Automatic recognition of underwater acoustic signature for naval applications

E. Artusi*[a, b], F. Chaillan[b],
[a]MINES ParisTech PSL Research University, 1 rue Claude Daunesse, 06904 Sophia Antipolis, France; [b]NAVAL Group, NAVAL Research, 83190 Ollioules, France

ABSTRACT

In a military context, where human capabilities are no longer sufficient to process quickly and reliably an ever-increasing amount of data, the implementation of algorithms based on Artificial Intelligence (AI), through the computing power of modern infrastructure, increases the ability to interpret and correlate massive heterogeneous data. This article will present an original automatic underwater acoustic signature recognition technique. The experiments are carried out from public underwater acoustic dataset. Besides, the performance of three architectures based on Mel-frequency cepstral coefficients (MFCCs), Machine Learning techniques and neural networks are compared.

Keywords: military naval application, mission, underwater acoustic signature, signal processing, classification, Machine Learning, Deep Learning

1. INTRODUCTION

The French Navy's surface warships and submarine vessels are designed to carry out various missions as for instance, maritime surveillance, infrastructure protection, participation in interoperability exercises and intervention on a theatre of operation. Thus, the conduct of a naval mission requires the simultaneous integration of a large number of information related to the ship and to its environment, such as sea state, meteorology, AIS data [1], [2], kinematic data and so on.

In addition, a mission in an operational context is obviously subject to the notion of risk and must therefore be reconfigurable in order to being able of managing many contingencies [3]. It is in this context that the ship's commander, in order to make the mission a success [4] must quickly analyze large quantities of massive and heterogeneous data in order to take the right decision depending on the situation. Nowadays, the indicators leading to decision-making are estimated by human beings based on their discernment and expertise. However, human capabilities are no longer sufficient to reliably and quickly process the amount of data collected by the fleet and its respective environment's many sensors. In other words, data volumes are forever increasing while the operational constraint like speed or efficiency used to achieve the mission remains.

For example, during naval missions, the capability to recognize underwater acoustic signatures is essential to be aware in real time of the evolution of the fleet's environment. The operator dedicated to sound analysis identifies and lists underwater acoustics noises of interest among all the detections. Then, as soon as possible he advises the Commander in case of classification of acoustic signature interpreted as a threat. Besides, this is an important function as it allows distinguishing between mechanical noise from a foe's vessel and normal biological activities issued from the underwater landscape; in practice it is difficult to automatically separate similar sounds as well as a human ear.

This fact, associated with the need of rapid classification might drive this operator into a mental overload and a very stressful discomfort. In order to provide him efficient assistance for identifying potential threats despite of fatigue, mental overload and stress it is necessary to design an automatic underwater acoustic signature recognition system.

[*] eva.artusi@mines-paristech.fr, eva.artusi@naval-group.com

Despite automatic sound recognition being an important aspect in emerging civilian applications, recognizing efficiently sounds in noisy underwater conditions for military applications and classifying these sound events from a passive device remains a serious difficulty. Actually, the recognition of some everyday real-life sounds such as ice breaking, the barking of a dog, music, or more generally any other acoustic event swamped by noise as well as several sound sources contained within the same recording [5] or bird species based on their song [6], are some of the emerging civilian applications. They ensue from the sound propagation' specificities, the diverse interesting signals' representability like for example noises issued from a submarine. Classification methods inspired by the field of speech recognition using audio characteristics such as MFCCs coefficients [7], [14], which effectiveness for sound classification have been positively shown for civilian applications, should be of interest for military purposes, modulo some upgrades.

This article focuses on the creation of a database, the establishment of classes and on the selection of architectures including machine learning (ML) and deep learning (DL) techniques. It also focuses on a performance assessment to quantify the relevance of our suggested solutions regarding underwater sound recognition designed for military naval applications. This paper is organized as follows: in section 2, the three architectures we will experiment to classify underwater sound recognition are introduced. In section 3, we describe our dataset made up of public sounds, the different configurations we have studied, and the results obtained for each case. Our conclusions and future works are presented in section 4.

