International Journal of Embedded and Real-Time Communication Systems
Volume 11 • Issue 1 • January-March 2020

A Concept Drift-Aware DAG-Based Classification Scheme for Acoustic Monitoring of Farms

Stavros Ntalampiras, Università degli studi di Milano, Milan, Italy

Ilyas Potamitis, Technological Educational Institute of Crete, Heraklion, Greece

ABSTRACT

Intelligent farming as part of the green revolution is advancing the world of agriculture in such a way that farms become dynamic, with the overall scope being the optimization of animal production in an eco-friendly way. In this direction, this study proposes exploiting the acoustic modality for farm monitoring. Such information could be used in a stand-alone or complimentary mode to monitor the farm constantly at a great level of detail. To this end, the authors designed a scheme classifying the vocalizations produced by farm animals. More precisely, a directed acyclic graph was proposed, where each node carries out a binary classification task using hidden Markov models. The topological ordering follows a criterion derived from the Kullback-Leibler divergence. In addition, a transfer learning-based module for handling concept drifts was proposed. During the experimental phase, the authors employed a publicly available dataset including vocalizations of seven animals typically encountered in farms, where promising recognition rates were reported.

KEYWORDS

Acoustic Farm Monitoring, Directed Acyclic Graph, Echo State Network, Hidden Markov Model, Mel-Frequency Cepstral Coefficients, Smart Farming

INTRODUCTION

The area of Computational Bioacoustic Scene Analysis has received increasing attention by the scientific community in the last decades (Stowell, 2018; Blumstein et al., 2011; Towsey, Truskinger, & Roe, 2015; Dong, Towsey, Zhang, & Roe, 2015; Li, Zhou, Zou, & Li, 2012). Such interest is motivated by the potential benefits that can be acquired towards addressing major environmental challenges including invasive species, infectious diseases, climate and land-use change, etc. Availability of accurate information regarding range, population size and trends is crucial for quantifying the conservation status of the species of interest. Such information can be obtained via classical observer-based survey techniques; however, these are becoming inadequate since they are a) expensive, b) subject to weather conditions, c) cover a limited amount of time and space, etc. To this end, autonomous recording units (ARUs) are extensively employed by biologists (Grill & Schlter, 2017; Ntalampiras, 2018a). This is also motivated by the cost of the involved acoustic sensors which is constantly decreasing due to the advancements in the field of electronics.

One of the first approaches employed for classifying animal vocalizations is described in (Mitrovic, Zeppelzauer, & Breiteneder, 2006). The authors extracted Linear predictive coding coefficients, cepstral coefficients based on the Mel and Bark scale, along with time-domain features describing the

DOI: 10.4018/IJERTCS.2020010104

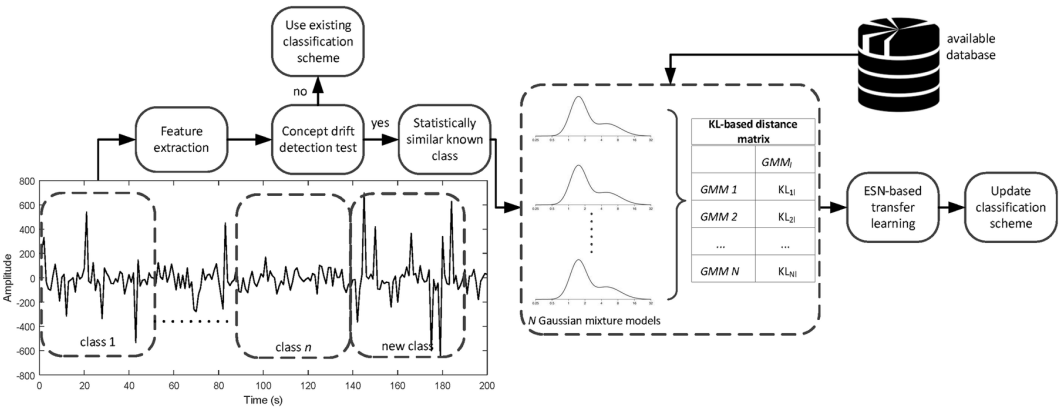
peaks and silence parts of the waveform. The classifier was a Support Vector Machine, while three kernels were considered, i.e. polynomial, radial basis function, and sigmoid. These were compared with nearest neighbor and linear vector quantization schemes. The specific dataset included sounds of four animal classes, i.e. birds, cats, cows, and dogs. The literature further includes several approaches which concentrate on specific species, classification of Australian anurans (Han, Muniandy, & Dayou, 2011), interpretation of chicken embryo sounds (Exadaktylos, Silva, & Berckmans, 2014), classification of insects (Noda, Travieso, Snchez-Rodrguez, Dutta, & Singh, 2016), etc. However, a systematic approach addressing the specific case of farm monitoring, is not present in the literature. This work wishes to cover exactly this gap (Figure 1).

Indeed, the acoustic modality could provide complementary information to monitor the health as well as population of animals. For example, it could be used in combination with solutions such as (Kumar & Hancke, 2015; Nagpal & Manojkumar, 2016; Anu, Deepika, & Gladance, 2015) which record physiological parameters of the animals, such as rumination, body temperature, and heart rate with surrounding temperature and humidity. The valuable information that can be obtained via the acoustic modality could assist an overall assessment of the current status of the animals as well as the farm in general. More precisely, acoustic farm environment monitoring could assist in the following applications:

- Tracking of similar breed animals and parturitions
- Identification of specific animal(s) for several reasons (vaccination, medication, diseases, diet, etc.)
- Animal health monitoring
- Population monitoring
- Detect animals missing from the farm
- Intruder detection and identification

Of course, this is a non-exhaustive list of the potential applications, while the overall aim is to optimize animal production in an eco-friendly way. The specific area is an emerging new topic comprising an intersection of several disciplines starting from fundamental biology all the way to the current trends in computer science including internet of things, signal processing over networks and advanced fault diagnosis methods. Towards an integrated solution, the operation of each of these components should be recognized by the rest as it directly influences the stability and efficacy of the overall system.

