Journal of Computers Vol. 34 No. 4, August 2023, pp. 127-143
doi: 10.53106/199115992023083404011

Animal Vocal Recognition-Based Breeding Tracking and Disease Warning

Yinggang Xie[1,2*], Yangpeng Xiao[1], Xuewei Peng[1], Qijia Liu[1]

[1] Key Laboratory of Information and Communication Systems, Ministry of Information Industry,
Beijing Information Science and Technology University, Beijing, 100101, China
{xieyinggang, yangpeng.xiao, xuewei.peng}@bistu.edu.cn

[2] Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument,
Beijing Information Science and Technology University, Beijing, 100101, China

Received 1 July 2023; Revised 20 July 2023; Accepted 25 July 2023

Abstract. In this study, we investigate the application of a deep learning framework for the recognition of pig vocalizations. This innovative approach aims to actively monitor and evaluate the diverse states of pigs, with an overarching objective to improve the efficiency of pig farming through prompt identification and resolution of issues. In our comprehensive data collection effort, we carefully gathered a vast assortment of vocal samples from 50 pigs, representative of four distinct states: normal, frightened, coughing, and sneezing. We then meticulously analyzed this vocal data using Mel Frequency Cepstral Coefficients (MFCC). For accurate recognition of pig vocalizations, we devised a fusion model that combines the strengths of Residual Networks (ResNet) and Long Short-Term Memory Networks (LSTM). This model was subsequently tailored, trained, and optimized to meet our specific requirements. Upon rigorous evaluation, we found our model to exhibit exceptional performance in pig vocal recognition tasks, thereby reinforcing the potential of deep learning methodologies in revolutionizing the livestock industry. This research notably underscores the potential of deploying efficient real-time health monitoring systems, offering a promising avenue towards modernizing livestock management practices.

Keywords: deep learning, pig vocal recognition, MFCC, ResNet, real-time health monitoring

1 Introduction

With the rapid advancement of technology, modern agriculture is undergoing continuous transformation and upgrading. Particularly within animal production, the application of intelligent technology has become a critical tool for enhancing productivity and reducing operational costs. As China's swine industry stands at the dawn of a transformative phase [1], improving breeding techniques, reducing cultivation costs, and promoting the intelligent and scalable growth of the livestock industry have emerged as pressing issues to address [2].

Traditional pig farming methods often require substantial labor and material resources for regular monitoring and inspection [3]. Moreover, there are notable challenges concerning disease prevention and control, as well as monitoring the conditions of the pigs. The emergence of intelligent pig farming technology has the potential to revolutionize the industry. By leveraging a variety of sensors and equipment, it is possible to monitor the physiological conditions of pigs and their production environment in real-time, significantly elevating the precision of management practices in pig farming.

However, the current state of intelligent pig farming technology has its shortcomings [4], such as the unidimensional nature of monitoring methods and the underutilization of collected data. Against this backdrop, we recognize the immense potential of voiceprint recognition in animal farming. By effectively employing this cutting-edge technology, we foresee an opportunity to enhance not only the efficiency and accuracy of monitoring practices but also to unlock untapped potentials in the modern livestock industry.

Voiceprint recognition is a technology that identifies individuals by analyzing the characteristics of their voice, which has been extensively employed in human speech recognition [5, 6] and identity verification [7, 8]. Research in voiceprint recognition commenced in the 1950s, yet in the realm of animal production [9], notably pig farming, the research and application of voiceprint technology has been significantly delayed and remains in its nascent stage. Schlegel Patrick et al. [10] studied quantitative voice outcome tracking in three Yucatan minipigs. A.T. Kavlak et al. [11] predicted the health status (sick or healthy) of pigs based on raw feeding behavioural

* Corresponding Author

features and features derived using a machine learning algorithm (Xgboost), with a final detection accuracy of 80%. Yanling Yin et al. [12] improved the previously proposed feature fusion algorithm by selecting acoustic and image features for fusion and proposed a new classifier fusion algorithm that fused support vector machine classifiers trained from acoustic and depth features for pig cough prediction by soft voting, with a final accuracy of 97.47%. Weizheng Shen et al. [13] proposed a feature fusion method combining acoustic and depth features of audio clips, and this fusion achieved 97.35% accuracy in pig cough recognition. Jian Zhao et al. [14] proposed a new method based on Deep Neural Network Hidden Markov Model (DNN-HMM) to construct an acoustic model for continuous pig cough recognition, and compared the conventional acoustic model Gaussian Mixture Model Hidden Markov Model (GMM-HMM) with DNN-HMM. It was found that the word error rate of all groups in DNN-HMM was lower than that of GMM-HMM, with an average word error rate of 3.45% lower. Ji, Nan et al. [15] proposed a method that combines acoustic and visual features to improve the recognition rate of pig coughs. It is shown that the fused acoustic features (Acoustic) combined with LBP and HOG (A-LH) achieved 96.45% pig cough accuracy. Haonan Sun et al. [16] took the output of the time regularization algorithm as the input of the support vector machine, STE-M established the DTW-SVM tandem model for the sound feature parameters, optimized the DTW-SVM tandem model, and used the particle swarm optimization algorithm to select the most suitable penalty factor and kernel function width, and established the PSODTW-SVM recognition model, with an average recognition rate of 85.17%.

Although these researchers have achieved many important results in the field of pig vocal recognition, however, these traditional methods often face some challenges when dealing with complex pig vocal classification tasks. With the development of deep learning techniques, we have new opportunities to improve this situation. Therefore, the aim of this paper is to implement deep learning techniques to classify and identify pig sounds in order to promote welfare farming and improve health monitoring of pigs. The MFCC feature parameters of pig sound signals were extracted by analysing the sounds collected in the field, after sound pre-emphasis, windowing and framing, and endpoint detection for multiple categories of sounds. A ResNet-LSTM network model was selected to train a classification model using the extracted sound features to build a pig sound recognition system that effectively identifies the sounds of different states of different pigs.

2 Technical Principles

2.1 Digital Model of the Speech Signal

In vocal recognition studies, understanding the process of sound production is pivotal [17]. Firstly, it aids us in discerning the unique characteristics of a voice [18]. Secondly, the mechanisms of sound generation have implications for extracting vocal characteristics, a key aspect of vocal recognition [19, 20]. Lastly, gaining deeper insights into the process of sound generation can steer improvements and optimizations in our voiceprint recognition techniques [21], such as crafting more precise voiceprint recognition models or devising innovative feature extraction methods. Fig. 1 illustrates a discrete time-domain model of speech signal generation:

