Project Title

Optional Subtitle

Candidate Number: YPDR0[1]

MSc Data Science & Machine Learning

Supervisor: Dr. Ifat Yasin

Submission date: 11 September 2023

---

[1]**Disclaimer:** This report is submitted as part requirement for the MSc Data Science & Machine Learning at UCL. It is substantially the result of my own work except where explicitly indicated in the text. The report may be freely copied and distributed provided the source is explicitly acknowledged

Abstract

Summarise your report concisely.

Contents

Chapter 1

Introduction

Birds play a crucial role in ecosystems around the world. They form an important link in the food chain, pollinate plants, and even plant trees [5]. The quantity and diversity of birds observed in an area can therefore be seen as a key indicator of the strength of its ecosystem.

Aside from being important ecological agents, birds also colour our lives with their sights, sounds and behaviours. The community of birdwatchers, known colloquially in the U.K. as 'birders' or 'twitchers', has grown considerably over the past few years. The Royal Society for the Protection of Birds (RSPB) reported a 70% increase in their website views over the first lockdown, with more than 50% of those views on pages looking at bird identification. The Bird Bird Garden Watch, an annual event which encourages people to note bird sightings in their own residences, brought in 1 million participants in January 2021, more than double the previous year's tally. As such there is a growing commercial demand for accurate and easily accessible bird identification tools.

Birds are typically shy and protective creatures and tend to reside out of harm's way in shrubs, trees and nests, and therefore are usually heard but not seen. The majority of their communication is done through distinctive vocalizations, known as birdsong, that are usually unique to an individual species. The consequence of this is that birds are typically identified through their birdsong rather than their visual sighting. Birdsong used for identification is usually recorded on microphones that may be running for a long time and/or located in a place that isn't necessarily optimum to capture the vocalization. The recordings may be corrupted from a wide range of sources, such as ambient background noise, large changes in birdsong amplitude, long periods of silence, and vocalizations from other birds. Recordings captured by long running microphones may be several hours in

duration so it would be impractical to have a human expert identify the birds manually. There is therefore a need for robust birdsong identification tools that can handle these corruptions and that require minimal human intervention in order to classify unknown birdsong.

As with most machine learning classification problems, the work boils down to two key components. The first is feature extraction, that is, given an input signal usually in the form of an amplitude varying over time, how can it be transformed to a vector or matrix representation that captures the key information of the signal. The second is classification, i.e. given this representation in the feature space, how can it be used to train a model that can then perform classification on unseen samples of birdsong. There are further steps that can be taken to improve this process, such as pre-processing the signal in order to remove or reduce the influence of background noise [31].

There has been numerous research into audio classification problems both in birdsong and in other types of audio, such as speaker identification from human speech [19]. However, there remain novel methods that have been tested and have been shown to have good results in wider audio classification problems that have yet to be tried out in birdsong identification problems. These methods include both feature extraction and classification. The broad aim of this thesis is to attempt to evaluate these methods. In more detail, the aims of this thesis can be summarised as follows:

1. Explore the birdsong classification performance of feature representations shown to have promising results for non-birdsong related audio classification problems using simple classification models such as Support Vector Machines (SVM). The hypothesis is that feature representations shown to have good results in audio classification problems will have good results for birdsong identification.

2. Explore some of the hyperparameters available during the feature extraction process. The hypothesis is that using hyperparameter values more suited to birdsong classification problems will yield a higher classification accuracy.

3. Explore the performance of deep learning architectures such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) and compare the results with simpler statistical models. The hypothesis is that more complex and flexible architectures such as these will have superior performance when compared to simpler statistical models, such as SVMs.

Chapter 2

Background & related work

Describe here work that is connected to your thesis. This should include references to published work. There is no fixed rule, but I would expect a student to have read around 50 published research papers and reference them in a thesis.

2.1 Evaluation metrics

For multiclass classification problems, e.g. identifying a sample as a particular bird species from a last of $n > 2$ species, a simple classification accuracy is often used as an evaluation metric ([7], [32]). This is calculated as the percentage of correctly identified samples from a set of labelled samples previously unseen by the model.

