Manatee Call Detection

Govind Rathore
University of Florida
govinds@ufl.edu

***Abstract*—A manatee call detector can help warn the boats of their presence in the nearby waters. This way the boats can slow down to avoid any injury to endangered manatee species. However, detecting manatee call in the presence of background noise is a challenging task. We have explored application of LMS adaptive filters along with Time Delayed Neural Networks on the manatee call detection problem. We show that LMS, despite being simple, performs well on this task. In the end, we have also shown results of a CNN applied on the time delayed audio signal. This architecture performs overall best and should be explored further.**

Keywords*—*LMS, Time Delayed Neural Networks, Time Delayed Convolutional Neural Networks, Audio Signal Processing, Time-Series Analysis, Voice Detection

I. INTRODUCTION

Manatees are often injured or killed by the watercrafts. The number of manatee deaths caused by accidents is estimated to be as high as 24% of all manatee deaths [1]. In order to protect manatees and their habitats, the state of Florida has established more than quarter of million acres of manatee protection zones [2]. Some of these zones are round-the-clock idle speed zones. Such protection zones impact the economy by affecting the fishing and recreational water activities, sale of boating gear, limiting the construction of boating ramps, tourism related income and directly affecting the monetary value of surrounding residential property.

The purpose of this project is to detect manatee calls from real hydrophone recordings taken in an estuary. The proposal is to plant an array of hydrophones in the areas frequented by manatees at a fixed distance intervals. The detector would trigger a signal as soon as a manatee call is detected. This signal can then be interpreted by the boaters, who then should slow down to idle speed limit and maintain an attentive lookout until clear of the nearby area. This way the number of idle speed zones can be cut down while maintain or lowering the manatee accident rate. As the study of Cathy Steel [3] shows, the manatee calls are generated when manatees approach one another, submerge from breathing, and especially during play activities. The rate of manatee vocalization, approximately 1 to 5 times within a period of five minutes [4], can be used as a reliable source of detection.

We have used LMS adaptive filters and neural networks to detect the manatee calls. The rest of the work is outlined as follows. First, we explain the dataset and the problem statement along with the compute resources available to us in Section 2. We then provide a brief introduction of the methods used in this work in Section 3. We then, present results in Section 4 with conclusion in Section 5. We have listed possible future direction in Section 6 and References are present in Section 7.

II. DATASET/RESOURCES

We have three audio signals: train, noise and test of lengths 25, 2 and 30 seconds, all sampled at 48kHz. The data used for this project was collected by the Department of Biology at University of Florida. The training signal contains 10 sample manatee calls. Length of calls have a mean and standard deviation of 0.03 and 0.05 seconds. The test set contains manatee calls with background noise. All these manatee sounds are sharp high frequency squeaks. The background noise maybe of (fish, shrimp, marine mammals, and wave motion), boat noise (engine and propeller). In our particular case, the noise sounds like flowing water.

The frequency spectrum analysis shows that manatee squeaks can be distinguished from background noise given that the amplitude of such squeaks is higher than the background noise.

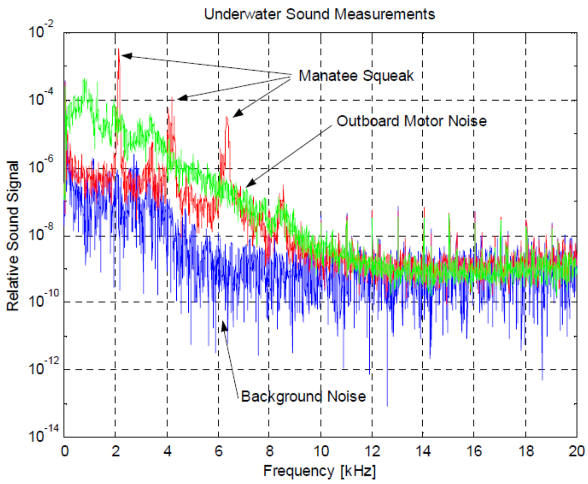

Fig.1: Underwater hydrophone sound spectrum showing distinct calls [5]

The ground truth for the given test data is created by us by manually listening to the audio. We have identified 16 manatee calls, out of which we have categorized 3 calls to be 'hard' to detect (we ourselves aren't sure if these are calls or noise). Hence ignoring these, our test set contains 13 manatees calls to be detected. We would like to clarify that the test data is never used for hyperparameter optimization as usual in machine learning world. We have created a new validation signal is created for tuning hyperparameters.

