Received: 7 February 2020 | Revised: 27 October 2020 | Accepted: 7 December 2020 | CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12007

ORIGINAL RESEARCH PAPER

Acoustic detection of unknown bird species and individuals

Stavros Ntalampiras[1] | **Ilyas Potamitis**[2]

[1]Department of Computer Science, University of Milan, Milan, Italy

[2]Department of Music Technology and Acoustics, Hellenic Mediterranean University, Rethymno, Greece

Correspondence

Stavros Ntalampiras, via Celoria 18, 20133, Milan, Italy.
Email: stavros.ntalampiras@unimi.it

Funding information

Università degli Studi di Milano, Grant/Award Number: INSPIRE

Abstract

Computational bioacoustics is a relatively young research area, yet it has increasingly received attention over the last decade because it can be used in a wide range of applications in a cost-effective manner. This work focuses on the problem of detecting the novel bird calls and songs associated with various species and individual birds. To this end, variational autoencoders, consisting of deep encoding–decoding networks, are employed. The encoder encompasses a series of convolutional layers leading to a smooth high-level abstraction of log-Mel spectrograms that characterise bird vocalisations. The decoder operates on this latent representation to generate each respective original observation. Novel species/individual detection is carried out by monitoring and thresholding the expected reconstruction probability. We thoroughly evaluate the proposed method on two different data sets, including the vocalisations of 11 North American bird species and 16 *Athene noctua* individuals.

1 | INTRODUCTION

Acoustic monitoring of bird activity is vital for various environmental, research, and scientific goals [1]. Most existing works focus on detecting birds by their sounds, which is the primary step in a series of applications: biodiversity monitoring, detection of endangered species etc. [2,3]. Indeed, the area of computational bioacoustics has gained attention in recent years, especially after the mass diffusion of automated recording units, that is, devices that can record, store, and potentially transmit audio recorded in the wild to remote locations where further processing is usually carried out. These devices have become popular owing to the audio mode, an attractive and suitable characteristic for bird monitoring—many bird species are much easier to detect by their audio patterns than by data captured through other modes, such as video. Importantly, the audio mode is not affected by occlusions, lighting conditions etc., placing it in a unique position for monitoring the activities of bird species.

In general, applications of computational bioacoustics concern the analysis of a habitat's health including population densities/trends of target species, migration patterns, protection of endangered species etc. For example, seabird colonies are currently being analysed using bird call activity [4,5], by tracking general animal population density acoustically, [6] by identifying calls of *black-rumped flameback* in real-time etc.

Let us define the set $\mathcal{S}$, which includes all species a priori known to reside in a specific habitat. To the best of our knowledge, the existing research has analysed species while assuming complete knowledge of $\mathcal{S}$, that is, its size and composition [7,8]. Thus, the outcome is the presence/absence of species included in $\mathcal{S}$. However, this assumption might not always be true, as both new species (even if only for migration purposes, i.e. for a limited time) and new individuals may appear among an existing species. Changes in $\mathcal{S}$ alter species diversity, and their analyses could be useful to track migratory movements, record seasonal changes, detect invasive species etc. [9] with the overall goal being the preservation of habitat quality and balance. The consequences depend on the characteristics of the new species, and they range from biodiversity loss to the normal continued operation of the altered ecosystem [10]. This could be particularly useful in long-term monitoring scenarios, where automatic recording units can be employed in a standardised manner under potentially harsh environmental conditions to deliver accurate biodiversity indices [11].

In such a scenario (see Figure 1]), the first step in processing new species/individuals is detecting them. This work is concentrated on this exact problem—learning in non-stationary environments where $\mathcal{S}$ is not static during system operation—that is, its composition and cardinality are subject to change. Such a problem is addressed in the literature on novelty and

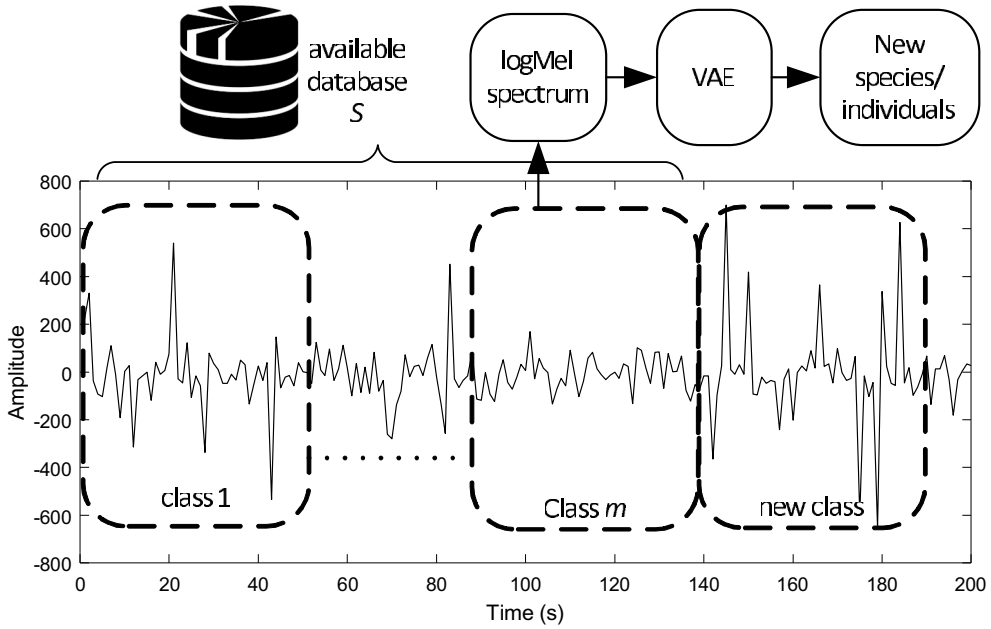

FIGURE 1 Proposed method's pipeline including (a) signal windowing, (b) log-Mel spectrum extraction, and (c) acoustic detection of novel bird vocalisations of species and individual birds

concept drift detection [12], while solutions that address the field of audio signal processing are limited. The case most similar to this work is presented in [13], where a typical home environment is considered. There, the solution is based on a change detection test consisting in a hidden Markov model (HMM) that characterises the available set of classes by operating in the feature space formed by Mel-frequency cepstral coefficients.

