Classifying dolphin whistles using convolutional neural networks

Genevieve Flaspohler, Tammy Silva, Aran Mooney, Yogesh Girdhar

Massachusetts Institute of Technology and the Woods Hole Oceanographic Institute Joint Program

Acoustics ’17 Boston - 28 June 2017
Human speech recognition systems like Siri and Google Now can achieve ~95% accuracy using neural network-based recognition.

Hypothesis: Neural network methods from Automatic Speech Recognition (ASR) can be adapted to classify dolphin whistles.
Machine learning as function approximation

$$f: \text{input} \to \text{category}$$

$$f(\text{dog}) = \text{dog}$$

$$f(\text{cat}) = \text{cat}$$

How do we come up with $f(\cdot)$?

1. Task-specific features, learn $f(\cdot)$
   (Previous work)

2. Standard features, learn $f(\cdot)$
   (SVMs, Feed-forward NN, etc.)

3. Learn features and $f(\cdot)$ directly
   (Convolutional neural methods)
Task-specific Features:
Based on 10 features*:
1) Start frequency
2) End frequency
3) Minimum frequency
4) Maximum frequency
...

*Oswald, Julie N., et al. (2007) [3]

MFCC Features*:
A 26-dimensional vector that represents the power in each frequency band
Filter bank designed to emulate human hearing physiology


Full Spectrogram data:
Calculated using a sliding window FFT
1024-pt FFT on sliding Hanning window with 50% overlap.
Contains “all” data available in signal
Convolutional neural network (CNN)
Machine learning dataset

8,632 total vocalizations
339 “clear” whistles

Data generously provided by Tammy L. Silva ad T. Aran Mooney [1]
Machine learning dataset

- 339 total whistle clips
- Classified into 11 categories
- Random shuffle x 10
- 80% training clips
- 20% testing clips
CNNs attain maximum classification accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>Feature Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear discriminant analysis</td>
<td>17.2% in 11-way classification*</td>
<td>Task specific</td>
</tr>
<tr>
<td>Decision trees</td>
<td>20.6%*</td>
<td>MFCC</td>
</tr>
<tr>
<td>Linear discriminant analysis</td>
<td>59.7%</td>
<td>Spectrogram</td>
</tr>
<tr>
<td>Decision trees</td>
<td>61.5%</td>
<td></td>
</tr>
<tr>
<td>SVM polynomial kernel</td>
<td>82.8%</td>
<td></td>
</tr>
<tr>
<td>Feed forward NN</td>
<td>83.7%</td>
<td></td>
</tr>
<tr>
<td>Logistic regression</td>
<td>76.8%</td>
<td></td>
</tr>
<tr>
<td><strong>Convolutional NN</strong></td>
<td><strong>85.9% average accuracy</strong></td>
<td></td>
</tr>
</tbody>
</table>

*vs 33.5 and 33.6 reported for 8-way classification in Oswald, Julie N., et al. (2007)
CNN performance is label dependent

- High classification accuracy in classes with many examples
- Low accuracy for classes with few examples

Performance should continue to improve with additional training examples
Code:  https://gitlab.com/warplab/dolphin-lang
Website: https://warp.whoi.edu
References


Feed-forward neural network