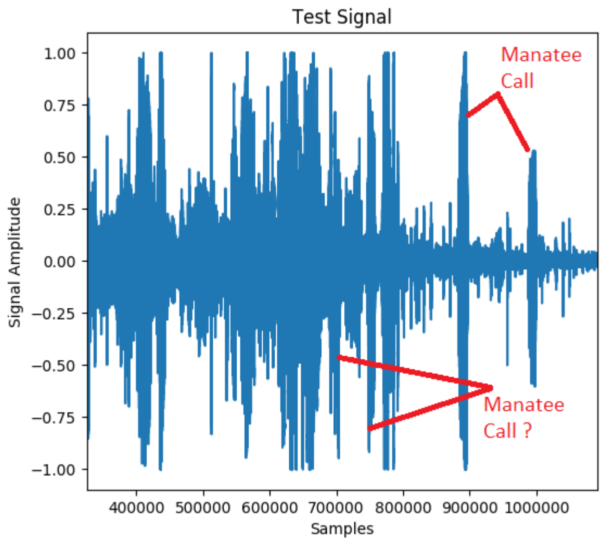

Fig.2: Close-up of test signal. Noise makes it hard to discern a Manatee call

The validation signal is created from the given training and noise signal. We have enlengthened the noise signal by appending it with itself multiple times and then overlaid it on the training signal. The validation signal generated this way, is 25 seconds long, and appears a lot like the test signal. This validation signal is used for selecting hyperparameters for LMS algorithm. For training the neural networks, we have used this validation signal as training signal.

All our experiments were performed on Intel i5-6300HQ 2.3 GHz Quad-Core machine with 8GB RAM and NVIDIA GeForce GTX 960M 4GB GDDR5 GPU.

III. METHODS

LMS

Least mean squares (LMS) algorithms [6] are a class of adaptive filter which adapt its filter coefficients to best be able to predict the desired signal. The filter coefficients are modified by stochastic gradient descent by computing the mean square between the true value vs predicted value of signal. The figure below shows the block diagram of an LMS adaptive filter being trained to estimate the next sample in each signal.

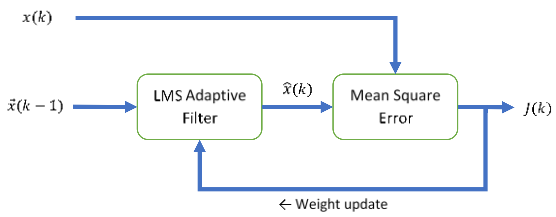

Fig.3: Block diagram of predicting a signal using LMS

The cost at any time step k is given by

$$J(k) = e^2(k) = \frac{1}{2}\left(x(k) - \vec{w}(k-1)^T \vec{x}(k-1)\right)^2$$

And the weight update equations are

$$\vec{w}(k) = \vec{w}(k-1) + \mu e(k)\,\vec{x}(k)$$

Where

$\vec{x}(k)$ Input vector $[x(k), x(k-1), \dots, x(k-m-1)]^T$

m Filter order

μ Step size

$\vec{w}$ Weight vector

From this, we trained two different LMS filters to predict the next sample in manatee calls that are noise-free and with-noise respectively. We then applied these two filters to predict the next sample in the given test sound. If the sound contained a manatee call then the LMS-noise would give higher error in prediction and if the sound contained noise then LMS-manatee would give higher error. Looking at the relative magnitude of errors produced by these two models for the given test sound, we can determine if the sound contains a manatee call. The block diagram of the adopted approach is given below:

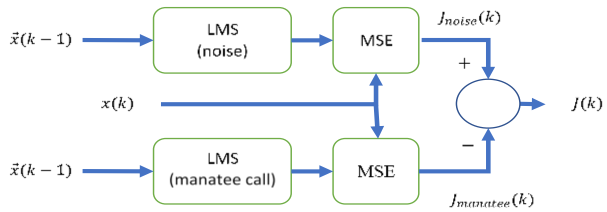

Fig.4: Proposed method for detecting calls using two LMS adaptive filters

Time Delayed Neural Networks

Time delayed neural networks have been shown to be able to model the temporal dynamics of a time-series and has been employed for phoneme recognition [7]. The main idea is to tap the sequential signal at some time intervals and feed it as a multi-dimensional input to neural network.

