BI-CLASS CLASSIFICATION OF HUMPBACK WHALE SOUND UNITS AGAINST COMPLEX BACKGROUND NOISE WITH DEEP CONVOLUTION NEURAL NETWORK

Dorian Cazau & Riwal Lefort
ENSTA Bretagne, Lab-STICC (UMR CNRS 6285)
2 rue François Verny, 29806 Brest Cedex 09, France
dorian.cazau@ensta-bretagne.fr

Julien Krywyk
logiCells
55, rue Labourdonnais - Apt 303
97400 St Denis, La Réunion, France

Jean-Luc Zarader
Institut des Systemes Intelligents et de Robotique
Université Pierre et Marie Curie (Paris VI), F-75005, Paris, France

Olivier Adam
Sorbonne Universités, UPMC Univ Paris 06, UMR 7190
Institut Jean Le Rond d'Alembert, F-75005, Paris, France

ABSTRACT

Male humpback whales produce complex structured songs during the winter-spring breeding season. These songs are recorded within a marine soundscape that melt different biological, environmental and anthropenic sources. Developing computational tools to automatically provide an insight into these song structures, and allow for quantitative analysis over a long time period, is a current challenge for scientists, with unsatisfied and dataset-dependent detection results. In this paper, we explore the applicability of Convolution Neural Network (CNN) method to identify the inherent structure of whale sound units and extract them from the complex time-varying soundscape. In the evaluation stage, we present 6 bi-class classification experimentations of whale sound detection against different background noise types, such as rain, wind, marine traffic and song chorus. In comparison to classical FFT-based representation, we showed that the use of image-based pretrained CNN features brought higher performance to classify spectrograms (FFT-based features) of whale sounds and background noise. Furthermore, the observed rates of correct classification remained very robust and satisfactory with varying signal-to-noise ratio and background noise types.

1 INTRODUCTION

1.1 UNDERWATER SOUNDSCAPES

In the frequency band from few tens of Hz up to 50 kHz, dominant sources of ambient noise in the ocean can be broadly divided into sounds resulting from geophony (i.e., sounds from natural physical processes, e.g. wind-driven waves, rainfall, seismicity, breaking waves, current), biophony (i.e. sounds from biological activities, e.g., whale vocalizations, snapping shrimp beds, fish, coral reef), and anthropophony (i.e. man-made sounds, e.g. commercial shipping, sonar, seismic prospecting, oil and gas surveys) (Knudsen et al., 1948; Wenz, 1962). All these sources contribute conjointly to the noise spectrum characteristics (e.g. pressure level, spectral slope) in varying degrees, depending on their strength and conditions prevailing at the measurement location. Thus, the underwater ambient sound field contains quantifiable information about the physical and biological marine environment. To extract these information, passive acoustic systems have been used for monitoring, recording and interpreting in a continuous and autonomous way the underwater acoustic signal, facilitating an all-weather and all-season ocean monitoring.

1.2 HUMPBACK WHALE SONGS

Vocal repertoires of humpback whales range widely, with a great variety of bandwidths, durations and intensities of their sounds. Scientists have particularly been interested in the complex stereotyped songs that male individuals of humpback whales emit during the winter-spring breeding season. Such a complex acoustic display is indeed thought to play an important role in both the mating ritual and male to male interaction (see e.g. Tyack (1981)). Hence, the need to detect, and further classify, the unit constituents of a song objectively and systematically has become crucial to analyse the songs and to allow processing large data set.

However, some major difficulties are raised with this objective. First of all, during recording sessions on the field, capturing the song of a single whale is a difficult task, as during breeding season all whales gather in a same narrow geographical area, called "hot-spot", and sing altogether. For example, the Sainte Marie Channel in Madagascar (the 60 km-long Ste Marie Island is less than 30 km far from Madagascar Island), which is our study area, gathers a few hundreds of whales every breeding season. The resulting song chorus is composed of many overlapping vocalizations, which is further complicated due to underwater acoustic propagation, including geometrical spreading loss, dispersion due to underwater structures, channel inhomogeneities, stratification. These phenomena are reinforced in shallow water environment, like the Sainte Marie channel with a maximal column water depth of 80 m. Second, in addition to this biological activity, as stated earlier, underwater ambient noise is also dominated by geophonic and anthropic sources Wenz (1962). These sources make the background acoustic environment highly variable during whale songs. Third, the acoustic richness and variability of humpback whale sound units make them difficult to classify into discrete categories, as performed by human experts.

