FAST MOSQUITO ACOUSTIC DETECTION WITH FIELD CUP RECORDINGS: AN INITIAL INVESTIGATION

Yunpeng Li[1], Ivan Kiskin[1], Marianne Sinka[2], Davide Zilli[1,3], Henry Chan[1], Eva Herreros-Moya[2], Theeraphap Chareonviriyaphap[4], Rungarun Tisgratog[4], Kathy Willis[2,5], Stephen Roberts[1,3]

[1] Machine Learning Research Group, Department of Engineering Science, University of Oxford, UK
{yli,ikiskin,dzilli,sjrob}@robots.ox.ac.uk, tsunhenry@gmail.com
[2] Department of Zoology, University of Oxford, UK
{marianne.sinka,eva.herreros-moya,kathy.willis}@zoo.ox.ac.uk
[3] Mind Foundry Ltd., UK
[4] Department of Entomology, Faculty of Agriculture, Kasetsart University, Bangkok, Thailand
aasthc@ku.ac.th, rungarun_tis@hotmail.com
[5] Royal Botanic Gardens, Kew, UK

ABSTRACT

In terms of vectoring disease, mosquitoes are the world's deadliest. A fast and efficient mosquito survey tool is crucial for vectored disease intervention programmes to reduce mosquito-induced deaths. Standard mosquito sampling techniques, such as human landing catches, are time consuming, expensive and can put the collectors at risk of diseases. Mosquito acoustic detection aims to provide a cost-effective automated detection tool, based on mosquitoes' characteristic flight tones. We propose a simple, yet highly effective, classification pipeline based on the mel-frequency spectrum allied with convolutional neural networks. This detection pipeline is computationally efficient in not only detecting mosquitoes, but also in classifying species. Many previous assessments of mosquito acoustic detection techniques have relied only upon lab recordings of mosquito colonies. We illustrate in this paper our proposed algorithm's performance over an extensive dataset, consisting of cup recordings of more than 1000 mosquito individuals from 6 species captured in field studies in Thailand.

***Index Terms*—** Mosquito detection, acoustic signal processing, multi-species classification, convolutional neural networks

1. INTRODUCTION

Malaria results in half a million deaths each year and mosquitoes are the only vector for malaria [1]. Among more than 3500 mosquito species, only around 60 out of the 450 *Anopheles* species can transmit malaria parasites to infect humans, i.e. are vectors [2]. Therefore, detailed mosquito surveying in areas of endemic malaria is crucial to identify the distribution of malaria-vectoring mosquitoes.

Standard mosquito sampling approaches, including human landing catches, odour-baited traps and cow-baited tents, can be effective in sampling malaria vectors [3, 4]. However, they expose volunteers to potentially infectious bites or are not sufficiently efficient for large-scale and frequent monitoring of mosquito distributions. An alternative solution, using mosquito flight tones to distinguish species, has been researched for some 60 years [5, 6]. In recent years, proof-of-concept mosquito acoustic sensing paradigms, based on embedded devices such as mobile phones, have been proposed [7, 8, 9]. Embedded devices provide a compelling platform for such environmental acoustic sensing tasks due to their cheap and efficient sensors, wide availability and built-in storage and wireless connectivity [10].

Research in the signal processing aspect of mosquito acoustic sensing has often focused on two areas. Firstly the use of domain knowledge to extract hand-crafted features to then allow high-quality detections and secondly the construction of machine learning frameworks which are well-suited to not just detect mosquitoes but importantly also to distinguish species. In much work, fundamental frequencies and associated harmonics form the basis for models which identify mosquito species [11, 9]. However, these low-dimensional features suffer from high intra-species variances and significant overlaps between different species [11, 12], hence limiting their application in multi-species classification. Alternative approaches look to avoid such feature construction and instead allow machine learning algorithms to extract relevant information direct from e.g. the spectrogram. Promising detection results have been reported [8, 13], though we note that the datasets used in evaluations of most previous work are limited in their sample sizes and were usually collected with mosquitoes raised in lab environments.