This work proposes the use of variational autoencoders (VAEs) for detecting changes in $\mathcal{S}$, as VAEs have been proved effective in similar tasks such as network attack detection [14], anomaly detection in energy time series [15], images [16], machine acoustics [17,18] etc. VAEs are a family of powerful generative statistical modelling techniques that fit the current problem's specifications. Importantly, they encompass two principal processing stages [19,20]:

(a) an *encoder* able to learn a non-linear projection of the input signal space—the so-called latent space (encoding)—that is typically characterised by a relatively small number of dimensions, and
(b) a *decoder* for inverse non-linear transformation of the latent coefficients into the original signal space.

Keeping in mind the considerations expressed in [21], we used suitably normalised log-magnitude spectra to characterise the audio signal space. . In sequence, the encoder and decoder consist of a series of convolutional and transposed convolutional layers. The network is trained to minimise reconstruction loss and Kullback–Leibler (KL) divergence between the input and output log-spectrograms. Detection of novel audio signals is carried out by monitoring the respective reconstruction probability based on a threshold determined on a validation set during the training phase.

The proposed solution is thoroughly evaluated in two data sets: [(a)] the first comprises 11 North American bird species [22], while [(b)] the second comprises 16 individuals of the little owl (*Athene noctua*) species [23]. The experimental protocol follows leave-one-species-out/leave-one-individual-out logic, and we report excellent results in terms of false positive and false negative rates, thus improving the current state of the art for HMMs. We also present an analysis of the latent spaces learned by the constructed VAE.

The following section formalises the present problem, while section 3 describes the VAE and the change detection process as well as a representation of the audio signal employed. Section 4 details the experimental set-up including a brief description of the data sets and the parameterisation of the approaches as well as the results analysis. Finally, we draw our conclusions in section 5.

2 | PROBLEM DEFINITION

This work assumes availability of a data set including monophonic audio signals, denoted as y_t. We further assume a single dominant sound source at each time t, leaving the problem of composite sound scenes for future work. The sources come from a known but unbounded set of classes $\mathcal{S} = \{S_1, \ldots, S_m\}$, where S_i denotes the i-th class, with $i \in \mathbb{N}^+, 1 < i < m$, and m representing the total number of known classes. At the same time, each class is stationary over time, that is, characterised by a consistent yet unknown probability density function, P_i.

In contrast to the limits described in the vast majority of related literature, $\mathcal{S}$ is unbounded for this system, meaning that new classes may appear during system operation. These may correspond to either new species or individuals belonging to an a priori known species. In this case, y_t becomes y_t', that is,

$$y_t = \begin{cases} y_t & t < t^* \\ y_t', & t \geq t^*, \end{cases} \tag{1}$$

where t^* is the starting time instance of the manifestation of a new sound class. In order to address such an increasingly complex auditory scene, the current analysis method should be adapted, or a new one should be designed from scratch.

This formulation assumes the availability of an initial training sequence, $TS = y_t$, $t \in [1, T_0]$, where the involved classes are known via labelled pairs (y_t, S_i), $i \in [1, m]$. No assumptions are made about the number of new classes or the properties associated with the probability density functions. The overall goal is to detect changes in $\mathcal{S}$ with the smallest false positive and false negative rates.

Algorithm 1 Proposed variational auto-encoder–based detection of new bird species/individuals

Input: set $\mathcal{S}$, test audio signal y^t;
Output: New/known species/individual detection ;
1. Divide data in $\mathcal{S}$ into training TS and validation sets VS;
2. Extract log-Mel spectrograms F^{TS} and F^{VS} out of TS and VS respectively ;
3. Learn VAE V as per section 3.1 using F^{TS};
4. Determine threshold as follows:;
5. **for** *i=1:*$|F^{VS}|$ **do**
   6. Apply V on F_i^{VS} and get its reconstruction as follows $\hat{F}_i^{VS} = V(F_i^{VS})$;
   7. Compute error $E(i) = mse(F_i^{VS}, \hat{F}_i^{VS})$;
end
8. Determine change detection threshold as $T_h = max(E)$;
9. Extract log-Mel spectrogram F^y from test audio signal y^t ;
10. Generate its reconstruction $\hat{F}^y = V(F^y)$;
11. Compute error $E^y = mse(\hat{F}^y, F^y)$;
12. Compare to threshold T_h as follows ;
13. **if** $E^y > T_h$ **then**
   14. New species/individual detected
else
   15. Known species/individuals
end

3 | VARIATIONAL AUTOENCODERS FOR ACOUSTIC CHANGE DETECTION

Data modelling algorithms can be broadly divided into two categories, namely, *discriminative* and *non-discriminative*. The first aims to discover the boundaries discriminating classes that exist within the available data, while the second typically models the characteristics of each class independently from the rest. The generative models form an attractive solution for estimating such characteristics in a probability density form, $\mathcal{P}(\mathbf{x})$, describing a given data set. This density estimation typically includes an additional set of random variables, the so-called *latent* variables, denoted as $\mathbf{z}$. Such high-level representation operating in the data domain can be employed to suitably control the data generation process as well.

Generative models are formalised using a joint probability function, such as

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{x}|\mathbf{z})p(\mathbf{z}), \tag{2}$$

where $p(\mathbf{z})$ represents the Bayesian *prior* in the latent space. When the data-generating process is positioned in latent space $\mathbf{z}$, it corresponds to a probability density function, $p(\mathbf{x}|\mathbf{z})$, in the data domain. At the same time, we wish to compute the *posterior* distribution $p(\mathbf{z}|\mathbf{x})$ with respect to the latent one associated with a sample in the data space $\mathbf{x}$. Bayesian classification frameworks to calculate the specific posterior distribution during their operation form a rigid inference framework. Unfortunately, complex highly non-linear data distributions are typically intractable without strong assumptions about their occupancy, shape etc. To attempt to overcome this obstacle, a *variational inference* (VI) framework transforms the distribution estimation problem to one of optimisation [24]. To this end, VI starts from a distribution formalised as $q(\mathbf{z}|\mathbf{x})$ and parameterised in $\mathbf{z}$. Such a distribution is freely constructed and updated to reach the true posterior $p(\mathbf{z}|\mathbf{x})$. During this process, the next bound is respected:

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})] + D[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] = \mathcal{L}_{\text{ELBO}}(q) \tag{3}$$

with D being the KL divergence. In this inequality (Equation 3), our model $p(\mathbf{x})$ is intrinsically optimised via maximisation of the evidence $\mathbb{E}$. The above-defined bound, a so-called *evidence lower-bound* (represented as ELBO), is essentially the sum of the likelihood $p(\mathbf{x}|\mathbf{z})$ and the KL divergence imposed on the estimated distribution $q(\mathbf{z}|\mathbf{x})$ to shift it towards the prior $p(\mathbf{z})$.