1.3 PAPER OBJECTIVES & CONTRIBUTION

Passive underwater acoustic data offer a great opportunity for the development of non-invasive integrative continuous monitoring systems of anthropogenic, biological and environmental sources. This integrative characteristic presents a precious advantage over classical monitoring systems, that need one specific sensor per source category, but also raises important challenges to extract useful information and disambiguate between the different sources. In this paper, we are interested in extracting vocalizations of singing whales from a complex marine environment. Most computational tools (e.g. PAMguard Gillespie (2008), Ishmael Mellinger (2001)) for this task are based on hand-engineered features optimized for one specific source (e.g. a single whale in a given geographical location). However, in highly time-variable marine environments, the inconceivable feature dimensionality of the sources to identify and the ambivalent nature of latent variables, that are poorly apprehended by these methods, make them fail to generalize to other acoustic scenarios. In other words, the extraneous complexities of the ocean ambience prevent the feature engineering methods to capture the invariant features. To deal with this processing complexity, probabilistic approaches have also been used with the adaptive extraction of feature dictionaries. Halkias et al. (2013) applied a deep network built with Restricted Boltzman machines on FFT-based images, boosted by a sparse auto-encoder. Cazau & Adam (2016) used a Bayesian statistical learning approach based on Probabilistic Latent Component Analysis. In this paper, we explore the applicability of the deep learning method (Hinton et al., 2006) to the task of identifying properly the sound units of a singing humpback whale within a complex multi-source underwater soundscape. The deep learning approach, with architectural resemblance to the mammalian neocortex, is deemed to capture the useful information hidden deeply in the actual observation by automatically learning features in a hierarchical manner. In particular, Convolutional Neural Networks (CNNs) have proven to be very successful frameworks for image recognition. In the past few years, variants of CNN models have achieved an impressive suite of results on standard recognition datasets and tasks, especially on the renowned ImageNet dataset for object classification, starting from AlexNet (A. Krizhevsky & Hinton, 2012), OverFeat (P. Sermanet & LeCun, 2013), GoogLeNet (C. Szegedy & Rabinovich, 2014), and a model by (K. He & Sun, 2015) with classification accuracy surpassing human-level performance. Based on these state-of-the-art performance, such a learning process is expected to be robust enough to tolerate soundscape and SNR variations within our task. In the evaluation stage, we present 6 bi-class classification experimentations of whale sound detection against different background noise types, such as rain, wind, marine traffic and song chorus. Correct recognition rates are reported and discussed. In particular, we are interested in quantifying the performance of image-based pretrained

CNN features to classify spectrograms (FFT-based features) of whale sounds and background noise, in comparison to classical FFT-based representation.

2 METHODS

Figure 1 shows the block diagram of our classification task. The input of this processing chain is a temporal slice m of a time series x. The windowed slice of signal is named $x^{(m)}$. The output result is a binary decision classifying this slice m into a whale sound (i.e. 1) or background noise (i.e. 0). We used image-based pretrained CNN features and a SVM classifier for a bi-class classification of spectrograms (FFT-based features) of whale sounds and background noise. In the following, we detail briefly each processing block.

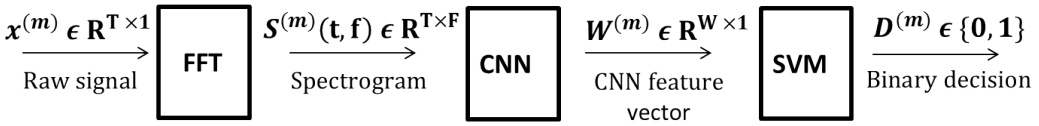

Figure 1: Block diagram of our classification task.