Figure 1. The logical flow of the proposed method encompassing a) signal windowing, b) feature extraction, c) concept drift detection, d) statistical affinity calculation, e) ESN-based transfer learning, and f) update of the classification scheme



This works is an extension of (Ntalampiras, 2018b) while the aim is to construct a comprehensive classification scheme, the operation of which does not follow the black-box logic, i.e. where one is able to 'open' the classifier, and by inspecting the misclassifications, obtain clear insights on how its performance can be boosted. At the same time, the proposed system is designed keeping in mind that it may have to operate under non-stationary conditions (Ditzler, Roveri, Alippi, & Polikar, 2015; Dargie, 2009), where distributions followed by the known classes may evolve over time (e.g. due to noise, reverberation effects, etc.), new classes may appear (e.g. new species), etc. Such obstacles require a scheme able to incorporate changes during its operation and address the evolving phenomena by appropriately altering its structure.

Keeping these in mind, we employed a well-known feature set (Dargie, 2009) in combination with a classification scheme adopting a directed acyclic graph structure. There, the topological ordering problem is addressed by means of an approach based on the Kullback-Leibler divergence measured among the different sound classes. During the experimentations, we used part of the dataset called Environmental Sound Classification-10 described in (Piczak, 2015b) which includes the animals typically encountered in a farm environment, i.e. dog, rooster, pig, cow, cat, hen, and sheep. There, a preliminary classification analysis on the entire dataset provided a recognition rate of 72.7%, while a more recent effort (Piczak, 2015a) based on convolutional neural networks achieved approximately 80%. Moreover, we performed a comparison with other classification schemes (echo state network, class-specific and universal hidden Markov models, support vector machines, and random forest) and feature sets (MPEG-7 audio standard and perceptual wavelet packets), a process which demonstrated the superiority of the proposed approach.

Importantly, the directed acyclic graph is accompanied by the framework responsible for handling concept drifts, i.e. the appearance of novel audio data emitted from sources not existing in the available database. To this end, we employ the existing HMM-based concept drift detection test described in (Ntalampiras, 2016) complemented by a data augmentation module based on transfer learning. We argue that the main problem in addressing concept drifts is the unavailability of data coming from the new source. Towards addressing this point, we propose to find statistically similar data in the available corpus and transform them to represent the new source. The specific module was evaluated by keeping out of the available set, data belonging to a class representing the novel one. This procedure was carried out for all available classes in a rotational manner.

The rest of this article is organized as follows: section 2 formulates the problem, while section 3 details the proposed sound classification framework including the formalization of the DAG-HMM and its topological ordering. Section 4 provides information on the dataset we employed, the contrasted approaches, and presents and analyzes the experimental results. Finally, section 5 concludes this work.

PROBLEM FORMULATION

In this paper we suppose a single channel audio datastream, y_t the duration of which is unknown. y may be emitted by various sources which are known only to an extent, i.e. $C=\{C_1, \ldots, C_m\}$, where m is the number of known sources. It is further assumed that each source follows a consistent, yet unknown probability density function P_i in stationary conditions, while at a specific time instance one sound source dominates (operating for example after a source separation framework e.g. (Gao, Woo, & Dlay, 2011)).

However, in the concept drift environment several obstacles might be encountered, e.g. change of the recording conditions, reverberation, appearance of sound events produced by sources which are not a-priori known, y_t might be corrupted by non-stationary noise, alterations in the realization of known sound events, etc. Thus, y_t becomes y_t' at time t^*, where t^* is the starting time instance of the concept drift. Such obstacles change the data generation process P_i, thus either a new classification method should be designed or the already constructed system should be adapted.

We assume that an initial training sequence $TS = y_t$, $t \in [1, T_0]$ is available characterized by stationary conditions and containing supervised pairs (y_t, C_i), where t is the time instant and $i \in [1, m]$. No assumptions are made with respect to the way the probability density functions (pdf) P_i's might alter for $t > T_0$.

THE PROPOSED SOUND CLASSIFICATION FRAMEWORK

The proposed framework relies on the Directed Acyclic Graph logic (Ntalampiras, 2014), i.e. the classification scheme is a graph denoted as $G = \{N, L\}$, where $N = \{n_1, \ldots, n_m\}$ represents the nodes and $L = \{l_1, \ldots, l_k\}$ the links associating the nodes. Each node in N is responsible for a binary classification task conducted via a set of hidden Markov models (HMM) which fit well the specifications of audio pattern recognition tasks, thus the DAG-HMM notation.

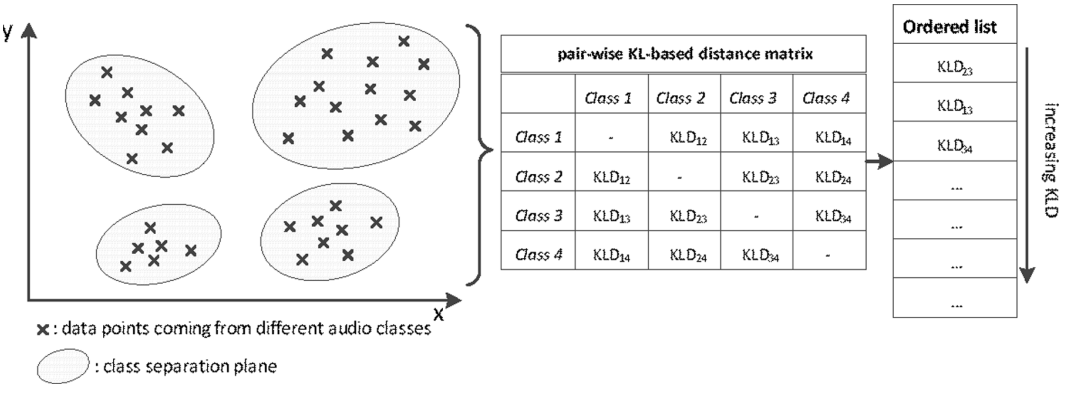
The motivation behind creating such a graph-based classification system is that in this way, one is able to limit the problem space and design classification algorithms for two mutually exclusive classes than having to deal with the entirety of the different classes at the same time. Essentially, the proposed methodology breaks any C_m-class classification problem to a series of 2-class classification problems.