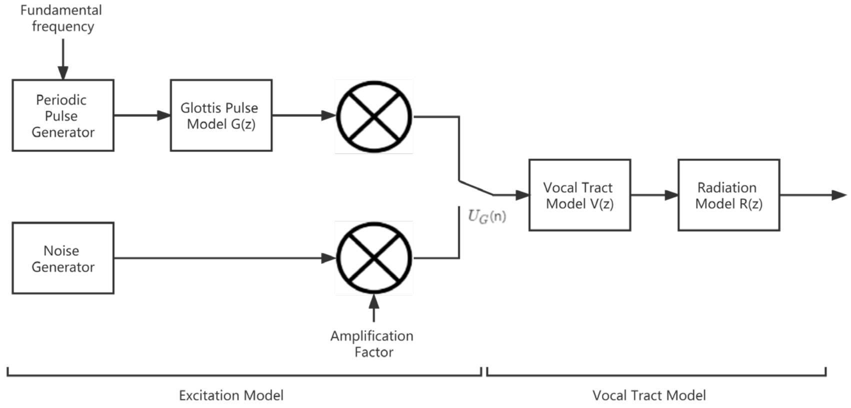


Fig. 1. Discrete time-domain model for speech signal generation

As the diagram reveals, the speech signal is a product of the excitation signal $U_G(n)$ processed through a linear system V(z). Here, the vocal tract model V(z) is a discrete time-domain vocal tract transfer function, typically approximated by an all-pole function. As different speakers have unique vocal tract shapes, they correspondingly have distinct vocal tract models. Among the speech generation models presented, V(z) most aptly captures the differences between speakers [22], and V(z) is expressed as:

$$V(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{i=1}^{p} a_i z^{-i}}, \tag{1}$$

where p is the order of the all-pole filter and a(i=1, 2, ..., p) are the coefficients of the filter. Higher-order models can more accurately fit the transfer function; however, an excessively high order can lead to overfitting, increased computational cost, and complex parameter estimation. Therefore, striking a balance between model complexity and generalization capability is essential when selecting the order [23], typically within the range $8<p<12$. The fundamental concept of the presented speech generation model is to disentangle the excitation from the system, allowing the speech signal to be deconstructed and described separately, as opposed to directly analyzing the characteristics of the signal waveform.

2.2 Selection of Feature Parameters

The Mel Frequency Cepstrum Coefficient (MFCC) [24] is an acoustic feature derived from the perceptual characteristics of the human ear, providing distinct advantages over other acoustic features such as the Linear Predictive Cepstrum Coefficient (LPCC), which is derived through studies of the human vocal mechanism [25, 26]. The MFCC parameters are distinguished by their performance in recognition and immunity to noise, as they employ the Euclidean distance as the measure of distance, faithfully mimicking human auditory properties without any presupposed assumptions. The MFCC parameters are computed by transforming the original frequency domain signal to the Mel frequency domain via a set of critical band filters, and subsequently to the inverse spectral domain through a discrete cosine transform. The exact sequence of frequency conversion is depicted in Fig. 2.

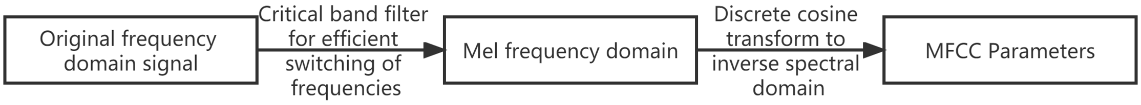

Original frequency domain signal
Critical band filter for efficient switching of frequencies
Mel frequency domain
Discrete cosine transform to inverse spectral domain
MFCC Parameters

Fig. 2. Frequency conversion flow chart

MFCC features are cepstrum parameters extracted from the Mel scale frequency domain, with their computation based on the filter bank analysis of the speech signal. The Mel scale [27] encapsulates the non-linear nature of frequency perception in the human ear, i.e., the intensity of frequency perception in the human ear is proportional to the logarithm of the frequency, while the Mel frequency depicts the relationship between speech frequency and the frequency perceived by the human ear, maintaining a linear relationship with human perception of pitch in the Mel frequency domain. If the Mel frequencies of two speech segments differ by a factor of two, the human ear perceives these two segments with a pitch difference of the same factor. The formula for conversion between Mel frequencies and linear frequencies is presented below:

$$f_{\mathrm{Mel}} = 2595 \log_{10}(1 + f / 700), \tag{2}$$

where f_{Mel} represents the Mel frequency and f signifies the linear frequency.

Fig. 3 portrays the relationship between Mel frequency and linear frequency:

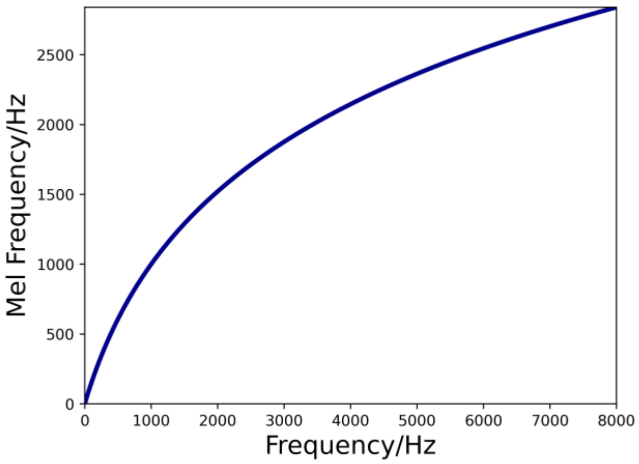


Fig. 3. Mel frequency and linear frequency relationship

When two tones of similar frequency are emitted simultaneously, humans perceive only one tone. It is only when the two frequency components differ by a certain bandwidth that humans can distinguish between them. This bandwidth is referred to as the Critical Bandwidth and is computed as:

$$BW_c = 25 + 75\left[1 + 1.4\left(f_c / 1000\right)\right]^{0.69}, \quad \textbf{(3)}$$

where BW_c stands for the critical bandwidth and is the centre frequency.

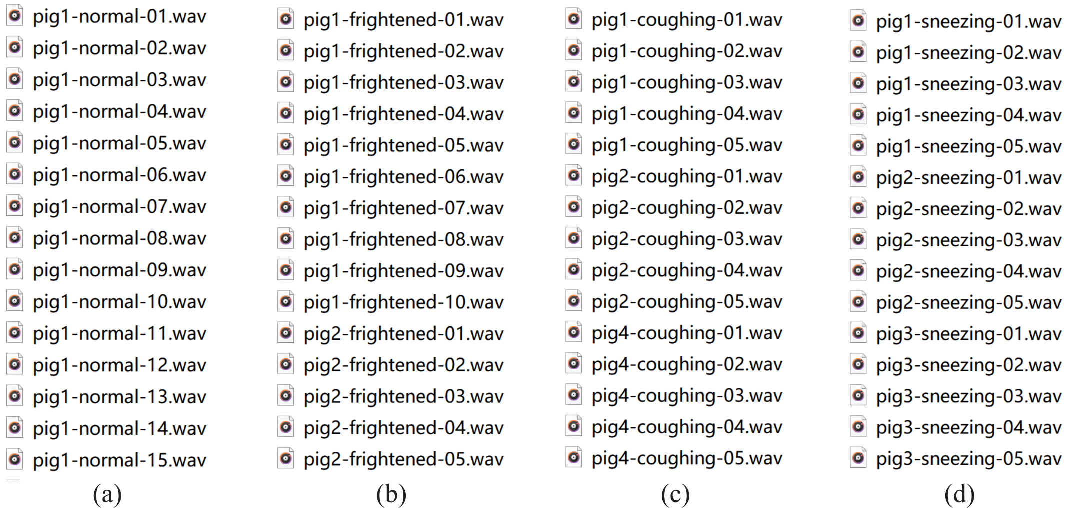
3 Data Collection and Processing

3.1 Data Collection

Data collection for this study was conducted using a microphone and recording software. The sampling frequency was set at 8kHz, and the samples were encoded using 8 bits before they were stored in a wav format on a storage device.

