DNN-HMM based acoustic model for continuous pig cough sound recognition

Jian Zhao[1], Xuan Li[1*], Wanghong Liu[2], Yun Gao[1], Minggang Lei[2], Hequn Tan[1], Di Yang[1]

(1. *College of Engineering, Huazhong Agricultural University, Wuhan 430070, China*;
2. *College of Animal Science, Huazhong Agricultural University, Wuhan 430070, China*)

Abstract: To detect the respiratory disease through pig cough sound in the early stage, a novel method based on Deep Neural Networks-Hidden Markov Model (DNN-HMM) was proposed to construct an acoustic model for continuous pig cough sound recognition. Noises in the continuous pig sounds were eliminated by the Wiener algorithm based on wavelet thresholding the multitaper spectrum, and the experimental corpus was obtained from the denoised continuous pig sounds. The 39-dimensional Mel Frequency Cepstral Coefficients (MFCC) extracted from the corpus were considered as feature vectors. Sounds in pig farms were divided into pig coughs, non-pig coughs, and silence segments. In the HMM, the number of hidden states of pig cough, non-pig cough and silence segments were 5, 5 and 3 respectively, and the observation states represented the feature vectors of the continuous pig sound signal. Based on experiments and empirical theory, the DNN model with 3 hidden layers and 100 nodes per layer was used to describe the correspondence between hidden states and observation serials. Through experiments, the context frames of DNN input were set to 5. Under the condition of optimal parameter setting, the traditional acoustic model Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) was compared with DNN-HMM through a 5-fold cross-validation experiment. It was found that the Word Error Rate (WER) of each group in DNN-HMM was lower than that in GMM-HMM, and the average WER was 3.45% lower. At the same time, the best result of the DNN-HMM model was obtained with the lowest WER of 7.54%, and the average WER was 8.03%. The results showed that the method of DNN-HMM based acoustic model for continuous pig cough sound recognition was stable and reliable.

Keywords: DNN-HMM, acoustic model, continuous pig coughs, recognition, pig industry
DOI: 10.25165/j.ijabe.20201303.4530

Citation: Zhao J, Li X, Liu W H, Gao Y, Lei M G, Tan H Q, et al. DNN-HMM based acoustic model for continuous pig cough sound recognition. Int J Agric & Biol Eng, 2020; 13(3): 186–193.

1 Introduction

The global demand for meat products is expected to increase steadily[1-3], while pork accounts for a huge part of the meat products. At present, there is a growing trend towards large-scale, intensive, and standardized breeding in the global pig industry. However, pig respiratory diseases have become one of the most common and harmful diseases in pig farms. Even though there are numerous kinds of pig respiratory diseases, pig cough sounds are the most dominant symptoms, containing vibration information when pig's throat or airway is stimulated. Thus, the monitoring of pig cough sound could be used to build an intelligent alarm system for the early detection of pig respiratory diseases[4-7]. The method for pig cough detection used nowadays is the simple manual detection, which not only has high labor costs but also can't guarantee the recognition results. With the rapid development of modern information, digital signal processing, sensing techniques and other technology, the combination of computer science technology and speech characteristics analysis technology is proved to be more efficient. Thus, applying the automatic speech recognition method to the pig cough sound recognition field is beneficial to the pig industry.

By studying different characteristics of pig cough sound, Mitchell et al.[8] studied the cough sounds of Belgian Landrace× Duroc hybrid pigs, and found that the duration of cough sounds in sick and healthy pigs was 0.3 s and 0.21 s, respectively. Sara et al.[9] studied the cough sounds of Landrace×Large White hybrid pigs and found that the duration of cough sounds in sick and healthy pigs was 0.67 s and 0.43 s, respectively, at the same time, the Root Mean Square and peak frequency of the cough sounds of sick pigs were lower than those of healthy pigs. In the early study, the methods of pig cough sound recognition almost relied on isolated-word speech recognition methods. Moshou et al.[10] used a neural network as a classifier to distinguish pig cough sounds from other sounds like grunts, metal clanging and noise. The correct recognition rates of the 4 kinds of sounds were desirable, around 90.00%, but the experiments were carried out in laboratory conditions. Van Hirtum and Berckmans[11,12] used a fuzzy c-means clustering algorithm to recognize pig cough sounds in laboratory conditions as well. Although their results were as high as 92%, the overall error rate reached 21%. Further study was done by Guarino et al.[13] who applied the dynamic time warping

Received data: 2018-07-17 **Accepted data:** 2020-03-13

Biographies: Jian Zhao, Master, research interest: intelligent detection and control technology, deep learning, Email: 1597254919@qq.com; **Wanghong Liu,** MD, Associate Professor, research interest: animal genetic and reproduction, Email: whliu@mail.hzau.edu.cn; **Yun Gao,** PhD, Associate Professor, research interest: environmental control confinement livestock housing, artificial intelligent in livestock production, Email: angelclouder@mail.hzau.edu.cn; **Minggang Lei,** PhD, Professor, research interest: animal genetics and breeding, pig healthy production, Email: leimg@mail.hzau.edu.cn; **Hequn Tan,** Professor, research interest: livestock and poultry health farming facilities and equipment, Email: thq@mail.hzau.edu.cn; **Di Yang,** Master, research interest: intelligent detection and control technology, Email: 1055467933@qq.com.

***Corresponding author: Xuan Li,** PhD, Associate Professor, research interest: livestock and poultry information intelligent perception and intelligent equipment. Mailing address: College of Engineering, Huazhong Agricultural University, Wuhan 430070, China. Tel: +86-27-87282120, Email: lx@mail.hzau.edu.cn.

algorithm to conduct pig cough sounds recognition in pig farm conditions with the 85.5% of pig cough sound and 86.6% non-cough sounds being correctly identified. Liu Zhenyu et al.[14] realized the recognition of pig cough sound based on HMM, and the recognition rate reached 80.0%.

Continuous speech recognition technology is more practical and efficient than the isolated word recognition method. An acoustic model can describe the physical changes of the speech and the appropriate choice of acoustic modeling unit is very crucial for continuous speech recognition. As the most common acoustic models, the HMM can deal with sequences of variable length. In the HMM, by setting the causes of the speech signal as hidden states and the characteristics of the speech signal as observation states. The speech signal can be described through the transitions of hidden states. The wide use of HMM in human speech segmentation and classification builds its foundation role in modern speech recognition technology[15,16]. More and more scholars apply the HMM acoustic model to animal speech recognition. In 2012, Milone et al.[17] proposed an acoustic model based on HMM for continuous cattle ingestive sounds recognition, which divided the continuous cattle ingestive sounds into three syllables: “chews”, “bites” and “chewbites”. These three syllables were used as the HMM acoustic modeling units. The HMM hidden states were the components of ingestive events and HMM observation states were used to describe the spectral characteristics of continuous cattle sounds. More similar applications have been done in other kinds of animal sound recognition field. Based on HMM, Reby, et al.[18] proposed a method for the recognition of the bouts in red deer. Milone et al.[19] proposed the sheep ingestive sounds recognition and Trifa et al.[20] proposed the antbird sounds recognition.

