Deep Learning Locally Trained Wildlife Sensing in Real Acoustic Wetland Environment

Clement Duhart[1], Gershon Dublon[1], Brian Mayton[1], and Joseph Paradiso[1]

Responsive Environment Group
MIT Media Lab
{duhart,gershon,bmayton,joep}@mit.de

Abstract. We describe 'Tidzam', an application of deep learning that leverages a dense, multimodal sensor network installed at a large wetland restoration performed at Tidmarsh, a 600-acre former industrial-scale cranberry farm in Southern Massachusetts. Wildlife acoustic monitoring is a crucial metric during post-restoration evaluation of the processes, as well as a challenge in such a noisy outdoor environment. This article presents the entire Tidzam system, which has been designed in order to identify in real-time the ambient sounds of weather conditions as well as sonic events such as insects, small animals and local bird species from microphones deployed on the site. This experiment provides insight on the usage of deep learning technology in a real deployment. The originality of this work concerns the system's ability to construct its own database from local audio sampling under the supervision of human visitors and bird experts.

Keywords: Wildlife Acoustic Identification, Signal Processing, Deep Learning, Wetland Environment.

1 Introduction

In an era of increasingly ubiquitous sensing, Paradiso et al. [PP16] discuss how it is now possible to document natural ecosystems and record ecological change over longer periods and at significantly higher resolutions than ever before. These new capabilities are of particular interest to restoration scientists and practitioners, who seek to create conditions for complex ecosystems to flourish where human activity previously eradicated them. The restoration context offers researchers the unique opportunity to embed sensor networks directly into new ecosystems as they form, and the resulting data gives the public a chance to learn about ecological functions and environmental impacts. With these dual goals in mind, we developed comprehensive network and sensing infrastructure on a wetland restoration site in southern Massachusetts, called Tidmarsh. Once a 600+ acre industrial cranberry farm, Mass Audubon's recently opened Tidmarsh Wildlife Sanctuary and the Town of Plymouth's Foothill Preserve now host the largest freshwater wetland restoration in New England. We deployed a large number of

custom-designed wireless sensor devices, microphones, and cameras on the site to monitor its transition from industrial farm to protected wetland. The data from our sensors are recorded and streamed in real-time for use in scientific studies, as well as for new immersive experiences for the remote and visiting public. Those experiences include both traditional web applications and augmented reality tools for landscape exploration. A significant challenge we face in this work is in the automated analysis and classification of our data, particularly of the streaming audio and video. To process the audio, we developed a system called *Tidzam* that analyzes large numbers of live streams, recognizes ambient acoustic scenes, and identifies the sources of transient sonic events from an array of wildlife (including dozens of bird species, frogs, and insects), vehicles, and visitors. Recognizing the enormous potential for visitors to submit audio from their mobile devices, our system can also flexibly process temporary streams. In both cases, the resultant classifications are made available to end-user applications in real time. This paper presents the end-to-end Tidzam acoustic wildlife sensing system, its technical underpinnings, and its novel applications to both environmental science and public outreach. A review of the broader vision, as well as details of the sensing and environmental interaction projects, can be found in [MM18]. The locations of the different sensors, microphones and the network infrastructure can be found in Figure 1.

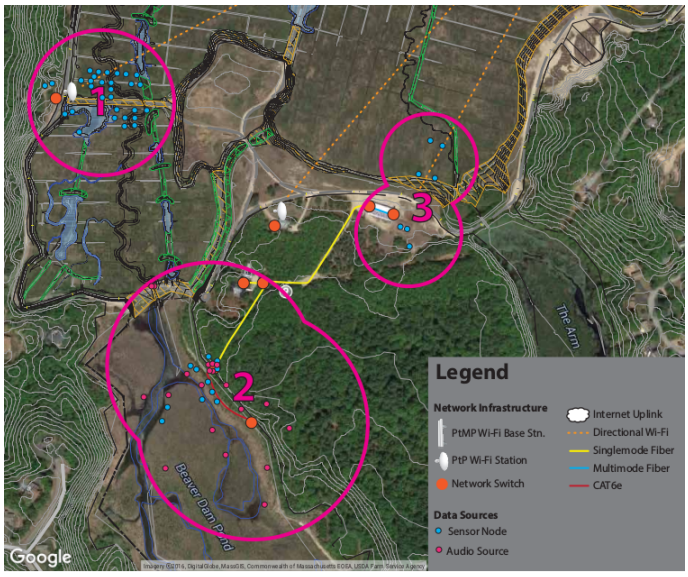

Fig. 1: Locations of the deployed sensors and microphones on Tidmarsh site.

