Astronomy & Astrophysics manuscript no. output
October 31, 2019

AADHAAR

Adaptive Acoustic DSP Help for Avian Audio Recognition

Ajay Vikram Singh[1,⋆]

National University of Singapore
e-mail: e034189@u.nus.edu **

Report Submitted: 31 October, 2019

ABSTRACT

Context. Various interest groups including ecologists require a robust, mass deploy-able, low cost solution to identify bir presence in space-time dimensions. Signal processing and advances in Machine learning have the potential to address above need.
Aims. Explore and mature an accurate, automated, and generalized Avian presence classifier based on avian audio signal detection. Classifier aim is binary: Presence or Absence of avian presence, with no species specificity or avian density estimation
Methods. 3 iterative step exploration: Signal Validation Transformation, Feature Extraction, Machine learning Model Architecture. Performance evaluation using a mixture of confusion matrix and Receiver Operator Characteristic (ROC)-Area Under Curve(AUC).
Results. Achieved an Accuracy of 84% and AUC of 74% on unseen data. Industry best AUC 89%. Thoughtful model design to audio signal and problem definition found beneficial vs a complex neural network model (# of params).
Conclusions. Audio based Avian detection classifier has improvement potential for possible field deployment. Approach explored in project opens path to other novel applications in avian ecology management, environment conservation, education and machine auditory perception based industrial applications

Key words. Avian Audio – Binary Classification – Frequency-Temporal Power Spectrum – MFCC CNN – ROC AUC

1. Introduction

Avian population has reduced due to rapid urbanization and pollution. Avian Population reduction has primary and secondary derivative effects on ecology, human society, and business. Protection and conservation groups need verifiable data to take effective policy and monetary decisions. Current market solutions based on standard signal processing techniques, as investigated by Marques (Marques 2013) render a data based decision making ineffective. Further, Borker (Borker 2014) recommends that vocal activity i.e. presence of avian audio signals, is a viable avian population metric. In Summary, a solution to above problem should aim to satisfy multiple or all of following constraints:

- cost effective - sensor/processing/human costs,
- highly automated - needing no manual calibration,
- Generalized - species agnostic,
- Noise agnostic - Wind, Rain, Human(imitation),
- bird songs - tonal, longer,
- bird calls - shorter, alerts, non-tonal/tonal,
- polyphonic/monophonic agnostic,
- distant sounds - signal:noise power spectrum

The project investigates addressing above needs through digital audio digital signal processing and subsequent deploying machine learning techniques.

2. Related Works

Related works exploration and investigation was with the intent of both understanding the best practices as well as well possible avoidance of wheel reinvention given the short time duration of the project. Project setup steps being:

- Data sets and data transformation
- Audio feature extraction,
- Machine learning models

Stowell (2018) conducted an open challenge for Bird audio detection. The challenge website provided with a rich dataset base. PaulDavid (2019) Praat is a helpful tool for manual audio spectral appreciation and analyisis. Review of Audio DSP modelling at Adams (2019) Grill (2017) Dan (2017) to understand the computing optimization feasibility in Digital Audio transformation. lyons (2019) provided necessary thinking for audio feature selection for data preperation. Schlüter (2017) Cakir (2016) Adavanne (2017) works was investigated to understand alternate approaches to integrated Digital Signal processing and usage of machine learning possibilities. Hinton (2012) Li (2019) Shaobo Li (2018) works provided rationale for Deep Neural network

⋆ part-time: https://www.linkedin.com/in/ajaynetwork/
** personal id: ajayvsingh@gmail.com

based exploration as well as nuaunces for model structuring. Lastly and importantly Stowell (2018) summary on the challenge participants techniques, relative performances provided a glimpse on model exploration dimensions as well as solution evaluation techniques deployed in the (contextual) challenge.

3. Proposed Approach

Exploration and investigation spread in following dimensions, that collectively contribute to project effectiveness:

- Representative Data sets,
- Digital Signal (Audio) transformation,
- Audio feature extraction and summarizations,
- Machine learning models,
- evaluation mechanisms considering the solution constraints discussed in Introduction section

First, we had a well curated dataset due to challenge conducted at Stowell (2018). Of the two annotated datasets, we chose the ff1010 data-set. Key reasons for data selection are:

- sampling rate uniformity at 44.1 kHz,
- limited corrupt files,
- uniform temporal sample size 10,000 milli seconds

PaulDavid (2019) was a handy tool for visually understanding data quality including power spectral features.