At present, there is no reported research of continuous pig sound recognition based on the acoustic model. In this paper, the HMM[21,22] is used to construct the acoustic model of continuous pig sound in farm conditions. Since the pig cough is the subject of study, other sounds in the pig houses can all be classified as non-coughs. Therefore, pig cough and non-cough constitute the HMM acoustic modeling units. The factors affecting pig cough and non-cough are set as hidden states of HMM, and feature vectors of continuous pig sound as observation states of HMM. DNN[23-25] is powerful in functional expression and has a strong ability to learn the essences of the high dimensional feature vectors. Its deep structure helps to extract more abstract and more discriminative characteristics[26,27]. DNN can be used to model feature vectors of continuous pig sound and to describe the corresponding relation between HMM observation states and hidden states. In this paper, we try to take the feature vectors of the continuous pig sounds as inputs of DNN and take the probability distributions of hidden states of HMM as its outputs. Thus, we propose a new method for continuous pig cough recognition based on DNN-HMM acoustic model[28-30].

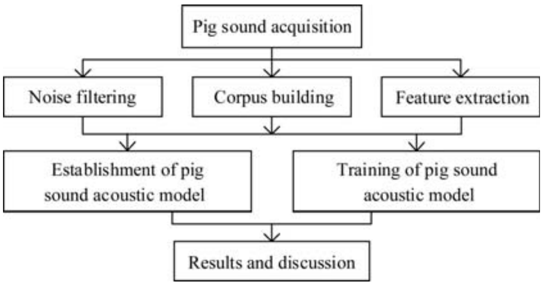


Figure 1 Flowchart of the structure of this paper

The rest of the paper is organized in Figure 1. The second part includes pig sound acquisition, noise filtering, corpus building and feature extraction. The third part introduces the establishment and training of a continuous pig sound acoustic model. The experimental results and discussion are presented in the fourth part, followed by the conclusions in the fifth part.

2 Corpus building and feature extraction

2.1 Pig sound acquisition

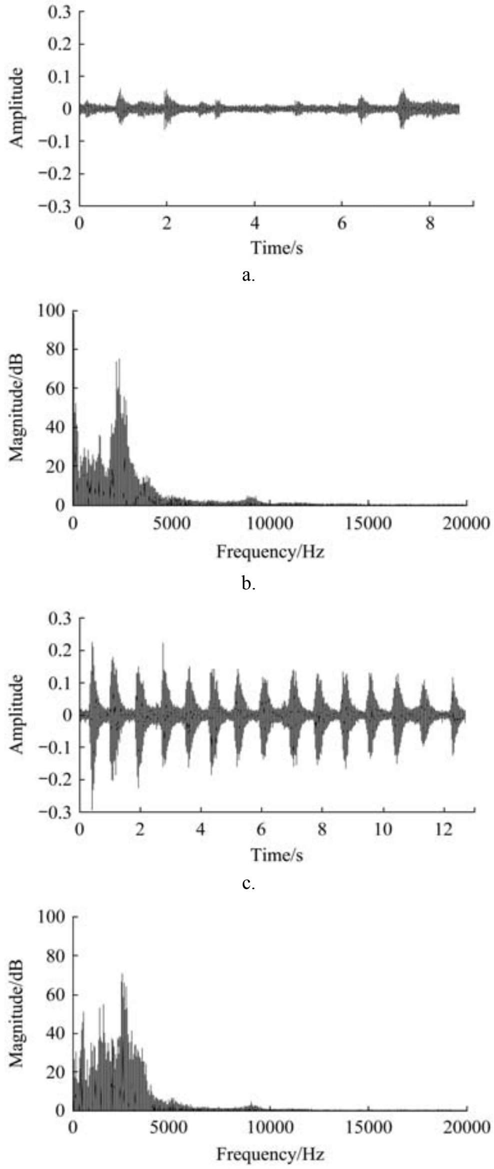
Pig sound acquisition was conducted at the quality pig farm of Huazhong Agricultural University using a recording pen (China, Mrobo M66) with a sampling frequency of 48000 Hz and the continuous working time of 24 h. The acquisition experiment was carried out in March and April, the epidemic phase of pig respiratory diseases. The pig sounds were registered from 10 Landrace pigs, 5 of which have a respiratory disease with apparent coughs diagnosed by veterinarians. And the sounds include pig sounds of cough, sneeze, eating, scream, hum, shaking ears, and sounds of dogs, metal clanging and some other background noise. Then the recorded signals with frequent pig coughs for 30 h were selected for experiments.

2.2 Pig sound preprocessing

Due to the complex environmental conditions in pig farms, the suitable filtering algorithm is vital for the recognition of pig cough sounds. Figure 2a is a farm noise time-signal with duration of 8.68 s and Figure 2b shows its frequency content, while Figure 2c is a continuous pig cough sound time-signal with a length of 12.71 s and Figure 2d shows its frequency content. From Figure 2b and Figure 2d, it can be seen that the frequency band of farm noise mainly concentrates below 5000 Hz and overlaps with the frequency band of pig cough sounds (300-8000 Hz), which means the traditional digital filter (low-pass, high-pass or band pass) is not effective in denoising pig sounds. In this paper, the speech enhancement algorithm is used to eliminate the noise in continuous pig sounds.

The Wiener algorithm based on wavelet thresholding the multi-taper spectrum proposed in the paper of Hu et al.[31] was used to enhance the speech signals of pig sound. The wavelet transform can get multi-scale subdivisions and focus on any details through scaling and translation operations. And the threshold of the suppression noise level is obtained according to the difference between the speech and the noise. The multi-taper spectrum is a nonparametric spectral estimation method that applies many mutually orthogonal windows to the estimated sequence of the pig sound sample and then obtains the average frequency spectrum. The algorithm can be summarized in five steps: (1) Use multi-taper spectrum to calculate the multi-taper power spectrum of original pig sounds; (2) Smooth multi-taper power spectrum of original pig sounds based on wavelet thresholding; (3) Calculate noise power spectrum based on the power spectrum of silence segment of the original pig sounds; (4) Calculate the ratio of the power spectrum of original pig sound to the noise power spectrum to obtain the a priori Signal to Noise Ratio (SNR); (5) Using the prior SNR to obtain the transfer function of the Wiener filter, and with the help of the transfer function, the denoised pig sound signals were obtained.

For easier processing, the 30-hour original pig sound signal was divided into 360 segments with 5 min each denoised individually. The following Figure 3 shows a denoised time-signal of the continuous pig cough sounds from Figure 2c processed by the Wiener algorithm based on wavelet thresholding the multitaper spectrum. Compared with the time-signal before and after speech enhancement, it can be seen that the noise in the pig sound signal has been significantly reduced after the enhancement even with almost no distortion.



a.

b.

c.

d.