As a part of the HumBug project[1], a two-month mosquito survey was conducted in rural Thailand. A total of 1256 individual mosquitoes of 9 different mosquito species were captured and the flight tones of these mosquitoes were recorded for each captured individual. We here present the development of a machine learning algorithm that is computationally efficient (as it needs to be for implementation on low-powered embedded devices) and report in this paper on its performance over this field-recorded dataset.

The rest of this paper is organised as follows. We describe in Section 2 the dataset and the proposed mosquito acoustic detection algorithm. In Section 3 we report detection performance and discuss results. We conclude the paper and discuss future directions in Section 4.

[1] humbug.ac.uk

2. DATA AND METHODS

2.1. Field experiments and data summary

A two-month comprehensive survey of mosquito fauna was conducted at Pu Teuy Village, Sai Yok District, Kanchanaburi Province, Thailand. The survey was conducted within the peak mosquito season (May to October), and ran from the 12th of June until the end of July in 2018. Three methods of capture were used: human baited nets (HBN), cow baited nets (CBN) and larval collections. The HBN and CBN were run for 12 hours over night with collections made each hour throughout. All adult mosquitoes were placed into sample cups large enough for them to fly freely and their flight was recorded the morning following capture.

Recording was conducted using two microphone set-ups: 'budget' and 'high spec'. The 'budget' set-up used an Alcatel One-Touch Pixi smartphone with a TIE 19-90003 condensor microphone. The budget setup also used our 'Mozzwear' app on Alcatel smartphones to perform data capture and digitisation. The 'high-spec' set-up used a high specification field microphone (Telinga EM-23) plugged into a digital sound recorder (Olympus LS-14). Monophonic recordings were collected in both set-ups. Larval collections were made along a small river with known anopheline larval sites and the larvae/pupae were placed into rearing trays. The emerging adults were placed into individual sample cups and provided with $10\% w/v$ sucrose solution and recorded as above. As of the 20th July 2018, a total of 1256 individual mosquitoes had been captured. The detailed number of individuals for each species is reported in Table 1. A total of 127 mosquito individuals died before recording, 92 were lost, 46 did not fly and 21 individuals are as yet unidentified.

After recording the mosquito flight tones of these captured mosquito individuals, data tagging was required to mark segments of recordings with mosquito flight tones. Our project research team labelled a subset of the recordings and obtained more than 1 hour of mosquito flight tones from the field-captured mosquitoes, in addition to background recordings. The number of mosquito individuals and durations of flight tones of each species are shown in Table 1.

Table 1: Number of captured mosquito individuals and durations of recorded mosquito flight tones for different species.

Mosquito species	# individuals	Recorded time
Aedes sp.	256	954 seconds
An. maculatus	105	486 seconds
An. dirus	110	474 seconds
An. harrisoni	150	612 seconds
Armigeres sp.	261	1084 seconds
Culex sp.	67	386 seconds
Mansonia sp.	4	14 seconds
An. minimus	8	61 seconds
An. barbirostris	9	13 seconds

2.2. Mel-frequency spectrum-based convolutional neural networks

In this paper we propose a computationally efficient multi-species classification pipeline. As our goal is to port the model onto limited resource hardware, we base the feature encoding on the mel-frequency spectrum (MFS) and use convolutional neural networks (CNNs) as the decision engine. The mel-frequency is chosen due to its effectiveness as well as its interpretability. Although the CNN incurs well-known computational costs during training, running the trained CNN on new data is efficient. We detail below the feature encoding approach taken as well as the classification algorithm.

2.2.1. Feature representation

Previous work [8, 13] identified either mel-frequency cepstral coefficients (MFCCs) or wavelets as effective features for multiple machine learning algorithms. MFCCs are computationally efficient and have been a popular choice in mosquito detection and other acoustic scene classification tasks [8, 10]. However, the discrete cosine transform (DCT) step leads to less human-interpretable features than MFCCs, as shown in Figure 1, where the mel-frequency spectrum (Figure 1(d)) better preserves the harmonic structure in the spectrogram (Figure 1(b)) in comparison to the MFCC (Figure 1(c)). Further, the mel-frequency spectrum is also fast to compute and it forms a compact representation of the spectrum. Our experiments, including that presented in Section 3, have shown that the mel-frequency spectrum leads to detection performance with no statistically significant difference to results obtained with the MFCC.