Acevedo et al [1] used true positive (TP) and false positive (FP) rates to evaluate their models used to classify bird and amphibian calls. This evaluation method is sometimes preferred when analysing long recordings which may include several different species to be classified. TP gives an indication as to how well the model correctly identifies species present in the recording. FP indicates how often the model erroneously identifies species absent in the recording. A high performing model will have high values for TP and low values for FP.

Potamitis et al [31] used an alternative form of evaluation presented in terms of precision (P) and recall (R) which are defined as

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.1}$$

where FN is the number of false negatives. Informally, P and R can be thought of as

$$P = \frac{\text{relevant retrieved instances}}{\text{all \textbf{retrieved} instances}}, \quad R = \frac{\text{relevant retrieved instances}}{\text{all \textbf{relevant} instances}} \qquad (2.2)$$

It's well known that there is usually a trade-off between P and R, so usually the F-score is reported along with the P and R metrics. The F-score is defined as

$$F = \frac{(1 + \beta^2)PR}{\beta^2 P + R} \qquad (2.3)$$

for some $\beta \in \mathbb{R}$.

For binary classification problems, i.e. classifying a sample as one of two classes (bird species), the Area Under Curve (AUC) metric is sometimes preferred [23]. The AUC ranges from 0 to 1, with a higher value indicating a better performing model. The AUC is sometimes desirable because it is scale-invariant (it measures how well predictions are ranked, rather than their absolute values) and classification-threshold-invariant (it measures the quality of the model's predictions regardless of what classification threshold is chosen.)

2.2 Feature extraction

As mentioned in Chapter 1, birds use vocalizations as a way to communicate with others. Birds have evolved over many thousands of years to make this form of communication very efficient in that it can travel long distances and cut through local ambient noise frequencies so that it can be received by other birds clearly. Some birds have even adapted their vocalizations so that they can be heard in local environments that have rapidly changed over the past decades, such as urban areas with increasing anthropogenic noise [25].

Bird vocalizations can be broadly divided into two main categories, calls and songs. Calls are typically short vocalizations that carry some specific function, such as warning others to the presence of a predator or calling others to flight [26]. Songs are usually longer and more acoustically complex and occur more spontaneously. Songs are typically employed as breeding calls or territorial defence. While all birds produce calls, in many species of birds only the males utilise songs and often only during breeding season. The song and call of a Eurasian Wren can be seen and contrasted in figure 2.1.

Doupe et al [12] showed that there were striking similarities between birdsong and human language. Similar to human language, birdsong can be thought of as comprised of hierarchical levels of phrases, syllables, and elements [6]. ¡Figure here¿. Raw recordings

