UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE MATEMATICHE,
FISICHE E NATURALI

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

Michelangelo Acconcjaioco and Stavros Ntalampiras
University of Milan
Department of Computer Science

Outline

- Introduction and the main problem
- The proposed Siamese Neural Network
- Feature set and Datasets
- Experimental Set-up and Results
- Conclusions

Introduction and the 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).

Novelties of this work

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.

UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

The proposed SNN

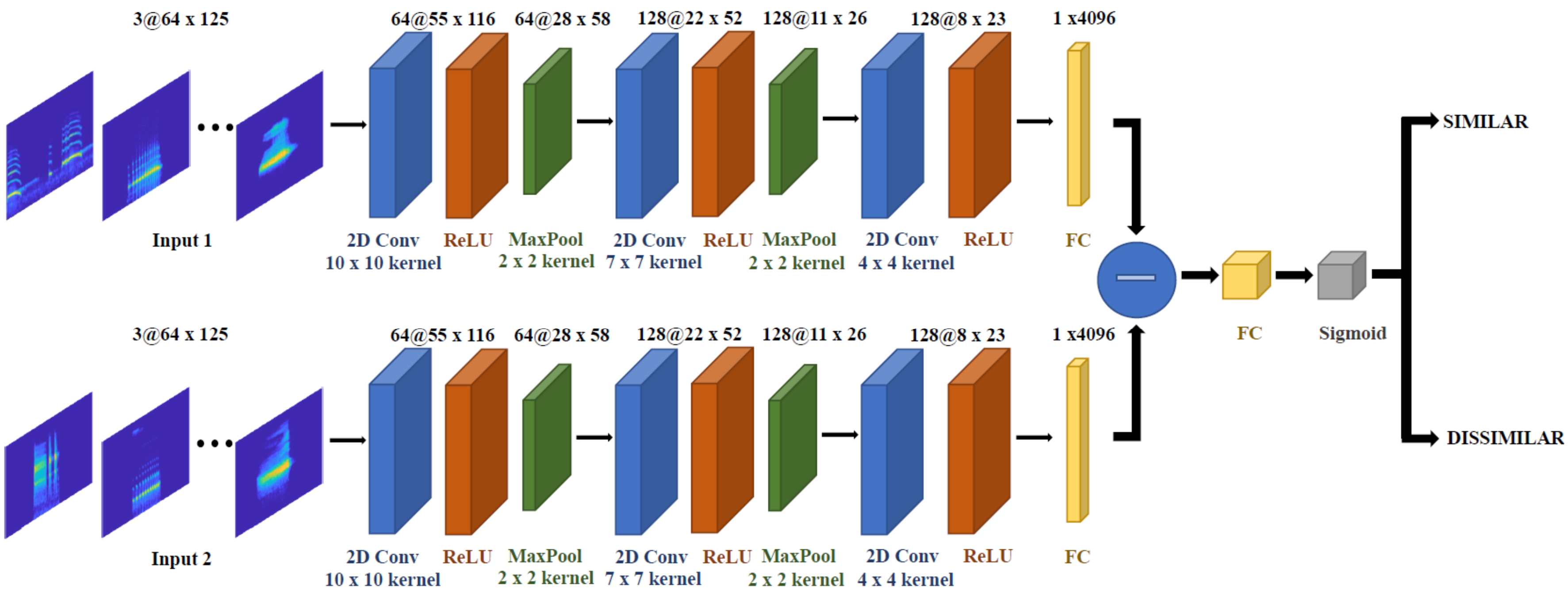


Fig. 1. The pipeline of the proposed one-shot learning scheme using Siamese neural networks.

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.

