COWTALK: DEEP LEARNING-BASED ACOUSTIC MONITORING OF CATTLE FOR VOCAL ACTIVITY DETECTION

Ignasi Nou-Plana[1], Ester vidaña-Vila[1], Marc Freixes[1], Rosa Ma Alsina-Pagès[1]

[1] *Human-Environment Research (HER), La Salle - Universitat Ramon Llull, Barcelona, Spain*

Abstract

Effective management of respiratory diseases in cattle (Bos taurus), particularly within the context of Bovine Respiratory Disease Complex (BRDC), relies on early, non-invasive monitoring strategies to safeguard animal health and welfare. Among the earliest observable indicators, vocal and respiratory-related sounds such as coughs and vocalizations provide valuable insight into physiological and behavioural states. These acoustic cues offer strong potential for continuous, automated surveillance in precision livestock farming systems. In this study, we present a deep learningbased framework for the detection of cow vocal activity using environmental audio recordings collected from multiple farm settings. The proposed system employs convolutional neural networks operating in real time to identify vocal events. Unseen recordings were used to evaluate the robustness and generalization capabilities of the approach. Results indicate that incorporating data from heterogeneous farm environments enhances model generalization, achieving an F1-score of 57.40% and a recall of 74.05%. While models trained on single-farm data can reach higher peak performance under matched conditions, cross-farm training yields more stable behaviour across varying acoustic contexts. Additionally, detection performance is shown to depend strongly on temporal segmentation strategies, highlighting the importance of parameter selection for deployment. This work demonstrates the feasibility of scalable, noninvasive acoustic monitoring systems for livestock, supporting early detection of health and behavioural changes. Such approaches contribute to the development of autonomous tools for improving animal welfare and farm management practices.

Corresponding author: *ignasi.nou@salle.url.edu.*

Keywords: *acoustic monitoring, animal welfare, sound event detection*

1. Introduction

Animal welfare is increasingly recognized as both an ethical priority and a determinant of animal growth, reproductive performance and overall farm productivity [1]. Consumer demand for welfare-conscious animal production has grown in parallel, further motivating the search for monitoring methods that do not compromise the physical integrity of the animals being observed [2]. Among the indicators available to assess welfare non-invasively, cattle vocalizations can be particularly informative: cows and calves vocalize in response to pain, stress, hunger and respiratory discomfort, providing information regarding their condition without the need for physical handling [2], [3].
Manually reviewing continuous farm audio to identify these events is, however, extremely labour-intensive: a single microphone can generate many hours of recording per day, only a small fraction of which typically contains events of interest. This has motivated a growing body of work on automatic acoustic event detection for livestock, ranging from handcrafted-feature classifiers to deep neural networks [3], [4]. Prior work on automatic vocalization detection has generally been validated on a single farm, which limits confidence in how such systems generalize once deployed elsewhere [4]. Differences in barn acoustics, herd composition, background noise sources (feeding equipment, ventilation, human activity) and microphone placement can all shift the acoustic distribution a model encounters at deployment time relative to what it was trained on, and this domain shift is rarely characterized quantitatively in the existing literature.
This work is carried out under the CowTalkPro project, a collaboration between engineers at La Salle Campus Barcelona (Ramon Llull University) and veterinarians and researchers at AWEC Advisors S.L., which aims to build a single acoustic sensor that can be deployed across multiple farms for real-time welfare monitor-

ing. The system targets three periods of particular veterinary relevance: the first weeks of a calf's life, when clusters of coughing may signal the onset of respiratory illness and warrant early intervention; the dry-off period, when cows may vocalize due to pain or discomfort as milk production is halted ahead of calving; and calving itself, when vocal activity can indicate whether farmer assistance is needed. Meeting these goals requires an algorithm that is both accurate across heterogeneous farm environments and light enough to run in real time on low-cost embedded hardware, since a commercially viable sensor cannot depend on continuous connectivity to an off-site server for inference.

This paper condenses and updates our earlier evaluation of a convolutional neural network (CNN) for cow vocalization detection [5], focusing on (i) the effect of temporal segmentation strategy on detection performance, (ii) the extent to which the model generalizes to unseen farm environments, and (iii) the practical requirements for real-time, on-farm deployment. The remainder of this paper is organized as follows: Section 2 describes the data collection campaign, the detection model, and the experimental design; Section 3 reports the results of both experiments; and Section 4 discusses the implications of these findings and outlines directions for future work.

2. Materials and Methods

2.1. Data collection

Audio was recorded on two commercial dairy farms in Spain: Girona and Valencia. At each site, a mains-powered Zoom H5 recorder, housed inside a protective box, was connected to an omnidirectional microphone via a 30 m XLR cable and suspended from the ceiling over the animals' yard, sampling continuously at 44.1 kHz. Two microphones were installed at each farm; in Valencia, both were placed in a single large calf yard roughly 50 m apart, while in Girona one covered a calves' yard and the other covered dairy cows during the dry-off period. The hardware ran continuously for approximately one year at each site; due to the 32 GB storage limit of the recorder's SD card, each card held around four and a half days of continuous audio before requiring manual replacement.