Such a VI-based problem formulation does not impose Bayesian inference constraints, because it simply demands tractability of $p(\mathbf{x}|\mathbf{z})$ and KL divergence $D[q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})]$. As a result, such a VI framework accounts for both sets of parameterised distributions, $p_\theta(\mathbf{x}|\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$, and models complex relationships among them. At the same time, the advantages of the Bayesian inference formulation are maintained [25,26].

3.1 | Variational auto-encoder

Based on the above-described VI framework, both generative and inference models can be formulated as normal distributions, that is,

$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\boldsymbol{\mu}_q(\mathbf{x}), \boldsymbol{\sigma}_q^2(\mathbf{x})\right), \tag{4}$$

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\boldsymbol{\mu}_p(\mathbf{z}), \boldsymbol{\sigma}_p^2(\mathbf{z})\right). \tag{5}$$

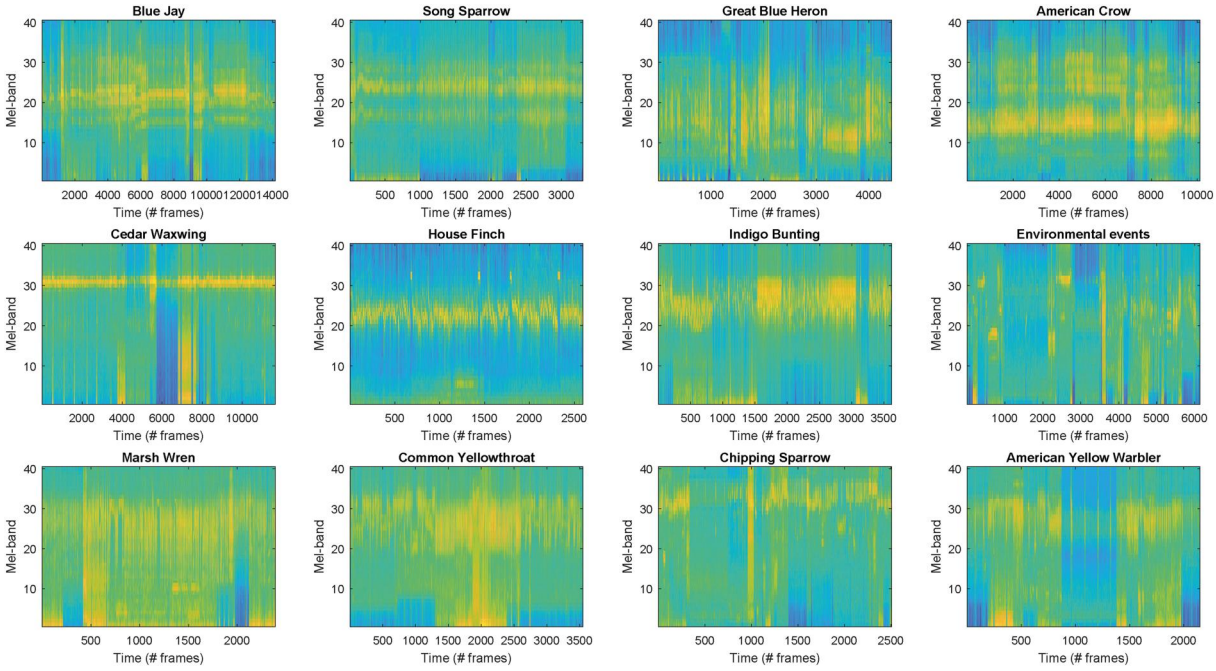

FIGURE 2 Log-Mel spectrograms of the considered bird species

Such normal distributions are parameterised in $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q^2)$ and $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$, which can be calculated using $f_\theta(\mathbf{x}; \theta)$ and $g_\phi(\mathbf{z}; \phi)$. The original composition of the VAE suggests approximating f_θ and g_ϕ as neural networks [27]. The prior distribution is typically assumed to be an isotropic normal, $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, ensuring the independence of latent dimensions. Following the terminology of auto-encoder approaches, functions $f_\theta(\mathbf{x}; \theta)$ and $g_\phi(\mathbf{z}; \phi)$ constitute the corresponding *encoder* and *decoder*. Such neural network-based architectures are collectively learned based on back-propagation until convergence occurs on $\{\theta, \phi\}$ or a stopping criterion is met. Notwithstanding such a simple framework, this approach offers distribution encoding and generation that is quite descriptive. At the same time, the latent space is well regulated by the D_{KL} divergence term.

3.2 | Anomaly detection

After the encoder and the decoder models have been defined, an anomaly detection framework is needed to identify deviations from the normal/known patterns available in the training set. This is fundamentally different logic concerning the way that neural networks are typically used, that is, to achieve a specific outcome given a specific input, such as supervised classification. A straightforward solution would be to assess an observation $\mathbf{x}$ by evaluating $p_\theta(\mathbf{x})$, potentially using Monte Carlo methods, that is, $p_\theta(\mathbf{x}) = \mathbb{E}_{p_\theta(\mathbf{z})}[p_\theta(\mathbf{x}|\mathbf{z})]$. However, as described in [28], sampling the prior distribution is not a practical solution.