A unique challenge of our acoustic classification task in the restoration context is the appearance and disappearance of numerous different sound sources both seasonally and over the long environmental recovery period. As a result, Tidzam is scalable to new classification tasks to accommodate newly resident

and migrating wildlife; the system is able to detect when a new kind of sound appears and builds new classifiers as needed. A second distinguishing feature of the Tidzam system is our focus on real-time processing for online use with minimal latency. Finally, in addition to identifying bird calls and other wildlife, it has also been designed to record/playback requested samples for expert scientists to review/use.

Automated Recorded Systems (ARS) are crucial tools for wildlife monitoring, estimating bird species abundance and diversity as discussed in Celis et al. [CMDA09]. However, while ARS have evolved from manually triggered to time scheduling systems as presented in Acevedo et al. [AVR06], automated stream segmentation and real-time signal identification remain difficult in noisy and unpredictable outdoor environments. Even in a wildlife sanctuary, human activity appears across the spectrum, with noise caused by cars, aircraft, visitors, and abutters. In real deployment, a wildlife identification system must be coupled with an acoustic scene analysis system.

Acoustic Scene Classification (ASC) is an active research area that has seen significant recent advancement with Deep Learning algorithms. Li et al. [LDM$^+$17] compare such approaches with classical ones, and conclude that temporal specialized models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNN) produce better results than resolution-specialized models like Gaussian Mixture Model (GMayton) and i-vector. A combination of both model types also improves the results by some percent points. A number of researchers are focused on efficient deep learning architectures that are robust to noise or improve diversity of learned features, such as Han et al. [HP17] and Xu et al. [XHW$^+$17]. Those improvements allow expert classifiers for bird calls to be cascaded in realistic outdoor deployments.

Acoustic Bird Identification (ABI) systems in outdoor deployments face two main challenges: first, calls from same and different species frequently overlap in time, and second, many species sound similar to one another in addition to have multiple calls. Kojima et al. [KSH$^+$17] propose a source separation approach which allows the classifier to be processed on independent sources for each possible target bird. This promising approach requires multiple microphones, and we intend to experiment with it in future work. The issue of large output classes is investigated by Hershey et al. [HCE$^+$17], who propose a bottleneck model similar to a de-noising auto-encoder. Their approach forces the system to learn efficient kernels of features in CNNs, and the feature compression may increase diversity and robustness. Cakir et al. [CAP$^+$17,ADCV17] present a Recurrent Convolutional Neural Networks (RCNN) approach, which allows the classifier to learn on both the acoustic texture and its evolution over time, with possible long term dependencies.

Our system has a deep learning expert architecture combining an ASC with an ABI. Section 2 presents the overall architecture as applied to our real-time classification task on 24 live microphone streams and visitor-submitted audio at Tidmarsh, as well as the sample extraction strategy we developed to build our database. Section 3 presents an evaluation of classifier accuracy, which progres-

sively improves with sample extraction during database construction. Section 4 discusses current limitations, diverse applications, and planned future work, leading into our concluding remarks in Section 5.

2 Tidzam

The Tidzam architecture is composed of several processing pipelines presented in Figure 2. A stream manager receives the different input audio streams produced by the fixed microphones onsite (illustrated in Figure 3) or from visitor-submitted mobile audio streams. To reduce bandwidth requirements, the microphone streams are compressed in software running on an embedded server in the marsh; by bundling the channels together into a single Opus-encoded stream, that application also maintains sample-accurate synchronization. The Tidzam server decodes the multichannel stream and sends individual channels to the classifier, which in turn sends its analysis to remote clients and a database server for logging. To grow the sample database, a rules-based engine extracts samples automatically according to classifier confidence. Samples are automatically forwarded to a web interface used by bird identification experts to label unknown samples and further refine the classifier for the next cycle.

