Deep Learning and Domain Transfer for Orca Vocalization Detection

Paul Best*, Maxence Ferrari*†, Marion Poupard*‡, Sébastien Paris*, Ricard Marxer*,
Helena Symonds § and Paul Spong §, Hervé Glotin *

*Univ. Toulon, Aix Marseille Univ.
CNRS, LIS, DYNI Marseille, France
†LAMFA, CNRS Amiens France
‡BIOSONG SARL France
§OrcaLab Alert Bay
Email: paul.best@univ-tln.fr

***Abstract*—In this paper, we study the difficulties of domain transfer when training deep learning models, on a specific task that is orca vocalization detection. Deep learning appears to be an answer to many sound recognition tasks in human speech analysis as well as in bioacoustics. This method allows to learn from large amounts of data, and find the best scoring way to discriminate between classes (e.g. orca vocalization and other sounds). However, to learn the perfect data representation and discrimination boundaries, all possible data configurations need to be processed. This causes problems when those configurations are ever changing (e.g. in our experiment, a change in the recording system happened to considerably disturb our previously well performing model). We thus explore approaches to compensate on the difficulties faced with domain transfer, with two convolutionnal neural networks (CNN) architectures, one that works in the time-frequency domain, and one that works directly on the time domain.**

***Index Terms*—Deep Convolutionnal Neural Networks, Orca Vocalizations, End-to-end sound recognition, Spectral sound recognition**

I. Introduction

A. Deep learnging and bioacoustics

In recent years, with the collection of large datasets and the democratization of powerful distributed computation systems, deep learning has proven great performances in modelling complex tasks, such as image or sound recognition. Bioacousticians, that often have to deal with complex sound recognition tasks themselves (e.g from species to acoustic unit classification) are now taking hold of this method. Deep learning has already been successfully applied to tasks such as bird classification [1], or orca vocalization detection [2]–[4]. However bioacousticians that use deep learning often still have to cope with the lack of broad-domain clean datasets. The question of the tranferability of the learned models to different acoustic conditions is still relevant, we will try to approach it in this paper.

For this task of orca vocalization detection using deep learning, certainly the main paper in the literature is Orca-spot [3]. In this paper, several sources of data have been used. Some efforts have been made to dissociate the training data from the test data, by putting excerpts from a single file into a single set. However, samples from every sources of data are found in every set (train, validation, test) in similar proportions. One could argue that the domains in test are similar to the domains encountered during training, and thus that the performance (0.95 of AUC score on the test set) does not reflect the generalization capabilities of the models. In this paper, we study this problematic of domain transfer, in the case of having no access to the foreign domain in advance (even unlabelled). Techniques such as context-adaptive neural networks [5] [6] or unsupervised domain adaptation [7] are thus out of the scope of this paper.

B. The orcas of British Columbia

Orca (*Orcinus orca*, also called the killer whale) is a top-predator of the marine food chain [8]. The Northern Resident Killer Whale community is composed of several "pods" of matrilines [9]. Some of them visit the area surrounding Hanson Island (Canada, north of Vancouver) during summer to feed on the migrating salmons [10]. This odontocete can produce 3 different types of signals : clicks, whistles and pulsed calls [11]. This study focuses only on pulsed calls (referred to as vocalizations or calls, see Fig.1).

II. Material

A. Continuous recording at OrcaLab

Since 1970, the NGO OrcaLab developed an in situ laboratory around Hanson Island. It is equipped with 5 hydrophones, 2 watchtowers and 4 webcam to study the local population of orcas visually and acoustically. The acoustic coverage extends over 50 km^2 (Fig. 2), in a river mouth where communities of orcas come to feed on salmon every summer.

A database called Orchive [13] was built in collaboration with this laboratory, aggregating manually segmented orca calls from 1980 up till today. It is the only large scale open corpus of orca's acoustic emissions, and can be used to train a deep learning model [3]. However, the fact that this database consists only of manually selected calls induces a bias (e.g. researchers often select mostly high sound to noise ratio (SNR) calls).