Figure 2 Pig farm environment noise analysis

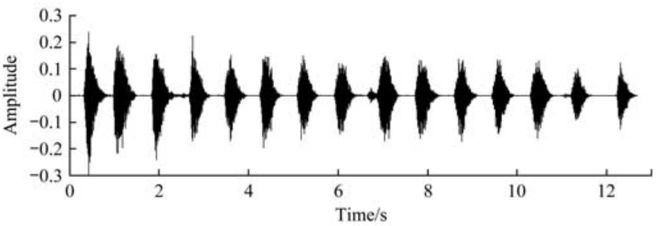


Figure 3 Time-signal of enhanced continuous pig cough sounds in Figure 2c

2.3 Selection of experimental corpora

Even after denoising, there are still so many different kinds of sounds in the continuous pig signal, posing a difficulty for recognizing pig cough sound. Subsequently, based on previous studies on the characteristics of the pig cough sound[8,9], this paper study the duration and energy of individual pig cough sound in the experimental corpus to get the threshold ranges. The threshold ranges were then set according to these two characteristics, and if the speech samples from the denoised continuous speech were not within this range, they would be removed. After this step, we got the final experimental corpus. The following describes the endpoint detection algorithm for detecting speech samples in denoised continuous speech, and then introduces the method of threshold selection.

Pig sound signal endpoint detection means finding the start and end frame of all speech samples from a signal containing pig sounds. As the noise in the 360-segment pig sound was significantly reduced after the speech enhancement, we performed the detection of speech samples by the single-parameter double-threshold endpoint detection method based on short-time energy. For the pig sound segment $y(n)$, the vth frame is expressed as $y_v(n)$ after framing, and the short-time energy of the pig sound signal E_v is defined as

$$E_v = \sum_{n=1}^{L} y_v^2(n) \tag{1}$$

where n is the sampling point number, and L is the frame length, which is taken as 25 ms (1200 sampling points) according to the short-time stationary characteristic of the speech signal.

The single parameter in the algorithm is the normalized short-time energy e_v. The formula is as follows:

$$e_v = \frac{E_v}{\sum_{v=1}^{V} E_v} \tag{2}$$

where, V is the total frames of the sound segment.

Thresholds T_1 and T_2 are calculated as:

$$T_1 = 1.5\max\{e_1, e_2, \cdots, e_{NIS}\} \tag{3}$$

$$T_2 = 1.1\max\{e_1, e_2, \cdots, e_{NIS}\} \tag{4}$$

where, NIS is the frame length of the front part of the pig sound segment.

Calculated from Equation (3) and (4), T_1 and T_2 are approximately 0.02 and 0.002 respectively. When e_v is higher than T_1, it is judged as speech frames. When it is lower or higher than T_2, it is determined as the start and end frames of the pig speech samples.

100 pig cough samples were randomly selected from the denoised continuous pig signal by the endpoint detection method. The start and end frame of the ith pig cough sample are v_{begin_i} an v_{end_i}, thus the duration of the ith pig cough sample is defined as

$$t_{dur_i} = \frac{(v_{end_i} - v_{begin_i})inc + L}{F_s} \tag{5}$$

where, inc represents the overlap, which is chosen to be 10 ms.

Then calculate the average duration of the 100 pig cough samples by the following equation

$$t_{ave} = \frac{\sum_{i=1}^{100} t_{dur_i}}{100} \tag{6}$$

The average duration of the 100 pig cough samples was calculated to be 0.529 s, the longest was 0.688 s, and the shortest was 0.388 s. Since this threshold is not calculated from all of the pig cough samples from the denoised 30-hour continuous pig sound signal, we extended the duration of the cough sample ranged from 0.338 to 0.738 s.

Then, the short-time energy E_v of the vth frame of the ith pig cough sample is calculated by Equation (1), then the energy of the ith pig cough sample is obtained as

$$E_{en_i} = \sum_{v=v_{begin_i}}^{v_{end_i}} E_v \tag{7}$$

Further, calculate the average energy of these 100 pig cough samples as

$$E_{ave}=\frac{\sum_{i=1}^{i=100}E_{en_i}}{100} \tag{8}$$

The calculated average energy of 100 pig cough samples is 189.60, the maximum is 822.87, and the minimum is 40.15. We only consider the lower limit energy of the pig cough samples, so this threshold is set to 35.15.

The endpoint detection algorithm is applied to the 360-segment continuous pig signal to obtain all the speech samples in the corpus. And then calculate the duration and energy of each sample, and remove those speech samples whose duration and energy are not within the set thresholds. After that the speech still contains a large number of silence segments, so most of them are deleted manually. At last, we cut 5 continuous speech samples from the 360 segments as one sentence, and 610 sentences are obtained. Thus, there are 3050 speech samples in the experimental corpus, of which 2032 are pig cough samples and 1018 are non-pig cough samples. Then label the 610 sentences, for example, the sentence "Sneeze, Cough, Cough, Hum, Scream" in Figure 5 can be labeled as "non-pig cough, pig cough, pig cough, non-pig cough, non-pig cough".

2.4 Extraction of signal characteristics

The Mel Frequency Cepstral Coefficients (MFCC)[32,33] is based on the auditory mechanism of the human ear. In MFCC, the linear spectrum is mapped into a non-linear Mel spectrum, and the spectral characteristics of the sound are analyzed based on human auditory experiments. After being framed and windowed, the spectral energy of the pig sound signal is calculated by the Fast Fourier Transform (FFT). And then apply the Mel filter to the energy to get the Mel filter energy and calculate its Discrete Cosine Transform (DCT). As a result, we obtain a 13-dimensional MFCC which can reflect the static characteristics of pig speech. At last, we add first-order and second-order differential coefficients which reflect the dynamic characteristics of the speech to obtain a 39-dimensional MFCC. The specific steps of MFCC extraction are shown in Figure 4.

