International Journal of Automation and Computing
www.ijac.net

18(6), December 2021, 947-962
DOI: 10.1007/s11633-021-1283-2

Research on Voiceprint Recognition of Camouflage Voice Based on Deep Belief Network

Nan Jiang[1] Ting Liu[2]

[1] College of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 110854, China

[2] College of Light Industry, Liaoning University, Shenyang 110036, China

Abstract: The problem of disguised voice recognition based on deep belief networks is studied. A hybrid feature extraction algorithm based on formants, Gammatone frequency cepstrum coefficients (GFCC) and their different coefficients is proposed to extract more discriminative speaker features from the original voice data. Using mixed features as the input of the model, a masquerade voice library is constructed. A masquerade voice recognition model based on a depth belief network is proposed. A dropout strategy is introduced to prevent overfitting, which effectively solves the problems of traditional Gaussian mixture models, such as insufficient modeling ability and low discrimination. Experimental results show that the proposed disguised voice recognition method can better fit the feature distribution, and significantly improve the classification effect and recognition rate.

Keywords: Disguised voice recognition, deep belief network, feature extraction, Gammatone frequency cepstrum coefficients (GFCC), dropout.

Citation: N. Jiang, T. Liu. Research on voiceprint recognition of camouflage voice based on deep belief network. *International Journal of Automation and Computing*, vol.18, no.6, pp.947–962, 2021. http://doi.org/10.1007/s11633-021-1283-2

1 Introduction

In recent years, biometrics, which is based on fingerprint, face, iris and other physiological features, has developed rapidly in many fields and been widely used[1]. Because these physiological characteristics are relatively stable for the same person and have relatively unique characteristics for different people, the recognition effect is better. Compared with other biometric recognition technology, voiceprint recognition has been gradually applied to many fields because of its unique advantages[2]. Voiceprint technology from the original voice to extract the unique characteristics of the individual, only need to collect voice without direct contact with people, the user is more acceptable. And it requires less equipment, only a device with the function of recording. However, face, fingerprint and other identification technologies are usually more expensive, because these technologies need to use professional scanning equipment, need to be certified by professional institutions.

Therefore, voiceprint recognition has obvious advantages over other biometric technologies, and it can be applied in public security and judicial departments. For example, if the police obtain a recording of a criminal at the scene of a crime, they can compare the voiceprint information of the recording with the voiceprint information in the trained database to find the identity information of the suspect. By using voiceprint recognition method, law enforcement agencies can quickly and efficiently arrest criminal suspects.

The voiceprint recognition technology facilitates our daily life[3]. But unlike biometric technologies such as fingerprints, irises and DNA, voice is not immutable. Due to the influence of internal and external factors such as background noise interference, channel transformation, disguised, excitement and pressure, sound will change. Voice variation makes it difficult to use voice for identity authentication, and when it is used by illegal elements, it will bring trouble and crisis to our life[4].

In voice cases, more and more criminals use whispers, falsetto, imitation of other people′s voice and other means to disguise their voices in order to conceal their identity and avoid arrest[5]. For the normal voice, due to the influence of non-human factors, the voice is often distorted, which brings some difficulties to the voice identification, and the appearance of disguised voice makes the identity authentication more difficult[6,7]. Therefore, improving the performance of voiceprint recognition system under the condition of disguised voice is of great significance for identity recognition and forensic evidence[8, 9].

Matveev[10] investigated the impact of age-related

Research Article
Manuscript received August 15, 2020; accepted January 22, 2021; published online March 12, 2021
Recommended by Associate Editor Shuang-Hua Yang
Colored figures are available in the online version at https://link.springer.com/journal/11633

voice changes on voiceprint recognition performance based on voice data collected from 2006 to 2010, and found that the performance of automatic voiceprint recognition system tended to decline over a four-year period. A statistical feature and support vector machine (SVM) classifier algorithm based on mean and correlation coefficient can separate the disguised voice from the original voice[11]. In 2015, Wu et al.[12] held the first disguised voice test contest and released the first statistical analysis system (SAS) database designed for disguised voice recognition research[12]. One algorithm successfully uses pitch estimation scale factor and improved Mel frequency cepstrum coefficient (MFCC) extraction algorithm to eliminate the disguised effect, and verifies the identity of the speaker by transforming the disguised voice[13]. By analyzing the recognition results of human ear to disguised sounds and the disguised effects of different disguised sounds, the most difficult disguised to recognize is deduced[14]. Using normal voice and ten kinds of disguised voice, human auditory experiments of a familiar and unfamiliar speaker′s disguised voice are carried out, and it is found that the recognition of whispers is the most difficult and the disguised effect is the best. A method for recognizing an electronically disguised voice is proposed[15]. A Gaussian mixture model (GMM) model is established to construct the combination features of the mean vector. Then the SVM classifier is used for training and recognition.

At present, the achievements of disguised speech recognition are almost all in the form of electronic disguised speech, while the achievements of physical disguised speech recognition are mainly based on feature analysis. Zhou et al.[16,17] have studied voiceprint recognition with mixed features of Gammatone frequency cepstrum coefficients (GFCC) and MFCC, but the recognition model is a GMM method using shallow network. Cao et al.[18,19] mainly studied the classical GMM model for voiceprint recognition, but GMM is a shallow network structure model, and its ability to represent complex functions is limited in the case of not enough samples. Lv and Pan[20,21] have studied speaker recognition based on deep neural networks, but only a single MFCC feature is considered in feature extraction.

Therefore, we discuss the effect of disguised voice on voiceprint recognition performance from two aspects: feature extraction and model building. In order to obtain more robust and better voice features, a hybrid feature parameter method based on the combination of GFCC and formant is proposed, which can effectively improve the recognition accuracy of a disguised voice. In order to solve the problem of low modeling ability and low discrimination of traditional models, a disguised voice recognition model based on a depth belief network is proposed. Meanwhile, a dropout strategy is introduced. Compared with the GMM model, the proposed depth model can effectively improve the poor performance of over-fitting and disguised voice recognition system.

2 Voiceprint feature extraction of camou-flage voice based on GFCC and formant

The technical difficulty of disguised voice voiceprint recognition is that when the voice is disguised, some characteristic parameters of the voice can be changed greatly. It is a key step to improve the performance of speaker recognition system by extracting more discriminative feature parameters to reduce the influence of the voice.

We focus on feature extraction to solve the problem of low performance of the disguised voice recognition system. Considering the combination of formant and feature parameters, a robust and more distinctive voice feature is extracted from the limited original voice data, which can effectively improve the poor performance of the traditional disguised voice voiceprint recognition system.

2.1 Detection of formant based on cepstrum

Formant is one of the important characteristic parameters of voiceprint recognition, and its parameters include formant frequency and bandwidth. The spectral envelope of vocal tract information is approximately the same as that of voice information, so the formant extraction is to obtain the spectral envelope of voice, and the maximum of the spectral envelope is regarded as the formant parameter[22].

In this paper, the formant of the language is calculated based on cepstrum. First, the homomorphic analysis method is used to eliminate the influence of the excitation, and the information of the vocal tract is obtained.

Based on the homomorphic deconvolution technique, the pitch information is separated from the vocal tract information in the cepstrum domain. It is more accurate and effective to extract the formant of the voice by using the information of vocal tract.

The voice $x(n)$ is obtained by filtering the glottal pulse $e(n)$ through the channel response $h(n)$ as shown in (1).

$$x(n) = e(n) \times h(n). \tag{1}$$

The cepstrum calculation for voice signals is

$$\hat{x}(n) = \hat{e}(n) + \hat{h}(n). \tag{2}$$

Therefore, it can be concluded that pitch information $\hat{e}(n)$ and vocal tract information $\hat{h}(n)$ in the cepstrum domain are relatively independent. $e(n)$ and $h(n)$ can be separated by cepstrum, and then the resonance peak can be obtained according to the excitation $h(n)$ and the characteristics of the cepstrum. The specific steps are as follows:

1) By pre-emphasizing, windowing and framing the audio signal $x(n)$ (frame length N), we can get $x_i(n)$ and i represents the i-th frame of the sound signal.