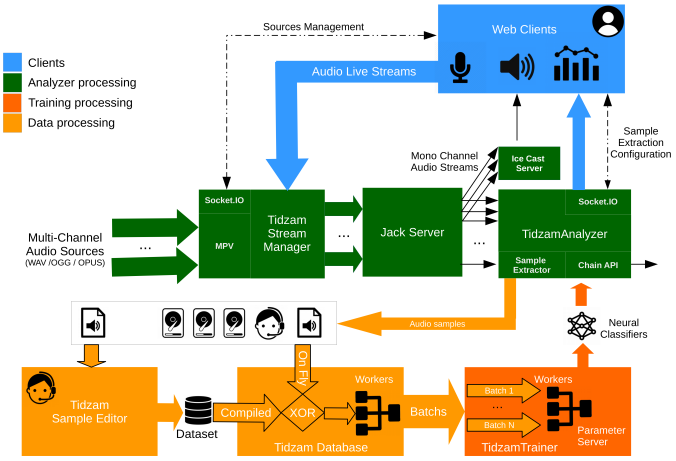

Fig. 2: Overview of Tidzam architecture with the three main pipelines. The green modules represent the classification flows of the different microphone streams loaded by the StreamManager, dispatched by the JACK server to the classifiers in the Analyzer. The results are transmitted to subscribed data consumers in the blue module, which also configure the input streams. The orange modules represent the processing chain for the classifier training, from extraction of a poorly identified sample through the upgrade of the classifier.

2.1 Audio Live Capture Infrastructure

On site A set of 24 microphones, specially designed for harsh environmental conditions, were deployed at Tidmarsh (Figure 3). The audio channels are synchronized and compressed before their transmission to the Tidzam server. Additional audio channels are collected from the streaming wildlife cameras.

Fig. 3: Fixed audio capture at Tidmarsh: 24 microphones are connected to a 32-channel mixer and audio interface, whose output is compressed into a Ogg/Opus stream.

Data Sampling By default, input audio streams are split into segments of 500 ms with a half-overlapping window. A Mel-Filter Banks (MFB) spectrogram is computed on each sample, followed by a median filter for background noise reduction in the presence of brief sonic events. A 50 Hz to 12 KHz bandpass filter is then applied to select the frequency range of the major local bird vocalizations. Finally, samples are normalized for the classifier. Some result examples can be observed in Figure 4. Our choice of a simple preprocessing step over the state of the art reflects a trade-off between the real-time constraints and available compute resources.

2.2 Convolutional Expert Based Classifier

The classifier is a Convolutional Neural Network (CNN) with an expert architecture, as illustrated in Figure 5. The classifier layer is separated into an Acoustic Scene Classification (ASC) and an Acoustic Bird Identification (ABI). The ASC computes a general classification that weights the inhibition of the ABI according

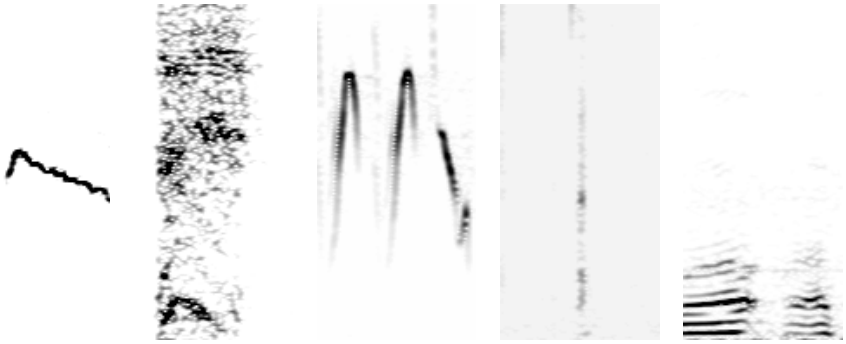

Fig. 4: Illustration of preprocessed sample spectra before their transmission to the classifier: (a) Blue Jay (b) American Crow (c) Fox Sparrow (d) Frog (e) Human Voice.

