Use of deep features for the automatic classification of fish sounds

Marielle Malfante, Omar Mohammed, Cedric Gervaise, Mauro Dalla Mura, Jerome I. Mars

Use of deep features for the automatic classification of fish sounds.

Marielle MALFANTE[1],
Omar MOHAMMED[1], Cédric GERVAISE[2], Mauro DALLA MURA[1], Jérôme I. MARS[1].
[1] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab*, 38000 Grenoble, France
[2] Chorus, Fondation Grenoble INP, 38000 Grenoble, France

***Abstract*— The work presented in this paper focuses on the environmental monitoring of underwater areas using acoustic signals. In particular, we propose to compare the effectiveness of various feature sets used to represent the underwater acoustic data for the automatic processing of fish sounds We focus on the detection and classification tasks. Specifically, we compare the use of features issued from signal processing presented and validated in [15], [16] to the use of features obtained through deep convolutional neural networks. Experimental results show that the use of signal processing features outperform the deep features in terms of classification accuracy.**

I. INTRODUCTION

The vitality assessment of environmental areas is of great interest to the scientific community [11], [13]. The monitoring of sensitive marine zones (which are often difficult to access) is of tremendous importance. Sensors can be deployed and record continuous acoustic signals which once processed, help gather information about the area of interest. The potentially huge amount of data acquired has triggered the development of autonomous analysis procedures for example based on machine learning approaches [10], [13]. In this study, we focus on tools dedicated to the automatic detection and classification of such data, focusing on in situ recordings of many fish sounds acquired on the Mediterranean coasts.

In previous a work [15], we showed the effectiveness of a machine learning based approach to continuously process underwater recordings for the automatic classification of fish sounds. The effectiveness of the method in processing continuous recordings is of particular interest with respect to the monitoring of underwater areas. In this study, we propose to compare the use of different sets of descriptors for extracting features of the recorded signals. Those feature are then input to the learning algorithm to train a classification model. In particular, we compare the discriminative power (in terms of classification accuracy) of the feature set we proposed in [16] (based on descriptors computed on the signals in the time, frequency and cepstral domains) with features based on deep learning.

In the latest years, deep learning methods have been of an increasing interest given the remarkable results achieved in image classification [12], [19] or speech processing [8]. Given the success of those methods, they are now being applied for processing other types of data, including underwater signals [10].

*Institute of Engineering Univ. Grenoble Alpes

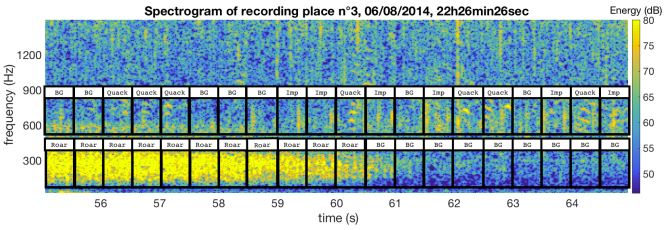

Fig. 1. Spectrogram of underwater recordings and illustration of the labeling process. Recording place #3: posidonia/ sand boarder, depth 38m.

Convolutional Neural Networks (CNNs), a classification technique based on the deep learning paradigm, have proven their ability to learn features from a dataset of images and capture their properties [12], [19]. Similarly, they are now used with underwater signals [2], [5]. In this applicative scenario, spectrograms are computed from the recorded signals and then processed with a CNN as they are images. However, the major limitation of CNN - and of deep neural network in general - is the amount of labeled data needed for training. Such large labeled datasets are very difficult to gather given the amount of labor and cost needed to perform the labeling task. Nevertheless, one might benefit from the CNN ability to represent data in the case of a limited labelld dataset by considering features computed by a CNN trained on another dataset as input to a standard classifier. This procedure is within the scope of transfer learning, and has proven its effectiveness when training a full CNN is not feasible due to limited number of training observations [17].