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

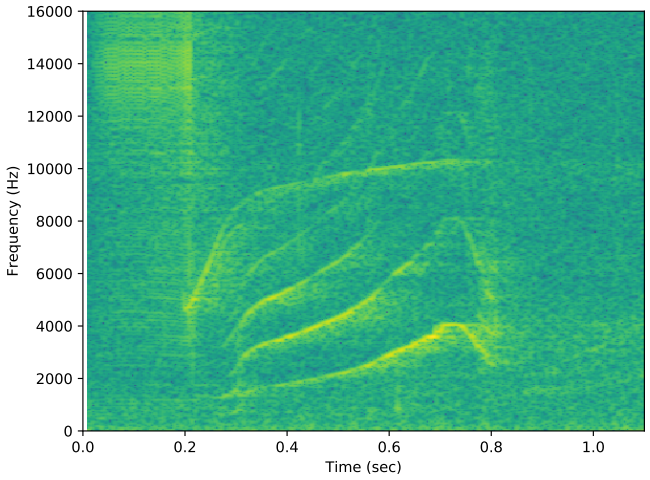

Fig. 1. Spectrogram of an orca pulsed call

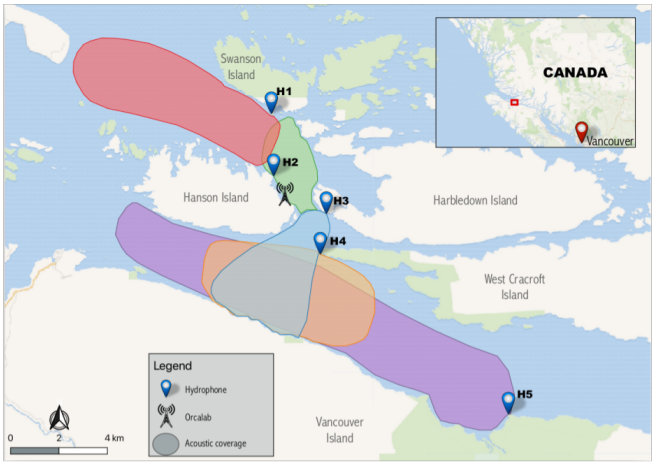

Fig. 2. Map of the area and the listening range of the 5 hydrophones. Map pins with H1 to H5 in bold print denote the hydrophone locations. Detection zones indicate which hydrophones can capture orca calls in a particular area, according to experience of ten years of audio-visual observations of the orcas by the OrcaLab team. Map generated by QGIS software [12] (version 2.14 Essen)

Our ultimate goal is to study the orcas' acoustic emissions with all of their context, such as the environmental conditions (e.g. tides, temperature) as well as the acoustic conditions (nearby ferry boats). For this purpose, we need a model to detect any orca whistle, even the ones potentially dismissed by the monitors.

To help with solving this paradigm, in 2015, we have set up a continuous recording of all the hydrophones of this station. It allows observation and modelling of bioacoustic activities at large temporal scales, in all acoustic and environmental contexts.

The architecture of the global system is shown on Fig.3. The hydrophones record the soundscape continuously and transmit to the OrcaLab station in real time via a very high frequency

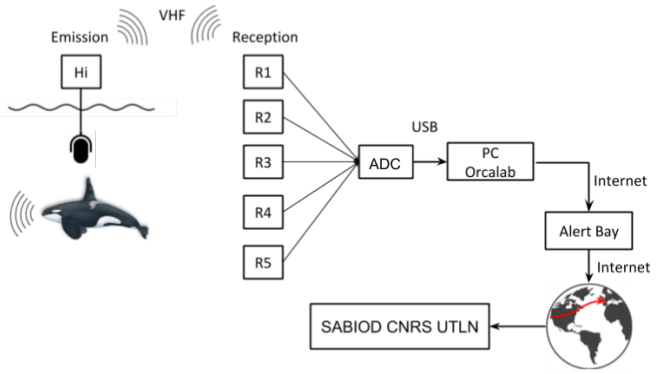

Fig. 3. Data acquisition system, from recording with hydrophones Hi with i from 1 to 5, through the analog to digital converted (ADC), until storage in Toulon (France)

(VHF) radio. The analog signal is received by a radio receiver, digitized at 22050Hz, and sent through internet to be stored in our servers in UTLN Toulon. In total, from July 2015 to 2017, around 50 TB of sound (about 14,500 h) was stored on our server at UTLN.