2.1 PROCESSING BLOCKS

FFT The power spectrum is used to measure the amplitude of discrete frequency components. The power spectra for each of the data segments t are then combined to form an array in frequency and time $S^{(m)}(t, f)$ that we call spectrogram,

$$S^{(m)}(t, f) = 10log_{10}\frac{|X_t^{(m)}(f)|}{p_{ref}^2} \tag{1}$$

where $X_t^{(m)}(f) = \sum_{n=0}^{N-1} x_t^{(m)}[n]e^{\frac{-i2\pi fn}{N}}$ is the FFT transformed of the time series signal x, and p_{ref} is a reference pressure of 1 μPa. The spectrogram is then used to measure the amplitude of discrete frequency components against time.

CNN CNN is a type of artificial neural networks inspired by visual information processing in the brain. To recognize complex features from the visual information, CNN consists of several layers which extract and repeatedly combine low-level features for the composition of high-level features. The composed high-level features are used for CNN to classify an input image. In our processing chain, FFT-based spectrograms $S(f, t)$ were converted into 256 $\times$ 256 pixels .jpg images in uint-8 format. Then, each image is used as input of the pre-trained CNN, giving a feature vector called the CNN code. For the process and the simplicity of computation, many CNN structures are described by combinations of convolutional, pooling, and fully-connected layers. Convolutional layer (C-layer) extracts higher-level features by convolving received feature maps from the previous layer and activating the convolved features. By supposing that j-th layer has N_j nodes (neurons) whose output feature map is h_i^j, then :

$$h_i^j = g(\sum_{n=1}^{N_{j-1}} \boldsymbol{\phi_{in}} * h_n^{j-1} + b_i^j) \tag{2}$$

for $j = 1, ..., N_j$, where $\boldsymbol{\phi_{in}}$ is the convolutional filter connecting i-th node (neuron) on the j^{th} layer and n-th node (nueron) on the j 1-th layer, b_i^j is the bias for the i-th output on the j^{th} layer, and f is the activation function. Given its efficiently of computation and common use, ReLU $g(x) = max(0, x)$ is the used activation function. A C-layer usually is followed by a pooling layer (P-layer) which reduces the dimension of feature maps by "max pooling". The max pooling downsamples the input feature maps by striding a rectangular receptive field and taking the maximum in the field. For

example, we use 2×2 receptive field with stride 2 in our research, which reduces the dimension of feature map by 1/4. After couples of pairs of C-layer and P-layer, a fully-connected layer (F-layer) integrates high-level features and produces compact feature vectors:

$$h_i^j = g(\sum_{n=1}^{N_{j-1}} < \boldsymbol{\phi_{in}}, h_n^{j-1} + b_i^j) \tag{3}$$

for the j^{th} F-layer. Note that we use inner-product rather than convolution between the filter Win and the input feature map h_n^{j-1}. Like the C-layers, ReLU is used as the activation function of the F-layers. On the final layer, say J-th layer, the output layer produces the posterior probability for each class by the softmax function. The idea of exploring CNN features is motivated by their usefulness on a wide variety of tasks. The activations which are the output $W^{(m)}$ of CNN layers can be interpreted as visual features.

SVM Given a set of training examples of CNN features, each marked as belonging to whale sounds or background noise, an SVM training algorithm builds a model that assigns new examples from a test dataset to one category or the other, making it a non-probabilistic binary linear classifier (R.-E. Fan & Lin, 2008).

2.2 Implementation details

Each slice of time series $x^{(m)}$ is 2-s long. The spectrogram parameters are as follows: segment size: 22 ms (1024 samples with a sample frequency f_s=44.1 kHz) ; 50% overlap (temporal resolution: 11 ms) ; F = 1024 samples (i.e. FFT size, providing a spectral resolution of 11 Hz) ; Hamming window.