A total of 50 pigs were incorporated into the study. Each pig was isolated for the recording session, conducted in a quiet environment. We recorded each pig's vocalizations in four different states: normal, startled, coughing, and sneezing. The categorization of the recorded sounds was carried out using an expert-based labeling method. For each pig, we collected 20 samples of normal sounds, 10 samples of startled sounds, 5 samples of coughing sounds (only for the subset of pigs that exhibited coughing), and 5 samples of sneezing sounds.

The dataset was partitioned into training and testing sets at an 8:2 ratio. The audio files were named following the format "pignum-state-num.wav", as illustrated in Fig. 4.



pig1-normal-01.wav
pig1-normal-02.wav
pig1-normal-03.wav
pig1-normal-04.wav
pig1-normal-05.wav
pig1-normal-06.wav
pig1-normal-07.wav
pig1-normal-08.wav
pig1-normal-09.wav
pig1-normal-10.wav
pig1-normal-11.wav
pig1-normal-12.wav
pig1-normal-13.wav
pig1-normal-14.wav
pig1-normal-15.wav

(a)

pig1-frightened-01.wav
pig1-frightened-02.wav
pig1-frightened-03.wav
pig1-frightened-04.wav
pig1-frightened-05.wav
pig1-frightened-06.wav
pig1-frightened-07.wav
pig1-frightened-08.wav
pig1-frightened-09.wav
pig1-frightened-10.wav
pig2-frightened-01.wav
pig2-frightened-02.wav
pig2-frightened-03.wav
pig2-frightened-04.wav
pig2-frightened-05.wav

(b)

pig1-coughing-01.wav
pig1-coughing-02.wav
pig1-coughing-03.wav
pig1-coughing-04.wav
pig1-coughing-05.wav
pig2-coughing-01.wav
pig2-coughing-02.wav
pig2-coughing-03.wav
pig2-coughing-04.wav
pig2-coughing-05.wav
pig4-coughing-01.wav
pig4-coughing-02.wav
pig4-coughing-03.wav
pig4-coughing-04.wav
pig4-coughing-05.wav

(c)

pig1-sneezing-01.wav
pig1-sneezing-02.wav
pig1-sneezing-03.wav
pig1-sneezing-04.wav
pig1-sneezing-05.wav
pig2-sneezing-01.wav
pig2-sneezing-02.wav
pig2-sneezing-03.wav
pig2-sneezing-04.wav
pig2-sneezing-05.wav
pig3-sneezing-01.wav
pig3-sneezing-02.wav
pig3-sneezing-03.wav
pig3-sneezing-04.wav
pig3-sneezing-05.wav

(d)

Fig. 4. Schematic representation of data names

3.2 Preprocessing of Speech Signals

Preprocessing the sound signal before feature extraction significantly aids subsequent feature extraction and recognition [28]. For pig vocal signals, the preprocessing steps resemble those used for typical sounds, which include pre-emphasis, framing, windowing, and endpoint detection of the sound signal.

Pre-emphasis Treatment. As the average power spectrum of the speech signal is influenced by vocal gate excitation and muzzle radiation, the high-frequency band experiences an attenuation of approximately 6 dB/octave above roughly 800 Hz. To counter this, we perform pre-emphasis in the preprocessing stage. Pre-emphasis aims to enhance the signal-to-noise ratio in the high-frequency segment of the pig vocal signal and to flatten the spectrum for spectral or parametric analysis of the vocal tract. This is achieved with a pre-emphasis digital filter that boosts high frequencies by 6dB/octave, typically first-order.

The pre-emphasis function is calculated as follows:

$$H(z) = 1 - \alpha z^{-1}, \tag{4}$$

where H(z) signifies the transfer function in the z domain; α denotes the pre-emphasis factor, usually $0.9 < \alpha < 1.0$.

Let the sampled value of the pig vocal signal at time n be x(n), then the pre-emphasized signal expression is represented in equation (5)

$$Y(n) = x(n) - \alpha x(n-1), \tag{5}$$

where Y(n) is the pre-emphasized signal sequence; $\alpha = 0.95$ is employed in this paper.

Framing. The sound signal of a pig, being a non-smooth process signal, is often more challenging to process directly and presents difficulties in feature extraction. Therefore, we analyze each frame by dividing the speech signal into short-time frames. This approach permits us to consider the speech signal characteristics within each frame as static, or at least approximately static.

A collection of N samples into an observational unit is termed a frame. Typically, N equals 256 or 512, covering a duration between 20 and 30ms. This is due to the pig speech signal being considered static within this timeframe. To prevent introducing significant distortion at frame boundaries, some overlap between frames is often allowed, usually between 1/3 and 1/2 of a frame. Fig. 5 illustrates the principle of frame splitting.

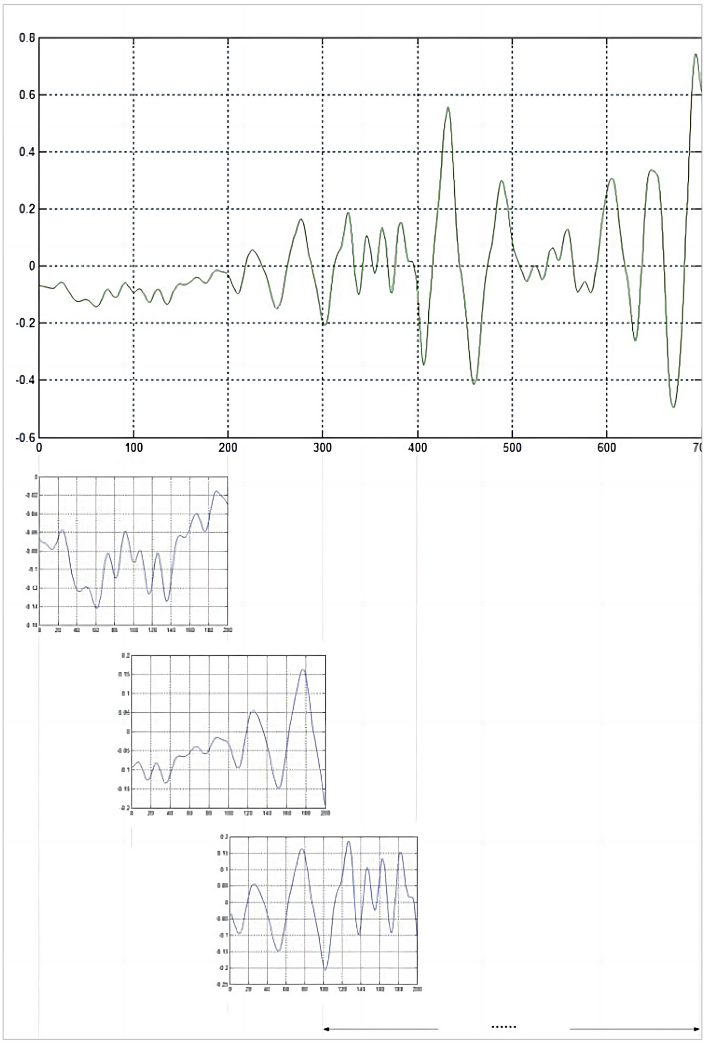


Fig. 5. Frame-splitting schematic

Windowing. Each frame is substituted into the window function, with values outside the window set to 0. This step mitigates potential signal discontinuity at both ends of each frame. The Hamming window [29], superior in suppressing side lobes and reducing spectral leakage in the time domain while demonstrating smoother amplitude characteristics in the frequency domain, is used in this study. It provides improved dynamic range and frequency resolution. Its mathematical expression is shown in equation (6), and its weighting factor enables smaller sidelobes and slower sidelobe decay. Fig. 6 compares the speech signal before and after applying the Hamming window.

$$W(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right), 0\le n\le N-1 \\ 0, \qquad\qquad\qquad\qquad\qquad others\end{cases}, \tag{6}$$

where: N is the window length and signifies the number of signal points within that window.