To deal with this massive amount of data for the analysis of orcas' acoustic behaviour, automatic detection of orca vocalizations is necessary. To build a robust detection model with the amount of available data, deep learning seems to be the best approach.

III. Dataset Construction

A. Orcalab recordings

To train our deep learning model from this continuous recording collected since 2015, we first grossly selected strong acoustic events that could correspond to orca vocalizations, to then manually annotate them. We thus designed an automatic acoustic event extractor (based on [14]). The main steps of the algorithm (shown in Fig. 4) are: (i) calculating the spectrogram (time frequency representation) of the recording using an STFT with a Tukey window of 1024 samples and 20% overlap; (ii) computing a binary image by comparing each pixel against the median over its frequency and time bands: if the energy of a pixel is greater than 3 times the median plus 3 times the standard deviation of its row, and greater than 3 times the median plus 3 times the standard deviation of its column, it is set to 1, otherwise to 0; (iii) applying a "closing" and "dilation" filter for each pixel to remove the noise; (iv) finding connected components and removing small components and isolated pixels; (v) computing bounding boxes for remaining components.

Merging nearby boxes (with a gap of at most 0.2 s), and filtering out irregular ones (boxes with impossible ranges or with a maximum spectral magnitude too high to be orcas) helped us to get rid of a large amount of non-orca acoustic events. Running the detector on 2 days of august 2017 with lots of orca vocalizations output 14k boxes, from which the latter filtering kept 3.5k (1.2 box per minute). Those were annotated

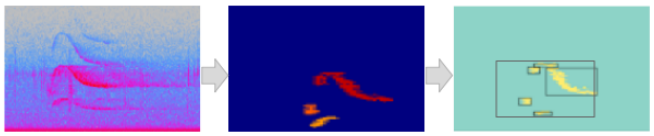

Fig. 4. Main steps for the acoustic events extraction: Binarization and detection of connected components. The spectrogram shows frequencies from 0 to 6.5 kHz during 2.5 s.

TABLE I
SUMMARY OF THE DIFFERENT DATASET

Set	Hydrophones	Sampling rate (Hz)	Depth (m)	Year
Training set	unknown	22050	2-6	2017
Test set OrcaLab	unknown	44100	2-6	2019
Test set JASON	SQ26-01	192000	23	2019

as orca or noise manually, to build a dataset composed of 851 orca vocalization samples (5 seconds long) and 4114 noise samples (boats, rain, void...). Those samples will be referred to as the OrcaLab recordings.

B. JASON antenna recordings

A special recording session took place during summer 2019 at OrcaLab. Equiped with 4 hydrophones and the JASON sound card [15], an antenna was placed next to OrcaLab's main station at 23m of depth. Recording at 192kHz, this setup allows time difference of arrival (TDOA) computation on orca clicks and vocalizations, and thus acoustic source localization, and individual call attribution. The orca vocalizations within these recordings were annotated to then study the potential link between calls, individuals, and behaviour. The calls were thus manually segmented with a good precision (0.1sec approximately). These vocalisations labels are created in version 2.1.2 of Audacity with access to the audio recording and a spectrogram visualization [16].

We used these annotations as an extension to our dataset. Sound files were cut in 5 seconds excerpts, with a positive label if there is an overlap bigger than 1 second with a call annotation, and as a negative label otherwise. Those new annotations hold 411 positive (orca calls) and 682 negative samples (5 second excerpts). The samples extracted from these recordings will be referred to as the JASON antenna recordings.