of birdsong often contain periods of silence that occur between phrases which are unlikely to provide useful information in order to train a machine learning model. Fagerlund [15] showed that a good level of accuracy can be achieved by training models on features extracted from segmented syllables which were used as training samples. An algorithm for robust syllable segmentation was proposed in [14] and is used in this thesis. The algorithm is described in more detail in Section.
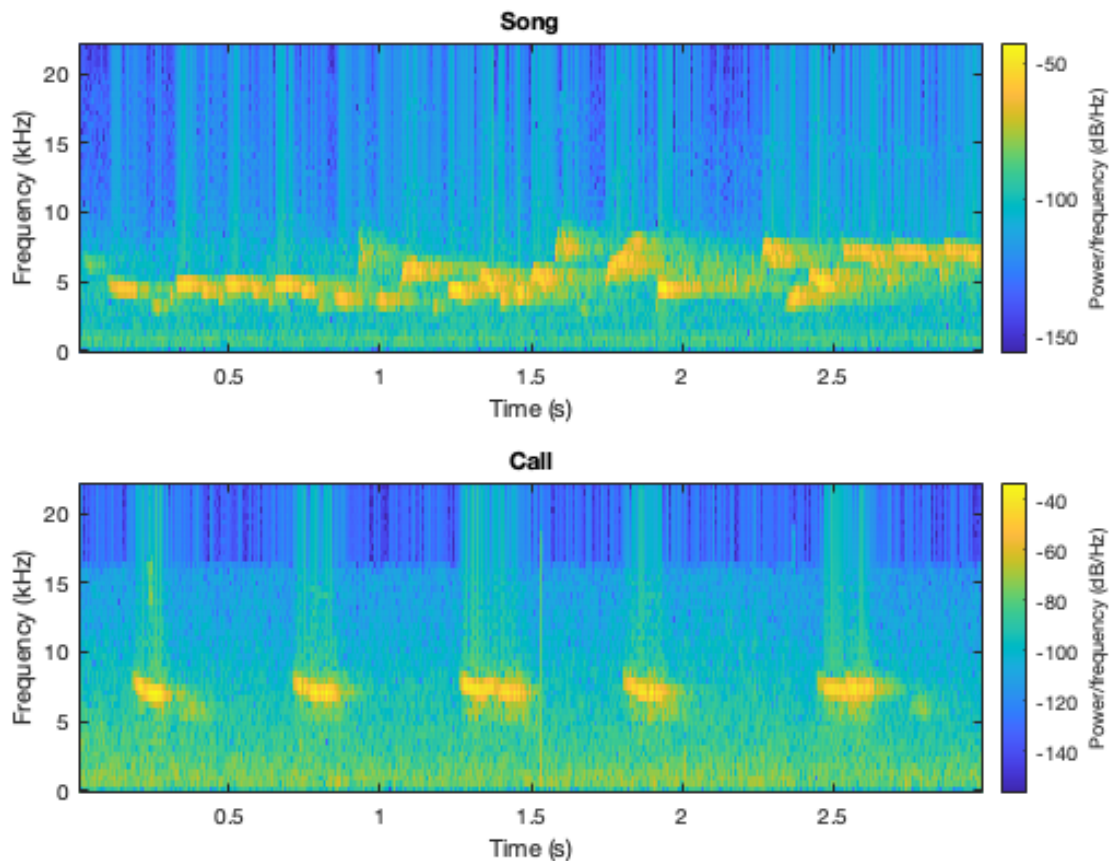

Figure 2.1: Spectrograms of the call and song of the Eurasian Wren (*Troglodytes troglodytes*). As can be seen, the song is much more complex in terms of pitch and acoustic structure compared to the call

Once the syllables have been segmented, there exists an abundance of options to turn the raw syllable input signal into features that might be used as training samples. In the following sections some of the more popular feature representations are described, along with some novel methods that have yet to be tried on birdsong. Note that the following list is certainly not exhaustive and emphasis has been placed on feature representations

that are relevant to this thesis.

2.2.1 Features inspired by the human auditory system

Research has shown that birdsong and human language share many similarities in terms of the neural mechanisms employed to form the language/song, the impact of social contact in learning the language/song, and so on [12]. Give that the human auditory system has evolved other thousands of years to best process human speech, it seems reasonable to suggest that using features based on the human auditory system may be effective when it comes to birdsong.

At a high level, the human auditory system works by translating changes in air pressure originating from a source and reaching the outer ear of a listener into vibrations that travel along an internal organ known as a cochlear. The vibrations trigger electric signals that move along auditory nerves to the brain, where they are interpreted as sounds. The physiology of the cochlear means that certain parts of the organ are more sensitive to certain frequencies of vibrations, so different frequencies will lead to different electrical signals moving to the brain. This allows the brain to learn to differentiate between frequencies.

Mel frequency cepstrum coefficients

The makeup of the human auditory system means that humans have a greater ability to differentiate pitch at lower frequencies then they do at higher frequencies. In other words, humans perceive pitch non-linearly. This has led to the development of a logarithmic scale, known as the mel scale, such that equal distances on the scale have the same *perceptual* distance.