to its probabilistic estimation of bird presence. The final probabilities are fused by multiplying ABI outputs with the ASC bird class probability. The classifiers share the same stack of convolutional layers responsible for the acoustic feature learning. The cost function is a regular multi-class cross-entropy with parameter regularization as defined in Eq. 1. The ASC was trained on 15 different acoustic scene samples (e.g. rain, crickets, aircraft) as well as brief generic sonic events for inhibition control (e.g. human voice, frog, birds, etc). The ABI has been trained on 50 local bird species.

$$C = -\sum_{i} y'_i \log y_i + (1 - y'_i) \log (1 - y_i) + \gamma \sum_{k} ||w_k||_2^2 \text{ with } \gamma \in]0, 1] \qquad (1)$$

Eq. 1: Cross-entropy cost function used during classifier training with L2 regularization on the weights w. y_i and y'_i are the neural network response probability and the expected value for the class i, respectively.

Note: The design of our expert architecture is a result of early experiments, which demonstrated the difficulty of learning both ambient sounds and punctual sonic events in the same classifier. We observed ambient sounds stimulating a large number of neurons, which can overwhelm the subset of activations produced by smaller patterns of punctual sonic events. In the expert architecture, the separation of the classification task into two independent layers allows the ABI classifier to retain details in the information flow for bird species identification without being overwhelmed by ASC sensitivity.

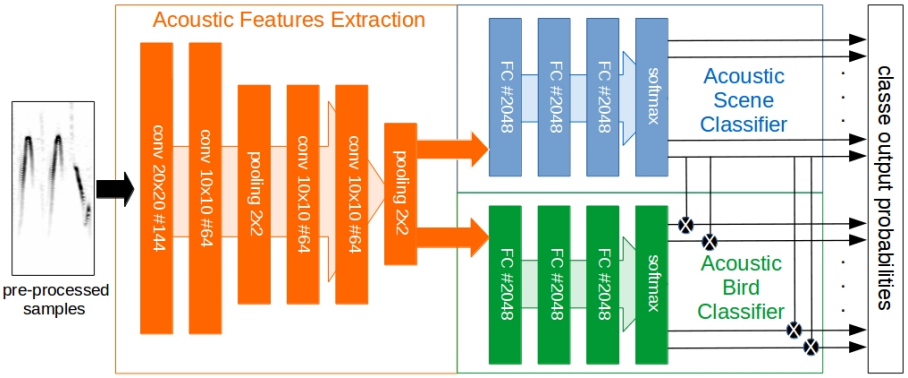

Fig. 5: Neural Expert Architecture

2.3 Semi-Supervised Database Generation

The design of a relevant ASC and ABI database is strongly dependent on the site-specific acoustic scene and the local bird species. Even if a list of bird species is provided by experts, the specific acoustic environment of Tidmarsh must be learned to isolate wildlife sounds from rustling tree leaves, rain, wind, etc. In addition, new species appear over time and throughout the year, requiring constant classifier updates. Semi-supervised database generation from a combination of local ambient sound and off-site samples provides a flexible framework for iterative classifier updates. Based on a confidence rule, the classifier extracts current local samples in order to refine its database. If a sound is labeled as unknown, it is brought to human attention, or can become a candidate for a database augmentation if found to be misclassified as such.

Classifier Confidence Function The confidence function F defined in Eq. 2 evaluates the reliability of the classifier response for the current sample. Based on the comparison between the confidence value and two threshold hyper-parameters $d_u, d_a \in [0; 1]$, the system can decide to extract the sample for a human consideration if lower than d_u (unidentified sample), or for database augmentation (with a preset label) if lower than d_a.

Note: This confidence function tends to extract samples which do not produce enough sparse responses between the output classes. In practice, the confidence thresholds start with low values, which are manually and progressively increased according to the classifier versions. As the diversity of samples in the database increases, the output classifier probabilities tend to become more saturated.