Eventually, some time-stamped annotations of the JASON recording sessions that occurred in the range of the OrcaLab hydrophones were also used to add samples of OrcaLab recordings (288 samples, 97 positives and 191 negatives).

C. Dataset splitting

It is common in the field to encounter evolution in the recording setups used. Quite often this leads to a mismatch between the audio systems models have been trained on and that on which they are deployed. We use the data collected in our two different configurations (see Table I) as a way to test the effect of this mismatch on our models. The effect of the different recording configurations are shown in Fig.5

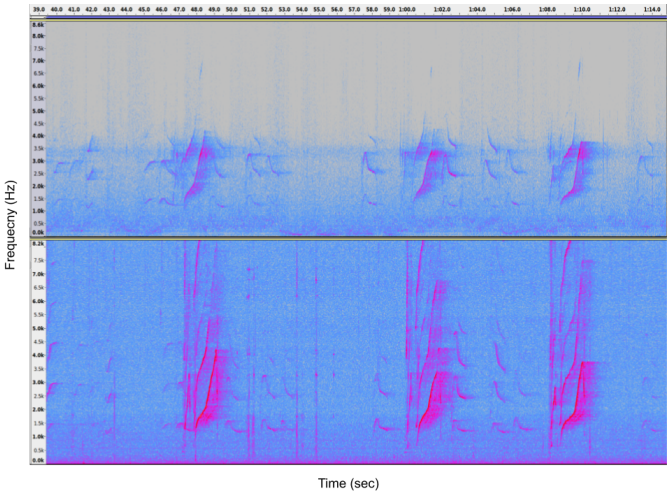

Fig. 5. Spectrograms of the same vocalizations from the two recording configurations (top : OrcaLab continuous recorders, bottom : JASON antenna)

TABLE II
SPLIT OF THE TEST AND TRAINING SETS

Set	Positives	Negatives	Size
Training set	851 (20%)	3424 (80%)	4275 (76%)
OrcaLab test set	97 (34%)	191 (66%)	288 (5%)
JASON test set	411 (38%)	682 (62%)	1093 (19%)
Total	1359 (24%)	4297 (76%)	5656

in the time frequency domain, and in Fig.6 as the average energy received on each frequency bins (as computed for the spectral model, see IV-C1). Both figures reveal the fact that the frequency response of the JASON antenna is more evenly distributed than OrcaLab's continuous recorders. This distribution mismatch supports our assumption that the two recording configuration induce two different "data domains". Another domain mismatch comes from the samples source. The OrcaLab recordings dataset construction III-A relying on the detection of strong acoustic events, it won't include calls below a certain SNR. The JASON recordings samples however, coming from human annotations, include low SNR calls.

In this paper, when we mention the generalization capabilities of our models, we mean their performance on a domain they have not encountered in training. For example, in this experiment, this means having good performances in orca vocalization detection with the JASON antenna recordings when the model was trained only on the OrcaLab recordings. Our train / test split reflects our desire to measure the generalization capabilities of our models, as we put only OrcaLab recordings in the training set. The test set is composed of OrcaLab recordings (from a different year than those in training), and of JASON antenna recordings (resampled to 22050Hz).

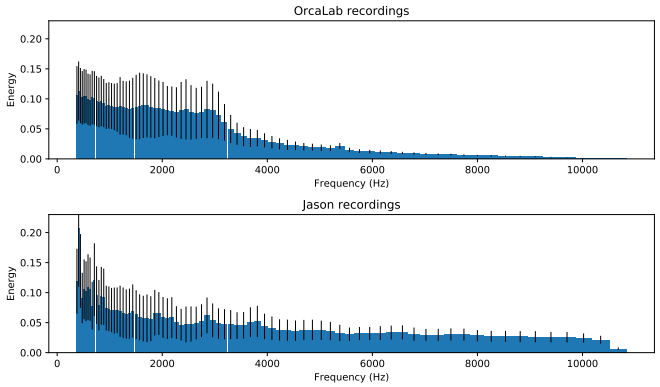

Fig. 6. Average spectrum for the two recording configurations. The x axis denotes the frequency bin in Hz and the y axis the energy level. The blue bars denote the mean energy for each bin, with a black indicator of the standard deviation.