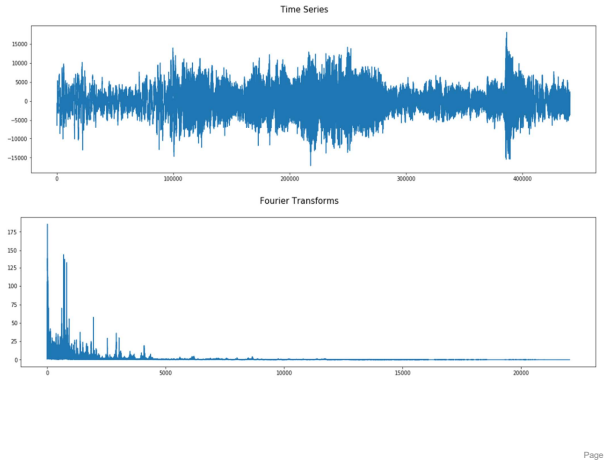

Fig. 1. Sample 10,000 ms, 44.1 kHz sampling rate Audio D-Fast Fourier Transform

FFT transformation inspection revealed that 0-8/10 kHz was the frequency range of interest (figure 1). Hence, considering Nyqst criteria, we down-sampled the data to 2x10=20kHz. This has a beneficial effect on model experimentation as well as implementation.

Basis review of Cakir (2016) Grill (2017) Li (2019) MFCC based feature extraction was chosen. Here we chose a triangular 40 filter band over 0-10kHz range. We took 40 ms window size, 20 ms overlap, leading in a 40(MFCC

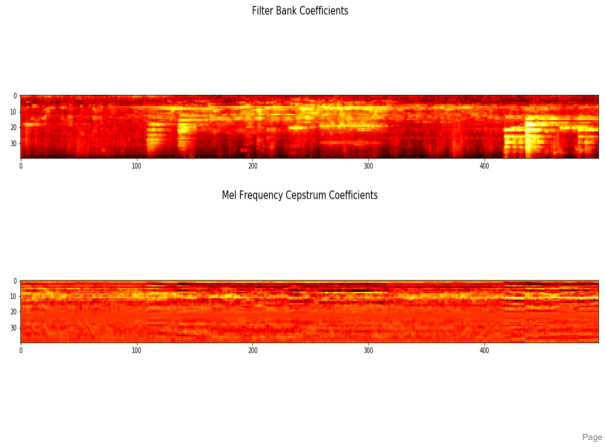

Fig. 2. Sample 10,000 ms audio 40 triangular Filter MFCC features

filters), 499 (window slots) data frame (40x499) per audio sample. See figure two for a representative random audio file sample. See figure 2.

Reviewing multiple literatures including Hinton (2012), Cakir (2016), Adavanne (2017) brought the benefits of exploring convolution neural architecture in feature extraction. Additionally, esp for bird songs presence, a temporal continuity brings sequential learning based feature extraction. Highest performing models in Stowell (2018) use varying combinations of MFCC based power spectrum summarization followed by CNN/RNN/CRNN and other permutations.

Net.pdf

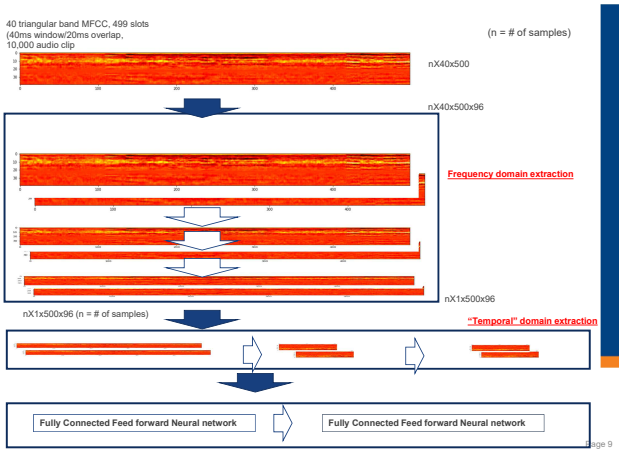

Fig. 3. Model Architecture Concept

I experimented with multiple variants of Convolution neural networks and CRNNs. Here, following considerations were made:

- number of parameters,
- model training complexity,
- over-under fitting
- model convergence

– capturing frequency temporal features

Figure 3 captures the finalised model concept, while Figure 4, captures the best performing model architecture details. Total number of parameters is approximately 300,000 in the final model.