Following the line of reasoning explained in [29], this work proposes to use the reconstruction probability, defined as $\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]$, for assessing the stationarity of the audio stream. Novel audio classes are expected to bring a bias to the mapped $\mathbf{z}$ exhibiting low reconstruction probabilities. Based on the findings reported in [30], each recording $\mathbf{x}$ in the validation set is given to model for reconstruction by the encoding–decoding process. There, we compute the largest reconstruction error that can make up the anomaly detection threshold. During testing, each record is reconstructed, and the produced error is checked against the threshold. If the error surpasses the threshold, an anomaly is signalled; in the opposite case, the testing record is considered generated by the normal distribution. The encoding–decoding process is demonstrated in Figure 2, while the change detection algorithm is formalised next.

3.3 | Change detection algorithm

The proposed VAE-based change detection algorithm is outlined in Algorithm 1. Its inputs are the set $\mathcal{S}$ and a test audio signal y^t, while its output is whether a new or known species/individual is detected in y^t. Initially, the algorithm divides the data in $\mathcal{S}$ into training sets, TS, and validation sets, VS (line 1, Algorithm 1). Then, the log-Mel spectrograms are extracted (line 2, Algorithm 1), and a VAE V approximates the distribution exhibited in F^{TS} (line 3, Algorithm 1). Subsequently, V is applied on F^{VS}, and the corresponding reconstruction errors are computed (lines 5–7, Algorithm 1). The maximum mean squared error is set as the detection threshold T_h (line 8, Algorithm 1).

Then, to check the test audio signal y^t, we first extract its log-Mel spectrogram (line 9, Algorithm 1) and feed it to V

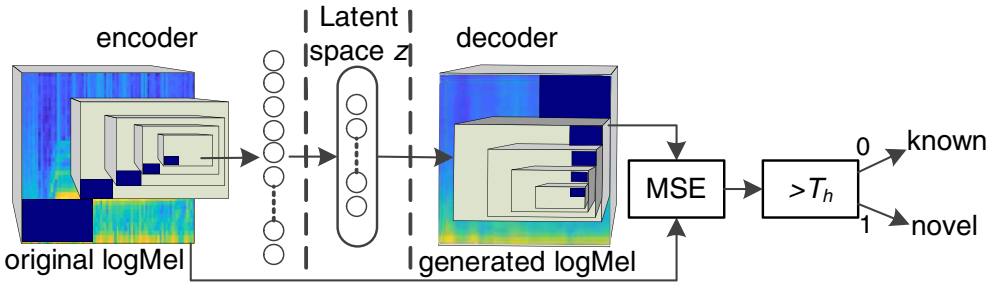

FIGURE 3 Variational auto-encoder (VAE)-based detection of novel acoustic events. The VAE network is applied on the test log-Mel spectrogram, and the reconstruction error is computed and compared with threshold T_h to determine whether it comes from the known distribution

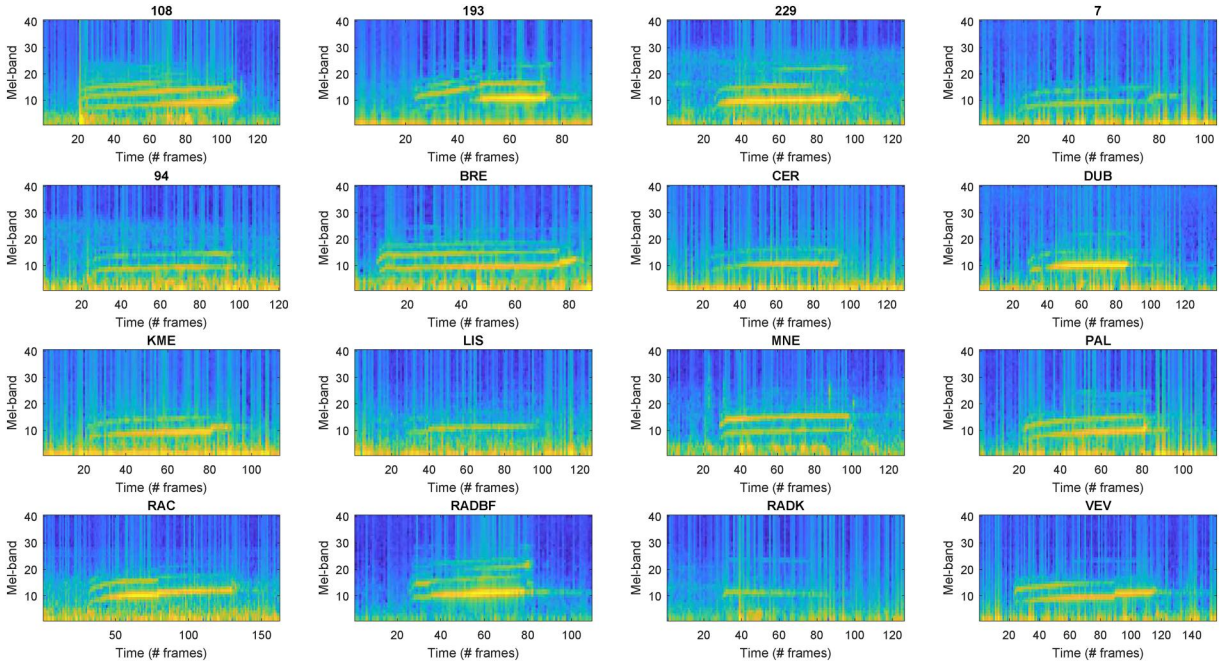

FIGURE 4 Log-Mel spectrograms of the considered individuals (little owl species)

(line 10, Algorithm 1). The reconstruction error is computed (line 11, Algorithm 1) and compared against T_h (lines 12–15, Algorithm 1). If the reconstruction error is larger than T_h, the algorithm signals the detection of a new species/individual. Conversely, a species/individual existing in $\mathcal{S}$ is included in y^t.

3.4 | Feature set

This section briefly describes the considered feature set, which is a simplification of the Mel-frequency cepstral coefficients wherein the final dimensionality reduction step based on the discrete cosine transform is omitted, as is typically performed in deep learning solutions that target audio signals [31,32].

Initially, the audio signal is windowed using the Hamming function, and the short-time Fourier transform (STFT) is computed for each frame. The outcome of the STFT passes through a triangular Mel-filterbank of 23 filters. Consecutively, we obtain the logarithm to adequately space the data and derive a vector of 23 log-energies per frame [33].