2) The discrete Fourier transform of $x_i(n)$ is carried out to obtain:

$$X_i(k) = \sum_{n-0}^{N-1} x_i(n)\,\mathrm{e}^{-\frac{\mathrm{j}2\pi kn}{N}}. \tag{3}$$

3) Take the amplitude of $X_i(k)$ and then take the logarithm to obtain:

$$\hat{X}_i(k) = \log(|X_i(k)|). \tag{4}$$

4) An inverse Fourier transform is performed on $\hat{X}_i(k)$ to obtain a cepstrum sequence.

$$\hat{x}_i(n) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}_i(k)\,\mathrm{e}^{\frac{\mathrm{j}2\pi kn}{N}}. \tag{5}$$

5) Set a low-pass window function *window*(n) on the inverted frequency domain axis, which can generally be set as a rectangular window:

$$window(n) = \begin{cases} 1, \text{ if } n \le n_0 - 1 \text{ and } n \ge N - n_0 + 1 \\ 0, \text{ if } n_0 - 1 < n < N - n_0 + 1 \end{cases} \tag{6}$$

where n_0 is the width of the window function, and then the window function is multiplied by the cepstral sequence $\hat{x}(n)$ to obtain:

$$h_i(n) = \hat{x}_i(n) \times window(n). \tag{7}$$

6) After Fourier transformation of $h_i(n)$, the envelope of $X_i(k)$ is obtained:

$$H_i(k) = \sum_{n=0}^{N-1} h_i(n)\,\mathrm{e}^{-\frac{\mathrm{j}2\pi kn}{N}}. \tag{8}$$

7) The formant parameters can be obtained by searching for the maximum on the envelope.

As shown in Fig. 1, the envelope (black thick line) calculated by the cepstrum is used to show the location of the formant peak with black dots, and the corresponding frequency of the formant is marked with dotted lines. By calculating the four resonance peak, frequency is: 1 593.75, 3 062.50, 4 312.50, 7 187.50.

2.2 Extraction of GFCC parameters

1) Gammatone filter

Our perception of sound is mainly through the cochlea. The basement membrane is the most important part of the cochlea that receives voice signals. The basement membrane has not only frequency selective characteristics, but also spectral analysis characteristics. It can match different frequency components with different positions of the basement membrane, and transform the frequency intensity into the amplitude of the basement membrane.

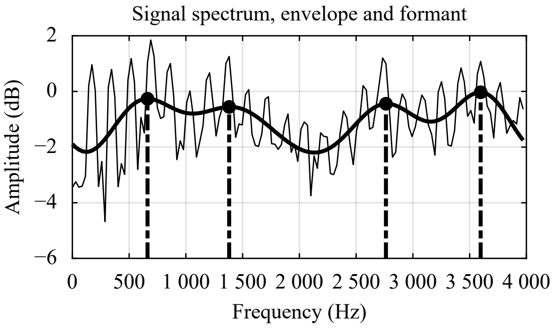


Fig. 1 Formant frequency of speech

The time-domain expression of the Gammatone filter is as follows:

$$h(t) = kt^{n-1}\mathrm{e}^{-2\pi bt}\cos(2\pi f_c t + \phi), \mathrm{t} \ge 0 \tag{9}$$

where ϕ is the phase, f_c is the center frequency, n is the order of the filter. When $n = 3, 4, 5$, the Gammatone filter can better simulate the auditory characteristics of the human ear basement membrane. k is the filter gain. b is the attenuation factor, which depends on the filter bandwidth. It controls the rate of decay of the impulse response. Its relation to the center frequency f is

$$b = 1.019 \times 24.7 \times (4.37 \times f_c/1\,000 + 1). \tag{10}$$

Equation (9) consists of two parts: the filter envelope $kt^{n-1}\mathrm{e}^{-2\pi bt}$ and the amplitude $\cos(2\pi f_c + \phi)$ of the frequency f_c. Fig. 2 shows the frequency response of a Gammatone filter.

A number of Gammatone filters with different center frequencies can be combined to form a filter bank. Signals using this filter bank can represent the response characteristics of the original voice signal at different frequency components. Fig. 3 is a simulated cochlear model composed of 24 Gammatone filters.

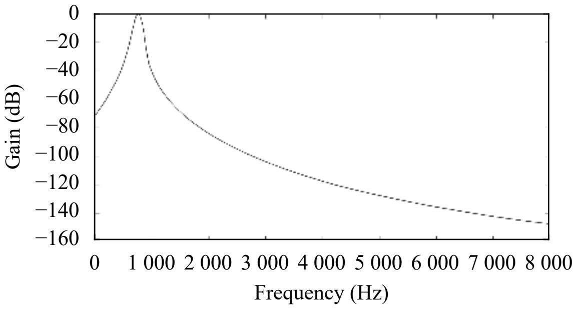


Fig. 2 Frequency response of Gammatone filter

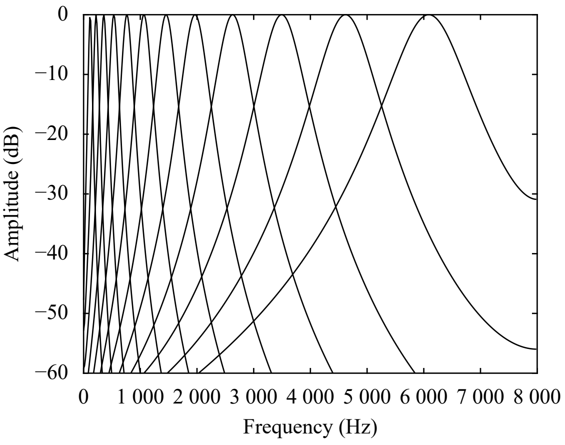


Fig. 3 Gammatone filter bank

2) GFCC feature extraction

After the voice signal is preprocessed, a set of cepstral feature parameters can be obtained through the Gammatone filter bank based on the auditory characteristics of the human cochlea. The parameter is recorded as GFCC (Gammatone frequency cepstrum coefficients), which can be further used in a speaker recognition system. In the presence of noise, the recognition rate and robustness of this feature parameter is better than the traditional feature parameter MFCC, and it can have more advantages in the case of low SNR.

Although GFCC can reflect the static characteristics of sound signals, human ears are more sensitive to the dynamic characteristics of sound. The system can achieve a higher recognition rate by adding the difference parameter which represents the dynamic characteristics of voice into the feature parameters and combining the static features and dynamic features.

The first order difference and the second order difference are selected as dynamic features. Combining GFCC with the first order difference and the second order difference, we can get the eigenvector of GFCC. Fig. 4 shows the characteristic parameters of GFCC, first-order difference and second-order difference of a segment of voice.

