How Much Attention Should we Pay to Mosquitoes?

Original
How Much Attention Should we Pay to Mosquitoes? / Vaiani, Lorenzo; Koudounas, Alkis; LA QUATRA, Moreno; Caglieri, Luca; Garza, Paolo; Baralis, ELENA MARIA. - ELETTRONICO. - (2022), pp. 7135-7139. (Intervento presentato al convegno Computational Paralinguistics Challenge 2022 (ComParE 2022) tenutosi a Lisbon (PT) nel October 10-14, 2022) [10.1145/3503161.3551594].

Availability:
This version is available at: 11583/2971157 since: 2022-09-09T10:15:13Z

Publisher:
Association for Computing Machinery

Published
DOI:10.1145/3503161.3551594

Terms of use:
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
ACM postprint/Author's Accepted Manuscript

(Article begins on next page)
How Much Attention Should we Pay to Mosquitoes?

Moreno La Quatra*  
moreno.laquatra@polito.it  
Politecnico di Torino  
Turin, Italy

Lorenzo Vaiani*  
lorenzo.vaiani@polito.it  
Politecnico di Torino  
Turin, Italy

Alkis Koudounas*  
alkis.koudounas@polito.it  
Politecnico di Torino  
Turin, Italy

Luca Cagliero  
luca.cagliero@polito.it  
Politecnico di Torino  
Turin, Italy

Paolo Garza  
paolo.garza@polito.it  
Politecnico di Torino  
Turin, Italy

Elena Baralis  
elena.baralis@polito.it  
Politecnico di Torino  
Turin, Italy

ABSTRACT
Mosquitoes are a major global health problem. They are responsible for the transmission of diseases and can have a large impact on local economies. Monitoring mosquitoes is therefore helpful in preventing the outbreak of mosquito-borne diseases. In this paper, we propose a novel data-driven approach that leverages Transformer-based models for the identification of mosquitoes in audio recordings. The task aims at detecting the time intervals corresponding to the acoustic mosquito events in an audio signal. We formulate the problem as a sequence tagging task and train a Transformer-based model using a real-world dataset collecting mosquito recordings. By leveraging the sequential nature of mosquito recordings, we formulate the training objective so that the input recordings do not require fine-grained annotations. We show that our approach is able to outperform baseline methods using standard evaluation metrics, albeit suffering from unexpectedly high false negatives detection rates. In view of the achieved results, we propose future directions for the design of more effective mosquito detection models.

CCS CONCEPTS
• Computing methodologies → Artificial intelligence: Supervised learning.

KEYWORDS
Mosquito detection, Audio event detection, Transformer models, Audio sequence modelling

1 INTRODUCTION
The identification of mosquitoes is of great importance for monitoring the risk of mosquito-borne diseases and for the development of mosquito control strategies. However, recognizing the mosquito buzzing sounds is a challenging task. Aside from the difficulty of detecting faint sounds, they can be also easily confused with similar background or buzzing sounds. In recent years, data-driven methodologies have shown state-of-the-art performance on Acoustic Event Detection (AED). Deep neural networks are capable of learning complex representations of raw audio data, which can be particularly beneficial to tackle AED. However, most of the aforementioned approaches are designed for the audio classification tasks, e.g., urban sound classification [13]. Conversely, the identification of mosquitoes events is a frame-based event detection task, which requires the correct identification of intermittent and variable-length buzzing sounds.

The most common approach to identify acoustic events in long-lasting recordings is to train a deep learning model to classify short audio segments. In [8, 9] each audio frame is processed independently, and the classification results are combined to make global decisions on the whole audio recording. As a drawback, the aforesaid approaches ignore the temporal dependencies among frames, which can be crucial for the identification of variable-length mosquitoes buzzing sounds.

To address the above issue, we propose a methodology that leverages transformer-based architectures to analyze long-range dependencies among audio frames in the whole audio recording. Transformer-based models have proven to be useful for modeling dependencies among sequential data in a variety of tasks, related to both text [4, 16] and audio processing [1, 3]. They leverage a self-attention mechanism to model the relationships between tokens/frames, thereby building a global representation of the input sequence. Notice that acoustic events can occur at different temporal locations and have arbitrary durations. This makes their representations particularly well-suited for their identification of mosquitoes acoustic events. To our purposes, we generate audio signals containing acoustic events at different positions, which are then used to train the model for the prediction of frame-level labels. To the best of our knowledge, this is the first attempt to use transformer-based models to automatically identify mosquitoes acoustic events in audio recordings.

2 WAVEFORM ANALYSIS FOR MOSQUITO DETECTION
Detecting mosquito audio events in an audio recording entails the identification of the time intervals in which each event occurs. Hence, it entails modeling the audio sequence to take into account the sequential dependencies among the audio recordings.

