A Deep Spiking Machine-Hearing System for the case of Invasive Fish Species

Konstantinos Demertzis, Lazaros Iliadis, Vardis-Dimitris Anezakis
Democritus University of Thrace,
Lab of Forest-Environmental Informatics and Computational Intelligence
Pandazidou st., 68200 N Orestiada, Greece
[kdemertz, liliadis, danezaki]@fmenr.duth.gr

***Abstract*—Prolonged and sustained warming of the sea, acidification of surface water and rising of sea levels, creates significant habitat losses, resulting in the proliferation and spread of invasive species which immigrate to foreign regions seeking colder climate conditions. This is happening either because their natural habitat does not satisfy the temperature range in which they can survive, or because they are just following their food. This has negative consequences not only for the environment and biodiversity but for the socioeconomic status of the areas and for the human health. This research aims in the development of an advanced Machine Hearing system towards the automated recognition of invasive fish species based on their sounds. The proposed system uses the Spiking Convolutional Neural Network algorithm which cooperates with Geo Location Based Services. It is capable to correctly classify the typical local fish inhabitants from the invasive ones.**

Keywords—Deep Learning, Spiking Convolutional Neural Networks, Machine Hearing, Bioacoustics, Underwater Acoustics, Invasive Species, Big Data

I. INTRODUCTION

The invasion of alien species is considered the second biggest threat to local biodiversity worldwide (after habitat destruction) and it has been named "biofouling". The impacts of invasive species (INSP) and their intrusion in the Mediterranean Sea are primarily economic, such as: risk of extinction of native species of economic value which implies costs to restore the natural balance, risk of introducing organizations harmful to the human health and potential reduction in regional tourism development. Additionally, they can have environmental dimension namely: disturbance of the food web by acting either as predators or as competitors, risk of introducing new diseases that can destroy sensitive native species, change in biodiversity and in the relative abundance of species [1].

All these effects are combined to result in the reduction of the number of native species in the Mediterranean and in their replacement by "opportunistic" ones, thus causing homogenization of the regional ecosystem [1], [2]. The INSP can be extremely toxic (e.g. the Lagocefalous family) containing a very dangerous substance, the tetrodotoxin, capable of causing serious health problems and even death to the potential consumer. Timely and accurate identification of the INSP by employing innovative and intelligent algorithms can be a potential partly solution to the problem [2]. The recognition of only phenotypic markers is an extremely difficult and uncertain process, as no major differences in morphology, nor significant similarities, reflect the affinity or not of organisms [3]. The identification by using genetic approaches such as the DNA Barcoding analysis [4] or comparisons of biochemical or molecular markers can be implemented relatively easily without requirements in high-tech devices and costs.

II. THEORETICAL BACKGROUND

A. Machine Hearing and Big Data

Machine Hearing is an Artificial Intelligence scientific field which attempts to algorithmically reproduce the sense of hearing. It is related to intelligent algorithms and to data analysis technologies accepting as input digitally recorded sounds and sensor data. The whole audio signal analysis process is related to knowledge elicitation by considering the content and the nature of the incoming signals. This process performs classification, segmentation, retrieval and synthesis. Initially, some basic features are determined whose values can be differentiated based on the content and the structure of the corresponding signals. After obtaining the audio features that describe the respective audio signal, a pattern recognition method is chosen to be employed depending on the application case.

According to the Nyquist theorem [5], an analog signal can be reproduced by the respective discrete, when the sampling frequency used is at least twice the maximum frequency of the

978-1-5090-5795-5/17/$31.00

original signal. Obviously high sampling frequencies, though they improve the quality of the digital signal, they can cause high computational complexities to the analysis algorithms. This is due to the fact that they produce much more samples per second which must be processed. The algorithmic resolution of a classification problem with Machine Hearing methods requires a high availability of resources. Time complexity of the algorithm and availability of memory (according to the size of input data) must be examined. Under this perspective, this problem is a case of big data modeling, as data extracted from audio clips require large storage space (especially if the sounds have high sampling frequencies). A proper solution must employ a deterministic automaton in polynomial time. Dynamic assignment of computational resources and coordination of complex data analysis procedures are required.