2.3 Gaussian mixture model

The Gaussian mixture model is formed by a linearly weighted combination of a plurality of Gaussian probability density functions, as shown in (11).

$$p(x_i) = \sum_{j=1}^{M} \phi_j N_j \left(x_i; \mu_j, \sum_j \right) \tag{11}$$

where M represents the degree of mixture of the model, i.e., the number of Gaussian components. ϕ_j is the weight corresponding to the j-th Gaussian component, and $\sum_{j=1}^{M} \phi_j = 1$, N_j are used to represent the j-th single

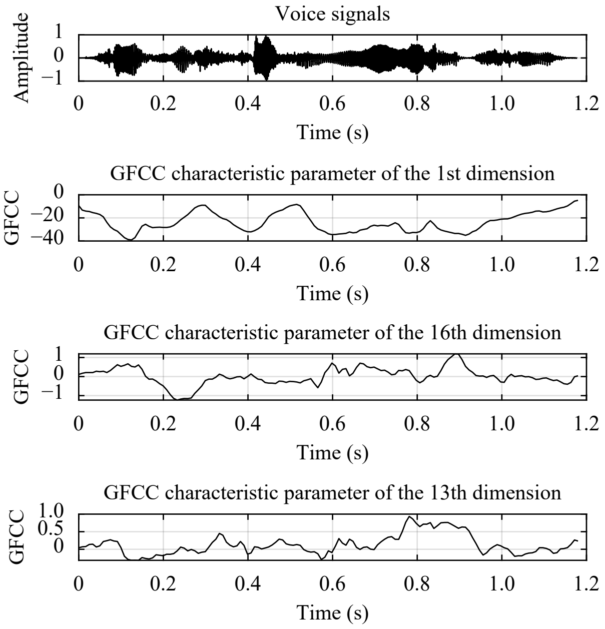


Fig. 4 GFCC characteristic parameter of different dimensions

Gaussian probability density function, see below.

$$N_j \left(x_i; \mu_j, \sum_j \right) = \frac{1}{\sqrt{(2\pi)^n \left| \sum_j \right|}} \exp\left[-\frac{1}{2}(x-\mu_j)^{\mathrm{T}} \sum_j^{-1} (x-\mu_j) \right]. \tag{12}$$

Because the EM algorithm is an iterative method to solve the model parameters in the case of incomplete data and loss data, the EM algorithm is applied to the parameter estimation of Gaussian mixture model, and (13) and (14) are obtained.

E-Step:

$$w_j^{(i)} = Q_i \left(z^{(i)} = j \right) = P\left(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma \right) \tag{13}$$

M-Step:

$$\begin{aligned} \phi_j &= \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)} \\ \mu_j &= \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \\ \sum_j &= \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^{\mathrm{T}}}{\sum_{i=1}^{m} w_j^{(i)}}. \end{aligned} \tag{14}$$

In Step E, we treat $\phi, \mu, \sum$ as a constant and compute the probability of $z^{(i)}$, i.e., the probability that sample i belongs to class j. In step M, ϕ_j is the ratio of

$z^{(i)} = j$ in the sample class, i.e., the weight of the Gaussian component. μ_j is the mean of the sample features of class j. $\sum_j$ is the covariance matrix of the example of class j.

2.4 Improvement of feature extraction algorithm based on hybrid parameters

The Gammatone filter bank for extracting GFCC features is based on the human cochlear auditory model. Each filter has a steep edge on both sides of the center frequency, which can better simulate the characteristics of frequency selection and spectral analysis of the basement membrane, and suppress the interference of background noise, so this paper uses GFCC as a feature parameter.

Formant is one of the important parameters for describing the vocal tract in voice signal processing, and GFCC is a kind of auditory characteristic simulating the human ear. The two kinds of voice feature parameters are weighted and combined by different weight coefficients, which can not only reflect the mechanism of human pronunciation, but also reflect the perceptual characteristics of human ears. The combination of human voice and hearing, as the basis of acoustic modeling, can better reflect the personality characteristics of the speaker, as shown in Fig. 5.

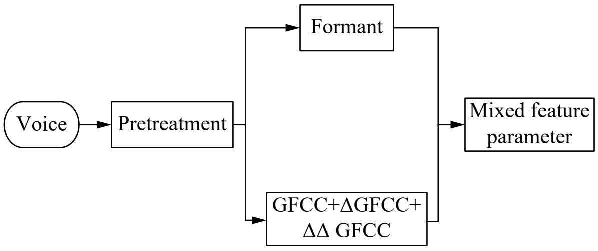


Fig. 5 Block diagram of combining feature parameter

Firstly, the voice signal is preprocessed to extract the vocal tract impulse information $\hat{h}(n)$ and G_m (GFCC and the first-order difference and second-order difference of GFCC) obtained by Gammatone filter. Then, these two kinds of feature parameters are combined into a hybrid feature as the input of the acoustic model. Thereafter, $\hat{h}(n)$ and G_m are normalized as shown in (15) and (16).

$$\hat{h}(n)' = \frac{\hat{h}(n)}{\hat{h}(n)_{\max}} \tag{15}$$

$$G_m{}' = \frac{G_m}{G_{m-\max}} \tag{16}$$

where $\hat{h}(n)_{\max}$ is the maximum value of formant characteristic parameter, $G_{m-\max}$ is the maximum value of GFCC and its difference characteristic parameter. In this way, both $\hat{h}(n)'$ and $G_m{}'$ are data between 0 and 1. And then,

$$d_1 = \hat{h}(n)' \tag{17}$$

$$d_2 = G_m{}'. \tag{18}$$

The impact factors of the two methods in (19) and (20) can be expressed by the average value of the test set.

$$C_1 = \frac{ave\,(d_1)}{ave\,(d_1) + ave\,(d_2)} \tag{19}$$

$$C_2 = \frac{ave\,(d_2)}{ave\,(d_1) + ave\,(d_2)} \tag{20}$$

where C_1 and C_2 represent the influence of the two feature parameters on the recognition results. The resulting mixed feature parameter is a weighted combination of the two feature parameters, as shown in (21).

$$S = C_1\hat{h}(n)' + C_2 G_m{}'. \tag{21}$$

2.5 Experiment and result analysis

Construction of disguised voice database

1) The choice of pronouncers and the design of the language material

The general principle of the construction of the disguised speech database is as follows: The study objects are the students who speak standard mandarin and enunciate clearly, in order to eliminate the interference of other factors on the phonetic variation. The self-built disguised speech database consists of 49 students, both male and female, all postgraduates in the school, whose ages are between 24 and 26.

The pronouncers come from different parts of the same country, and their mandarin proficiency is relatively good. It is not excluded that some pronouncers still have obvious dialect characteristics. From the perspective of phonetics, the normal speech database and the physical disguised speech database are established respectively. The normal speech database is used for training, and the physically disguised speech database is used for testing. According to the characteristics of physically disguised speech, the speaker should have certain performance ability, and be able to make corresponding disguised speech according to the way of disguised[23].

The selected physical disguises were fast, slow, high, low, whispering, biting a pencil and pinching a nose. Each speaker repeats the pronunciation three times for each physical disguised pronunciation mode and one time for the normal pronunciation. 49 normal speech samples of 49 speakers and 343 speech samples of different physically disguised modes can be collected.

As for the selection of the content of the corpus, we mainly consider two aspects: i) The words or phrases included in the prediction do not have an obvious emotional tendency. ii) The corpus should include main vowels, monophthongs and diphthongs. The above considerations are to exclude the influence of emotion on disguised speech, and to ensure that the speech conforms to the norms of phonetics research. The content of the corpus is I am XX, the school number is XX.

2) Hardware and software equipment for voice recording of physical disguised

In order to eliminate the interference of background noise, the recording location is selected in a quiet studio, using laptop computers and a Newsmy voice recorder, the voice processing software is Adobe Audition CS6. Before recording, let the speaker be familiar with the voice text and practice pronunciation. The voice is recorded by monophonic channel, the storage format is 16 bits, the sampling rate is 32 kHz, and the storage format is WAV.

3) Speech processing, annotation and classification

Before annotating the speech, it needs to preprocess, standardize and unify the length and format. The voice will be recorded into the computer, with Adobe Audition CS6 to cut all the voice, remove the blank part and the part of the effect that is not good, and the voice length is standardized at 3 s. The output is saved as an 8 000 Hz sample rate, 16-bit WAV file.

After the voice processing is completed, it is labeled, i.e., each voice is named according to certain rules. For the sake of convenience, the file names of normal speech are named by digital numbers, and the naming rules corresponding to different disguised modes are the same as those of normal speech.

After each voice is labeled, it is classified, i.e., according to its different way of speaking, it is placed in different folders, which are named for its corresponding way of speaking. This results in one normal voice folder and seven physically disguised voice folders, each containing 49 voices.

Experimental design

This part focuses on the influence of the characteristic parameters on the disguised voice recognition system. In order to show that the hybrid feature parameters can effectively improve the performance of the system, we compare the hybrid feature parameters with GFCC. The training voices used in the experiment are normal voice and disguised voice, and the test voice is disguised voice. An improved map based on feature extraction is shown in Fig. 6.