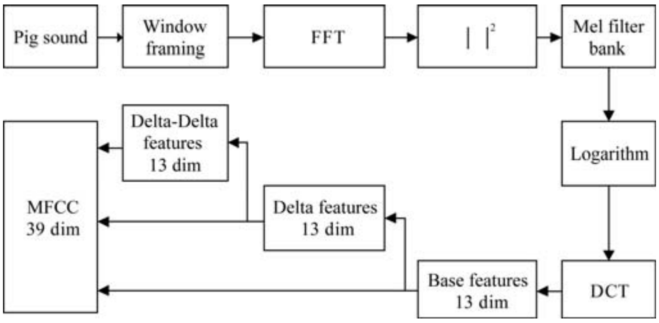


Figure 4 Flowchart of MFCC extraction

3 Establishment and training of pig sound acoustic model

3.1 Establishment of acoustic model of pig sound

HMM describes two stochastic processes, one is the short-time stationary properties of a non-stationary signal, called the observable process, and the other is how each short-time stationary property changes to the next short-time stationary property, called hidden dynamic process. The sound signal of the pig farm is an observable sequence $X=\{x_1, x_2, x_3, \cdots, x_T\}$, where T is their length. And this sequence is made by factors from pig itself or external conditions. These factors can be regarded as the hidden state sets $S=\{s_1, s_2, s_3, \cdots, s_K\}$, where K is the number of hidden states. There are three key parameters in HMM, including the initial hidden state probability distribution $\pi(q_1)=\{p(q_1=s_i)\}$, where q_t is the hidden state at time t, s_i is the ith hidden state in S; the transition probability $a_{s_is_j}=\{p(q_t=s_j|q_{t-1}=s_i)\}$ of the hidden state from s_i to s_j; and a model to estimate the observation probabilities $p\{X/S\}$.

Since the object of this recognition target is pig cough, all other sounds could be classified as non-pig cough. And in this paper, the pig cough and non-pig cough are named as pig cough syllable and non-cough syllable, and another silence syllable is introduced when the silent segments in the sentence are taken into consideration. Thus, these three kinds of syllables are respectively represented as "ks", "nks" and "sil".

In order to describe the three syllables, we use 5 hidden states to model the pig cough syllable or non-cough syllable, and 3 hidden states to model the silence syllable. So there are altogether 13 hidden states in HMM, the hidden states of "ks" is expressed as $\{s_1, s_2, s_3, \cdots, s_5\}$, the hidden states of "nks" is expressed as $\{s_6, \cdots, s_{10}\}$, and the hidden states of "sil" is expressed as $\{s_{11}, s_{12}, s_{13}\}$. Each hidden state has two transitions (self-loop or transition to the next hidden state), which we call transition-id after they are numbered. So there are a total of 26 transition-ids, expressed as $S^{tr}=\{s_1^{tr},\ldots,s_{26}^{tr}\}$, $\{s_i^{tr},s_{2i}^{tr}\}$ is the expression of the transition-id of s_i, where s_i^{tr} represents the self-loop transition of s_i, s_{2i}^{tr} indicates the transition from s_i to s_j. Because the continuous pig sound signal can be viewed as observable sequences resulting from the transition between hidden states, every frame from the sentence can be represented as a transition-id, which reflects the correspondence between the transitions and the hidden states. Figure 5 shows the correspondence between one sentence, 3 syllables, 13 hidden states, and 26 transition-ids.

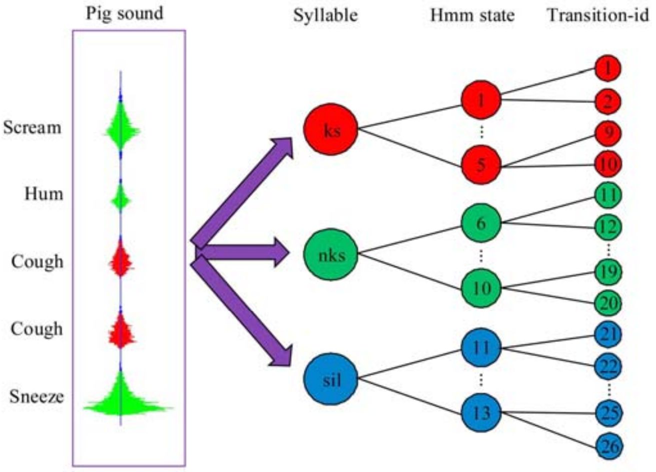


Figure 5 Correspondence between one sentence, 3 syllables, 13 hidden states, and 26 transition-ids

The DNN-HMM acoustic model structure is shown in Figure 6. From the paper [34], the traditional acoustic model GMM-HMM could be described as the following equation:

$$p(X/w)=\sum_q p(X,q/w)p(q/w)\cong \max \pi(q_1)\prod_{t=2}^{T}a_{q_{t-1}q_t}\prod_{t=1}^{T}p(x_t/q_t) \tag{9}$$

note that the observation probability is $p(x_t/q_t)=p(q_t/x_t)p(x_t)/p(q_t)$. where, w is a possible recognition sequence obtained by the Viterbi decoding algorithm[35,36]; x_t is the observation state at time t; $p(q_t)$ is the prior probability of each hidden state estimated from the training sentences, and $p(x_t)$ is independent of the word sequence and thus can be ignored[34].

Then Formula (9) could be further simplified into the expression of the DNN-HMM acoustic model.

$$p(X/w) \cong \max \pi(q_1)\prod_{t=2}^{T} a_{q_{t-1}q_t}\prod_{t=1}^{T} p(q_t/x_t)/p(q_t) \qquad (10)$$

From the above Equation (10), the DNN-HMM based acoustic model is mainly determined by the initial hidden state probability distribution $\pi(q_1)$ and the transition probability $a_{q_{t-1}q_t}$ of the HMM model, and the posterior probability $p(q_t/x_t)$ of DNN model.

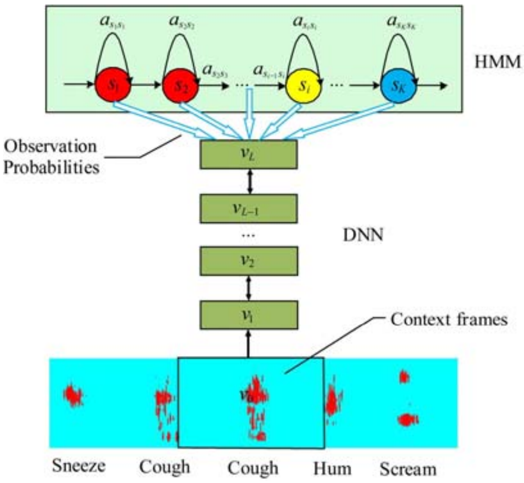


Figure 6 DNN-HMM acoustic model structure

3.2 GMM-HMM model training

The purpose of DNN-HMM based acoustic model training is to obtain an HMM model and a DNN model. Before training the DNN-HMM model, we should first train a GMM-HMM model[37,38] to generated alignment information in the training procedure of the DNN model. As each hidden state of the GMM-HMM model is modeled by a probability density function, we then get 13 probability density functions in the Gaussian mixture model[39]. The joint probability density function of GMM is expressed as:

$$p(X) = \sum_{k=1}^{K} c_k G(X;\mu_k,\Sigma_k) \qquad (11)$$

where, k is the kth probability density function; c_k is the weight of the kth probability density function; μ_k and $\sum_k$ denote the mean and covariance matrix of the kth probability density function respectively.