2. METHODOLOGY

Historically, acoustic event detection (AED) was processed with characteristics such as MEL Frequency Cepstrum Coefficient (MFCC) combined with classifiers based on Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) or Support Vector Machine (SVM) techniques [8], [9], [10]. More recent approaches use deep neural networks including convolutional networks [11] and recurrent networks [12]. In this article, we will describe the following three architectures:

- MFCCs and SVM,
- VGGish [19],
- An original enhanced version: VGGish and one dense layer.

In the following, after describing the three of them we will evaluate their performance.

2.1. Architectures description

In accordance with the objectives of this study, instead of classically setting up the hyperparameters, which is done, in practice with 5-fold cross validations to improve the reliability of the estimated hyperparameters, here we simply want to show that at least one of the three following architectures would be suitable for military naval applications.

2.2.1. MFCCs features and SVM method

For sound recognition, a feature extraction technique that extracts both linear and non-linear features is required. That is why here the Mel-frequency Cepstral Coefficients (MFCC) are used [23]. The MFCC is a type of frequency representation of the signal, in which any linear frequency f is mapped to the MEL scale according to the non-linear transformation:

$$m(f) = 2595\ \log_{10}(1 + f/700) \tag{1}$$

Inversely, each MEL frequency is mapped to the frequency scale according to the following relation:

$$f(m) = 700\left(10^{m/2595} - 1\right) \tag{2}$$

Consequently, instead of a linear scale we have now a new MEL scale providing rescaled frequencies in accordance with human ear behavior:

- An almost linear scale for frequencies less than 1 kHz, as in this case the transformation can be reduced to:

$$m(f) \approx 3.7f \tag{3}$$

- A log scale for frequencies above 1 kHz.

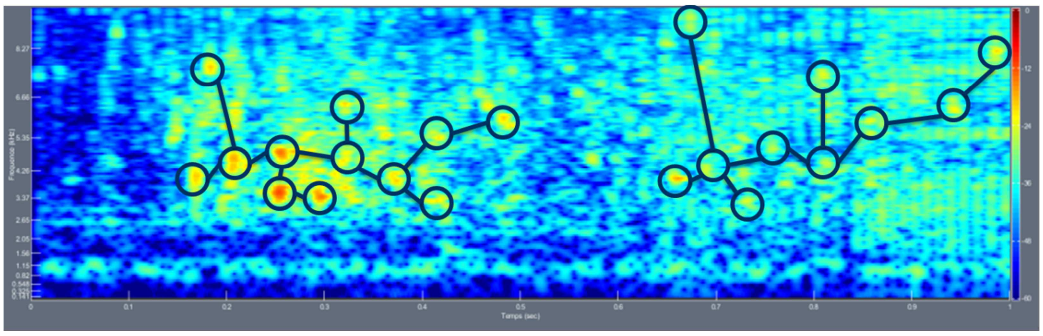

Figure 1: Illustration of the capability of MFCC's coefficient to identify thin details of underwater acoustic signal. X-axis correspond to Time (horizontal) and Y-axis corresponds to Mel frequencies (vertical).

As shown by Figure **1**, the MFCC transform is capable to capturing the important characteristic of audio signals. MFCC is widely used in speech recognition and in sound event recognition [13]. Thus, from any audio signal the MFCC's coefficients are computed according to the processing described by Figure 2.

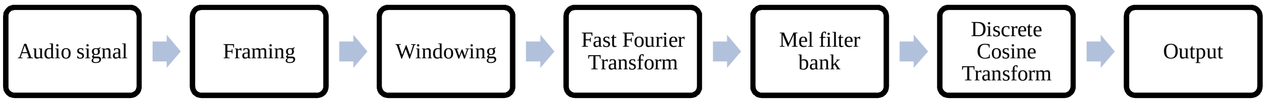

Figure 2 : MFCC block diagram [22]

