A tristimulus-formant model for automatic recognition of call types of laying hens

Xiaodong Du [a,b,c], Guanghui Teng [a,*], Chaoyuan Wang [a], Lenn Carpentier [b], Tomas Norton [b,*]

[a] College of Water Resources & Civil Engineering, China Agricultural University, Key Laboratory of Agricultural Engineering in Structure and Environment, 100083 Beijing, China
[b] Division Animal and Human Health Engineering, Department of Biosystems, KU Leuven, Kasteelpark Arenberg 30, 3001 Heverlee, Belgium
[c] Shandong New Hope Liu he Co., Ltd., 266102 Qingdao, China

ARTICLE INFO

Keywords:
Animal vocalisation
Sound recognition
Chicken
MFCC
Tristimulus-formant

ABSTRACT

An essential objective of Precision Livestock Farming (PLF) is to use sensors that monitor bio-responses that contain important information on the health, well-being and productivity of farmed animals. In the literature, vocalisations of animals have been shown to contain information that can enable farmers to improve their animal husbandry practices. In this study, we focus on the vocalisation bio-responses of birds and specifically develop a sound recognition technique for continuous and automatic assessment of laying hen vocalisations. This study introduces a novel feature called the "tristimulus-formant" for the recognition of call types of laying hens (i.e., vocalisation types). Tristimulus is considered to be a timbre that is equivalent to the colour attributes of vision. Tristimulus measures the mixture of harmonics in a given sound, which grouped into 3 sections according to the relative weights of the harmonics in the signal. Experiments were designed in which calls from 11 Hy-Line brown hens were recorded in a cage-free setting (4303 vocalisations were labelled from 168 h of sound recordings). Then, sound processing techniques were used to extract the features of each call type and to classify the vocalisations using the LabVIEW® software. For feature extraction, we focused on extracting the Mel frequency cepstral coefficients (MFCCs) and tristimulus-formant (TF) features. Then, two different classifiers, the back-propagation neural network (BPNN) and Gaussian mixture model (GMM), were applied to recognise different call types. Finally, comparative trials were designed to test the different recognition models. The results show that the MFCCs-12+BPNN model (12 variables) had the highest average accuracy of 94.9 ± 1.6% but had the highest model training time (3201 ± 119 ms). At the same time, the MFCCs-3+TF+BPNN model had fewer feature dimensionalities (6 variables) and required less training time (2633 ± 54 ms) than the MFCCs-12+BPNN model and could classify well without compromising accuracy (91.4 ± 1.4%). Additionally, the BPNN classifier was better than the GMM classifier in recognising laying hens' calls. The novel model can classify chicken sounds effectively at a low computational cost, giving it considerable potential for large data analysis and online monitoring systems.

1. Introduction

In the past, decisions on the management of farm animals have traditionally been based on the observation, judgment, and experience of farmers. However, the consolidation of livestock production throughout the world has made it increasingly difficult for farmers to monitor and manage their animals at the level of detail that was once possible. Currently, it has become possible for cameras, microphones, and sensors to take the place of farmers' eyes and ears to monitor animal houses effectively (Vandermeulen et al., 2013; Kashiha et al., 2013). Moreover, such technology can provide further benefits by monitoring animals continuously for 24 h per day and 365 days per year to provide more complete information on livestock (Guarino et al., 2017).

Automatic and continuous sound analyses can provide more detailed information about the state of farm animals. Sound analysis has been used to estimate the thermal comfort of chicks in different thermal environments (de Moura et al., 2008; Du et al., 2020). Sound analysis has also been used to monitor drinking behaviour by analysing pecking sounds (Pluk et al., 2010; Kashiha et al., 2013). Similarly, sound analysis has been used to predict broiler feed intake by acquiring pecking sounds

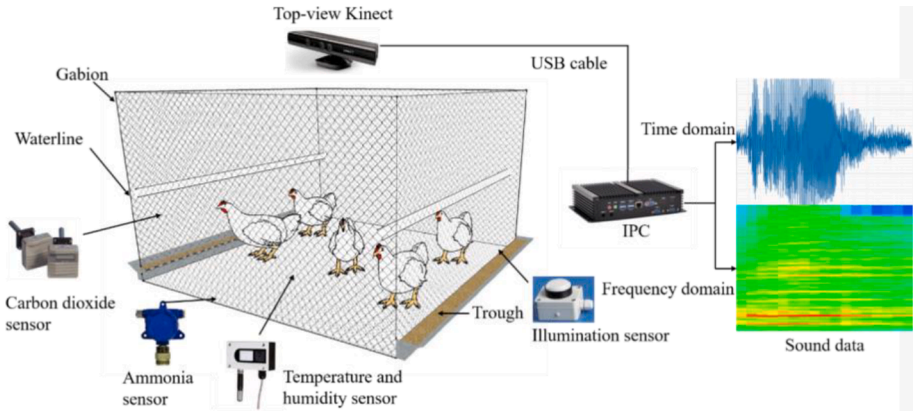

Fig. 1. Schematic of the experimental platform.

during feeding (Aydin et al., 2014; Aydin and Berckmans, 2016). Moreover, sound analysis has recently formed the basis of a pig cough monitoring system based on the fact that coughing can serve as a biomarker for respiratory disease and aerial pollution in livestock houses (Van Hirtum and Berckmans, 2004; Exadaktylos et al., 2008; Berckmans et al., 2015). Extensions of sound analysis include sound source localisation analysis, which has also been applied to detect the source of pig respiratory diseases and chicken nocturnal vocalisations (Silva et al., 2008; Du et al., 2018).

Animal vocalisations are a fundamental component of animal behaviour, and they can be used to provide information on animal health and welfare (Manteuffel et al., 2004). Capturing animal sounds precisely and quickly on the farm has the capacity to help farmers improve their husbandry practices. In animal vocalisation studies, information about a specific vocalisation is often extracted manually from its spectrogram, while the choice of parameters is often driven by the intuition of the researcher. This manual extraction makes the process unsuitable for online and real-time large data analysis (Mielke and Zuberbühler, 2013). Moreover, given the high level of mechanisation and the large population of animals in typical livestock buildings, a large quantity of different sounds can be heard throughout the day and night. The goal of designing a suitable classifier is a particular challenge in the case of poultry buildings (containing approximately 50 K broilers or 80 K laying hens) because it is difficult to realise accurate sound detection algorithms that can be implemented in such buildings (Cao et al., 2014). Therefore, to date, it has been almost impossible to realise accurate algorithms for monitoring laying hen sounds and vocalisations. To improve the current state-of-the-art in this field, two key improvements are required for accurate animal vocalisation detection, namely, feature extraction and classification.

For feature extraction, the source-filter theory of vocal production is a robust framework for studying animal vocal communication (Taylor and Reby, 2010). According to this theory, calls are produced through the syrinx, which is regarded as a 'source' and located at the base of the trachea. Syringeal constriction functionally overlaps the role of the larynx in mammalian phonation, and the trachea acts as a 'filter' to remove certain frequencies or leave others unchanged (Fletcher, 1988; Beckers et al., 2003; Taylor and Reby, 2010; Favaro et al., 2015). Source-related vocal features (the fundamental frequency, F_0), which are related to the vibrating mass in the syrinx, are stable (laying hens, F_0: 400–2500 Hz) (Cao et al., 2014), while filter-related features (formants), which are related to the supra-syringeal vocal tract, are dependent on different vocalisations. Specifically for the latter, the first three formants of each vocalisation and the tristimulus values of these formants contain the most energy and variance (Yeon et al., 2006; Favaro et al., 2017). These tristimulus values were first introduced as a timbre equivalent to the colour attributes in vision analysis, and they represent three different types of energy ratio, which allow a fine description of the first harmonic of the spectrum (Pollard and Jansson, 1982). Various studies have also shown that the widely used features of Mel frequency cepstral coefficients, MFCCs, can be employed when classifying animal sounds with good effect (Cheng et al., 2010; Chung et al., 2013; Noda et al., 2016; Bishop et al., 2019). However, MFCCs often have more feature dimensionalities than other features, which can slow the computational rate. For this reason, an optimal feature combination from tristimulus values and MFCCs with fewer feature dimensionalities can perform better than individual features.