Features extracted from pretrained CNN have been successfully used in computer vision tasks such as scene recognition, object attribute detection and achieves better results compared to handcrafted features (A. S. Razavian & Carlsson, 2014). Given the usefulness of CNN features, our current research shares common efforts to further assess the features and demonstrate how we can use the features for other tasks (Athiwaratkun & Kang, 2015). CNN models such as AlexNet, GoogLeNet that are pre-trained on ImageNet can been used as generic feature extractors for other tasks. This is done by removing the top output layer and using the activations from the last fully connected layer (CNN codes) as features. To setup a pretrained network, a first downloading of the weights file is required, and then a setup of the architecture, and finally the loading the downloaded weights. It turns out that this off-the-shelf feature extractor give features that yield better results than handcrafted features (A. S. Razavian & Carlsson, 2014). Pre-trained CNNs from the Vlfeat project[1] were used. We tried the following networks from the imagenet framework (Chatfield et al., 2014): imagenet-vgg-f, imagenet-vgg-m, imagenet-vgg-s. After several experimentations, imagenet-vgg-f showed the best performance and was consequently used in our study. The output dimension of the CNN features is set to 4096.

The two-class linear SVM used in this study is from LibLinear[2] (R.-E. Fan & Lin, 2008). Default parameters of the toolbox have been used, with in particular a radial basis function $e^(- \gamma * |u - v|^2)$ for the kernel of degree 3. We did not try to optimize such parameters for better classification performance.

2.3 Evaluation dataset

2.3.1 Humpback whale songs

Since 2007, we have been recording humpback whales songs in the Sainte Marie Island Channel, North East Madagascar (16 °S , 49 °E), during the month of August. In this paper, we chose three songs from different whale singers. To ensure uniqueness of these singers, these songs were selected from different years, in 2007, 2010 and 2013. Recordings were done from the boat (motor off) using the COLMAR Italia GP280 hydrophone and digitalized by the Tascam HD-P2 recorder

[1]Website link: `http://www.vlfeat.org/matconvnet/pretrained`

[2]Website link: `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`

at a frequency sample (f_s) of 44.1 kHz and 16 bits. The hydrophone was at 20 m depth, around the middle of the water column, and was taken as an average depth estimation of the singing whale position. Recordings were done close to the singers (closer than 200 m) to improve signal-to-noise ratio. Great care was taken to record only the singing focal animal, in order to prevent any confounding effect of background noise and overlapping vocalizations. It is also assumed that these songs were recorded in a known and unique behavioral context that was identified as the mating song emitted during breeding season.

In each of the three songs, 500 sound units were manually selected through a visual inspection of the spectrogram outcomes under Adobe Audition software, for a total of 1500 sound units. All suspect features (e.g., spray whale splashes, boat noise, features from non-vocal social sounds) affecting sound units were removed. Spectrogram examples of humpback whale sound units are provided in figure 2.

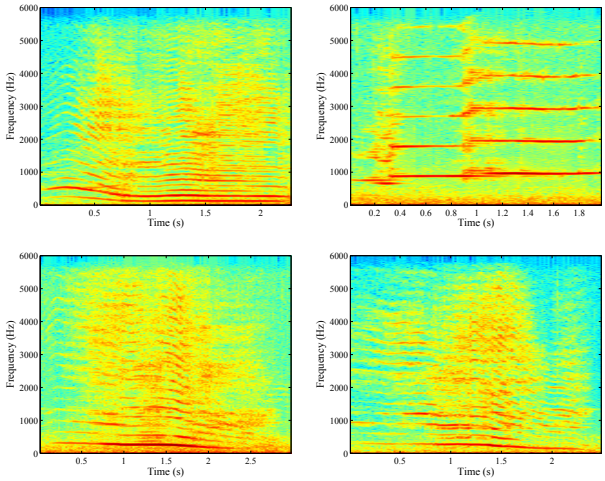

Figure 2: Spectrogram examples of a few sound units of a singing whale.