TABLE III
ARCHITECTURE OF THE CONVOLUTIONAL NEURAL NETWORK, FOR THE MEL-SPECTROGRAM OF A 5SEC AUDIO EXCERPT AS INPUT

Input	1x346x80
Conv2D(3x3)	32x345x78
Conv2D(3x3)	32x343x76
MaxPool(3x3)	32x114x25
Conv2D(3x3)	32x112x23
Conv2D(3x3)	32x110x21
Conv2D(3x19)	64x108x3
MaxPool(3x3)	64x36x1
Conv2D(9x1)	256x28x1
Conv2D(1x1)	64x28x1
Conv2D(1x1)	1x28x1
GlobalMax	1

IV. METHOD

A. Data augmentation

Several data augmentation techniques were used, as they helped in generalizing to the test set domain and avoiding overfitting. Temporal transformations such as random offset translation and mirroring were used, as well as the addition of white noise and pink noise. We are aware that mirroring sound samples doesn't presumably make sense as it does for images. Nonetheless, it conserves the harmonic structure of the call and helps us double the diversity of call shapes. In addition, the mixup method [17] was used. These techniques happened to be necessary to obtain a better balance between train set accuracy and test set accuracy (data augmentation typically induced a gain of 10 points of AUC on the validation set).

B. Hyper-parameters

Hyper parameters were empirically set to maximise the generalization performance :

- Batch size : 16
- L2 regularization [18] : 0.002
- Dropout probability [19] : 0.25
- Learning rate : 0.0005 (decrease by of 10% at each epoch)
- Optimizer [20] : Adam
- Number of epochs : 30
- Loss function : Binary Cross Entropy
- Mixup beta distribution [17] : $\alpha = \beta = 0.2$

C. Spectral model

A quite common approach to sound analysis and classification using deep learning is to first compute a mel-spectrogram and to use it as a 2-dimensional image, input for a convolutionnal neural network. We thus experimented this approach by training a binary classifier alike the one that won the EUSIPCO 2017 bird classification challenge [1] (sparrow submission) to distinguish orca calls from other sounds (e.g. void, ferry boat, rain, humpback whales).

1) Spectral Model Input: For each audio file under analysis, we first compute an STFT magnitude spectrogram with a window size of 1024 samples a hop size of 315 samples, and a mel-scaled filter bank of 80 triangular filters from 50 Hz to 10 kHz. The features are normalized per frequency band to zero mean and unit variance. Each audio segment of 5 seconds thus results in an input image of 345 by 80 pixels.

2) Spectral Model Architecture: We kept the same architecture as in [1], shown in Table III. Except for output layer, each convolution and dense layer is followed by the leaky rectifier non-linearity (with a negative slope of 0.01). Each convolution layer is also followed by a batch normalisation [21] and a drop out [19] layer. The total number of network parameters is 309,825.

D. End-to-end model

Here we investigate an approach of treating the first convolution layers' features as new dimensions (similar techniques are used in [22] [23]). The intuition is that this would help the model to build the best suitable representation of the signal for the given task (instead of fixing a representation by computing a spectrogram with selected window size and hop size). We also envision that treating the features output by a layer as a new dimension for the next layers forces some continuity and consistency into them.

1) End-to-end Model Architecture: Except for the output layer, each convolution and dense layer is followed by the leaky rectifier non-linearity (with a negative slope of 0.01). The convolution layers are also followed by a batch normalisation [21] and a drop out [19]. The total number of network parameters is 294,497.