In details, the audio signal is segmented into small duration blocks known as frame. Each of the above frames is multiplied with a hamming window. Then, the Power Spectral Density (PSD) of each frame is obtained by taking the squared modulus of its Fast Fourier Transform (FFT). Then, each PSD is multiplied elementwise by a set of triangular band pass filter in order to get filtered magnitude spectrum. It also reduces the size of features involved as there is generally less frequency channel considered than for classical Fourier analysis. Then, the Discrete Cosine Transform (DCT) is applied on the log energy obtained from the triangular band pass filters to finally provide Mel-scale cepstral coefficients. More precisely, $\forall n = 0 \ldots N-1$, $x_l[n]$ is the N-signal containing the samples of the frame l. Its PSD is computed for channels $k = 0 \ldots N_{fft}/2 - 1$. Besides, the MEL M-filter bank $H_{MEL}[k, m]$ is defined $\forall m = 0 \ldots M-1$ by

$$\mathrm{H}_{\mathrm{MEL}}[k,m] = \begin{cases} 0 & \text{if} \quad k < f(m-1) \text{ or } k > f(m+1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & \text{if} \qquad f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & \text{if} \qquad f(m) \le k \le f(m+1) \end{cases} \tag{4}$$

This relation shows MEL filters are nothing else than triangular filters centered on each considered MEL frequency m. Finally, according to relation (4) and to the flow graph presented in Figure 2, $\forall m = 0 \ldots M-1$, the $MFCC_l[m]$ coefficients of the l^{th} audio frame $x_l[n]$ sampled at F_s Hz, windowed by $h[n]$, with are given analytically by the following relation:

$$MFCC_l[m] = \sum_{p=0}^{M-1} \log_{10}\left(\sum_{k=0}^{N_{fft}/2-1} \frac{2}{F_s N} \left| \sum_{n=0}^{N-1} x_l[n] h[n] e^{-2i\pi nk} \right|^2 \mathrm{H}_{\mathrm{MEL}}[k,p] \right) \cos\left(\frac{\pi}{N}(p + \frac{1}{2})m\right) \tag{5}$$

Hence, let suppose we have L lags, then $\forall l = 0 \ldots L-1$ if we consider each M −vector $MFCC_l[m]$ given by relation (5) as a column of an (M, L) −image, it allows us to build the MFCC representation of the audio signal, such as the one illustrated by the Figure 3. MFCC's are commonly used as feature extraction technique in speech recognition system because it approximates the human system response more closely than any other system; nevertheless, it suffers of a lack of robustness in case of low signal-to-noise ratio observation.

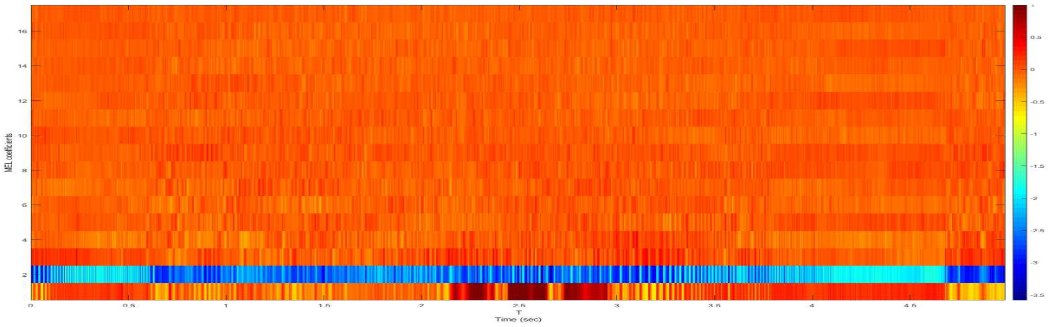

Figure 3 : Mel frequency Cepstrum of a biological signal.

In our case, our automatic underwater acoustic signature recognition must attend the operator dedicated to sound analysis, that's why the use of MFCCs seems to be pertinent. Following the extraction of MFCCs, we use the SVM algorithm as a classifier [15].

2.2.2. The VGGish network

Recently, high-performance Neural Networks for image classification such as AlexNet [16], VGG [17], Inception [18] and others are being tested for audio classification problems. These techniques seem to perform better as the classical one with MFCCs extraction. The architecture we have selected is based on a recent publication [19] trained on millions of YouTube videos. Hence, we take a VGG network slightly modified called VGGish to match with the sound classification technique found in [20].
First, Mel spectrogram is used as input features, computed from the spectrogram of each audio file. Figure 4 shows an example of the Mel spectrogram of a biological signal.