2.3.2 BACKGROUND NOISE DATA

We collected data from four different background noise types: wind (geophony), rain (geophony), marine boat traffic (anthropogeny) and humpback whale song chorus (biophony), listed in table 1 and detailed in the following.

Clean background Clean background data were extracted from the songs detailed in Sec. 2.3.1, and correspond to the time intervals between each sound unit manually labelled. As already mentioned, these song recordings were performed in clean meteorological conditions ans without observed external sources.

Wind Wind-driven background data samples were extracted from two different underwater sound datasets recorded in shallow water environments, namely off Saint-Pierre-et-Miquelon (French archipelagos close to Newfoundland, Canada, 47 °N , 56 °W) and in the Sainte-Marie channel (from other recordings and time periods of the songs used previously). We only collected wind events belonging to the see state of Beaufort scale 6 or higher (identified to high and gale winds with a speed higher than 7 m.s^{-1}) (Knudsen et al., 1948). Sea state 1 has not been used as noise produced by weak wind speed is too low to be detected (Pensieri et al., 2013). These wind speed values were extracted from a moored meteorological buoy in Saint-Pierre-et-Miquelon (from the National Climatic Data Center[3]) and the European Center for Medium range Weather Forecasting (ECMWF) analysis, that provide gridded daily-averaged wind and wind stress fields over global oceans through satellite imaging and assimilation models (Bentamy & Croize-Fillon, 2011).

[3]Website link: `https://www.ncdc.noaa.gov/`

Rain As for wind, rain-driven background data samples were extracted from the Saint-Pierre-et-Miquelon and the Sainte-Marie channel. We also only collected rain events superior to 10 mm/h, that have already been proved to have a significant effect on ambient noise (Pensieri et al., 2015). These precipitation rates were extracted from a moored meteorological buoy in Saint-Pierre-et-Miquelon (from the National Climatic Data Center[4]) and the European Center for Medium range Weather Forecasting (ECMWF) analysis.

Marine boat traffic Our data samples for background noise from marine boat traffic were extracted from the Saint-Pierre-et-Miquelon dataset in recordings from 19/08/2010 to 02/11/2010 and in the Sainte Marie channel. Most encountered ships are motor boats (from 15m to 25m long) used for material shipping or passenger transport. Ship tracklines around the recording hydrophone ($\approx$ 40 km^2 searching area) were taken from the World Meteorological Organization Voluntary Observing Ship Scheme (VOS) climate project [5]. This project is based on ships of many countries that voluntarily transmit their location updates multiple times a day under this program.

Song chorus Song chorus were extracted from the Sainte Marie channel (Madagascar), from other recordings and time periods of the songs than the ones used previously.

2.3.3 Construction of evaluation datasets

Evaluation datasets were built automatically by combining sound units of the focal whales with a noise sample taken successively from the different noise background types. Such a generative process allows generating a large set of labelled training/test data with a minimal amount of human labour, allowing the inclusion of an important variety of vocalizations melted with different ambient noise backgrounds. One may criticize that they are less realistic than real recordings, but actually a similar process has already been usefully used in various application domains (e.g., in music information retrieval (Lee & Slaney, 2008) and acoustic scene analysis (Lafay et al., 2016)).

Each sound unit file was first unitary normalized, and buffered into non-overlapped 2-s time windows. Each time window was then added to some background noise, randomly selected from the 5 background types previously described, respecting a given Signal-to-Noise Ratio (SNR). This SNR was defined as a sensitivity parameter for our numerical experimentations, ranging from -10 dB to 10 dB. SNR is defined as

$$SNR = 10log_{10}\frac{< x_s^2 >}{< x_n^2 >} \tag{4}$$

where $< x_s^2 >= \frac{1}{T}\int_0^T < x_s^2 >^2 (t)dt$ and where x_s represents the recorded pressure of the sound unit time series, x_n the background noise, and T= 2s is the duration of the signal. Note that negative SNR in the time domain does not imply negative SNR for individual frequencies following a transformation into the Fourier domain.