The classification algorithm is the second key component of sound recognition algorithms that operates on the feature output. Researchers have mainly used classifiers such as the Decision Tree (DT) (Digby et al., 2013; Moi et al., 2014; Mcgrath et al., 2017), Gaussian Mixture Model (GMM) (Cheng et al., 2010; Alonso et al., 2017; Jahn et al., 2017;), Neural Network (NN) (Mielke and Zuberbühler, 2013; Khunarsal et al., 2013; Favaro et al., 2014; González-Hernández et al., 2017), and Support Vector Machine (SVM) (Steen et al., 2012; Chung et al., 2013; Bishop et al.,2019). Although there has been no agreement on which classifier is the most suitable for poultry vocalisation classification, classifiers for chicken calls should be carefully considered (Ramachandran et al., 2002). The major challenge is that some laying hens' sounds overlap in the frequency domain. The GMM and NN methods can potentially solve this problem because they both have the ability to differentiate overlapping features, which are already well known in ASR (Automatic speech recognition) and animal vocalisation recognition.

Given the above rationale, this study aims to explore and develop an optimal recognition model for classifying hens' call types (including drinking, laying, twitter and grunt calls). The objectives of this study are as follows: (i) sound feature extraction, (ii) sound classification, and (iii) modelling analysis and comparison.

2. Materials and methods

2.1. Animal and housing

Experiments were conducted on a pilot farm (Shangzhuang Experimental Station of China Agricultural University, Beijing, China). Eleven Hy-Line brown hens were reared to an age of 35–36 weeks. The floor-rearing area was 1.5 m L × 1.35 m W × 1.8 m H (Fig. 1). The birds were given ad libitum access to food and water, and a timer-controlled light schedule (light period: 6:00 a.m. to 10:00 p.m.) was applied during the experimental period (35–36 weeks). The room environment was suitably controlled to maintain a good level of thermal comfort.

2.2. Data collection

A top-view Kinect camera for Windows V1 (Microsoft Corp., Redmond, WA, USA) was installed at a height of 1.8 m above the ground and

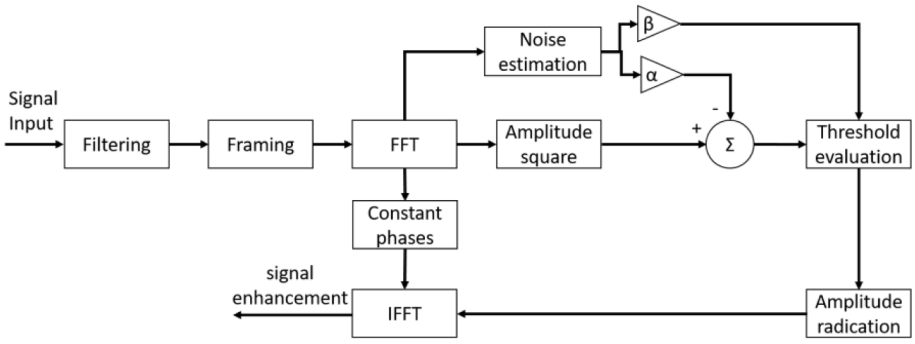

Fig. 2. Flowchart of the improved spectral subtraction algorithm.

used to continuously acquire sound data in WAV format (1 channel, 32 bit, 16,000 Hz, recording at approximately 55 s of each file) (Fig. 1). The Kinect was connected to a mini industrial personal computer (IPC) via a USB cable. A 2 TB storage USB 3.0 mobile hard disk drive (HDD) was used to store recorded sound data. Data were recorded for approximately 24-h a day for seven days (a total of 168 h). NI LabVIEW 2015 (American National Instrument Corp., Austin, TX, USA) and were used to pre-process and extract the sound features with the toolkits SVM (sound and vibration module) and MLT (machine learning toolkit), and the implementation was developed as part of the vocalisation classifier.

2.3. Sound signal pre-processing

An improved spectrum subtraction algorithm was used to pre-process raw sound data by filtering background noise. The algorithm transformed the sound signal from the time domain to the frequency domain, and squared difference values between the sound signal and noise were regarded as the estimated signal power spectrum (Fig. 2). Then, the results were retransformed into time-domain signals using the Fourier inverse transformation for the subsequent processing operation (Berouti et al., 1979; Upadhyay and Karmakar, 2013). This method has been proven to be a suitable de-noising approach with a low computational cost in practical situations (Du and Teng, 2017).

The improved spectrum subtraction method was calculated following Berouti et al. (1979) and Upadhyay and Karmakar (2013):

$$P(w) = P_s(w) - \alpha P_n(w) \tag{1}$$

$$P(w) = P_s(w) - \alpha P_n(w) \tag{2}$$

$$P_s(w) = \begin{cases} P(w), P(w) > \beta P_n(w) \\ \beta P_n(w), P(w) \leqslant \beta P_n(w) \end{cases} \tag{3}$$

where $P_s(w)$ is the amplitude of the noisy signal power spectrum; $P_n(w)$ is the amplitude of the noise power spectrum; α is the subtraction factor; and β is the spectral floor parameter ($\alpha \geqslant 1, 0 < \beta \leqslant 1$).

2.4. Labelling

After filtering, the sound data were labelled by manual audio-visual inspection (Tullo et al., 2017) performed by the first author, who is an experienced researcher in animal sound analysis. Audacity® software version 2.3.0 was used to label the data by human observers to inspect each recording and annotate the start and end time of the call events. In the process of replaying and visualising the sound recordings using spectrograms, overlapped sounds were not selected as testing samples for further analysis due to their complex acoustic features and the current limitations in sound source separation technology. Finally, there were approximately 15 min and 22 s of data being labelled for subsequent analysis (0.15% of the original data).

Table 1
Description of feature parameters.

Feature parameter	Description
F_1 (Hz)	The lowest frequency band with substantial energy is regarded as the first formant. This is the first harmonic of resonance
F_2 (Hz)	The second harmonic of resonance
F_3 (Hz)	The third harmonic of resonance
TF_1 (%)	F_1 energy ratio derives from the tristimulus values
TF_2 (%)	F_2 energy ratio derives from the tristimulus values
TF_3 (%)	F_3 energy ratio derives from the tristimulus values
MFCCs-12	12-dimensional MFCCs feature

2.5. Feature extraction

2.5.1. Mel frequency cepstral coefficients

One of the most popular sources of features that is widely used in animal vocalisation recognition is Mel frequency cepstral coefficients (MFCCs), which are short-term spectrum-based features. The extraction of MFCCs includes the following steps:

(1) Pre-emphasis

Usually, the system function is given by $H(z) = 1 - az^{-1}$, where $a \in [0.95, 0.98]$.