We therefore use the mel-frequency spectrum to construct time-frequency representations of audio recordings in the same spirit of the short-time Fourier transformation (STFT). For a short audio clip, e.g. 0.1 second, we can compute the mel-frequency spectrum for smaller segments within the clip, and combine them to form a time-frequency matrix. This resultant matrix forms the input space of the subsequent machine learning algorithm (Figure 2).

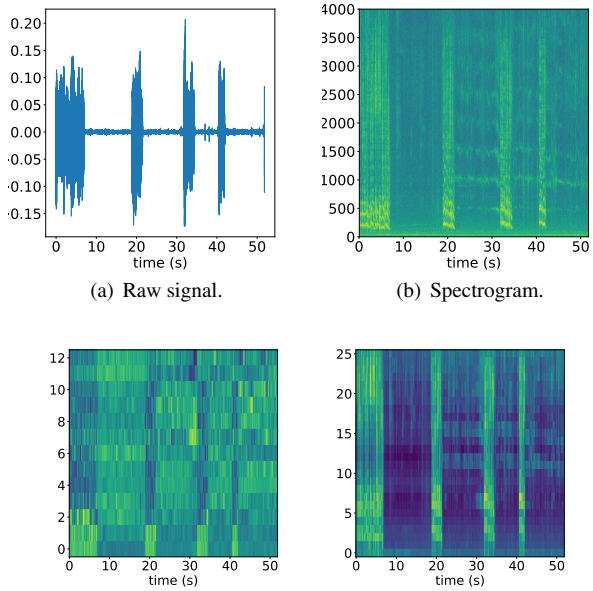
Compared with the spectrogram or the wavelet, this mel-frequency, spectrum-based representation has much smaller dimension - making it efficient for model training with datasets of small sample sizes, such as those often encountered in our application. We note that the wavelet transformation has been shown to provide informative time-frequency features that are well-suited to subsequent use, particularly by convolutional neural networks (CNNs) [13]. However, the wavelet transformation used in this latter work is computationally demanding and unsuited for the task of real-time mosquito species classification on low-cost phones.

2.2.2. Classification algorithm

Convolutional neural networks (CNNs) are a subset of (deep) neural networks that have been widely used in processing two-dimensional inputs, e.g. images [14, 15]. The first layer in a CNN consists of a set of convolutional filters, whose parameters (the filter coefficients) are learned as part of the training process. These filters act so as to create latent representations of the observations which are passed upwards to layers in the CNN which create discriminants associated with the classes of interest. This approach, crucially, does not prespecify the form of patterns in the data which are highly informative. CNNs have been widely used in the machine vision literature and we exploit their excellent performance over images by treating the mel-frequency spectrum as a 2d image.

The convolutional layer takes an input tensor $X \in \mathbb{R}^{h_1 \times w_1 \times c}$ where $c = 1$ for the mel-frequency spectrum-based features. N_k kernels $K_p \in \mathbb{R}^{k,k}$ for $p \in \{1, \ldots, N_k\}$ are applied to the input tensor which produces the convolution output Y_p:

$$Y_p(i,j) = (X * K)(i,j) = \sum_m \sum_n X(i-m, j-n) K(m,n) \ . \quad (1)$$



(a) Raw signal. (b) Spectrogram.

(c) MFCC (plotted over time). (d) Mel-frequency spectrum (plotted over time).

Figure 1: An example recording and associated spectral features. Mosquito flight tones are present from 21 s to 31 s for cow capture #242, 35 s to 40 s from cow capture #243, 42 s to 52 s from cow capture #251. Other segments of the recording (including the high amplitude sections) are either background noise or human voices.

These feature maps, i.e. convolution outputs, are then passed through some non-linear activation function, such as a Rectified Linear Unit (ReLU), before being flattened and connected to fully connected hidden layers. The last layer consists of N nodes where N is the number of classes.

