Short-term feeding behaviour sound classification method for sheep using LSTM networks

Guanghui Duan[1], Shengfu Zhang[2], Mingzhou Lu[1*], Cedric Okinda[1], Mingxia Shen[1], Tomas Norton[3]

(1. *College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China*;
2. *Computer College, Qinghai Nationalities University, Xining 810007, China*;
3. *M3-BIORES- Measure, Model & Manage Bioresponses, KU Leuven, Heverlee/Leuven 3001, Belgium*)

Abstract: A deep learning approach using long-short term memory (LSTM) networks was implemented in this study to classify the sound of short-term feeding behaviour of sheep, including biting, chewing, bolus regurgitation, and rumination chewing. The original acoustic signal was split into sound episodes using an endpoint detection method, where the thresholds of short-term energy and average zero-crossing rate were utilized. A discrete wavelet transform (DWT), Mel-frequency cepstral, and principal-component analysis (PCA) were integrated to extract the dimensionally reduced DWT based Mel-frequency cepstral coefficients (denoted by PW_MFCC) for each sound episode. Then, LSTM networks were employed to train classifiers for sound episode category classification. The performances of the LSTM classifiers with original Mel-frequency cepstral coefficients (MFCC), DWT based MFCC (denoted by W_MFCC), and PW_MFCC as the input feature coefficients were compared. Comparison results demonstrated that the introduction of DWT improved the classifier performance effectively, and PCA reduced the computational overhead without degrading classifier performance. The overall accuracy and comprehensive F1-score of the PW_MFCC based LSTM classifier were 94.97% and 97.41%, respectively. The classifier established in this study provided a foundation for an automatic identification system for sick sheep with abnormal feeding and rumination behaviour pattern.

Keywords: sheep behaviour, short-term feeding behaviour, acoustic analysis, Mel-frequency cepstral coefficients, long-short term memory networks
DOI: 10.25165/j.ijabe.20211402.6081

Citation: Duan G H, Zhang S F, Lu M Z, Okinda C, Shen M X, Norton T. Short-term feeding behaviour sound classification method for sheep using LSTM networks. Int J Agric & Biol Eng, 2021; 14(2): 43–54.

1 Introduction

Healthy sheep has a stable daily rhythm of grass intake at pasture[1]. Given that abnormalities in this rhythm can indicate health disorder, accurate measure of a sheep's daily grass intake can thus be utilized to assess animal's healthy status. Previous studies have reported a good correlation between the grass intake of individual ruminant animals and different short-term feeding behaviours[2-7]. The short-term feeding behaviour is that accomplished with discrete jaw movements, such as a chew, bite, or simultaneously chew and bite (abbreviated as chew-bite) on the same jaw opening-closing cycle[5].

It was suggested that chewing energy per bite and total amount of energy in chewing sounds during ingestion were the most important predictors of the dry matter intake (DMI) of sheep[2]. It was also found that the best estimation of grass intake of cows can be achieved when grazing time and bite frequency were used as predictors[3]. On the other hand, rumination chewing frequency (chews per minute during rumination) were suggested to be the most significant explanatory variable of feed intake of cows[4]. In a later study[5], researchers reinforced the idea of applying generalized sound-based predictions of DMI, using the chewing sound energy as the main predictor. A fibre intake prediction model was established for goats with R^2 higher than 0.867, using a signal feature termed slope sign change, which could express signal frequency characteristics that indicate physiological aspects of bite and chewing[6]. Thus, in summary, it is evident in previous research that the most relevant explanatory variables of intake estimation for ruminants can be extracted from measurements of short-term feeding behaviour, including ingestion bite, ingestion chew, and rumination chew.

It is a challenge to monitor short-term feeding behaviour of ruminants by manual observation for animal behaviour analysis[8]. Alternatively, wearable pressure sensors[4], accelerometers[3,9], electromyography signals[10], and so on, have been applied to measure ingestion bites and ingestion chewing for ruminants in a more automatic way. However, position of these wearable sensors had a significant impact on the accuracy of the feeding related behaviour identification. Sensor position shift could bring unstable, or even wrong data of the bite occurrence or time spent in feeding, which would in turn lead to a further reduction of the feed intake estimation accuracy.

The sound produced by short-term feeding behaviour of a

Received date: 2020-08-14 **Accepted date:** 2021-01-30
Biographies: **Guanghui Duan**, MS candidate, research interest: acoustic signal analysis, Email: 2018812092@njau.edu.cn; **Shengfu Zhang**, Professor, research interest: machine learning, Email: zhangsf@qhmu.edu.cn; **Cedric Okinda**, PhD candidate, research interest: data analysis and modelling, Email: cedsean@hotmail.com; **Mingxia Shen**, Professor, research interest: digital agriculture, Email: mingxia@njau.edu.cn; **Tomas Norton**, Professor, research interest: precision livestock farming, Email: tomas.norton@kuleuven.be
***Corresponding author: Mingzhou Lu**, Professor, research interest: intelligent computing of animal husbandry data. College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210031, China. Tel: +86-13813841336, Email: lmz@njau.edu.cn.

ruminant animal contains a wealth of valuable information, such as grass biting, chewing, and rumination[1]. Moreover, a collar-mounted microphone can record all the acoustic signal around an animal mouth during the ingestion and rumination process without altering the animal's normal behaviour[11]. Slight shifts in the position of a collar do not make much difference to the quality of an animal's ingestion and rumination audio signal. As a result, intake estimation for ruminants based on acoustic features in short-term feeding behaviour attracted great attention among researchers[2,12].

Identifying the sound produced by short-term feeding behaviour has previously been reported in the literatures[13-18]. Various methods, such as hidden Markov models[13], support vector machine[16], decision tree[17], random forest[17], radial basis function network[17], linear discriminant analysis[18], and so on, were employed as the foundation techniques to classify short-term feeding behaviour. However, to the best of the authors' knowledge, distinguishing features of the sound signals associated with both ingestion chew and rumination chew were not yet defined. However, distinguishing ingestion chew from rumination chew is an important prerequisite in the development of an automatic intake estimation system based on acoustic features in short-term feeding behaviour. Therefore, this study aims to develop a classifier which can classify different sound produced by short-term feeding behaviour, including ingestion bite (*IB*), ingestion chew (*IC*), bolus regurgitation (*BR*), and rumination chew (*RC*). The former two and latter two respectively belong to ingestion and rumination.

An improved version of Recurrent Neural Network (RNN), called Long-Short Term Memory (LSTM) networks approach has a strong processing capability for sequence-type data, such as audio signals. Therefore, LSTM networks approach has been widely used in various kinds of audio analysis tasks, including speech recognition[19], acoustic modelling[20], sentence embedding[21], correlation analysis[22], and so on. Inspired by the advantage of LSTM networks in acoustic analysis, the target classifier of this study was trained using LSTM networks.

2 Materials and methods

2.1 Feeding acoustic signal data collection

Four Qinghai semi-fine wool female sheep, each being 2 years old and with a body weight of 40±2 kg, were randomly selected from a sheep farm for the short-term feeding behaviour sound data collection from 3rd July to 26th July, 2019. The farm was located in Heka Town, Xinghai County, Tibetan Autonomous Prefecture of Hainan, Qinghai Province, China. Four head-mounted collars were custom-made according to the head shapes of the selected sheep. An audio recorder (brand: Whislon, model: H29-1, sampling rate 44.1 KHz, resolution 16 bits, recording save format WAV) was fixed in one side of each collar, as shown in Figure 1. The recorder has a built-in noise reduction circuit to reduce the noise generated by the recorder itself and ensure the validity of the data.

Figure 1 Photo of a sheep with a collar and image of the audio recorder

The sheep were transferred to 4 fenced plots on 3rd July, 2019. Each plot had a dimension of 1.5 m×1.5 m×1.2 m (long × wide × high). There was no naturally growing pasture inside the plots. Four cameras (brand: SARGO, model: A8, digital pixels 20.1 million, recording save format mp4), one per plot, were employed to record sheep behaviour inside the plot. Each sheep had free access to water, which was provided in a plastic basin by the experimenters. For acclimatization purpose, each sheep started to wear the collar for four days before the audio data collection. Each sheep was fed three times every day. The duration of each meal was 9:00 to 10:00, 14:00 to 15:00, and 19:00 to 20:00, respectively. The feeding process of each meal was carried out through the following two steps:

1) An experimenter presented a handful of oat grass close to the sheep mouth.

2) The sheep then accepted the grass and began to eat. The audio recorder, which was fixed in the collar, collected and stored the original ingestion audio data during this feeding period.

The original rumination audio data was collected automatically by the audio recorder while the sheep was ruminating. At the end of data collection, 4 sheep × 25 d/sheep × 24 h/d of original audio data, including feeding (ingestion and rumination) and non-feeding audio, along with the behaviour video, were obtained and stored in a computer for the further processing.