(2) Framing

Overlapping frames with a 50% overlap were recommended in each 0.2 s sound clip to avoid losing information, and the Hamming window was used to reduce the edge effects and spectral leakage in each frame.

(3) Discrete Fourier transform (DFT)

Every frame passed through a DFT, and the frequency band was filtered using a filter-bank of triangular filters spaced on the Mel-scale (approximately linear below 1 kHz and logarithmic above 1 kHz) (Noda et al., 2016).

$$Mel(f) = 2595 log\left(1 + \frac{f}{700}\right) \tag{4}$$

(4) Discrete cosine transformation (DCT)

The spectral envelope in the decibel unit was obtained by applying a logarithm to the amplitude spectrum. Then, the signal was processed via DCT.

The zeroth coefficient of the MFCCs is usually dropped because its value is the average log-energy. The first and second order coefficients of the MFCCs are often used as feature parameters. At the same time, this

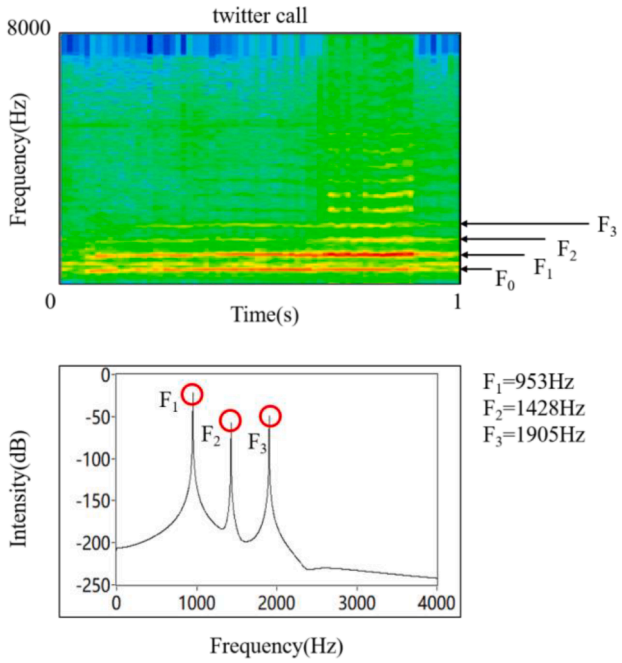

Fig. 3. Example of the extraction of the locations of the formants (twitter calls).

study only chose a 12-dimensional MFCC (12 vectors) because of the smaller number of feature dimensionalities.

2.5.2. *Tristimulus-formant feature*

The first three formants (F_1-F_3) of each call can represent the main timbre information (Table 1). A popular autocorrelation approach was used for tracking formants (Rabiner and Schafer, 2007). Fig. 3 shows an example of how to locate and label the formants (F_1-F_3). In the spectrogram setting, each frame consisted of N = 512 samples to comply with the size of the short-time Fourier transform (STFT). Overlapping frames with a 50% overlap were recommended to avoid losing information, and the Hamming window was used to reduce the edge effects and spectral leakage in each frame. To highlight the variability of the dominant formants, the tristimulus values of each of the formants were transformed into energy ratios of the first three formants as new tristimulus values (TF_1, TF_2, and TF_3).

$$TF_i = \frac{F_i}{F_1 + F_2 + F_3}, i = 1, 2, 3 \tag{5}$$

2.6. *Classification*

Two classifiers, BPNN and GMM, were chosen to recognise different call types, and a total of 4304 samples were labelled as training data and testing data. To avoid overfitting, k-fold cross-validation was used to estimate the accurate performance of the algorithm (Refaeilzadeh et al., 2009). Due to the limitations in the data size, the sound data were split into five smaller sets, and the algorithm was trained using four sets (80% of the data) and validated on the remaining set (20% of the data). The average of the five validation sets was used to measure the performance of the algorithm as demonstrated by Carpentier et al. (2018).

2.6.1. *BPNN*

The network selected for this study was a variation of a multilayer, backpropagation neural network, which is a commonly used NN for vocalisation recognition. The network consists of three parts: (1) the input layer; (2) hidden layer(s); and (3) the output layer (Mielke and Zuberbühler, 2013). The basic principle of the BPNN algorithm is that the learning process consists of two processes—information forward propagation and error back propagation. When information is propagated forward, the input sample is passed into the input layer and then transmitted to the output layer after processing in each hidden layer. If the actual output does not match the expected output, there is back propagation of the errors. In the back-propagation phase, the output is transmitted backward step by step through the hidden layers in a certain form, and the errors are distributed to all the elements of each layer to correct the weights according to the error signal. The definition of the error function is the sum of the square of the difference between the expected output and the actual output (Theodoridis, 2010).

$$e = \frac{1}{2}\sum_{p=1}^{m}\left(y_p - q_p\right)^2 \tag{6}$$

where e is the error function; y_p is the actual output; q_p is the expected output; p is the index of the output vectors; and m is the number of output vectors.

The training data set was used to minimise the error between the predicted call type (the output of the network) and the actual call type (the known sounds in the training set). The weights were adjusted using a gradient descent function with momentum and an adaptive learning rate (Khunarsal et al., 2013). The maximum iteration and tolerance were set at 1000 and 0.0001, respectively. Moreover, the accuracy, sensitivity and precision rates were chosen to assess the performance of the BPNN classifier (Carpentier et al., 2018).

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of total samples}} \times 100\% \tag{7}$$

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} \times 100\% \tag{8}$$

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}} \times 100\% \tag{9}$$

2.6.2. *GMM*

Feature classification methods developed for human speech recognition have been applied to species, individual and call type recognition in animals (Mielke and Zuberbühler, 2013; Jahn et al., 2017). GMM is also widely used because any probability distribution model can be approximated by a weighted combination of multiple Gaussian distributions. The parameters of the Gaussian Mixture Model were calculated by maximising the likelihood function and iteratively using the expectation-maximisation algorithm (Alonso et al., 2017). The mathematical expression of the GMM for the probability density function is shown as follows (Reynolds and Rose, 1995):

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{10}$$

where x is a d-dimensional random vector, $p_i(x)$, i = 1,2,…,M, is the component density and w_i, i = 1,2,…,M, is the mixture weight. The component densities are d-variate Gaussian functions given by (Reynolds and Rose, 1995):

$$p_i(x) = \frac{1}{\sqrt{(2\pi)^d det(\Sigma_i)}} exp\left[-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)\right] \tag{11}$$

where μ_i is the mean, Σ_i is the covariance matrix, and d is the number of features incorporated into every feature vector. The weights w_i must satisfy the following relation (Reynolds and Rose, 1995):

Table 2
Description of different sound types.

Sound type	Description	Number of sound clips
Drinking	Pecking sound for water, contains a short duration (<0.3 s) and a wide range of frequency band (1–8 kHz)	353
Twitter	Normal chirp call, contains a long duration (<1.0 s) and a distinct harmonic structure (0.1–2 kHz)	744
Laying	Sounds in the process of egg laying, contain a succession of shorter notes (>1.0 s) and a distinct harmonic structure	1984
Grunt	Sounds of hens snoring at night, contain a long duration (>0.5 s) and a narrow range of frequency band (0.1–5 kHz, concentrated energy within 0.1–2 kHz)	893
Fans	Sounds of mechanical noise, contain a random signal and a stable and narrow fundamental frequency band (<1000 Hz)	330
Total		4304

$$\sum_{i=1}^{M} w_i = 1 \tag{12}$$

Each model can be expressed as a function of the following parameters:$\lambda = (w_i, \mu_i, \Sigma_i)$, $i = 1,2,\ldots,M$.