Representative log-Mel spectrograms extracted from the considered species and individuals are shown in Figures 3 and 4, respectively. In Figure 3, it is worth noting the inter-class variance in terms of frequency content (e.g. differences between American crow and house finch) and time evolution (e.g. differences between common yellowthroat and American yellow warbler). By contrast, log-Mel spectrograms at the individual level are not as diverse (see Figure 4) when examining both time and frequency content. From this point of view, we can expect that change detection in the first case would be easier because large differences are anticipated between the known and unknown species. At the same time, the second task is expected to be characterised by a higher degree of difficulty. Importantly, the use of such a standardised feature extraction mechanism removes the need to conceptualise and implement handcrafted features specifically designed to address a given problem.

4 | EXPERIMENTS

This section includes details about (a) the data sets employed, (b) the parameterisation of the proposed and contrasted approaches, and (c) the analysis of the obtained results.

4.1 | Data sets

To validate the proposed method, we employed two data sets satisfying the requirements presented in section 1. The first serves new species detection, as it includes the following 11 North American bird species: *bluejay*, *song sparrow*, *great blue heron*, *American crow*, *cedar waxwing*, *house finch*, *indigo bunting*, *marsh wren*, *common yellowthroat*, *chipping sparrow*, and *American yellow warbler*. There are 2762 bird acoustic events adequately distributed among the available bird species, with the audio signals sampled at 32 kHz. More information is available in [8,22], and the data set can be downloaded at https://zenodo.org/record/1250690#.XfOmzOhKhww.

The second data set serves new individual bird detection, as it encompasses 16 individuals of the Little owl (*Athene noctua*) species. It consists of 952 bird acoustic events evenly distributed among individuals, with the audio signals sampled at 44.1 kHz. More information is available in [23], and the data set can be downloaded at https://zenodo.org/record/1413495#.XfOnMehKhww.

4.2 | Figures of merit and parameterisation

To assess the performance of the proposed and contrasted approaches, we employed the following two figures of merit:

- *False positive index* (FPI): This index counts the times a test detects a nonexistent novel acoustic event (percentage).
- *False negative index* (FNI): This index counts the times an existing novel acoustic event is not detected as such (percentage).

The feature set was extracted using an STFT size of 512 samples, while the signals were windowised in frames of 30 ms overlapped by 50%.

The VAE encoder comprises three convolutional layer sizes [32, 64, and 64], while the decoder network is symmetric. The kernel size is 3 × 3 with a stride equal to 2. The rest of the parameters are (a) latent dimension, 40; (b) number of epochs, 50; (c) batch size, 10; and (d) learning rate, 0.001.

The first contrasted approach is based on an HMM trained, validated, and tested on identical sets of data. The explored number of states ranges from two to seven, while the Gaussian functions composing each state come from the following set: {2, 4, 8, 16, 32, 64}. The probability threshold between subsequent iterations of the Baum–Welch algorithm is 0.001 with a limit of 50 iterations. The combination providing the most accurate modelling in terms of log-likelihood was chosen [13]. The second comparison concerned a

TABLE 1 Mean and standard deviation of FPI and FNI offered by the proposed and contrasted approaches

	Species				Individuals			
Fig. of merit	***VAE***	***HMM***	***GMM***	***SVM***	***VAE***	***HMM***	***GMM***	***SVM***
FPI (%)	**1.6 (1.8)**	7.3 (2)	9.7 (2.2)	9 (2.4)	**2.5 (1)**	9.5 (2.1)	9.4 (2.1)	9 (2)
FNI (%)	**0.9 (0.4)**	4.5 (0.7)	6.6 (1.1)	5.9 (0.8)	**2.9 (1.9)**	8.1 (1.8)	8.7 (1.9)	8.2 (1.8)

Note: The lowest rates per task are shown in bold.

Abbreviations: FNI, false negative index; FPI, false positive index; GMM, Gaussian mixture model; HMM, hidden Markov model; SVM, support vector machine; VAE, variational auto-encoder.

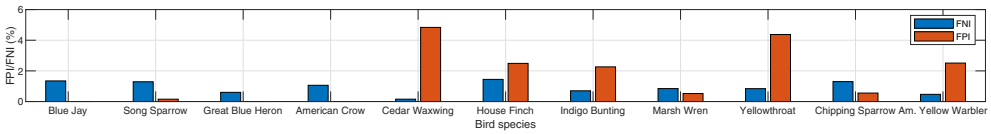

FIGURE 5 Detection results for each bird species following the leave-one-species-out learning paradigm

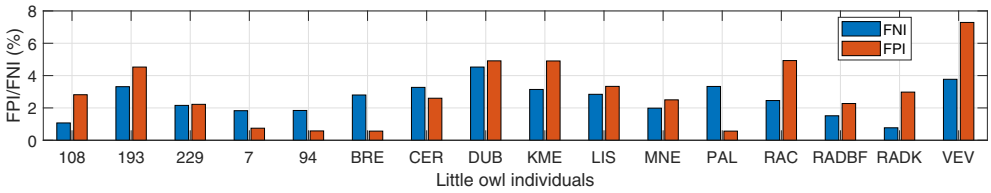

FIGURE 6 Detection results for individual birds following the leave-one-individual-out learning paradigm

Gaussian mixture model (GMM), and the third a one-class support vector machine (SVM) with a radial basis function kernel [34].

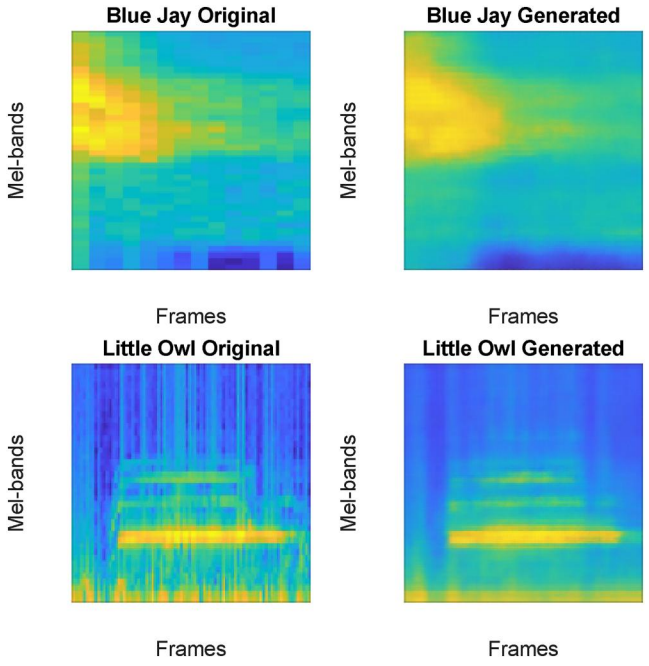

FIGURE 7 Two examples of original and generated log-Mel spectrograms of a bird species and individual

5 | RESULTS

We present the following experimental results: (a) comparison of HMM and SVM, (b) FPI and FNI rates per bird species/individual class, (c) log-Mel spectrogram generation examples, and (d) latent space visualisation. All results were obtained following the leave-one-species-out/leave-one-individual-out experimental protocol.