DAGs can be seen as a generalization of the class of Decision Trees, while the redundancies and repetitions that may occur in different branches of the tree can be observed more efficiently since different decision paths might be merged. In addition, DAGs are able to collect and conduct a series of tasks in an ordered manner, subject to constraints that certain tasks must be performed earlier than others. The sequential execution of tasks is particularly important and directly related to the efficacy with which the overall task is addressed (VanderWeele & Robins, 2010).

The DAG-HMM architecture used in this paper includes $_m\,(_m - 1)/2$ nodes, each one associated with a two-class classification problem. The connections between the different nodes in G have only one orientation without any kind of loop(s). As a result, each node of a such a so-called rooted DAG has either 0 or 2 leaving arcs.

The following subsections provide a detailed analysis of the way the DAG-HMM is constructed and subsequently operates. The principal issue associated with the design of every DAG is the topological ordering, i.e. ordering the nodes in a way that the starting endpoints of every edge occur earlier than the corresponding ending endpoints. In the following, we describe how such a topological ordering is discovered based on the Kullback-Leibler divergence.

Figure 2. The determination of the topological ordering (for simplicity, only four classes are considered)



Determining the Topological Ordering of the DAG-HMM

Naturally, one would expect that the performance of the DAG-HMM depends on the order in which the different classification tasks are conducted. This was also evident from early experimentations. This observation motivated the construction of the DAG-HMM so that "simple" tasks are executed earlier in the graph. In other words, these are placed in the top nodes of the DAG-HMM, in a way that classes responsible for a high number of misclassifications are discarded early in the graph operation. In order to get an early indication of the degree of difficulty of a classification task, we employed the metric representing the distance of the involved classes in the probabilistic space, i.e. the Kullback-Leibler Divergence (KLD). The basic motivation is to place early in the DAG-HMM tasks concerning the classification of classes with large KLD, as they could be completed with high accuracy. The scheme determining the topological ordering is illustrated in Figure 2. The KLD between two J-dimensional probability distributions A and B is defined as in (Taylor, 2006).

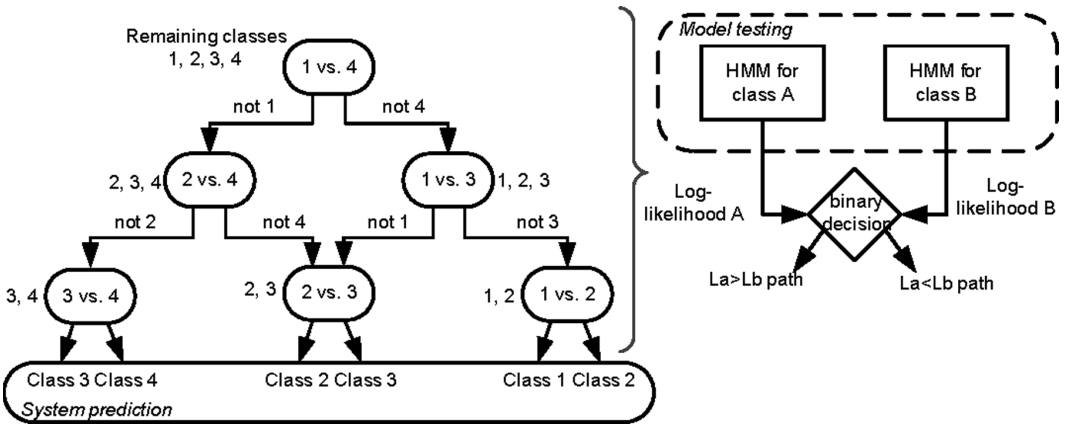
It should be noted the KLD between HMMs was not used since computing distances between HMMs of unequal lengths, which might be common in this work as HMMs representing different classes might have different number of states, can be significantly more computationally demanding without a corresponding gain in modeling accuracy (Zhao, Zhang, Soong, Chu, & Xiao, 2007; Liu, Soong, & Zhou, 2007).

After computing the KLD for the different pairs of classes, i.e. reach the second stage depicted in Figure 2, the KLD distances are sorted in a decreasing manner. This way the topological ordering of the DAG-HMM is revealed placing the classification tasks of low difficulty on its top. Each node removes a class from the candidate list until there is only one class left, which comprises the DAG-HMM prediction. The elements of the distance matrix could be seen as early performance indicators of the task carried out by the corresponding node. The proposed topological ordering places tasks likely to produce misclassifications at the bottom of the graph. This process outputs a unique solution for the topological sorting problem, as it is usually met in the graph theory literature (Cook, 1985).

The DAG-HMM Operation

The operation of the proposed DAG-HMM scheme is the following: after extracting the features of the unknown audio signal, the first/root node is activated. More precisely, the feature sequence is fed to the HMMs, which produce two log-likelihoods showing the degree of resemblance between the training data of each HMM and the unknown one. These are compared and the graph flow continues on the larger log-likelihood path. It should be stressed out that the HMMs are optimized (in terms of number of states and Gaussian components) so that they address the task of each node optimally.

Figure 3. An example of a DAG-HMM addressing a problem with four classes along with operation carried out by each node



That said, it is possible that a specific class is represented by HMMs with different parameters when it comes to different nodes of the DAG-HMM.

An example of a DAG-HMM addressing a problem with four classes is illustrated in Figure 3. The remaining classes for testing are mentioned beside each node. Digging inside each node, Figure 3 shows the HMM-based sound classifier responsible for activating the path of the maximum log-likelihood.