From this year-long deployment, a subset of 199 recordings of 15 minutes each - 40 files from cows and 79 from calves in Girona, and 80 calf files from Valencia, roughly 50 hours in total - was manually annotated by two annotators under veterinary supervision, using the Audacity software to mark the onset and offset of each event. Two label categories were used: vocalizations and coughs. Twenty Valencia recordings, chosen for their higher density of vocalizations relative to Girona, were reserved as a held-out test set; the remaining files formed the training pool (Table 1). An example of the microphone installation is shown in Fig. 1.

As Table 1 shows, the test set is dominated by vocalizations (1,756 instances) relative to coughs (129 instances). This imbalance reflects the underlying biology of the recorded animals: cows and calves vocalize considerably more often than they cough, particularly when the herd is healthy, and coughing episodes tend to cluster around specific health events rather than occurring uniformly throughout a recording.

Figure 1: Microphone installed over the cows' yard in Girona.

Table 1: Label counts per dataset.

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

2.2. Detection model

Vocalizations were detected with a MobileNet CNN [6], chosen for its lightweight architecture and consequent suitability for future deployment on low-cost embedded devices such as a Raspberry Pi. As inputs to the network, 1-second log-mel spectrograms were used, following the window length adopted in our previous acoustic event classification work [7]. The network was trained as a multilabel classifier, since more than one acoustic event can occur within a single 1-second window - for instance, one animal vocalizing while another coughs in the background - and the two output units of the network correspond to the presence or absence of a vocalization and a cough, respectively. Training was carried out for 15 epochs, with early stopping used to retain the model checkpoint with the lowest validation loss.

At inference time, and unlike the training stage (in which non-overlapping 1-second windows were used), the CNN was concatenated with a post-processing stage responsible for delimiting the onset and offset of each detected vocalization. This was achieved by sliding an overlapping analysis window of 1 second across the recording and aggregating the per-window predictions into continuous event segments, with the degree of overlap between consecutive windows treated as a tunable parameter, evaluated in Section 2.3.

2.3. Temporal segmentation and cross-farm evaluation

Two experiments assessed the system. The first compared three inference-time window overlaps (0.1 s, 0.25 s and 0.5 s), training on data pooled from both farms and evaluating on the Valencia test set. Performance was quantified using segment-based sed_eval metrics [8], with a configuration of t_collar = 0.9 and percentage_of_length = 0.1. The t_collar parameter defines a tolerance, in seconds, around the ground-truth event boundary within which an estimated boundary is still considered correct, while percentage_of_length expresses this tolerance as a fraction of the event's duration; together they determine how strictly a predicted onset/offset pair must match the annotated event to be scored as a true positive.

The second experiment used the same Valencia test set - entirely unseen during training in two of the three configurations - to quantify how well the model generalizes across farms, comparing training on (a) both farms, (b) Girona only (a different farm from the test set, and therefore a genuine test of cross-farm generalization) and (c) Valencia only (the same farm as the test set, representing the best-case, farm-matched scenario). Between the two experiments, the inference-time overlap that performed best in Section 2.3's first experiment was retained, so that the second experiment isolates the effect of training-data provenance alone.

2.4. Towards real-time deployment

A key design constraint for CowTalkPro is that the final sensor must run unattended on-farm, which motivated the choice of MobileNet over heavier architectures such as full-scale convolutional networks originally designed for large-scale image classification. A comparable low-cost single-board setup was previously benchmarked within our group for multilabel acoustic event classification under physical sensor redundancy in urban settings [7], indicating that this class of lightweight architecture is compatible with the computational budget of low-cost embedded hardware such as a Raspberry Pi. This supports the feasibility of integrating the MobileNet-based pipeline presented here into a standalone, real-time CowTalkPro sensor, which is the final goal of the wider project.

3. Results

3.1. Effect of temporal segmentation

Table 2 shows that the overlap used at inference time strongly shapes the precision/recall trade-off, while F1-score is comparatively stable across the three configurations tested. The narrowest overlap (0.1 s) maximizes recall (74.05%) at the cost of precision, favouring the detection of more candidate vocalizations for later expert review; the widest overlap (0.5 s) instead filters false positives, raising precision to 77.74% but lowering recall to 40.99%. The intermediate overlap (0.25 s) achieves the best balance between the two, with the highest F1-score of the three (61.7%).

Table 2: Effect of inference-time overlap on detection performance.