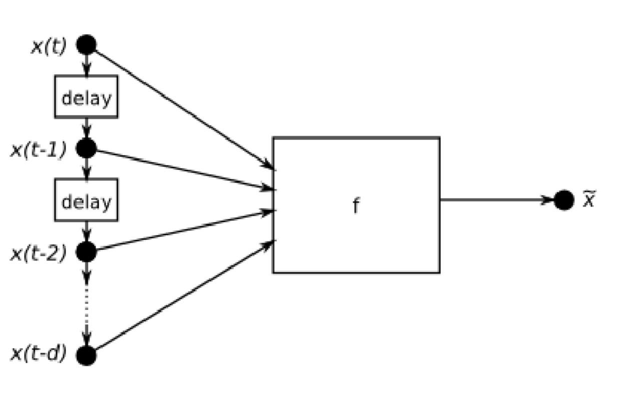

Fig.5: Time delayed signal to create a multidimensional input

We have experimented with neural network with varying number of hidden layers from 1 to 4 while also varying the number of units in each layer from 32 to 256. All hidden layers had ReLU activation. RMS Prop was used as optimizer and binary-crossentropy was used as cost function. The network was trained for 10 epochs with batch size 32 while saving only the weights with least training loss; no validation set was used.

The proposed network hence, is made to behave like a classifier, given a time window, it predicts if the window represents a manatee call. We could generate one label corresponding to each sample, but since the number of such samples would have been greater than 1 million, it would have taken a long time to train and test. To get past this issue, we have taken non-overlapping windows in the audio signal rather than sliding window. This way we have one label for 100 samples, representing if this window is part of a manatee call or not. This speeds up the run-time and still is a good emulation of real world detection since a manatee call is way lager in length than 100 samples.

Time Delayed Convolutional Neural Networks

Vijayaditya et. al. [8] have applied a unique time delayed neural networks architecture to Switchboard speech recognition task.

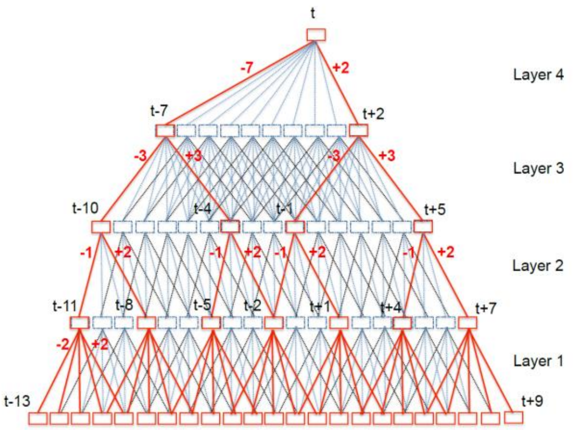

Fig.6: Time delayed NN used for Speech Recognition used by [8]

Our Time Delayed CNN architecture for manatee call detection is inspired from their work. However, their architecture has very specific interconnections and is not implemented by popular deep leaning frameworks. Hence, we have made a compromise with the number of weights by sharing weights for spatial locations between two layers. This can be achieved by applying a 1-D Convolution filter to the input feature map, a popular and easily available feature in all major deep leaning frameworks.

We have used four 1-D convolutional layers followed by two fully connected layers. The ReLU are used as activations for conv layers to avoid vanishing gradient problem, and sigmoid for fully connected layers towards the end to be able to classify the input as one of the two classes. Just like our Time Delayed Neural Network, this network was also trained for 10 epochs, batch size 32 with RMS Prop and binary-crossentropy as cost function. Only the weights with least training loss were saved for testing; no validation set was used. We couldn't test variations of this architecture in the limited time available to us and hence cannot comment on the effect of specific components on the performance.