The operation of the DAG-HMM may be parallelized with that of investigating a list of classes, where each level eliminates one class from the list. More in detail, in the beginning the list includes all the potential audio classes. At each node the feature sequence is matched against the respective HMMs and the model with the lowest log-likelihood is erased from the list, while the DAG-HMM proceeds to the part of the topology without the discarded class. This process terminates when only one class remains in the list, which comprises the system's prediction. Hence, in case the problem deals with m different classes, the DAG's decision will be made after the evaluation of m-1 nodes.

1. Input: Novel sound S_u after concept drift detection (Ntalampiras, 2016), parameter k;
2. Partition the dataset into *TS* and *VS*;
3. Build GMM G_u and find the k closest models in *TS*;
4. Majority voting in k and discover the class c closest to S_u;
5. Employ *TS* of class c to learn and optimize the transformation T: $TS_c -> S_u$ based on the minimum reconstruction error on *VS*;
6. Apply T on VS_c and augment TS_u;

Algorithm 1: The algorithm for dataset augmentation based on Transfer Learning.

The Feature Set

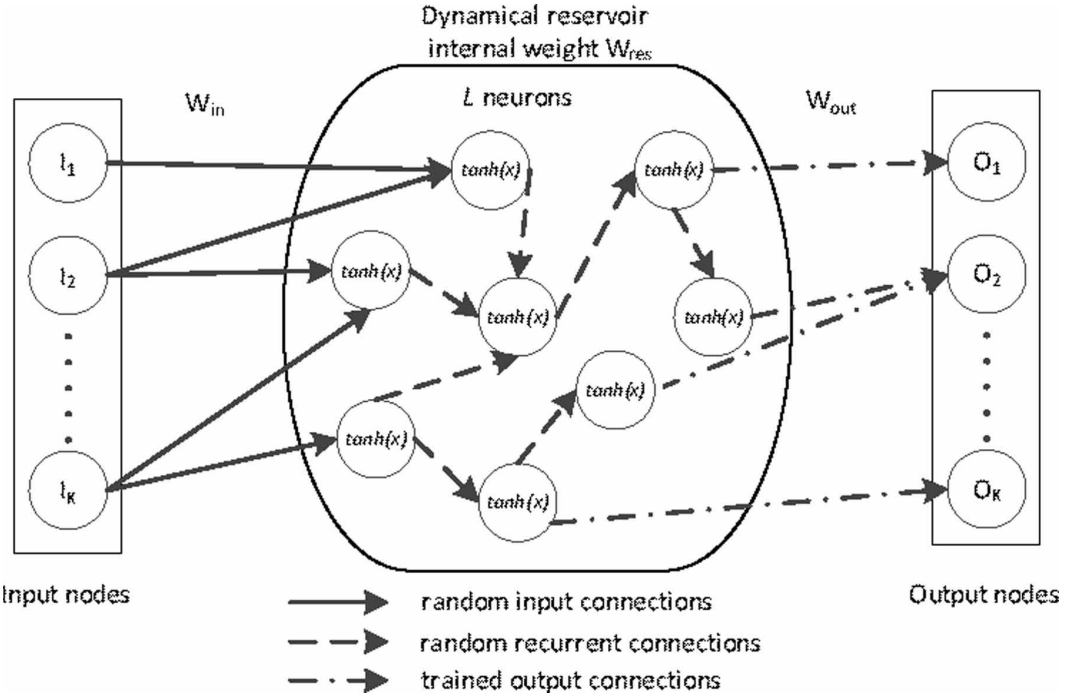
This feature set is composed of the first thirteen Mel frequency cepstral coefficients including the 0-th coefficient which reflects upon the energy of each frame. For MFCC's derivation we compute the power of the short time Fourier transform with respect to every frame and pass them through a triangular Mel scale filterbank. Subsequently, the log operator is applied and the energy compaction properties of discrete cosine transform are exploited in order to decorrelate and represent the majority of each energy band with just a few coefficients. Lastly, a thirteen-dimension vector is formed by the most important coefficients. Three derivatives of the initial vector are appended resulting to 52 dimensions. The processing stage was based on the openSMILE feature extraction tool (Eyben, Weninger, Gross, & Schuller, 2013).

Data Augmentation Based on Transfer Learning

Unlike deformation-based efforts, e.g. time stretching, pitch shifting, etc. (Salamon & Bello, 2017), we propose to transfer knowledge existing in the available dataset to augment the data of the novel sound source. The proposed algorithm first finds statistically close recordings included in the assumed to be known dataset, and subsequently selects the closest class.

Let the novel sound source, after concept drift detection as described in (Ntalampiras, 2016), be S_u (line 1, Algorithm 1). After portioning the dataset into training *TS* and validation *VS* sets (line 2, Algorithm 1), the algorithm creates G_u and finds the k closet models in *TS* using Equation 3 (line 3, Algorithm 1). Subsequently, we discover the class c closest to S_u based on majority voting (line 4, Algorithm 1). Finally, the algorithm learns the transformation T by employing TS_c (line 5, Algorithm 1), and augments TS_u by applying T on VS_c (line 6, Algorithm 1).

Figure 4. The Echo State Network used for feature space transformation



The Proposed ESN Transfer Learning Module

In this work, the transfer learning transformation T is a multiple-input multiple-output Echo State Network. ESN modeling, and in particular Reservoir Network (RN), was employed at this stage as it is able to capture the non-linear relationships existing in the data (Lukosevicius & Jaeger, 2009; Verstraeten, Schrauwen, & Stroobandt, 2006; Jalalvand, Triefenbach, Verstraeten, & Martens, 2011).

The typical topology of an RN is demonstrated in Figure 4. In is composed of neurons including non-linear activation functions with two possibilities: a) connection with the input data (so-called input connections), and b) connection to each other (so-called recurrent connections). Both of them are assigned randomly generated weights during the learning stage and remain constant during the operation of the RN. Lastly, each output node holds a connection to a linear function.