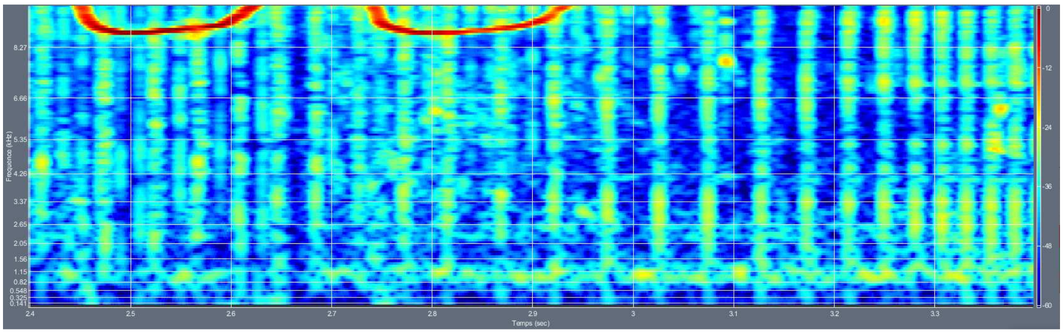

Figure 4: example of a Mel spectrogram of a biological signal

The VGGish we take is a variant of the VGG model described in [17]. In particular, its architecture is modified so that it uses Configuration A with 11 weight layers. Furthermore, the input size was changed to 96x64 for log Mel spectrogram audio inputs. Then, the last group of convolutional and maxpool layers is removed, so there are only four groups of convolution-maxpool layers instead of five, and instead of a 1000-wide fully connected layer at the end, there is a 128- wide fully connected layer. This acts as a compact embedding layer. We use the provided VGGish model that is pre-trained on a large YouTube dataset, which is a preliminary version of what later became YouTube-8M, and then we fine-tune the VGGish model parameters for our application we just add a classifier layer at the end, consisting of parallel logistic for classifier, one per class, which allows multi-class task

Our approach can be considered belonging to Transfer Learning domain [25], as the original VGGish model is modified in order to fit with the new targets we want to recognize, where these new targets are nothing else than classes of underwater acoustics noises.

2.2.3. The VGGish network and one dense layer

Starting from the classical approach [19], we propose an original architecture composed of VGGish followed by one fully connected layer. Here, the VGGish network is considered as a “warm start” for the lower layers of a model that takes audio features as input and adds more layers on top of the VGGish embedding. It is used to fine-tune VGGish because our dataset is different from the typical made up with YouTube video clips.

2.2.4. Synthesis of the three architectures

In summary, performance will be evaluated on:

- A first architecture based on the classification using an SVM classifier of the MFCCs of each audio signal.
- A second architecture using the VGGish network pre-trained that we will re-train on our data.
- A third architecture combining VGGish pre-trained and a dense layer to match our data.

3. EXPERIMENTS AND RESULTS

In this section, the database and the performance metrics we have selected are presented. Then, the three architectures and their performances are described and then discussed.

3.1. Database composition

Public data were collected to train acoustic event recognition architectures. These are from the San Francisco National Park Association [24], which works on maritime history and refers to acoustic events collected during military naval missions of the WWII. Hence, these records constitute a great opportunity to train our models with public but representative data. Nevertheless, despite they are realistic these records are old ones, even if biological noises remain the same, some mechanical noises have slightly changed or are depreciated, so we have to carefully select only up to date representative noises.

Records were divided into 5-second intervals, set up after human observation process as the typical duration of the acoustic phenomena studied. This task is followed by labelling of each 5-second audio frame, manually annotated and checked by professional expertise. Here, a class is dynamically constituted when we have at least 5 examples from the same. In the context of this study, 3 classes belonging to the military field have been retained, Biological, submarine and vessels, respectively labelled as BIO, SM, and SS. The base has initially 478 signals, 80% signals are dedicated to the training set and 20% to the validation set, for a total duration of 39 minutes. Here, the test set is currently only composed of 12 signals including 6 SM, 3 BIO and 3 SS. The training dataset is summarized below:

Classes	BIO	SS	SM
Occurrences	134	158	174

Table 1: Repartition of training dataset

The table 1 allows us to see how classes are balanced, which is important to obtain good performance during learning step and fine-tuning step, avoiding overlearning.