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 $\langle \mathcal{F}_{\mathcal{S}}^{i} \rangle_{i=1}^{i=|S|}$;
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

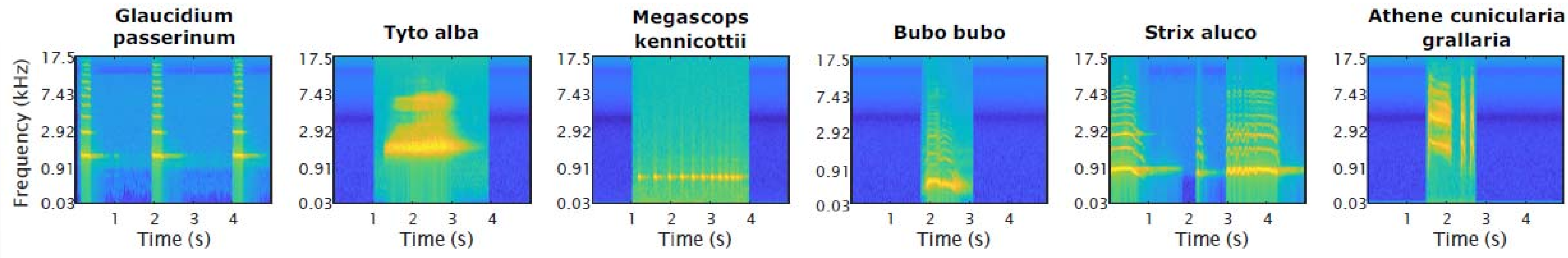
7. Predict the class maximizing the similarity score $S^* = \arg\max_{\mathcal{S}} \{V(:,i)\}$;
8. Assign S^* to y^t ;

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

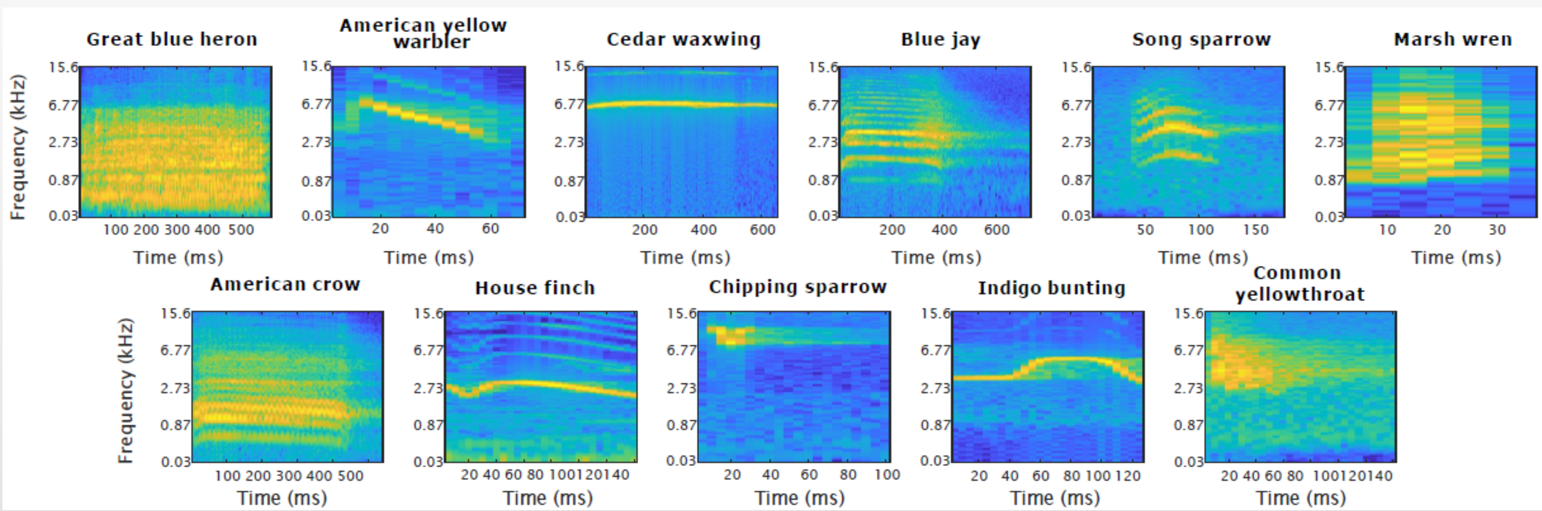
- A change in stationarity is signaled when a new log-Mel spectrogram is predicted as **dissimilar with respect to all** sound classes in dictionary *S*.
- The class **maximizing the similarity score** is the algorithm's prediction.