Extraction Rules The extraction rules produce samples for both *human consideration* and *database augmentation*, according to the confidence function and the distribution of samples among the classes as defined in Eq.3. The well-identified samples are used for the next classifier training (after validation by

$$F(y_1, ..., y_n) = b_1 - \sum_{j=2}^{n} b_j \tag{2}$$

Eq. 2: Confidence function evaluating how distant the predicted class output b_1 is from the others, where b_j is the j^{th} largest value of the classifier outputs y_i.

expert operator), whereas unidentified samples are extracted for manual labeling. Unidentified samples are considered as a class, so their extraction probability depends on the number of samples awaiting human consideration. *Note:* At the same time, the system favors samples for classes that are not well represented in the database, regulating the flow of samples for manual labeling.

$$P(s_i) = \lambda_1(1 - \lambda_2^i \frac{|D^i|}{\max_{j \neq i} |D^j|}) \text{ with } s \in D, \lambda_1, \lambda_2 \in [0; 1[\tag{3}$$

Eq. 3: Extraction probability for a sample s of class i according to the database composition D. λ_1 is a flow control parameter used to smooth the number of extracted samples over time, whereas λ_2 controls a probabilistic margin of extraction dependent on class maturity in the database.

2.4 Novel Class Bootstrapping & Training

The classifier is periodically retrained from scratch according to the database augmentation. The training and testing sets, respectively 80% and 20% of the database, are composed of downloaded samples from online databases, mixed with samples extracted on the Tidmarsh site. The bootstrapping of a new class is based on downloaded or manually extracted samples for a rough pretraining and setup of the confidence function. At the first database update, the sample diversity is low, which in turn favors very similar samples by the confidence function. At the same time that the classifier is trained with similar samples (which tends to saturate its output probability), the confidence function threshold is reduced in order to increase the diversity of possible extracted samples. Hence the system extracts samples increasingly diverse from its initial bootstrapped sample references.

Note: After running for one year, Tidzam has generated a database of approximately 300,000 samples, composed of a set of acoustic scenes such as rain noise, water flowing, aircraft, etc, and a set of bird calls from local species. Bootstrapping a new class requires at least 500 samples, which are selected to cover the largest sound diversity for each species' vocalizations or each acoustic scene. The class is then progressively augmented with extracted samples cross-validated with human oversight.

3 Evaluation

The following experimentation has been conducted in order to evaluate whether the self-extraction of samples directly from the acoustic environment can produce effective sample databases. Over one year, a new classifier is trained from scratch each month on a new dataset composed of the previous dataset version augmented with the collected samples of the current month.

3.1 Datasets

During the database construction, new classes were added over one year. For an objective comparison in terms of classifier improvements due to database augmentation, this study has been reduced from 75 to 25 classes (10 acoustic scenes and 15 bird species) that were present in the first version of the database. The samples were added in order of their timestamps of extraction and in order to conserve the balance between the different classes of each training and testing dataset versions.

3.2 Experimental Results

The curves in Figure 6 illustrate the classifier improvements on the testing datasets through the sequence of database updates each two months. The curve V0 is the initial classifier trained on the bootstrap dataset composed of samples from the online Cornell database and some audio recordings from the acoustic scene of Tidmarsh e.g. quiet sound, rain, wind, etc. Each classifier version has been tested on each database version in order to evaluate the improvement resulting from the addition of the newly extracted samples. Parallel trainings are operated in order to get the best classifier depending of the regularization hyper-parameter γ. for a given database version, usually with a success rate greater than 93%. It can be observed that each generation of the classifier outperforms the previous one. Indeed each classifier version gets lowest scores on following augmented databases. Hence the database augmentation by sample extraction improves the inference generalization. The histograms represent the sample extraction frequencies F_u for unidentified samples and F_a for database augmentation. At the beginning, the system tends to extract mainly the unidentified samples, but after several generations of training, this proportion starts to invert. The average frequency of extraction is maintained using the flow control parameter of λ_1. The Figure 7 presents the confusion matrices of the ASC and ABI of the last classifier version on the testing dataset.

Note: The inversion of extraction frequency from unidentified samples to data augmentation results in an improvement of the confidence function, potentially due to an improvement in sound diversity of the database. The system has learned more different kinds of bird calls, but still needs to improve its sensitivity according to the extraction rules in the confidence function.