II. DATASET

The data used in this study were recorded in coastal areas of the Mediterranean sea, in France (Corsica) [14]. Several days of continuous recordings sampled at 156kHz are available. Ten minutes of recordings containing several fish sounds were labeled into five classes displayed in Figure 1: four fish sounds and one background class. A non-overlapping sliding window of size $\Delta_t = 0.5s$ is used and two bandwidths (50-450Hz and 400-900Hz) are considered, leading to 913 labeled observations.

III. METHODS

The automatic classification method follows the architecture presented in [15], [16] and uses machine learning algorithms to build a prediction model from the learning data. The key point of the architecture lies in the descriptors

leading to the representation of the acoustic signals in feature space. We showed in [15], [16] the effectiveness in terms of class discrimination when using a feature set composed of 84 features for the characterization of the acoustic signals shape and properties (more detail on the feature set will be given in the full paper). In particular, general shape descriptor features are extracted from various representation of the observations: temporal, spectral, and cepstral domains. Cepstral domain is usually used in speech processing, to describe the harmonic properties of a signal by computing the Fourier transform on the signal spectrum. This approach lead to excellent classification results and has since then been tested on other datasets [16].

The aim of this study is to compare the discrimination performance when using features based on deep learning and referred as deep features. In particular, four pre-trained deep CNNs architectures are considered: `VGG16` [19], `ResNet50` [7], `MobileNet` [9] and `InceptionV3` [20]. The top dense layer of each CNN is removed (since it is the classification layer and is adjusted only for data similar to the training dataset) and the output features can be directly used into a classic machine learning algorithm. Each CNN was train on ImageNet dataset [18], and the top layer was removed. Those networks are among the most effective ones that are already trained, and are publicly available.

Some studies have also reported that features extracted from a CNN are more and more complex with the network depth [21], [6]. It can therefore be interesting to extract features at different layers of a CNN and to compare the associated results. This issue will be investigated with the best performing CNN feature set.

IV. Results & Discussion

In the following section, we describe the obtained results and compare them to the use of signal processing features. Spectrograms are computed using Kaiser windows of size $n = 1024$, with on overlap of 90%. The fast Fourier transform is computed on $N = 1.5 * n$ points and a decibel scaling is used on the spectrograms is used before generating the images. Keras implementation of the networks were used to extract features, and all tests were run using on a M2000 NIVIDIA Quadro GPU card with 768 cores and 4 GB of memory. Cross validation process was use to estimate the models performances, with 70% of the data in training and 30% in testing. Ten trials were performed, and mean and standard deviation of the overall accuracy are considered. Both Random Forest (RF) [4] and Support Vector Machine (SVM) [3] learning algorithms are used and compared. Hyper-parameters of both algorithms are set to optimized the cross validation results. SVM was considered with a linear kernel for the deep features and rbf kernel with the signal processing features. RF was computed with 100 decision trees that were not pruned in both configurations.

A. Comparison between the various feature sets

Cross-validation results for SVM (linear kernel) and RF are summarized in table I. Several remarks can be made on those results. First, among the four different CNNs, `ResNet50` systematically performs better than the others, with $89.5 \pm 1.15\%$ for RF and $93.9 \pm 0.97\%$ for SVM. The lowest results are obtained with `VGG16`, with $75.0 \pm 1.99\%$ for RF and $86.8 \pm 1.40\%$ for SVM. Secondly, accuracy is repeatedly lower with deep features than with the signal processing features ($96.9 \pm 2.0\%$ for RF and $96.5 \pm 1.6\%$ for SVM). However, it is worth noticing that the dimension of deep features is very high compared to the 84 proposed features. At this point, it is not possible to say if the lower results obtained with deep features are related to a lesser informative content or to their high dimension. The curse of dimensionality [1] could influence the results. It therefore rises the issue of the feature vectors dimension and their impact on the accuracy. Thirdly, results are systematically higher with SVM (with linear kernel) than RF. The accuracy gain when using SVM can vary from 4.4% with `ResNet50` to 11.9% with `VGG16`. It is worth noticing that the network leading to the smaller accuracy difference (`ResNet50`) is also the one with the smaller output dimension. This third point also rises the question of the feature vectors dimension.