Firstly, the training voice and the test voice are preprocessed, the length of the frame is 256 points, and the frame is shifted to 80 points. In order to reduce the edge effect of the voice frame, a Hamming window is used to add a window. Then the endpoint detection method is used.

In the training process, the training voice signal is preprocessed, and then the feature parameters of the training voice are extracted. The formant coefficients and $\mathrm{GFCC}+\Delta\mathrm{GFCC}+\Delta\Delta\mathrm{GFCC}$ coefficients are extracted and linearly combined with different weight coefficients. The feature parameters of each frame are 39 dimensions (3 dimensions formant, 12 dimensions Gammatone filter output, 12 dimensions first order difference coefficients and 12 dimensions second order difference coefficients). Thus, the feature vectors of 49 speakers can be obtained. The extracted feature parameters are used as the input of the Gaussian mixture model. The Gaussian mixture model used in this experiment is composed of 32 Gaussian

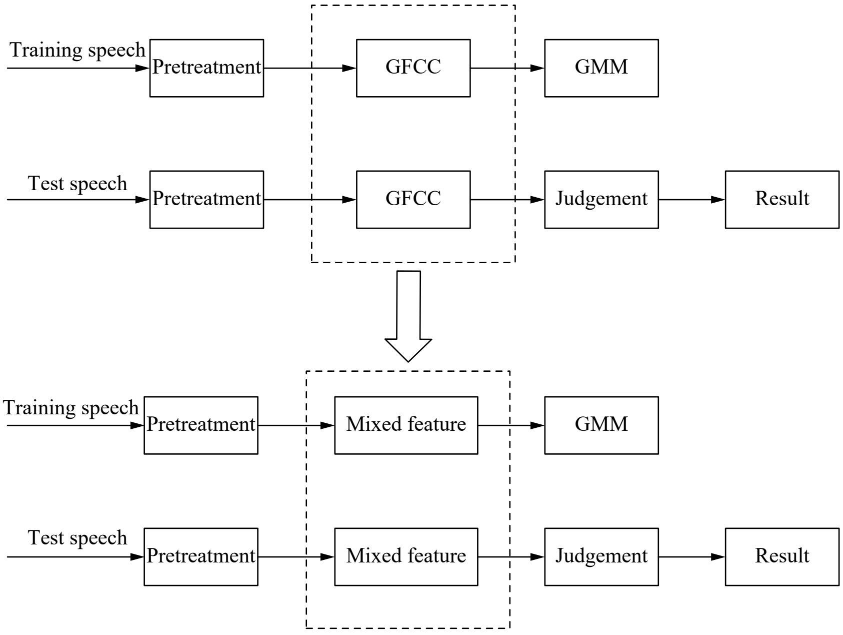


Fig. 6 Graph based feature extraction improvement

models, and then the parameters of each GMM can be obtained at the end of the training process.

In the test session, the same preprocessing and feature extraction methods are used as in the training session, the obtained mixed feature parameters are matched with each GMM model, the similarity score with each GMM is calculated, and the label with higher score is the one to be recognized. The correct label of the voice to be recognized is compared with the recognition label obtained in the testing phase; if the two labels are the same, it shows that the result obtained in the testing phase is correct, otherwise, the result obtained is wrong, so the accuracy of the voiceprint recognition system is obtained by calculating the number of correct recognition results.

In the voiceprint recognition system, accuracy is undoubtedly the most direct and important performance evaluation index, if a higher recognition rate cannot be guaranteed, then the recognition system will be of no great use. For a speaker recognition system, the accuracy usually represents the probability that the system will recognize the correct sample, and is calculated mathematically as (22).

$$C_{ID} = \frac{n_{correct}}{n_{total}} \times 100\% \tag{22}$$

where C_{ID} is the accuracy rate, and $n_{correct}$ is the number of correctly identified samples, and n_{total} is the total number of samples to be identified.

Experimental results and analysis

Zhou et al.[16,17] have studied the voice print recognition based on the GFCC and MFCC hybrid features, so the recognition rate of the two methods is compared with the GFCC + Formant hybrid feature method proposed in this paper.

Based on the parameters of MFCC, GFCC, GFCC + MFCC and GFCC + Formant mixed features, in the case of GMM model, using normal speech as training data and different disguised speech as test data, the performance of the speaker recognition system is judged by the accuracy. Thus, the accuracy of different feature extraction methods for different disguised voices can be obtained, as shown in Table 1.

Table 1 System recognition rate of different disguised voices based on different features

Disguised voice	MFCC	GFCC	GFCC+ MFCC	GFCC+Formant
Fast	79.65	89.80	91.50	95.92
Slow	77.38	81.63	94.33	93.88
Treble	57.30	59.18	65.38	67.35
Bass	65.15	75.51	87.55	89.80
Whispers	33.45	38.78	38.78	40.82
Nose pinching	39.75	44.90	52.70	55.10
Pencil biting	59.35	65.31	66.45	65.31

In order to clearly see the accuracy of the system using different feature extraction algorithms under different disguised voice tags, the form of a histogram is used, as shown in Fig. 7. The first bar of each set of histograms represents the MFCC as the feature parameter, the second represents the GFCC as the feature parameter, the third represents the use of GFCC and MFCC mixed feature parameter, and the fourth represents the use of GFCC and Formant mixed feature parameter.

From Table 1 and Fig. 7, we can see the influence of different feature parameters on voiceprint recognition systems. In the case that all models are GMM, the recognition rate of the system based on mixed feature parameters is higher than that of the system using only GFCC or MFCC feature parameters, except for biting pencils. Compared with the acoustic system only using GFCC or MFCC features, the acoustic system based on mixed feature parameters combines the vocal characteristics and auditory characteristics, which can better reflect the information of the speaker and have a better classification effect on the disguised voice. The recognition rate of the hybrid feature is better than that of the single feature method, and the recognition rate of the GFCC + Formant hybrid feature is also higher than that of the GFCC + MFCC hybrid feature.

When biting a pencil to pronounce, because the teeth and one corner of the mouth cannot be completely closed, it sounds like a leak. The articulator′s tongue, lips, teeth and other vocal organs are inhibited, which cannot pronounce normally, thus affecting the formant frequency to a certain extent. Therefore, under the masquerade label of biting pencil, the recognition rate of the acoustic system based on mixed features is not improved compared with that based on GFCC.

It can also be seen that among the seven disguised methods, fast and slow have least influence on speaker recognition, while whisper and nose pinch have the greatest influence on speaker recognition. Although the recognition rate of the system based on mixed feature parameters is improved under the disguised labels of whisper and nose pinching, the recognition rate is only 48.2% and 55.1%, which is similar to human auditory perception. Whispering is mainly caused by the friction between the

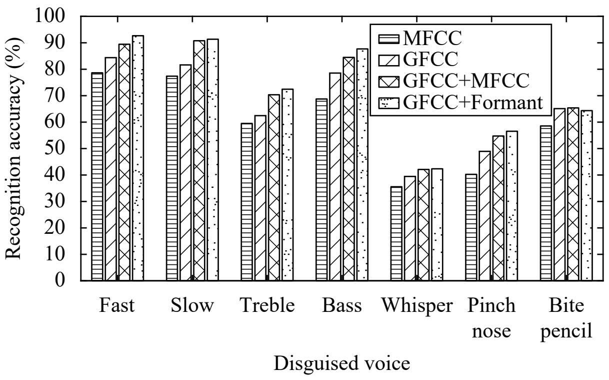


Fig. 7 Recognition rate of different features

air flow and the vocal organs. The vocal cords do not vibrate during articulation, and pinching the nose leads to nasal obstruction, which greatly changes the resonance characteristics of the voice cavity. As these two kinds of disguised for the pronouncer are relatively easy to do, the pronouncer does not need to change the habit of pronunciation, so the disguised effect is better and the system recognition rate is relatively low.