B. Literature Review

Recently, a new data modeling direction has emerged, known as deep learning [6]. Through a series of new learning architectures and algorithms, domains such as object recognition [7] and machine translation [8], [9] have been transformed; deep learning methods are now the state-of-the-art in many of these areas. In particular, deep learning has been the driving force behind large leaps in accuracy and model robustness in audio related domains like speech recognition [10], [11]. Also, spiking deep neural networks have become an increasingly active field of research. This has been driven both by the interest to build more biologically realistic ANN models and by the availability of larger-scale neuromorphic computing platforms which are optimized for emulating brain-like spike-based computation in dedicated analog or digital hardware [12], [13], [14], [15].

Also, there have been many studies on automatic audio classification and segmentation using several features and techniques. The most effective models of audio classification are related to speech/music cases. High levels of accuracy have been achieved, especially when the segmentation information is known beforehand. In [16], wavelets are first applied to extract acoustical features such as sub-band power and pitch information. The method uses a bottom-up Support Vector Machine (SVM) over these acoustic features and additional parameters, such as frequency cepstral coefficients, to accomplish audio classification and categorization.

A new approach towards high performance speech/music discrimination on realistic tasks related to the automatic transcription of broadcast news is described in [17], employing a hybrid ANN-Hidden Markov Model. In [18], a generic audio classification and segmentation approach for multimedia indexing and retrieval is described. In [19], a method is proposed for speech/music dis-crimination based on root mean square and zero-crossings. The approach of [20] investigates the feasibility of an audio-based context recognition system, where simplistic low-dimensional feature vectors are evaluated against more standard spectral ones. Using discriminative training, competitive recognition accuracies are achieved with very low-order hidden Markov models.

III. DATA

The Sea_Audio_Dataset includes sounds gathered by the University of Rhode Island's Graduate School of Oceanography. They are available at http://www.dosits.org/ [21] by the Ocean Conservation Research center. This is a non-profit research and policy development organization focusing in finding solutions to the growing threat of ocean noise pollution and its impact on marine habitat http://ocr.org/ [22].

Other datasets were obtained by the Macaulay Library of Cornell Lab of Ornithology - Cornell University http://macaulaylibrary.org/ [23]. They are the most comprehensive and authoritative Bioacoustics and hydro acoustics repositories globally. Bioacoustics is an interdisciplinary scientific branch that combines biology with acoustics, aiming in studying sound generation and propagation through elastic means and also in analyzing its perception by animals. It also deals with the objective electrophysiological measurements, carried out on animals in order to study the organ of hearing and especially bioelectric potentials.

A. Sound in the Sea

The following four basic types of sounds were included in our dataset in order to obtain high complexity scenarios including the sounds that are more likely to be found under the sea. These datasets will be used to train Spiking Convolutional Neural Networks (SCNN) [5] developed herein, capable to generalize.

- Anthropogenic (Anthr) Sounds: This category includes 684 sounds that belong to 9 classes namely: Ship, Sonar, Zodiac, Torpedo, Wind Turbine, Scuba Noise, Bubble Curtain, Personal Water Craft, Airgun.
- Natural Sounds: Totally 477 sounds belong to this category and they are classified in the following 6 classes: Earthquake, Hydrothermal Vents, Ice Cracking, Rainfall, Lightning, Waves.
- Fishes: Many fish species produce sounds with various parts of their body, e.g. with their teeth, pharynx, fins and swim bladder. This group contains 1076 sounds classified as follows: Bidyanus Bidyanus, Epinephelus Adscensionis, Cynoscion Regalis, Carassius Auratus, Cyprinus Carpio, Rutilus Rutilus, Salmo Trutta, Oreochromis Mossambicus, Micropterus Salmoides, Oncorhynchus Mykiss.
- Mammals: Sea mammals are producing and using sounds in order to orientate and to communicate. Totally 836 sounds are used in this group classified in the following 8 classes of mammals. (Delphinus Delphis, Erignathus Barbatus, Balaena Mysticetus, Phocoena Phocoena, Neophocaena Phocaenoides, Trichechus, Tursiops Truncates, Phoca Hispida).

B. Audio Feature Extraction

The main role of the feature extraction process is to properly reflect the characteristics that identify exclusively the uniqueness of each sound and to distinguish between the acoustic categories. The distinction between categories is based on 34 characteristics mainly related to statistical measures received from the information of the signal frequencies. The feature extraction methodology is followed two specific stages [24]:

- Short-term feature extraction: It divides the input signal in short-term windows (frames). In this way a number of features are calculated for each frame, driving and so discovering, the sequence of short-term feature vectors for the entire signal.
- Mid-term feature extraction: In many cases, the signal is represented by the relevant statistics of the features exported by the Short-term feature extraction operation described above. That is why the Mid-term feature extraction process uses a series of statistics (e.g. mean and standard deviation) over a short-term feature sequence.