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]



Experimental set-up and results - stationary conditions

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

- SNN achieves significant rates outperforming k-NN in all percentage splits.
- The proposed method achieved similar rates with the state of art even though SNN is trained only on assessing similarities of logMel spectrograms, unlike [27] which employs *handcrafted* features and SVM trained for *classification*.

Experimental set-up and results - stationary conditions

$\mathcal{M}^s$ (IN %) ACHIEVED BY 3CONVSNN EMPLOYING 60% OF TRAINING DATA OF $D2$

Input 1 \ *Input 2*	*A-C*	*A-Y-W*	*B-J*	*C-W*	*C-S*	*C-Y*	*G-B-H*	*H-F*	*I-B*	*M-W*	*S-S*
A-C	**95.31**	-	-	-	-	-	7.25	-	-	-	1
A-Y-W	-	**95.04**	-	-	-	41.86	-	-	5.88	4.71	-
B-J	-	-	**93.24**	-	1.14	-	1.72	2.53	-	2.11	1.32
C-W	-	-	-	**93.43**	11.65	-	-	-	2.67	-	-
C-S	-	-	1.14	11.65	**88.18**	4.48	-	-	-	4.12	-
C-Y	-	41.86	-	-	4.48	**93.85**	-	-	1.87	16.3	-
G-B-H	7.25	-	1.72	-	-	-	**91.14**	-	-	-	-
H-F	-	-	2.53	-	-	-	-	**97.67**	30.12	-	-
I-B	-	5.88	-	-	-	1.87	-	30.12	**85.75**	-	-
M-W	-	4.71	2.11	2.67	4.12	16.3	-	-	-	**86.22**	5.26
S-S	1	-	1.32	-	-	-	-	-	-	5.26	**98.41**

- We observe that the best recognized species are House finch (97.67%) and Song sparrow (98.41%).
- SNN recognizes better similarities than dissimilarities.

Experimental set-up and results - non-stationary conditions

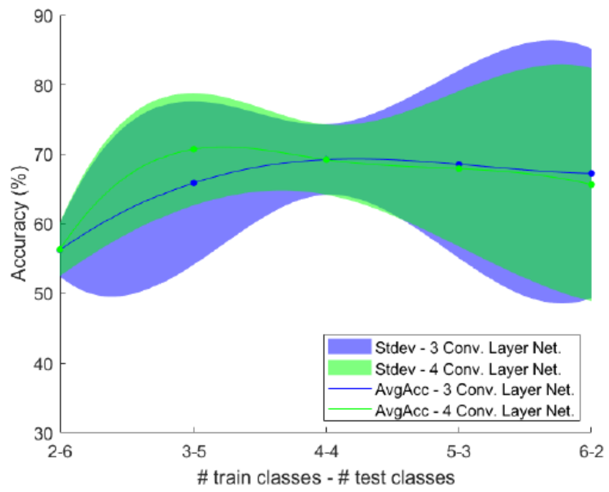


Fig. 4. Average recognition accuracy (%) in non-stationary conditions considering dataset $D1$.

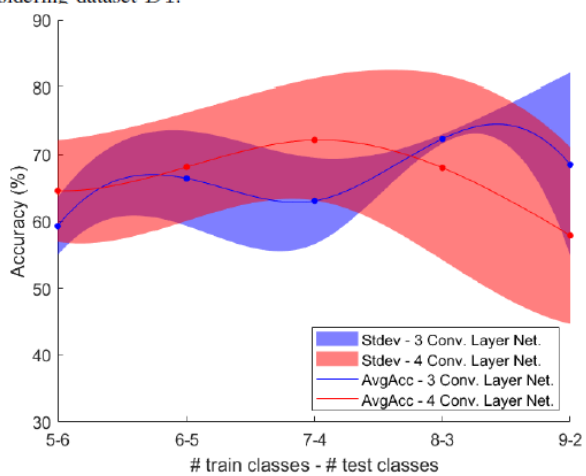


Fig. 5. Average recognition accuracy (%) in non-stationary conditions considering dataset $D2$.