3 Voiceprint recognition of disguised voice based on DBN model

The key to improving the accuracy of disguised voiceprint recognition is to mine the hidden speaker information from the voice data. In the above content, we study the methods to improve the performance of voiceprint recognition from the feature extraction level. The research of the voiceprint recognition model is another important part of the speaker recognition system, the acoustic model has an important impact on the performance of the system. This part starts with the acoustic model of the system to solve the problem of poor performance of disguised voice recognition. We can construct a powerful deep model to deal with voiceprint recognition. At the same time, in order to simulate the way of thinking of the human brain, we can take the depth belief network as the acoustic model of the recognition system to realize the recognition of disguised voice speakers.

2006 is the first year with deep learning, and the research boom of deep learning is the deep belief network (DBN) proposed by Hinton, which is one of the first successful applications of a deep network model training non-convolution model.

DBN is a deep generative network composed of a set of restricted boltzmann machines (RBMs), which is a generative model with multiple hidden variable layers. Hidden layer neurons are usually binary (0 or 1), while explicit layer neurons can use binary or real numbers. Although DBNs with relatively sparse connections can be constructed, in most cases, all neurons in different layers are connected, and there is no connection between neurons in different layers. The structure of a DBN is shown in Fig. 8.

The core of a DBN is a greedy, layer-by-layer learning algorithm. The parameters obtained by pre-training in an unsupervised way can provide good initial points, and the results are usually better than those obtained by random initialization[24]. Then, the parameters are fine-tuned by the supervised back-propagation algorithm, which can effectively solve the local optimal situation and under-fitting problem of the deep network.

A DBN is composed of many RBMs in series, in which the hidden layer of the former RBM is the visual layer of the latter RBM, and the output of the former RBM is the input of the latter RBM. When the model is trained, the parameters of the former RBM are kept unchanged after the former RBM is fully trained, and the latter RBM is trained until all RBMs are trained.

3.1 Training of the DBN

The parameters of the deep confidence network are directly obtained by RBM unsupervised training. An important feature of DBNs is that their hidden States can be efficiently and correctly inferred by bottom-up passing, and that up-bottom generated weights are used inversely. Another important feature is that the new DBN has a lower bound on the logarithmic probability of the training data when the DBN adds an additional feature learning layer, which is better than that of the previous DBN.

1) Gauss-Bernoulli restricted boltzmann machine

In the simplest RBM, the explicit unit and the hidden unit are binary and random, and the values are only 0 and 1. For voice signal, it needs to have the ability to represent the probability distribution. In order to deal with the real input data, we use a Gaussian-Bernoulli RBM. The visible cells use Gaussian distribution and the hidden cells use Bernoulli distribution. Its energy function is defined as

$$E_\theta\left(v,h\right)=\sum_{i=1}^{n_v}\frac{\left(v_i-a_i\right)^2}{2{\sigma_i}^2}-\sum_{j=1}^{n_h}b_jh_j-\sum_{i=1}^{n_v}\sum_{j=1}^{n_h}h_jw_{ji}\frac{v_i}{\sigma_i}\tag{23}$$

where w_{ji} is the weight of the i-th neuron in the visual layer and the j-th neuron in the hidden layer. a_i is the bias value of neurons in the visual layer. b_j is the bias value of hidden layer neurons. σ_i is the variance of neurons in the visual layer. n_v and n_h are the number of neurons in the visual layer and the number of neurons in the hidden layer, respectively.

According to (23), the conditional probabilities of V and H are obtained as follows:

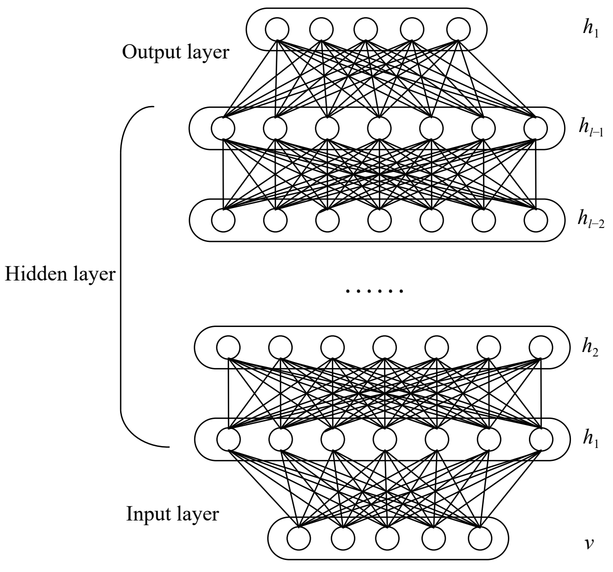


Fig. 8 Schematic diagram of deep belief network

$$p\left(v_i|h\right) = N\left(a_i + \sigma_i \sum_{j}^{n_h} W_{ji} h_j, {\sigma_i}^2\right) \tag{24}$$

$$p\left(h_j = 1|v\right) = \delta\left(\sum_{i=1}^{n_v} W_{ji} \frac{v_i}{\sigma_i} + b_j\right) \tag{25}$$

$$\delta\left(x\right) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{26}$$

where $N\left(\mu, \sigma\right)$ is a Gaussian distribution with mean μ and variance σ.

2) Softmax regression

A softmax classifier is often used in the field of deep learning, which is an extension of logistic regression classification. Logistic regression is a binary nonlinear classifier, which is extended to multiple classifiers by softmax. It takes the most posterior probability of all categories as the recognition object, so it is very suitable for the task of speaker recognition. After the training of the unsupervised restricted Boltzmann machine, the softmax classifier is added to the top level to classify the samples. The specific classification process is as follows:

$$S_i = SoftMax\left(f\right) = \frac{\mathrm{e}^{f_t}}{\sum_{i=1}^{d} \mathrm{e}^{f_t}} \tag{27}$$

where $f_\theta\left(x\right) = WX + b, \theta = \{W, b\}$. X is the neural unit of the input layer. W is the weight coefficient of the model. b is the offset of the model.

Assume that $t = [0,1]^d$ denotes the classification of the sample, when the i-th sample is correctly classified, $t_i = 1$, otherwise, $t_i = 0$. The form of cross entropy is used to calculate the loss function. As shown in (29),

$$J\left(t, S\right) = -\frac{1}{d}\left[\sum_{i=1}^{d} \left(t_i \log S_i + \left(1 - t_i\right) \log\left(1 - S_i\right)\right)\right]. \tag{28}$$

Adjusting the model parameter θ to minimize the loss function of (29).

$$\theta^* = \arg\min_{\theta} J\left(t, S\right). \tag{29}$$

The partial derivative of the model parameter θ can be obtained:

$$\frac{\partial J\left(t, S\right)}{\partial \theta} = -\frac{1}{d} \sum_{i=1}^{d} \left(t_i - S_i\right) \frac{\partial f_i}{\partial \theta}. \tag{30}$$

The gradient descent method was used to update the model parameter θ. The gradient descent method was used to update the model parameter θ as

$$\begin{cases} W' = W - \eta\left(\left(S - t\right)^{\mathrm{T}} X + \lambda W\right) \\ b' = b - \eta\left(S - t + \lambda b\right) \end{cases} \tag{31}$$

where λ is the weighting factor and η is the learning factor.

3) Layer-by-layer pretraining and fine tuning

The training of a deep network can be divided into two stages: pre-training and fine-tuning. Pre-training is to train each layer of the network in an unsupervised way. When one layer is trained, the parameters of all other layers remain unchanged, and the input of the next layer is the output of the previous layer. The fine-tuning is trained by the supervised BP algorithm until convergence after the parameters of all layers are determined. The depth belief network is based on these two steps to complete the training parameters, as shown in Fig. 9.