For the Gaussian Mixture Model, different numbers of Gaussian components were selected. The EM algorithm was implemented with a maximum of 1000 iterations, and the value for tolerance was set to 0.0001. The Rand index was used to assess the clustering effect, which is expressed as (Theodoridis, 2010):

$$RI = \frac{a+b}{C_2^{n_{samples}}} \times 100\% \tag{13}$$

where C represents the actual call category, K represents the clustering result, a means that both C and K are elements of the same call type, b means that both of C and K are in different categories, and $C_2^{n_{samples}}$ represents the number of coupled samples from the data set. The value range of RI in Eq. (13) was 0–100%, which indicates the performance of the clustering effect of GMM.

3. Results and discussion

Comparative trials were conducted to determine the best features and classifier suitable for recognising the call types of laying hens. Table 2 and Fig. 4 present the descriptions of the different sound types and their spectrograms, respectively.

3.1. Using BPNN with different features

Comparative trials were designed, such as MFCCs-12+BPNN, MFCCs-3+TF+BPNN, Formants+TF+BPNN, and MFCCs-3+BPNN. As a result, MFCCs-12 had the highest accuracy, which was 94.9 ± 1.6% (Table 3). However, MFCCs-12 also had the longest training time. For this reason, dimensionality reduction is a must for practical online identification for analysing large data sets. After a series of tests, the 1st, 2nd and 5th dominant vectors of the MFCCs were extracted and reassembled into one 3-dimensional vector (MFCCs-3), which gave an acceptable recognition rate of 87.3 ± 3.3% (Tables 3 and 4). Next, a combination of multiple features was explored to determine whether the recognition rates might be improved, as suggested in the literature (Scheumann et al., 2012; Fukushima et al., 2015). Two combined features, formants-TF and MFCCs-3+TF, were selected for training and testing. Although formants-TF exhibited a non-ideal recognition rate, the joint-feature (MFCCs-3+TF) approach worked well. Both Figs. 5 and 6 give information about the visual classification results, which were evaluated by observing a distinct block class of each call type. As shown in the two figures, the feature points of MFCCs-12 are much closer than those of MFCCs-3+TF and the latter testing points have a more dispersed distribution, which means that MFCCs-3+TF can better disperse the feature points of different animal call types. In summary, compared with using MFCCs-3 or MFCCs-12 as single, independent features, the combined MFCCs-3+TF feature can not only increase the recognition effect of MFCCs-3 (from 87.3 ± 3.3% to 91.4 ± 1.4%) but also significantly reduce the model training time (from 3201 ± 119 ms to 2633 ± 54 ms).

Two confusion matrices were created to analyse the classification performance and difference between MFCCs-3+TF and MFCCs-12 (Tables 5 and 6). As shown in Table 3, fan noise was easily distinguished from the total 4304 sound clips because of the difference between the signal of a dynamic sound system (animal) and a static sound system (machine) (Du and Teng, 2017). In terms of the classification rates of drinking vs. a twitter call, MFCCs-3+TF outperforms MFCCs-12. The reason for the difference in performance could be due to the different vocal productions of the two call types because MFCCs-3+TF can better differentiate based on the TF feature. In contrast, MFCCs-12 is superior to MFCCs-3+TF in the recognition of laying vs. grunt calls. The main reason for the low precision and sensitivity of the grunt call might be due to the similarity of the first three formants in the laying call. Another reason might be the difficulty of tracking accurate formants in the grunt call.

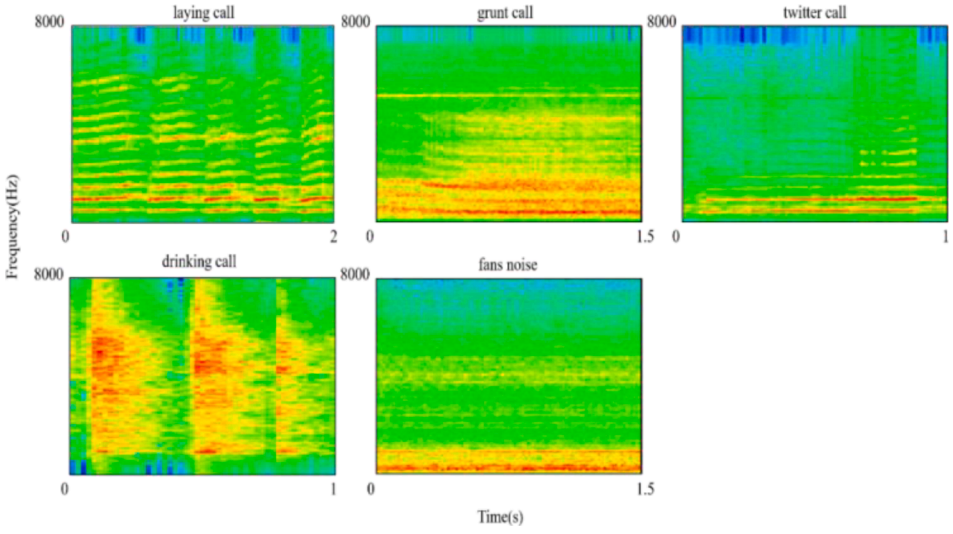

Fig. 4. Spectrograms of four hens' vocalisations and one mechanical noise.

Table 3
Classification performance using the BPNN classifier.

Feature category	Model training time ± SD (ms)	Call type	Classification performance		
			Sensitivity ± SD (%)	Precision ± SD (%)	Accuracy ± SD (%)
MFCCs-12-12D	3201 ± 119	Drinking	84.2 ± 4.9	83.3 ± 4.3	–
		Twitter	90.1 ± 4.1	92.1 ± 2.0	–
		Laying	97.5 ± 2.9	97.3 ± 1.9	–
		Grunt	95.4 ± 5.5	95.2 ± 5.4	–
		Fans	100.0 ± 0.0	100.0 ± 0.0	–
		Total	93.4 ± 1.4	93.6 ± 1.7	94.9 ± 1.6
MFCCs-3+TF-6D	2633 ± 54	Drinking	84.3 ± 3.7	89.3 ± 2.3	–
		Twitter	93.8 ± 1.1	91.6 ± 1.5	–
		Laying	92.3 ± 0.6	94.5 ± 0.2	–
		Grunt	87.3 ± 2.4	81.1 ± 3.3	–
		Fans	97.2 ± 0.9	100.0 ± 0.0	–
		Total	91.0 ± 1.5	91.3 ± 1.7	91.4 ± 1.4
Formants+TF-6D	2667 ± 28	Drinking	77.6 ± 12.1	79.1 ± 5.6	–
		Twitter	18.7 ± 3.9	59.0 ± 8.5	–
		Laying	89.6 ± 4.8	67.3 ± 1.9	–
		Grunt	61.5 ± 4.8	67.5 ± 5.8	–
		Fans	61.2 ± 9.6	82.5 ± 16.1	–
		Total	61.7 ± 1.5	71.7 ± 4.5	68.4 ± 1.6
MFCCs-3-3D	2395 ± 38	Drinking	78.7 ± 7.4	75.6 ± 8.2	–
		Twitter	83.5 ± 4.1	86.8 ± 4.5	–
		Laying	89.5 ± 8.0	92.0 ± 5.6	–
		Grunt	84.5 ± 12.6	82.3 ± 13.4	–
		Fans	100 ± 0.0	97.6 ± 5.4	–
		Total	87.2 ± 2.3	86.8 ± 3.1	87.3 ± 3.3

Note: D is an abbreviation for dimension. – means null value.