To extract the short term feature sequences for an audio signal, a frame size of 50 msecs and a frame step of 25 msec (50% overlap) was used. All of the sounds had a sampling ratio equal to 44.1 kHz with a 16-bit stereo analysis, whereas their average duration was 10.3 seconds.

IV. The Proposed System

In recent years, the use of Deep Learning techniques and especially Convolutional Neural Networks (CNN) has achieved great progress in solving problems related to robotics and particularly in recognizing visual patterns. They were developed and applied for many cases, like speech recognition, multimodal learning. They are characterized by their ability to handle more powerful computing hardware, larger datasets and improved training algorithms. [25].

The CNN design was inspired by the visual cortex of the cat that has a complicated layout and an interface of optical cells that are sensitive to specific subregions of the visible field, functioning as local 'filters'.

The aim of the CNNs [26] is to model a high degree of abstraction in imported samples, by using architectures based on multiple levels of nonlinear transformations. Their logic is based on the replacement of pre-designed characteristics (filter extraction mechanisms) with effective learning algorithms.

The differentiation of the CNNs compared to the Feed Forward ones is in the proceedings neurons' levels which are considered as filters for specific subsets of the data under study.

A. Spiking Convolutional Neural Network (SCNN)

The proposed method of converting a CNN to Spiking Neural Network (SNN) overcomes the greatest problem which is the unexpected fall of the classification accuracy (accuracy loss) when converting, for reasons explained in detail in [27]. This conversion can be achieved if the following assumptions are made:

Firstly, the Rectified Linear Units (ReLU) can be considered at firing rate approximation of an Integrate-and-Fire neuron with no refractory period [27], whereby the output of the ReLU is proportional to the number of spikes produced by an IF neuron within a given time window. ReLUs are also advantageous during training as their piecewise constant derivative leads to weight updates of a particularly simple form. Secondly, for classification tasks, only the maximum activation of all units in the output layer is of importance, allowing the overall rate to be scaled by a constant factor. Finally, without a bias to provide an external reference value, the relative scale of the neuron weights to each other and to the threshold of the neuron are the only parameters that matter. The whole process works as follows:

1) Rectified Linear Units (ReLUs) are used for all units of the network. ReLUs are a type of nonlinearity which is applied to the weighted sum of inputs, and is described by the following equation 1 [28].

$$x_i = max\left(0, \Sigma_i w_{ij} x_j\right) \quad (1)$$

where x_i is the activation of unit i, w_{ij} is the weight connecting unit j in the preceding layer to unit i in the current layer, and x_j is the activation of unit j in the preceding layer. By successively updating all the activations of a current layer based on the activations of the previous layer, the input is propagated through the network to activate the output label neurons. Training proceeds according to standard error backpropagation, successively propagating an error gradient backwards through the layers by computing local derivatives to update individual weights and minimize the error. In these networks, the training process adjusts the randomly-initialized weight matrix describing the connections between the layers to minimize the overall error through stochastic gradient descent [28].

2) Fix the bias to zero throughout training, and train with backpropagation.

3) Directly map the weights from the ReLU network to a network of Integrate-and Fire (IF) units. The evolution of the membrane voltage υ_{mem} is given by equation 2

$$\frac{d\upsilon_{mem}(t)}{dt} = \sum_i \sum_{s \in S_i} w_i \delta(t - s) \quad (2)$$

where w_i is the weight of the ith incoming synapse, $\delta(\cdot)$ is the delta function, and $S_i = \{t_i^0, t_i^1, \dots\}$ contains the spike-times of the ith presynaptic neuron. If the membrane voltage crosses the spiking threshold υ_{thr}, a spike is generated and the membrane voltage is reset to a reset potential υ_{res}. In our simulations, this continuous-time description of the IF model is discretized into 1 ms timesteps.