TABLE I
Comparison between the use of Signal Processing features and features computed from various CNNs. Results are proposed using SVM and RF algorithms and cross-validation process ($\alpha = 0.7$ on the learning set).

Feature set	**Dim**	**X-validation**	
		RF	SVM
`Signal Processing`	84	$96.9 \pm 2.0\%$	$96.5 \pm 1.6\%$
`InceptionV3`	131072	$80.6 \pm 0.93\%$	$88.4 \pm 1.72\%$
`MobileNet`	50176	$81.9 \pm 1.04\%$	$91.5 \pm 1.37\%$
`ResNet50`	2048	$89.5 \pm 1.15\%$	$93.9 \pm 0.97\%$
`VGG16`	25088	$75.0 \pm 1.99\%$	$86.8 \pm 1.40\%$

B. Influence of the feature vectors dimension

We here study the impact of the feature vectors dimension on the accuracy results. Two main questions are raised. Compared to the other three networks, is `ResNet50` leading to the best results because of its shorter dimension? and Is SVM performing better than RF because of the high dimension, or is SVM more adapted for this dataset? To answer both questions, Principal Component Analysis (PCA) is use to compress the feature vectors output by the various CNNs. Five to 500 components are kept as the new feature vectors, and the comparison between the four CNNs is run again. Results are presented in Figure 2. The top graph presents RF results, and the lower one SVM results. Accuracy values depending on the feature vectors dimensions (CNNs and PCA) are represented (color dots), and accuracy values from the previous part (CNNs without PCA) are displayed for reference (colored triangles).

First, `ResNet50` still performs well with PCA (red dots), but `MobileNet` tends to have better results (green dots). The difference is generally small, but is more pronounced

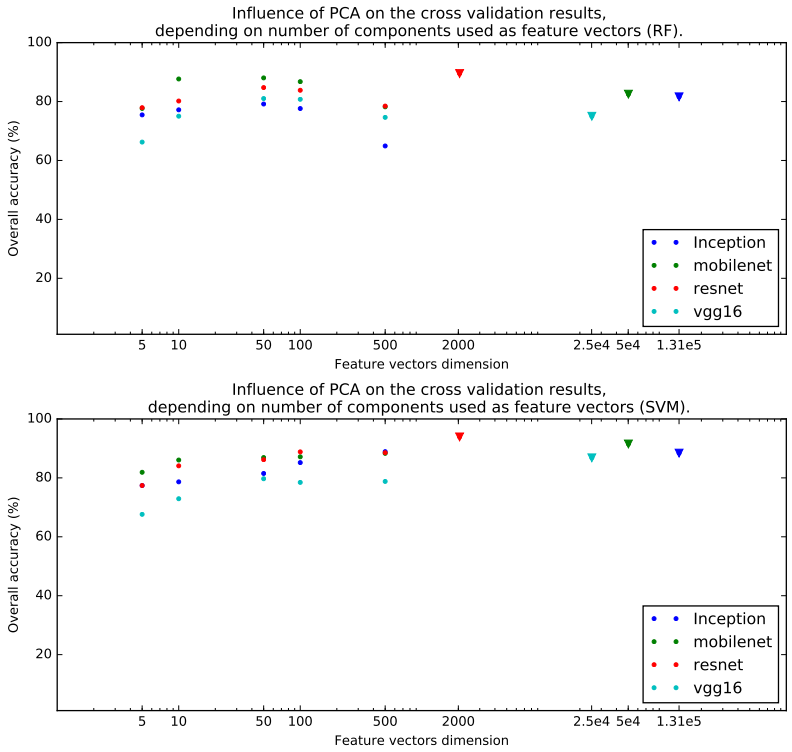

Fig. 2. Use of PCA on the deep feature vectors, with different components kept as input of the learning algorithm. RF is consider on the top graph, SVM with linear kernel on the lower graph. Triangles display results when using the full deep feature vectors (no PCA).

when using RF compared to SVM. When using PCA, `ResNet50`, `Inception` and `MobileNet` perform better with SVM while results with `VGG16` are higher with RF.