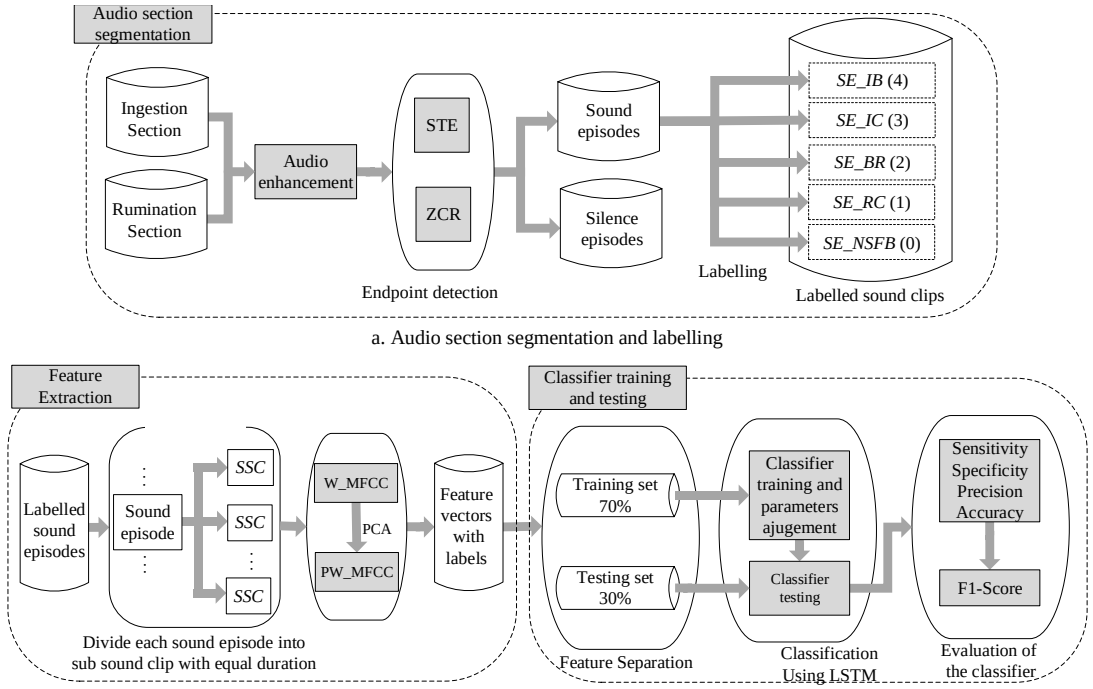
With the help of the video and an animal behaviour expert, the start and end points of each audio section corresponding to a long-term feeding behaviour were determined manually using the open source software Audacity 2.1.2[23]. Here, long-term feeding behaviour referred to an ingestion or rumination process. The former included one bite and several ingestion chews to grind up the grass mass before swallowing it. The latter included one bolus regurgitation and several rumination chews. As a result, 204 and 100 pieces of ingestion and rumination audio sections were obtained. The lengths of ingestion and rumination audio sections were 87.829±23.123 s (median ± standard deviation) and 61.752±13.103 s, respectively.

2.2 Overall structure of the proposed method

Each ingestion or rumination audio section in the dataset comprised segments (episodes) of sound and silence. Sound episodes (*SE*) consisted of audio signal produced by ingestion bite (*IB*), ingestion chew (*IC*), bolus regurgitation (*BR*), and rumination chew (*RC*), denoted by *SE_ IB*, *SE_ IC*, *SE_ BR*, and *SE_ RC*, respectively, as well as sounds of behaviour unrelated to the feeding process (denoted by *SE_NSFB*), such as vocalizations (bleating) and others sounds (caused by, e.g., collisions between sheep and the fence). The short-term feeding behaviour sound classifier established in this study was developed to classify each sound episode into one of the following categories: *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, or *SE_NSFB*. The pipeline of the classifier establishment can be structured into audio section segmentation, feature extraction and classification, as shown in Figure 2.

All the ingestion and rumination sections were split into sound and silence episodes using an endpoint detection method based on short-time energy (STE) and average zero-crossing rate (ZCR). Each obtained sound episode was assigned a label manually to indicate its category. In the next stage, each labelled sound episode was segmented into several sub-sound clips with equal duration. Then, discrete wavelet transform (DWT), Fast Fourier transform (FFT), Mel filter bank and Discrete Cosine Transform (DCT) were utilized to extract the DWT based MFCC, denoted by

W_MFCC, for each sub-sound clip. The dimension of each W_MFCC was reduced by using principal-component analysis (PCA), and the resultant coefficients matrix was denoted by PW_MFCC. Finally, the feature vectors were divided into training and testing set, which were respectively used for classifier training and testing.



Note: STE and ZCR are the abbreviation of short-time energy and average zero-crossing rate, respectively. *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB* represent the sound episodes produced by ingestion bite, ingestion chew, bolus regurgitation, rumination chew, and behaviour unrelated to feeding, respectively. The numbers in the parentheses behind *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB* are the assigned labels. Each labelled sound episode in Figure 2a was segmented into several sub-sound clips (abbreviated as *SSC*) with equal duration (be set to 4096 sample points in this study). W_MFCC and PW_MFCC represent the original and dimensionally reduced wavelet transform based Mel-Frequency Cepstral Coefficients (MFCC) extracted from each sub-sound clip.

Figure 2 Pipeline of the classifier establishment

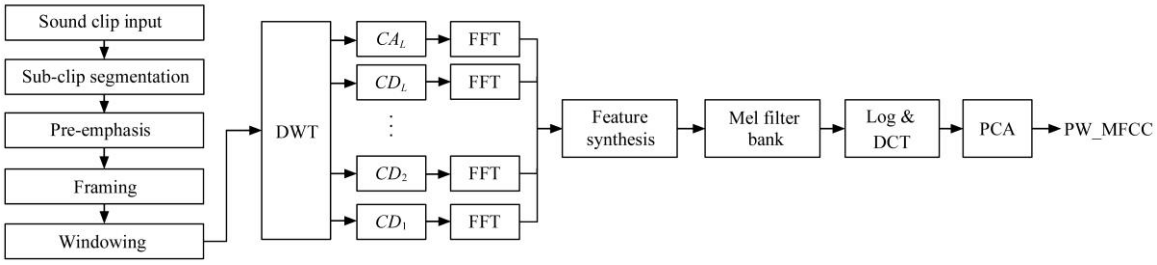
2.3 Audio section segmentation

Audio section segmentation consisted of three steps. Firstly, ingestion and rumination sections were enhanced by using Minimum Mean-Square Error Log-Spectral Amplitude Estimator (MMSE-LSA)[24]. Secondly, a double thresholds endpoint detection method, which was described in detail by Sheng et al.[25], was applied to split each enhanced audio section into sound and silence episodes. The former ones covered *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB*. Finally, each obtained sound episode in *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB* after the endpoints detection operation was respectively assigned the label 4, 3, 2, 1, and 0.

All the ingestion and rumination audio sections obtained in feeding behaviour acoustic data collection step were segmented and the number of the obtained sound episodes belonging to *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB* were 450, 6043, 100, 8826, and 1000, respectively. 70% and 30% of the sound episodes in each category respectively formed the training and testing dataset for classifier establishment. These two datasets were denoted by SE_{tr} and SE_{te}, respectively.

2.4 Feature extraction

In audio processing, Mel-Frequency Cepstral Coefficients (MFCC)[26] are the most commonly used features[27-29]. Typical step for time-frequency transformation in MFCC extraction is short time FFT (STFT). The disadvantage of STFT is that it keeps the length of the analysis window fixed for all frequencies, which leads to a resolution trade-off between time and frequency[30]. DWT offers a remedy to this difficulty by providing well localized time and frequency resolution. Therefore, the Mel-frequency cepstral and DWT were integrated to generate a feature vector for the subsequent task of sound episode classification, and the resultant feature vector was denoted by W_MFCC. As coefficients in W_MFCC may be redundant or highly correlated, PCA was employed to reduce the dimension of W_MFCC. Therefore, a PCA module was added in the end of pipeline of the extraction of the dimensionality-reduced W_MFCC, which was denoted by PW_MFCC, as shown in Figure 3.



Note: L is the number of the levels of the DWT. CD_t (t=1, 2, …, L) is the detail coefficient at level t, CA_L is the approximation coefficient at level L.

Figure 3 Block diagram of PW_MFCC extraction.

2.4.1 Sub-sound clip segmentation

Each of the sound episodes obtained in section 2.3 was segmented into sub-sound clips with *len_ssc* (set to be 4096 in this study) sample points. The last sub-sound clip in each sound episode was discarded if its length was less than *len_ssc*.

2.4.2 Pre-emphasis

Pre-emphasis was applied to compensate for the high frequency part using the following filter:

$$SSC'(n) = SSC(n) - \alpha \cdot SSC(n-1) \quad (1)$$

where, *SSC* and *SSC′* are the sub-sound clip signal before and after the pre-emphasis operation; *n* is the index of each sample in *SSC*; α is an adjustable parameter which was normally assigned a value in the range of [0.9, 1].

2.4.3 Framing and windowing

Each *SSC′* was divided into frames, each of which had *npf* samples, by using a window whose size was normally set to 20 to 40 ms. The shift between consecutive windows was typically set to 1/3-1/2 of the window size[31]. The number of the obtained frames and the q^{th} frame of *SSC′* were denoted by *Num_frame* and SCF_q, respectively.

Each obtained frame was multiplied by the Hamming window, using Equation (2), to avoid truncation of the continuity fames.

$$SCF'_q(n') = SCF_q(n') \times w(n') \quad (2)$$

where, n' is the index of each sample in the q^{th} frame, $0 \le n' \le npf-1$; and $w(n')$ is the Hamming window[32].