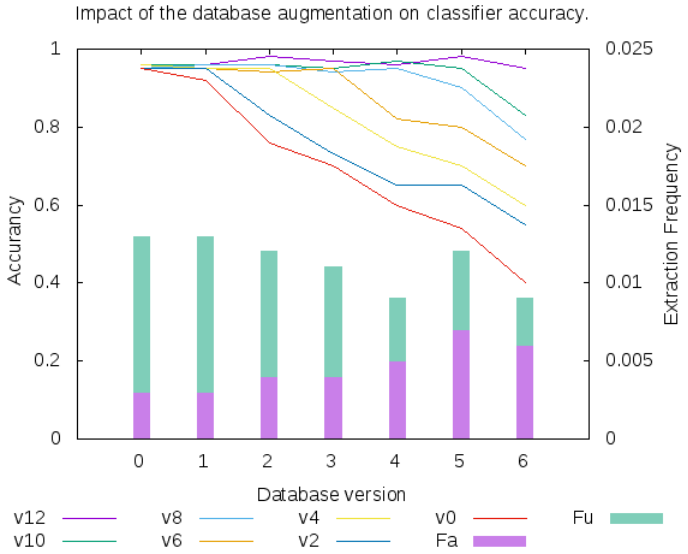

Fig. 6: Evaluation of the impact of extraction strategy on classifier accuracy. Curves represent the accuracy, on the test dataset, of the different upgraded versions of the classifier according to the iterations of database augmentation each two months. Bars represent the extraction frequencies of samples.

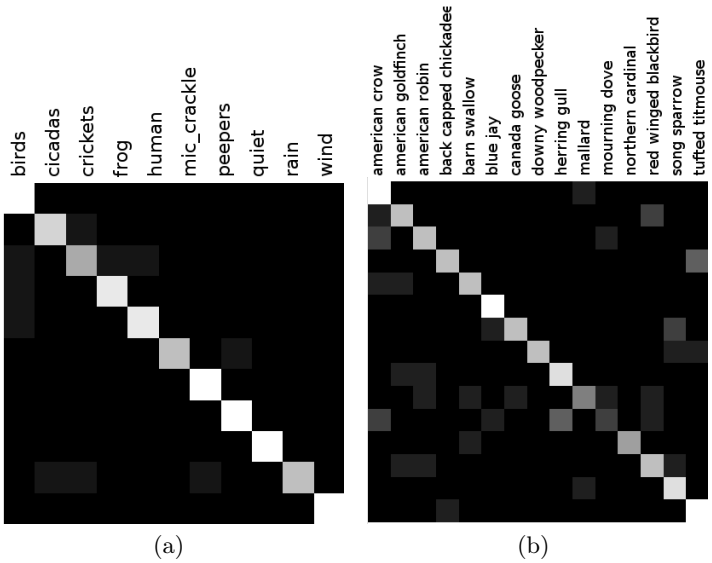

Fig. 7: Confusion matrices of the ASC and ABI classifiers on version V12 of the testing dataset with $\gamma = 0.01$. Main errors occur in the bird species classification.

4 Limitations, Discussions and Applications

As Tidzam is an online learning system which builds its database over time, it is not yet possible to draw firm conclusions about the choices made regarding the neural network architecture for this application. However, we can observe the generic sample extraction strategy constantly improving the database by examining the classifier accuracy over time. This improvement has been achieved thanks to the confidence function, which allows the system to automatically extract samples cross-validated by a human using the Tidzam web interface.

In the interest of expediency, the web application was designed to be as simple as possible, consisting of an automatic sound player, spectrogram visualization, and validation buttons. Given our 24/7 monitoring, validation can be extremely time consuming, with as many as 700 cross-validation samples per day. In the future, this web interface will be replaced by a more engaging, gamified interface, in which both remote experts and Tidmarsh visitors would be able to learn more about the wildlife sanctuary. At the same time, they would be able to correct misclassifications and aid in the labeling of new samples. We see this closed learning loop as beneficial both to amateur enthusiasts interested in learning about bird calls and to experts investigating animal behavior. Crowd-sourcing the validation and labeling process would help Tidzam refine its database.