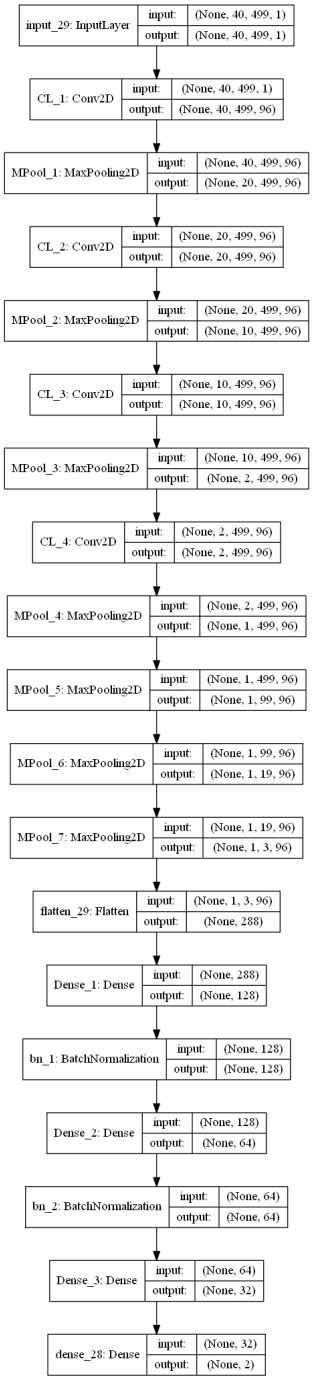

Fig. 4. Finalized Convolution Neural network

4. Experimental results discussion

The problem is a two class classification. We chose a combination of Performance evaluation using a mixture of confusion matrix and Receiver Operator Characteristic (ROC)-Area Under Curve(AUC).

On confusion matrix, Precision and recall are both desirable characteristics depending on the end user focus and objective. ROC-AUC on the other hand provides a robust performance metric. The problem suffers from natural class imbalance which has both space and time (e.g. of the year) variance .

ROC at 74% compares with best figure of 89%

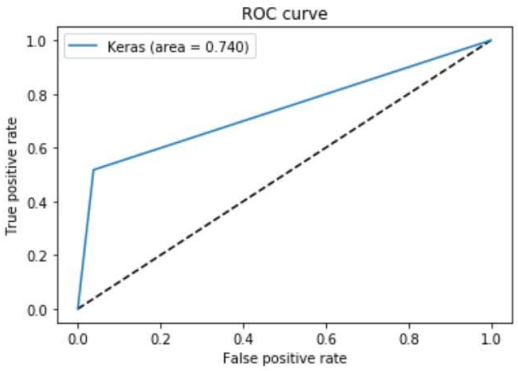

Fig. 5. ROC-AUC Curve & Scores

Overall accuracy figures are satisfactory on both the classes. The accuracy figures are generated on "unseen data". Overall, as both recall and precision are of interest for different possible users, we have to work on improve Type 1 and Type 2 erros. See Fig 6.

```
Best accuracy (on testing dataset): 84.42%
               precision    recall  f1-score   support

No Bird Audio     0.8469    0.9620    0.9008       552
   Bird Audio     0.8306    0.5176    0.6378       199

  avg / total     0.8426    0.8442    0.8311       751

[[531  21]
 [ 96 103]]
```

Fig. 6. Confusion Matrix

The model training achieved saturation at 50-100 epochs. And no further improvements was possible even after changing initialization as well as learning rate progressions. See figure 7

Initial training was finding the local loss minima quite quickly and across a wide variety of hyper parameters. In conclusion, the model responded well to SGD optimizer with a reasonably slow learning rate. Traditionally well per-

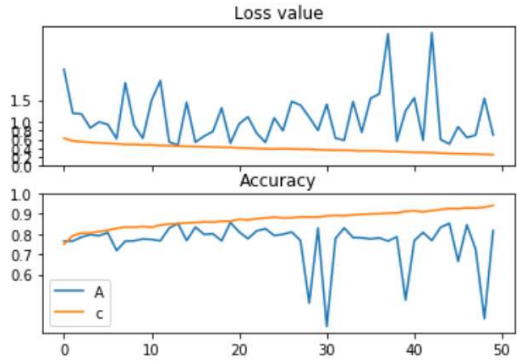

Fig. 7. Accuracy training

forming adam optimization and also to a range of learning rate setups didn't yield.