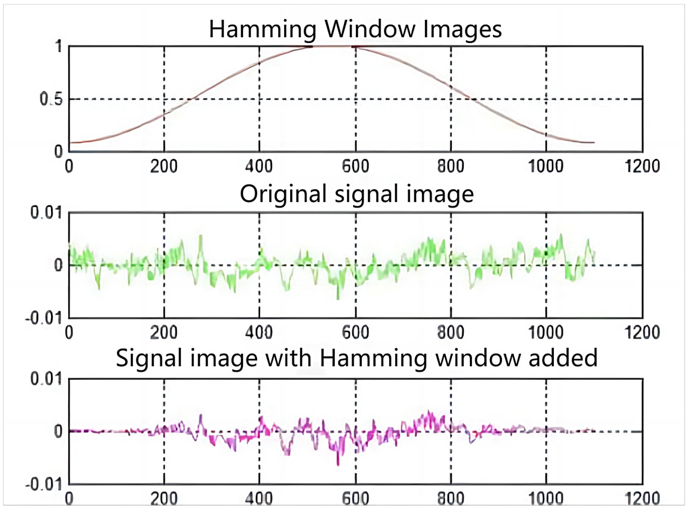


Fig. 6. Comparison before and after the addition of the Hamming window signal

End-point Detection. Despite controlling the recording to capture an audio segment each time a sound is made, the resulting pig audio signal is still replete with irrelevant segments, including ambient noise and silent segments during the pig's silence. Endpoint detection of the captured pig audio signal is then necessary to ascertain the start and end points within the audio signal, which reduces data processing requirements and suppresses silent or noisy segments. Short-time energy and short-time zero-crossing rate are chosen for double threshold endpoint detection due to their straightforward and effective computation, adaptability to different speech environments and characteristics, more precise detection of speech start and end points, and some suppression effects for non-speech noise and background sound.

Short-time energy, a crucial characteristic of sound, is the focus of sound energy analysis. It is defined as follows:

$$\boldsymbol{E}_n = \sum_{m=0}^{N-1}\left[X_{n(m)}\right]^2 . \quad \textbf{(7)}$$

Zero-crossing in a continuous sound signal implies the time-domain waveform crossing the coordinate axis, while in a discrete signal sequence, it manifests as a change in the sign of adjacent sample values. It is defined as follows:

$$Z_n = \frac{1}{2}\sum_{m=0}^{N-1}\left|\operatorname{sgn}[x_n(m)] - \operatorname{sgn}[x_{(m-1)}][X_{n(m)}]^2\right|, \quad \textbf{(8)}$$

where: x_n(m)=w(m)x(k+m), x(m) is the pig vocal signal, w(m) is the window function, N is the window length and sgn is the sign function.

3.3 Extraction of Speech Feature Parameters

The extraction of speech feature parameters can transform complex speech signals into a simplified representation that can be used for analysis and recognition [30]. By extracting speech feature parameters, the dimensionality of the speech signal can be reduced, redundant information can be removed, key acoustic features can be extracted, and the efficiency and accuracy of speech processing can be improved.

The calculation process of Mel inversion coefficients is shown in Fig. 7:

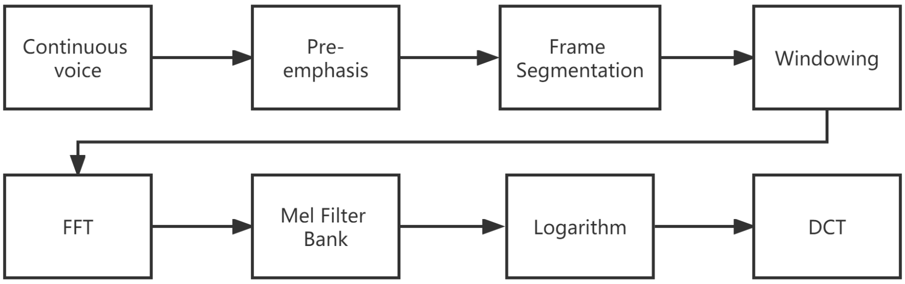


Fig. 7. Calculation process of Mel's inverse spectral coefficient

Mel Spectrum Transformation. To extract key perceptually relevant features from speech signals, Mel Frequency Cepstral Coefficients (MFCCs) are commonly used for speech feature representation. The process involves transforming the speech signal into the Mel frequency domain and applying the discrete cosine transform to obtain discriminative feature parameters. The signal is preprocessed and then subjected to fast Fourier transform to convert the time-domain signal into the frequency-domain signal. Next, Mel filter banks are applied to transform the spectral signal into the Mel frequency domain. The Mel filter banks consist of a series of triangular filters that filter the obtained discrete spectrum to generate a set of coefficients, denoted as m1, m2, ..., mp. These coefficients provide a representation of the speech signal in the Mel frequency domain and capture the perceptually significant information for subsequent analysis and classification tasks.

$$m_i = \ln\left(\sum_{k=0}^{N-1} \left|X(k) \cdot H_i(k)\right|\right), i = 1, 2, \dots, p, \tag{9}$$

where $H_i(k)$ is the response of the i-th Mel filter at frequency k.

$$H_i(k) = \begin{cases} 0, & k < f[i-1] \text{ or } k > f[i+1] \\ \dfrac{2(k - f[i-1])}{(f[i+1] - f[i-1])(f[i] - f[i-1])}, & f[i-1] \le k \le f[i]. \\ \dfrac{2(f[i+1] - k)}{(f[i+1] - f[i-1])(f[i+1] - f[i])}, & f[i] \le k \le f[i+1] \end{cases} \tag{10}$$

f[i] is the centre frequency of the triangular filter, which satisfies:

$$Mel(f[i+1]) - Mel(f[i]) = Mel(f[i]) - Mel(f[i-1]). \tag{11}$$

The Mel spectrum is transformed from a linear natural spectrum to a Mel spectrum that reflects the characteristics of human hearing.

