AUTOMATIC DETECTION OF COW VOCALIZATIONS USING CONVOLUTIONAL NEURAL NETWORKS

*Ester Vidaña-Vila[1] *, Jordi Malé[1], Marc Freixes[1], Mireia Solís-Cifré[1], Miquel Jiménez[1], Cristian Larrondo[2,3], Raúl Guevara[2], Joana Miranda[2], Leticia Duboc[1], Eva Mainau[2], Pol Llonch[2], Rosa Ma Alsina-Pagès[1]*

[1] HER – Human-Environment Research, La Salle – Ramon Llull University, Barcelona, ES
[2] AWEC Advisors S.L., Edifici Eureka, Parc de Recerca de la Universitat Autònoma de Barcelona
[3]Center for Applied Research in Veterinary and Agronomic Sciences, Faculty of Veterinary Medicine and Agronomy, Universidad de Las Américas, Chile
* Corresponding author: ester.vidana@salle.url.edu

ABSTRACT

The well-being of animals holds significant importance in our society. Apart from the ethical concerns, recent studies have highlighted the correlation of animal growth, reproductive potential, and overall productivity with animal welfare. In this context, the vocalizations of cows have emerged as a valuable indicator of their well-being for veterinary researchers, but gathering and labelling the vocalizations for their in-depth study is time-consuming and labour-intensive. For this reason, in this work, we present an acoustic event detection algorithm that has been trained and validated with different setups using acoustic data collected from two different farms. The experimental set-up consists of a Convolutional Neural Network followed by a post-processing stage for the detection of vocalizations, so veterinary researchers can easily analyze them. The experimental evaluation assesses the importance of selecting the convenient post-processing and overlapping acoustic window for finding new vocalizations. Furthermore, the study evaluates the significance of using data collected specifically from the same farm for acoustic event detection, as opposed to employing data from a different farm. Results show that by merging training data from different farms, including the farm that is being evaluated, an F1 score of 57.40% and a recall of 74.05% can be achieved.

Index Terms— Acoustic event detection, Cow vocalization, Deep learning, Bioacoustics, Cow monitoring

1. INTRODUCTION

Animal welfare has gained significant importance in our society, both for its ethical consideration and because it can affect animal growth, reproductive potential, and overall productivity [1]. For this reason, society is demanding welfare-monitoring methodologies that do not affect the physical integrity of the animals [2]. Among various animals, cows have gained particular attention from veterinary researchers due to the potential insights that can be gained from monitoring and interpreting their vocalizations (thus, avoiding animal manipulation). This vocal information is key, as it can provide details about the animals' conditions, such as pain, stress and hunger, among others [3, 4].

In order to respond to this need, recent contributions in the field have focused on developing algorithms for both automatically detecting and classifying the vocalizations of cows [3, 5] and analysing them for welfare monitoring [6, 7]. These automatic techniques can help farmers, veterinarians and researchers to gain valuable insights into the conditions and well-being of cows. However, most of these studies were conducted on single farms, which limits the ability to evaluate the performance of the algorithm in various environments and farm setups.

The work presented in this paper tackles this problem; that is, it seeks to develop an algorithm that can detect cow vocalizations in multiple farm environments. This research has been carried out under the umbrella of the project *"CowTalkPro: Desarrollo de un Sensor de Sonido en vacas para evaluar la salud y el bienestar animal"* (in English: Development of a Sound Sensor in cows to assess animal health and welfare.). Its interdisciplinary team is composed of engineers from La Salle Campus Barcelona (Ramon Llull University) and veterinarians and researchers from AWEC Advisors S.L..

The CowtalkPro project aims to create a single sensor that can be deployed in multiple farms—not only one—for real-time monitoring of the welfare of cows. More specifically, this project is concerned with three particular periods in the cows' lives:

First, during the initial weeks of life, monitoring calves can help support their health and, consequently, their wellbeing. If many coughs are heard within a short period, it might indicate that there are sick calves in the yard. For veterinarians and farmers, early detection of respiratory illness in calves is crucial to avoid spreading virus and because late treatment of such conditions could affect the production of that cow in its adult life.