3.2. Performance assessments

The following metrics were selected to estimate performances of the 3 architectures: train accuracy, validation accuracy, confusion matrix, ROC-curve and AUC [21].

Classification accuracy is the ratio number of correct predictions with the total number of input samples. It works well only if there is almost the same number of samples belonging to each class. A confusion matrix is a summary of the results of predictions about a classification problem. The matrix shows for each class the number of correct predictions on its diagonal entries and the number of incorrect predictions on its extra-diagonal entries. In practice, for each example each row of the table corresponds to the actual class, while each column corresponds to the expected class. Even if the confusion matrix is intuitive and so easy to interpret, to take into account the influence of misclassification costs we introduce as an extension of the confusion matrix the Receiver operating characteristic curve. The ROC curve is a probability curve showing the true positive rate in function of the false positive rate. Finally, in order to have a simple scalar metric, easy to interpret, cost-insensitive, summarizing the global performance level of the network, we integrate the ROC curve and we obtain the area under curve (AUC), which represents the measure of separability of one class versus all others. Higher the AUC, better the model is, and conversely. Because of its global cost-insensitive behavior, AUC allows to compare two different classifiers.

3.3. Performances of the three architectures

The train accuracy, the validation accuracy, confusion matrix and AUC-ROC curves are computed from the test set of the 3 architectures and illustrated in Table 2:

Architecture	Train_accuracy	Validation_accuracy
MFCCs and SVM	1.0	0.8696
VGGish	0.9380	0.9130
VGGish and one dense layer	**0.9903**	**0.9203**

Table 2: Accuracy of the train and validation dataset for the 3 architectures

These results allow first to ensure the network well learned and then it generalizes well its knowledges on the validation set. We simply verify overfitting of the networks, this is the case with the VGGish and VGGish and one dense layer architectures, which learned well about the training set and are able to generalize about the validation set. However, the MFCCs and SVM architecture seems to overfit to ensure this, it would be necessary in future work to use cross-validation and add regularization in the event of overfitting.

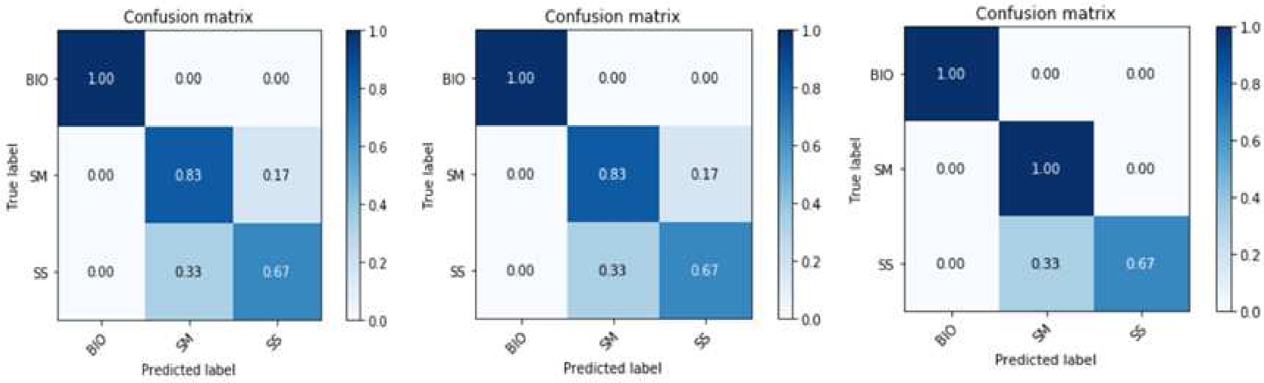

Figure 5: Confusion matrices of the 3 architectures: MFCCs + SVM (left), VGGish (center), VGGish + 1 Dense Layer (right)

Performances obtained on MFCCs+SVM and VGGish architecture are similar, which is quite disappointing in regard of the AI techniques state of the art. However, the Google network pre trained on millions YouTube videos is used; so it therefore seems regular that the network fully recognizes biological signals and has more difficulty to classify SM and SS signals. Nevertheless, the third architecture shows better performances to classify SM and SS by adding one dense layer and by reentering the network to fine-tune VGGish to adapt to the specificity of the input data.