We present Waveform analysis for Mosquito Detection (WavMoDe). It extends the WavLM model [3] to address the task of mosquito event detection. WavLM [3] is a transformer-based model for modeling the sequential dependencies in an audio sequence. It has been originally proposed for speech-related tasks such as automatic speech recognition, speaker identification, and speaker

*All authors contributed equally to this research.

The project source code is available for research purposes https://github.com/MorenoLaQuatra/ComParE2022_MED (Latest access: June 2022)
When the frames are tagged at the frame level, the problem can be reformulated as a sequence labeling task in which each frame is a binary label indicating whether or not the corresponding frame contains an acoustic mosquito event. Given an audio recording, the goal is to identify the start and end timestamps of mosquito audio events. In the former case, the position of the mosquito is selected according to the following heuristics:

- In one third of the cases, the mosquito is added at the beginning of the recording (e.g., $p_{\text{back}} = 0$).
- In one third of the cases, the mosquito is added at the end of the recording (e.g., $p_{\text{back}} = \text{len}(r_{\text{back}}) - \text{len}(r_{\text{mos}})$).

Where $\text{len}(\cdot)$ denotes the length of the corresponding recording. The frames of the generated sequence are labeled as $y_j = 0$ for the ones that contain the mosquito and as $y_j = 0$ otherwise. More details about the labeling procedure for the generated recordings are given in Section 2.2. The background signals for the data augmentation process are randomly selected from the recordings of the training dataset containing background noise. Furthermore, additional noise samples are selected from:

- FSDnoisy18k [7], a public-available dataset characterized by 42.5 hours of data belonging to 20 different classes. It contains real-world noisy audios and their corresponding manual annotations.
- Sound Events for Surveillance Applications (SESA) [15], a freely-accessible dataset containing 585 audios belonging to 4 different noise classes.
- Glasgow Isolated Sound Events (GISE-51) [17], an open dataset, based on FSD50k dataset [6], which includes 16'357 audios of different duration covering 51 different classes.
2.2 Data labeling
We combine audio signals containing mosquitoes events with background sounds to create new signals annotated with mosquito buzzing sounds. The audio frame classification task is treated as a sequence labeling problem. The label assigned to each audio frame indicates either the presence or the absence of a mosquito event. The proposed endeavor closely resembles the Named Entity Recognition (NER) task in the domain of Natural Language Processing: the key idea is to assign the label 0 to every frame that is not part of any event, and assign labels B, I, E to the frames at the beginning, inside, and at the end of the event, respectively.

In the experiments we test the following three different labeling options (see Figure 2):

- **2-label schema:** It is the most straightforward labeling schema. The label 0 is assigned to every frame that is not part of any event whereas all the frames containing mosquito sounds are assigned to the class I.
- **3-label schema:** The label 0 is assigned to every frame that is not part of any event. The frames containing mosquito sounds are assigned to the classes B if they are at the beginning of an event, I otherwise (i.e., if they are internal or at the end of the mosquito event).
- **4-label schema:** It is similar to the 3-label schema, but the end of an event is explicitly highlighted using label E.

The 2-label schema is potentially an over-simplification and does not allow to distinguish between frames at the beginning, in the middle and at the end of the mosquito event. The 4-label schema, on the other hand, is a more fine-grained labeling schema and allows us to distinguish frames at the beginning, in the middle and at the end of the mosquito event. The 3-label scheme is a trade-off in which only the beginning of an event is denoted by the label B and the end is not explicitly indicated. During inference, we aggregate the predictions of all mosquito-related events, i.e., we sum the class probabilities of frames not labeled with 0 into one class probability (i.e, the overall probability of having a mosquito event). An empirical comparison between the labeling options can be found in Section 3.

3 EXPERIMENTS

We evaluate the Wav-MoDe performance on the official dataset (HumBugDB) of the ComParE 2022 mosquito sub-challenge [14]. However, notice that the proposed solution is general and can be applied to address similar tasks.

**HumBugDB dataset characteristics.** The dataset proposed for the mosquito sub-challenge is a large-scale collection of mosquito sounds recorded with mobile phones [9, 10]. It is characterized by ~ 20 hours of recordings containing mosquito sounds and ~ 15 hours of recordings of non-mosquito sounds. The dataset is split into train, dev-a, dev-b and test sets. While train and development sets are made available to the participants, the test set is redacted for a blind evaluation process. Both training and development sets contain only coarse-grained labels that indicate the presence of mosquito sounds in each audio sample.

**Evaluation metrics.** The ComParE 2022 mosquito sub-challenge relies on the SDS [2] metric to compare the performance of the proposed systems. It is specifically designed to evaluate sound event detection systems. In light of the similarity between the mosquito event detection and the speaker diarization problem, we also consider the Detection Error Rate (DER) [5], which is commonly used in evaluations by binary classification tasks such as speech activity detection. It is defined as:

\[
\text{DER} = \frac{F + M}{N}
\]

where \(F\), \(M\), and \(N\) represent the duration of false positives, misses, and the total duration of each audio event, respectively.

