Benchmarking neural network models for acoustic behavior with vak and VocalPy

David A. Nicholson[1], Yarden Cohen[2]

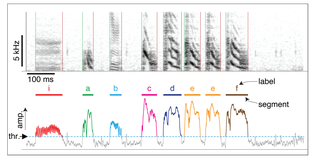
[1]Independent researcher, Baltimore, Maryland, United States of America

[2]Department of Brain Sciences, Weizmann Institute of Science, Rehovot, Israel

Neural Syntax Lab

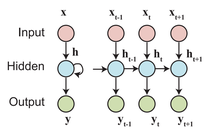
HH22

Introduction

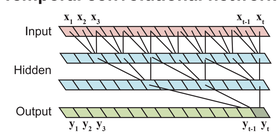


- Understanding the neural basis of acoustic behavior requires automated analyses.
- Recent work automates segmentation and annotation of animal sounds with neural network models.
- Remains unclear which models perform best for this task:

Recurrent networks



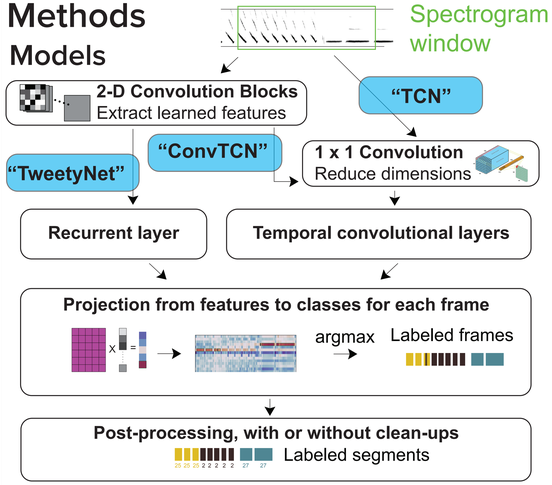
Temporal convolutional networks



Here we compare them with a neural network framework, vak, and VocalPy, a core package for acoustic communciation in Python.

