The 11th Asian-Australasian Conference on Precision Agriculture (ACPA 11)
October 14-16, 2025, Chiayi, Taiwan

Deep Learning-Based Vocal Signature Analysis for Rumination Belching Identification

Yu-Hsuan Shen[1], Kai-Feng Cheng[1], Jin-Chern Chiou[2], Yi-Chich Chiu[1], Chin-Cheng Wu[1,*]

[1] Department of Biomechatronic Engineering, National Ilan University, Taiwan.
[2] Department of Electronics and Electrical Engineering, National Yang Ming Chiao Tung University, Taiwan

* Corresponding Author: chincheng@niu.edu.tw

ABSTRACT

Animal vocalizations provide important information about various behaviors, including courtship, parental care, foraging, and vigilance, as well as physiological states such as estrus, parturition, stress, and disease. In ruminants, eructation-related sounds are primarily produced when methane, generated by microbial fermentation of plant fibers in the rumen, is released orally. The acoustic characteristics of these sounds—such as frequency, intensity, and duration—are closely associated with the volume and flow rate of gas emissions.
This study developed a cattle vocalization classification method based on convolutional neural networks (CNNs) and Mel-frequency cepstral coefficient (MFCC) features to identify the intensity and duration of rumination behaviors. The model demonstrated high performance across different configurations, achieving a maximum classification accuracy of 97%. The proposed system has potential applications in precision livestock farming (PLF) for non-invasive behavior monitoring and the development of low-carbon feeding strategies.

Keywords: deep learning, rumination, methane, eructation

INTRODUCTION

The rapid development of sustainable agriculture and precision livestock farming (PLF) technologies has fostered a critical research focus on the non-invasive monitoring of animal behaviors and physiological states. For ruminants, behaviors such as chewing and eructation are accompanied by specific acoustic signatures. Notably, when methane produced by microbial fermentation of fibrous feed in the rumen is emitted through eructation, distinct sound patterns are generated. Recent findings indicate that the frequency distribution, intensity, and duration of these sounds strongly correlate with the volume and flow rate of gas release. Therefore, effective interpretation of such acoustic signals has the potential to facilitate rumination behavior recognition and serve as a proxy for greenhouse gas emissions, thereby supporting the design and monitoring of low-carbon feeding strategies.
Recent studies have adopted deep learning and advanced feature extraction techniques to overcome the limitations of traditional threshold- and rule-based approaches under noisy conditions. For example, Li et al. (2021) have classified three feeding behaviors in dairy cows (bite, chew, and chew-bite) using acoustic signals while comparing various deep learning architectures [1]. These results show that LSTM models effectively capture the temporal features of behavioral sounds, achieving classification performance above 0.93; nevertheless,

accuracy varies with forage type and height. Chelotti et al. (2023) have proposed the JMFAR method, which integrates statistical and spectral features of jaw movement sounds to address the shortcomings of prior RAFAR and BUFAR algorithms in distinguishing grazing and rumination, demonstrating enhanced real-time monitoring capabilities [2]. Furthermore, Ferrero et al. (2023) developed Deep-Sound, an end-to-end deep learning approach that processes raw audio inputs using CNNs and bidirectional GRUs for acoustic sequence modeling and incorporating data augmentation significantly improved model generalization, resulting in an F1-score of 79.82%, and outperforming ResNet and other advanced algorithms [3].

This study presents a CNN-based rumination behavior recognition method utilizing Mel-frequency cepstral coefficients (MFCCs) as acoustic features. The system systematically analyzes model performance and parameter effects in identifying bite, chew, and rumination behaviors, thereby providing a foundation for future applications in low-carbon livestock management and non-invasive behavioral monitoring.

MATERIALS AND METHODS

1. Cattle Vocalization Dataset and Preprocessing

A publicly available dataset, Acoustic Monitoring of Dairy Cow Feeding and Rumination Behaviors [4] , which was released in 2022 by the Politecnico di Torino research team, was employed in this study. The dataset comprised two long-duration audio recordings of dairy cow activities: JM_grazing.wav (approximately 1 hour, labeled as grazing) and JM_rumination.wav (approximately 30 minutes, labeled as rumination). Corresponding .csv annotation files were used to provide the onset times and categories of feeding events. The annotated behavior classes included bite (b), chew during grazing (c), chew during rumination (r), and chew-bite (x). The raw audio recordings were segmented using 1-second windows with 50% overlap (stride = 0.5 s), resulting in a total of 6,454 audio clips: chew-bite (2,683), chew during rumination (2,347), chew during grazing (953), and bite (471). Each segment was labeled according to the center timestamp within the annotated interval, and unannotated segments were discarded.