**Experimental settings.** Wav-MoDe is trained on the training set given by the challenge organizers. The model is fine-tuned using AdamW optimizer [11] with an initial learning rate of \(10^{-5}\) and a linear decay schedule with a decay rate of 0.01. We train the model for a maximum of 20 epochs using early stopping strategy. At each epoch, the model is evaluated on the development set provided by organizers (i.e., the evaluation is performed combining both dev-a and dev-b sets) to choose the best checkpoint.

All the experiments were performed on a machine running Ubuntu 21.10 and equipped with AMD® Ryzen 9th 3950X CPU, Nvidia® RTX 3090 GPU, and 128 GB of RAM.

**Attention window.** Transformer-based models learn long-range dependencies among audio frames by encoding the entire sequence using an attention mechanism. However, the complexity of the attention mechanism is quadratic to the sequence length and impede the model to process long audio recordings. To overcome the aforesaid limitation, Wav-MoDe is trained setting the maximum duration of the attention window to 60 seconds. During the inference process, the model is then applied to the recording by splitting it into chunks of different lengths \(L_w \in \{60, 6, 0.6, 0.06\}\) seconds. The length of a single frame in Wav-MoDe is 0.02 seconds. Specifically, the highest value \(L_w = 60\) seconds corresponds to the training time window, whereas \(L_w = 0.06\) seconds corresponds to lowest resolution in terms of context that can be used to classify an audio frame.

3.1 Results
The results of the evaluations performed on the development and test sets are reported in Table 1. It indicates the DER and SDS values achieved by Wav-MoDe in different evaluation settings, as well as for the baseline proposed in [8]. The results on the test set are limited by the maximum number of system submissions allowed by the challenge organizers.

We separately evaluate the results obtained by Wav-MoDe with different attention windows to assess the impact of contextual information on the results. For both metrics, the optimal results are obtained with \(L_w = 0.6\). This indicates that considering a larger context is not always beneficial. Indeed, the \(L_w = 60s\) setting is likely to provide too much context for the classification task of a single frame, introducing noise and impeding the model to focus on the relevant context. On the other hand, by setting \(L_w = 0.06s\) is likely to provide a too limited context for the classification task of a single frame, thus reducing the ability of the model to leverage the local context. The empirical evaluation shows that \(L_w = 0.6s\) (i.e., the best performing configuration for both metrics
in the development sets) is a good trade-off, as provides enough context without introducing significant irrelevant information.

Among different labeling schemes, the best results are obtained with the 3/4-label scheme. By explicitly indicating the beginning and ending of an event, the model is better trained to recognize local patterns. Comparing their results in the development sets, 3-label scheme provide better results with larger attention windows, while 4-label scheme provide better results with smaller ones.

**Pitfall in evaluation metrics.** The official evaluation metric (PSDS) provides a quantitative assessment of the ability of the model to recognize mosquito events. However, when evaluating the model from a qualitative perspective, we found that it is likely to predict short-lasting background within long-lasting mosquito events. The PSDS metric, in its standard setting overlooks this behaviour and, as a result, both model fine-tuning and evaluation are likely to lead to sub-optimal results in real-world applications. To overcome the aforementioned limitation, contest organizers provided an alternative PSDS scoring function, which accounts for short incorrect detections (see the PSDS column with MaxEFPR=3600 in Table 1). These modifications penalizes Wav-MoDe, whose results, due to the model’s tendency to produce short-lasting background detections, are significantly lower compared to the original scoring system (see DER column).

4. CONCLUSIONS AND FUTURE WORKS

In this work, we proposed a novel approach, namely Wav-MoDe, to detect and localize mosquito events from audio recordings. It is based on an end-to-end transformer architecture that is trained to predict the mosquito event label for each frame of the input waveform. Although the model achieves very good performance according to the standard evaluation metrics, unfortunately, it is susceptible to predicting short-lasting background within long-lasting mosquito events. This behaviour is partly highlighted by the alternate PSDS scoring provided by the contest organizers, which penalizes systems providing such prediction pattern.

The future research activities call for the study and adoption of new and ad hoc evaluation metrics able to effectively manage short-lasting events and thus better reflecting human expectations. Furthermore, we aim at overcoming the limitations enforced by the complexity of the standard attention mechanism by leveraging time-restricted self-attention models [12].

Finally, seeking for a more efficient utilization of temporal correlations, we will investigate the design of streaming architectures able to provide online prediction of mosquito events.

**ACKNOWLEDGMENTS**

The research leading to these results has been partly funded by the SmartData@PoliTO center for Big Data and Machine Learning technologies.

**REFERENCES**


How Much Attention Should we Pay to Mosquitoes?  

Short Papers. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423