Second, during the dry-offs, which are transitional phases between milk production and their dry phase before the get inseminated, cows may vocalize because they are experimenting pain or discomfort. Detecting these feelings can help cows' welfare by indicating the need to apply pain mitigation actions.

Finally, monitoring cows vocalizations during calving may inform whether the cow needs the assistance of a farmer.

Therefore, the resulting sensor can benefit farmers, veterinarians, and veterinary researchers interested in the assessment and monitoring of cows welfare.

Prior to the development of the sensor, it is important to determine which vocalizations give important insight to determine animal well-being. This normally requires the collection and interpretation of a significant number of cow vocalizations by the veterinary researchers, which is a complex and time-consuming task. In order to support this work, we have developed an automatic detector of vocalizations over audio recordings. The algorithm takes an audio

file recorded on a farm and detects two types of sounds: vocalizations and coughs.

At the current stage of the project, the algorithm, described in this paper, focuses on detecting vocalizations for the veterinarian researchers to analyse. More specifically, an acoustic event detection algorithm has been trained and validated using acoustic data collected from two distinct farms, with the aim of improving its adaptability and reliability in monitoring cow vocalizations in real-world scenarios. The presented algorithm utilizes a Convolutional Neural Network (CNN) as the primary detection model, which is then followed by a post-processing stage to refine the results.

The experimental evaluation of our approach encompasses two key aspects: on one hand, we investigate the significance of selecting the appropriate post-processing techniques and overlapping acoustic window for effectively detecting vocalizations. These parameters play a crucial role in uncovering new vocalizations that might otherwise go unnoticed. On the other hand, we explore the implications of using farm-specific data for acoustic event detection, as opposed to employing data from a different farm. This analysis allows us to assess the impact of dataset heterogeneity on the algorithm's performance.

The paper is organised as follows. Section 2 presents the experimental evaluation pipeline. Next, Section 3 details the obtained results. Finally, the conclusions and future work are presented in Section 4.

2. EXPERIMENTAL EVALUATION

This section provides an overview of the experimental evaluation pipeline, which includes the following components: data collection campaigns conducted in two farms, the utilization of a CNN-based algorithm for automated vocalization detection, post-processing techniques employed to determine the onset and offset of each vocalization, and the utilization of data from multiple farms to assess the algorithm's generalization capabilities.

2.1. Data collection

For the experimental evaluation, audio files recorded in two different farms have been used. The first farm is located in Girona (Spain), and the second farm is located in Valencia (Spain). In both cases, a similar recording setup was used. That is, a mains powered audio recorder Zoom H5 [8] placed inside of a box, and connected to an omnidirectional microphone via a long XLR wire (about 30 m). The microphone hung on the ceiling of the cows' yard. An example of set-up is shown in Figure 1. Two microphones were placed on each farm. In Valencia, both microphones were in a big yard for calves, with a separation of about 50 m between them. In Girona, one microphone was over a calves yard and the other one covered dairy cows at the dry-off period.

The hardware set-up was installed in the farm collecting continuous data for about one year. Due to the limitation of the SD card that can be placed on the Zoom recorder, which cannot hold more than 32 GB, and using a sample rate of 44,100 Hz, each week, we have recorded for about four days and a half. After that, the SD card had to be manually replaced.

A small proportion of this audio data has been manually labelled and used for the experiments. Specifically, for this work, the following audio files of 15 min each have been used:

- **Girona:** 40 audio files from cows and 79 audio files from calves.

Figure 1: Installation of a microphone over the calves yard in Girona.

- **Valencia:** 80 audio files from calves.

This makes a total of 199 files, which represent almost 50 hours of labelled acoustic data. The annotation process was carried out by two different annotators under the supervision of veterinary experts using the Audacity software. The annotation taxonomy had two different categories: vocalizations and coughs.

The test set was built with 20 audio files from Valencia, as this farm has many more calves than Girona — and therefore more vocalizations per audio file.