2.4.4 DWT

For SCF'_q, its approximation and detail coefficients at level t after the DWT operation are denoted by $a_q(t+1,k)$ and $d_q(t+1,k)$, respectively, where $k \in \{0, 1, \ldots, N_t - 1\}$. N_t is the number of approximation or detail coefficients at level t. Denoted $a_q(t, \cdot)$ as the set of all the approximation coefficients at level t, which is defined as the following:

$$a_q(t, \cdot) = \{a_q(t,0), a_q(t,1), \ldots, a_q(t, N_t-1)\} \quad (3)$$

where, $t \in \{1,2,\ldots,L\}$, and L is the number of the decomposition levels. When $t=0$, $a_q(0, \cdot)$ yields SCF'_q itself.

2.4.5 FFT and feature synthesis

The power spectral estimate for each of the detail coefficients sets of SCF'_q was obtained by using discrete FFT, as shown in the following equation:

$$FD_q(t,r) = \frac{1}{N_t}\left|\sum_{k=1}^{N_t} d_q(t,k) e^{-j2\pi rk/N_t}\right|^2 \quad (4)$$

where, t and k ranges over 1 to L and 1 to N_t, respectively; r is an integer value in the range of $[1,R]$, where R is the length of the discrete FFT. SCF'_q has only one approximation coefficient set, whose power spectral is calculated by Equation (5).

$$FA_q(r) = \frac{1}{N_L}\left|\sum_{k=1}^{N_L} a_q(L,k) e^{-j2\pi rk/N_L}\right|^2 \quad (5)$$

where, N_L is the number of the approximation coefficients at level L; All the $FD_q(t,r)$ for a given level t formed a set, which is denoted by $FD_q(t)$ and defined as:

$$FD_q(t) = \{FD_q(t,1), FD_q(t,2), \ldots, FD_q(t,(N_t/2)+1)\} \quad (6)$$

Similarly, all the $FA_q(r)$ formed a set FA_q, which is defined as:

$$FA_q = \{FA_q(1), FA_q(2), \ldots, FA_q((N_L/2)+1)\} \quad (7)$$

There are two steps in the feature synthesis module to assemble the $FD_q(t)$ and the FA_q of all the frames in a sub-sound clip.

Step 1: All the $FD_q(t)$ and the FA_q were concatenated using Equation (8) to form a power spectrum array, which is denoted by arr_coef_q.

$$arr\_coef_q = FA_q \| FD_q(L) \| FD_q(L-1) \| \ldots \| FD_q(1) \quad (8)$$

Step 2: All the arr_coef_q of a sub-sound clip were concatenated in row direction to form a matrix, which was denoted by *arr_coef*.

2.4.6 Mel filter bank

Mel and the linear frequency are related, namely, $\phi_f = 2595 * \log10(1 + l_f / 700)$, where ϕ_f is the Mel-frequency and l_f is the linear frequency. Each filter in the filter bank is a triangular having a response of 1 at the centre frequency and decreases linearly towards 0 till it reaches the centre frequencies of the two adjacent filters where the response is 0. Therefore, the Mel filter bank, denoted by $H_m(k)$, can be modelled by the following equation[33].

$$H_m(k) = \begin{cases} 0 & l_f(k) < l_{fc}(m-1) \\ \dfrac{l_f(k) - l_{fc}(m-1)}{l_{fc}(m) - l_{fc}(m-1)} & l_{fc}(m-1) \le l_f(k) < l_{fc}(m) \\ \dfrac{l_f(k) - l_{fc}(m+1)}{l_{fc}(m) - l_{fc}(m+1)} & l_{fc}(m) \le l_f(k) < l_{fc}(m+1) \\ 0 & l_f(k) \ge l_{fc}(m-1) \end{cases} \quad (9)$$

where, $l_{fc}(m)$ was the centre frequency of the m^{th} filter. Denoted the number of the filters in a Mel filter bank F, then $m \in \{1,F\}$. The filter bank came in the form of a matrix, which was denoted by H.

2.4.7 Mel-frequency Cepstrum

The logarithm of the filter bank outputs of a sub-sound clip, denoted by *MS*, was calculated by:

$$MS = \ln(H \cdot arr\_coef) \quad (10)$$

where, operator $\cdot$ represents the matrix multiplication. According to the original MFCC extraction method, the W_MFCC was obtained by applying DCT to its *MS*. W_MFCC considered only the static characteristics of a sound episode, but did not reflect its dynamic characteristics[34]. In order to introduce the dynamic characteristics, the first and second order difference of W_MFCC, denoted by W_MFCC′ and W_MFCC″ respectively, were calculated and appended to W_MFCC using Equation (11). The resultant coefficients matrix was denoted by W_MFCC‴.

$$\text{W\_MFCC}''' = \begin{bmatrix} \text{W\_MFCC} \\ \text{W\_MFCC}' \\ \text{W\_MFCC}'' \end{bmatrix} \quad (11)$$

As a result, the number of the feature coefficients of a given frame in a sub-sound clip was $3F$. These coefficients may be redundant or highly correlated, PCA was utilized to map a high dimensional W_MFCC‴ into a lower dimensional one. The final obtained feature matrix for a sub-sound clip was named by PW_MFCC. The q^{th} column vector in PW_MFCC was the dimensionality-reduced DWT based MFCC of the q^{th} fame, which was denoted by PW_MFCC_q.

2.5 Classifier training and testing

2.5.1 LSTM

RNN considers current and past input data simultaneously, which is a highly appropriate algorithm for modelling time series data[35]. The LSTM network is a more robust subclass of RNN that solves the RNN long term dependencies problem[36]. Sheep ingestion and rumination audio signals are typical sequence (time series) data. Therefore, LSTM was utilized in this study to classify sound episode produced by different short-term feeding behaviour. The structure of the LSTM model used in this study is shown in Figure 4.

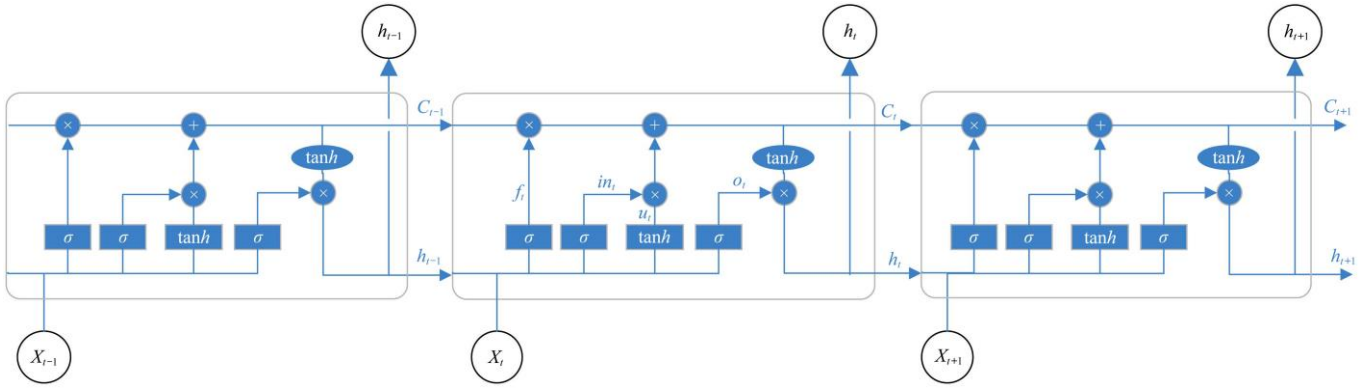


Figure 4 Structure of the LSTM used in this study.

The LSTM model depicted in Figure 4 can be formulated mathematically as follows:

$$f_t = \sigma(W_{hf} * h_{t-1} + W_{pf} * X_t + b_{fo}) \quad (12)$$

$$in_t = \sigma(W_{hi} * h_{t-1} + W_{pi} * X_t + b_{in}) \quad (13)$$

$$u_t = \tan h(W_{hu} * h_{t-1} + W_{pu} * X_t + b_u) \quad (14)$$

$$C_t = f_t \odot C_{t-1} + in_t \odot u_t \quad (15)$$

$$o_t = \sigma(W_{ho} * h_{t-1} + W_{po} * X_t + b_o) \quad (16)$$

$$h_t = o_t \odot \tan h(C_t) \quad (17)$$

where, in_t, f_t, o_t are input, forget, and output gate, respectively. C_{t-1} and C_t represent the previous and current cell state, respectively; W_{hf}, W_{pf}, W_{hi}, W_{pi}, W_{hu}, W_{pu}, W_{ho}, W_{po} are weights; and b_{fo}, b_{in}, b_u, b_o are biases to be computed during training; u_t is a vector of new candidate values that could be added to the state; h_{t-1} and h_t represent the previous and current hidden state; $\odot$ denotes pointwise multiplication; σ and tanh denote sigmoid and hyperbolic tangent functions, respectively; X_t is the input vector at time t.