Table 1 tabulates the FPI and FNI rates achieved by VAE and HMM for bird species/individual detection. The best indices are shown in bold. As shown, the proposed method outperforms the contrasted one in both classes for both types of rates. VAE is able to encode and decode from a structured compact latent representation of the input log-Mel spectrograms providing low reconstruction error/high reconstruction probability. As such, it is able to provide reliable results in terms of FPI and FNI rates. At the same time, the HMM modelling the temporal evolution of the Mel-frequency cepstral coefficients cannot identify novel acoustic patterns with the same degree of accuracy. We also observe improved performance at the species level, which was expected because differences are more evident in that case than for individuals. Finally, the SVM offers rates that are slightly worse than those of the HMM, which may result from its inability to capture existing temporal patterns. However, the SVM performs better than the GMM-based solution.

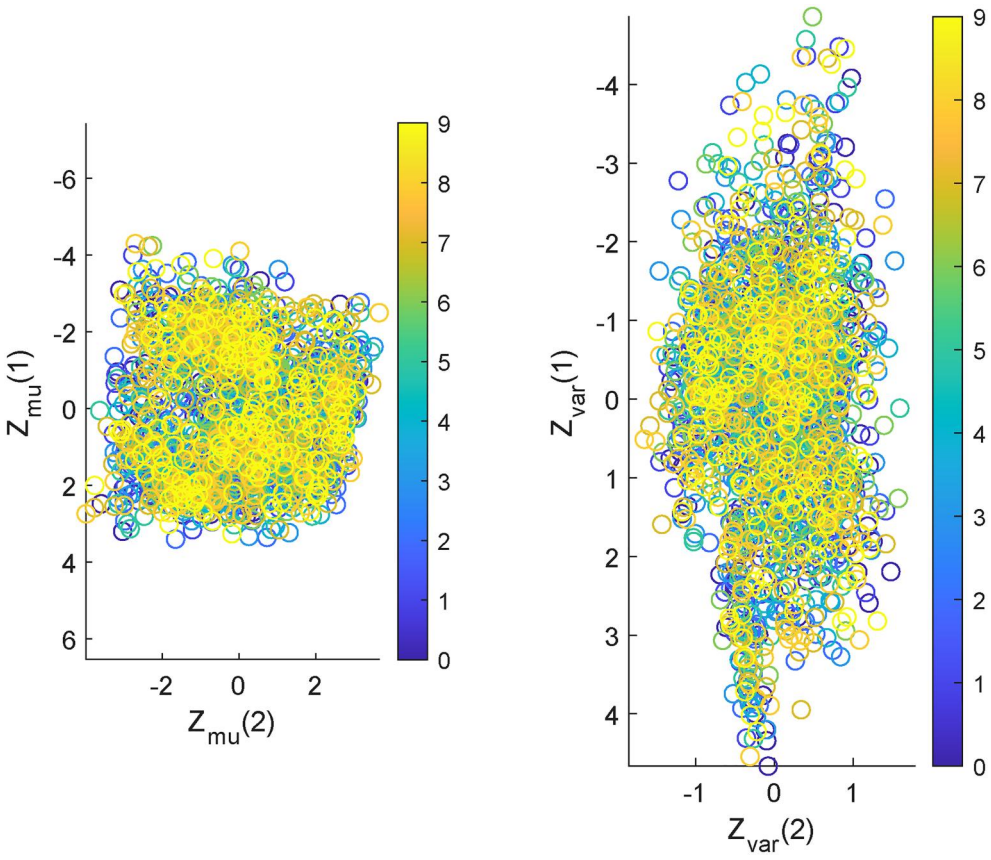

FIGURE 8 Visualization of the latent space learned by the VAE trainedon bird species vocalizations