This approach is in line with the goals of the overall Tidmarsh project, in which immersive technologies are used to augment the visitor experience (see Mayton et al. [MM18]). In one example, called HearThere, a custom-designed smart headphone allows users to hear a dynamic spatial rendering of real-time or recorded sound from all the microphones on the landscape as they walk through it. Output from Tidzam is used to adjust the mix of microphones to promote channels where interesting wildlife has been observed, and suppress channels carrying undesirable sound, such as wind or the voices of other visitors. In another example, called DoppleMarsh, a 3d model of the Tidmarsh terrain is used as the basis of a virtual reality (VR) world driven by the sensing on the physical site. Users can navigate the site remotely in VR, and Tidzam classifications are used to render virtual wildlife and other scenic dimensions. However Tidzam is not able to localize precisely the acoustic sources which would be a very interesting assets in terms of rendering for DoppleMarsh as well as ecological studies. In future works, geo-localization based on source separation on the microphone array will be investigated after an acoustic propagation study on site. This will lead a location update / new deployment of microphones on Tidmarsh.

5 Conclusion

This contribution presents the technique used by the Tidzam project for acoustic scene analysis, wildlife detection, and bird species identification in the outdoor acoustic environment of a wetland. Based on 24 microphones deployed on site, a classifier expert architecture based on deep learning techniques is used to analyze in real-time multiple audio streams. The system is able to automatically

extract the samples in which its confidence is too low so they can be identified by human experts on a web interface. Then the classifier is refined, incorporating the new samples. The preliminary results of this contribution are promising - along 12 training iterations, the system performance has significantly improved. The system has now been building its own database of 300,000 samples over the past year. It is currently used as a wildlife tracker in a large wetland restoration project.

Additional Information: Tidzam is is running on two NVIDIA Titan X GPUs and available online at http://tidzam.media.mit.edu/. All source code can be downloaded from https://github.com/mitmedialab/tidzam

Acknowledgement

The authors would like to acknowledge Living Observatory and the Mass Audubon Tidmarsh Wildlife Sanctuary for the opportunity to realize the audio deployment at this location. The NVIDIA GPU Grant Program has provided the two TITAN X which are used by Tidzam. Clement DUHART has been supported by the PRESTIGE Fellowship of Campus France and the Pôle Léonard de Vinci. We also thank the Elements Collaborative and the sponsors of the MIT Media Lab for their support of this work.

References

[ADCV17] Sharath Adavanne, Konstantinos Drossos, Emre Cakir, and Tuomas Virtanen. Stacked convolutional and recurrent neural networks for bird audio detection. In *25th European Signal Processing Conference (EUSIPCO)*, page 1729–1733, Aug 2017.

[AVR06] Miguel A. Acevedo and Luis J. Villanueva-Rivera. From the field: Using automated digital recording systems as effective tools for the monitoring of birds and amphibians. *Wildlife Society Bulletin*, 34(1):211–214, 2006.

[CAP+17] Emre Cakir, Sharath Adavanne, Giambattista Parascandolo, Konstantinos Drossos, and Tuomas Virtanen. Convolutional recurrent neural networks for bird audio detection. In *25th European Signal Processing Conference, EUSIPCO 2017, Kos, Greece, August 28 - September 2, 2017*, pages 1744–1748, 2017.

[CMDA09] Antonio Celis-Murillo, Jill L. Deppe, and Michael F. Allen. Using soundscape recordings to estimate bird species abundance, richness, and composition. *Journal of Field Ornithology*, 80(1):64–78, March 2009.

[HCE+17] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. Cnn architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 131–135, 2017.

[HP17] Yoonchang Han and Jeongsoo Park. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017.

[KSH+17] Ryosuke Kojima, Osamu Sugiyama, Kotaro Hoshiba, Kazuhiro Nakadai, Reiji Suzuki, and Charles E. Taylor. Bird song scene analysis using a spatial-cue-based probabilistic model. *Journal of Robotics and Mechatronics (JRM)*, 29:236–246, 2017.

[LDM+17] J. Li, W. Dai, F. Metze, S. Qu, and S. Das. A comparison of deep learning methods for environmental sound detection. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, page 126–130, March 2017.

[MM18] MM. Nn. *Presence*, 2018.

[PP16] J. PP. Oo. *TT*, 1(1):47–75, 2016.

[XHW+17] Y. Xu, Q. Huang, W. Wang, P. Foster, S. Sigtia, P. J. B. Jackson, and M. D. Plumbley. Unsupervised feature learning based on deep models for environmental audio tagging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6):1230–1241, June 2017.