Methods

Models



Window Size

- 176 frames (Cohen et al. 2022)
- 2000 frames

Loss functions

Classification

Cross entropy

$$\mathcal{L}_{ce} = \frac{1}{T}\sum_{t} -\log\left(y_{t,c}\right)$$

Smoothing losses

Dataset

Bengalase Finch Song Repository

4 birds, 36 syllable classes

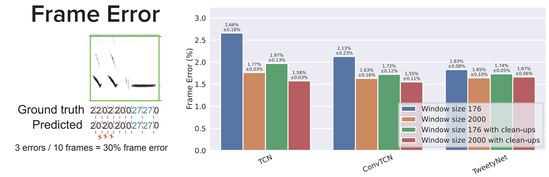
Training set: 3600s total

5 subsets of 1800s

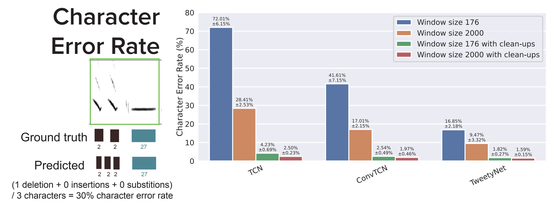
Test set : 1600s

Increased window size improves perfomance

Frame Error

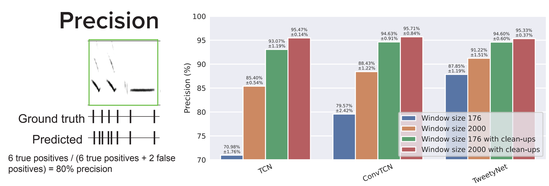


Character Error Rate

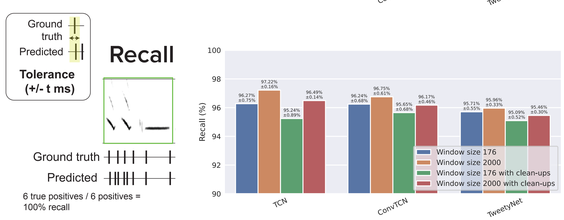


Segmentation metrics

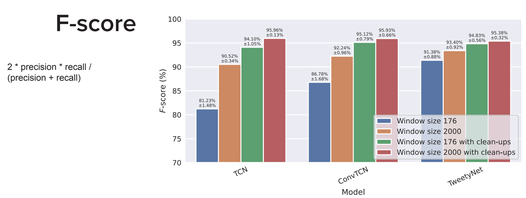
Precision



Recall



F-score



Smoothing losses mitigate oversegmentation

Smoothing losses

Temporal MSE

$$\mathcal{L}_{TMSE} = \frac{1}{TN}\sum_{t,c}\tilde{\Delta}^2_{t,c}$$

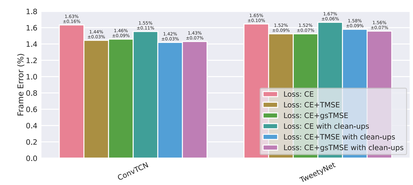
$$\tilde{\Delta}_{t,c} = \begin{cases} \Delta_{t,c} & : \Delta_{t,c} \leq \tau \\ \tau & : otherwise \end{cases}$$

$$\Delta_{t,c} = |\log y_{t,c} - \log y_{t-1,c}|$$

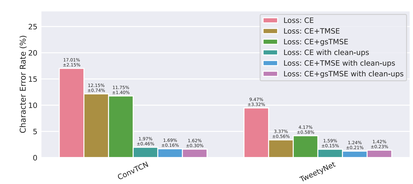
Gaussian Similarity Temporal MSE

$$\mathcal{L}_{GS-TMSE} = \frac{1}{TN}\sum_{t,c}\exp\left(-\frac{\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2}{2\sigma^2}\right)\tilde{\Delta}^2_{t,c}$$

Frame Error

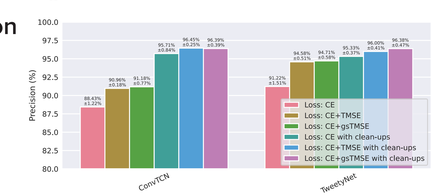


Character Error Rate

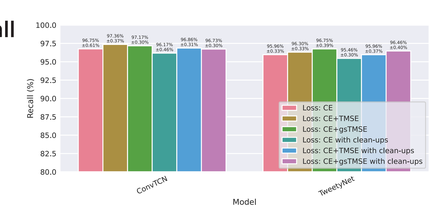


Segmentation metrics

Precision



Recall



Conclusions and Future Work

- Our results provide a case study of how vak and vocalpy make it easier to compare different neural network models for automated analysis of acoustic behavior within a single framework
- Results here suggest that researchers using these models to automate annotation of animal sounds should use much larger window sizes to train models
- Intuitively, increasing window size provides additional context to the neural newtork model
- Improvement from larger window size comes at a cost of slowing down model training and inference (not demonstrated here)
- Analysis with segmentation metrics shows that temporal convolutional networks tend to oversegment
- Adding a smoothing term to the loss function mitigates this oversegmentation, and further improves character error rate for both recurrent and temporal convolutional networks.
- Future work will test these results on multiple species and extend our understanding of the role of different architectural features
- We encourage other researchers to replicate and build on our results using vak and VocalPy

vak VocalPy

For more results and analysis, please see:

https://github.com/vocalpy/Nicholson-Cohen-SfN-2023-poster

References

- Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018).
- Farha, Yazan Abu, and Jurgen Gall. "Ms-tcn: Multi-stage temporal convolutional network for action segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- Guirguis, Karim, et al. "SELD-TCN: Sound event localization & detection via temporal convolutional networks." 2020 28th European Signal Processing Conference (EUSIPCO). IEEE, 2021.
- Ishikawa, Yuchi, et al. "Alleviating over-segmentation errors by detecting action boundaries." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.
- Steinfath, Elsa, et al. "Fast and accurate annotation of acoustic signals with deep neural networks." Elife 10 (2021): e68837.
- Cohen, Yarden, et al. "Automated annotation of birdsong with a neural network that segments spectrograms." Elife 11 (2022): e63853.
- Nicholson, David, and Yarden Cohen. "vak: a neural network framework for researchers studying animal acoustic communication." Python in Science Conference. 2023

Acknowledgements

Supported by Israeli Science Foundation grant ISF-2401/2022

ISRAEL SCIENCE FOUNDATION