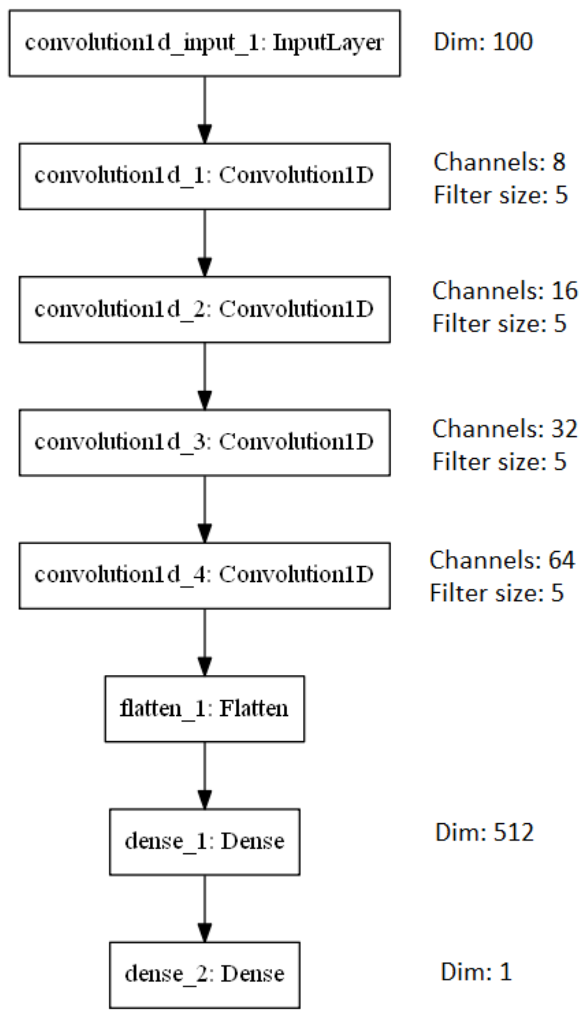

Fig.7: Our proposed Time Delayed CNN architecture

IV. RESULTS

LMS

To tune the best filter order for LMS, we have applied it on detecting manatee calls in our self-created validation signal. As expected the performance (measured in terms of AUC) improves with increasing filter size and saturates then onwards. We have tested LMS on validation set with a filter size of 1 to 50 at fine intervals. Our best suitable filter order for test set is around 10 which gives 0.88 AUC. This result is better than expected since LMS is a simple linear adaptive filter and is not expected to learn the highly non-linear relationships between audio samples.

The performance rises sharply between 1 to 3 filter order, and beyond which the improvements slow down. This could be attributed to the fact the LMS has the imited ability to represent the highly varying frequency characterizations