Results are generally speaking quite different from RF to SVM. The general trend for RF accuracy evolution would be to increase with the PCA dimension, with a maximum around 50 dimensions, and then to decrease. This observation could have been related to the curse of dimensionality, but the comparison with results from the previous part rules out this hypothesis. Results without PCA are equivalents (`VGG16`) or superiors (`ResNet50`, `MobileNet` and `InceptionV3`) without PCA than with PCA and 500 components. There is no clear tendency between accuracy without PCA and PCA with 50 dimensions (best results when considering RF and PCA): `ResNet50` and `Inception` perform better without PCA, `VGG16` has similar results in both configuration, while `MobileNet` performs notably better with the use of PCA. There is no clear interpretation on RF and the input data dimension, but the use of PCA with RF would not necessarily be recommended. If used, the choice of the PCA dimensions to keep should be considered as a hyper-parameter of the problem.

This interpretation with SVM as learning algorithm is quite different, since the general trend is toward better results with higher dimensions, whether PCA is used or not. This observation could be explained by the use of a linear kernel: the more separable the input data are, the better for classification results. With this configuration, the use of PCA which compacts the data informative content would not be recommended.

C. Layer selection for deep features extraction

CNN features get more and more complex with the layer depth. In those conditions, it might be interesting to use features from the bottom layers (simple shape features) rather than from the top layers (complex shapes and features). This is the object of this experiment in which we compare features extracted at different layers from `ResNet50`. In particular, `ResNet50` architecture is made of four blocks of layers, and features extracted from each of those four blocks are considered and will be referred as `block1` to `block4` feature sets. Feature set `block1` being the simplest and `block4` the most complex. Results comparing those four feature vector sets are reported in Figure 3. The figure displays accuracy levels for each of the four feature sets and put them in relation with the feature vectors dimensions. `block1` and `block4` features vectors have similar dimensions, $193,600$ and $200,704$ respectively. They lead to similar accuracy levels when used with SVM ($93.8 \pm 1.17\%$ and $94.2 \pm 1.14\%$ respectively) but `block1` leads to better results with RF ($90.5 \pm 1.18\%$ and $83.1 \pm 2.07\%$ respectively). The second and third feature sets lead to lower accuracy levels, but also have larger dimensions, 774,400 and 401,408 respectively. To compare the effect of the different feature vectors regardless of their dimensions, the same experiment is conducted with the use PCA. `block1` features perform better than the other feature sets regardless of the learning algorithm and the