2. Acoustic Feature Extraction

The Mel-frequency cepstral coefficients (MFCCs) were adopted as the primary acoustic features. MFCCs have been widely used in speech recognition because of their effectiveness in capturing the perceptual characteristics of the human auditory system by mapping the frequency spectrum onto the Mel scale. They were considered efficient, robust, and generalizable, and have become a standard representation for deep learning models in audio processing.

In this study, audio signals were sampled at 16 kHz, with a pre-emphasis coefficient of 0.97, a frame length of 512 samples (approximately 32 ms), and an FFT size of 1024. A 13-dimensional MFCC representation was extracted, and zero-padding was applied to standardize tensor shapes. Figure 1 illustrated the 13-dimensional MFCC features of a chew-bite instance. To evaluate the effect of pre-emphasis filtering on recognition performance, two configurations were compared: with and without pre-emphasis. Pre-emphasis was applied to the time-domain signal prior to feature extraction using a first-order high-pass filter to amplify high-frequency energy and reduce spectral tilt, defined as follows.

$$y[n] = x[n] - \alpha\, x[n-1], \alpha = 0.97 \quad (1)$$

Where $x[n]$ is the raw signal and $y[n]$ is the pre-emphasized signal.

The pre-emphasis condition employed α=0.97, while the non-pre-emphasis condition computed MFCCs directly from the raw signal. All other feature parameters were kept identical across both conditions.

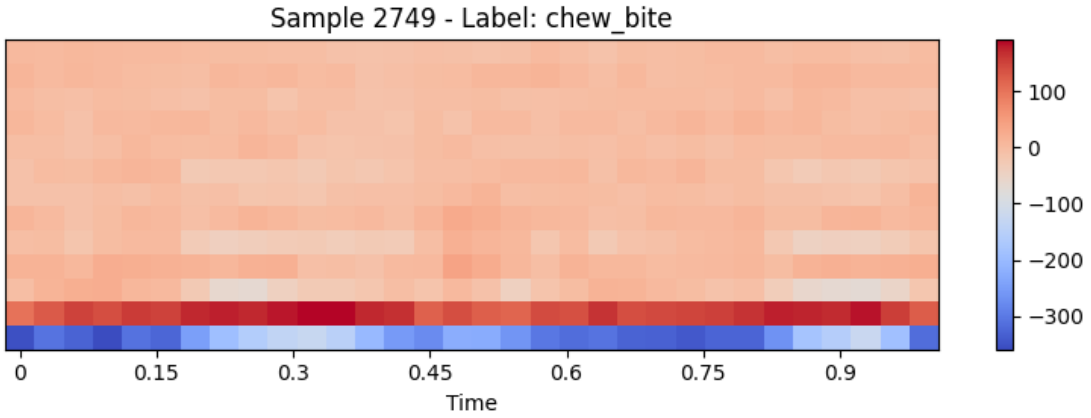

Figure 1、13-dimensional MFCC features of a chew-bite instance.

3. Model Architecture

A convolutional neural network (CNN) was employed for the classification task. The network consisted of two 2D convolutional layers with 32 and 64 filters, respectively. Each convolutional layer was followed by a max-pooling layer to reduce dimensionality while retaining salient features. The output of the convolutional blocks was flattened and passed to a fully connected layer with 128 neurons for higher-level feature integration. A dropout layer with a rate of 0.3 was applied after the fully connected layer to mitigate overfitting. The output layer used the softmax activation function to produce probability distributions across the four behavior classes.

4. Model Training and Evaluation

The dataset was split into training, validation, and testing sets with an 80/10/10 ratio. The model was trained for 30 epochs using the Adam optimizer and the categorical cross-entropy loss function. Model performance was evaluated on the test set using overall accuracy and class-wise F1-scores as primary metrics.

Class Weighting Comparison:

To address class imbalance in the dataset, experiments were conducted both with and without class weighting. The class weights were computed using the class_weight parameter set to "balanced" in Scikit-Learn [5], which automatically assigns weights W_C based on class frequencies:

$$Wc = \frac{N}{K \cdot n_c} \tag{2}$$

where N is the total number of samples, K is the number of classes, and n_c is the number of samples in class c. The weights were incorporated into the cross-entropy loss, effectively yielding a weighted cross-entropy function.