The model parameters include $\pi(q_1)$, $a_{s_is_j}$ from the HMM and Θ={c_k, μ_k, $\sum_k$} from the GMM. The HMM model parameters can be updated through the alignment information obtained by the Viterbi decoding algorithm, and the update of GMM model parameters is conducted by the Expectation-Maximization(EM) algorithm[40,41]. The GMM model parameters after the tth iteration are defined as $\Theta^{(t)}$={$c_k^{(t)}$, $\mu_k^{(t)}$, $\sum_k^{(t)}$}. Then the observation sequence $X^{(t)}$ is obtained after the process of Viterbi decoding. According to the correspondence between the transition-id and the hidden state, we can count the number of the transition-id s_{tr_i} and $s_{tr_{2i}}$ which belong to hidden states s_i. So the self-loop probability of hidden state s_i can be calculated by $a_{s_is_i} = \frac{s_i^{tr}}{s_i^{tr} + s_{2i}^{tr}}$, and the transition probability from s_i to s_j is defined as $a_{s_is_j} = \frac{s_{2i}^{tr}}{s_i^{tr} + s_{2i}^{tr}}$. In addition, hidden state initial probability indicates the probability of being in a hidden state at the initial moment, so the initial probability of the hidden state $\pi(q_1) = \frac{s_i^{tr(1)} + s_{2i}^{tr(1)}}{\sum_{i=1}^{K}(s_i^{tr(1)} + s_{2i}^{tr(1)})}$, where $s_i^{tr(1)}$ and $s_{2i}^{tr(1)}$ are the two transition-ids corresponding to the hidden state s_i at the initial time.

Since the hidden states are in one-to-one correspondence with the probability density functions, all frames corresponding to the kth probability density function can be obtained, which means, all the frames with the transition-id s_k^{tr} or s_{2k}^{tr} in the observation sequence $X^{(t)}$ are easy to be found, defined an $X_k^{(t)}$, the number is N_k. So from the E-step of the EM algorithm, we can get the following iterative formula.

$$h_{nk}^{(t)} = \frac{c_k^{(t)} G(X_k^{(t)};\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{k=1}^{K} c_k^{(t)} G(X_k^{(t)};\mu_k^{(t)},\Sigma_k^{(t)})} \qquad (12)$$

where, n=1, 2, 3, $\cdots$, N_k.

The updated parameters can be obtained by M-step:

$$\begin{aligned} c_k^{(t+1)} &= \frac{1}{N_k}\sum_{n=1}^{N_k} h_{nk}^{(t)} \\ \mu_k^{(t+1)} &= \frac{\sum_{n=1}^{N_k} h_{nk}^{(t)} X_k^{(t)}}{\sum_{n=1}^{N_k} h_{nk}^{(t)}} \\ \Sigma_k^{(t+1)} &= \frac{\sum_{n=1}^{N_k} h_{nk}^{(t)} (X_k^{(t)} - \mu_k^{(t)})(X_k^{(t)} - \mu_k^{(t)})^{\mathrm{T}}}{\sum_{n=1}^{N_k} h_{nk}^{(t)}} \end{aligned} \qquad (13)$$

In this procedure, the iterations are set to 40. When the procedure is done, the HMM and GMM model parameters are combined to construct a trained GMM-HMM model.

3.3 Unsupervised training of DBN

Deep Belief Network (DBN) is a generative model proposed by Hinton[42], forming by Restricted Boltzmann Machines (RBM)[43]. Its training is performed by the training of RBM, whose training method is the unsupervised layer-wise greedy pre-training algorithm. RBM is an energy model with only two layers, called visible layer $\boldsymbol{v}$ and hidden layer $\boldsymbol{h}$, the visible and hidden units form a bipartite graph with no visible-visible or hidden-hidden connections.

The hidden units take a binary value and obey the Bernoulli distribution $\boldsymbol{h}\in\{0,1\}^{N_h\times 1}$ where N_h indicates the number of hidden units. The RBM can be divided into Gauss-Bernoulli RBM and Bernoulli-Bernoulli RBM depending on the input of the visible layer, if the input feature vectors are continuous pig sound characteristics, it will be a Gaussian-binary RBM, and if the inputs are binary values, it will be a binary-binary RBM.

In this DNN-HMM model, the input is a feature vector of a continuous pig sound sentence which determines the first layer of RBM is Gaussian-binary RBM and the other layers are binary-binary RBM. For Gaussian-binary RBM, $\boldsymbol{v}\in R^{N_v\times 1}$ where N_v is the number of visible units. The energy function is defined as

$$E(\boldsymbol{v},\boldsymbol{h}) = \frac{1}{2}(\boldsymbol{v}-a)^{\mathrm{T}}(\boldsymbol{v}-a) - b^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{h}^{\mathrm{T}} w\boldsymbol{v} \qquad (14)$$

where, a and b represent the visible unit bias and hidden unit bias respectively, $w\in R^{N_h\times N_v}$ is the matrix of the visible and hidden connection weights.

For binary-binary RBM, $\boldsymbol{v}\in\{0,1\}^{N_v\times 1}$ and the energy function is defined as

$$E(\boldsymbol{v},\boldsymbol{h}) = -a^{\mathrm{T}}\boldsymbol{v} - b^{\mathrm{T}}\boldsymbol{h} - \boldsymbol{h}^{\mathrm{T}}w\boldsymbol{v} \tag{15}$$

Introduce a regularization factor $Z = \sum_{\boldsymbol{v},\boldsymbol{h}} e^{-E(\boldsymbol{v},\boldsymbol{h})}$ and define the joint probability of RBM as

$$P(\boldsymbol{v},\boldsymbol{h}) = \frac{e^{-E(\boldsymbol{v},\boldsymbol{h})}}{Z} \tag{16}$$

The RBM model parameters include w, a and b. The special network structure of RBM can be used to get the conditional probability $P(\boldsymbol{v}/\boldsymbol{h})$ and $P(\boldsymbol{h}/\boldsymbol{v})$, that from the hidden layer calculates the visible layer and from the visible layer calculates the hidden layer, as shown in Equation (17) and Equation (18) respectively.

$$P(\boldsymbol{v}/\boldsymbol{h}) = \prod_{i=1}^{N_{\boldsymbol{v}}} P(\boldsymbol{v_i}/\boldsymbol{h}) \tag{17}$$

$$P(\boldsymbol{h}/\boldsymbol{v}) = \prod_{j=1}^{N_{\boldsymbol{h}}} P(\boldsymbol{h_j}/\boldsymbol{v}) \tag{18}$$