Cepstrum Analysis Process. The main process of cepstrum analysis is shown in Fig. 8:

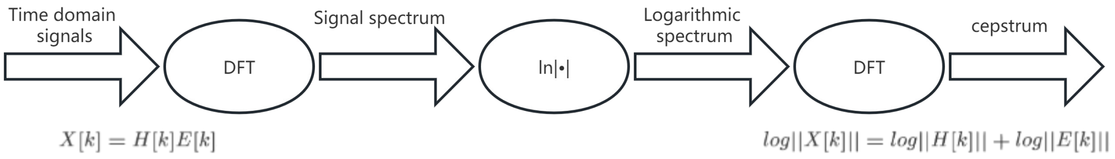


Fig. 8. Flow chart of cepstrum analysis

1) First, the original speech signal x(t) is transformed by the Fourier transform into a spectrum X[k], which can be considered as the product of the vocal tract transfer function H[k] and the excitation signal E[k]:

$$\mathrm{X[k] = H[k]E[k]}, \tag{12}$$

where H[k] represents the frequency response of the vocal tract filter, and E[k] is the frequency response of the source signal (usually vocal fold vibration).

We take the mode of X[k] to obtain |X[k]| and then take the logarithm of the result. In this step, we focus only on the amplitude information and ignore the phase information:

$$\mathrm{| X[k] |=| H[k] || E[k] |}. \tag{13}$$

2) Since human perception is proportional to the logarithm of the frequency, we take the logarithm on both sides of:

$$\log \| X[k] \| = \log \| H[k] \| + \log \| E[k] \| . \tag{14}$$

3) Finally, we perform an inverse Fourier transform on the logarithmic operation to obtain the inverse spectral sequence x[k]. x[k] can be considered as the sum of the channel characteristic h[k] and the excitation characteristic e[k]:

$$\mathrm{x[k] = h[k] + e[k]}. \tag{15}$$

The Mel frequency cepstrum coefficient MFCC can be obtained by performing an inverse spectrum analysis on top of the Mel spectrum.

Extraction of Dynamic Differential Parameters. The standard cepstrum parameter MFCC reflects mainly the static characteristics of speech parameters. In order to capture the characteristics of speech more comprehensively, we introduce the modelling of dynamic characteristics [31]. This is achieved by performing differential spectral analysis on static features to produce dynamic features. Our model therefore combines both static and dynamic features as a way to improve the performance of speech recognition. The equation for calculating the differential parameters is shown below:

$$dt = \begin{cases} C(t+1) - C(t), t < K \\ \dfrac{\sum_{k=1}^{K} k\left(c_{t+k} - c_{t-k}\right)}{\sqrt{2\sum_{k=1}^{K} k^2}}, \text{Others} \\ C(t1) - C(t-1), t \geq Q - K \end{cases}, \tag{16}$$

where dt denotes the tth first-order difference, Ct denotes the tth inverse spectral coefficient, Q denotes the order of the inverse spectral coefficient, and K denotes the time difference of the first-order derivative, which can be taken as 1 or 2. The result of the above equation is then substituted to obtain the parameters of the second order difference.

The composition of the MFCC is shown in Fig. 9.

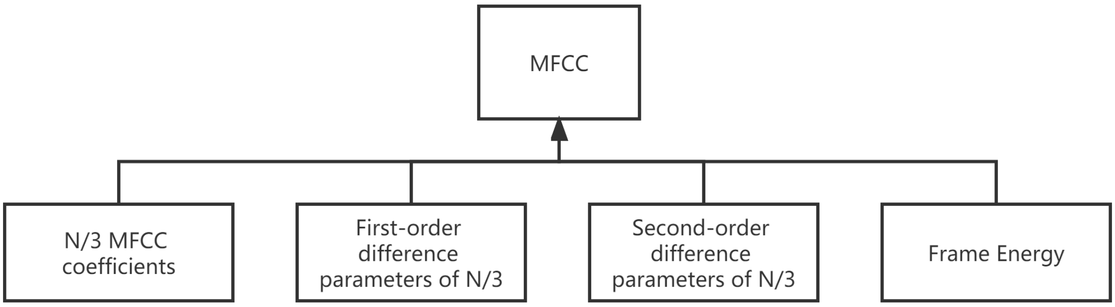

MFCC
N/3 MFCC coefficients
First-order difference parameters of N/3
Second-order difference parameters of N/3
Frame Energy

Fig. 9. MFCC complete composition

As can be seen from the above diagram, the coefficients of MFCC consist of four components [32]: MFCC coefficients, first order differential parameters (Delta MFCC), second order differential parameters (Delta-Delta MFCC) and frame energy. The MFCC coefficients are the main part of the MFCC and are obtained by filtering the power spectrum of the audio signal through a Mel filter bank and a logarithmic operation, followed by a discrete cosine transform (DCT). MFCC coefficients effectively represent the spectral shape of speech signals, capturing the main characteristics of the speech. The first-order differential parameters are the first-order temporal derivatives of the MFCC coefficients and reflect the dynamic variation characteristics of the speech signal, thereby enhancing the performance of speech recognition. The second-order differential parameters are the second-order temporal derivatives of the MFCC coefficients, which better reflect the acceleration changes in the speech signal, further improving the performance of speech recognition. Frame energy refers to the energy value of each frame of the audio signal, reflecting the intensity variation of the speech. In speech recognition, frame energy is commonly used to detect the beginning and end of speech segments, as well as to distinguish speech from noise.

4 Voiceprint Recognition Method Based on ResNet-LSTM

4.1 ResNet

Deep residual networks [33] generally consist of multiple residual blocks, of which the standard residual block is shown in Fig. 10, usually consisting of a stack of convolutional (Conv), batch normalisation (BN) and non-linear activation (ReLU) layers. In normal neural network training, increasing the depth of the network often means that the network fits the data better, but it can also introduce problems of overfitting and gradient disappearance or gradient explosion, which can lead to a reduction in the accuracy of the network or even the loss of its ability to detect the test set. Residual neural networks can solve this problem through their jump connection mechanism, where the original input feature information is added to the output information as the input to the next neuron.

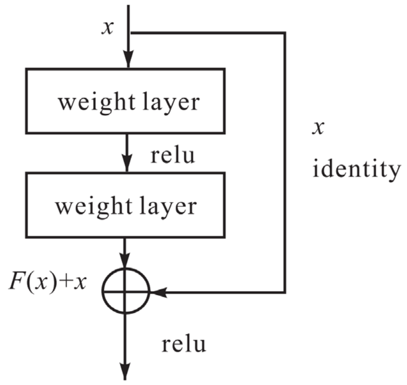


Fig. 10. Residual fast structure diagram

4.2 Long Short Term Memory Network (LSTM)

The LSTM introduces the concept of gates and has a more complex hidden cell structure [34], where the hidden cell typically consists of an input gate i, an oblivion gate f and an output gate o. The storage and updating of information in the LSTM is achieved by the gating part [35], which can be seen as a fully connected layer containing Sigmoid activation functions and dot product operations. The gating operation can be formulated as:

$$g(x) = \sigma(Wx + b), \quad \textbf{(17)}$$

where σ(x)=1/(1+exp(-x) is the Sigmoid activation function, one of the common non-linear activation functions in deep learning. The Sigmoid activation function in the LSTM is used to describe the proportion of the message that passes, when the output of the gate is 0, it means that no data passes, when the output is 1, it means that all data passes [36].