2.5.2 Evaluation

To evaluate the classifier's performance, recall, specificity, precision, accuracy, and F1-score were used, which were defined as follows:

$$recall = \frac{TP}{TP + FN} \times 100\% \quad (18)$$

$$specificity = \frac{TN}{FP + TN} \times 100\% \quad (19)$$

$$precision = \frac{TP}{TP + FP} \times 100\% \quad (20)$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (21)$$

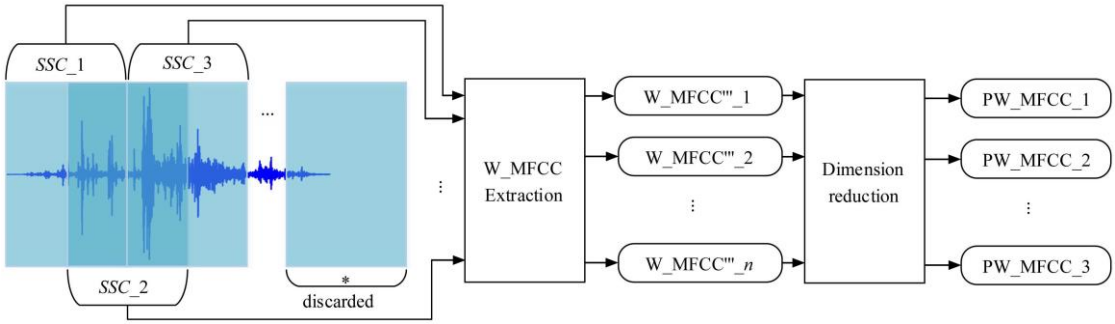
$$F1 - score = 2 \times = \frac{precision * recall}{precision + recall} \times 100\% \quad (22)$$

where, *TP* means a sample is actually positive and is predicted as a positive one; *FN* means a sample is actually positive but is predicted as a negative one; *FP* means a sample is actually negative but is predicted as a positive one; *TN* means a sample is actually negative and is predicted as a negative one.

3 Results

3.1 PW_MFCC extraction

The length of each sub-sound clip split from the sound episodes in SE_{tr} and SE_{te} was set to 4096 sample points in this study. The last sub-sound clip whose length was less than 4096 points was discarded. Then, PW_MFCC was extracted for each sub-sound clip using the methods described in section 2.4, as shown in Figure 5.



Note: The last sub-sound clip labelled by a star point was discarded, whose length was less than the specific value (4096 sample points in this study). The *SSC_i* ($1 \le i \le n$) was the i^{th} sub-sound clip of the example sound episode in this figure, which was split into *n* sub-sound clips.

Figure 5 Sound episode splitting for generating the input PW_MFCC of LSTM classifier

As shown in Figure 5, each two consecutive sub-sound clips were overlapped by 50%. Each sub-sound clip, denoted by *SSC_i* ($1 \le i \le n$), was sent to the W_MFCC extraction module one by one. Here, parameter α in filter (1) was set to 0.97. Frame length and shift in framing and windowing stage were respectively set to 1024 and 512 sample points. Therefore, each sub-sound clip had 7 frames. Daubechies 2 was chosen as the wavelet in the DWT described in section 2.4.4, where the transform level (L) was set to 2. Three set of DWT coefficients, namely $a_q(2, \cdot)$, $d_q(2, \cdot)$, and $d_q(1, \cdot)$ were obtained for the q^{th} ($1 \le q \le 7$) fame in a sub-sound clip. Due to each frame having 1024 samples, the number of the elements in $a_q(2, \cdot)$, $d_q(2, \cdot)$, and $d_q(1, \cdot)$ were 256, 256, and 512, respectively. Therefore, N_1 and N_2 in Equation (4) and (5) were 512 and 256, respectively. As a result, the size of the power spectrum matrix *arr_coef* was 515 by 7.

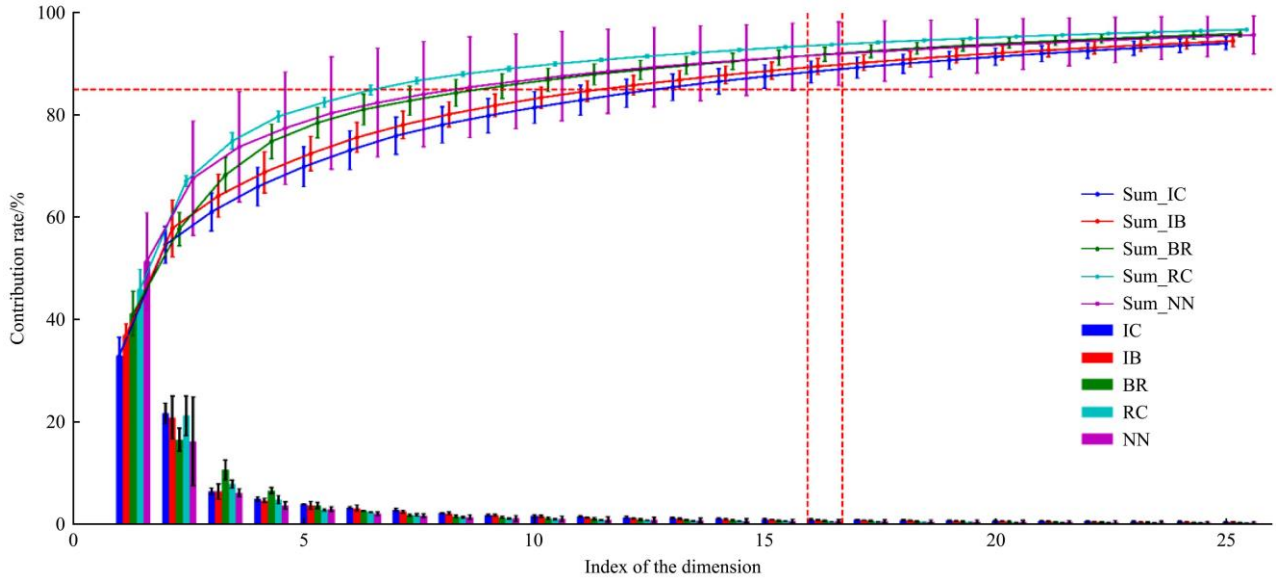
The typical number of the Mel filters described in section 2.4.6 was 26-40, and 35 was chosen in this study. As a result, the size of H obtained by Equation (9) was 35 by 515. After the logarithmic operation was carried out to the matrix product of H and *arr_coef*, as expressed in Equation (10), the obtained *MS* had

the size of 35 by 7, which was the same as the size of W_MFCC. By appending the first and second difference of W_MFCC, as shown in Equation (11), the size of the resultant W_MFCC‴ was 105 by 7.

To determine the dimensionality that needs to be retained in W_MFCC‴, 100 sub-sound clips obtained from the sound episodes respectively belonging to *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB*, 20 in each category, were randomly selected. Contribution rates of all the 105 dimensional feature coefficients were calculated, and the first 25 dimensional features with the highest average contribution rate are shown in Figure 6 in the form of bars. The standard deviation of each average contribution rate was depicted in the top of the corresponding bar. The average and standard deviation of the cumulative contribution rate of each dimensional feature were also depicted in Figure 6 in the form of line chart.

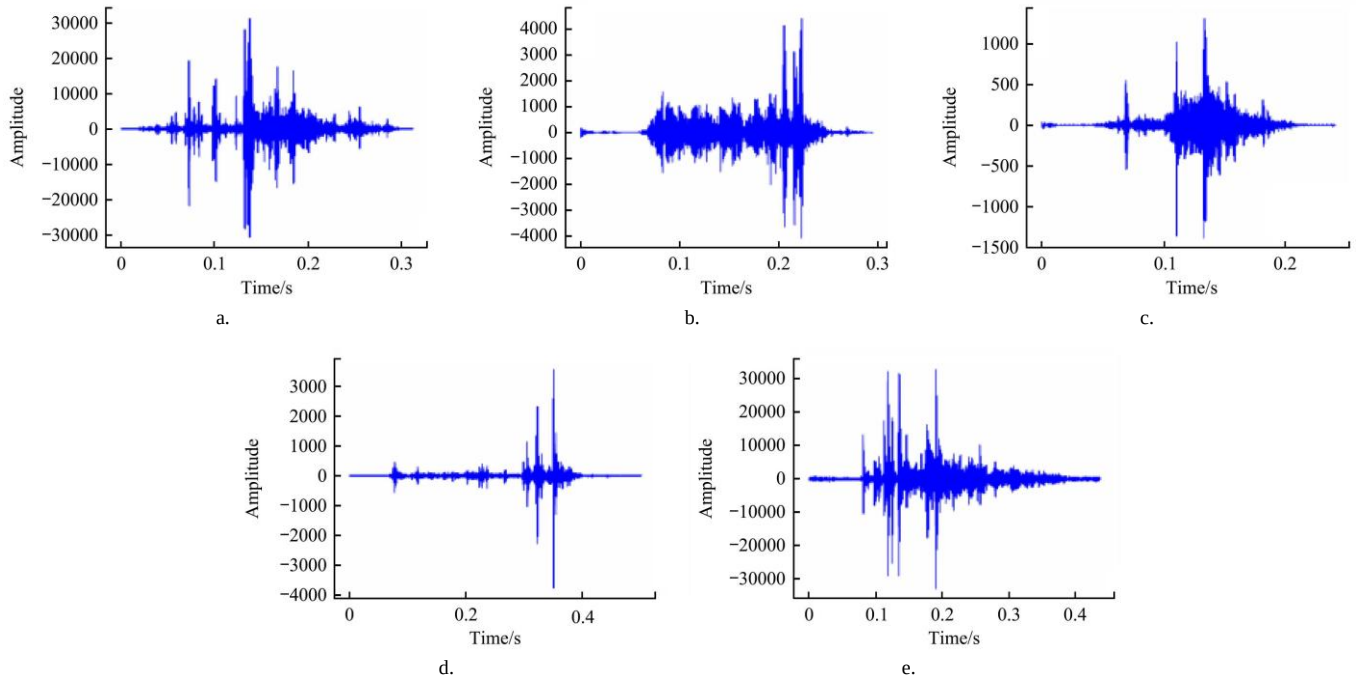
The cross region formed by the horizontal and vertical dashed red lines in Figure 6 indicated that after the feature coefficients reconstruction by PCA, the first 16-dimensional features accounted for more than 85% of the total variance for the original feature matrices. Therefore, the dimension of each W_MFCC‴ was reduced from 105×7 to 16×7.