2.3. Training strategy

In training and evaluating mosquito acoustic detection algorithms, we first randomly remove 50% of recordings described in Section 2.1 to create the hold-out test dataset. For the remaining recordings, we divide recordings into audio clips with a length of 0.1 seconds, thus creating a relatively large number of samples with which we train the CNN and other benchmark algorithms.

As shown in Table 1, the dataset is highly imbalanced. To avoid issues with very small data sample sizes, we only use samples from the *Aedes* sp., *An. Maculatus*, *An. dirus*, *An. harrisoni*, *Armigeres* and *Culex* sp. to evaluate the detection performance of the algorithms, as there is less than 2 minutes of recordings for the other species. Three of these species are known malaria vectors, including *An. maculatus*, *An. dirus* and *An. harrisoni*. Following [8], we randomly sample the training samples, without replacement, to produce a balanced training set. In our application, this creates a data set of close to 2000 samples for each mosquito species. A total of 100 randomised trials were performed so that different training and test data sets were produced among different simulation trials.

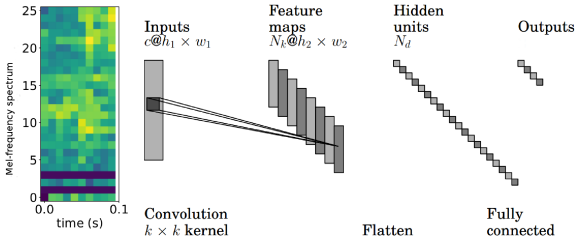


Figure 2: Example CNN architecture. Input to the CNN is a mel-frequency spectrum computed from a 0.1 second audio clip with $c = 1$ channel and dimension $h_1 \times w_1$. The CNN has N_k filters with kernel $K \in \mathbb{R}^{k \times k}$ thus it reduces the input dimension to $h_2 \times w_2$ following convolution with each filter. These feature maps are flattened (i.e. vectorised) before being fully connected to N_d hidden units in a dense layer. The last fully-connected layer produces the classification output.

3. EXPERIMENT RESULTS

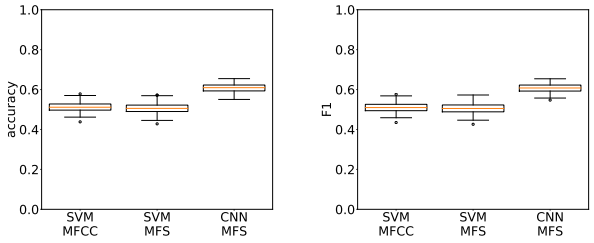
3.1. Parameter setup and benchmarking

The input image to the CNN is of dimension 26×10, where 26 is the dimension of the mel-spectrum and 10 is the number of 0.01 second windows within a 0.1 second audio clip. In this initial investigation we adopt a network structure inspired by [16] which was used for MNIST handwriting digit recognition. There are three convolutional layers: the first layer consists of 8 filters, the second one has 32 filters, and the last one is made up of 64 filters. All filters have 3×3 kernel size. They are followed with one hidden layer of 128 nodes. The ReLU activation unit and a dropout rate of 0.3 are used. The neural network is trained using the Adam algorithm [17] with a batch size of 256 for 100 epochs. A full cross-validation of these default parameter values, and the optimisation of network structure, will be performed in follow-up studies.

We choose as benchmark classifier a support vector machine (SVM) using a one-versus-one multi-class classification strategy [18]. The one-versus-one multi-class classification strategy has a simpler data balancing requirement compared to one-versus-rest. As discussed in [8], MFCC features combined with a SVM obtains the best multi-species classification accuracy among several common acoustic features and off-the-shelf detection algorithms. However, subsequent studies showed that the mel-frequency spectrum leads to similar detection performance and more human-interpretable features.