The mel frequency cepstrum coefficients (MFCC) are part of a family of cepstrum coefficients that capture information about the rate of change in different spectrum bands. MFCC differs from other cepstrum coefficients in that it uses the mel scale to transform the spectrum of an input signal, thus utilising the human auditory system in the calculation of its coefficients.

The i-th mel cepstral coefficient is computed as [9]

$$\text{MFCC}_i = \sum_{k=1}^{K} X_k \cos\left(\frac{i(k-0.5)\pi}{K}\right) \tag{2.4}$$

where X_k is the logarithmic energy of the k-th mel-spectrum band, and K is the total

number of the mel-spectrum bands. Usually 8–13 MFCC coefficients are used as the feature vector representing one time frame of the signal. The 0th coefficient is often excluded as it represents the average log energy of the signal and is unlikely to carry any relevant information to help with classification.

MFCCs are often presented with their delta (Δ) and double-delta ($\Delta\Delta$) values that capture the local temporal dynamics and temporal changes of the delta values respectively.

MFCCs are used as feature representations since they are simple to compute and have been shown to have good performance in a wide range of audio classification tasks, such as speaker identification [28] and emotion recognition [24]. MFCCs have also been shown to lead to good classification accuracy for birdsong identification problems ([15] and [32]).

¡include graph of filterbank¿

Gammatone cepstrum coefficients

Gammatone cepstrum coefficients (GTCCs) are similar to MFCCs except their calculation uses gammatone filterbanks instead of mel scale filterbanks. Gammatone filterbanks are designed to simulate the motion of the membrane inside the cochlear, known as the basilar membrane, when it is exposed to vibrations transmitted by the outer and middle ear [30].

A gammatone filter with a centre frequency f_c is defined as

$$g(t) = at^{n-1}e^{-2\pi bt}\cos(2\pi f_c + \psi) \tag{2.5}$$

where t is the time in seconds, ψ is the phase in radians (usually set to 0), $a \in \mathbb{R}$ controls the gain, n is the filter's order and b is the filter's bandwidth in Hz. To simulate the human auditory system, the centre frequencies are uniformly spaced on the equivalent rectangular band-width (ERB) scale.

Similar to MFCCs, GTCCs are typically used with their Δ and $\Delta\Delta$ values and have also been shown to have a good performance in non-speech audio classification problems [36].

¡include graph of filterbank¿

Multi resolution cochleagram

Chen et al [8] proposed a novel feature known as a multi-resolution cochleagram (MRCG) that is formed by combining four cochleagrams at different resolutions, designed to capture both local and contextual information.

A cochleagram is formed by passing an input signal through a gammatone filterbank,

similar to the first steps of the formation of the GTCCs. Each response signal from the gammatone filterbank is then divided into frames of a certain length with a certain overlap or frame shift. The cochleagram is then generated by calculating the power of each frame at each channel.

¡example cochleagram¿

A MRCG is formed by first taking one cochleagram with a frame length of 20ms and frame shift of 10ms, using a gammatone filterbank with 64 output channels. A log operation is applied to each time-frequency (T-F) unit of the output cochleagram, denoted CG1. CG2 is calculated in the same way, but with a frame length of 200ms and the same frame shift. CG3 is calculated by averaging CG1 across a square window of 11 frequency channels and 11 time frames centred at the given T-F unit. Zero padding is used here. CG4 is calculated in the same way as CG3, but with a 23×23 square window. CG1–4 are then stacked vertically to obtain the full MRCG feature, which will be a $256 \times p$ matrix, where p is the number of time frames.

¡example mrcg¿

The MRCG feature has been shown to outperform both MFCC and GTCC at separating human speech at various SNR levels with different types of background noise [8]. Abdullah et al [4] showed that the MRCG outperforms the Auditory Image Model (AIM) when classifying noise. Wang et al [37] showed that improved speech enhancement for hearing impaired listeners was achieved when the MRCG feature was added to the feature set.