Table 4
Accuracy rate of each MFCC vector using the BPNN classifier.

Vector	1	2	3	4	5	6
Accuracy ± SD (%)	75.4 ± 0.6	62.8 ± 1.4	54.6 ± 2.0	56.5 ± 1.1	66.3 ± 0.8	58.4 ± 1.0
Vector	7	8	9	10	11	12
Accuracy ± SD (%)	60.7 ± 0.7	54.2 ± 0.9	54.1 ± 1.2	61.6 ± 0.3	56.5 ± 2.1	48.0 ± 1.0

3.2. Using GMM with different features

Comparative trials were also designed based on the GMM model, such as MFCCs-12+GMM, MFCCs-3+TF+GMM, Formants+TF+GMM, and MFCCs-3+GMM. As shown in Table 7, MFCCs-12 still shares the highest *RI* index of 91.7 ± 5.3%, with the longest model training time. Both MFCCs-3 and MFCCs-12 outperform MFCCs-3+TF in regard to the clustering effect, but MFCCs-3+TF is superior to other features at the model training time. Both Figs. 7 and 8 show the visual clustering results that can be intuitively evaluated by observing the matching degree between the training set and testing set. Unfortunately, many feature points cannot be classified correctly by using MFCCs-3+TF, but it can better disperse the feature points of different call types (Fig. 7). In short, compared with MFCCs-3 and MFCCs-12 that take a single feature at a time, the combined feature MFCCs-3+TF using GMM can decrease the model training time but has an inferior classification rate.

3.3. Comparison of BPNN and GMM performances

To identify an optimal recognition model for recognising hen call types, the accuracy rate, *RI* and model training time of all of the call types were calculated. Figs. 9 and 10 show the differences in the performance between the BPNN and GMM classifiers. As shown in these two figures, the BPNN classifier obviously outperforms the GMM classifier, and the former also has a shorter model training time. Moreover, the MFCCs-12 feature shares the longest model training time in spite of its high classification rate, which is not suitable for big data analysis. In contrast, the novel MFCC-3+TF feature is more competent for big data analysis as well as for real-time monitoring because it can effectively recognise hen call types at a low computational cost (a 12.8–22.3% decrease in the execution time).

Artificial Neural Networks (ANNs) were first introduced in animal behavioural studies in the early 1990s of the past century, and today, they have been widely used as a valuable acoustic classification tool (Reby et al., 1997; Pozzi et al., 2009). Compared with traditional statistical approaches, the largest advantage of ANNs is their ability to model complex and non-linear relationships among acoustic parameters (Favaro et al., 2014). In this paper, the proposed method can be used to recognise the five call types of laying hens with a high accuracy of 94.9 ± 1.6% (MFCCs-12+BPNN model) and 91.4 ± 1.4% (MFCCs-3+TF+BPNN model). The average precision rates are 93.6 ± 1.7% (MFCCs-12+BPNN model) and 91.3 ± 1.7% (MFCCs-3+TF+BPNN model). Other similar animal sound recognition rates are the following: 98% for blue monkeys (2 call types: 'pyow' and 'hack' calls) (Mielke and Zuberbühler, 2013), 92% for geese (an average accuracy for 3 behaviours) and 84% (an average precision for 3 behaviours) (Steen et al., 2012), 80.4–92.5% for birds (Cheng et al., 2010), 90% for marine mammals (three call types: whistles, calls and squeaks) (González-Hernández et al., 2017), 84% for cattle (three ingestive behaviours: chews, bites and composite chew-bites) (Chelotti et al., 2016) and 92.5–95.6% for black lemurs(Pozzi et al., 2009). Favaro demonstrated that ANNs are a powerful tool for studying goat kid contact calls. For each call, 27 spectral and temporal acoustic parameters (including formant parameters) were measured, and the accuracy rates were 71.1 ± 1.2% (vocal individuality, 10 goats), 79.6 ± 0.8% (3 social groups), 91.4 ± 0.8% (maturation, 2 classes) (Favaro et al., 2014). Similarly, the proposed Formants+TF (5 classes) also show a low accuracy of 68.4 ± 1.6%. At the same time, the novel combined feature MFCCs-3+TF has a high accuracy of 91.4 ± 1.4%, which overmatches the classification performance found in previous research with fewer feature dimensionalities. In previous research, fuzzy logic values and a feedforward network was used to classify alarm call barks, with 21 neurons in the input layer and 50 neurons in the hidden layer. For different predator species, the lowest accuracy of 79% was obtained when classifying all four species together (Placer and Slobodchikoff, 2000). Compared with previous studies, the proposed method can better classify 5 classes with a smaller number of neurons and features. It is very difficult to theoretically estimate the number of hidden layers due the possibility of overfitting and additional training time. Increasing this number can further enhance the risk of overfitting. Training time can be saved by avoiding overfitting (Reby et al., 1997). The problem of overfitting the training set (overlearning) can be overcome using cross-validation sets

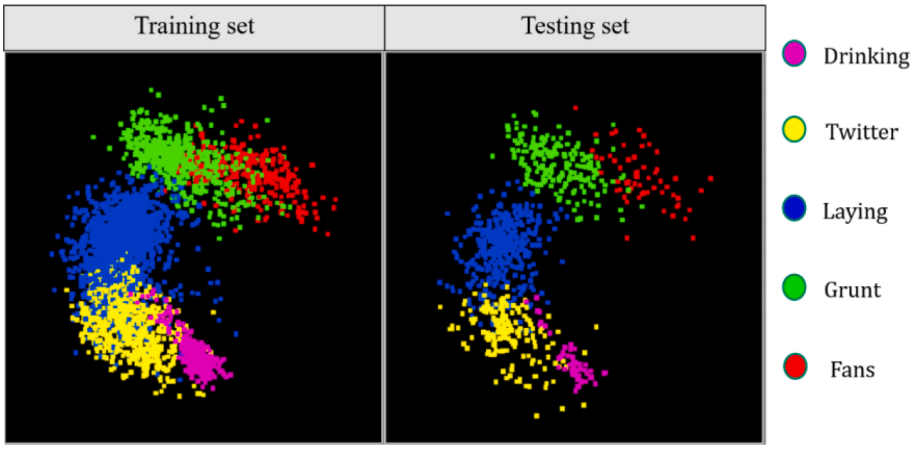

Fig. 5. Classification results using the BPNN classifier (MFCCs-3+TF feature).

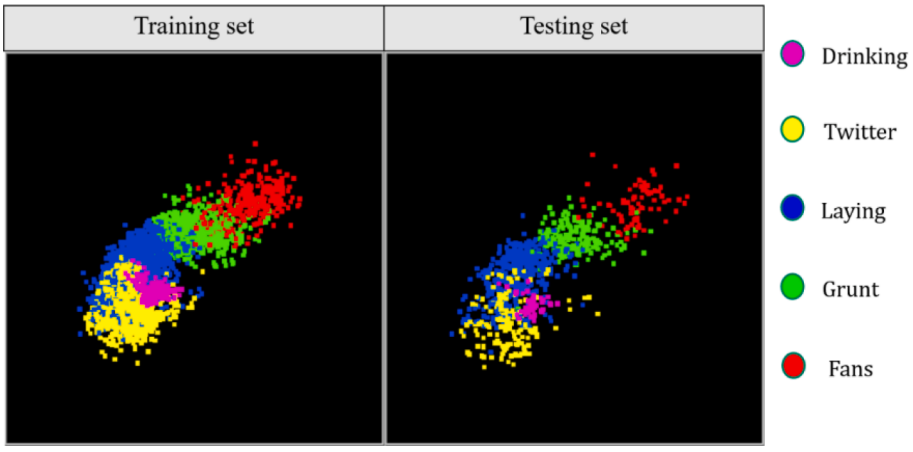

Fig. 6. Classification results using the BPNN classifier (MFCCs-12 feature).