4.3 ResNet-LSTM

Table 1. Structure of ResNet-LSTM network

Layer	Structure	Stride
Conv	3×3, 64	2×2
ResNet block	$\begin{bmatrix}1\times1,\ 64\\ 3\times3,\ 64\\ 1\times1,\ 256\end{bmatrix}\times 2$	—
Conv	3×3, 128	2×2
ResNet block	$\begin{bmatrix}1\times1,\ 128\\ 3\times3,\ 128\\ 1\times1,\ 512\end{bmatrix}\times 2$	—
LSTM (512)	—	—
LSTM (512)	—	—
Average Pooling	512	—
Dense	512×512	—
Dropout (0.2)	—	—
Dense (Softmax)	512×4	—

In this paper, the ResNet part and the LSTM part are fused to form a ResNet-LSTM and text-independent voice recognition method, in which the ResNet part and the LSTM part are used to extract the spatial and temporal features of voice patterns respectively, combining the advantages of ResNet and LSTM. The ResNet part consists of two convolutional layers and four standard residual blocks, and the LSTM part consists of two LSTM layers. The detailed network structure is shown in Table 1.

5 Experimental Results and Analysis

5.1 Experimental Environment

This experimental simulation environment and configuration is as follows: operating system is Windows 10-64 bit, running memory is 16G, graphics card is GTX 1060, development language is Python 3.7.16, development environment is anaconda virtual environment. The ResNet runtime environment including PyTorch 1.8.1+cu111, CUDA 11.1, and cuDNN8005 was built on this basis.

The specific parameters set for the model in the experiments are shown in Table 2.

Table 2. Specific parameter settings

Parameters	Selection
Optimizer	Adam
Learning Rate	0.001
L2 Weight Decay	0.0001
Batch Size	64

5.2 Data Pre-processing

The following text utilizes the audio file pig7-normal-03.wav as an illustrative case to demonstrate the process of pre-processing a dataset. This approach is presented sequentially, detailing each operation and its effects on the audio data.

Initially, the audio file is read and its corresponding waveform graph is generated, as shown in Fig. 11.

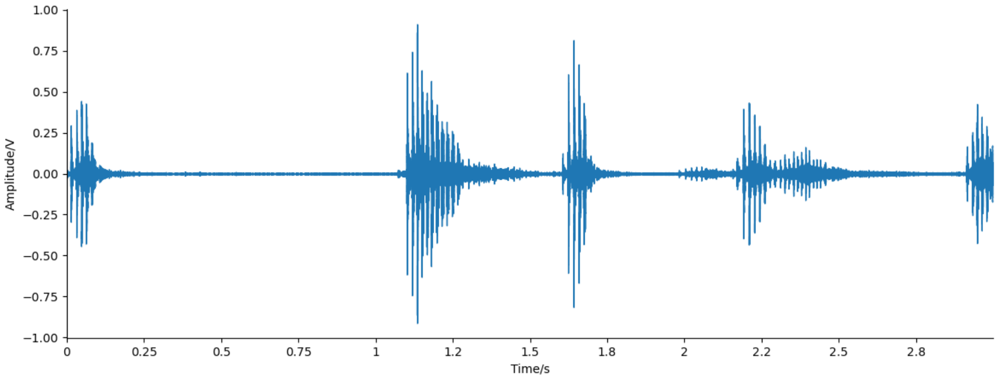


Fig. 11. Waveform diagram

Following this initial step, a pre-emphasis operation is applied. This operation accentuates the high-frequency components of the signal, resulting in a waveform that appears more “sharp”. The intensified visibility of the "sharp" peaks of the waveform is attributed to the increased representation of high-frequency signals that correspond to rapid changes in the waveform, in contrast to the more gradual changes associated with low-frequency signals. The results of pre-emphasis operation can be seen in Fig. 12.

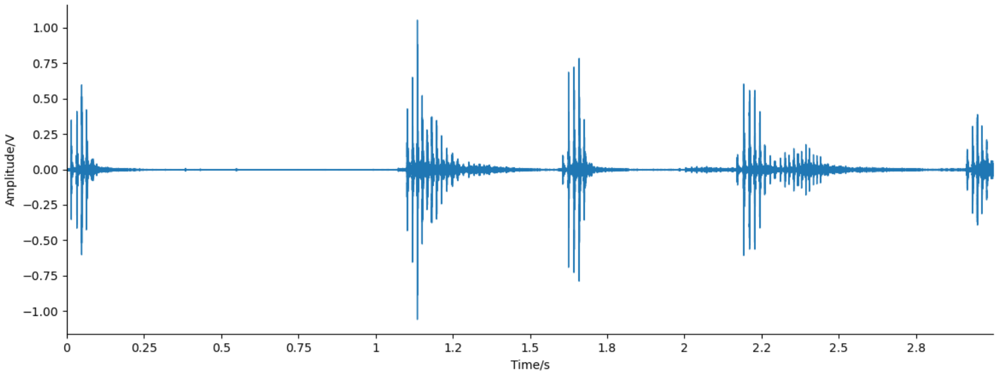


Fig. 12. Pre-weighting

After pre-emphasis, a framing operation is executed on the audio signal. This process divides the continuous audio signal into a series of brief frames, each containing a discrete segment of the audio data. Fig. 13 depicts the waveform of the first frame post-framing operation.

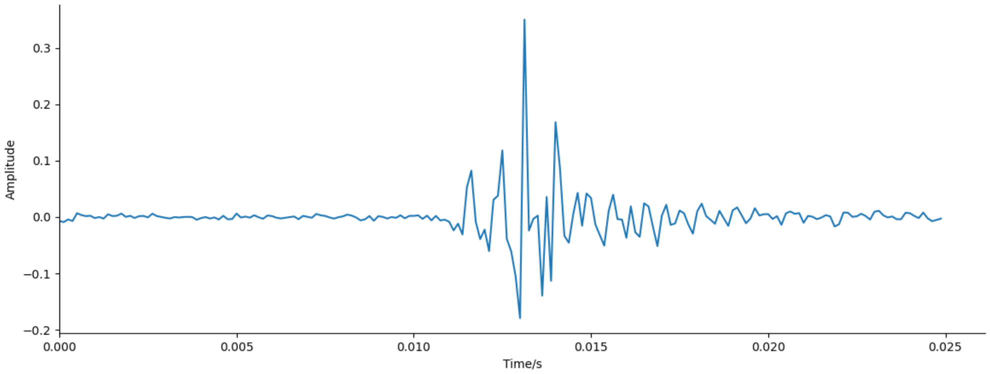


Fig. 13. First frame after frame-splitting

The application of the framing operation reveals that the waveform edges are somewhat rough at this stage. This observation justifies the subsequent windowing operation. The application of the window smoothens the amplitude fluctuations at the beginning and end of the signal, which can be clearly seen in Fig. 14. The result is a signal with smoothed edges, and a gradual transition to zero amplitude at both ends.

This step-by-step demonstration illustrates the transformation of the audio signal throughout the pre-processing operation. Each operation contributes to the overall objective of preparing the audio signal for subsequent analysis or processing tasks.