As detailed in table 1, we designed 6 different numerical experimentations based on the dataset of sound units from the focal whale, combined with the previous evaluation background types. The sixth evaluation dataset was built by randomly selecting different noise samples from all possible background types, similarly to a one-against-all classification scenario.

2.4 Evaluation procedure

Our numerical experimentations consist of a two-class classification task, that aims to detect focal whale sound units against noise background. Training and test datasets were defined with sound samples from different years, and were set to 300 and 200 samples, respectively. For each year couple, final classification accuracy resulted from the average scores provided by a Monte Carlo simulation with 100 iterations. We then also averaged over years. Results are then displayed as 2D confusion matrices of Whale sound vs Noise background.

[4]Website link: `https://www.ncdc.noaa.gov/`

[5]Download link: `http://www1.ncdc.noaa.gov/pub/data/vosclim/`

Experimentation names	Evaluation background types
E_1	Clean background
E_2	Wind
E_3	Rain
E_4	Marine traffic
E_5	Song chorus
E_6	Wind + Rain + Marine traffic + Song chorus

Table 1: Details of numerical experiments.

3 Results and Discussions

Figure 3 represents confusion matrices of the experiments E_1 to E_6, where a SNR of 0 dB was set. Experiment E_1 is the baseline performance, with a correct recognition rate of 95 %, and a false alarm rate of 1%. Standard deviations over the Monte-Carlo iterations never exceeded 2 % for all experimentations, neither over the different years, showing a great consistency of our results throughout our complete dataset. All other background noise types that have been used to corrupt whale sounds decreased baseline performance, from - 9 % to -36 % in the correct recognition rate, and from + 3 % to + 36 % in the false alarm rate.

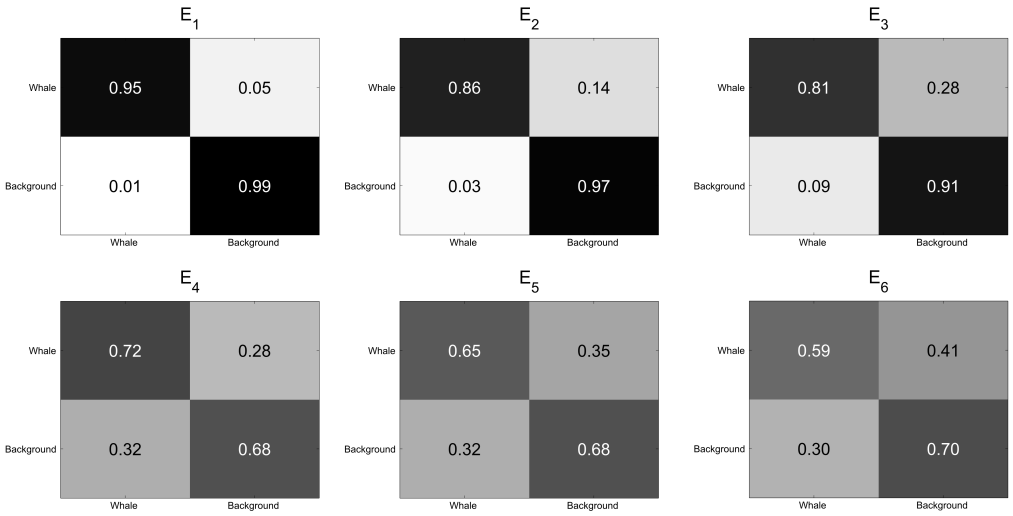

Figure 3: Confusion matrices of the experiments E_1 to E_6.

Environmental noises like wind and rain had the smallest impact on classification performance. These acoustic sources have broadband frequency spectra, with a rather uniform distribution of the energy in the spectrum (Nystuen et al., 2015), which does not distort that much whale sound spectra. On the contrary, marine traffic (especially at short distance from the hydrophone) generates a great variety of sound spectra that are more similar to whale sounds, with most often a time-varying harmonic content. Naturally, song chorus is the background noise type that decreases the more classification performance, as it is composed of a multitude of whale sounds from different singers.