The 3 architectures have more difficulty predicting the SS class which is often confused with the SM class. One possible explanation would be that SM are old and therefore noisy, so their signatures would look like those of an SS.

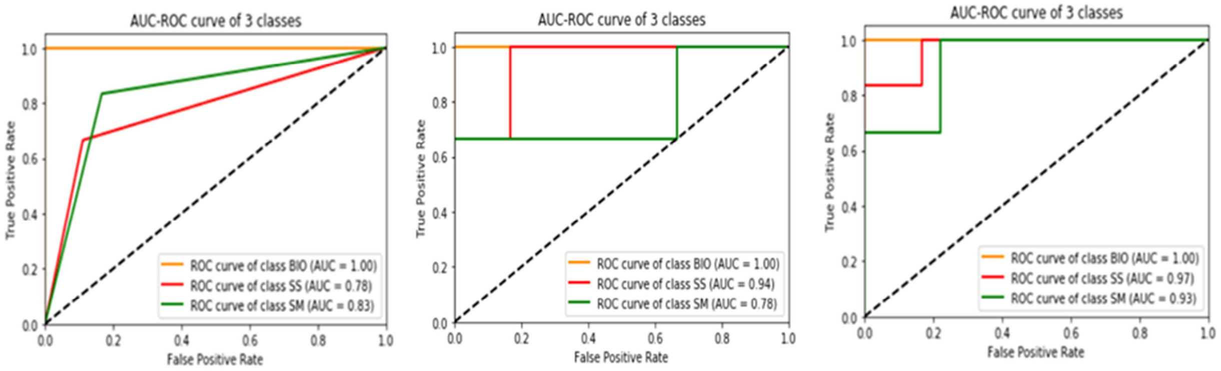

Figure 4 : AUC-ROC curves of the 3 architectures: MFCCs + SVM (left), VGGish (center), VGGish + 1 Dense Layer (right)

Classifiers that give curves closer to the top-left corner indicate a better performance. ROC-curve of class BIO is the same for the 3 architecture, but the ROC-curve of class SM is above that of class SS for the architecture MFCCs + SVM contrary to the 2 others. This may be explained by the fact that VGGish-based architectures use a network pre-trained on YouTube videos with relatively low submarines and surface vessels. We also notice that the performance of VGGish + one dense layer is better than VGGish alone, the fine-tuning of the network worked. The AUC values show that the best performance is obtained on the latest architecture, which is in accordance with the literature. The 3 architectures perform well, but VGGish + one dense layer architecture remains the best taking into account the different metrics.

4. CONCLUSION

Performances obtained via the three architectures are satisfying and it would seem that the last VGGish and one dense layer architecture is the most efficient. Nevertheless, experiments must be carried out with more data, trying to separate them among different classes, in order to take all the benefit deep neural networks which are not much more efficient contrary to the conclusions drawn from the bibliography.

However, the main goal which was the creation of a first database with operational classes, the selection of architectures and performance assessment is achieved, and we can think about future works. In order to see if better performances are possible through deep learning, we consider first the increase and the improvement of the database. Indeed, the current dataset is too old and is not representative anymore, especially for the SS; a new database should be created. Moreover, the recovery of public acoustic data characteristic of naval military field is complicated, that’s why an artificial increase in data is being considered. A first step is to create new signals by adding Wenz-type sea noise [19] combined with other techniques from signal processing. Moreover, in order to be as representative as possible of reality, we cannot be satisfied with only 3 classes and will increase them. The neural network architectures that have been used are very deep, it would then be interesting to encourage shallower networks to avoid possible overfitting and observe if performances are better.