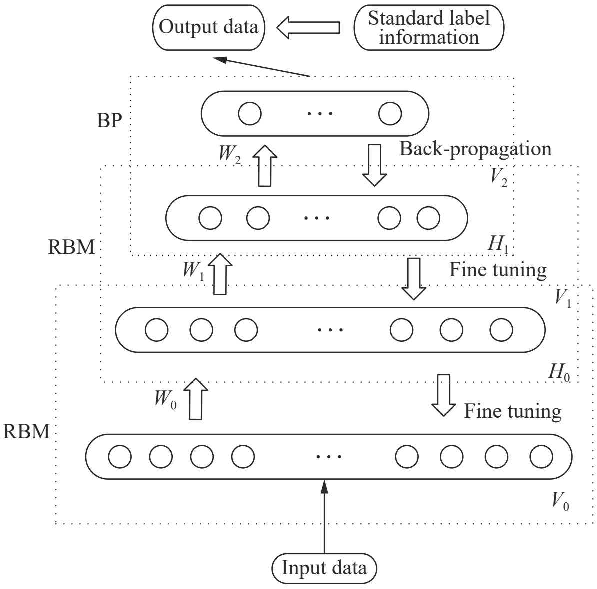


Fig. 9 DBN training

i) Greedy pre-training layer by layer

The k-layer RBM fast learning method based on the greedy algorithm is comprised of the following steps:

Take the training data X as input, i.e., $v_i = x$. The contrastive divergence algorithm (CD-k, k=1) is used to train the first RBM parameter θ_i, and calculate h_i.

$k = 2, 3, \cdots, K$. The hidden layer h_{k-1} of the last trained RBM is used as an input, i.e., $v_k = h_{k-1}$, and the parameter θ_k of the k-th RBM are trained.

After the parameters of all the K RBMs are obtained, the network parameters of the whole K layer can be obtained by adding $\theta_1, \theta_2, \cdots, \theta_k$, which is used as the initial value θ of the depth confidence network.

ii) BP reverse fine-tuning

The BP network is added to the last layer of the DBN network, and the output of the last layer RBM is used as its input. We can add tag information for supervised training. Because the parameters obtained by each layer RBM of unsupervised training can only ensure that the feature mapping in this layer is optimal, and cannot ensure that the feature mapping in the whole DBN is optimal, so supervised back propagation is used to transmit the deviation obtained from the normal label from top to bottom to each layer RBM, so as to realize the fine-tuning of the whole DBN. The pre-training process can be regarded as the parameter initialization process of the deep BP network. Compared with the traditional BP network, this method effectively solves the problem that the network falls into a local optimum due to the random initialization of parameters.

The non-output layer uses the sigmoid function as the activation function, and the parameter value is updated as (32).

$$a_j^l = \delta\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right) \tag{32}$$

where K means that there are K units in layer l–1, and w_{jk} means the weight of the k-th unit in layer l–1 and the j-th unit in layer l. To be written in matrix form as follows:

$$a^l = \delta\left(w^l a^{l-1} + b^l\right). \tag{33}$$

Equation (33) can be written as $a^l = \delta\left(z^l\right)$ if the intermediate quantity $w^l a^{l-1} + b^l$ is calculated and designated separately as z^l.

In order to find the error of the reverse transmission, it is assumed that the error of the j-th neural unit of the l-th layer is

$$\zeta_j^l = \frac{\partial J}{\partial z_j^l} \tag{34}$$

where J is the loss function of cross entropy.

L represents the last layer of the network (output layer). Since the last layer is a softmax layer, the error of the output layer is obtained according to (35).

$$\zeta_j^L = S_j - t_j. \tag{35}$$

Error of non-output layer is

$$\left\{\begin{aligned} &\zeta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \frac{\partial z_k^{l+1}}{\partial z_j^l} = \sum_k \zeta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_j^l} \\ &z_k^{l+1} = \left(\sum_i w_{ki}^{l+1} a_i^l\right) + b_k^{l+1} = \left(\sum_i w_{ki}^{l+1} \delta(z_i^l)\right) + b_k^{l+1} \end{aligned}\right. \Rightarrow$$

$$\left\{\begin{aligned} &\zeta_j^l = \sum_k \zeta_k^{l+1} \frac{\partial z_k^{l+1}}{\partial z_j^l} \\ &\frac{\partial z_k^{l+1}}{\partial z_j^l} = w_{kj}^{l+1} \delta'\left(z_j^l\right) \end{aligned}\right. \Rightarrow$$

$$\zeta_j^l = \sum_k \zeta_k^{l+1} w_{kj}^{l+1} \delta'\left(z_j^l\right). \tag{36}$$

Partial derivative of the loss function with respect to an arbitrary weight is

$$\left\{\begin{aligned} &\frac{\partial C}{\partial w_{jk}^l} = \sum_i \frac{\partial C}{\partial z_i^l} \frac{\partial z_i^l}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{jk}^l} \\ &z_j^l = \left(\sum_m w_{jm}^l a_m^{l-1}\right) + b_j^l \end{aligned}\right. \Rightarrow$$

$$\frac{\partial C}{\partial w_{jk}^l} = \frac{\partial C}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{jk}^l} = \zeta_j^l a_k^{l-1}. \tag{37}$$

Partial derivative of the loss function with respect to any offset is

$$\left\{\begin{aligned} &\frac{\partial C}{\partial b_j^l} = \sum_k \frac{\partial C}{\partial z_k^l} \frac{\partial z_k^l}{\partial b_j^l} = \frac{\partial C}{\partial z_j^l} \frac{\partial z_j^l}{\partial b_j^l} \\ &z_j^l = \left(\sum_k w_{jk}^l a_k^{l-1}\right) + b_j^l \end{aligned}\right. \Rightarrow$$

$$\frac{\partial C}{\partial b_j^l} = \zeta_j^l\ . \tag{38}$$

In this way, the output can be obtained from the input along the forward direction. We can solve the parameter differential in reverse direction and obtain the final network parameters by parameter optimization.

The training process of DBN can be completed through supervised pre-training and unsupervised fine-tuning:

According to the CD-1 algorithm, the first RBM network is trained through several iterations, and the weights and offsets of the first network are obtained.

The weight and bias of the first RBM are kept unchanged, and the output vector of the first RBM is used as the input vector of the second RBM.

The CD-1 algorithm is used to train the second RBM through multiple iterations, so that a superimposed RBM structure can be obtained, as shown in Fig. 10.

The above process is repeated until the RBM network of the last layer is reached. The optimal weights and parameters of each layer of RBM are taken as the initial parameters of the whole DBN network, and the softmax classifier and label are added. The output vector of the last layer of RBM is taken as the input vector of

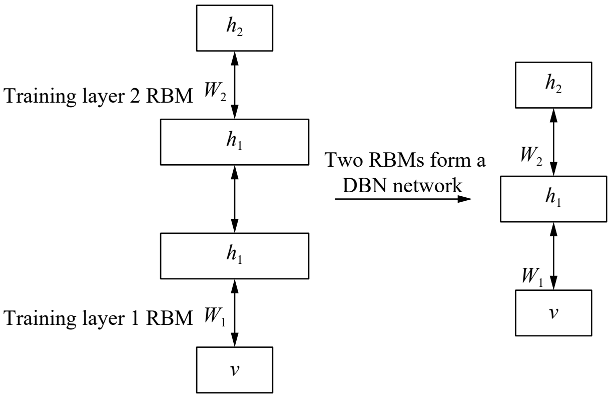


Fig. 10 RBM additive process

the softmax classifier, and then the error between the predicted label and the real label is calculated to fine-tune the reverse BP.

3.2 Improved model algorithm based on DBN

1) Model structure based on DBN

GMM is essentially a shallow network structure model, which uses the Gaussian probability density function to quantify things accurately and decomposes a thing into several models based on the Gaussian probability density function. That is to say, no matter how the distribution of the observation data sets and what rules they present, they can be fitted by a mixture of multiple single Gaussian models. When the number of samples is not enough, GMM usually uses the entire sample and feature information to predict, so it cannot fully describe the distribution of features and its ability to represent complex functions is limited. At the same time, if the number of features is more than dozens, the high-dimensional space model will be invalid. Meanwhile, GMM is modeled by likelihood, although discriminant training can simulate the discrimination between some sample classes, the discrimination ability is relatively limited.