V. RESULTS

A. Models' performances

10 identical training procedures were ran for each model architectures. The results presented in V display mean $\pm$ standard deviation of the 10 scores. To measure the classification performances of the models, their output being a real value,

TABLE IV
END-TO-END MODEL'S ARCHITECTURE FOR A 5 SECOND EXCERPT, WITH DIMENSION AUGMENTATION IN THE FIRST LAYERS

Input	1x1x1x110250
Conv1D(5)	1x1x32x55125
Conv2D(3,5)	1x32x32x27563
MaxPool	1x32x32x13781
Conv3D(3,3,5)	8x16x16x3446
Conv3D(3,3,5)	32x8x8x862
Conv3D(3,3,5)	64x4x4x431
Conv3D(2,2,5)	128x3x3x216
Conv3D(1,1,1)	128x1x1x216
MaxPool	128x1x1x1
Linear	64
Linear	1

TABLE V
METRICS OF PERFORMANCE OF THE MODELS ON EACH SET (MEAN SCORE OF THE 10 RUNS ± THE STANDARD DEVIATION)

Spectral Model			
	Precision	Recall	AUC
Training	0.91 ± 0.017	0.97 ± 0.005	0.99 ± 0.001
Test OrcaLab	0.91 ± 0.105	0.90 ± 0.044	0.98 ± 0.010
Test JASON	0.51 ± 0.04	0.87 ± 0.030	0.74 ± 0.027

End-to-end Model			
	Precision	Recall	AUC
Training	0.63 ± 0.004	0.87 ± 0.002	0.95 ± 0.002
Test OrcaLab	0.50 ± 0.019	0.96 ± 0.005	0.94 ± 0.008
Test JASON	0.63 ± 0.023	0.70 ± 0.032	0.79 ± 0.010

a discrimination threshold needs to be chosen. To emphasis the domain transfer problematic, thresholds were chosen to optimize the score on the training set (threshold for which the true positive rate equals the true negative rate). All metrics except the AUC highly depend on that threshold.
As expected, for both models, the closer the data is to the training domain, the better the performances. For samples that come from the same recording configuration as in training (test set of OrcaLab recordings), the scores show a small but significant decrease of approximately 1 point of AUC. The domain mismatch between the OrcaLab and JASON recordings is confirmed by the important decrease in AUC scores for both architectures. Focusing on the AUC metric, the scores shown in Table V demonstrate that, the spectral model performs better on the known domain, both in training and test sets. However, it scores 5 points less than the end-to-end model for the JASON recordings.

Fig. 7 shows the receiving operator curves (ROC) for each architecture and for each set. These ROC also support the previous observation that the spectral model learns better than the end-to-end model on the training domain, but has a more unstable behaviour on the foreign domain. To try to better understand the models' behaviour facing domain mismatch, we will study their inner representation of the data using a multi layer perceptron (MLP) domain classifier.

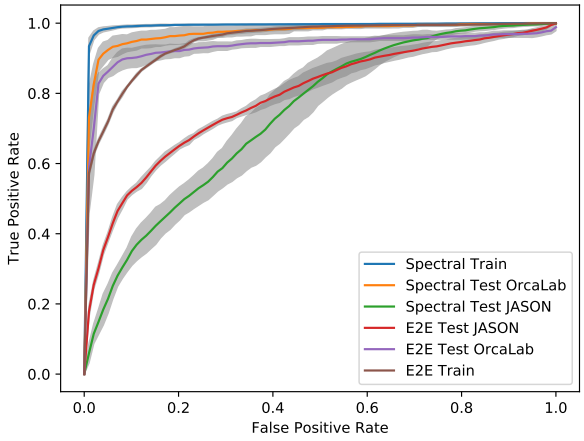

Fig. 7. Receiving operator curves for each set, for the end-to-end (E2E) and spectral architectures. The mean curve of the 10 runs curves are plotted ± their standard deviation

TABLE VI
PERFORMANCE OF THE DOMAIN CLASSIFIER ON THE TEST SET

	Accuracy	Area under curve
Spectral architecture	0.95 ± 0.01	0.96 ± 0.01
End-to-end architecture	0.65 ± 0.07	0.77 ± 0.07