RESULTS & DISCUSSION

The performance of the presented CNN-MFCC model was summarized in Table 1. Overall test accuracy ranged from 96% to 97%, while macro F1-scores were between 0.94 and 0.96, which indicated that the proposed framework exhibited strong classification capability.

Table 1. The performance of the CNN-MFCC model

Condition	Bite F1	Chew-bite F1	Grazing F1	Rumination F1	Macro F1	Acc
No class weighting, no pre-emphasis	0.93	0.97	0.93	1.00	0.96	0.97
No class weighting, with pre-emphasis	0.88	0.96	0.92	1.00	0.94	0.96
With class weighting, no pre-emphasis	0.91	0.96	0.93	1.00	0.95	0.97
With class weighting, with pre-emphasis	0.90	0.96	0.91	0.99	0.94	0.96

Performance by Behavior Class

Among the four classes, chew during rumination (r) achieved the most consistent results, as both precision and recall were close to 1.00 under all experimental settings. This finding suggested that rumination sounds possessed distinctive structural and spectral patterns, which made them highly recognizable to the model. The chew-bite (x) class also exhibited high classification accuracy, with F1-scores ranging from 0.96 to 0.97, indicating that its mixed characteristics did not lead substantial confusion. In contrast, bite (b) and chew during grazing (c) showed slightly lower F1-scores, between 0.88–0.93 and 0.91–0.93, respectively. This reduction could be attributed to the predominance of high-frequency components and similar waveform distributions, together with the relatively small sample size of the bite class. Consistent with this, the confusion matrix indicated some degree of misclassification between bite and chew behaviors.

Effects of Class Weighting and Pre-emphasis

Class weighting improved performance on underrepresented categories, such as the bite class, particularly with respect to recall and F1-score. However, because the dataset was not severely imbalanced (bite accounted for ~7%), the overcompensation slightly reduced the macro F1-score, resulting in only marginal differences in overall performance. Although pre-emphasis was widely used in speech recognition to enhance high-frequency components, it did not produce consistent improvements for bovine chewing sounds. In some cases, pre-emphasis slightly degraded performance in the classification of bite and grazing behaviors.

CONCLUSIONS

This study developed a CNN-based cattle sound classification model using MFCC features to identify rumination-related behaviors, including bite, chew during grazing, chew-bite, and chew during rumination. The model achieved an overall accuracy of 96–97% and macro F1-scores ranging from 0.94 to 0.96. Experimental results showed that chew during rumination exhibited the most distinct and stable acoustic patterns, yielding near-perfect classification performance. The chew-bite class was also highly distinguishable, whereas bite and chew during grazing achieved slightly lower performance due to smaller sample sizes and similar spectral characteristics. Class weighting provided some benefits for underrepresented categories but had a limited impact overall, while pre-emphasis did not produce consistent advantages.

In summary, the proposed system demonstrates the potential of deep learning for non-invasive

monitoring of animal acoustic behaviors. Future work may incorporate PCEN features and temporal sequence modeling methods (e.g., LSTM) to enhance temporal dynamics analysis and cross-domain generalizability, thereby facilitating the practical deployment of acoustic monitoring technologies in low-carbon livestock management.

ACKNOWLEDGEMENTS

This study is financially supported by the National Science and Technology Council of Taiwan, Republic of China, under Grant No. NSTC 114-2222-E-197-002 -.

REFERENCES

[1] G. Li, Y. Xiong, Q. Du, Z. Shi, and R. S. Gates, “Classifying Ingestive Behavior of Dairy Cows via Automatic Sound Recognition,” Sensors, vol. 21, no. 15, p. 5231, Aug. 2021, doi: 10.3390/s21155231.

[2] J. O. Chelotti, M. Ferrero, G. Comba, D. Milone, and G. Rufiner, “Using segment-based features of jaw movements to recognise foraging activities in grazing cattle,” Biosyst. Eng., vol. 230, pp. 100–112, May 2023, doi: 10.1016/j.biosystemseng.2023.03.014.

[3] M. Ferrero, J. O. Chelotti, G. Comba, D. H. Milone, and G. L. Rufiner, “A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle,” Eng. Appl. Artif. Intell., vol. 121, p. 106016, 2023, doi: 10.1016/j.engappai.2023.106016.

[4] L. S. Martinez-Rau et al., 'Daylong acoustic recordings of grazing and rumination activities in dairy cows,' Sci. Data, vol. 10, p. 782, 2023

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.