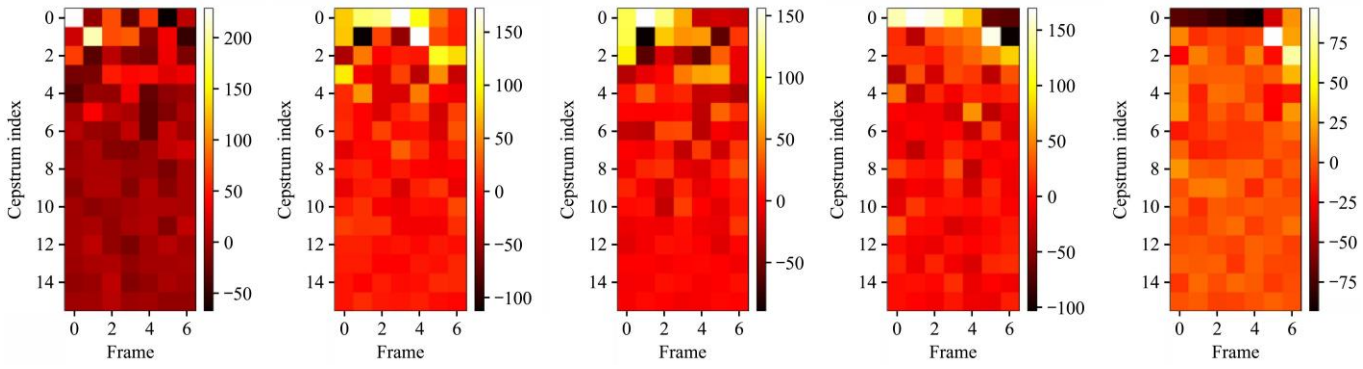
The resultant PW_MFCC after the dimensionality reduction was used as the input of the LSTM classifier to identify the category of each sub-sound clip. Respectively take a sound episode from *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB* as examples, the corresponding waveforms in time domain are shown in Figures 7a-7e. The dimensionality reduced PW_MFCCs (in form of 2D heat map) obtained from the first sub-sound clip of each example sound episode are shown in Figures 7f-7j.



Note: Average and the standard errors of the contribution rates in different dimension index are depicted in the forms of bars. Average and the standard errors of the cumulative contribution rates are depicted in the form of line charts. The horizontal dashed red line represented the contribution rate with the value of 85%.

Figure 6 First 25 dimensional feature coefficients with the highest average contribution rate after the feature coefficients reconstruction by using PCA





f. g. h. i. j.

Note: Five sound episodes were randomly selected from *SE_ IB*, *SE_ IC*, *SE_ BR*, *SE_ RC*, and *SE_NSFB*, respectively.

Figure 7 Original waveform in time domain (a-e) of the example sound episodes and the PW_MFCC (f-j) obtained from the first sub-sound clip of each sound episode

3.2 Implementation of LSTM

Denoted the sub-sound clips sets obtained from sound episodes in SE_{tr} and SE_{te} as SSC_{tr} and SSC_{te}, respectively. The number of the sub-sound clips in SSC_{tr} and SSC_{te} are shown in Table 1.

Each sub-sound clip was assigned a label in the form of one-hot code, which was generated from label of the sound episode where the sub-sound clip was derived. The mapping between the sound episode label and sub-sound clip one-hot code is depicted in Table 2.

Together with their respective one-hot code, the PW_MFCC matrices extracted from the sub-sound clips in SSC_{tr} were provided to the LSTM network in batches. The batch size, time steps, and input dimension were set to 50, 7, and 16, respectively, as shown in Figure 8.

Table 1 Number of the sub-sound clips in SSC_{tr} and SSC_{te}

Category	SSC_{tr}	SSC_{te}	Total
*SSC*s belonging to *SE_ IB*	1069	369	1438
*SSC*s belonging to *SE_ IC*	6900	3291	10191
*SSC*s belonging to *SE_ BR*	223	183	406
*SSC*s belonging to *SE_ RC*	13296	4985	18281
SSCs belonging to *SE_NSFB*	1935	795	2730
Total	23423	9623	33046

Table 2 Mapping between the sound episode label and sub-sound clip one-hot code

Sound episode category	*SE_ IB*(4)*	*SE_ IC*(3)*	*SE_ BR*(2)*	*SE_ RC*(1)*	*SE_ NSFB* (0)*
One-hot code of the sub-sound clips	[0,0,0,0,1]	[0,0,0,1,0]	[0,0,1,0,0]	[0,1,0,0,0]	[1,0,0,0,0]

Note: * The number in the parentheses was the label assigned to the corresponding sound episode category

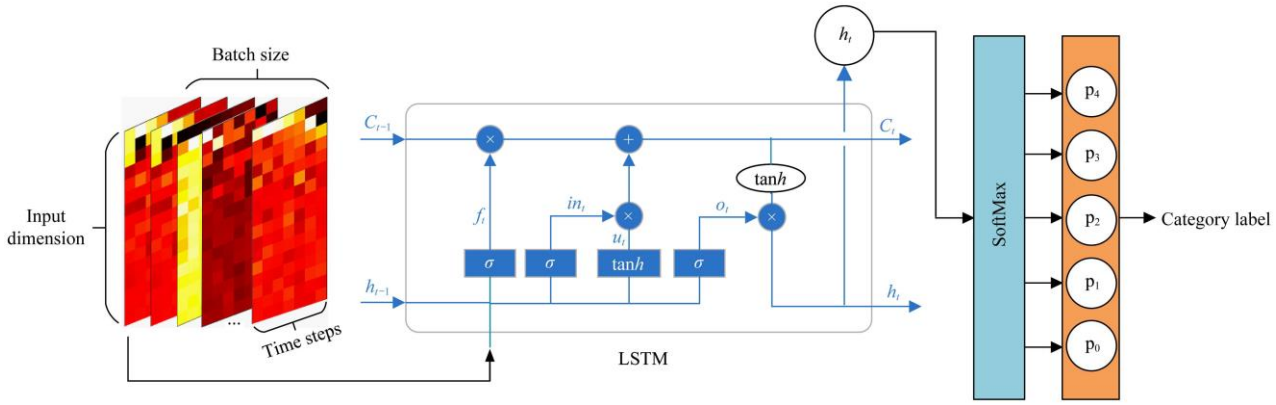


Figure 8 Block diagram of the implementation of the LSTM for sound episode classification

The input dimension and time steps were decided by the size of the PW_MFCC of each sub-sound clip. The parameters used to define the LSTM network were set as the following. The number of the hidden layers and the iterations for each input were respectively set to 500 and 1000. The learning rate was set to 0.0001. As shown in Figure 8, sigmoid (σ) and hyperbolic tangent (tanh) were respectively employed as the recurrent activation and activation functions. A SoftMax layer was used to convert the LSTM output (a vector with 5 elements) to a probabilities vector. The index of the maximum probability in the vector was the category label predicted by the LSTM. That was, 4 for *SE_IB*, 3 for *SE_IC*, 2 for *SE_BR*, 1 for *SE_RC*, and 0 for *SE_NSFB*.

The Adam optimizer[37] was chosen in this study to fine-tune the model parameters, such as the weights matrices and bias, by optimizing the loss function. The cross entropy was chosen as the loss function, as expressed in Equation (23), which can reduce the risk of vanishing gradient during the process of stochastic gradient descent.

$$J(\theta) = -\frac{1}{m}[\sum_{i-1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{i}) \log(1-h_\theta(x^{(i)}))] \quad (23)$$

where, m represents the number of categories, which is 5 in this study; $y^{(i)}$ and $\log h_\theta(x^{(i)})$ respectively represents the label value and its log probability. The training and validation losses obtained by the loss function in each iteration are shown in Figure 9.

It can be observed from Figure 9 that no gradient explosion appeared. The two loss curves began to entangle with each other

and both converged in the interval [0, 20] from around 500 iterations. Therefore, it was inferred that the classifier was not prone to be overfitting or under-fitting. This meant that after 500 iterations, the classifier was beginning to stabilize.