B. Evaluating inner model's representation robustness

The observed scores and ROCs suggest that the end-to-end model builds an inner representation of the input data that is more stable against strong domain change compared to the spectral model. To verify this hypothesis, we studied the capacity of a small binary classifier to infer the domain of the input data from hidden layers of the previous models. The two classes to predict are OrcaLab and JASON recordings, and the classifier takes the form of a 2 linear layers neural network (with 64 and 32 input features) with a non linear activation function in between. For both models, the domain classifier takes as input the 64 features of the before last layer. The whole dataset II was split randomly (with conservation of the domain proportions) into training and test sets (67% and 33% of the data respectively). For each of the 20 models (10 runs per architecture), the domain classifier was trained for one epoch, and then tested on the test set. The means and standard deviations of the scores for the test set are reported in VI. The domain classifier predicts the domain of the input data with much greater accuracy when given the inner representation of the spectral model than with the inner representation of the end-to-end model (despite the latter's great variability). This result supports our previous hypothesis that the end-to-end architecture builds an inner representation that is more robust against strong domain change.

VI. DISCUSSION

The generalization problematic is central to the deep learning field. To study it in this experiment with the available data,

we split our test set into two, one that is relatively close to the training set, and one that is quite far (due to the different recording configurations with different frequency responses). We used common deep learning techniques to help the models generalize better, such as strong data augmentation, small batch size, regularization via weight L2 loss, and dropout. We then implemented two very different architecture, to study their effect on learning and domain transfer. It occurs that the spectral model learns better than the end-to-end model, as long as the input data is relatively close to the training domain. When given a foreign domain, the spectral model loses 24 points of AUC. Future work will study whether per channel energy normalization (PCEN) [24] suffices to tackle this issue. On the other hand, the end-to-end model has lower AUC scores on the training domain, but faces a smaller decrease of performance when given the foreign domain (15 points of AUC loss). The performances of a small domain classifier on the models' last layer support the hypothesis that the end-to-end model's inner representation of the data is more stable against the strong domain change, since the domain of the input data is harder to predict (compared to the spectral model). The main difference between the two architectures being the input layer (one having a spectral representation of the signal, and the other building its own multi-dimensional representation), we suggest that it is responsible for the difference in behaviour.

VII. Conclusion

One of the best ways of studying animals that produce signals in underwater environments is to use passive acoustic monitoring. Automated analysis for captured sound is essential because of the large quantity of data. Labelling data in sufficient amounts to train deep neural networks demands significant efforts. Thus, being able to apply a model trained on a single source to other recording configurations could help the community. In this paper, we explored methods to optimize the transferability of models, applied to the orcas' vocalization detection task. We revealed the benefits and down sides of the choice of model architecture. The results suggests that the end-to-end approach builds an inner representation that is more stable against strong domain change, whereas the spectral model discriminates better on its training domain.

Acknowledgments

We thank first the OrcaLab direction Paul Spong and Helena Symonds and collaborators for their incredible inspired work. We thank MARITTIMO EUR GIAS projects, Région SUD Provence Alpes Côte d'Azur, and FUI 22 Abysound for P. Best PhD funding. We thank l'Agence de l'innovation de défense and Région Hauts-de-France for M. Ferrari's PhD grant. We thank Biosong SAS France for the PhD funding of M. Poupard. We thank ANR for connected granted projects : ANR-18-CE40-0014 SMILES, ANR-17-MRS5-0023 NanoSpike, and Chaire Artificial Intelligence in Bioacoustics ADSIL (PI Glotin, 2020-2024). We thank Institut Universitaire de France for H. Glotin's chair 2011-16 during which he designed and installed the long term full recording of OrcaLab to UTLN DYNI.

References

[1] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1764–1768.