The basic motivation behind reservoir computing lies behind the computational complexity of the back-propagation algorithm. During its application, the internal layers are not altered significantly, thus it is not included in RN learning. On the other hand, the output layer is associated with a linear problem and as such, of relatively low degree of perplexity. Nonetheless, the stability of the network is ensured by constraining the weights of the internal layers. Linear regression is employed to learn output weights, so-called read-outs in the literature. A detailed analysis of this process is out of the scope of this work, while the interested reader is directed at (Lukosevicius & Jaeger, 2009; Jaeger & Haas, 2004) for more information.

EXPERIMENTAL EVALUATION

In this section, we analyze the: a) dataset used to acoustically simulate a farm environment, b) parametrization of both DAG-HMM and feature extraction module, c) contrasted approaches, and d) we present and comment the achieved results.

Dataset

We collected data associated with the following typical farm animals: dog, rooster, pig, cow, cat, hen, and sheep. These are taken from the Environmental Sound Classification-10 described in (Piczak, 2015b), while they are sampled at 44.1 KHz. Each class includes 40 recordings, each one with a duration of 5 seconds.

System Parametrization

Following the MPEG-7 standard recommendation, the low-level feature extraction window is 30 ms with 10 ms overlap, so that the system is robust against possible misalignments. The sampled data are hamming windowed to smooth potential discontinuities while the FFT size is 512. Standard normalization techniques, i.e. mean removal and variance scaling, were applied.

The HMMs of each node are optimized in terms of number of states and nodes following the Expectation-Maximization and Baum Welch algorithms (Rabiner, 1989). As the considered sound events are characterized by a distinct time evolution, we employed HMMs with left-right topology, i.e. only left to right states transitions are permitted. Moreover, the distribution of each state is approximated by a Gaussian mixture model of diagonal covariance, which may be equally effective to a full one at a much lower computational cost (Reynolds & Rose, 1995).

The maximum number of k-means iterations for cluster initialization was set to 50 while the Baum-Welch algorithm used to estimate the transition matrix was bounded to 25 iterations with a threshold of 0.001 between subsequent iterations. The number of explored states ranges from 3 to 7 while the number of Gaussian components used to build the GMM belongs to the {2, 4, 8, 16, 32, 64, 128, 256, and 512} set. The final parameters were selected based on the maximum recognition rate criterion. The machine learning package Torch (freely available at http://torch.ch/) was used to construct and evaluate GMMs and HMMs.

Contrasted Approaches

The proposed approach was contrasted to the following ones: class-specific HMM (Kim & Sikora, 2004), universal HMM with a KLD based data selection scheme (Ntalampiras, 2013), support vector machine (SVM) with radial basis function kernel (Chen, Gunduz, & Ozsu, 2006), random forest (Al-Maathidi & Li, 2015), and echo state network (Scardapane & Uncini, 2017). The parameters of these classification schemes were optimized on *TS*. As for the feature set, we experimented with the descriptors from the MPEG-7 audio protocol (Casey, 2001) and the Perceptual Wavelet Packets (Ntalampiras, Potamitis, & Fakotakis, 2009), which have shown encouraging performance in generalized sound recognition tasks. The ESN implementation is based on the Echo State Network

Table 1. The recognition rates achieved by the proposed and contrasted approaches. The approach providing the highest rate is emboldened.

Classifier	Average recognition rate (%)
DAG-HMM	**93.1**
DAG-HMM (concept drift)	*90.8*
Class-specific HMMs	77.1
Universal HMM	68.6
SVM	52.3
ESN	60
RF	54.3

toolbox (freely available at https://sourceforge.net/projects/esnbox/) and the SVM on the libsvm library (Chang & Lin, 2011).

As far as the transfer learning module is concerned, we performed a comparison with Stacked AutoEncoders (SAE) which have been used for analyzing emotion manifestations across music and speech signals (Coutinho, Deng, & Schuller, 2014). The experiment was conducted as described in Algorithms 1 and 2 while the log-likelihood is the evaluation metric.

1. Input: Augmented TS_u^t from Algorithm 1
2. Learn HMMs H, where states $s \in \{3, 4, 5, 6\}$ and components $g \in \{2, 4, 8, 16, 32, 64, 128, 256\}$
3. Compute the log-likelihoods $L = P(VS_u|H)$
4. Find the maximum log-likelihood in L
5. Identify the HMM providing the highest modeling accuracy of S_u

Algorithm 2: The algorithm for evaluating the augmented feature set and selecting the optimum HMM to represent the novel class.

Experimental Results

Algorithm 2 is designed to assess the performance of the transfer learning-based dataset augmentation explained in Algorithm 1. The distribution of the augmented feature set TS_u^t (line 1, Algorithm 2) is learnt by means of HMM H constructed using states $s \in \{3, 4, 5, 6\}$ and components $g \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ (line 2, Algorithm 2). In assessing the constructed models we compute the log-likelihoods on VS_u (line 3, Algorithm 2) and discovering the highest one (line 4, Algorithm 2). The respective HMM is identified and stored, while the associated log-likelihood constitutes a measure of compatibility between the augmented feature set TS_u^t and the actual data of the novel class VS_u.