The remaining audio files were chosen to be used as Training set with different splits, to evaluate whether using data from different farms improves or impairs the metrics of the vocalizations detection model explained in the following subsection.

Set	Farm	Vocalizations	Coughs
Train	Girona	2 289	1 107
Train	Valencia	3 107	1 579
Total train	Both	5 396	2 686
Test	Valencia	1 756	129

Table 1: Amount of labels found in every dataset.

As it can be observed in Table 1, the test set contains 129 cough instances and 1,756 vocalizations. This class imbalance is due to the nature of the audio files, as cows tend to vocalize more than cough, especially when they are not sick.

2.2. Automatic detection of vocalizations

The model used to automatically detect the cow vocalizations is a MobileNet [9] architecture. This model was chosen because its light architecture could be applied in the future to real-time detection of vocalisations on farms using low-cost devices (e.g., Raspberry Pi [10]), as tested by a subset of the authors of this paper in other domains [11], which is the final goal of the CowTalkPro project.

In all experiments, the MobileNet was trained for 15 epochs, using early stopping to obtain the best model (lowest validation loss) out of the 15. As inputs of the CNN, spectrograms were used. In line with previous studies [11], a window size of 1 second was selected to sample the audio file for training.

The CNN was configured as a multilabel classifier, as there might be more than one acoustic event present in a 1-second fragment (e.g., one cow is vocalizing while another cow is coughing). The two possible outputs of the model are vocalizations or coughs.

At the inference stage, the CNN was concatenated with a post-processing algorithm, which is in charge of delimiting the starting and ending point of every vocalization (on-set and off-set times). To achieve this, at inference time, and contrarily to the training stage (in which the audio files were split in windows of 1 second without overlap), the audios were split in overlapping windows.

2.2.1. Post-processing technique

The selection of the overlap time plays a decisive role for an accurate detection of vocalizations. For this reason, we present the classification results for three different overlapping times: 0.1 seconds, 0.25 seconds and 0.5 seconds. Figure 2 illustrates the different overlapping times.

For this experiment, all the data except for the one selected as test set was used for training, meaning that it incorporated data from both farms.

The metrics were calculated using the *"sed_eval" - Evaluation toolbox for Sound Event Detection* [12]. More specifically, segment-based metrics were used, with a configuration of a t_collar of 0.9 and percentage_of_length of 0.1. The first parameter is a tolerance with respect to the ground truth event duration, and the second one is the percentage of the length within which the estimated offset has to be in order to be considered a valid estimation.

2.2.2. Using data from different farms for training

After the previous experiment, and once a convenient post-processing overlapping time was selected, another set of experiments was carried out. In this case, the aim of the experiment was to quantify how the training data affected the results.

Three training sets were configured, each one used for a different experiment:

1. **Experiment 1**: Using the complete dataset of Girona (cows and calves) and the 60 audio files from Valencia that were not used as test set.
2. **Experiment 2**: Using only the dataset from Girona (cows and calves). Therefore, in this experiment, the training set consists of data recorded in a different farm than the one used for testing.
3. **Experiment 3**: Using only the dataset from Valencia. This means that the data used for training comes from the same farm as the data used for testing.