4) Use weight normalization to obtain near-loss accuracy and faster convergence. The safest, most conservative method to normalize the network weights, which ensure that activations are sufficiently small to prevent the ReLU from overestimating output activations, is to consider all possible positive activations that could occur as input to a layer, and rescale all the weights by that maximum possible positive

input. If the maximum positive input can only cause one spike, then the network will never need to produce more than one spike at once from the same neuron. By doing so, the resulting spiking networks become robust to arbitrarily high input rates and completely eliminate losses due to too many inputs. Unfortunately, this means that evidence integration (in order to generate a spike) might require much more time.

The architecture of the SCNN is a 34x34-12c5-2s-68c5-2s-4o Convolutional Network. The input audio is 34x34, followed by 12 convolutional kernels of size 5x5, followed by a 2x2 averaging subsampling window. This convolution process is repeated in a second stage with 68 maps of size 5x5, followed by a 2x2 averaging of the network. These final features are vectorized and fully connected to a 4-node output layer, where each of the 4 nodes represents one of the 4 audio classes. The training process used a fixed learning rate of 1, a batch size of 50, no momentum, 50% dropout of the kernels, zero bias, and 50 epochs of training. The intensity values of the Sea_Audio_Dataset were normalized to values between 0 and 1. Based on those intensity values, Poisson distributed spike trains were generated for each audio with firing rates proportional to the feature's intensity value [27].

B. GeoLocation Country Based Services

A GPS device provides the user with specific location coordinates. The accurate determination of the country where these coordinates belong is achieved by employing a process that considers the global borders of the states as they appear in the shapefile that can be found in the following address: http://thematicmapping.org/downloads/world_borders.php

The following Python code is used to import the above shapefile and to check the coordinates 39.35230, 24.41232 that belong to Greece [29]:

```
import countries
  cc = countries.CountryChecker('TM_WORLD_BORDERS-0.3.shp')
  print cc.getCountry(countries.Point(39.35230, 24.41232)).iso
print Greece
```

V. Description of The Proposed Method

The Audio Feature Extraction (AFE) process is the first stage of the proposed system's algorithm. It is used to extract the proper characteristics of every sound included in the Sea_Audio_Dataset. In the second phase these features are input to the optimal SCNN that performs the classification. This is done to determine if the sound spotted by the Machine Hearing system can be assigned to an invasive species fish or not. If the sound is characterized as noise coming from human sea activity, then it is rejected and the process is terminated. If it is related to some kind of fish or mammal, then the system performs identification of the species with pattern recognition. Its coordinates are obtained by the GPS and they are correlated to the country of origin. Then it is checked if it is native species in this country and if not it is recorded as an invasive one. The lists with the native and the invasive species are extracted by the Invasive Species Compendium (http://www.cabi.org/isc/) [30] the most comprehensive database on the subject globally. The algorithm for the species identification is presented below:

Input:
Recognized_Species;
Country;
Country_Native_Species;
1: Read Recognized_Species, Country, Country_Native_Species;
2: **for** i=1 to Country_Native_Species [max] **do**
3: **if** Country_Native_Species [i]= Recognized_Species **then**
4: Recognized Species=Native_Species
5: **else**
6: Recognized Species=Invasive_Species
7 **end if**
8: **end**
Output:
Species Identity;

VI. Results And Comparative Analysis

Given that the dataset employed presents high complexity and considering the specificities resulting from the process of the audio feature extraction, it is promising that the system managed to solve a particularly complex realistic engineering hearing problem, with high accuracy. The classification accuracy for all the examined cases is quite satisfactory with high "Precision" and "Recall" indices values something that shows a good generalization ability. The testing was related to classification of Machine Hearing sounds included in the Sea_Audio_Dataset that consisted of Anthropogenic_Sounds, Natural_Sounds, Fishes and Mammals. The evaluation indices for this case were very high and the differentiations between the sounds were traced by the system.

The overall classification accuracy is presented in table 1, whereas Table 2 presents the Confusion Matrix (CM) where the main diagonal values (top left corner to bottom right) correspond to correct classifications and the rest of the numbers correspond to very few cases that were misclassified. The numbers of misclassifications are related to the False Positive (FP) and False Negative (FN) indices appearing in the confusion Matrix of Table 2. A FP is the number of cases where you wrongfully receive a positive result and the FN is exactly the opposite. On the other hand the True Positive (TP) is the number of records where you correctly receive a Positive result. The True Negative (TN) is defined respectively. The True Positive rate (TPR) also known as Sensitivity, the True Negative rate also known as Specificity (TNR) and the Total Accuracy (TA) are defined by using equations 3, 4, 5 respectively.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \ (3) \quad \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \ (4) \quad \mathrm{TA} = \frac{\mathrm{TP+TN}}{\mathrm{N}} \ (5)$$

The Precision (PRE) the Recall (REC) and the F-Score indices are defined as in equations 6, 7 and 8 respectively.