As a deep network model, DBN has many advantages by simulating the mechanism of neurons in human brain for nonlinear learning.

i) DBN can construct a deep nonlinear network model and realize the approximation of complex function, and its generalization ability is relatively strong.

ii) DBN can reduce the number of hidden units by nonlinear transformations of the network, reducing the high dimension of feature representation to low dimension, effectively reducing the amount of calculation, and making the features more compact, get better details of features.

iii) All feature data share the same network structure, which is more conducive to extracting deeper features and enhancing the memory ability of the network.

iv) After supervised pre-training and unsupervised fine-tuning, DBN can not only build a multi-level generation model to discover the features themselves, but also adjust the boundaries of classes based on the limited amount of information in the tags.

With the deepening of research and the improvement of parallel computing ability, it is found that using more layers of a neural network has a better representation effect than a single layer network. A deep network has strong feature memory ability because it contains a lot of parameters, and the classification effect will be significantly improved by the model constructed by a deep network.

Lv[20] and Pan[21] have studied speaker recognition based on deep layer neural networks, but only a single feature is considered in feature selection, and the mixed feature considered in this paper is more discriminative. At the same time, the over-fitting problem has not been deeply discussed in the literature[20].

In order to improve the performance of voiceprint recognition for disguised speech, the GMM model in [18, 19] is replaced by the DBN model in deep learning. Model improvements to the voiceprint recognition system are shown in Fig. 11.

2) Dropout Strategy

Because the training sample data is limited, and the number of layers and neurons of the deep network model is large, it is easy to occur the phenomenon of over-fitting that the training set is very good, but the test set is not good. In order to suppress the over-fitting problem, we often regularize or reduce the network size based on L1 and L2. Scholar Hinton proposed that a part of the feature detectors can be stopped every time the samples are trained, which will make the generalization ability of the network stronger. Hinton called it dropout[22]. We choose to introduce a dropout strategy to suppress the over-fitting phenomenon. Fig. 12 shows a comparison of the network before and after the dropout policy is applied.

The image above is a visual representation of dropout, with the network on the left before using dropout, and the same network on the right already using dropout.

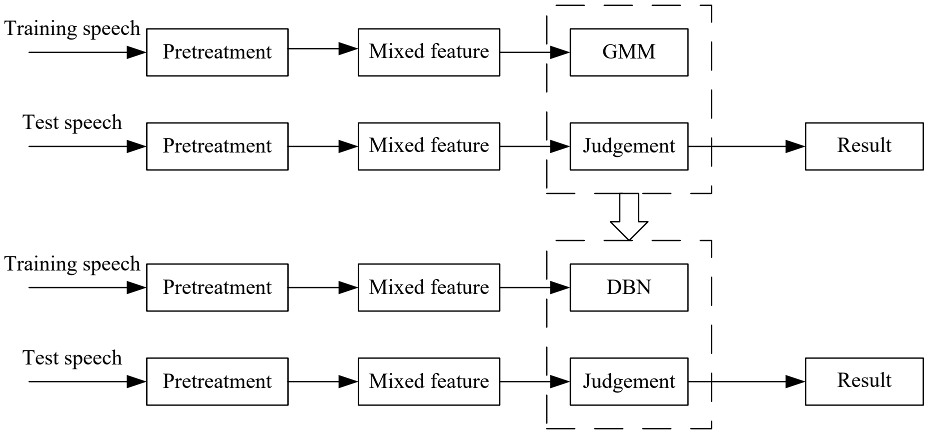


Fig. 11 Graph based model improvement

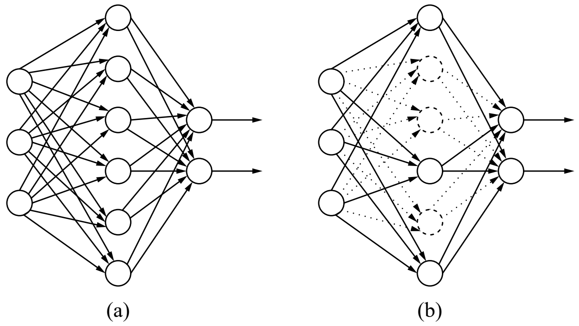


Fig. 12 Visual representation using dropout

Although dropout can achieve the effect of regularization, its principle is completely different from L1 and L2 regularization. L1 and L2 regularization are adjustment cost functions, while dropout is the adjustment of the depth network itself. The dropout network is composed of sub-networks which are formed by removing non-output units from the base network. A unit can be effectively deleted by multiplying some output units by zero.

There are some changes in the training and testing of deep neural networks with dropout strategy:

i) In the stage of training model, the ratio of dropout is set as p, i.e., the probability of a unit being abandoned is p, and the probability of being left is $1{-}p$. In the trained network, a probability step is added to each neuron, as shown in Fig. 13. The presence of offsets is not considered in the Fig. 13.

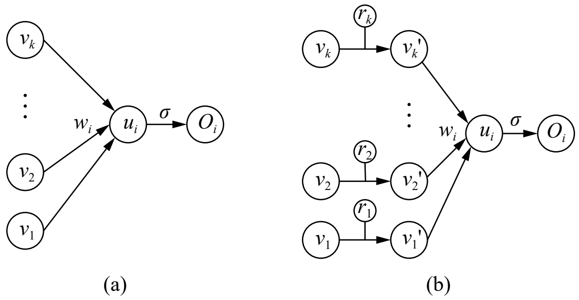


Fig. 13 Dropout network

The network without dropout is shown in Fig. 13(a) and is calculated as follows:

$$u_i = \sum\nolimits_i w_i v_i \tag{39}$$

$$O_i = \sigma\left(u_i\right) \tag{40}$$

where σ is the sigmoid function. The network with dropout is shown in Fig. 13(b) and is calculated as follows:

$$r_i \sim Bernoulli\left(p\right) \tag{41}$$

$$v_i{}' = r_i v_i \tag{42}$$

$$O_i = \sigma\left(\sum\nolimits_i w_i v_i{}'\right) \tag{43}$$

where r_i is a Bernoulli function, and $P\left(r_i = 0\right) = p$, generating a zero vector with probability p at random. It samples r_i and multiplies the input of that layer one by one to create fewer outputs. These outputs are then used as inputs to the next layer. This process is applied at each layer and is equivalent to sampling a subnetwork from the larger network.

ii) In the testing phase, the integrated network model of the training phase is simulated. The geometric mean of the ensemble members can be used to approximate the prediction of the whole ensemble, and only one forward propagation is needed as the cost.

The idea of dropout is actually to train the model to be optimized as an integrated model, and then average the output value, not just train the individual model. Thus, the output of the hidden layer unit may be expressed as

$$O = \frac{1}{1 + \mathrm{e}^{-u}} \tag{44}$$

where $u = \sum\nolimits_i w_i v_i$ is the linear combination of all input elements. Assigning the input units randomly with probability p, N different kinds of network structures will be obtained, and the non-standardized probability distribution directly defined by the geometric mean can be obtained by the following equation:

$$G\left(O\right) = \prod\nolimits_{n=1}^{N} {O_n}^{\frac{1}{N}} \tag{45}$$

where $G\left(O\right)$ represents the probability that the output unit O is activated, and we can also find the probability that the unit O is not activated, as shown in (47).

$$G'\left(O\right) = \prod\nolimits_{n=1}^{N} \left(1 - O_n\right)^{\frac{1}{N}}. \tag{46}$$