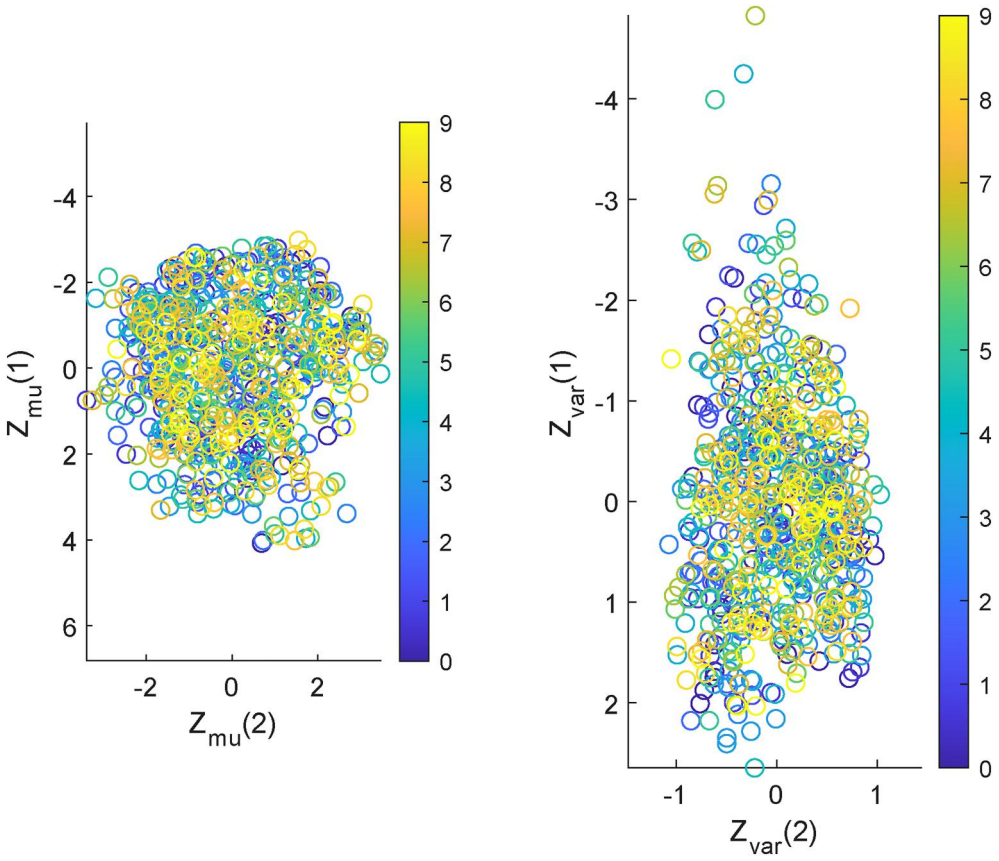

FIGURE 9 Visualization of the latent space learned by the VAE trained on individual bird vocalizations.

Figures 5 and 6 present the FPI and FNI rates achieved at the species and individual levels, respectively. In Figure 5, we observe a relatively large FPI for the *cedar waxwing*, which may be due to the similarities it exhibits to the rest of the North American bird species. At the same time, FNI rates show a more consistent behaviour across species. Finally, it is worth mentioning that the *great blue heron* is detected quite effectively with very low FPI and FNI rates.

Figure 6 depicts results at the individual level. We observer that individuals VEV, DUB, and KME are associated with high FPI rates. All individuals are acoustically similar in terms of what a human listener can assess. The remaining individuals are characterised by consistent FPI/FNI rates.

Subsequently, we visualised the latent space by capturing the mean and variance encodings (each with a dimension of 40) extracted by feeding the encoder network with test log-Mel spectrograms. Principal component analysis was carried out on the matrices comprising the encodings for the log-Mel spectrogram associated with each bird species and individual. Finally, the latent space defined by the means and variances in the first two principal component dimensions is visualised [35]. Figures 7 and 8 visualise the latent space produced for the corresponding cases of bird species and individuals. We see that both latent spaces offer a compact representation of the features associated with the classes of interest.

TABLE 2 List of notations

Symbol	Meaning
y_t	Audio signal
$\mathcal{S}$	Class dictionary
S_i	i-th class
P_i	i-th class pdf
t^*	Time instance of the manifestation of a new sound class
TS	Training set
VS	Validation set
M	Number of known classes
$p(\mathbf{z},\mathbf{x})$	Joint probability function
$p(\mathbf{z})$	Bayesian *prior* in the latent space
D	Kullback–Leibler divergence
θ, ϕ	Distribution parameters
$q(\mathbf{z}\|\mathbf{x})$	Generative model
$p(\mathbf{x}\|\mathbf{z})$	Inference model
$f_\theta(\mathbf{x};\theta)$	Encoder
$g_\phi(\mathbf{z};\phi)$	Decoder
$\mathbb{E}$	Reconstruction probability
F	Feature sequences
T_h	Detection threshold

Figure 9 demonstrates generated log-Mel spectrograms of test recordings and their corresponding original. The top row refers to the *bluejay* and the bottom to a *little owl* individual. We observe that the log-Mel space has been smoothed, providing a good basis for analysing reconstruction probabilities. In our future work, we intend to elaborate on the generated space in an effort to reverse them into the audio signal space while focusing on interpretable sound intelligibility.

A list of notations used in the data modelling algorithm is provided in Table 2.

6 | CONCLUSIONS

This work formalised the problem of detecting novel bird calls/songs by means of a VAE-based change detection algorithm. Log-Mel spectrograms of testing sounds are reconstructed by the proposed encoding–decoding networks, and their novelty is assessed in terms of reconstruction error. The superiority over the contrasted HMM-based change detection algorithm was proven on two use cases, detecting (1) bird species and (2) individuals.

Interestingly, the VAE latent space can be sampled and as such generate new bird vocalisations that come from the distribution exhibited by the training data. Currently, such generated vocalisations present perceptible artefacts. This would be the focus of our future work, that is, user-defined and adjustable generation of a realistic bird repertoire [36]. At the same time, we intend to analyse the present algorithms from the point of view of computational complexity with the aim to evaluate hardware requirements for real-time applications. Another fruitful path would be the exploration of few-shot learning techniques [37] so that new sound classes can be learned and incorporated in the class dictionary on the fly as soon as a change is detected.

ACKNOWLEDGMENTS

This work was carried out within the project automatIc aNalySis of comPlex evovlIng auditoRy scEnes (INSPIRE) funded by the Piano Sostegno alla Ricerca of University of Milan. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