According to the CD1 algorithm[44,45], when the hidden layer is calculated from the visible layer input, only one Gibbs Sampling is needed, that is, the hidden layer is used to estimate the visible layer, and then the hidden layer is estimated by the visible layer. Then RBM can perfectly perform the extraction of input characteristic parameters. For the feature vector $\boldsymbol{v}$, we first calculate the characteristic distribution of hidden units by Equation (18), and apply Gibbs sampling to the probability distribution to obtain $\boldsymbol{h}$; then generate $\boldsymbol{v}'$ from $\boldsymbol{h}$ by Equation (17), and finally generate $\boldsymbol{h}'$ by Equation (18). Here we get RBM parameter updating equations as follows

$$w^{(t+1)} = w^{(t)} + \eta_1(\boldsymbol{v}\boldsymbol{h}^{\mathrm{T}} - \boldsymbol{v}'\boldsymbol{h}'^{\mathrm{T}}) \tag{19}$$

$$a^{(t+1)} = a^{(t)} + \eta_1(\boldsymbol{v} - \boldsymbol{v}') \tag{20}$$

$$b^{(t+1)} = b^{(t)} + \eta_1(\boldsymbol{h} - \boldsymbol{h}') \tag{21}$$

where, η_1 is the RBM learning rate; $w^{(t)}$, $a^{(t)}$ and $b^{(t)}$ are the RBM model parameters obtained in the tth training, and $w^{(t+1)}$, $a^{(t+1)}$ and $b^{(t+1)}$ are the RBM model parameters obtained in the t+1th training.

3.4 Supervised training of DNN

In this DNN-HMM model, the DNN is used to simulate the posterior probability $p(q_t/x_t)$ of the HMM hidden state q_t under given input observation state conditions x_t, which is a classification problem. Therefore, a DNN model with L+1 layers is formed by adding a softmax layer to the trained l-layer DBN network where layer 0 is the characteristic input layer, and layer L is the softmax layer. DNN model parameters include w_l, b_l and l={1, 2, 3,…, L}. Here, the DNN uses the error back propagation algorithm[46,47] for training, and the DNN training parameters label y_{labels} are generated by the trained GMM-HMM model. The objective function J$_{\mathrm{CE}}$ is selected as the cross-entropy loss function and its function is

$$J_{CE} = -\sum_{i=1}^{K} y_{labels_i} \log o_{L_i} \tag{22}$$

where, y_{labels_i} is the ith element of $y_{labels} = \{y_{labels_1}, \cdots, y_{labels_K}\}$, o_{L_i} is the ith element of the output layer characteristic vector, $O_L = \{o_{L_1}, \cdots, o_{L_K}\}$.

After minimizing the objective function J_{CE} by the stochastic gradient descent method, we can get the neuron weight matrix $w_l^{(t)}$ of lth layer and l-1th layer and the lth layer neuron threshold vector $b_l^{(t)}$ after the tth iteration which are as follows

$$\frac{\partial J_{CE}}{\partial w_l^{(t)}} = e_l(O_l)^{\mathrm{T}} \tag{23}$$

$$\frac{\partial J_{CE}}{\partial b_l^{(t)}} = e_l \tag{24}$$

In the above formula $e_L = O_L - y_{labels}$, $e_{l-1} = (w_l^{(t)})^{\mathrm{T}}[f'(Z_l)e_l]$, $f(Z_l) = \dfrac{1}{1+e^{-Z_l}}$.

As $f'(Z_l) = f(Z_l)(1-f(Z_l))$, where Z_l represents the neuron input characteristic vector in the lth layer, O_l is the neuron output characteristic vector in the lth layer, e_l is the error in the lth layer, e_L is the error in the output layer.

So the DNN optimization formula is as follows

$$w_l^{(t+1)} = w_l^{(t)} + \eta_2 \frac{\partial J_{CE}}{\partial_{w_l^{(t)}}} \tag{25}$$

$$b_l^{(t+1)} = b_l^{(t)} + \eta_2 \frac{\partial J_{CE}}{\partial b_l^{(t)}} \tag{26}$$

where, η_2 is learning rate; $w_l^{(t+1)}$ and $b_l^{(t+1)}$ are the DNN model parameters obtained in the t+1th training.

Figure 7 shows the training flowchart of the DNN-HMM acoustic model. In the process of training, a GMM-HMM model is trained first, then the unsupervised DBN model is trained, based on which a softmax classification layer is added to form the DNN model, and then the supervised training is carried out with the help of the output y_{labels} of GMM-HMM model, followed by the re-evaluating of the initial probability $\pi(q_1)$ and transfer probability $a_{q_{t-1}q_t}$ of HMM. Finally, the DNN-HMM model is constructed and the posterior probability $p(q_t/x_t)$ is obtained.

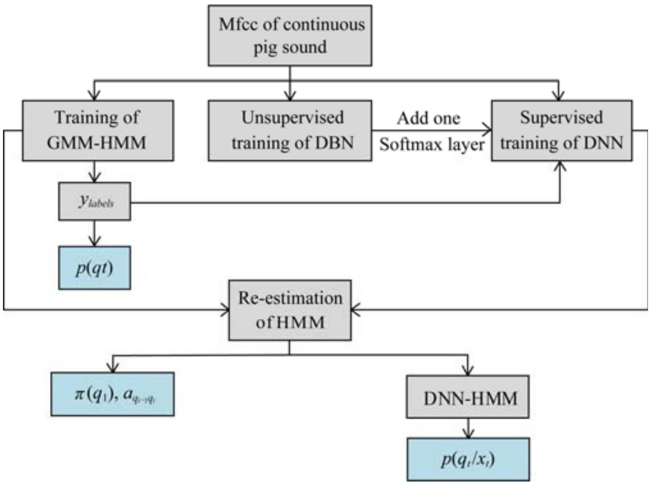

Mfcc of continuous pig sound
Training of GMM-HMM
Unsupervised training of DBN
Add one Softmax layer
Supervised training of DNN
y_{labels}
$p(qt)$
Re-estimation of HMM
$\pi(q_1)$, $a_{q_{t-1}q_t}$
DNN-HMM
$p(q_t/x_t)$

Figure 7 Flowchart of the training of DNN-HMM acoustic model

4 Results and discussion

4.1 Evaluation metrics

The recognized sequence in continuous pig voice is continuous. In this study, Word Error Rate (WER) is used to evaluate the performance of the continuous pig cough sound recognition system. Compare the recognition result with the reference sequence of test corpora, and get the sum of substitution error S (Substitution), insertion error I (Insertion) and deletion error D (Deletion), and divide by N, the total number of speech samples in the corpus, which bring the WER, that is, $\mathrm{WER}=\dfrac{S+I+D}{N}$.

The following is the interpretation of the three errors: Assume Reference sequence is (nks, ks, ks, nks, nks), and the Recognized sequence is (nks, nks, ks, nks, nks). By comparing the test results with the standard labels, we can see that the second non-pig cough sound in the recognition result is the insertion error, the third non-pig cough sound is the substitution error, and the unidentified last non-pig cough sound in the standard labels is classified as the deletion error.