One potential drawback when comparing the MRCG with MFCC and GTCC is the higher dimensionality of the MRCG feature. When including the Δ and $\Delta\Delta$ features, as is standard practice ([4], [37]), the MRCG feature vector for a single time frame has a dimensionality of 768. For MFCC or GTCC the equivalent dimensionality is 24–39, depending on the number of the coefficients used.

2.2.2 Feature stacking

It has been shown that combining features to be used as training inputs can lead to improved classification accuracy. Ramashini et al [32] showed that a combination of GTCC and MFCC yielded higher birdsong classification accuracy than MFCC alone. However, the combination did not improve on the accuracy of GTCC alone. Yan et al [38] demonstrated that a combination of MFCC with two other feature types (Log-mel spectrogram and Chroma) yielded higher birdsong classification accuracy than combinations of two of the features alone. Fagerlund [15] showed that combining MFCC with spectral and tem-

9

poral features, such as frequency range and zero crossing rate, can lead to higher birdsong classification accuracies.

A consequence of feature stacking is that feature vectors will have increased size. This can lead to problems such as longer training times and overfitting, especially when used in conjunction with deep learning (DL) architectures. As a result, dimensionality reduction techniques, such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) [33], are often used to lower the dimensionality of the feature vectors while still capturing enough information to be able to train a model effectively.

2.3   Classification

Once a suitable feature representation has been selected and performed, the remaining task is to pick an appropriate model and train the model using the set of training samples generated from the feature extraction.

There exist many different options in terms of models, and no one model is superior to the others. Each problem must be considered and a model which fits the problem at hand must be selected, or indeed a selection of suitable models tested and the evaluated for performance.

In the literature, birdsong classification has been tackled with a wide variety with models. Ramashini et al [33] used the simple Nearest Centroid (NC) classifier to assign birdsong samples to classes. It was compared to a K Nearest Neighbours (KNN) classifier and shown to have superior accuracy. Lasseck [21] used decision trees with bagging to identify birds. Trifa et al. [35] used Hidden Markov Models (HMM) to identify species of birds based on their vocalizations. They considered each sample as a discrete-time dynamical system, where an unobserved state generates observed features, such as pitch and MFCC. A different HMM can be learned for each class and a new sample can be assigned to a class by calculating which HMM gives the highest likelihood of the observed features. Kwan et al. [20] used a Gaussian Mixture Model (GMM) to classify birds. In their experiments, a Gaussian was learned for each class of bird. A new sample was then classified according to whichever class best describes the new sample.

The following sections describe some models that are relevant to this thesis in more detail

2.3.1 SVMs

SVMs are widely used in machine learning applications as a classification tool. They are well-established due to their high accuracy results [15] and relative simplicity to implement. Most implementations of SVMs also require little or no tuning of hyperparameters.

At its core, a SVM is a binary classifier that separates two classes by finding a hyperplane that maximizes the margin from the nearest vectors in the feature space from both classes. Classifications are then made by computing in which side of the hyperplane a test feature vector lies.

Binary classification

Let $\mathbf{x}_i \in \mathbb{R}^m$ be a feature vector of dimensionality m. Let $y_i \in \{+1, -1\}$ be its class label. For linearly separable data, the separating hyperplane satisfies

$$y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) \geq 1, \quad i = 1, \ldots, n \tag{2.6}$$

where $\mathbf{w} \in \mathbb{R}^m$ is a vector of weights and $b \in \mathbb{R}$ is a bias term. The margin between the hyperplane and the nearest feature vectors from each class is given by

$$\mathrm{d}(\mathbf{w}, b) = \min_{\mathbf{x}_i, y_i = 1} \frac{|\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} + \min_{\mathbf{x}_j, y_i = -1} \frac{|\langle \mathbf{w} \cdot \mathbf{x}_j \rangle + b|}{\|\mathbf{w}\|} \tag{2.7}$$

$$= \frac{2}{\|\mathbf{w}\|}. \tag{2.8}$$

The optimal hyperplane can now be found by maximizing (2.8) subject to (2.6). This can be solved using Lagrange multipliers.