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \ (6) \quad \mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \ (7)$$

$$\mathrm{F-Score} = 2\mathrm{X}\frac{\mathrm{PRE\ X\ REC}}{\mathrm{PRE\ +\ REC}} \ (8)$$

Finally the ROC in Table 1 is an abbreviation for the Receiver Operating Characteristic, a standard technique for

summarizing classifier performance over a range of trade-offs between TP and FP error rates. ROC curve is a plot of *Sensitivity (the ability of the model to predict an event correctly)* versus *1-Specificity* for the possible cut-off classification probability values π_0 [31]. In our case, the classification success rate was a little lower but still very high and satisfactory considering the complexity of the classification scenarios. The system had very high accuracy for the recognition of mammals (detailed results and the respective CM in Tables 3 and 4). The most difficult classification case was the one of the fish (the evaluation results with the CM appear in Tables 5 and 6). The 10-fold cross validation (10-Fold CV) approach was employed for every case to improve generalization.

TABLE I. PERFORMANCE METRICS OF CATEGORIES_DATASET

Classifier	**Classification Accuracy (ACC) & Performance Metrics**					
	ACC	RMSE	Precision	Recall	F-Score	ROC
SCNN	**96.25%**	**0.1368**	**0.963**	**0.963**	**0.963**	**0.975**

TABLE II. CONFUSION MATRIX OF CATEGORIES_DATASET

Fishes	Mammals	Anthr_Sounds	Natural_Sounds	
1045	13	11	7	Fishes
15	**797**	16	8	Mammals
5	7	**659**	13	Anthr_Sounds
6	5	9	**457**	Natural_Sounds
TPR				
0.971	0.969	0.963	0.958	
TNR				
0.975	0.953	0.948	0.942	

TABLE III. PERFORMANCE METRICS OF MAMMALS_DATASET

Classifier	**Classification Accuracy (ACC) & Performance Metrics**					
	ACC	RMSE	Precision	Recall	F-Score	ROC
SCNN	**92.10%**	**0.1577**	**0.921**	**0.921**	**0.921**	**0.952**

TABLE IV. CONFUSION MATRIX OF MAMMALS_DATASET

a	b	c	d	e	f	g	h	
142	2	0	1	1	0	1	0	a = Delphinus Delphis
1	**101**	3	0	0	5	0	4	b = Erignathus Barbatus
1	2	**121**	2	0	0	0	2	c = Balaena Mysticetus
1	1	2	**61**	1	1	0	3	d = Phocoena Phocoena
2	2	3	1	**51**	0	2	2	e = Neophocaena Phocaenoides
3	0	2	0	1	**81**	0	2	f = Trichechus
2	0	0	0	0	1	**51**	1	g = Tursiops Truncates
2	2	1	1	1	0	1	**162**	h = Phoca Hispida
TPR								
0.965	0.885	0.945	0.871	0.809	0.91	0.927	0.952	
TNR								
0.922	0.918	0.916	0.924	0.927	0.92	0.927	0.924	

TABLE V. PERFORMANCE METRICS OF FISHES_DATASET

Classifier	**Classification Accuracy (ACC) & Performance Metrics**					
	ACC	RMSE	Precision	Recall	F-Score	ROC
SCNN	**89.03%**	**0.1481**	**0.892**	**0.890**	**0.890**	**0.939**

TABLE VI. CONFUSION MATRIX OF FISHES_DATASET

a	b	c	d	e	f	g	h	i	j
141	5	0	2	0	0	1	0	1	3
4	**92**	4	0	3	0	0	2	4	0
0	3	**117**	0	2	0	2	4	2	2
1	0	0	**89**	1	0	0	1	0	1
0	3	1	0	**101**	0	0	2	1	2
1	0	0	0	0	**59**	0	2	0	3
1	0	1	0	0	0	**95**	1	0	0
3	5	6	1	3	1	0	**104**	1	8
0	3	2	0	3	0	0	0	**81**	3
1	0	3	1	3	1	0	3	1	**79**
0.921	0.844	0.823	0.956	0.918	0.907	0.969	0.787	0.880	0.858
0.927	0.760	0.873	0.956	0.870	0.967	0.969	0.873	0.890	0.782

a = Bid/nus Bidyanus, b = Epin/lus Adscen/nis, c = Cyn/cion Regalis, d = Cara/us Auratus
e = Cyp/nus Carpio, f = Rutilus Rutilus, g = SalmoTru, h = Oreo/mis Mossambicus
i = Micr/rus Salmoides, j = Oncor/hus Mykiss