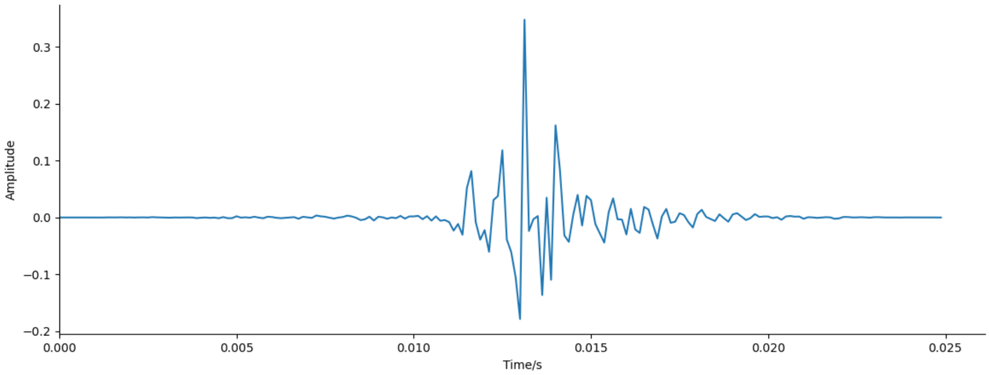


Fig. 14. The first frame after adding the window

5.3 Feature Extraction

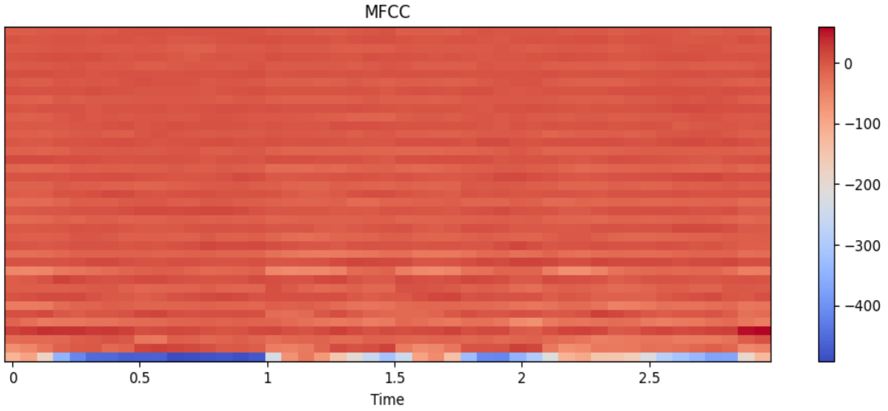
Upon completing the pre-processing of the audio data, the subsequent phase of feature extraction ensues. The key feature being extracted in this case are Mel Frequency Cepstral Coefficients (MFCCs), a commonly utilized representation for audio and speech signals.

The MFCCs are data in their own right, manifesting as vectors. Each of these feature vectors encapsulates the spectral properties of a particular audio frame. They present a comprehensive representation of the audio signal by capturing the salient characteristics while reducing dimensionality and retaining essential information.

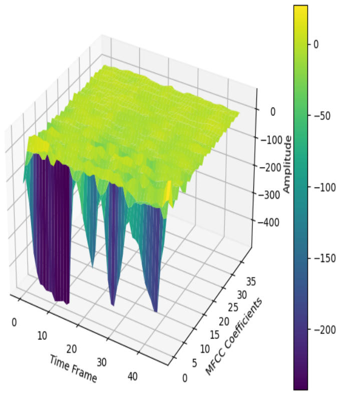
To better understand these feature vectors, visualization proves helpful. This can be achieved by plotting the MFCCs in two dimensions. Fig. 15(a) provides such a visual representation in a 2D plan view. This graph elucidates the relationship between each of the MFCCs across the different audio frames, providing a bird's-eye view of the spectral properties of the entire audio signal.

Alternatively, a surface view (or 3D plot) can offer an enhanced understanding of the distribution and interaction of the MFCCs. Fig. 15(b) exhibits this surface view, providing a tangible perspective of the topographical distribution of the MFCCs across various audio frames. This perspective can offer invaluable insights into the spectral dynamics of the audio signal over time.

In essence, both these visualizations, (a) the 2D plan view and (b) the surface view, provide insightful explorations of the MFCC features extracted from the pre-processed audio data. They serve as beneficial tools for understanding and interpreting the inherent spectral characteristics and temporal dynamics of the audio signal.



(a) The 2D plan view



(b) The surface view

Fig. 15. 39-dimensional MFCC feature map

5.4 Experimental Results

In this paper, we compare four different models: BP (back propagation) neural networks, LSTM (long and short-term memory) neural networks, ResNet (residual networks) and a hybrid model combining ResNet and LSTM.

Based on the long short-term memory (LSTM) model, we achieved an accuracy of 96.1%, thanks to its remarkable time-dependent capture capability in continuous Mel frequency cepstral coefficient (MFCC) data. This result significantly outperforms the performance of the back propagation (BP) model, which has an accuracy of only 43.7%, reflecting the limitations of traditional neural networks in dealing with serial data. The Residual Network (ResNet) model, which is stronger in dealing with spatial correlation within the data, achieved an accuracy of 89.7%.The complex design of the ResNet model confirms its excellent ability to analyse the complex features of MFCC by introducing skip connections to mitigate the problem of gradient disappearance.

When we fused the respective strengths of LSTM and ResNet into a hybrid model, the performance of the model was further improved, with an accuracy rate of 98.3%. This result emphasises the importance of fully considering temporal and spatial information in the learning process of complex audio signal classification tasks. By combining the sequence processing capability of LSTM with the spatial feature extraction capability of ResNet, the ResNet-LSTM model excels in identifying different vocal states of pigs, further revealing its potential application in animal welfare monitoring and precision animal husbandry. In future studies, we will explore the generality of this hybrid model for identifying vocal states in a wider range of animal species and different environmental contexts.

The comparative results are shown in Table 3.

Table 3. Comparison of results from different models

Models	Accuracy/%
BP	43.7
LSTM	96.1
ResNet	89.7
ResNet-LSTM	98.3

6 Conclusion

This paper presents a comprehensive methodology for voiceprint recognition, utilizing audio recordings from pigs in differing states. The data collection process involves recording audio samples from pigs under varying conditions, which are then classified based on expert evaluation. These recorded samples undergo a series of preprocessing steps, encompassing pre-emphasis, framing, windowing, and endpoint detection. The highly significant Mel Frequency Cepstral Coefficients (MFCC) are then extracted from these preprocessed samples, providing a robust representation of voiceprint features. Beyond static features, we incorporate dynamic differential parameters to capture the dynamic aspects of the voice signals. This is achieved through the analysis of the differential spectra derived from the static features. In the classification phase, we implement a model founded on the Residual Network (ResNet) and Long Short-Term Memory (LSTM) architectures to monitor the state of pigs. Our experimental results corroborate the efficacy of the proposed algorithm in correctly identifying the states of pigs with a high degree of accuracy. As we look to the future, our research endeavors will concentrate on exploring the integration of this technique with other models to amplify the performance of the voiceprint recognition method even further.

References

[1] X.-X. Zhou, Z.-R. Gao, X.-T. Yi, An improved chicken swarm optimization algorithm based on adaptive mutation learning strategy, Journal of Computers 33(6)(2022) 1-19.

[2] W.-J. Wang, Research and application of pig sound state recognition based on CNN, [dissertation] Harbin: Harbin Engineering University, 2019.

[3] M. Silva, V. Exadaktylos, S. Ferrari, M. Guarino, J.-M Aerts, D. Berckmans, The influence of respiratory disease on the energy envelope dynamics of pig cough sounds, Computers and Electronics in Agriculture 69(1)(2009) 80-85.