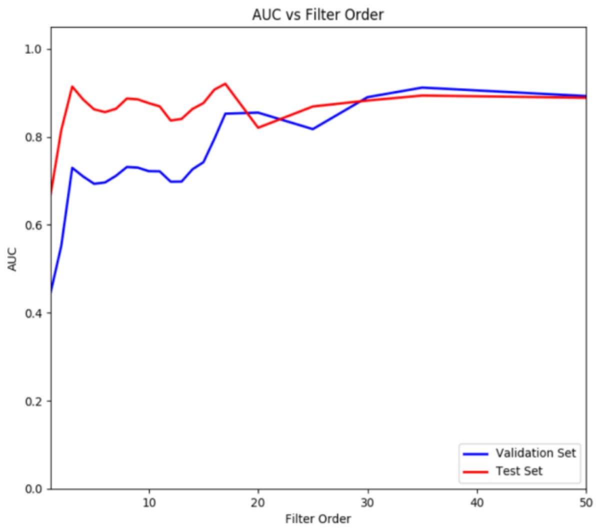

Fig.8: AUC vs LMS filter-order. On validation and test signals

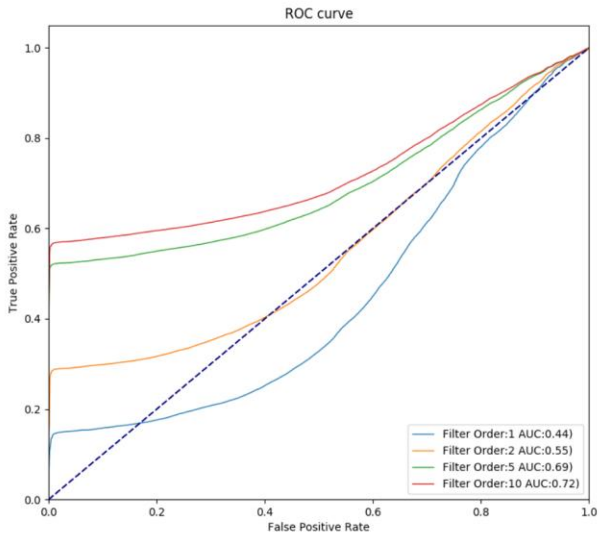

Fig.9: ROC curve vs filter order on validation set

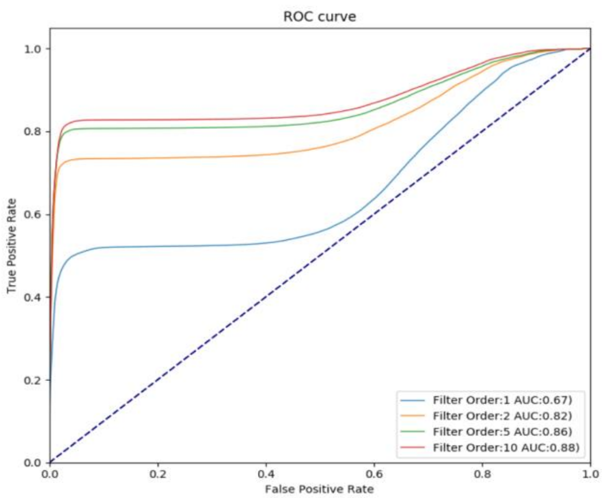

Fig.10: ROC curve vs filter order on test set

caused by the background noise overlaid on top of manatee calls.

For a same filter order, the LMS performance is lower on the validation set, this can be attributed to the small length of noise signal that is appended with itself before being overlaid on top of training signal. Thus, the validation signal has higher amount of noise than actual test set.

The AUC on the validation set improves with increments in filter-order. However, LMS could only achieve an AUC of 0.72 with the filter order of 10. When the same weights are applied on test set, we get a better detection with an AUC of 0.88 for filter order of 10. The performance stagnates beyond this limit.

From the ROC curve, the LMS can remove the noise from test signal (true positive rate is > 0.5 even with near zero false positive rate), however it is unable to detect the manatee calls (true positive does not improve significantly despite increasing false positive rate).

Time Delayed Neural Networks

We expected the TDNN to perform at least as good as LMS. The neural network does detect manatee calls as can be concluded from the high true positive rate achieved even with 0.5 false positive rate.

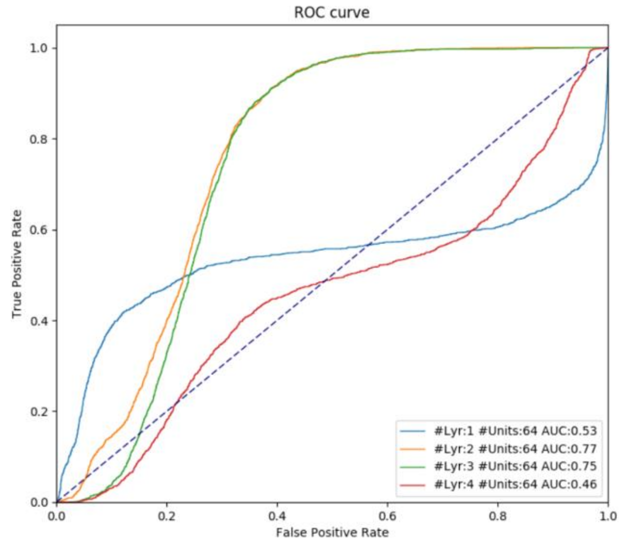

Fig.11: Effect of increasing layers while fixing number of units to 64

However, it mistakenly identifies noise samples as manatee calls when the threshold is lowered. This behavior is consistent irrespective of the capacity of the network. The best network was only 2 to 3 layers deep, and beyond which the performance decreased. This could be caused by the newly introduced free parameters with each additional layer. With this increased capacity, the network must have overfit on the training data and could not generalize well on the test data.

When we increased the capacity of each layer, while keeping number of layers fixed, we observed that a higher number of units was better for the network. Unlike our previous experiment where each layer added a new non-linear layer, we don't see the effect of overfitting here. Although, we're certain that if we had trained the network for more epochs, the network would be able to overfit.