Often with real world data, classes are not linearly separable. This can remedied by projecting the data using a nonlinear mapping into a new feature space where the data are linearly separable. Equation (2.6) can then be re-written as

$$y_i \left(\langle \mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}_i) \rangle + b \right) \geq 1, \quad i = 1, \ldots, n \tag{2.9}$$

where $\mathbf{\Phi}$ is the nonlinear mapping. Instead of explicitly finding $\mathbf{\Phi}$, (2.9) can be re-written in dual form

$$y_i \left(\sum_{j=1}^{l} \alpha_j y_j \langle \mathbf{\Phi}(\mathbf{x}_j) \cdot \mathbf{\Phi}(\mathbf{x}_i) \rangle \right) + b \geq 1, \quad i = 1, \ldots, n \tag{2.10}$$

and replacing the inner product with a kernel function $K(\mathbf{x}_j, \mathbf{x}_i) = \langle \mathbf{\Phi}(\mathbf{x}_j) \cdot \mathbf{\Phi}(\mathbf{x}_i) \rangle$.

In practice, a hyperplane that separates the classes perfectly may suffer from poor generalization ability. To improve this, nonnegative slack variables ξ_i are introduced to (2.9) to allow for some missclassification of training samples in order to improve generalization ability. The slack variables are introduced like so

$$y_i \left(\langle \mathbf{w} \cdot \mathbf{\Phi}(\mathbf{x}_i) \rangle + b \right) \geq 1 - \xi_i, \quad i = 1, \dots, n \tag{2.11}$$

The amount of regularization is controlled by the constant C such that the maximization problem (2.8) becomes

$$\frac{2}{\|\mathbf{w}\|} - C \sum_{i=1}^{n} \xi_i \tag{2.12}$$

Therefore for large values of C the classifier will behave like a hard-margin SVM and attempt to perfectly classify all training samples.

The last remaining question is around what determines a valid kernel function. A function in the input space is a kernel function if its kernel matrix $\mathbf{K} = [K(\mathbf{x}_j, \mathbf{x}_i)]_{i,j=1}^{n}$ is positive semidefinite. Typical kernel functions include

- linear: $K(\mathbf{x}_j, \mathbf{x}_i) = \langle \mathbf{x}_j, \mathbf{x}_i \rangle$

- polynomial: $(\langle \mathbf{x}_j, \mathbf{x}_i \rangle + c)^p$ for some $c \in \mathbb{R}$

- Gaussian or RBF: $\exp\left(-\gamma \|\mathbf{x}_j - \mathbf{x}_i\|^2\right)$ for some $\gamma \in \mathbb{R}, \gamma > 0$.

2.3.2 Neural Networks

Although neural networks have existed in some form since the 1950s with inception of the perceptron algorithm [34], they have experienced a huge surge in popularity over the past decade or so. This has largely been due to remarkable progress being made thanks to deep neural networks in fields such as image classification [18] and language models [27].

Another reason for DL's increasing popularity in recent years is due to the availability of more performant hardware and software. The 'deep' aspect to deep learning refers to the fact that the software architecture depends on large amounts of parameters and needs massive amounts of training data in order to learn. In order to run training routines in a reasonable time and store large amounts of data in memory, access to powerful hardware and/or machine parallelism tools can be an essential part of the process.

The progress of audio problems related has also been accelerated due to deep neural networks. Hinton et al. [17] showed improved performance in speech recognition tasks using feed-forward neural networks when compared to more classical approaches like HMMs and GMMs. Speech enhancement [2] and speech separation [13] problems leveraging DL paradigms have also been shown to outperform more classical statistical models.