ORCID

Stavros Ntalampiras https://orcid.org/0000-0003-3482-9215

REFERENCES

1. Stowell, D. et al.: Bird detection in audio: a survey and a challenge. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. Salerno, Italy (2016)
2. Digby, A. et al.: A practical comparison of manual and autonomous methods for acoustic monitoring. Methods Ecol. Evol. 4(7), 675–683 (2013) https://doi.org/10.1111/2041-210x.12060
3. Ntalampiras, S.: Bird species identification via transfer learning from music genres. Ecol. Inf. 44, 76–81 (2018) https://doi.org/10.1016/j.ecoinf.2018.01.006
4. Borker, A.L. et al.: Vocal activity as a low cost and scalable index of seabird colony size. Conserv. Biol. 28(4), 1100–1108 (2014) https://doi.org/10.1111/cobi.12264
5. Marques, T.A., et al.: Estimating animal population density using passive acoustics. Biol. Rev. 88(2), 287–309 (2012) https://doi.org/10.1111/brv.12001
6. Aravinda, S.P.P., Gunawardene, S., Kottege, N.: An acoustic wireless sensor network for remote monitoring of bird calls. In: 2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), pp. 1–4. Vienna, Austria (2016)
7. Florentin, J., Dutoit, T., Verlinden, O.: Detection and identification of european woodpeckers with deep convolutional neural networks. Ecol. Inf. 55, 101023 (2020) https://doi.org/10.1016/j.ecoinf.2019.101023
8. Zhao, Z., et al.: Automated bird acoustic event detection and robust species classification. Ecol. Inf. 39, 99–108 (2017) https://doi.org/10.1016/j.ecoinf.2017.04.003
9. Colautti, R.I., MacIsaac, H.J.: 'A neutral terminology to define 'invasive' species'. Divers. Distrib. 10(2), 135–141 (2004) https://doi.org/10.1111/j.1366-9516.2004.00061.x
10. Ehnes, M., Dech, J.P., Foote, J.R.: Seasonal changes in acoustic detection of forest birds. Journ. of Ecoacoust. 2, QVDZO7 (2018) https://doi.org/10.22261/jea.qvdzo7
11. Kułaga, K., Budka, M.: Bird species detection by an observer and an autonomous sound recorder in two different environments: forest and farmland. PLoS One. 14(2), e0211970 (2019) https://doi.org/10.1371/journal.pone.0211970
12. Ntalampiras, S., Potamitis, I., Fakotakis, N.: Probabilistic novelty detection for acoustic surveillance under real-world conditions. IEEE Trans. Multimed. 13(4), 713–719 (2011)
13. Ntalampiras, S.: Automatic analysis of audiostreams in the concept drift environment. In: 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. Salerno, Italy (2016)
14. Wiewel, F., Yang, B.: Continual learning for anomaly detection with variational autoencoder. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3837–3841. Brighton, United Kingdom (2019)
15. Pereira, J., Silveira, M.: Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1275–1282. Orlando, Florida, USA (2018)
16. Wang, X. et al.: adVAE: a self-adversarial variational autoencoder with Gaussian anomaly prior knowledge for anomaly detection. Knowl. Based Syst. 190, 105187 (2019) https://doi.org/10.1016/j.knosys.2019.105187
17. Koizumi, Y. et al.: 'Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma'. IEEE/ACM Trans. Audio Speech Language Process. 27(1), 212–224 (2019)
18. Kawachi, Y., Koizumi, Y., Harada, N.: Complementary set variational autoencoder for supervised anomaly detection. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2366–2370. Calgary, Canada (2018)
19. Vincent, P. et al.: Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res. 11, 3371–3408 (2010) http://dl.acm.org/citation.cfm?id=1756006.1953039
20. Kingma, D.P., Welling, M.: 'Auto-encoding variational bayes' (2013) http://arxiv.org/abs/1312.6114
21. Girin, L. et al.: Notes on the use of variational autoencoders for speech and audio spectrogram modeling. In: DAFx 2019 - 22nd International Conference on Digital Audio Effects, pp. 1–8. Birmingham (2019) https://hal.archives-ouvertes.fr/hal-02349385
22. Zhang, S. et al.: Automatic bird vocalization identification based on fusion of spectral pattern and texture features. In: 2018 IEEE

International Conference on acoustics, Speech and signal processing (ICASSP), pp. 271–275 (2018)
23. Stowell, D. et al.: Automatic acoustic identification of individuals in multiple species: improving identification across recording conditions. J. R. Soc. Interface. 16(153), 20180940 (2019) https://doi.org/10.1098/rsif.2018.0940
24. Jaakkola, T.S., Jordan, M.I.: Bayesian parameter estimation via variational methods. Stat. Comput. 10(1), 25–37 (2000) https://doi.org/10.1023/A:1008932416310
25. Bishop, C.M.: Pattern recognition and machine learning, Springer ed. Springer-Verlag New York, (2006)
26. Romeu, C. et al.: Cross-modal Variational Inference for bijective signal-symbol translation. In: DAFx 2019 - 22nd International Conference on Digital Audio Effects, pp. 1–8. Birmingham, UK (2019)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations, (ICLR) 2014, Banff, AB, Canada, April 14–16, 2014, Conference Track Proceedings (2014)
28. Xu, H., et al.: 'Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications'. In: Proceedings of the 2018 World Wide Web Conference. Geneva: International World Wide Web Conferences Steering Committee, 2018. pp. 187–196. https://doi.org/10.1145/3178876.3185996
29. Xu, H., et al.: Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. CoRR, Lyon, France (2018) https://arxiv.org/abs/1802.03903
30. Rezende, D.J., Mohamed, S., Wierstra, D.: Stochastic backpropagation and approximate inference in deep generative models. In: Xing, E.P., Jebara, T. (eds.), Proceedings of the 31st International Conference on Machine Learning. vol. 32 of Proceedings of Machine Learning Research, pp. 1278–1286. PMLR, Bejing (2014) http://proceedings.mlr.press/v32/rezende14.html
31. Ntalampiras, S.: Automatic acoustic classification of insect species based on directed acyclic graphs. J. Acoust. Soc. Am. 145(6), EL541–EL546 (2019) https://doi.org/10.1121/1.5111975
32. Ntalampiras, S.: Hybrid framework for categorising sounds of mysticete whales. IET Signal Process. 11(4), 349–355 (2017)
33. Ntalampiras, S.: 'Deep learning of attitude in children's emotional speech'. In: 2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), pp. 1–5. Tunis, Tunisia (2020)
34. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 271–2727 (2011) software http://www.csie.ntu.edu.tw/?cjlin/libsvm
35. Tokozume, Y., Ushiku, Y., Harada, T.: Learning from between-class examples for deep sound recognition. CoRR, Vancouver, Canada (2017) http://arxiv.org/abs/1711.10282
36. Kumar, K., et al: Enleadertwodots 'Melgan: generative adversarial networks for conditional waveform synthesis' (2019) http://arxiv.org/abs/1910.06711
37. Schönfeld, E. et al.: Generalized zero- and few-shot learning via aligned variational autoencoders. CoRR, Long Beach, California (2018) http://arxiv.org/abs/1812.01784

How to cite this article: Ntalampiras S, Potamitis I. Acoustic detection of unknown bird species and individuals. *CAAI Trans. Intell. Technol.* 2021;1–10. https://doi.org/10.1049/cit2.12007