Figure 4 shows the dependency of correct recognition rates of whale sound units on SNR for the different background noise types. Globally, decrease in SNR between whale sounds and background noise degrades classification performance. Variations in SNR appear to affect to a higher extent performance of experiments E_4 and E_5. AS already commented, the acoustic content of these noise types are more similar to whale sounds, and consequently lower SNR vales reinforce ambiguity between them.

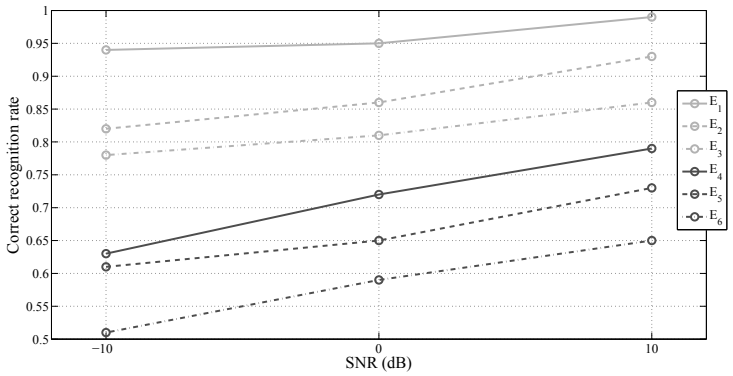

Figure 4: Correct recognition rates against SNR for each experiment, i.e. corresponding to the different background noise types.

We eventually compared detection performance of our method with two other systems: PamGuard[6] Gillespie (2008) and a PLCA-based system that shows promising performance for humpback whale sound detection (Cazau & Adam, 2016). Figure 5 represents confusion matrices of the experiments E_6, where a SNR of 10 dB was set. In comparison to classical FFT-based representation, we observe that the use of image-based pretrained CNN features brought higher performance to classify spectrograms (FFT-based features) of whale sounds and background noise, with gains in correct classification rates of + 12 % and + 7 %, in reference to PamGuard and PLCA systems, respectively.

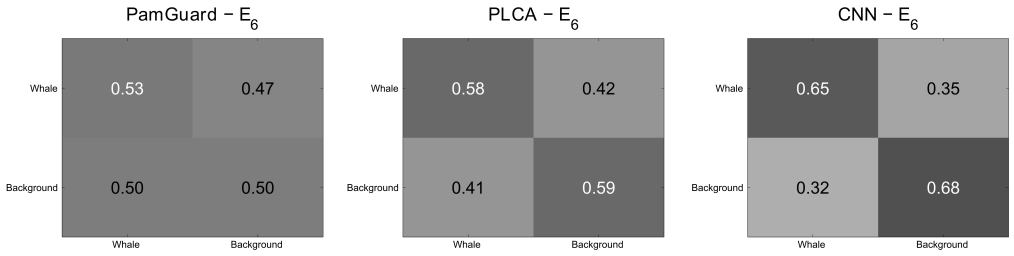

Figure 5: Confusion matrices of the experiments E_6 with different classification systems: CNN, PLCA and PamGuard.

4 CONCLUSION

In this workshop paper, we show that the inherent learning process of deep learning approach could greatly serve the analysis of humpback whale songs. This learning process includes a sequential abstraction process, where the lowest layer represents the original observed data, and higher levels represent more abstract representations less susceptible to the variations in input space, while they preserve the salient features. Properly identifying the sound units of singing humpback whales is a complex task in the realms of underwater environment, due one side to the multiple overlapping singers and filtering environment characteristics. In comparison to classical FFT-based representation, pretrained CNN features are shown to bring higher performance to classify spectrograms (FFT-based features) of whale sounds and background noise. The background noise types of marine traffic and song chorus induced the highest degradation in classification performance.

ACKNOWLEDGMENTS

This work was in part financially supported by the Dirac association (Paris, France).

[6]Publically available sofwtare that has been widely used in the marine mammal community for the task of vocal sound detection in particular. Downloading links of this software is `http://www.pamguard.org/`.