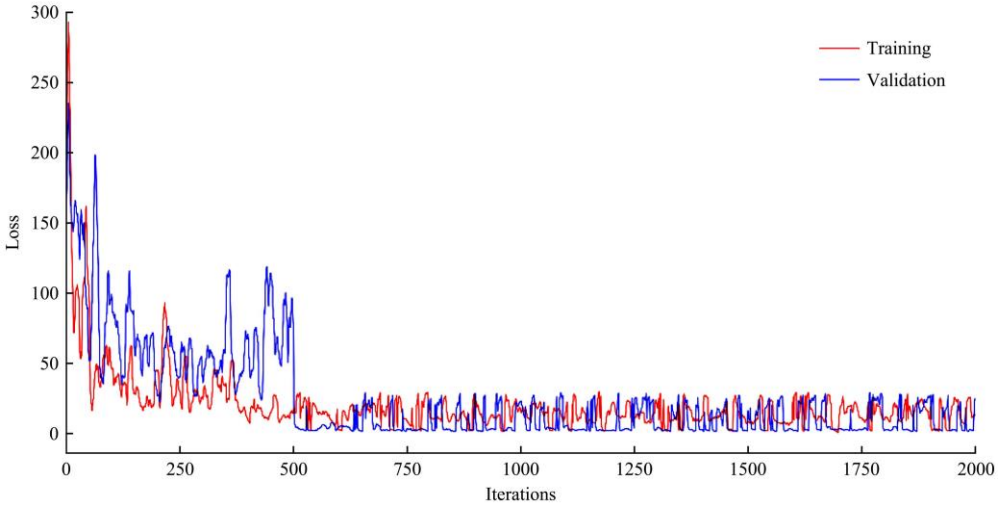
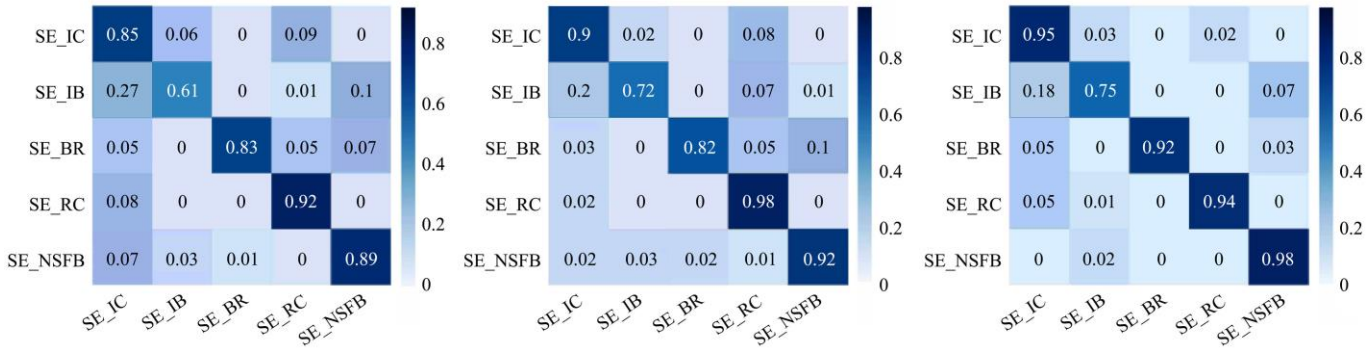


Figure 9 Convergence diagram of the training and validation loss

3.3 Classification performances

Evaluation criteria described in section 2.5.2 and confusion matrix were utilized to evaluate the performance of the classifier established in this study with the input of PW_MFCC, which was denoted by C_{PW_MFCC}. At the same time, comparisons between the performance of C_{PW_MFCC} and another two LSTM based classifiers with the input feature coefficients of MFCC and W_MFCC were also conducted. The latter two classifiers, respectively denoted by C_{MFCC} and C_{W_MFCC}, were trained by using the MFCC and W_MFCC extracted from the sub-sound clips in SSC_{tr}.

MFCC, W_MFCC, and PW_MFCC, extracted from the 9623 sub-sound clips which were obtained from the sound episodes in SE_{te}, were respectively provided to C_{MFCC}, C_{W_MFCC}, and C_{PW_MFCC}. The obtained confusion matrix of the three classifiers is given in Figure 10.



a. MFCC as input feature coefficients b. W_MFCC as input feature coefficients c. PW_MFCC as input feature coefficients

Note: Each row and column of the confusion matrices respectively represented the ground truth and the predicted sound episode category to whom a sub-sound clip was belonging.

Figure 10 Confusion matrix obtained for different classifier with the test data

Results in Figure 10 indicated that C_{PW_MFCC} outperformed C_{W_MFCC} and C_{MFCC} when it was applied to identify sub-sound clips in *SE_IC*, *SE_IB*, *SE_BR*, and *SE_NSFB*, where the proportions of the correct identification were 95%, 75%, 92%, and 98%, respectively. While C_{W_MFCC} performed best in classifying sub-sound clips in *SE_RC*. The lowest proportion of the correct identification obtained by C_{PW_MFCC} appeared when it was used to identify sub-sound clips in *SE_IB*, where nearly 20% of the sub-sound clips in *SE_IB* were misclassified as *SE_IC*. Possible reason for this may have something to do with sheep's feeding behaviour characteristics. One ingestion bite was followed by several ingestion chews. This caused that the total number of the sound episodes of ingestion bite in the dataset to be 1438, which was much less than those of *SE_IC* and *SE_RC*. Furthermore, sometimes, the gap between an ingestion bite and the following first ingestion chew was very small, and high similarity may exist between the signal characteristics of ingestion bite and chew. All of these may also be a reason that caused some *SE_IB* to be misclassified as *SE_IC*.

For a better understanding of the performance of different classifiers, Equations (18) to (22) were carried out for each sound episode category to calculate the recall, specificity, precision, accuracy, and F1-score of each classifier. The results are presented in Table 3.

Generally speaking, the results in Table 3 indicated that the performance of C_{W_MFCC} was close to C_{PW_MFCC} in terms of the five evaluation criteria. The overwhelming majority of the performance values of both classifiers were better than C_{MFCC}. As shown in Equation (22), the F1-score was interpreted as a weighted average of the precision and recall, where the relative contribution of precision and recall to the F1-score were equal. It was represented in Table 3 that F1-score of C_{PW_MFCC} was better than C_{W_MFCC} and C_{MFCC} in classifying sub-sound clips in *SE_IB*, *SE_BR*, and *SE_NSFB*. However, C_{W_MFCC} outperformed C_{PW_MFCC} and C_{MFCC} in F1-score when it was used to identify sub-sound clips in *SE_IC* and *SE_RC*.

Table 3 Evaluation criteria of the trained LSTM classifier with different feature coefficients for sound episodes in different categories

Evaluation criteria	Feature coefficients	*SE_IC*	*SE_IB*	*SE_BR*	*SE_RC*	*SE_NSFB*
Recall	MFCC	85.00%	41.29%	83.33%	91.99%	89.00%
	W_MFCC	90.02%	71.85%	83.33%	98.00%	92.00%
	PW_MFCC	92.01%	71.11%	93.33%	95.02%	99.00%
Specificity	MFCC	91.29%	98.90%	99.94%	92.71%	99.65%
	W_MFCC	97.21%	99.06%	99.88%	93.06%	99.89%
	PW_MFCC	94.82%	99.56%	100.00%	93.94%	99.98%
Precision	MFCC	85.04%	61.48%	89.29%	93.62%	94.35%
	W_MFCC	97.21%	68.31%	80.65%	94.26%	98.22%
	PW_MFCC	91.19%	82.05%	100.00%	94.80%	99.66%
Accuracy	MFCC	88.98%	96.55%	99.84%	92.33%	99.01%
	W_MFCC	94.56%	98.32%	99.78%	95.72%	99.41%
	PW_MFCC	93.78%	98.78%	99.59%	94.52%	99.91%
F1-score	MFCC	85.02%	49.40%	86.21%	92.80%	91.60%
	W_MFCC	92.41%	70.04%	81.97%	96.09%	95.01%
	PW_MFCC	91.60%	76.19%	96.55%	94.91%	99.33%

In order to make a more intuitive presentation of the performance comparison, the overall accuracy and specificity (respectively denoted by *overall_accu* and *overall_spec*) and comprehensive F1-score (denoted by *comp_F*1) of each classifier for all the sub-sound clips in SE_{te} were calculated and presented in Table 4.

Table 4 Overall accuracy, specificity, and comprehensive F1-score of the trained LSTM classifier with different feature coefficients extracted from the test data

Feature coefficients	*overall_accu*	*overall_spec*	*comp_F1*
MFCC	92.04%	97.41%	88.35%
W_MFCC	95.81%	98.47%	93.89%
PW_MFCC	94.97%	98.38%	97.41%

The *overall_spec* and *overall_accu* were also respectively calculated by using Equations (19) and (21), where *TP*, *TN*, *FP*, and *FN* were the statistics gathered from the entire data in SE_{te}. The *comp_F*1 was calculated by the following equation:

$$comp\_F1 = 2 \times \frac{overall\_prec * overall\_recall}{overall\_prec + overall\_recall} \quad (24)$$

where, *overall_recall* and *overall_prec* were respectively calculated by using Equations (18) and (20). Same as the calculation of *overall_spec* and over*all_accu*, *TP*, *TN*, *FP*, and *FN* used to calculate *overall_prec* and *overall_recall* were also the statistics gathered from the entire data in SE_{te}.

Data in Table 4 indicated that the best overall accuracy and specificity were obtained by C_{W_MFCC}. However, the gaps between C_{W_MFCC} and C_{PW_MFCC} in these two criteria were very small, both of which were less than 1%. On the other hand, C_{PW_MFCC} outperformed both C_{W_MFCC} and C_{MFCC} in terms of comprehensive F1-score. Therefore, it can be concluded that the introducing of wavelet transform improved the classifier performance effectively. At the same time, PCA reduced the computational overhead without degrading classifier performance.

4 Discussion

4.1 Performance comparison with previous methods

Some classifiers on feeding behaviour classification have been developed for ruminant animals in the past decade. Existing strategies could be divided into two categories: binary classifier and multi-class classifier. The former focuses on distinguishing two kinds of feeding behaviour, such as chewing and non-chewing[25], grazing and non-grazing[38], grazing and ruminating[39], and so on. The latter tries to distinguish multiple behaviours, such as biting, chewing, chew-bite, etc. The classifier established in this study belongs to the latter. Therefore, performance comparison was carried out between the developed classifier and the previous multi-class feeding behaviour classifiers, as shown in Table 5.