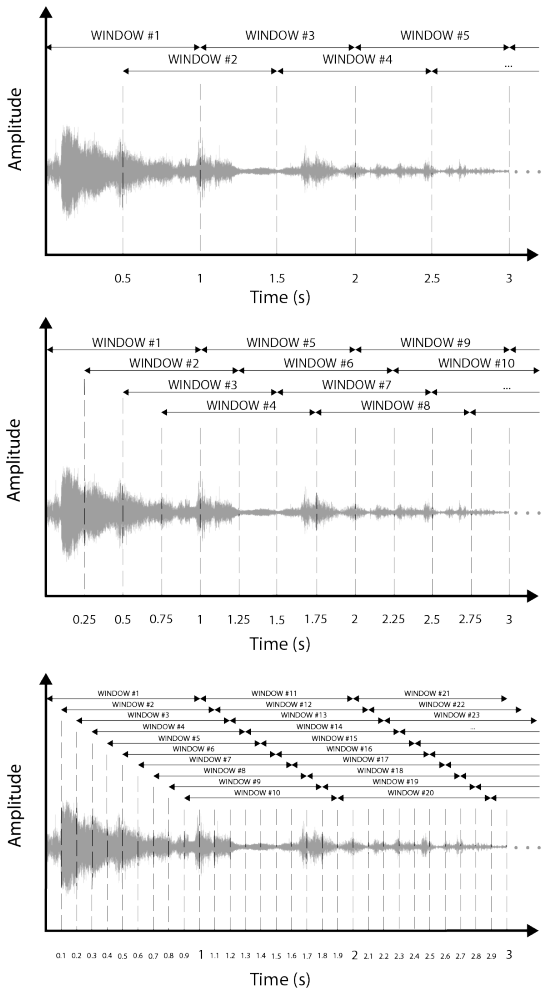

Figure 2: Three different overlaps at inference time to detect vocalizations. On top, an overlap of 0.5 seconds, in the middle, an overlap of 0.25 seconds, and in the bottom, an overlap of 0.1 seconds.

The motivation behind doing these three experiments was to evaluate whether incorporating data recorded in the same farm improve the scores of the classifier.

3. RESULTS

3.1. Post-processing technique

Table 2 shows the results of the experiment regarding the overlapping times. As it can be observed, selecting different overlapping has a huge impact on results. While the F1-score is more or less maintained (achieving its highest value with an overlap of 0.25 s), the Precision and Recall vary substantially. The biggest overlap (0.5 s) results in higher Precision and lower Recall, while the smallest overlap (0.1 s) results in lower Precision and higher Recall.

Having a more precise system means that the number of false positive events is lower. Therefore, the presented results show that a wider overlap filters more false positive events.

Analogously, having a system with a higher Recall suggests that

Overlap	F1-score	Precision	Recall
0.1 s	57.4%	46.87%	**74.05%**
0.25 s	**61.7%**	63.24%	60.24%
0.5 s	53.68%	**77.74%**	40.99%

Table 2: Precision, Recall and F1-score obtained by varying the overlapping window for the vocalizations detection.

there are fewer false negative events (i.e., that fewer vocalizations are missed). A smaller overlap, even if less precise, decrements the number of vocalizations that are mistakenly confused by noise.

As the aim of the presented algorithm is to detect vocalisations that can be further analysed by AWEC veterinary experts, the smallest overlap (0.1 s) was selected to detect all possible vocalisations, even if some of them are false positives that need to be manually removed. Therefore, for the following experiments, the post-processing stage was carried out with the overlap of 0.1 s.

3.2. Using data from different farms for training

Farm training data	F1-score	Precision	Recall
Both Farms	57.4%	46.87%	**74.05%**
Girona	50.58%	41.51%	64.72%
Valencia	**59.25%**	**54.16%**	65.39%

Table 3: F1-score, Precision and Recall of the three experiments.

Three different set-ups were evaluated, using 20 audio files recorded in Valencia as test set. As it can be seen in Table 3, the best F-score (59.25%) is obtained in the experiment that contains only audio files from Valencia. However, the highest Recall (74.05%) was obtained when using audio files from both farms for training.

Nevertheless, the results obtained when using data from Girona only are not very different from those in which Valencia audios are used.

This leads to the conclusion that using audio data from the same farm that is being evaluated is desirable, but not completely necessary to have moderatly good results (note that there is a difference of 8,67% of F1-score only between the best and the worst system).

4. CONCLUSIONS

This paper addresses the problem of automatically detecting the vocalizations of cows for further analysis by veterinary researchers, as these vocalizations can be an indicator of their welfare.

The developed algorithm uses a lightweight deep learning architecture that can run over a low-cost platform. Two experiments have been conducted, using data collected from two different dairy farms (Girona and Valencia) and manually labelling it.

The first experiment aimed at determining the optimal overlap time for vocalization detection. It was observed that the chosen overlap time correlated with the Precision and Recall metrics of the system. The system with the highest Recall was achieved when using the smallest overlapping time, resulting in more overlapped windows.

