One-shot learning for acoustic identification of bird species in non-stationary environments

Michelangelo Acconcjaioco and Stavros Ntalampiras

University of Milan, Italy – stavros.ntalampiras@unimi.it - https://sites.google.com/site/stavrosntalampiras/home

Introduction and main problem

➢**Computational bioacoustics** comprises a relatively recent scientific field placed on the crossroad of several disciplines including biology and computer science

➢**Acoustic automatic monitoring** of animals' populations can provide important information, such as a) monitoring of range shifts of animal species due to climate change, b) biodiversity assessment and inventorying of an area, c) estimation of species richness, and d) assessing the status of threatened species

➢The main problem is a **large** and **a-priori unknown** number of species, i.e. **composition** and **size** of **species dictionary *S*** are known only up to a certain extent, meaning that new species can appear at any point in time (unknown).

Novel aspects

The main novel points of this work are:

- ❑ **removes** the need of handcrafted features and domain knowledge,
- ❑ reaches **state of the art accuracy** with a very small amount of training data, and
- ❑ develops a reliable mechanism to **detect** and **react** to **changes** in the environment efficiently.

Algorithm for species identification

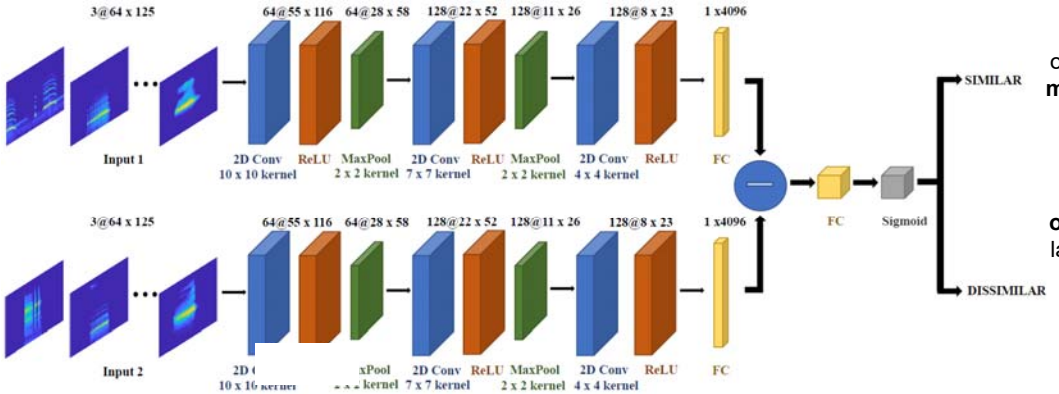
1. Input: test vocalization y^t, trained SNN $\mathcal{N}$, dictionary $\mathcal{S} = \{S_1, \ldots, S_m\}$, while each class is represented by extracted log-Mel spectrograms $\langle \mathcal{F}_S^i \rangle_{i=1}^{i=|S|}$;
2. Extract log-Mel spectrogram $logMel$ of y^t ;
3. Initialize similarity vector $V = []$;
4. **for** j=1:m **do**
   5. **for** i=1:$|S|$ **do**
      6. Query $\mathcal{N}$ with the pair $\{logMel, \mathcal{F}_j^i\}$ and get similarity score $V(j,i)$;
   end
end
7. Predict the class maximizing the similarity score $S^* = \arg\max_{\mathcal{S}} \{V(:,i)\}$;
8. Assign S^* to y^t ;

Algorithm 1: The proposed bird species identification algorithm based on one-shot learning ($|\bullet|$ denotes the cardinality operator).

❑A change in stationarity is signaled when a new log-Mel spectrogram is predicted as **dissimilar with respect to all** sound classes in dictionary *S*.

❑The class **maximizing the similarity score** is the algorithm's prediction.

The proposed Siamese Neural Network

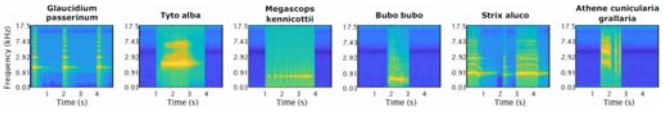


The structure includes N **convolutional** layers, each one followed by a **ReLu** and a **max-pooling** layer, except the last one where max-pool is substituted by a **fully-connected** one. The SNN is completed by a **distance operation**, a **fully-connected** layer and a **sigmoid function** responsible to decide on the inputs' affinity (similar/dissimilar) via thresholding its output.

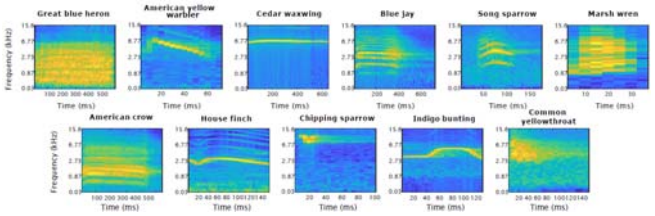
Fig. 1. The pipeline of the proposed one-shot learning scheme using Siamese neural networks.

Feature set and datasets

1) *D1* includes 6 nocturnal bird species, a task which is rather new for the computational bioacoustics community.



2) *D2* represents real-world conditions as it contains field recordings of 11 North American bird species [27].



Activation maps

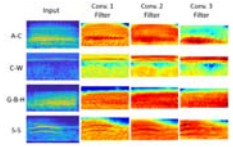


Fig. 7. Convolutional layer outputs to 4 different input spectrograms taken from dataset *D2* (3ConvSNN).

- We observe that each layer **simplifies** the received input and focuses on the most informative region of the spectrograms.
- We can assert that the most **distinctive** feature is the distribution of the **signal's energy in species-depended frequency bands**.

Experiments and Results

k-NN, SVM, 3ConvSNN and 4ConvSNN average recognition rates (in %) on dataset *D2*. The highest rate for each percentage split is emboldened.

Method \ split	10%	30%	50%	60%	70%
k-NN	80.21	87.82	90.56	91.72	92.46
handcrafted+SVM [27]	-	-	-	**96.7**	-
3ConvSNN					
Mean	88.61	92.09	93.96	94.92	96.02
Std	0.85	0.37	0.37	0.14	1.27
4ConvSNN					
Mean	88.12	**92.41**	93.60	94	95.74
Std	0.37	0.42	0.41	0.1	0.38

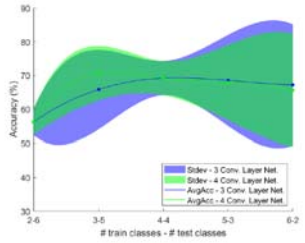


Fig. 4. Average recognition accuracy (%) in non-stationary conditions considering dataset *D1*.

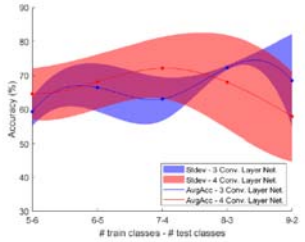


Fig. 5. Average recognition accuracy (%) in non-stationary conditions considering dataset *D2*.

- SNN achieves significant rates outperforming *k*-NN in all splits.
- The proposed method achieved state of art recognition rates
- In non-stationary conditions, the performance exhibited by SNN heavily depends on the **composition of the unknown class set** and their **similarity/dissimilarity with the classes composing the known one**.

Conclusions

✓The proposed solution, based on the one-shot learning paradigm, is able to **detect** changes in stationarity and **incorporate** unknown classes in the dictionary on the fly.

✓Furthermore, it employs a **standardized audio representation** eliminating the need of domain knowledge such as sophisticated features tailored to the problem at hand.