Table 5 Comparison between the developed PW_MFCC based classifier and previous classifiers

Source	Recorder	Behaviour	*overall_accu*	*comp_F*1
Milone et al.[13]	MP	IC, CB, IB	93.33% (cattle)	Not provided
Chelotti et al.[1]	MP	IC, CB, IB	84.0% (cattle)	Not provided
Deniz et al.[15]	MP	IC, CB, IB	76.4% (cattle)	Not provided
Giovanetti et al.[40]	Acc	GR, RE, RU	93% (sheep)	Not provided
Zehner et al.[42]	NP, Acc	EA, RU, DR, OB	94.5%# (cow)	Not provided
Chelotti et al.[17]	AR	IC, CB, IB	90.74% (cattle)	92.39%#
Decandia et al.[41]	Acc	GR, RU, OB	89.7% (sheep)	Not provided
Galli et al.[18]	MP	IC, CB, IB	85% (cow) 66% (sheep)	Not provided
Developed classifier	AR	IB, IC, BR, RC, NSFB	94.97%	97.41%

Note: Recorder: MP—Microphone; Acc—Accelerometer; NP—Noseband pressure sensor; AR—Audio recorder. Behaviour: IB—Ingestion bite; IC—Ingestion chew; CB—Chew-bite; GR—Grazing; RE—Resting; RU—Ruminating; EA—Eating; DR—Drinking; OB—Other behaviour; BR—Bolus regurgitation; RC—Rumination chew; NSFB—Behaviour not relating to feeding.

These values were calculated based on the data provided in the literatures.

Generally speaking, the previous multi-class classifiers were established based on the sensor data acquired by electronic sensors such as pressure sensors, accelerometers, and so on, or acoustic signal obtained by microphones or audio recorders. Most of the sensor data-based classifier focused on recognizing long-term activities (rumination and grazing) rather than individual jaw movements[17]. As shown in Table 5, in addition to the grazing and ruminating, resting and other behaviour were respectively taken into considered by the classifier established by Giovanetti et al.[40] and Decandia et al.[41], where multivariate statistical

techniques were adopted as the foundation method. The overall classification accuracy obtained by Giovanetti et al.[40] was better than that achieved by Decandia et al.[41], both of which were lower than that of C_{PW_MFCC} established in this study.

The commercial RumiWatch noseband sensor (Itin + Hoch GmbH, Liestal, Switzerland), together with two supporting software with different versions, were tested to distinguish eating, rumination, drinking, and other activities (e.g., idling) for stable-fed dairy cows[42]. Characteristic peak rates and peak intervals of the pressure sensor signal were employed by the supporting software to identify the mentioned long-term activities. Percentage of behaviour time and quantification of jaw movements and boluses within a 1-h interval obtained by the software were compared with the results obtained by direct observation. Comparison results indicated that the identification accuracy of one of the two versions software with better performance varying from 92% to 98% for different activities. The overall accuracy was not provided by Zehner et al.[42]. Assuming that the sample size of each activity category is similar, the overall accuracy obtained by Zehner et al.[42] was presumed to be 94.5%, which was calculated by averaging the classification accuracy of each activity. There was little difference in the overall accuracy obtained by the software and C_{PW_MFCC} established in this study. However, parameters in the supporting software that comes with RumiWatch noseband sensor has been optimized specifically for cows. Additional experiments are required to optimize the parameters of the software for sheep activity classification. In addition, more tests are needed to evaluate the classification performance of the RumiWatch noseband sensor system to distinguish short-term feeding behaviour for other ruminant animals.

Previous acoustic signal based jaw movement events classifiers mostly focused on recognizing bite, chew-bite, and chew. Experimental acoustic data in these studies was mostly obtained in a simulated grazing scenario, where ruminant animals could bite new pasture when the pasture mass obtained by the previous bite was still in its mouth. Therefore, chew-bite behaviour happened frequently. However, sheep in this study were fed by an experimenter, who provided new oat grass to the sheep when the pasture mass of the previous bite was swallowed. So, no sound segment corresponding to chew-bite was obtained. As a result, chew-bite was not taken into consider in the classifiers in this study.

Classifiers established by Milone et al.[13] and Galli et al.[18] were both aimed for identifying short-term feeding behaviour sound for sheep, where hidden Markov models and linear discriminant analysis were respectively used as the foundation methods. The overall accuracy achieved by Milone et al.[13] was higher than that of the classifier constructed by Galli et al.[18], both of which were lower than the overall accuracy achieved by C_{PW_MFCC} in this study.

Short-term feeding behaviour sound classifiers[1,17,18] were established for cattle. For computational overhead reduction, only acoustic features in time domain were utilized to classify sound signal produced by jaw movement events (including bite, chew, and chew-bite)[1]. The overall accuracy of the classification rules was 84%, which was much lower than that achieved in this study. However, the main goal of the study done by Chelotti et al.[1] was to develop a method which can achieve a balance between the classification performance and computational overhead. The rules developed in the research[1] were employed by Deniz et al.[15] to realize an embedded system for real time jaw movement events classification for cattle with an overall accuracy of 76.4%. The previous classifier based on acoustic signal with the best performance was established by Chelotti et al.[17], who combined de-trending technique (empirical mode decomposition) and support vector machine to classify jaw movements for cattle. The overall accuracy and comprehensive F1-score achieved by Chelotti et al.[17] were respectively 90.74% and 92.39%, which were a little lower than those achieved by the C_{PW_MFCC} in this study.

4.2 Potential usage of the developed classifier

Features in audio signal produced by grazing or ingestion chew have been already utilized by many researchers to estimate forage or dry matter intake for sheep and cattle[5,12,25]. It was found that the times of ingestion bite can also contribute to intake estimation for sheep[2]. However, to our best knowledge, existing studies have not yet solved the problem of distinguishing ingestion chew from rumination chew based on audio signals, which should be a premise of audio signal based intake estimation. The classifier established in this study can be used to separate rumination chew sound episodes from ingestion chew signal segments automatically. Based on this, a better accuracy could be achieved when an existing intake estimation model is used to estimate forage or dry matter intake for ruminant animals.

In addition, identification of the sound segments produced by bolus regurgitation and rumination chew can be used to keep statistics of daily rumination times and duration. It has been proved that decreasing of rumination times was an indicator of stress[43], anxiety[44], and disease[45]. Conversely, an increase in daily rumination duration is associated with more salivation and can improve rumen health[46]. Therefore, the real-time monitoring of daily rumination behaviour characteristics (times and duration), which can be obtained by the developed classifier in this study, can be used to identify sheep with health disorder in the future.

4.3 Future work for classifier optimization

As shown in Figure 10, 25% of the sub-sound clips of ingestion bite in testing dataset were misclassified into other categories. Like many other deep learning-based classifiers, enough data for training is essential for the improvement of the classifier performance. Therefore, more ingestion and rumination sound should be collected in the future and more sound episodes of ingestion bite and bolus regurgitation should be expanded in the training dataset. An approximate uniform quantity of the sound episodes in different categories could bring a better short-term feeding behaviour sound classifier.

In addition to the expansion of the dataset, the following measurements can be tried in the future to improve the classifier further.

1) Pressure sensors are suggested to be employed to monitor the pressure variation pattern of the jaws during the ingestion and ruminate behaviour. Comprehensive use of jaw movement pattern and double thresholds endpoints detection could contribute to improve the accuracy of sound sections identification and segmentation.

2) Besides the audio recorder used for feeding sound collection, an additional microphone can be introduced to collect environmental noise signal. High quality sound data for classifier establishment could be obtained by subtracting environmental noise spectral from the original feeding behaviour sound signal acquired by the audio recorder. Quality improvement of the feeding behaviour sound signal could in turn bring a better classification accuracy.

3) The length of each sub-sound clip was set to 4096 sample

points in this study, which was fit for the sound signal produced by the jaw movement or bolus regurgitation when oat grass was provided. Different kinds of forage or feed could bring different sound characters when sheep was ingesting or ruminating. Therefore, further efforts should be carried out to optimize the length of sub-sound clip for different forage or feed.

5 Conclusions

In this study, acoustic features were utilized to develop a short-term feeding behaviour sound classifier for sheep using LSTM network. A major contribution of this study was confirmation that deep learning methods, which have been widely used for natural language processing, could also be feasible for animal sound processing.

Three classifiers, respectively with MFCC, W_MFCC, and PW_MFCC as the input features, were established using LSTM networks. Performances of the established classifiers were compared and the results demonstrated that introducing of DWT and PCA improved the classifier performance effectively. This meant that W_MFCC brought a better characterization for non-stationary sound signal of ruminants, and PCA reduced not only the dimension, but also the redundant coefficients in W_MFCC. As a result, PW_MFCC based classifier was recommended, which was capable to classify different sound episodes produced by ingestion bite, ingestion chew, bolus regurgitation, and rumination chew with an overall accuracy and F1-score of 94.97% and 97.41%, respectively. These evaluation criteria were better than those achieved by the previous multi-class short-term feeding behaviour classifiers.

The sound episodes of different short-term feeding behaviour identified by the PW_MFCC based classifier can be used to estimate daily forage intake, rumination times and duration for sheep. Classifier established in this study is essential to sheep forage intake estimation and rumination behaviour analysis, which provides a foundation for an automatic identification system for sick sheep with abnormal feeding and rumination behaviour pattern.

Acknowledgements

This work was supported by the Basic Research Project of the Science and Technology Department of Qinghai province, China (Grant No. 2020-ZJ-716), the Key Research and Development Project of the Science and Technology Department of Jiangsu province, China (Grant No. BE2018433), and the Key Research and Development Project of the Science and Technology Department of Qinghai Province, China (Grant No. 2017-HZ-813).