The world of birdsong classification has also benefited from DL, but perhaps to a lesser extent than other audio related problems so far. Approaches have mainly focused on Convolutional Neural Networks (CNNs), applying convolutional layers to an image representation of an input signal, such as a (Mel-)spectrogram ([3], [29]). Further birdsong classification solutions have approached the problem in a similar fashion, but utilised transfer learning to fine-tune an existing model ([11], [22]), such as ResNet [16] and ImageNet [10].

Disabato et al. [11] went a step further with their research in that they considered the computational and memory demands of their proposed bird detection DL network, ToucaNet. As mentioned earlier, DL applications can be extremely demanding in terms of computational resources which can make them unsuitable for running on devices with limited resources, such as smartphones and autonomous recording units (ARUs). Their research showed that a level of accuracy in line with the literature can be achieved but with lower computational complexity and memory demands.

Although the architecture of DL tools for birdsong classification may vary widely, the final layer is typically the same. This consists of a fully connected layer of n units, where n is the number of classes, with a softmax activation so that the output of the network can be considered as a probability distribution over the classes.

Feedforward Neural Network

Convolutional Neural Network

Recurrent Neural Network

Chapter 3

Methodology

Describe your method in detail and with great clarity, distinguishing it from other works (if it is indeed a novel idea). It is very important to clearly motivate your method.

Describe the results of your method here in this chapter.

Chapter 4

Extensions of methodology

It is unlikely that everything you tried worked well, so in this chapter you may wish to describe a modified version of your method and the associated results. Explain why you were motivated to try this extension and how you think it might help to address some of the shortcomings of the method is Chapter 3.

Chapter 5

Conclusion

Summarise what you have achieved and evaluate honestly if you feel the approach has been largely successful. Explain what could be improved still and perhaps why the method is not working well (if that is the case).

Bibliography

[1] Miguel A Acevedo, Carlos J Corrada-Bravo, Héctor Corrada-Bravo, Luis J Villanueva-Rivera, and T Mitchell Aide. Automated classification of bird and amphibian calls using machine learning: A comparison of methods. *Ecological Informatics*, 4(4):206–214, 2009.

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. *arXiv preprint arXiv:1804.04121*, 2018.

[3] Franz Berger, William Freillinger, Paul Primus, and Wolfgang Reisinger. Bird audio detection-dcase 2018. *DCASE2018 Challenge, Tech. Rep.*, 2018.

[4] S Binti Abdullah, Andreas Demosthenous, and Ifat Yasin. Comparison of auditory-inspired models using machine-learning for noise classification. In *International Journal of Simulation Systems, Science & Technology Special Issue: Conference Procedings UKSim2020, 25 to 27 March 2020*. United Kingdom Simulation Society, 2020.

[5] Richard K Broughton, James M Bullock, Charles George, Ross A Hill, Shelley A Hinsley, Marta Maziarz, Markus Melin, J Owen Mountford, Tim H Sparks, and Richard F Pywell. Long-term woodland restoration on lowland farmland through passive rewilding. *PloS one*, 16(6):e0252466, 2021.

[6] Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003.

[7] Deep Chakraborty, Paawan Mukker, Padmanabhan Rajan, and Aroor Dinesh Dileep. Bird call identification using dynamic kernel based support vector machines and deep neural networks. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 280–285. IEEE, 2016.

[8] Jitong Chen, Yuxuan Wang, and DeLiang Wang. A feature study for classification-based speech separation at low signal-to-noise ratios. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1993–2002, 2014.

[9] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[11] Simone Disabato, Giuseppe Canonaco, Paul G Flikkema, Manuel Roveri, and Cesare Alippi. Birdsong detection at the edge with deep learning. In *2021 IEEE International Conference on Smart Computing (SMARTCOMP)*, pages 9–16. IEEE, 2021.

[12] Allison J. Doupe and Patricia K. Kuhl. Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience*, 22(1):567–631, 1999. PMID: 10202549.

[13] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[14] Seppo Fagerlund. Automatic recognition of bird species by their sounds. *Finlandia: Helsinki University Of Technology*, 2004.