VII. DISCUSSION-CONCLUSIONS

The advanced computing hearing application described herein, in conjunction with the promising results obtained, offers a reliable and innovative proposal towards the formulation and design of biosecurity and biodiversity protection methods.

The simplification of the detection-identification processes with a Machine Hearing intelligent model, considers the exact geographical position of the species in a cost effective manner. It also creates the conditions for studying the behavior of different species and the seasonal fluctuation of their populations. Finally, the accurate mapping of the species intrusion process is achieved. This effort can significantly slow the uncontrolled expansion of the INSP.

Based on the comparative analysis, the Spiking Convolutional Neural Network algorithm has proven to be effective with high generalization capacity. It combines the effectiveness and the speed of the Spiking ANN with the advantages of Deep Learning. An advanced audio feature extraction approach was used in order to achieve efficient, realistic and authoritative analysis and formulation of the identification data. An important advantage of the system is that except from the input features it also considers the geographical position of each species and it performs comparison to the local ones.

A future research direction that could be conducive to the proposed system is related to the choice of the optimization method, searching for the optimal parameters, plus the potential automatic calculation and consideration of the degree of significance of each independent variable.

The implementation of the system with the employment of online learning algorithms and with similar characteristics choice methods in the audio feature extraction process (e.g. Representational Similarity Analysis, Isoperimetry and Gaussian Analysis, Local Similarity Analysis) would lead to a potential improved performance. Using Deep Learning combined with Single Layer Spiking ANN (e.g. Deep Spiking Extreme Learning Machines) could offer a significant innovation.

REFERENCES

[1] A. Abdulla and O. Linden, "Maritime traffic effects on biodiversity in the Mediterranean Sea: Review of impacts, priority areas and mitigation

measures," Malaga, Spain: IUCN Centre for Mediterranean Cooperation. pp.184, 2008. ISBN: 978-2-8317-1079-2.

[2] F. Rahel and J. D. Olden, "Assessing the Effects of Climate Change on Aquatic Invasive Species," Conservation Biology, vol. 22, no.3, pp.521–533, 2008. doi:10.1111/j.1523-1739.2008.00950.x

[3] W. Miller, "The structure of species, outcomes of speciation and the species problem: Ideas for paleobiology, Palaeoclimatology Palaeoecology," vol. 176, no. 1-4 pp.1–10 2001. doi:10.1016/S0031-0182(01)00346-7

[4] K. Demertzis, L. Iliadis L, "Intelligent Bio-Inspired Detection of Food Borne Pathogen by DNA Barcodes: The case of Invasive Fish Species Lagocephalus Sceleratus," L. Iliadis, C. Jayne, Eds. Engineering Applications of Neural Networks. Communications in Computer and Information Science (CCIS), LNCS, vol.517, Springer, Cham 2015, pp.89-99, 2015. doi:10.1007/978-3-319-23983-5_9.

[5] A. Bruno Olshausen, "Aliasing, PSC 129 - Sensory Processes," 2000. http://redwood.berkeley.edu/bruno/npb261/aliasing.pdf

[6] L. Deng and D. Yu, "Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing," vol.7, no. 3–4, pp.197–387, 2014. http://dx.doi.org/10.1561/2000000039

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger, Eds. Proceedings of the 25th International Conference on Neural Information Processing Systems NIPS '12, Curran Associates, Inc., USA, pp.1097–1105, 2012.

[8] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," W. William, C. Mccallum, A. Mccallum, T.R. Sam, Proceedings of the 25th International Conference on Machine Learning ICML '08, ACM, New York, NY, USA, Helsinki, Finland, pp.160–167, 2008.

[9] T. Deselaers, S. Hasan, O. Bender, H. Ney. "A Deep Learning Approach to Machine Transliteration," Proceedings of the Fourth Workshop on Statistical Machine Translation StatMT '09, Association for Computational Linguistics, Stroudsburg, PA, USA, Athens, Greece, pp.233–241, 2009.