[References]

[1] Chelotti J O, Vanrell S R, Milone D H, Utsumi S A, Galli J R, Rufiner H L, et al. A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. Computers and Electronics in Agriculture, 2016; 127: 64–75.

[2] Galli J R, Cangiano C A, Milone D H, Laca E A. Acoustic monitoring of short-term ingestive behavior and intake in grazing sheep. Livest. Sci., 2011; 140(1-3): 32–41.

[3] Oudshoorn F W, Cornou C, Hellwing A L F, Hansen H H, Munksgaard L, Lund P, et al. Estimation of grass intake on pasture for dairy cows using tightly and loosely mounted di- and tri-axial accelerometers combined with bite count. Computers and Electronics in Agriculture, 2013; 99: 227–235.

[4] Leiber F, Holinger M, Zehner N, Dorn K, Probst J K, Neff A S. Intake estimation in dairy cows fed roughage-based diets: An approach based on chewing behaviour measurements. Applied Animal Behaviour Science, 2016; 185: 9–14.

[5] Galli J R, Cangiano C A, Pece M A, Larripa M J, Milone D H, Utsumi S A, et al. Monitoring and assessment of ingestive chewing sounds for prediction of herbage intake rate in grazing cattle. Animal, 2018; 12(5): 973–982.

[6] Campos D P, Abatti P J, Hill A G, Paula A D. Short-term fibre intake estimation in goats using surface electromyography of the masseter muscle. Biosystems Engineering, 2019; 183: 209–220.

[7] Rombach M, Sudekum K H, Munger A, Schori F. Herbage dry matter intake estimation of grazing dairy cows based on animal, behavioral, environmental, and feed variables. Journal of Dairy Science, 2019; 102(4): 2985–2999.

[8] Milone D H, Rufiner H L, Galli J R, Laca E A, Cangiano C A. Computational method for segmentation and classification of ingestive sounds in sheep. Computer and Electronics in Agriculture, 2009; 65(2): 228–237.

[9] Watanabe N, Sakanoue S, Kawamura K, Kozakai T. Development of an automatic classification system for eating, ruminating and resting behavior of cattle using an accelerometer. Grassl. Sci., 2008; 54(4): 231–237.

[10] Buchel S, Sundrum A. Technical note: Evaluation of a new system for measuring feeding behavior of dairy cows. Computers and Electronics in Agriculture, 2014; 108: 12–16.

[11] Sneddon J, Mason A. Automated monitoring of foraging behaviour in free ranging sheep grazing a bio-diverse pasture using audio and video information. In Proceedings of International Conference on Sensing Technology (ICST), Liverpool-UK, 2014; pp.170–173.

[12] Laca E A, Wallisdevries M. Acoustic measurement of intake and grazing behaviour of cattle. Grass Forage Sci., 2000; 55(2): 97–104.

[13] Milone D H, Galli J R, Cangiano C A, Rufiner H L, Laca E A. Automatic recognition of ingestive sounds of cattle based on hidden Markov models. Computers and Electronics in Agriculture, 2012; 87: 51–55.

[14] Navon S, Mizrach A, Hetzroni A, Ungar E D. Automatic recognition of jaw movements in free-ranging cattle, goats and sheep, using acoustic monitoring. Biosystems Engineering, 2013; 114: 474–483.

[15] Deniz N N, Chelotti J O, Galli J R, Planisich A M, Larripa M J, Rufiner H L, et al. Embedded system for real-time monitoring of foraging behavior of grazing cattle using acoustic signals. Computers and Electronics in Agriculture, 2017; 138: 167–174.

[16] Bishop J, Falzon G, Trotter M, Kwan P, Meek P. Sound analysis and detection, and the potential for precision livestock farming - a sheep vocalization case study. In Proceedings of the 1st Asian-Australasian Conference on Precision Pastures and Livestock Farming, Hamilton-New Zealand, 2017; October 1–7.

[17] Chelotti J O, Vanrell S R, Galli J R, Giovanini L L, Rufiner H L. A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. Computers and Electronics in Agriculture, 2018; 145: 83–91.

[18] Galli J R, Milone D H, Cangiano C A, Mart ńez C E, Laca E A, Chelotti J O, et al. Discriminative power of acoustic features for jaw movement classification in cattle and sheep. Bioacoustics, 2019; 29(5): 1–15.

[19] Hsu W N, Zhang Y, Glass J. A prioritized grid long short-term memory RNN for speech recognition. Proceedings of IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 2016; pp.467–473.

[20] Qu Z, Haghani P, Weinstein E, Moreno P. Syllable-based acoustic modeling with CTC-SMBR-LSTM. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Piscataway-NJ, 2017; pp.173–177.

[21] Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, et al. Deep sentence embedding using long short-term memory networks: Analysis and application to information retrieval. IEEE/ACM Trans. Audio Speech Lang. Process, 2016; 24: 694–707.

[22] Mallinar N, Rosset C. Deep canonically correlated LSTMs. arXiv:1801.05407, 2018.

[23] Audacity 2.1.2, 2016. Available: https://www.audacityteam.org/.

[24] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics Speech & Signal Processing, 1985; 33(2): 443–445.

[25] Sheng H, Zhang S F, Zuo L S, Duan G H, Zhang H L, Okinda C, et al. Construction of sheep forage intake estimation models based on sound analysis. Biosystems Engineering, 2020; 192: 144–158.

[26] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing, 1980; 28: 357–366.

[27] Ganchev T, Fakotakis N, Kokkinakis G. Comparative evaluation of various MFCC implementations on the speaker verification task. In Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005), Patras-Greece, 2005; 1: 191–194.

[28] Tang C P, Chui K L, Yu Y K, Zeng Z L, Wong K H. Music genre classification using a hierarchical long short term memory (LSTM) model. International Workshop on Pattern Recognition IWPR 2018, Jinan-China, 2018; pp.26–28.

[29] Li M, Chen N. A robust cover song identification system with two-level similarity fusion and post-processing. Applied Sciences, 2018; 8(8): 3383.

[30] Srivastava S, Bhardwaj S, Bhandari A, Gupta K, Bahl H, Gupta J R P. Wavelet packet based mel frequency cepstral features for text independent speaker identification. In: Abraham A, Thampi S (Ed.). Intelligent Informatics. Berlin: Springer, 2013; 182: 237–247.

[31] Patil J M, Desai P K. Word-based LID using HMM and bi-gram modeling. In: Chakravarthi V, Shirur Y, Prasad R (Ed.). Proceedings of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking (VCASAN-2013). India: Springer, 2013; 258: 373–383.

[32] Jain S, Gupta D. Feature extraction techniques based on human auditory system. In: Bhalla S, Bhateja V, Chandavale A, Hiwale A, Satapathy S (Ed.). Intelligent Computing and Information and Communication. Singapore: Springer, 2018; 673: 667–676.

[33] Sigurdsson S, Petersen K B, Schiler T L. Mel frequency cepstral coefficients: An evaluation of robustness of mp3 encoded music. Conference Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR), Vicoria-Canada, 2006; pp.286–289.

[34] Chelali F Z, Djeradi A. Text dependant speaker recognition using MFCC, LPC and DWT. International Journal of Speech Technology, 2017; 20(3): 725–740.

[35] Panakkat A, Adeli H. Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators. Comput. Civ. Infrastruct. Eng., 2009; 24(4): 280–292.

[36] Hochreiter S. Long short-term memory. Neural Computation, 1997; 9(8): 1735–1780.

[37] Kingma D P, Adam J B. A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations. San Diego: Springer Verlag, 2015; pp.1–15.

[38] Alvarenga F A P, Borges I, Palkovic L, Rodina J, Oddy V H, Dobos R C. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. Appl. Anim. Behav. Sci., 2016; 181: 91–99.

[39] Vanrell S R, Chelotti J O, Galli J R, Utsumi S A, Giovanini L L, Rufiner H L, et al. A regularity-based algorithm for identifying grazing and rumination bouts from acoustic signals in grazing cattle. Computers and Electronics in Agriculture, 2018; 151: 392–402.

[40] Giovanetti V, Decandia M, Molle G, Acciara M, Mameli M, Cabiddu A, et al. Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer. Livestock Science, 2017; 196: 42–48.

[41] Decandia M, Giovanetti V, Molle G, Acciara M, Mameli M, Cabiddu A, et al. The effect of different time epoch settings on the classification of sheep behaviour using tri-axial accelerometry. Computers and Electronics in Agriculture, 2018; 154: 112–119.

[42] Zehner N, Umstätter C, Niederhauser J J, Schick M. System specification and validation of a noseband pressure sensor for measurement of ruminating and eating behavior in stable-fed cows. Computers and Electronics in Agriculture, 2017; 136: 31–41.

[43] Herskin M S, Munksgaard L, Ladewig J. Effects of acute stressors on nociception, adrenocortical responses and behavior of dairy cows. Physiol. Behav., 2004; 83(3): 411–420.

[44] Bristow D J, Holmes D S. Cortisol levels and anxietyrelated behaviors in cattle. Physiology & Behavior, 2007; 90: 626–628.

[45] Welch J G. Rumination, particle size and passage from the rumen. Journal of animal science, 1982; 54(4): 885–894.

[46] Beauchemin K A, Farr B I, Rode L M. Enhancement of the effective fiber content of barley-based concentrates fed to dairy cows. Journal of Dairy Science, 1991; 74(9): 3128–3139.