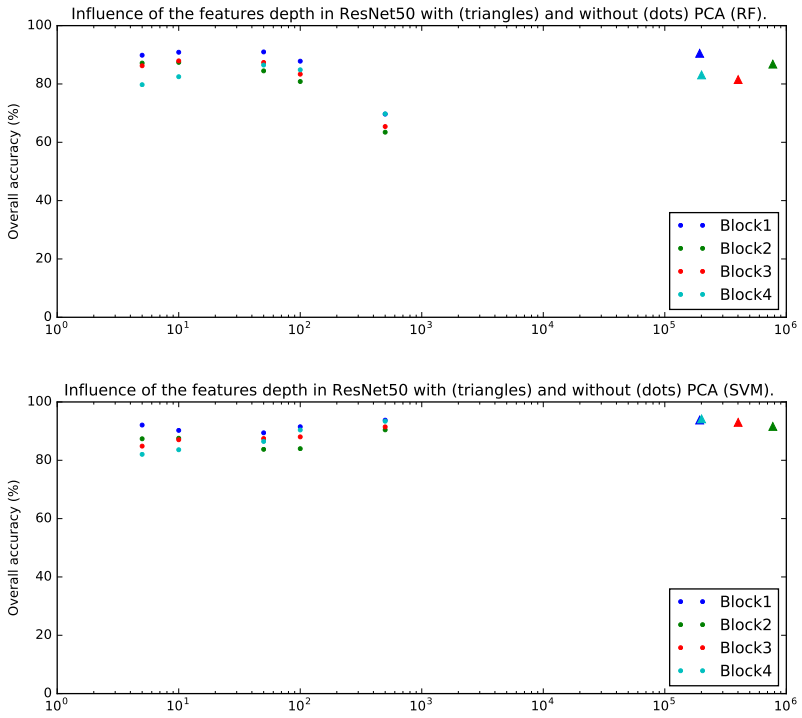

Fig. 3. Layer selection performed on ResNet50. Dimension reduction of the deep features vectors is also considered with the use of PCA (50 components). Results are proposed using SVM and RF algorithms and using cross-validation process.

number of components extracted from the PCA. Features from the bottom layers of `ResNet50` therefore seem to be interesting for this application, even if in this configuration, best results are obtained with `block4` features and the use of SVM. In this configuration, accuracy is almost as high as it is when using the proposed and handcrafted features.

V. Conclusion

The effectiveness of deep learning methods have originally been demonstrated in image and speech processing. Nowadays, these tools are considered for several other applications. In this work, we investigate the use of features issued from deep learning for the automatic classification of fish sounds, comparing them to more conventional features issued from signal processing. Results are in favor of signal processing features with overall accuracy reaching $96.9 \pm 2.0\%$ when using RF (similar results with SVM). However some configuration in which features are extracted from CNNs can reach close accuracy levels. Such results validate and encourage the use of machine learning methods for the process of continuous underwater recordings and the study of underwater areas.

ACKNOWLEDGMENT

This study was supported by a grant from Labex OSUG@2020 (Investissements d'avenir - ANR10 LABX56) and DGA/MRIS Geosciences. GIPSA-Lab SIGMAPHY is part of Labex OSUG@2020 (ANR10 LABX56).

References

[1] R. Bellman. Dynamic Programming and Lagrange Multipliers. *Proceedings of the National Academy of Sciences of the United States of America*, 42(10):767–769, 1956.

[2] C. Bentes, D. Velotto, and B. Tings. Ship classification in terrasar-x images with convolutional neural networks. *IEEE Journal of Oceanic Engineering*, 2017.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the fifth Annual Workshop on Computational Learning*, pages 144–152, 1992.

[4] L. Breiman. Random forests. *Machine learning*, pages 5–32, 2001.

[5] K. Denos, M. Ravaut, A. Fagette, and H.-S. Lim. Deep learning applied to underwater mine warfare. In *OCEANS 2017-Aberdeen*, pages 1–7. IEEE, 2017.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] 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*, 29(6):82–97, 2012.

[9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861*, 2017.

[10] IEEE. *NeXOS, developing and evaluating a new generation of in-situ ocean observation systems*, 2017.

[11] P. Johnston and M. Poole. Marine surveillance capabilities of the autonaut wave-propelled unmanned surface vessel (usv). In *OCEANS 2017-Aberdeen*, pages 1–46. IEEE, 2017.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C.

Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] H. Li, K. Yang, H. Xia, and Q. Yang. Model-independent depth classification of transient acoustic signal in deep water. In *OCEANS 2017-Aberdeen*, pages 1–4. IEEE, 2017.

[14] J. Lossent, C. Gervaise, and L. D. Iorio. Cartographie de la biophonie des écosystèmes côtiers. pages 1–5, 2015.

[15] M. Malfante, M. Dalla Mura, J. I. Mars, and C. Gervaise. Automatic fish sounds classification. *The Journal of the Acoustical Society of America*, 139(4):2115–2116, 2016.

[16] M. Malfante, M. Dalla Mura, J.-P. Métaxian, J. I. Mars, O. Macedo, and O. Inza. Machine learning for volcano-seismic signals: Challenges and perspectives. *IEEE Signal Processing Magazine*, 2017 in press.

[17] K. Nogueira, O. A. Penatti, and J. A. dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017.

[18] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[21] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.