We apply the 5-fold cross-validation method for experiments[48]. That is to say, the 610 sentence corpus is evenly divided into five mutually exclusive subsets of equal size, and then the union of four subsets is used each time as the training set, the rest one subset is used as a test set. So that trainings and tests can be performed five times, and the final result is the average recognition rate over the 5 tests.

4.2 Model parameter settings

The DNN-HMM acoustic model has a complex structure and many parameters, selecting a reasonable combination of parameters is of great significance to obtain a stable and reliable model. Considering the dimensions of the feature vectors and the amount of the training sentences, the layer number of depth neural network layer is chosen to be 3, and the number of units per layer is 100. For the first Gaussian-binary RBM, a learning rate of 0.01 was used for 40 epochs, while for the other binary-binary RBMs, a learning rate of 0.4 was fixed for 20 epochs.

In the acoustic model, DNN can use the information of adjacent frames to model the mutuality between context features. Here, through the 5-fold cross-validation experiments, the number of context frames was discussed in the range of 0-12 increased by 1. Figure 8 is the line chart of experimental results corresponding to context frame numbers. We can see that in the process of splice value increasing from 0 to 12, each group of WER shows a trend of decreasing first and then increasing. This trend can be better reflected by the average WER of the five groups. According to the line chart of mean change of WER, splice value is better in the range of 5-8. Considering that the larger the value of splice is, the higher the feature parameter dimension will be, to reduce network redundancy, select splice as 5, that is, the current frame is x_t, then the input of DNN model is $\{x_{t-5}, \cdots, x_{t-1}, x_t, x_{t+1}, \cdots, x_{t+5}\}$, a total of 11 frames correspond to the MFCC features of the continuous pig sound.

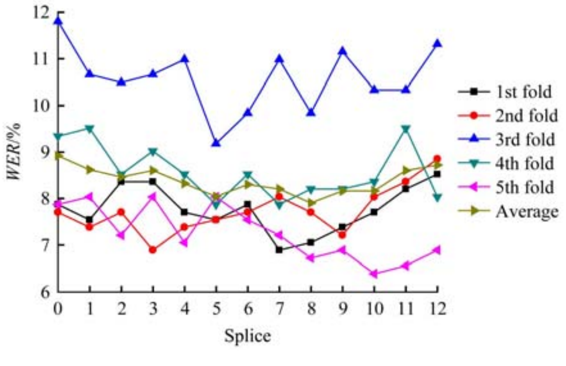


Figure 8 Line chart of WER of DNN-HMM acoustic model corresponding to context frames

4.3 Comparison of DNN-HMM and GMM-HMM acoustic model

Traditional acoustic model GMM-HMM uses GMM to describe continuous pig sound signals. Although enough Gaussian functions can simulate almost any data distribution, compared to DNN with a multi-layer neural network structure, shallow GMM modeling Capacity is still limited. In addition, DNN can not only capture high-order correlation characteristics in continuous pig sound signals but also consider the features of context frames by changing the number of input neurons, while GMM can only process one frame feature at a time. Furthermore, GMM, as a generative model, uses the EM algorithm for unsupervised training, aiming to simulate the original input continuous pig sound signal. As a classification model, DNN is trained with a supervised error back-propagation algorithm, which can better classify the input observation state into the corresponding hidden state.

In order to compare the traditional acoustic model GMM-HMM with DNN-HMM, the 5-fold cross-validation method was carried out, and the models were trained based on the model parameters selected in previous steps. The WER of the two models is shown in Table 1. According to the recognition results, the WER of all the five groups of DNN-HMM is lower than that of GMM-HMM, and the average WER is 3.45% lower. Specifically, among the three kinds of errors, the substitution error is obviously reduced, which shows the strong classification ability of DNN. At the same time, it can be seen from Table 1 that the WER of each group of DNN-HMM is maintained within 10.00%, the optimal WER is 7.54%, and the average WER is 8.03%, which shows that the model is stable and reliable.

Table 1 5-fold cross-validation WER between GMM-HMM and DNN-HMM/%

Species	GMM-HMM						DNN-HMM					
	ks	nks	*S*	*I*	*D*	WER	ks	nks	*S*	*I*	*D*	WER
1	416	194	63	3	3	11.31	416	194	39	4	3	7.54
2	418	192	47	3	2	8.52	418	192	42	2	2	7.54
3	364	246	86	5	5	15.74	364	246	46	7	3	9.18
4	399	211	51	3	3	9.34	399	211	41	5	2	7.87
5	435	175	70	3	3	12.46	435	175	45	3	1	8.03
Avg						11.48						8.03

5 Conclusions

(1) This paper proposed a new method for continuous pig cough sound recognition. The sounds from the pig house condition were divided into pig cough and non-pig cough. An acoustic model of continuous pig sounds was constructed based on DNN-HMM. The characteristic parameters corresponding to the continuous pig sound were considered as the observation sequences, while the factors that determine the causes of pig sounds were referred to as hidden states. So the continuous pig sounds were regarded as a signal generated by the transitions of the hidden states in the HMM.

(2) The DNN model which had a strong classification ability was introduced to describe the correspondences between hidden states and observation sequences in HMM. And DNN can not only consider multi-frame context features but also model nonlinear features of pig sounds.

(3) Through the 5-fold cross-validation experiments, the GMM-HMM model was compared with DNN-HMM. It turned out that the WER of DNN-HMM was lower than that of GMM-HMM, and the average WER was 3.45% lower. At the same time, the WER of each group of DNN-HMM was maintained within 10.00%, the optimal WER was 7.54%, and the average WER was 8.03%, which shows that the model was stable and reliable.

Acknowledgements

This work is supported by Project Supported by National Key Research and Development Program (Grant No. 2018YFD0500700) and Da Bei Nong Group Promoted Project for Young Scholar of HZAU (Grant No. 2017DBN005).

[References]

[1] van Zanten H H E, Bikker P, Meerburg B G, de Boer I J M. Attributional versus consequential life cycle assessment and feed optimization: alternative protein sources in pig diets. Int J Life Cycle Assess, 2018; 23(1): 1–11.

[2] Alexandratos N, Bruinsma A. World agriculture towards 2030/2050: The 2012 revision. Rome: Agricultural Development Economics Division. Food and Agriculture Organization of the United Nations, 2012; 147p.

[3] Huang W J, Zhu W X, Ma C H, Guo Y Z, Chen C. Identification of group-housed pigs based on Gabor and Local Binary Pattern features. Biosystems Engineering, 2018; 166: 90–100.

[4] Moura D J, Silva W T, Naas I A, Tolón Y A, Lima K A O, Vale M M. Real time computer stress monitoring of piglets using vocalization analysis. Computers and Electronics in Agriculture, 2008; 64(1): 11–18.

[5] Manteuffel G, Puppe B, Schön P C. Vocalization of farm animals as a measure of welfare. Applied Animal Behaviour Science, 2004; 88(1): 163–182.