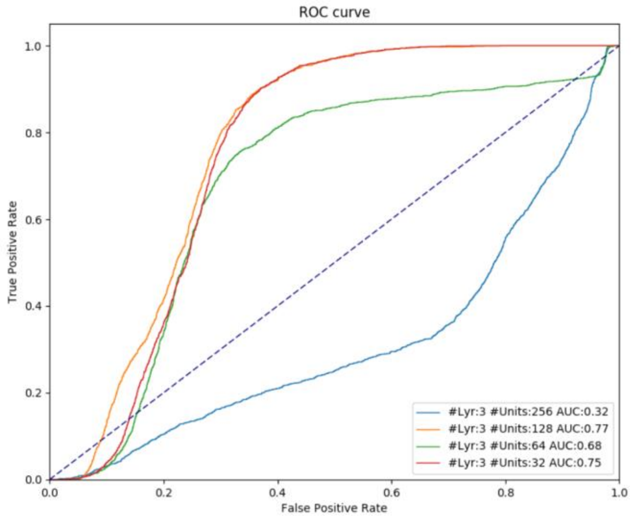

Fig.12: Effect of increasing number of units while keeping number of layers fixed to 3

Overall, although the neural network requires more compute resources than LMS, we did not see a significant improvement in the overall AUC score compared to LMS.

Time Delayed Convolutional Neural Networks

We have trained our network on mixed signal for 10 epochs, which achieved 99.2% accuracy on training data and took around 54 seconds. With these weights, we got an AUC of 0.98 on the test signal in 6 seconds.

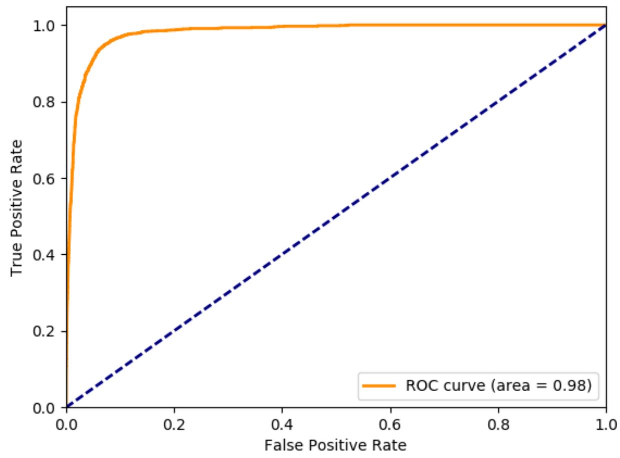

Fig.13: ROC curve for Time Delayed CNN

These results can be attributed to the fact that we rather than predicting the next signal (regression problem) we have converted the detection problem to a classification problem. And neural networks are known to perform well on classification tasks because of the limited span of solution space. Also, CNN's filters can model the temporal relationship in a sequence better than plain neural networks where we leave it to the network to learn that the input vector is a time series.

V. CONCLUSION

Time delayed CNN are superior to LMS in detecting manatee call in a noisy environment. However, LMS performance was found comparable to plain Time Delayed NN and it takes comparably less time to get the weights from LMS. For our case, the training time for TDNN could be significantly large if we had trained it on CPU. On the testing side, the time taken is almost same for both LMS and neural network methods. The comparable time of TDNN can be attributed to the fact that we are generating a label for a window of 100 samples, whereas in LMS we are predicting a class for each sample. In terms of performance, TDNN outperform the LMS by a huge margin.

VI. FUTURE WORK

We have only tested our filters on a test signal which contained just one type of background noise. However, in the real world the noise can be caused by other factors, such as other fishes, water splashing, watercraft motor sound etc. We should also look into other forms of neural networks such as LSTMs for detection as having a long memory would help based on a sample which is farther than our window size.

VII. REFERENCES

[1] Florida Fish and Wildlife Conservation Commission, 2006b, http://research.myfwc.com/features/view_article.asp?id=12084

[2] Farren, R., (2003), "State of Florida has Created More than One Quarter of a Million Acres of Manatee Protection Zones," Seawatch, Issue # 94, pp. 1-4.

[3] Steel, C. (1982), "Vocalization patterns and corresponding behavior of the West Indian manatee (Trichechus manatus)," Ph.D. Dissertation, Florida Institute of Technology, Melbourne Florida.

[4] Bengston J. L. and Fitzgerald, S. M., (1985), "Potential Role of vocalizations in West Indian manatees," J. Mamm. 66(4), pp. 816-819.

[5] A Manatee Avoidance System, Final Report 2006, State of Florida Fish and Wildlife Conservation Commission http://myfwc.com/media/201856/Final_Report_FFWCC_2006.pdf

[6] Least mean squares filter, Wikipedia. https://en.wikipedia.org/wiki/Least_mean_squares_filter

[7] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, Mar. 1989.

[8] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," INTERSPEECH, pp. 2–6, 2015.