3.2. Results

Figure 3 plots the out-of-sample classification accuracy and F1 score of the compared algorithms, respectively. The mel-frequency spectrum (MFS)-based CNN algorithm exhibits significantly better classification performance, in terms of both classification accuracy and F1 score over the SVM algorithms. We observe no significant difference between results obtained with MFCCs and the mel-frequency spectrum. As shown in Figure 1, the mel-frequency spectrum is able to better preserve the harmonic structure of the mosquito flight tone than MFCCs. Figure 4 and 5 plot confusion matrices for the MFS-based SVM and the MFS-based CNN, respectively. The CNN exhibits better mean sensitivities than the SVM for every mosquito species in this experiment.



(a) Classification accuracy. (b) F1 score.

Figure 3: Boxplots showing out-of-sample classification accuracy and F1 scores across 100 randomised trials.

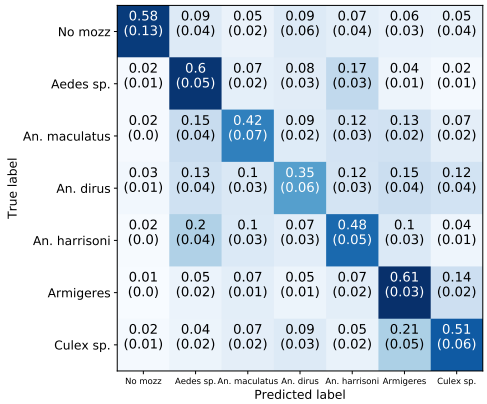


Figure 4: Confusion matrix of out-of-sample classification performance for the mel-frequency spectrum-based SVM. The first value in each entry is the corresponding mean value among 100 simulation trials, while the value in the parenthesis reports the standard deviation

Considering the fact that the dataset contains recordings from more than 1000 mosquito individuals from 6 species, an average classification accuracy of above 60% from the CNN (Figures 3 and 5), using half of the samples for training, is very encouraging. Once trained, both the SVM and the CNN are computationally efficient in their prediction step, allowing their real-time execution in low-cost low-power embedded devices.

Compared to the classification results reported in [8], where an average of 80% classification accuracy is achieved with a SVM, we notice a significant decrease of classification accuracy with this 'in the wild' dataset. It is important to note that the data size is significantly smaller in [8] where only 6.2 seconds of recordings were available for each species after data resampling, and only one mosquito individual per species was used to collect recordings. This suggests higher correlations between different samples in the lab-collected dataset of [8] and highlights the importance of performance evaluation with large-scale mosquito flight tone datasets.

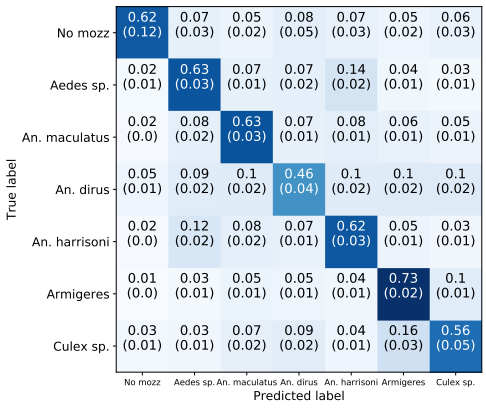


Figure 5: Confusion matrix of out-of-sample classification performance for the mel-frequency spectrum-based CNN. The first value in each entry is the corresponding mean value among 100 simulation trials, while the value in the parenthesis reports the standard deviation

4. CONCLUSION

Mosquito acoustic detection aims to provide an alternative solution to fast sample and update mosquito species distribution, which is crucial for control programmes and guiding intervention policies. The HumBug project uses mobile phones for mosquito acoustic detection and species classification. Our dataset, recently collected in field sites in Thailand, provides a valuable resource to develop and evaluate mosquito acoustic detection algorithms. The nature of the task requires an algorithm with a low computational cost while maintaining effectiveness with a small number of training samples.

We propose in this paper a computationally efficient classification pipeline, based on the mel-frequency spectrum and convolutional neural networks. The mel-frequency spectrum (MLS) is a computationally efficient, low-dimensional acoustic feature. We show that the proposed pipeline, with the CNN acting to construct latent features from the MLS, achieves impressive classification performance on a challenging field dataset.