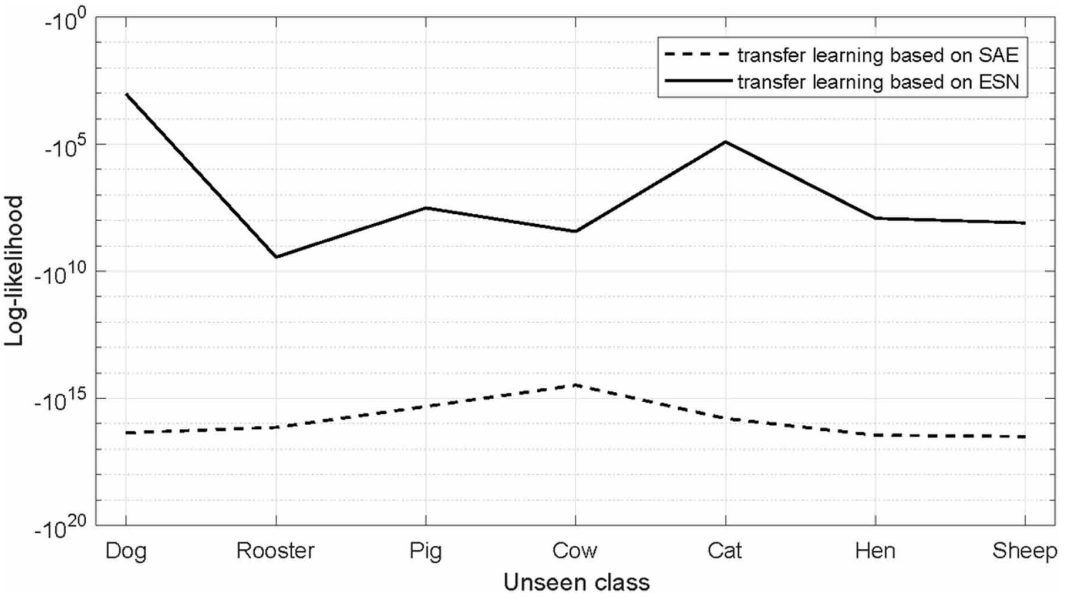
Figure 5 demonstrates the comparative results w.r.t the transfer learning capabilities of ESN and SAE. As we can see, the proposed ESN-based scheme outperforms the SAE one across all categories as it provides more accurate modeling as evaluated by the log-likelihoods produced on the *VS* (see Algorithm 1). As the log-likelihoods demonstrate significantly larger values, we argue that the ESN shows relevant ability in capturing the non-linear relationship existing in associating the features extracted from sound S_u and the features in *TS*.

Then, Table 1 includes the results achieved by the proposed DAG-HMM as well as the contrasted approaches. The data division protocol is the ten-fold cross validation one. Identically selected folds were used during the training and testing processes of all approaches, enabling a reliable comparison. A first observation is on the difficulty of the task which is relatively high since many classifications

Table 2. The confusion matrix (in %) with respect to the DAG-HMM. The average classification rate is 93.1%.

Responded / Presented	*Dog*	*Rooster*	*Pig*	*Cow*	*Cat*	*Hen*	*Sheep*
Dog	**99.4**	-	-	-	-	-	0.6
Rooster	-	**99.7**	-	-	-	0.3	-
Pig	14.4	-	**85.6**		-	-	-
Cow	-	-	-	**99.8**	0.2	-	-
Cat	13.9	-	-	-	**86.1**	-	-
Hen	-	0.9	-	-	-	**99.1**	-
Sheep	-	-	-	-	-	17.7	**82.3**

Figure 5. The comparative results of the transfer learning module when using ESN and SAE



schemes fail to provide a satisfactory recognition rate. Then, as we can see, the proposed DAG-HMM outperforms the rest of the approaches. The second one is based on class specific HMMs, while the ESN achieved the third best recognition rate. The UBM logic provides lower rate that the class-specific one showing the high degree of diversity characterizing the common feature space. Same conclusions can be derived for the SVM, which cannot find reliable boundaries between the classes and the RF, the rules of which do not classify the feature space in a reliable manner. It is worth to note the satisfactory recognition rate achieved by the DAG-HMM (90.8%) in the concept drift environment, i.e. when one class was unknown to the classifier. In the specific set of experiments only 5s of the novel class were presented to the system, while the corresponding feature space was augmented via the ESN-based transfer learning module. This experiment was performed for all recordings and classes in *TS*, and we report the averaged recognition rate.

We conclude that limiting the problem space using a DAG-HMM is particularly beneficial in the specific application scenario providing encouraging recognition rates. In a subsequent step, we experimented with the MPEG-7 and PWP feature sets: the DAG-HMM provides recognition rates of 75.7% and 68.9% respectively. As in (Kim & Sikora, 2004), the superiority of the MFCCs in a generalized sound recognition task, was confirmed.

The confusion matrix achieved by the DAG-HMM is tabulated in Table 2. We observe that the class recognized with the highest accuracy is *cow* one (99.8%), while the one presenting the worst rate is the *sheep* one (82.3%). The misclassifications' source is the great variability among sound samples of the same class as it is assessed by a human listener. Several sound clips are acoustically similar even though they belong to different categories. This is particularly evident in the cases of *sheep-hen*, *cat-dog*, and *pig-dog* pairs. We conclude that the DAG-HMM classification approach provides promising performance; even though the associated computational cost of the training phase is rather high, it is to be conducted only once and offline. At the same time, the testing phase includes simple log-likelihood comparisons and estimations carried out using the Viterbi algorithm, which is computationally inexpensive as it is based on recursive dynamic programming.

CONCLUSION

This paper presented a classification scheme addressing the novel scientific area of acoustic farm monitoring. We outlined a classification scheme based on a DAG composed of HMMs trained on and MFCCs feature set. The superiority of the proposed scheme over state-of-the-art classifiers was proven on a publicly available dataset encompassing vocalizations of seven farm animals. Importantly, the present framework is able to operate in a concept drift environment, i.e. being able to online evolve itself and increase the dictionary of animal vocalizations.

In the future work, we plan to enhance the proposed system so that it is able to operate under noisy conditions, and especially non-stationary noise, and evaluate it using real-world recordings.

ACKNOWLEDGMENT

This research was funded by the ELKE TEI Crete funds related to the domestic project: Bioacoustic applications.