[10] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury. "Deep Neural Networks for Acoustic Modeling in Speech Recognition," IEEE Signal Processing Magazine, vol.29, no.6, pp. 82–97, 2012. doi:10.1109/MSP.2012.2205597

[11] N.D. Lanez, P. Georgievy, L. Qendro, "DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments using Deep Learning," Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ACM,New York,NY,USA, Osaka, Japan pp.283-294, 2015. http://dx.doi.org/10.1145/2750858.2804262

[12] P.A. Merolla, J.V. Arthur, R. Alvarez-Icaza, A.S. Cassidy, J. Sawada, F. Akopyan, B.L. Jackson, N. Imam, C. Guo, Y. Nakamura, B.Brezzo, I. Vo, S.K. Esser, R. Appuswamy, B. Taba, A. Amir, M.D. Flickner, W.P. Risk, R. Manohar, D.S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," Science, vol. 345, no. 6197, pp. 668–673, 2014. doi:10.1126/science.1254642

[13] B.V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A.R. Chandrasekaran, J.M Bussat, R.Alvarez-Icaza, J.V Arthur, P.A. Merolla, K. Boahen, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," Proceedings of the IEEE, vol.102, no.5, pp. 699–716, 2014. doi: 10.1109/JPROC.2014.2313565

[14] S. Furber, F. Galluppi, S. Temple, L. Plana, "The SpiNNaker Project," Proceedings of the IEEE, vol.102, no.5, pp. 652–665, 2014. doi: 10.1109/JPROC.2014.2304638

[15] D. Neil and S.C. Liu, "Minitaur, an event-driven fpga-based spiking network accelerator," IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol.22, no.12, pp. 2621–2628, 2014. doi: 10.1109/TVLSI.2013.2294916

[16] C.C. Lin, S.H. Chen, T.K. Truong, Y. Chang, "Audio classification and categorization based on wavelets and support vector machine," IEEE Transactions on Speech and Audio Processing, vol.13 no.5 pp. 644–651 2005. doi: 10.1109/TSA.2005.851880

[17] J. Ajmera, I. McCowan, H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a HMM classification framework," Speech Communication, vol.40 no.3, pp.351–363, 2003. doi: 10.1016/S0167-6393(02)00087-0

[18] S. Kiranyaz, A.F. Qureshi, M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," IEEE Transactions on Audio, Speech and Language Processing, vol.14, no.3, pp.1062–1081 2006. doi: 10.1109/TSA.2005.857573

[19] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," IEEE Transactions on Multimedia, vol.7 no.1, pp.155–156. doi:10.1109/TMM.2004.840604

[20] A.J. Eronen, V.T. Peltonen, J.T. Tuomi, A.P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, J. Huopaniemi, "Audio-based context recognition," IEEE Transactions on Audio, Speech and Language vol.14, no.1, pp.321-329, 2006. doi:10.1109/TSA.2005.854103

[21] http://www.dosits.org/

[22] http://ocr.org/

[23] http://macaulaylibrary.org/

[24] T. Giannakopoulos, "pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis," PLoS ONE. vol. 10, no.12, 2015. doi:10.1371/journal.pone.0144610

[25] J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, vol. 61, pp. 85–117, 2015. doi: http://doi.org/10.1016/j.neunet.2014.09.003

[26] Y. Cao, Y. Chen, D. Khosla, "Spiking deep convolutional neural networks for energy-efficient object recognition," International Journal of Computer Vision, vol.113, no.1, pp. 54-66, 2014. doi: 10.1007/s11263-014-0788-3

[27] P.U. Diehl, D. Neil, J. Binas, M. Cook, S.C. Liu, M. Pfeiffer, "Fast-Classifying, High-Accuracy Spiking Deep Networks Through Weight and Threshold Balancing," IEEE International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, pp.1-8, 2015. doi: 10.1109/IJCNN.2015.7280696

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhut-dinov, "Dropout: A simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.

[29] https://github.com/che0/countries

[30] http://www.cabi.org/isc

[31] T. Fawcett, "An introduction to ROC analysis. Pattern Recognition Letters," Elsevier Science, vol 27, no.8, pp. 861-874, 2006. doi: http://doi.org/10.1016/j.patrec.2005.10.010