The second experiment focused on assessing the model's ability to generalize and classify vocalizations from different farms. Moderately improved results were observed when utilizing training data collected from the farm under monitoring. In fact, the best result (F-score of 59.25%) was obtained when using data solely from one farm (the same one used for both training and testing). However, the performance improvement was only 8.67% compared to the worst result, which involved training with data from one farm and testing on data from the other farm. These findings suggest that vocalisation detection generalisation is possible, even when operating in a farm without previously recorded samples.

In future research, we plan to incorporate data from additional farms to validate the conclusions drawn in this study in diverse environmental settings. In terms of the CowTalkPro project, once the veterinary researchers have analyzed the automatic vocalizations detected by the algorithm in multiple environments and the acoustic sensors are deployed in the farms, it will be necessary to study how can the real-time system assist both veterinary researchers and farmers to improve the welfare of cows.

5. ACKNOWLEDGMENT

This research is being supported by the project SNEO-20211301 "Desarrollo de un sensor de sonido en vacas para evaluar su salud y el bienestar animal", a NEOTEC funding awarded to AWEC Advisors together with HER-La Salle research team. The authors would also like to thank the Departament de Recerca i Universitats (Generalitat de Catalunya) under Grant Ref. 2021-SGR-01396 for the funding of Human-Environment Research (HER) research group.

6. REFERENCES

[1] P. Llonch, E. Mainau, I. R. Ipharraguerre, F. Bargo, G. Tedó, M. Blanch, and X. Manteca, "Chicken or the egg: the reciprocal association between feeding behavior and animal welfare and their impact on productivity in dairy cows," *Frontiers in veterinary science*, vol. 5, p. 305, 2018.

[2] B. Gołebiewska, M. Gebska, and J. Stefańczyk, "Animal welfare as one of the criterion determining polish consumers'decisions regarding their purchase of meat," *Acta Scientiarum Polonorum. Oeconomia*, vol. 17, no. 3, pp. 13–21, 2018.

[3] D.-H. Jung, N. Y. Kim, S. H. Moon, C. Jhin, H.-J. Kim, J.-S. Yang, H. S. Kim, T. S. Lee, J. Y. Lee, and S. H. Park, "Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering," *Animals*, vol. 11, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2076-2615/11/2/357

[4] V. Exadaktylos, M. Silva, and D. Berckmans, "Chapter automatic identification and interpretation of animal sounds, application to livestock production optimisation," 2014.

[5] S. Ntalampiras, A. Pezzuolo, S. Mattiello, M. Battini, and M. Brščić, "Automatic detection of cow/calf vocalizations in free-stall barn," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, 2020, pp. 41–45.

[6] G. Meen, M. Schellekens, M. Slegers, N. Leenders, E. van Erp-van der Kooij, and L. Noldus, "Sound analysis in dairy cattle vocalisation as a potential welfare monitor," *Computers and Electronics in Agriculture*, vol. 118, pp. 111–115, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169915002549

[7] G. Özmen, İ. A. Ozkan, I. Seref, S. Tasdemir, Ç. Mustafa, and E. Arslan, "Sound analysis to recognize cattle vocalization in a semi-open barn," *Gazi Mühendislik Bilimleri Dergisi*, vol. 8, no. 1, pp. 158–167, 2022.

[8] *H5 Handy Recorder - Operation Manual*, Zoom Corporation, 2014.

[9] D. Sinha and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," in *2019 IEEE 10th annual ubiquitous computing, electronics & mobile communication conference (UEMCON)*. IEEE, 2019, pp. 0280–0285.

[10] R. P. Foundation, "Raspberry pi," https://www.raspberrypi.com (accessed on 12 Jun 2023).

[11] E. Vidaña-Vila, J. Navarro, D. Stowell, and R. M. Alsina-Pagès, "Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors," *Sensors*, vol. 21, no. 22, p. 7470, 2021.

[12] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016. [Online]. Available: https://www.mdpi.com/2076-3417/6/6/162