- Class selection for both train and test sets is carried out in random way
- Average recognition rates tend to increase as more classes become available during training.
- In *D1* 4ConvSNN outperforms 3ConvSNN and the opposite for *D2*.
- In *D2* the specific experiment reaches higher rates with lower std's than *D1.*
- 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**.

Experimental set-up and results - activation maps

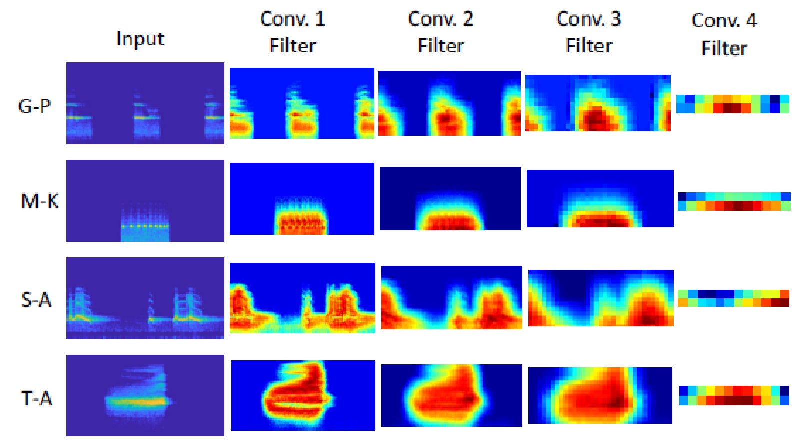


Fig. 6. Convolutional layer outputs to 4 different input spectrograms taken from dataset $D1$ (4ConvSNN).

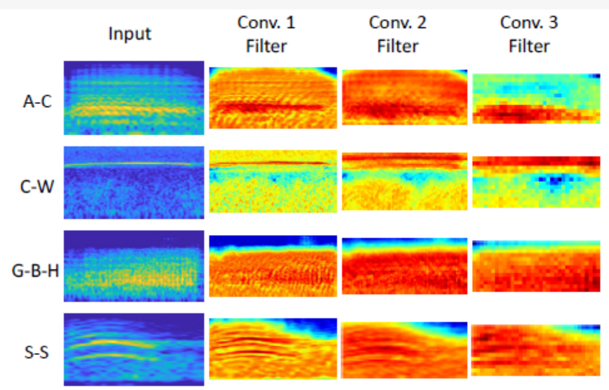


Fig. 7. Convolutional layer outputs to 4 different input spectrograms taken from dataset $D2$ (3ConvSNN).

- Analysis of 4 different samples taken from *D1* and *D2*
- 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**.

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.
- ✓ Experiments carried out on two datasets showed that the method **offers state of the art performance** in stationary conditions and, at the same time, it operates **quite satisfactorily in case of non-stationarities.**
- ✓ Furthermore, it employs a **standardized audio representation** eliminating the need of domain knowledge such as sophisticated features tailored to the problem at hand.
- ✓ We argue that a relevant part contributing to the success of this solution is its ability to **consider both similarities and dissimilarities** to known classes at the same time.

Future work

In the future, we wish to pursue the following directions

a) apply the present method to **other problems of similar constrains** examining the solution from the theoretical point of view,

b) examine the data quantity required by the system to improve the performance in **non-stationary environments**, and

c) after verifying that class selection during training heavily influences the performance, we wish to investigate **strategies enabling optimal selection of the classes composing the training set**.

UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI

The End!

Thank you for your attention!
Questions?

https://sites.google.com/site/stavrosntalampiras/home
stavros.ntalampiras@unimi.it