Table 5
Confusion matrix of one validation set (MFCCs-3+TF+BPNN model).

Actual call type	Classified by MFCCs-3+TF feature						
	Drinking	Twitter	Laying	Grunt	Fans	Total	Sensitivity (%)
Drinking	65	12	0	0	0	77	84.4
Twitter	5	158	2	0	0	165	95.8
Laying	0	2	382	30	0	414	92.3
Grunt	0	0	20	128	0	148	86.5
Fans	0	0	0	1	56	57	98.2
Total	70	172	404	159	56	861	–
Precision (%)	92.9	91.9	94.6	80.5	100.0	–	91.6*

Note: * means the accuracy rate. – means null value.

Table 6
Confusion matrix of one validation set (MFCCs-12+BPNN model).

Actual call type	Classified by MFCCs-12 feature						
	Drinking	Twitter	Laying	Grunt	Fans	Total	Sensitivity (%)
Drinking	57	13	0	0	0	70	81.4
Twitter	10	138	1	0	0	149	92.6
Laying	0	0	392	5	0	397	98.7
Grunt	0	0	5	174	0	179	97.2
Fans	0	0	0	0	66	66	100.0
Total	67	151	398	179	66	861	–
Precision (%)	85.1	91.4	98.5	97.2	100.0	–	94.0*

Note: * means the accuracy rate. – means null value.

Table 7
Classification performance using the GMM classifier.

Performance parameters	MFCCs-12-12D	MFCCs-3+TF-6D	Formants+TF-6D	MFCCs-3-3D
RI (%)	91.5 ± 5.3	73.0 ± 2.1	62.1 ± 0.7	81.9 ± 0.5
Model training time (ms)	3458 ± 151	2587 ± 107	2689 ± 41	2876 ± 55

Note: D is an abbreviation for dimension.

(as employed in this study). The proposed method chose 5 hidden neurons after optimised tests (Khunarsal et al., 2013).

The major disadvantage of machine learning algorithms is that they require large numbers of samples to train the model for high accuracy. Moreover, the training stage of most of the ML algorithms is computationally demanding due to the large number of features used as inputs (Acevedo et al., 2009). To overcome this problem, an optimal feature combination of TF and MFCCs with fewer feature dimensions can classify well without compromising precision or accuracy (training time reducing from 3201 ± 119 ms to 2633 ± 54 ms). In this paper, each call with only 6 feature variables is sufficient to obtain an acceptable classification.

Source-filter theory has been considered to be the commonly used theory for explaining the acoustic characteristics of bird vocalisations (Favaro et al., 2015). Williams concluded that the syrinx in birds can vary the harmonic amplitude output (Williams et al., 1989). Moreover, formants are completely independent of the fundamental frequency (F_0) (Fitch and Kelley, 2000). In this paper, the chosen filter-related features (F_1-F_3, TF_1-TF_3) are different among the five call types of laying hens, which is helpful to disperse the feature points for better recognition performance. Additionally, the formant parameters can be used to estimate the biological information of mammals, such as the vocal tract length (Reby and McComb, 2003). At the same time, the mammal vocal tract model might not be suitable for hens because the structure of their respiratory system is very different from that of mammals (Taylor and Reby, 2010). Unfortunately, we did not perform a physiological autopsy on the test chicken and were unable to verify the true vocal tract length. This matter remains to be fully investigated in future research.

Moreover, on-site machinery noise is an influencing factor that can reduce classification rates. Other researchers have suggested that ANNs may be very helpful to assign calls with high background noise (Pozzi et al., 2009). To date, it is still a challenge for researchers to implement sound algorithms in a commercial henhouse that stocks a large population of animals (approximately 50 K broilers, 80 K laying hens) because of the large quantity of sounds produced during the daytime. At the same time, it was found that hens' vocalisations during the night were less than those during the daytime and that most of the vocalisations were sounds that indicated animal health and production performance, such as the sound of egg laying, grunting and coughing. The application of sound source localisation (SSL) algorithms makes it possible to detect anomalous animal vocalisations at night by monitoring the number of concerned vocalisations and the area distributions

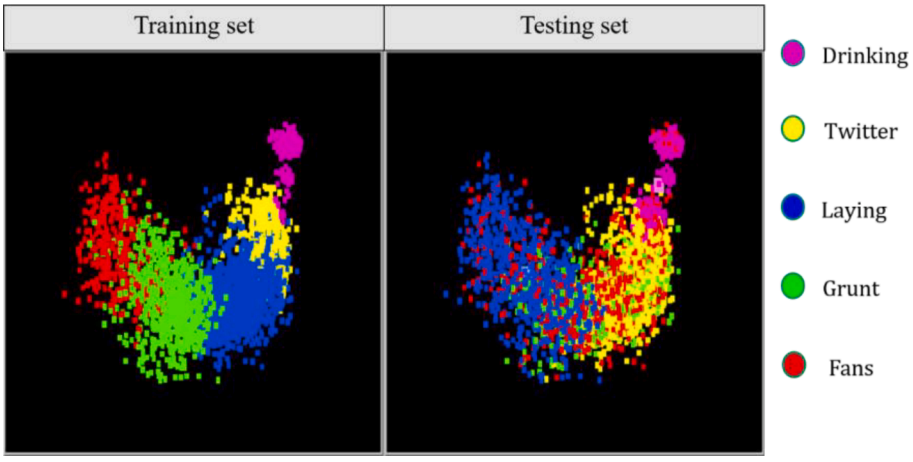

Fig. 7. Classification results using the GMM classifier (MFCCs-3+TF feature).

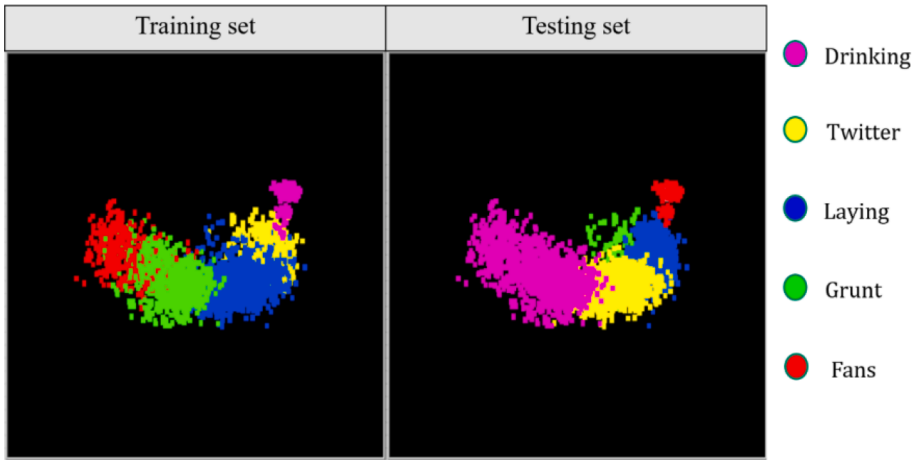

Fig. 8. Classification results using the GMM classifier (MFCCs-12 feature).