REFERENCES

Al-Maathidi, M. M., & Li, F. F. (2015, December). Audio content feature selection and classification a random forests and decision tree approach. *Proceedings of the 2015 IEEE international conference on progress in informatics and computing (PIC)* (p. 108-112). IEEE Press. doi:10.1109/PIC.2015.7489819

Anu, V. M., Deepika, M. I., & Gladance, L. M. (2015, February). Animal identification and data management using RFID technology. *Proceedings of the Int. conf. on innovation information in computing technologies* (p. 1-6). Academic Press. doi:10.1109/ICIICT.2015.7396069

Blumstein, D., Mennill, D., Clemins, P., Girod, L., Yao, K., Patricelli, G., Kirschel, A. (2011, 6). Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *Journal of Applied Ecology*, *48*(3), 758{767. doi:10.1111/j.1365-2664.2011.01993.x

Casey, M. (2001). Mpeg-7 sound-recognition tools. *IEEE Trans. on Circuits and Systems for Video Technology*, *11*(6), 737-747.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*. Retrieved from http://www.csie.ntu.edu.tw/cjlin/libsvm)

Chen, L., Gunduz, S., & Ozsu, M. T. (2006, July). Mixed type audio classification with support vector machine. *Proceedings of the 2006 IEEE international conference on multimedia and expo* (p. 781-784). IEEE Press. doi:10.1109/ICME.2006.262954

Cook, S. A. (1985). A taxonomy of problems with fast parallel algorithms. *Information and Control*, *64*(1), 2–22. doi:10.1016/S0019-9958(85)80041-3

Coutinho, E., Deng, J., & Schuller, B. (2014, July). Transfer learning emotion manifestation across music and speech. *Proceedings of the 2014 international joint conference on neural networks* (IJCNN) (p. 3592-3598). Academic Press. doi:10.1109/IJCNN.2014.6889814

Dargie, W. (2009, July). Adaptive audio-based context recognition. *IEEE Transactions on Systems, Man, and Cybernetics. Part A, Systems and Humans*, *39*(4), 715–725. doi:10.1109/TSMCA.2009.2015676

Ditzler, G., Roveri, M., Alippi, C., & Polikar, R. (2015, November). Learning in nonstationary environments: A survey. *IEEE Computational Intelligence Magazine*, *10*(4), 12–25. doi:10.1109/MCI.2015.2471196

Dong, X., Towsey, M., Zhang, J., & Roe, P. (2015, November). Compact features for birdcall retrieval from environmental acoustic recordings. *Proceedings of the 2015 IEEE international conference on data mining workshop (ICDMW)* (p. 762-767). doi:10.1109/ICDMW.2015.153

Exadaktylos, V., Silva, M., & Berckmans, D. (2014). Automatic identification and interpretation of animal sounds, application to livestock production optimisation. In H. Glotin (Ed.), *Soundscape semiotics - localization and categorization*. Rijeka: InTech; doi:10.5772/56040

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in opensmile, the munich open-source multimedia feature extractor. *Proceedings of the 21st ACM international conference on multimedia* (pp. 835-838). New York: ACM. doi:10.1145/2502081.2502224

Gao, B., Woo, W. L., & Dlay, S. S. (2011, September). Adaptive sparsity non-negative matrix factorization for single-channel source separation. *IEEE Journal of Selected Topics in Signal Processing*, *5*(5), 989–1001. doi:10.1109/JSTSP.2011.2160840

Grill, T., & Schlter, J. (2017, Aug). Two convolutional neural networks for bird detection in audio signals. *Proceedings of the 2017 25th European signal processing conference (EUSIPCO)* (p. 1764-1768). Academic Press. doi:10.23919/EUSIPCO.2017.8081512

Han, N. C., Muniandy, S. V., & Dayou, J. (2011). Acoustic classification of Australian anurans based on hybrid spectral entropy approach. *Applied Acoustics*, *72*(9), 639–645. doi:10.1016/j.apacoust.2011.02.002

Jaeger, H., & Haas, H. (2004). Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, *304*(5667), 78–80. doi:10.1126/science.1091277 PMID:15064413

Jalalvand, A., Triefenbach, F., Verstraeten, D., & Martens, J. (2011). Connected digit recognition by means of reservoir computing. *Proceedings of the 12th annual conference of the international speech communication association* (pp. 1725-1728). Academic Press.

Kim, H.-G., & Sikora, T. (2004, May). Comparison of mpeg-7 audio spectrum projection features and mfcc applied to speaker recognition, sound classification and audio segmentation. *Proceedings of the 2004 IEEE international conference on acoustics, speech, and signal processing* (Vol. 5, pp. 925-928). IEEE Press. doi:10.1109/ICASSP.2004.1327263

Kumar, A., & Hancke, G. P. (2015, January). A zigbee-based animal health monitoring system. *IEEE Sensors Journal*, *15*(1), 610–617. doi:10.1109/JSEN.2014.2349073

Li, B., Zhou, Z., Zou, W., & Li, D. (2012, September). Quantum memetic evolutionary algorithm-based low-complexity signal detection for underwater acoustic sensor networks. *IEEE Transactions on Systems, Man and Cybernetics. Part C, Applications and Reviews*, *42*(5), 626–640. doi:10.1109/TSMCC.2011.2176486

Liu, P., Soong, F. K., & Zhou, J. L. (2007, April). Divergence-based similarity measure for spoken document retrieval. Proceedings of the 2007 IEEE international conference on acoustics, speech and signal processing - ICASSP '07 (Vol. 4, p. 89-92). doi:10.1109/ICASSP.2007.367170

Lukosevicius, M., & Jaeger, H. (2009, August). Survey: Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.*, *3*(3), 127-149. doi:10.1016/j.cosrev.2009.03.005

Mitrovic, D., Zeppelzauer, M., & Breiteneder, C. (2006). Discrimination and retrieval of animal sounds. *Proceedings of the 2006 12th international multi-media modelling conference*. Academic Press. doi:10.1109/MMMC.2006.1651344