REFERENCES

I. Sutskever A. Krizhevsky and G. E. Hinton. *Advances in Neural Information Processing Systems*, chapter Imagenet classification with deep convolutional neural networks, pp. 1097–1105. F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. Curran Associates, Inc., 2012.

J. Sullivan A. S. Razavian, H. Azizpour and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *CoRR*, vol. abs/1403.6382, 2014.

Ben Athiwaratkun and Keegan Kang. Feature representation in convolutional neural networks. . *arXiv preprint arXiv:1507.02313*, 2015.

A. Bentamy and D. Croize-Fillon. Gridded surface wind fields from metop/ascat measurements. *Inter. Journal of Remote Sensing*, 2011. doi: 10.1080/01431161.2011.600348.

Y. Jia P. Sermanet S. Reed-D. Anguelov D. Erhan V. Vanhoucke C. Szegedy, W. Liu and A. Rabinovich. Going deeper with convolutions. *CoRR*, vol. abs/1409.4842, 2014.

D. Cazau and O. Adam. A plca model for detection of humpback whale sound units. *Advances in Applied Acoustics (AIAAS)*, 5:1–5, 2016. doi: 10.14355/aiaas.2016.05.001.

K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

D. Gillespie. Pamguard: semiautomated, open source software for real-time acoustic detection and localization of cetaceans. In *Proceedings of the Institute of Acoustics*, 2008.

X. C. Halkias, S. Paris, and H. Glotin. Classification of mysticete sounds using machine learning techniques. *J. Acoust. Soc. Am.*, 134:3496–3505, 2013.

G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

S. Ren K. He, X. Zhang and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. 2015.

V.O. Knudsen, R.S. Alford, and J.W. Emling. Underwater ambient noise. *J. of Marine Res.*, 7: 410–429, 1948.

G. Lafay, M. Lagrange, M. Rossignol, E. Benetos, and A. Roebel. A morphological model for simulating acoustic scenes and its application to sound event detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1854–1864, Oct 2016. ISSN 2329-9290. doi: 10.1109/TASLP.2016.2587218.

K. Lee and M. Slaney. Acoustic chord transcription and key extraction from audio using key-dependent hmms trained on synthesized audio. *IEEE Trans. Audio Speech Lang Proc.*, 16:291–301, 2008.

D.K. Mellinger. Ishmael 1.0 users guide. Technical report, Natl. Oceanogr. Atmos. Admin. Tech. Memo. OAR-PMEL-120 (NOAA PMEL, Seattle), 30 pp, 2001.

J. A. Nystuen, M. A. Anagnostou, E. M. Anagnostou, and A. Papadopoulos. Monitoring greek seas using passive underwater acoustics. *J. of Atmospheric and Oceanic Technology*, 32:334–349, 2015.

X. Zhang M. Mathieu R. Fergus P. Sermanet, D. Eigen and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013.

S. Pensieri, R. Bozzano, M.N. Anagnostou, E.N. Anagnostou, R. Bechini, and J.A. Nystuen. Monitoring the oceanic environment through passive underwater acoustics. In *OCEANS - Bergen, 2013 MTS/IEEE*, pp. 1–10, June 2013. doi: 10.1109/OCEANS-Bergen.2013.6607995.

S. Pensieri, R. Bozzano, J. A. Nystuen, E. N. Anagnostou, M. N. Anagnostou, and R. Bechini. Underwater acoustic measurements to estimate wind and rainfall in the mediterranean sea. *Hindawi Publishing Corporation, Advances in Meteorology*, pp. 1–18, 2015.

C.-J. Hsieh X.-R. Wang R.-E. Fan, K.-W. Chang and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

P Tyack. Interactions between singing hawaiian humpback whales and conspecifics nearby. *Behavioral Ecology and Sociobiology*, 8:105–116, 1981.

G. M. Wenz. Acoustic ambient noise in the ocean: Spectra and sources. *J. Acoust. Soc. Am.*, 34: 1936–1956, 1962.