[6] Marx G, Horn T, Thielebein J, Knubel B, von Borell E. Analysis of pain-related vocalization in young pigs. Journal of Sound and Vibration, 2003; 266(3): 687–698.

[7] Schön, P C, Puppe, B, Manteuffel, G. Automated recording of stress vocalization as a tool to document impaired welfare in pigs. Animal Welfare (South Mimms, England), 2004; 13(2): 105–110.

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

[9] Ferrari S, Silva M, Guarino M, Aerts J M, Berckmans D. Cough sound analysis to identify respiratory infection in pigs. Computers & Electronics in Agriculture, 2008; 64(2): 318–325.

[10] Moshou D, Chedad A, Van Hirtum A, De Baerdemaeker J, Berckmans D, Ramon H. Neural recognition system for swine cough. Mathematics & Computers in Simulation, 2001; 56(4-5): 475–487.

[11] Van Hirtum A, Berckmans D. Fuzzy approach for improved recognition of citric acid induced piglet coughing from continuous registration. Journal of Sound & Vibration, 2003; 266(3): 677–686.

[12] Van Hirtum A, Berckmans D. Objective recognition of cough sound as biomarker for aerial pollutants. Indoor Air, 2004; 14(1): 10–15.

[13] Guarino M, Jans P, Costa A, Aerts J M, Berckmans D. Field test of algorithm for automatic cough detection in pig house. Computers & Electronics in Agriculture, 2008; 62(1): 22–28.

[14] Liu Z Y, He X Y, Sang J, Li Y T, Wu Z Q, Lu Z M. Research on pig cough sound recognition based on hidden Markov model. The proceedings of the 10th Academic Seminar of Information Technology Branch of Chinese Society of Animal Husbandry and Veterinary Science, 2015; pp.99–104.

[15] Baker J K. Stochastic modeling for automatic speech understanding. In: Readings in Speech Recognition. San Francisco: Morgan Kaufmann Publishers, 1990; pp.297–307.

[16] Jelinek F. Continuous speech recognition by statistical methods. Proc IEEE, 1976; 64(4): 532–556.

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

[18] Reby D, André-Obrecht R, Galinier A, Farinas J, Cargnelutti B. Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (*Cervus elaphus*) stags. The Journal of the Acoustical Society of America, 2007; 120(6): 4080–4089.

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

[20] Trifa V M, Kirschel A N, Taylor C E, Vallejo E E. Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models. Journal of the Acoustical Society of America, 2008; 123(4): 2424–2431.

[21] Biagetti G, Crippa P, Falaschetti L, Orcioni S, Turchetti C. Learning HMM state sequences from phonemes for speech synthesis. Procedia Computer Science, 2016; 96: 1589–1596.

[22] Vidal E, Frinken V. Hmm word graph based keyword spotting in handwritten document images. Information Sciences, 2016; 370(C): 497–518.

[23] Kamilaris A, Prenafeta-Boldú F X. Deep learning in agriculture: A survey. Computers & Electronics in Agriculture, 2018; 147: 70–90.

[24] Schmidhuber, J. Deep learning in neural networks: an overview. Neural Networks, 2015; 61: 85–117.

[25] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015; 521: 436–444.

[26] Bengio Y. Learning deep architectures for AI. Foundations and Trends in Machine Learning, 2009; 2(1): 1–127.

[27] Bengio Y, Lecun Y. Scaling learning algorithms towards AI. Large-Scale Kernel Machines, 2007; pp.321–359.

[28] Dahl G E, Yu D, Deng L, Acero A. Large vocabulary continuous speech recognition with context-dependent DBN-HMMS. IEEE International Conference on Acoustics. IEEE, 2011; 125(3): 4688–4691.

[29] Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks. International Conference on International Conference on Machine Learning. Omnipress, 2012; pp.1–2.

[30] Maas A L, Qi P, Xie Z A, Hannun A Y, Lengerich C T, Jurafsky D, et al. Building DNN acoustic models for large vocabulary speech recognition. Computer Speech & Language, 2017; 41: 195–213.

[31] Hu Y, Loizou P C. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Transactions on Speech & Audio Processing, 2004; 12(1): 59–67.

[32] Ai O C, Hariharan M, Yaacob S, Chee L S. Classification of speech dysfluencies with mfcc and lpcc features. Expert Systems with Applications, 2012; 39(2): 2157–2165.

[33] Cao J, Zhao T, Wang J, Wang R, Chen Y. Excavation equipment classification based on improved MFCC features and elm. Neurocomputing, 2017; 216: 231–241.

[34] Dahl G E, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Transactions on Audio Speech and Language Processing, 2011; 20(1): 30–42.

[35] Sakoe, H, Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1978; 26(1): 43–49.

[36] Al-Naymat G, Chawla S, Taheri J. Sparse DTW: A novel approach to speed up dynamic time warping. In: Proceedings of the Eighth Australasian Data Mining Conference. Melbourne: Australian Computer Society, Inc, 2009; 101: 117–127.

[37] Deng L, Kenny P, Lennig M, Gupta V, Seitz F, Mermelstein P. Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition. IEEE Transactions on Signal Processing, 1991; 39(7): 1677–1681.

[38] Rabiner L R, Juang B H, Levinson S E, Sondhi M M. Recognition of isolated digits using hidden Markov models with continuous mixture densities. At and T Technical Journal, 1985; 64(6): 1211–1234.

[39] Zeng J, Xie L, Liu Z Q. Type-2 fuzzy gaussian mixture model. Pattern Recognition, 2008; 41(12): 3636–3643.

[40] Chuong B D, Serafim B. What is the expectation maximization algorithm? Nature Biotechnology, 2008; 26(8): 897–899.

[41] Hero A O. On the convergence of the EM algorithm. IEEE International Symposium on Information Theory, San Antonio: IEEE, 1993; 187p.

[42] Hinton G E, Osidero S, Teh Y W. A fast learning algorithm for deep belief nets. Neural Computation, 2006; 18(7): 1527–1554.

[43] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006; 313(5786): 504–507.

[44] Erhan D, Bengio Y, Courville A, Manzagol P A, Vincent P, Bengio S. Why does unsupervised pre-training help deep learning? Journal of Machine Learning Research, 2010; 11(3): 625–660.

[45] Hinton G E. Learning multiple layers of representation. Trends in Cognitive Sciences, 2007; 11(10): 428–434.

[46] Yu C C, Liu B D. A backpropagation algorithm with adaptive learning rate and momentum coefficient. International Joint Conference on Neural Networks. Honolulu: IEEE, 2002; pp.1218–1223.

[47] Hameed A A, Karlik B, Salman M S. Back-propagation algorithm with variable adaptive momentum. Knowledge-Based Systems, 2016; 114: 79–87.

[48] Michie D, Spiegelhalter D J, Taylor C C.Machine learning, neural, and statistical classification. Feb. 17, 1994.