This initial investigation reports classification results with labelled data from a subset of recordings in which labels were obtained by data tagging from project team members. Further work will include data which is being labelled via citizen scientists on the Zooniverse[2] citizen science platform. Working with such data will require for algorithms capable of handling crowdsourced labels which are at different resolutions to the original data frames - and this is a topic of active current research. Further directions also include optimisation of the network structure, identification of more effective feature extraction methods, as well as the incorporation of the mosquito field dataset with other DCASE Challenge datasets to form a large-scale acoustic sensing task in future DCASE Challenge events.

ACKNOWLEDGEMENTS

This work is part-funded by a Google Impact Challenge award.

[2]https://www.zooniverse.org/projects/yli/humbug

5. REFERENCES

[1] S. Bhatt et al., "The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015," *Nature*, vol. 526, no. 7572, pp. 207–211, 2015.

[2] D. E. Neafsey et al., "Highly evolvable malaria vectors: The genomes of 16 anopheles mosquitoes," *Science*, vol. 347, no. 6217, 2015.

[3] J. Wong et. al., "Standardizing operational vector sampling techniques for measuring malaria transmission intensity: evaluation of six mosquito collection methods in western Kenya," *Malaria Journal*, vol. 12, no. 1, p. 143, Apr 2013.

[4] B. St. Laurent et. al., "Cow-baited tents are highly effective in sampling diverse anopheles malaria vectors in cambodia," *Malaria Journal*, vol. 15, no. 1, p. 440, Aug 2016.

[5] W. H. O. Jr. and M. C. Kahn, "The sounds of disease-carrying mosquitoes," *The Journal of the Acoustical Society of America*, vol. 21, pp. 462 – 463, 1949.

[6] C. Pennetier, B. Warren, K. R. Dabir, I. J. Russell, and G. Gibson, "singing on the wing as a mechanism for species recognition in the malarial mosquito Anopheles gambiae," *Current Biology*, vol. 20, no. 2, pp. 131 – 136, 2010.

[7] I. Kiskin, B. P. Orozco, T. Windebank, D. Zilli, M. Sinka, K. Willis, and S. Roberts, "Mosquito detection with neural networks: the buzz of deep learning," *arXiv:1705.05180*, 2017.

[8] Y. Li, D. Zilli, H. Chan, I. Kiskin, M. Sinka, S. Roberts, and K. Willis, "Mosquito detection with low-cost smartphones: data acquisition for malaria research," in *NIPS Workshop on Machine Learning for the Developing World*, Long Beach, USA, Dec. 2017, arXiv:1711.06346.

[9] H. Mukundarajan, F. J. H. Hol, E. A. Castillo, C. Newby, and M. Prakash, "Using mobile phones as acoustic sensors for high-throughput mosquito surveillance," *eLife*, vol. 6, p. e27854, Oct. 2017.

[10] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Mag.*, vol. 32, no. 3, pp. 16–34, 2015.

[11] I. Potamitis and I. Rigakis, "Measuring the fundamental frequency and the harmonic properties of the wingbeat of a large number of mosquitoes in flight using 2d optoacoustic sensors," *Applied Acoustics*, vol. 109, pp. 54 – 60, 2016.

[12] S. M. Villarreal, O. Winokur, and L. Harrington, "The impact of temperature and body size on fundamental flight tone variation in the mosquito vector Aedes aegypti (diptera: Culicidae): Implications for acoustic lures," *Journal of Medical Entomology*, vol. 54, no. 5, pp. 1116–1121, 2017.

[13] I. Kiskin, D. Zilli, Y. Li, M. Sinka, K. Willis, and S. Roberts, "Bioacoustic detection with wavelet-conditioned convolutional neural networks," *Neural Computing and Applications*, 2018, accepted.

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Information Processing Systems*, Lake Tahoe, USA, Dec. 2012, pp. 1097–1105.

[16] P. Y. Simard, D. Steinkraus, and J. Platt, "Best practices for convolutional neural networks applied to visual document analysis," in *Proc. Int. Conf. Document Analysis and Recog. (ICDAR)*, Washington, DC, USA, Aug. 2003.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sept. 1995.