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%

Having a more precise system means that the number of false positive events is lower; the presented results therefore show that a wider analysis window overlap filters more false positive events, at the expense of missing a larger proportion of true vocalizations. Analogously, a smaller overlap, even if less precise overall, decrements the number of vocalizations that are mistakenly missed due to noise or partial occlusion by other sounds. Given the intended use of the system - flagging candidate vocalizations for subsequent manual review by veterinary experts, rather than fully automated diagnosis - the 0.1 s overlap was retained for the second experiment, since in this use case a higher recall (fewer missed vocalizations) is preferable even at the cost of a larger number of false positives that a human reviewer can subsequently discard.

3.2. Cross-farm generalization

Table 3 reports the cross-farm generalization experiment. Training on Valencia data alone gave the best F1-score (59.25%), while training on both farms gave the highest recall (74.05%). Training on Girona only - a farm entirely unseen relative to the test set - still reached an F1-score of 50.58%, only 8.67 percentage points below the best configuration.

Table 3: Effect of training-data provenance on detection performance for the Valencia test set.

Training data	F1	Prec.	Rec.
Both farms	57.4%	46.87%	74.05%
Girona only	50.58%	41.51%	64.72%
Valencia only	59.25%	54.16%	65.39%

This suggests that although farm-specific training data is preferable, the model retains moderately good performance on previously unseen recordings, supporting its use across new farm deployments without requiring site-specific retraining. Somewhat surprisingly, training on Girona data alone - despite Girona and Valencia differing in animal composition (Girona includes both cows and calves, whereas Valencia contains only calves) and in microphone placement - yields results reasonably close to those obtained with Valencia-only training. This indicates that the acoustic features distinguishing vocalizations and coughs from background noise are, to a meaningful extent, shared across farms and animal categories, rather than being highly farm-specific.

4. Discussion and Conclusions

This work condenses two experiments on CNN-based detection of cow vocalizations across multiple farms. First, the choice of inference-time overlap materially changes the precision/recall balance: a narrow 0.1 s overlap prioritizes recall, which is appropriate when the detector's output feeds into downstream veterinary review, whereas a wider 0.5 s overlap prioritizes precision, which may be preferable in settings where false alarms carry a higher operational cost (for example, if the system triggers an automated alert rather than a manual check). Selecting the operating point therefore depends on how the detector's output will ultimately be used, rather than being a purely technical choice.

Second, cross-farm evaluation shows that while same-farm training data yields the best F1-score, a model trained entirely on a different farm still generalizes reasonably well to unseen recordings, with only an 8.67% F1-score gap to the best configuration. This is a result for the CowTalkPro project's goal of a single, farm-agnostic sensor: it suggests that a model trained on a diverse, multi-farm dataset could be deployed to a new farm without requiring a costly site-specific data collection and annotation campaign, even though some accuracy is sacrificed relative to a fully farm-matched model. Combined with the lightweight MobileNet architecture, and the precedent for real-time deployment of comparable architectures on low-cost embedded hardware within our group [7], these results support the feasibility of a scalable, non-invasive CowTalkPro sensor for continuous, on-farm welfare monitoring.

Several limitations should be noted. The evaluation presented here involves only two farms, and both are located in Spain; validating the cross-farm generalization findings on a larger and more geographically diverse set of farms, with different barn constructions, herd sizes, and background noise profiles, would strengthen confidence in the approach's real-world robustness. Additionally, the current post-processing stage treats vocalizations and coughs as independent binary outputs per window; incorporating temporal context across consecutive windows, for instance through a recurrent or attention-based post-processing layer, could further improve boundary estimation without abandoning the lightweight per-window CNN backbone. Future work will extend validation to additional farms, explore such temporal post-processing refinements, and complete the integration of the detection pipeline into a real-time embedded prototype suitable for continuous deployment in commercial dairy operations.

Acknowledgments

This research is supported by the project SNEO-20211301 "Desarrollo de un sensor de sonido en vacas para evaluar su salud y el bienestar animal", NEOTEC funding awarded to AWEC Advisors together with the HER-La Salle research team, and by the Departament de Recerca i Universitats (Generalitat de Catalunya) under Grant Ref. 2021-SGR-01396.

References

[1] P. Llonch et al., "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] V. Exadaktylos and M. Silva, "Automatic identification and interpretation of animal sounds, application to livestock production," *Soundscape Semiotics: Localization and Categorization*, p. 65, 2014.

[3] D.-H. Jung et al., "Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering," *Animals*, vol. 11, no. 2, p. 357, 2021.

[4] 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)*, IEEE, 2020, pp. 41–45.

[5] E. Vidana-Vila et al., "Automatic detection of cow vocalizations using convolutional neural networks," *Tampere, Finland*, 2023.

[6] 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*, vol. 7, p. 8, 2019.

[7] 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.

[8] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.