[2] Jia jia Jiang, Ling ran Bu, Fa jie Duan, Xian quan Wang, Wei Liu, Zhong bo Sun, and Chun yue Li, "Whistle detection and classification for whales based on convolutional neural networks," *Applied Acoustics*, vol. 150, pp. 169 – 178, 2019.

[3] Christian Bergler, Hendrik Schröter, Rachael Xi Cheng, Volker Barth, Michael Weber, Elmar Nöth, Heribert Hofer, and Andreas Maier, "Orca-spot: An automatic killer whale sound detection toolkit using deep learning," *Scientific Reports*, vol. 9, no. 1, pp. 10997, 2019.

[4] M. Poupard, P. Best, J. Schlüter, JM. Prevot, P. Spong, H. Symonds, and H. Glotin, "Deep learning for ethoacoustics of orcas on three years pentaphonic continuous recording at orcalab revealing tide, moon and diel effects," in *IEEE OCEANS*, 2019.

[5] Vincent Lostanlen, Justin Salamon, Andrew Farnsworth, Steve Kelling, and Juan Pablo Bello, "Robust sound event detection in bioacoustic sensor networks," *PloS one*, vol. 14, no. 10, 2019.

[6] Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, Takuya Yoshioka, Dung T Tran, and Tomohiro Nakatani, "Context adaptive neural network for rapid adaptation of deep cnn based acoustic models.," in *INTERSPEECH*, 2016, pp. 1573–1577.

[7] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," *arXiv preprint arXiv:1409.7495*, 2014.

[8] TA. Jefferson, PJ. Stacey, and RW. Baird, "A review of killer whale interactions with other marine mammals: Predation to co-existence," *Mammal review*, vol. 21, no. 4, pp. 151–180, 1991.

[9] MA. Bigg, PF. Olesiuk, GM. Ellis, JKB. Ford, and KC. Balcomb, "Social organization and genealogy of resident killer whales (orcinus orca) in the coastal waters of british columbia and washington state," *Report of the International Whaling Commission*, vol. 12, pp. 383–405, 1990.

[10] MJ. Ford, J. Hempelmann, MB. Hanson, KL. Ayres, RW. Baird, CK. Emmons, JI. Lundin, GS. Schorr, SK. Wasser, and LK. Park, "Estimation of a killer whale (orcinus orca) population's diet using sequencing analysis of dna from feces," *Plos one*, vol. 11, no. 1, pp. e0144956, 2016.

[11] JKB. Ford, "Acoustic behaviour of resident killer whales (orcinus orca) off vancouver island, british columbia," *Canadian Journal of Zoology*, vol. 67, no. 3, pp. 727–745, 1989.

[12] QGIS Development Team, *QGIS Geographic Information System*, Open Source Geospatial Foundation, 2009.

[13] Steven Ness, *The Orchive: A system for semi-automatic annotation and analysis of a large collection of bioacoustic recordings*, Ph.D. thesis, University of Victoria, Canada, 2013.

[14] M. Lasseck, "Large-scale identification of birds in audio recordings.," in *CLEF (Working Notes)*, 2014, pp. 643–653.

[15] Manon Fourniol, Valentin Gies, Valentin Barchasz, Edith Kussener, Hervé Barthelemy, Remy Vauche, and Hervé Glotin, "Low-power wake-up system based on frequency analysis for environmental internet of things," in *2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA)*. IEEE, 2018, pp. 1–6.

[16] Team Audacity, "Audacity (2018) audacity(r): free audio editor and recorder (computer application), version 2.2.2," https://audacityteam.org/, 2018.

[17] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 13888–13899.

[18] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh, "L2 regularization for learning kernels," *arXiv preprint arXiv:1205.2653*, 2012.

[19] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[20] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[21] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015.

[22] Jonathan J Huang and Juan Jose Alvarado Leanos, "Aclnet: efficient end-to-end audio classification cnn," *arXiv preprint arXiv:1811.06669*, 2018.

[23] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[24] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard F Lyon, and Rif A Saurous, "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5670–5674.