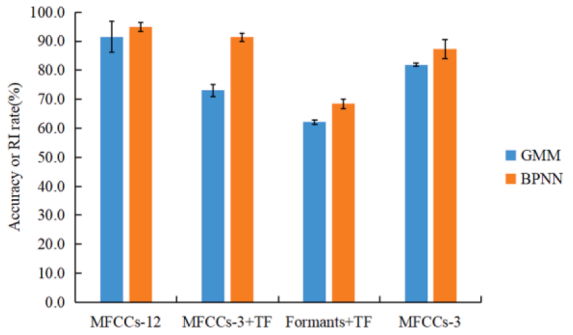

Fig. 9. Comparison of the classification performance between GMM and BPNN.

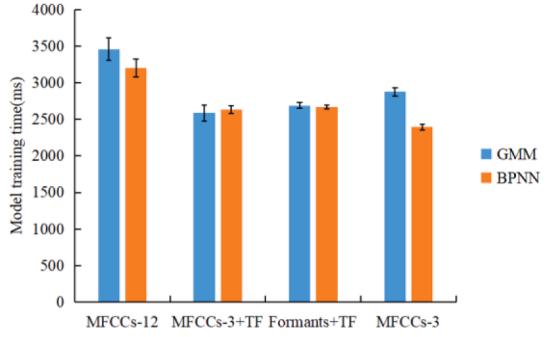

Fig. 10. Comparison of the model training time between GMM and BPNN.

for precision analysis (Du et al., 2018). Here, it is recommended to use the tristimulus-formant model to monitor birds when there are fewer than one thousand in a subarea and the SNR (Signal-to-noise ratio) >5 dB (Du and Teng, 2017). Additionally, the performance of the algorithm might be lower than expected in a chicken barn because the distance between a sound source and a microphone is an affecting factor. A long distance can lead to an inadequate sound quality and a low sound intensity. These problems have not yet been solved completely and remain to be fully investigated. Further studies can explore the possibility of combining call type recognition and SSL algorithms for the automatic detection of specific sounds in a sub-area, which can be considered to be one of the potential applications.

4. Conclusions

In this study, we determined which acoustic features and classifiers have the potential to better recognise each call type of laying hens. The novel model "MFCCs-3+TF+BPNN" performs well without compromising accuracy in recognising hen vocalisations. This model also has less training time and fewer feature dimensions (6 variables) than those of other models. Compared with other animal sound recognition approaches, the proposed model shows considerable potential for online identification and for large data analysis. Further research could be performed to study the relationship between animal behaviour recognition and animal sound recognition by using a multi-modality of video and sound streaming technology.

CRediT authorship contribution statement

Xiaodong Du: Data curation, Investigation, Methodology, Software, Writing - original draft. **Guanghui Teng:** Methodology, Project administration, Writing - review & editing. **Chaoyuan Wang:** Project administration, Writing - review & editing. **Lenn Carpentier:** Formal analysis, Software. **Tomas Norton:** Formal analysis, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by the National Key Research and Development Program of China (No. 2017YFD0701602 and No. 2016YFD0700204) and China Scholarship Council (CSC No. 201806350182).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compag.2021.106221.

References

Acevedo, M.A., Corrada-Bravo, C.J., Corrada-Bravo, H., Villanueva-Rivera, L.J., Aide, T. M., 2009. Automated classification of bird and amphibian calls using machine learning: a comparison of methods. Ecol. Inf. 4, 206–214.

Alonso, J.B., Cabrera, J., Shyamnani, R., Travieso, C.M., Bolaños, F., García, A., Villegas, A., Wainwright, M., 2017. Automatic anuran identification using noise removal and audio activity detection. Expert Syst. Appl. 72, 83–92.

Aydin, A., Bahr, C., Viazzi, S., Exadaktylos, V., Buyse, J., Berckmans, D., 2014. A novel method to automatically measure the feed intake of broiler chickens by sound technology. Comput. Electron. Agric. 101, 17–23.

Aydin, A., Berckmans, D., 2016. Using sound technology to automatically detect the short-term feeding behaviours of broiler chickens. Comput. Electron. Agric. 121, 25–31.

Beckers, G.J., Suthers, R.A., Ten, C.C., 2003. Pure-tone birdsong by resonance filtering of harmonic overtones. PNAS 100, 7372–7376.

Berckmans, D., Hemeryck, M., Berckmans, D., Vranken, E., van Waterschoot, T., 2015. Animal sound...talks! real-time sound analysis for health monitoring in livestock. In: Proceedings of Animal Environment and Welfare, Chongqing, China, 23-26 October, 215-222.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. ICASSP 79. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 208-211.

Bishop, J.C., Falzon, G., Trotter, M., Kwan, P., Meek, P.D., 2019. Livestock vocalisation classification in farm soundscapes. Comput. Electron. Agric. 162, 531–542.

Cao, Y., Yu, L., Teng, G., Zhao, S., Liu, X., 2014. Feature extraction and classification of laying hens' vocalization and mechanical noise. Trans. Chinese Soc. Agric. Eng. 18, 190–197 (in Chinese).

Carpentier, L., Berckmans, D., Youssef, A., Berckmans, D., van Waterschoot, T., Johnston, D., Ferguson, N., Earley, B., Fontana, I., Tullo, E., Guarino, M., Vranken, E., Norton, T., 2018. Automatic cough detection for bovine respiratory disease in a calf house. Biosyst. Eng. 173, 45–56.

Chelotti, J.O., Vanrell, S.R., Milone, D.H., Utsumi, S.A., Galli, J.R., Rufiner, H.L., Giovanini, L.L., 2016. A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. Comput. Electron. Agric. 127, 64–75.

Cheng, J., Sun, Y., Ji, L., 2010. A call-independent and automatic acoustic system for the individual recognition of animals: a novel model using four passerines. Pattern Recogn. 43, 3846–3852.

Chung, Y., Oh, S., Lee, J., Park, D., Chang, H.H., Kim, S., 2013. Automatic detection and recognition of pig wasting diseases using sound data in audio surveillance systems. Sensors 13, 12929–12942.

de Moura, D.J., Naeaes, I.D.A., de Souza Alves, E.C., Ridolfi De Carvalho, T.M., Do Vale, M.M., de Lima, K.A.O., 2008. Noise analysis to evaluate chick thermal comfort. Scientia Agricola 65, 438–443.

Digby, A., Towsey, M., Bell, B.D., Teal, P.D., 2013. A practical comparison of manual and autonomous methods for acoustic monitoring. Methods Ecol. Evol. 4, 675–683.

Du, X., Teng, G., 2017. Research on an improved de-noising method of laying hens' vocalization. Trans. Chinese Soc. Agric. Machinery 48 (12), 327–333 (in Chinese).

Du, X., Lao, F., Teng, G., 2018. A sound source localisation analytical method for monitoring the abnormal night vocalisations of poultry. Sensors 18, 2906.

Du, X., Carpentier, L., Teng, G., Liu, M., Wang, C., Norton, T., 2020. Assessment of laying hens' thermal comfort using sound technology. Sensors 20 (2), 473.

Exadaktylos, V., Silva, M., Aerts, J.M., Taylor, C.J., Berckmans, D., 2008. Real-time recognition of sick pig cough sounds. Comput. Electron. Agric. 63, 207–214.