[4] S. Hua, K.-Y. Han, Z.-F. Xu, M.-J. Xu, H.-B. Ye, C.-Q. Zhou, Image processing technology based on internet of things in intelligent pig breeding, Mathematical Problems in Engineering 2021(2021) 1-9.

[5] Y. Gu, A. Shi, R. Ma, Voiceprint recognition based on big data and Gaussian mixture model, in: Proc. 2021 6th International Conference on Smart Grid and Electrical Automation (ICSGEA), 2021.

[6] K. Luo, L. Fu, Research and application of voiceprint recognition based on a deep recurrent neural network, in: Proc. 2018 Automatic Control, Mechatronics and Industrial Engineering, 2019.

[7] Y. Wu, Z. Wu, H. Yang, A fingerprint and voiceprint fusion identity authentication method, in: Proc. 2019 Cyberspace Safety and Security: 11th International Symposium, 2019.

[8] S. Cheng, Y. Shen, D. Wang, Target speaker extraction by fusing voiceprint features, Applied Sciences 12(16)(2022) 8152.

[9] P. Liu, S.-B. Li, J.-G. Tang, An end-to-end macaque voiceprint verification method based on channel fusion mechanism, in Proc. 2022 Interspeech, 2022.

[10] P. Schlegel, K. Wong, M. Aker, Y. Alhiyari, J. Long, Objective assessment of porcine voice acoustics for laryngeal surgical modeling, Applied Sciences 11(10)(2021) 4489.

[11] A.-T. Kavlak, M. Pastell, P. Uimari, Disease detection in pigs based on feeding behaviour traits using machine learning, Biosystems Engineering 226(2023) 132-143.

[12] Y.-L. Yin, N. Ji, X.-P. Wang, W.-Z. Shen, B.-S. Dai, S.-L. Kou, C. Liang, An investigation of fusion strategies for boosting pig cough sound recognition, Computers and Electronics in Agriculture 205(2023) 107645.

[13] W.-Z. Shen, N. Ji, Y.-L. Yin, B.-S. Dai, D. Tu, B.-H. Sun, H.-D. Hou, S.-L. Kou, Y.-Z. Zhao, Fusion of acoustic and deep features for pig cough sound recognition, Computers and Electronics in Agriculture 197(2022) 106994.

[14] J. Zhao, X. Li, W.-H. Liu, Y. Gao, M.-G. Lei, H.-Q. Tan, D. Yang, DNN-HMM based acoustic model for continuous pig cough sound recognition, International Journal of Agricultural and Biological Engineering 13(3)(2020) 186-193.

[15] N. Ji, W.-Z. Shen, Y.-L. Yin, J. Bao, B.-S. Dai, H.-D. Hou, S.-L Kou, Y.-Z. Zhao, Investigation of acoustic and visual features for pig cough classification, Biosystems Engineering 219(2022) 281-293.

[16] H.-N. Sun, Study on extraction and recognition method of characteristic parameters of hogs' cough, [dissertation] Daqing: Heilongjiang Bayi Agricultural University, 2022.

[17] N. Mansouri, Z. Lachiri, Human laughter generation using hybrid generative models, KSII Transactions on Internet & Information Systems 15(5)(2021) 1590-1609.

[18] P.-P. Lu, Q. Li, H. Zhu, G. Sovernigo, X.-D. Lin, Voxstructor: Voice reconstruction from voiceprint, in: Proc. 2021 Information Security: 24th International Conference, 2021.

[19] C.-F. Cao, H.-R. Liang, G.-C. Liang, Research on information technology with voiceprint authentication algorithm and its implementation on mobile terminal, in: Proc. 2014 Advanced Materials Research, 2014.

[20] M. Skaf, M. Salah, M. Shalodi, Voiceprint authentication system, [bachelor thesis] Hebron: Palestine Polytechnic University, 2021.

[21] B. Gu, Deep speaker embedding with frame-constrained training strategy for speaker verification, in: Proc. 2022 Interspeech, 2022.

[22] J.-Y. Li, J.-M. Zhang, A study of voice print recognition technology, in: Proc. 2021 International Wireless Communications and Mobile Computing (IWCMC), 2021.

[23] M.-N. Kakade, D.-B. Salunke, An automatic real time speech-speaker recognition system: a real time approach, in: Proc. ICCCE 2019: Proceedings of the 2nd International Conference on Communications and Cyber Physical Engineering. Springer Singapore, 2019.

[24] Z.-K. Abdul, A.-K. Al-Talabani, Mel frequency cepstral coefficient and its applications: A review, IEEE Access 10(2022) 122136-122158.

[25] H.-Y. Yang, Y.-R. Deng, H.-A. Zhao, A comparison of MFCC and LPCC with deep learning for speaker recognition, in: Proc. 2019 Proceedings of the 4th International Conference on Big Data and Computing, 2019.

[26] I.-S. Areni, A. Bustamin, Improvement in speech to text for bahasa Indonesia through homophone impairment training, Journal of Computers 28(5)(2017) 1-10.

[27] Y.-J. Zhao, B.-F. Qin, Y.-H. Zhou, X.-Z. Xu, Bearing fault diagnosis based on inverted Mel-scale frequency cepstral coefficients and deformable convolution networks, Measurement Science and Technology 34(5)(2023) 055404.

[28] Y.-M. Shi, An improved machine learning model for pig abnormal voice recognition, Journal of Computers 33(6)(2022) 155-166.

[29] Q.-Y. Zhang, G.-L. Li, S.-B. Qiao, A retrieval algorithm of encrypted speech based on biological hashing, Journal of Computers 31(4)(2020) 126-140.

[30] G. Korvel, O. Kurasova, B. Kostek, Comparative analysis of spectral and cepstral feature extraction techniques for phoneme modelling, in: Proc. 2018 Multimedia and Network Information Systems: Proceedings of the 11th International Conference MISSI 2018, 2018.

[31] Q.-R. Chen, Z.-F. Wu, Q.-H. Zhong, Z.-W. Li, Heart sound classification based on mel-frequency cepstrum coefficient features and multi-scale residual recurrent neural networks, Journal of Nanoelectronics and Optoelectronics 17(8)(2022) 1144-1153.

[32] A.-R. Verma, S.-P. Singh, R.-C. Mishra, K.Katta, Performance analysis of speaker identification using Gaussian mixture model and support vector machine, in: Proc. 2019 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2019.

[33] Y.-X. Guo, W. Yang, Q. Liu, Y. Wang, Review of residual networks, Computer Application Research 37(5)(2020) 1292-1297. DOI: 10.19734/j.issn.1001-3695.2018.12.0922

[34] H. Yan, M. Xie, Y. Chen, X.-L. Li, Y.-Y. Dong, A new voiceprint recognition algorithm based on word embedding LSTM network, in: Proc. 2020 Basic & Clinical Pharmacology & Toxicology, 2020.

[35] Y.-A. Wubet, K.-Y. Lian, Voice conversion based augmentation and a hybrid CNN-LSTM model for improving speaker-independent keyword recognition on limited datasets, IEEE Access 10(2022) 89170-89180.

[36] L. Yang, Y.-X. Wu, J.-L. Wang, Y.-L. Liu, Review of research on recurrent neural networks, Journal of Computer Applications 26(2018) 1-6.