Detecting Bioacoustic Signals at Cal Poly Pier with Machine Learning

Anagha Sikha, Sophia Chung, Sucheen Sundaram, Maddie Schroth-Glanz, Hunter Glanz, Jonathan Ventura
Department of Statistics, California Polytechnic State University, San Luis Obispo

INTRODUCTION

- Passive acoustic monitoring allows scientists to listen in on marine animal behaviors and determine the sources of sound in marine ecosystems.
- The Marine Acoustics Research Team at Cal Poly SLO, led by Professor Schroth-Glanz, spends hours listening to audio files to manually detect and classify animal sounds along the Central Coast of California.
- To improve efficiency, our research enhances a machine learning pipeline to automatically detect marine animal signals from data collected at the Cal Poly Pier.
- By training on more data and refining post-processing methods, we aim to optimize model performance and advance understanding of marine animals.

PAST WORK

- Created a background noise file by removing marine animal signals from a single audio file and used this as training data
- Employed a Variational Autoencoder (VAE) that compresses and reconstructs background noise input to resemble the original audio file closely
- Applied a Short-Time Fourier Transform (STFT) to localize frequency over time and allow for visualization using spectrograms
- Used Per-Channel Energy Normalization (PCEN) to enhance the visibility of marine animal signals by balancing amplitude across different frequency bands

DATA

Hydrophone Recordings from Cal Poly Pier at Avila Beach

- 706 audio files - 30 minutes each

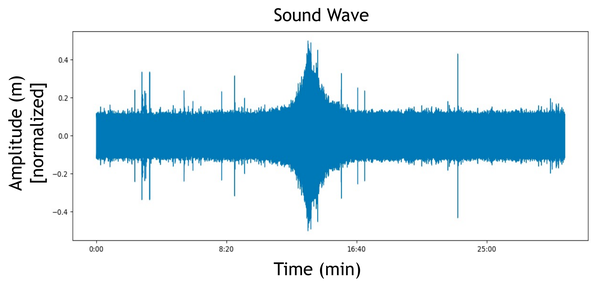

Fig 1. Sound wave of a single Cal Poly Pier audio file

- 39 annotation files
  - Manually recorded text files with each row containing the time and frequency ranges of a marine animal call
  - Each corresponds to one audio file
  - Use for training the ensemble
- Stored in Amazon Web Services (AWS) Simple Storage Service (S3)

METHODS

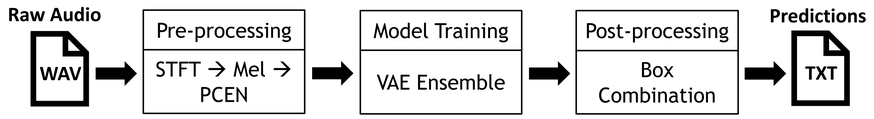

Fig 2. Pipeline of detecting signals from the inputted raw audio file

MEL SPECTROGRAM

- Convert frequencies from the linear hertz scale to the logarithmic mel scale to account for human auditory perception - not yet successfully implemented

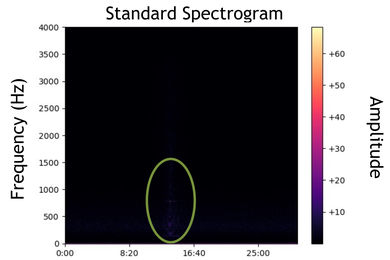

Fig 2. Standard spectrogram of a single Cal Poly Pier audio file

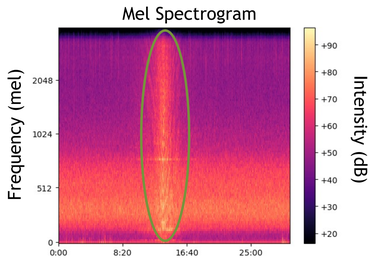

Fig 3. Mel spectrogram of a single Cal Poly Pier audio file

VAE ENSEMBLE

Fig 4. A single VAE model in the ensemble

- Divide training across 10 VAE models to overcome computing and time limitations
- Use all 39 background noise files for training
- Train each model on around two hours of data (three or four files)
- Output a text file with the predicted boxes' time and frequency ranges

BOX COMBINATION

1. Use an agglomerative hierarchical clustering algorithm with a 0.3 distance threshold to form similar clusters based on time and frequency ranges
2. Combine all boxes in a cluster into one box by taking the intersection: maximum beginning time, minimum end time, maximum low frequency, and minimum high frequency

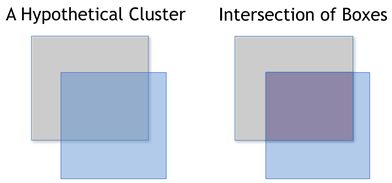

Fig 5. Box combination takes the intersection of the boxes

RESULTS

- Calculate intersection over union (IoU) to evaluate the overlap between all pairs of predicted boxes and ultimately obtain metrics

Table 1. Aggregate metrics for the 39 files

	Number of Predicted Boxes	Accuracy	Precision	Recall	F1
Before Box Combination	503.64103	0.00076	0.00077	0.00675	0.00137
After Box Combination	177.64103	0.00214	0.00216	0.00651	0.00302

CONCLUSIONS

- Our pipeline outputs suboptimal results.
- However, there is a large improvement from before the box combination to after:
  - The average number of predicted boxes was reduced by about 65%.
  - Accuracy: Out of all predicted boxes, 0.214% matched an annotated box.
  - Recall: The pipeline correctly identifies 0.651% of annotated boxes.
- We believe this suboptimal performance is due to the large number of predicted boxes, despite our efforts to combine and decrease the number of boxes.

FUTURE DIRECTIONS

- Refine the current pipeline to use the mel scale in pre-processing and further minimize the number of predicted boxes
- Alternatively, completely switch to Convolutional Variational Autoencoder (CVAE) or Convolutional Neural Network (CNN)
- Classify the detected signals into distinct groups (whale calls, other marine animal signals, human-generated noises, etc.)

ACKNOWLEDGEMENTS

- Thank you to Professor Maddie Schroth-Glanz for her guidance throughout this project.
- Thank you to our faculty advisors, Dr. Ventura and Dr. Glanz, for their support.
- Thank you to BCSM for hosting this research conference.

REFERENCES

Devin Levin, Jason Mulson, Nick Gammal (2023): Detecting Marine Acoustic Profiles with Deep-Learning Denoising

Doshi, K. (2021, February 18). Audio Deep Learning Made Simple (Part 2): Why Mel Spectrograms perform better. Medium. https://towardsdatascience.com/audio-deep-learning-made-simple-part-2-why-mel-spectrograms-perform-better-aad889a93505