Favaro, L., Briefer, E.F., McElligott, A.G., 2014. Artificial neural network approach for revealing individuality, group membership and age information in goat kid contact calls. Acta Acustica United with Acustica 100, 782–789.

Favaro, L., Gamba, M., Alfieri, C., Pessani, D., McElligott, A.G., 2015. Vocal individuality cues in the African penguin (Spheniscus demersus): a source-filter theory approach. Sci. Rep. 5.

Favaro, L., Gamba, M., Gili, C., Pessani, D., 2017. Acoustic correlates of body size and individual identity in banded penguins. PLoS ONE 12, e0170001.

Fitch, W.T., Kelley, J.P., 2000. Perception of vocal tract resonances by whooping cranes grus americana. Ethology 106, 559–574.
Fletcher, N.H., 1988. Bird song - a quantitative acoustic model. J. Theor. Biol. 135, 455–481.
Fukushima, M., Doyle, A.M., Mullarkey, M.P., Mishkin, M., Averbeck, B.B., 2015. Distributed acoustic cues for caller identity in macaque vocalization. R. Soc. Open Sci. 2, 150432.
González-Hernández, F.R., Sánchez-Fernández, L.P., Suárez-Guerra, S., Sánchez-Pérez, L. A., 2017. Marine mammal sound classification based on a parallel recognition model and octave analysis. Appl. Acoust. 119, 17–28.
Guarino, M., Norton, T., Berckmans, D., Vranken, E., Berckmans, D., 2017. A blueprint for developing and applying precision livestock farming tools: a key output of the EU-PLF project. Animal Frontiers 7, 12.
Jahn, O., Ganchev, T.D., Marques, M.I., Schuchmann, K.L., 2017. Automated sound recognition provides insights into the behavioral ecology of a tropical bird. PLoS ONE 12, e0169041.
Kashiha, M., Bahr, C., Haredasht, S.A., Ott, S., Moons, C.P.H., Niewold, T.A., Ödberg, F. O., Berckmans, D., 2013. The automatic monitoring of pigs water use by cameras. Comput. Electron. Agric. 90, 164–169.
Khunarsal, P., Lursinsap, C., Raicharoen, T., 2013. Very short time environmental sound classification based on spectrogram pattern matching. Inf. Sci. 243, 57–74.
Manteuffel, G., Puppe, B., Schön, P.C., 2004. Vocalization of farm animals as a measure of welfare. Appl. Animal Behav. Sci. 88, 163–182.
Mcgrath, N., Dunlop, R., Dwyer, C., Burman, O., Phillips, C.J.C., 2017. Hens vary their vocal repertoire and structure when anticipating different types of reward. Anim. Behav. 130, 79–96.
Mielke, A., Zuberbühler, K., 2013. A method for automated individual, species and call type recognition in free-ranging animals. Anim. Behav. 86, 475–482.
Moi, M., Naeaes, I.D.A., Caldara, F.R., Almeida Paz, I.C.D.L., Garcia, R.G., Cordeiro, A.F. S., 2014. Vocalization data mining for estimating swine stress conditions. Engenharia Agricola 34, 445–450.
Noda, J., Travieso, C., Sánchez-Rodríguez, D., 2016. Automatic taxonomic classification of fish based on their acoustic signals. Appl. Sci. 6, 443.
Placer, J., Slobodchikoff, C.N., 2000. A fuzzy-neural system for identification of species-specific alarm calls of Gunnison's prairie dogs. Behav. Process. 52, 1–9.
Pluk, A., Cangar, O., Bahr, C., Vranken, E., Berg, G.V.D., Berckmans, D., 2010. Impact of process related problems on water intake pattern of broiler chicken. In: Proceedings of the International Conference on Agricultural Engineering, Clermont-Ferrand, France. 6-8 September, 29.
Pollard, H.F., Jansson, E.V., 1982. A tristimulus method for the specification of musical timber. Acustica 51, 162–171.
Pozzi, L., Gamba, M., Giacoma, C., 2009. The use of artificial neural networks to classify primate vocalizations: a pilot study on black lemurs. Am. J. Primatol. 72, 337–348.
Rabiner, R.L., Schafer, W.R., 2007. Introduction to digital speech processing. Found. Trends Signal Process. 1 (1–2), 1–194.
Ramachandran, R.P., Ramachandran, R., Farrell, K.R., Mammone, R.J., 2002. Speaker recognition—general classifier approaches and data fusion methods. Pattern Recogn. 35, 2801–2821.
Reby, D., Lek, S., Dimopoulos, I., Joachim, J., Lauga, J., Aulagnier, S., 1997. Artificial neural networks as a classification method in the behavioural sciences. Behav. Process. 40, 35–43.
Reby, D., McComb, K., 2003. Anatomical constraints generate honesty: acoustic cues to age and weight in the roars of red deer stags. Anim. Behav. 65, 519–530.
Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-validation. In: Liu, L., Ozsu, M.T. (Eds.), Encyclopedia of database systems, Springer US, Boston, MA, pp. 532-538.
Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech & Audio Processing, 1995, 3 (1), 72-83.
Scheumann, M., Roser, A.E., Konerding, W., Bleich, E., Hedrich, H.J., Zimmermann, E., 2012. Vocal correlates of sender-identity and arousal in the isolation calls of domestic kitten (Felis silvestris catus). Front. Zool. 9, 36.
Silva, M., Ferrari, S., Costa, A., Aerts, J.M., Guarino, M., Berckmans, D., 2008. Cough localization for the detection of respiratory diseases in pig houses. Comput. Electron. Agric. 64, 286–292.
Steen, K.A., Therkildsen, O.R., Karstoft, H., Green, O., 2012. A vocal-based analytical method for goose behaviour recognition. Sensors 12, 3773–3788.
Taylor, A.M., Reby, D., 2010. The contribution of source-filter theory to mammal vocal communication research. J. Zool. 280, 221–236.
Theodoridis, S., 2010. Pattern Recognition, fourth ed. Publishing house of electronics industry, Beijing.
Tullo, E., Fontana, I., Diana, A., Norton, T., Berckmans, D., Guarino, M., 2017. Application note: Labelling, a methodology to develop reliable algorithm in PLF. Comput. Electron. Agric. 142, 424–428.
Upadhyay, N., Karmakar, A., 2013. Spectral subtractive-type algorithms for enhancement of noisy speech: an integrative review. Int. J. Image, Graphics Signal Process. 5 (11), 13–22.
Van Hirtum, A., Berckmans, D., 2004. Objective recognition of cough sound as biomarker for aerial pollutants. Indoor Air 14, 10–15.
Vandermeulen, J., Kashiha, M., Ott, S., Bahr, C., Moons, C. P. H., Tuyttens, F., Niewold, T. A., Berckmans, D., 2013. Combination of image and sound analysis for behaviour monitoring in pigs. In: Proceedings of the 6th European conference on Precision Livestock Farming, Leuven, Belgium, 10-12 September, pp. 62-67.
Williams, H., Cynx, J., Nottebohm, F., 1989. Timbre control in zebra finch (Taeniopygia guttata) song syllables. J. Comp. Psychol. 103, 366–380.
Yeon, S.C., Jeon, J.H., Houpt, K.A., Chang, H.H., Lee, H.C., Lee, H.J., 2006. Acoustic features of vocalizations of Korean native cows (Bos taurus coreanea) in two different conditions. Appl. Anim. Behav. Sci. 101, 1–9.