note. Hidden units are labeled with respect to how output units weighted their activities. The object name indicates the hidden unit to output unit connection. Plus and minus symbols indicate whether that connection was positively or negatively weighted. For instance, the Wood- unit is a unit whose connection to wood was inhibitory (high activities inhibited wood identification). Input features determining hidden unit activations are also shown.

GENERAL DISCUSSION

In this study we supplement a recent human listening experiment by using artificial neural networks to investigate the cues that could be used to identify objects using echoes. SOM analyses (Simulation 1) showed that frequency, amplitude, and combinations of those features within an echo train can distinguish some objects. However, the SOM grouped some echo trains from different objects as all being similar. Separating these similar echoes may require the extraction of higher-level features involving the combination of frequency and amplitude cues that can only be identified when specific information about object identities is available (e.g., in the form of feedback). Consistent with this hypothesis, multilayer perceptrons (Simulation 2) were able to identify objects from echo trains at levels comparable to human participants and generalize identification to novel aspects in similar ways.

The way in which ANNs learned to identify objects was in some manners consistent with human self-reports of the features they used. For instance, more humans reported using amplitude during transfer testing and the multilayer perceptron networks similarly weighted amplitude more after testing than after training. Also, more humans reported using frequency than amplitude and networks weighted mean frequency more than mean amplitude. However, analysis of ANN input weights did not show a bias for weighing frequency information over amplitude information for individual echoes. The simulations suggest that differentially weighing the importance of features from each echo facilitates object identification. Tracking the variations in features across echoes may also be useful for identification as ANNs weighed the standard deviation of frequency and amplitude higher than most other cues. Few human participants reported using the pattern of change over time to identify objects in the listening experiment (DeLong et al., 2013), but it is possible that more picked up on these cues and did not how to verbalize them.

We cannot know for sure whether humans or dolphins use the same acoustic features that multilayer perceptrons in simulation 2 used, but the simulations clearly show that when ANNs differentially weight acoustic cues in the ways shown in these simulations, then this leads to the levels of performance and generalization profiles that are observed in humans. One way to experimentally assess whether the acoustic features used by humans and ANNs to identify objects from their echoes are also used by dolphins is through phantom echo studies. Phantom echo experiments involve projecting acoustic signals that mimic the echo returns that would have occurred if an object had actually been present. The advantage of this approach is that specific acoustic cues can be selectively filtered from (or added to) the phantom echoes. So, for example, the echo train recorded from a ceramic object might be normalized so that all amplitude information is removed or so that all frequencies are equalized in terms of energy. Furthermore, the acoustic features of some objects can be gradually superimposed on or morphed into echo trains from other objects so that the threshold at which categorical perception switches from one object to another can be identified (as has been done in speech perception studies). Given the large amount of cues one could vary in such experiments, the cues found by ANNs used in this study (e.g., a combination of high peak frequency and high variability in peak frequency for the copper object) could be tested initially, since we now know that these cues are useful for object recognition.

ACKNOWLEDGMENTS

This project was supported by a Research Seed Funding Grant to CMD from the Rochester Institute of Technology Office of the Vice President for Research, a Faculty Development Grant to CMD from the Rochester

Institute of Technology College of Liberal Arts, and an NSF grant (SBE 0542013) to the Temporal Dynamics of Learning Center.

REFERENCES

Au, W.W.L. **(1993)**. *The sonar of dolphins* (Springer, New York, NY).

Au, W.W.L. and Martin, D.W. **(1989)**. "Insights into dolphin sonar discrimination capabilities from human listening experiments," J. Acoust. Soc. Am. **86**, 1662-1670.

DeLong, C.M., Au, W.W.L, and Stamper, S.A. **(2007)**. "Echo features used by human listeners to discriminate among objects that vary in material or structure: Implications for echolocating dolphins," J. Acoust. Soc. Am. **121**, 605-617.

DeLong, C.M., Au, W.W.L., Harley, H.E., Roitblat, H.L., and Pytka, L. **(2007)**. "Human listeners provide insights into echo features used by dolphins to discriminate among objects" J. Comp. Psychol. **121**, 306-319.

DeLong, C.M., Au, W.W.L., Lemonds, D.W., Harley, H.E., and Roitblat, H.L. **(2006)**. "Acoustic features of objects matched by an echolocating bottlenose dolphin," J. Acoust. Soc. Am. **119**, 1867-1879.

DeLong, C.M., Heberle, A.L., Mata, K., Harley, H.E., and Au, W.W.L. **(2013, June)**. "Recognizing objects from multiple orientations using dolphin echoes," Paper to be presented at the 21st International Congress on Acoustics and the 165th Meeting of the Acoustical Society of America, Montreal, Quebec, Canada.

Gorman, R.P., and Sawatari, T. **(1985)**. "The use of multidimensional perceptual models in the selection of sonar echo features," J. Acoust. Soc. Am. **77**, 1178-1184.

Gorman, R.P., and Sejnowski, T.J. **(1988)**. "Analysis of hidden units in a layered network trained to classify sonar targets," Neural Networks **1**, 75-89.

Guillette, L.M., Farrell, T.M., Hoeschele, M., Nickerson, C.M., Dawson, M.R.W., and Sturdy, C.B. **(2010)**. "Mechanisms of call note-type perception in black-capped chickadees (*Poecile atricapillus*): Peak shift in a note-type continuum," J. Comp. Psychol, **124**, 109-115.

Helweg, D.A., Au, W.W., Roitblat, H.L., and Nachtigall, P.E. **(1996)**. "Acoustic basis for recognition of aspect-dependent three-dimensional targets by an echolocating bottlenose dolphin," J. Acoust. Soc. Am. **99**, 2409-2420.

Helweg, D.A., H.L. Roitblat, and P.E. Nachtigall. **(1993)**. "Using a Neural Network to Model Dolphin Echolocation," In N. Kasabov (ed.), *Artificial Neural Networks and Expert Systems* (IEEE Computer Society Press, Los Alamitos, CA), pp. 247-251.

Helweg, D.A., Roitblat, H.L., Nachtigall, P.E., Au, W.W.L., and Irwin, R.J. **(1995)**. "Discrimination of echoes from aspect-dependent targets by a bottlenose dolphin and human listeners," In R.A. Kastelein, J.A. Thomas, & P.E. Nachtigall (Eds.), *Sensory systems of aquatic mammals* (De Spil Publishers, Woerden, The Netherlands), pp. 129-136.

Helweg, D.A., Roitblat, H.L., Nachtigall, P.E., and Hautus, M.J. **(1996)**. "Recognition of aspect-dependent three-dimensional objects by an echolocating Atlantic bottlenose dolphin," J. Exp. Psychol. Anim. B. **22**, 19-31.

Kohonen, T. **(2001)**. *Self-organizing maps.* (Springer, Berlin, Germany).

Moore, P. W., H. L. Roitblat, R. H. Penner, and Nachtigall, P.E. **(1991)**. "Recognizing successive dolphin echoes with an integrator gateway network," Neural Networks **4**, 701-709.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. **(1986)**. "Learning internal errors by error propagation," In D. Rumelhart, J. McClelland, & the PDP Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol. 1: Foundations* (MIT Press, Cambridge, MA), pp. 318–362.

Wisniewski, M.G., Radell, M.L., Guillette, L.M., Sturdy, C.B., and Mercado, E., III **(2012)**. "Predicting shifts in generalization gradients with perceptrons," Learn. Behav. **40**, 128-144.