REFERENCES

[1] Urakami M., Wakabayashi N. and Watanabe T. “A study on Location Information Screening Method for Ship Application Using AIS Recorded Data,”2018 International Conference on Broadband Communications for Next Generation Networks and Multimedia Applications (CoBCom), (2018).
[2] Li Y., Zhang Y. and Zhu F., “The method of detecting AIS isolated information based on clustering and distance, “2016 2nd IEEE International Conference on Computer and Communications (ICCC), (2016).
[3] Nishizaki C., Terayama M. Okazaki T., and R. Shoji, “Development of Navigation Support System to predict New Course of Ship,” 2018 World Automation Congress (WAC), (2018).
[4] Shahir H. Y., Glasser U., Shahir A. Y. and When H., “Maritime situation analysis framework: Vessel Interaction classification and anomaly detection,” 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC), (2017).

[5] Dawei Liang, Edison Thomaz, “Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale,” arXiv: 1810.08691v2 [cs.HC], (2018).

[6] Lukas Müller and Mario Marti,” Bird sound classification using a bidirectional LSTM,” CEUR-WS.org\vol-2125\paper_134.

[7] Nobuo Sato and Yasunari Obuchi,” Emotion recognition using Mel-Frequency Cepstral coefficients,” Journal of Natural Language Processing 14(4):835-84, (2007).

[8] Annamaria Mesaros, Toni Heittola, Antti Eronen, Tuomas Virtanen,”ACOUSTIC EVENT DETECTION IN REAL LIFE RECORDINGS,”18th European Signal Processing Conference (EUSIPCO-2010), (2010).

[9] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo, “CLEAR Evaluation of Acoustic Event Detection and Classification Systems,” R. Stiefelhagen and J. Garofolo (Eds.): CLEAR 2006, LNCS 4122, pp. 311 – 322, (2007).

[10] Xiaodan Zhuang, Xi Zhou, Mark A. Hasegawa-Johnson, Thomas S. Huang, “Real-world acoustic event detection,” Pattern Recognition Letters 31 1543–1551, (2010).

[11] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, Wojciech Samek,” Interpreting and explaining deep neural networks or classification on audio signals,” arXiv:1807.03418v1 [cs.SD], (2018).

[12] G. Keren, B. Schuller,” Convolutional RNN: an enhanced model for extraction features form sequential data,” arXiv: 1602.05875v3 [stat.ML], (2017).

[13] Satoshi IMAI,” CEPSTRAL ANALYSIS SYNTHESIS ON THE MEL FREQUENCY SCALE,” IEEE CH1841-6/83/000O0O93, (1983).

[14] Yixiong Pan, Peipei Shen and Liping Shen,“Speech emotion recognition using support vector machine,” International Journal of Smart Home, (2012)

[15] Christopher D.Manning, Prabhakar Raghavan and Hinrich Schütze,[Introduction to Information Retrieval], Online edition (c) Cambridge UP, (2009).

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Advances in neural information processing systems, pp. 1097–1105, (2012).

[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv: 1409.1556v6 [cs.CV], (2015).

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” arXiv: 1512.00567v3 [cs.CV], (2015).

[19] 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, Kevin Wilson, “CNN ARCHITECTURES FOR LARGE-SCALE AUDIO CLASSIFICATION,” arXiv:1609.09430v2 [cs.SD], (2017).

[20] Gordon M. Wenz, "Acoustic ambient noise in the ocean: Spectra and sources,” The Journal of the Acoustical Society of America 34, 1936, (1962).

[21] N. Japkowicz, M. Shah, [Evaluating Learning Algorithms: A classification Perspective], Cambridge University Press, 1 edition, (2014).

[22] Parwinder Pal Singh, Pushpa Rani, “An Approach to Extract using MFCC,” IOSR Journal of Engineering (IOSRJEN), ISSN (e): 2250-3021, ISSN (p): 2278-8719, Vol. 04, Issue 08, ||V1|| PP 21-25, (2014).

[23] Nilu Singh, R.A Khan, Raj Shree,” MFCC and Prosodic Feature Extraction Techniques: A Comparative Study,” International Journal of Computer Applications (0975 – 8887) Volume 54– No.1, (2012).

[24] https://maritime.org/sound/

[25] Sinno Jialin Pan; Qiang Yang, “A Survey on Transfer Learning,” IEEE Transactions on Knowledge and Data Engineering, Volume: 22, Issue: 10, Page s: 1345 – 1359, (2010).