Nagpal, S. K., & Manojkumar, P. (2016, February). Hardware implementation of intruder recognition in a farm through wireless sensor network. *Proceedings of the 2016 international conference on emerging trends in engineering, technology and science (ICETETS)* (p. 1-5). Academic Press. doi:10.1109/ICETETS.2016.7603012

Noda, J. J., Travieso, C. M., Snchez-Rodrguez, D., Dutta, M. K., & Singh, A. (2016, February). Using bioacoustic signals and support vector machine for automatic classification of insects. *Proceedings of the 2016 3rd int. conf. on signal processing and integrated networks* (pp. 656-659). Academic Press. doi:10.1109/SPIN.2016.7566778

Ntalampiras, S. (2013, February). A novel holistic modeling approach for generalized sound recognition. *IEEE Signal Processing Letters*, *20*(2), 185–188. doi:10.1109/LSP.2013.2237902

Ntalampiras, S. (2014). Directed acyclic graphs for content based sound, musical genre, and speech emotion classification. *Journal of New Music Research*, *43*(2), 173–182. doi:10.1080/09298215.2013.859709

Ntalampiras, S. (2016, September). Automatic analysis of audiostreams in the concept drift environment. *Proceedings of the 2016 IEEE 26th international workshop on machine learning for signal processing (MLSP)* (p. 1-6). IEEE Press. doi:10.1109/MLSP.2016.7738905

Ntalampiras, S. (2018a). Bird species identification via transfer learning from music genres. *Ecological Informatics*, *44*, 76–81. doi:10.1016/j.ecoinf.2018.01.006

Ntalampiras, S. (2018b, November). A classification scheme based on directed acyclic graphs for acoustic farm monitoring. *Proceedings of the 2018 23rd conference of open innovations association (FRUCT)* (p. 276-282). Academic Press. doi:10.23919/FRUCT.2018.8588077

Ntalampiras, S., Potamitis, I., & Fakotakis, N. (2009). Exploiting temporal feature integration for generalized sound recognition. *EURASIP Journal on Advances in Signal Processing*, (1).

Piczak, K. J. (2015a, September). Environmental sound classification with convolutional neural networks. *Proceedings of the 2015 IEEE 25th international workshop on machine learning for signal processing (MLSP)* (p. 1-6). IEEE Press. doi:10.1109/MLSP.2015.7324337

Piczak, K. J. (2015b). Esc: Dataset for environmental sound classification. *Proceedings of the 23rd acm international conference on multimedia* (pp. 1015-1018). New York: ACM. doi:10.1145/2733373.2806390

Rabiner, L. R. (1989, February). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. doi:10.1109/5.18626

Reynolds, D. A., & Rose, R. C. (1995, January). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*(1), 72–83. doi:10.1109/89.365379

Salamon, J., & Bello, J. P. (2017, March). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*(3), 279–283. doi:10.1109/LSP.2017.2657381

Scardapane, S., & Uncini, A. (2017, February). Semi-supervised echo state networks for audio classification. *Cognitive Computation*, *9*(1), 125-135. doi:10.1007/s12559-016-9439-z

Stowell, D. (2018). Computational bioacoustic scene analysis. In T. Virtanen, M. D. Plumbley, & D. Ellis (Eds.), Computational analysis of sound scenes and events (pp. 303-333). Cham: Springer International Publishing. doi:10.1007/978-3-319-63450-0_11

Taylor, P. (2006). The target cost formulation in unit selection speech synthesis. In *INTERSPEECH 2006 - ICSLP, ninth international conference on spoken language processing*, Pittsburgh, PA, September 17-21. Academic Press.

Towsey, M. W., Truskinger, A. M., & Roe, P. (2015, Nov). The navigation and visualisation of environmental audio using zooming spectrograms. *Proceedings of the 2015 IEEE international conference on data mining workshop (ICDMW)* (p. 788- 797). doi:10.1109/ICDMW.2015.118

VanderWeele, T. J., & Robins, J. M. (2010). Signed directed acyclic graphs for causal inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(1), 111-127. doi:10.1111/j.1467-9868.2009.00728.x

Verstraeten, D., Schrauwen, B., D'Haene, M., & Stroobandt, D. (2007). An experimental unification of reservoir computing methods. *Neural Networks*, *20*(3), 391–403. doi:10.1016/j.neunet.2007.04.003 PMID:17517492

Verstraeten, D., Schrauwen, B., & Stroobandt, D. (2006, July). Reservoir- based techniques for speech recognition. Proceedings of the international joint conference on neural networks IJCNN '06 (p. 1050-1053). Academic Press. doi:10.1109/IJCNN.2006.246804

Zhao, Y., Zhang, C., Soong, F. K., Chu, M., & Xiao, X. (2007). *Measuring attribute dissimilarity with hmm kl-divergence for speech synthesis.*

Stavros Ntalampiras is an Assistant Professor at the Department of Informatics of the University of Milan. He received the engineering and Ph.D. degrees from the Department of Electrical and Computer Engineering, University of Patras, Greece, in 2006 and 2010, respectively. He has carried out research and/or didactic activities at Politecnico di Milano, the Joint Research Center of the European Commission, the National Research Council of Italy, and Bocconi University. Currently, he is an Associate Editor of IEEE Access and a member of the IEEE Computational Intelligent Society Task Force on Computational Audio Processing, while his research interests include content-based signal processing, audio pattern recognition, machine learning, and cyber physical systems.

Ilyas Potamitis received the B.Eng. and Dr.Eng. degrees in electrical engineering from the Department of Electrical and Computer Engineering. University of Patras, Patras, Greece, in 1995 and 2002, respectively. He is a Full Professor at the Department of Music Technology and Acoustics of the Technological Educational Institute of Crete, Greece. His research interests are audio signal processing, biacoustics, speech enhancement, speech feature extraction for robust speech recognition, and one-channel signal separation.