Large kernel size based feature extraction didn't yield good results. A kerne size of (3,3) was found optimal to learning response.

Regularization improved the model performance both in terms of learning start as well as learning saturation in reasonable number of epochs (50-100 epochs was ideal with loss and accuracy on validation data saturating).

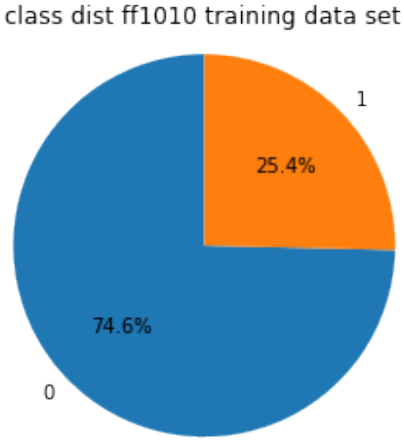

Fig. 8. ROC-AUC Curve & Scores

The data set presented has class imbalance to nudge the model to not penalize saturation with majority class. I didn't try under-sampling majority class considering the limited number of data. Synthetic data generation exploration as well as selective noise addition is a next step to improve the model learning outcomes.

5. Conclusions

1. The project got satisfactory results. The best model achieved accuracy in access of 84% on un-seen data. Development accuracy was in excess of 95
2. The pipeline created was intelligent in storage. Spectral characteristics by random manual sampling afforded down-sampling to 20kHz from 44.1 KHz with no loss in accuracy or robustness between the two sample sets.
3. The most efficient model is quite frugal with 15x less training parameters that most elaborate model trained, at the same time with much better AUC.
4. The model showed temporal resilience as the audio samples had short bursts of expected signal with no fixed pattern (e.g. bird calls).
5. The model showed power spectral resilience detecting amplitude as well as frequency spectrum ranges for the classes.
6. Emperical and mathematical investigation into optimizer algorithms. Adam never recovered from local minima and SGD performed only in a narrow learning rate.
7. However, the model needs to build robustness on long temporal sequences e.g. bird songs. Introducing temporal memory is an immediate next step. Initial results with RNN were discouraging both on outcome convergence as well as training duration.
8. The model can increase further resilience by careful synthetic data augmentation. Class imbalance was managed at model level only.

References

A. L. Borker et al., "Vocal activity as a low cost and scalable index of seabird colony size," Conservation biology,vol. 28, no. 4, pp. 1100–1108, 2014.

T. A. Marques et al., "Estimating animal population density using passive acoustics," Biological Reviews,vol. 88, no. 2, pp. 287–309, 2013.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29,no. 6, pp. 82–97, 2012

Hendrik Purwins, Bo Li, Tuomas Virtanen, Jan Schlüter, Shuo-yiin Chang, Tara Sainath Deep Learning for Audio Signal Processing JOURNAL OF SELECTED TOPICS OF SIGNAL PROCESSING, VOL. 13, NO. 2, MAY 2019, PP. 206–219

E. Cakir, E. C. Ozan, and T. Virtanen, "Filterbank Learning for Deep Neural Network Based Polyphonic Sound Event Detection," in IJCNN,2016.

Sharath Adavanne, Emre C¸ akır, Giambattista Parascandolo, Konstantinos Drossos, Tuomas Virtanen Convolutional Recurrent Neural network for bird audio detction, in EUSIPCO, 2017

Thomas Grill, Jan Schluter Two Convolutional Neural Networks for Bird Detection in Audio Signals in EUSIPCO, 2017

Praat: doing phoenetics by computer http://fon.hum.uva.nl/praat/, university of amsterdam

James Lyons http://www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs

Dan stowell https://github.com/danstowell/smacpy

Thomas Grill https://jobim.ofai.at/gitlab/gr/

Seth adams https://github.com/seth814/Audio-Classification

Shaobo Li, Yong Yao, Jie Hu, Guokai Liu, Xuemei Yao and Jianjun Hu Model for Environmental Event Sound Recognition in Applied Sciences July 2018

Dan Stowell Michael D. Wood, Hanna Pamula3, Yannis Stylianou, Hervé Glotin Automatic acoustic detection of birds through deep learning: The first Bird Audio Detection challenge British Ecological society September2018