[15] Seppo Fagerlund. Bird species recognition using support vector machines. *EURASIP Journal on Advances in Signal Processing*, 2007:1–8, 2007.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[19] Mineichi Kudo, Jun Toyama, and Masaru Shimbo. Multidimensional curve classification using passing-through regions. *Pattern Recognition Letters*, 20(11):1103–1111, 1999.

[20] Chiman Kwan, KC Ho, Gang Mei, Yunhong Li, Zhubing Ren, Roger Xu, Y Zhang, Debang Lao, M Stevenson, Vincent Stanford, et al. An automated acoustic system to monitor and classify birds. *EURASIP Journal on Advances in Signal Processing*, 2006:1–19, 2006.

[21] Mario Lasseck. Improved automatic bird identification through decision tree based feature selection and bagging. *CLEF (working notes)*, 1391, 2015.

[22] Mario Lasseck. Acoustic bird detection with deep convolutional neural networks. In *DCASE*, pages 143–147, 2018.

[23] Yi Ren Leng and Huy Dat Tran. Multi-label bird classification using an ensemble classifier with simple features. In *Signal and information processing association annual summit and conference (APSIPA), 2014 Asia-Pacific*, pages 1–5. IEEE, 2014.

[24] MS Likitha, Sri Raksha R Gupta, K Hasitha, and A Upendra Raju. Speech based human emotion recognition using mfcc. In *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*, pages 2257–2260. IEEE, 2017.

[25] David Luther and Luis Baptista. Urban noise and the cultural evolution of bird songs. *Proceedings of the Royal Society B: Biological Sciences*, 277(1680):469–473, 2010.

[26] PETER MARLER. Chapter 5 - bird calls: a cornucopia for communication. In Peter Marler and Hans Slabbekoorn, editors, *Nature's Music*, pages 132–177. Academic Press, San Diego, 2004.

[27] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[28] Lindasalwa Muda, Mumtaj Begam, and Irraivan Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques. *arXiv preprint arXiv:1003.4083*, 2010.

[29] Rajdeep Mukherjee, Dipyaman Banerjee, Kuntal Dey, and Niloy Ganguly. Convolutional recurrent neural network based bird audio detection. *DCASE challenge*, 2018.

[30] Roy D Patterson, KEN Robinson, John Holdsworth, Denis McKeown, C Zhang, and Michael Allerhand. Complex sounds and auditory images. In *Auditory physiology and perception*, pages 429–446. Elsevier, 1992.

[31] Ilyas Potamitis, Stavros Ntalampiras, Olaf Jahn, and Klaus Riede. Automatic bird sound detection in long real-field recordings: Applications and tools. *Applied Acoustics*, 80:1–9, 2014.

[32] Murugaiya Ramashini, Pg Emeroylariffion Abas, Kusuma Mohanchandra, and Liyanage C De Silva. Robust cepstral feature for bird sound classification. *International Journal of Electrical and Computer Engineering*, 12(2):1477, 2022.

[33] Murugiaya Ramashini, Pg Emeroylariffion Abas, Ulmar Grafe, and Liyanage C De Silva. Bird sounds classification using linear discriminant analysis. In *2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pages 1–6. IEEE, 2019.

[34] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[35] Vlad M Trifa, Alexander NG Kirschel, Charles E Taylor, and Edgar E Vallejo. Automated species recognition of antbirds in a mexican rainforest using hidden markov models. *The Journal of the Acoustical Society of America*, 123(4):2424–2431, 2008.

[36] Xavier Valero and Francesc Alias. Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification. *IEEE transactions on multimedia*, 14(6):1684–1689, 2012.

[37] Zhong-Qiu Wang and DeLiang Wang. A joint training framework for robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):796–806, 2016.

[38] Na Yan, Aibin Chen, Guoxiong Zhou, Zhiqiang Zhang, Xiangyong Liu, Jianwu Wang, Zhihua Liu, and Wenjie Chen. Birdsong classification based on multi-feature fusion. *Multimedia Tools and Applications*, 80:36529–36547, 2021.