In order to obtain the model, the normalized geometric mean of the activation probability of the cell is derived from equations (45) and (46) as follows:

$$\begin{aligned} NGM\left(O\right) = & \frac{G\left(O\right)}{G\left(O\right) + G'\left(O\right)} = \\ & \frac{\prod_{n=1}^{N} \sigma(u_n)^{\frac{1}{N}}}{\prod_{n=1}^{N} \sigma(u_n)^{\frac{1}{N}} + \prod_{n=1}^{N} (1 - \sigma(u_n))^{\frac{1}{N}}} = \\ & \frac{1}{1 + \prod_{n=1}^{N} (\frac{1-\sigma(u_n)}{\sigma(u_n)})^{\frac{1}{N}}} = \\ & \frac{1}{1 + \exp\left(-\sum\nolimits_{n=1}^{N} \frac{1}{N} u_n\right)} = \\ & \sigma\left(\frac{1}{N} \sum\nolimits_{n=1}^{N} u_n\right) = \sigma(E(u)). \end{aligned} \tag{47}$$

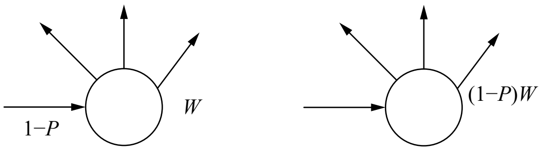
As can be seen from (47), the NGM value of the element O is equivalent to the desired nonlinear transformation after the input element is linearly weighted. Considering that the output of the first hidden layer before dropout is $u = \sum_i w_i v_i$, the expected value after dropout is

$$E(u) = \sum\nolimits_i (1-P)\, w_i v_i \tag{48}$$

where p is the discard rate, so (47) can be written as

$$NGM(O) = \frac{1}{1+\exp\left(-\sum_i (1-P)\, w_i v_i\right)}. \tag{49}$$

Since each neuron is present in the test phase, in order to maintain the same output expectations and get the same results for the next layer, the weights need to be multiplied by 1−p when testing, as shown in Fig. 14. Fig. 14(a) denotes the probability that a neuron exists at the training stage is 1−p and the weight of the next layer of neurons is W. Fig. 14(b) means that neurons are always present during that training phase. To ensure that the desired output is the same as the output of the training stage, the weight is multiplied by 1−p.



(a) Training stage (b) Test phase

Fig. 14 The weight transform of dropout network

In view of the above advantages, this chapter considers the introduction of dropout strategy to improve the system recognition rate. The improved model diagram with the dropout strategy is shown in Fig. 15.

3.3 Experiment and result analysis

1) Experimental design

The database used in the experiment is a self-built disguised voice database. Although the number of samples is limited, in order to reflect the strong modeling ability of deep belief network, the training samples and test samples are divided to get more samples. Each training data was divided into 50 pieces. Because the time of the fast voice is short, each test data is divided into 15 copies. The test data was divided into 20 parts under other disguised labels. The sample data after segmentation is shown in Table 2.

The voiceprint recognition system based on the DBN model can be divided into two parts: One is the training stage of speaker modeling. The other is the test phase of voice. Fig. 16 is a schematic diagram of speaker recognition based on a deep belief network.

In general, the state values of the visible layer unit and the hidden layer unit of RBM are 0 or 1. When applied to voice recognition, the generalization ability of RBM is severely limited. Therefore, by replacing the binary state of the dominant layer neurons of the first layer RBM with a Gaussian state, the first layer RBM becomes a Gauss-Bernoulli restricted Boltzmann machine. The other RBM layers use Bernoulli-Bernoulli-restricted Boltzmann machines.

In the training stage of voiceprint recognition, the voice sample is first preprocessed, then the mixed feature parameters are extracted, and then the mixed parameters are used as the input of the DBN acoustic model, as shown in Fig. 16(a) Then, Gibbs sampling and CD algorithm are used to train a single RBM layer. The optimal parameters of each layer are obtained by training layer by layer, and these parameters are used as the initial

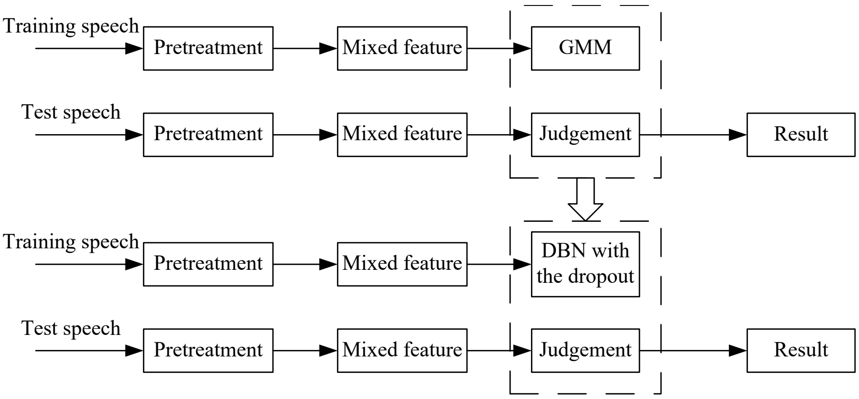


Fig. 15 Improved graph with dropout strategy

Table 2 Sample library of different disguised voices

	Fast	Slow	Treble	Bass	Whisper	Pinch nose	Bite pencil
Training data	2 450	2 450	2 450	2 450	2 450	2 450	2 450
Test data	735	980	980	980	980	980	980

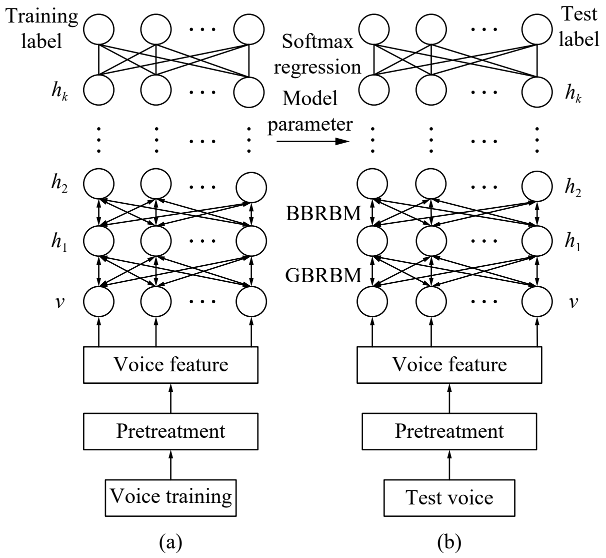


Fig. 16 Schematic of speaker recognition

parameters of the DBN acoustic model. When the unsupervised training is completed, the softmax classifier is added to the last layer of the DBN model for supervised fine-tuning.

In the recognition stage, as shown in Fig. 16(b), it is also necessary to preprocess the voice of the test set in the same manner and extract the mixed feature parameters. Taking the mixed feature parameters as the input vector of the trained DBN acoustic model, the label of the voice to be recognized can be obtained. The label obtained in the recognition stage is compared with the corresponding correct label. If the labels are the same, the result of recognition is correct, and if the result is wrong, the accuracy rate is calculated.

2) Experimental results and analysis

i) Network parameter setting

The number of input layer units of DBN is 780 (each frame corresponds to 39 dimensional mixed features, 20 frames, a total of 780 dimensions), and there are three hidden layers. The number of hidden layers is 400-200-100. The hidden layer uses the sigmoid activation function. The output layer is classified by the softmax function, and the cross entropy loss function is used.

ii) Network training

In this paper, the pre-training method is used to initialize the parameters of the network. The first RBM uses the Gauss-Bernoulli element, and the RBMs of the later layers are Bernoulli-Bernoulli elements. CD-k is 1, the number of iterations of RBM is 16, the number of iterations of DBN is 30, and the learning rate is 0.005.

For the DBN model with dropout policy, the method is the same as above, but the dropout policy is added when the model parameters are fine-tuned, and the discard rate is 0.2.

iii) Platform construction

The voiceprint recognition system is built on pycharm based on tensorflow framework. On this basis, the DBN model is established and the dropout strategy is added.

Fig. 17 shows the loss function values for DBN and each RBM layer. Fig. 17(a) is the loss function value of the DBN. Figs. 17(b)−17(d) are the loss function values of the first, second and third layer RBMs, respectively. The thin solid line in the Fig. 17 is the value of the loss function, and the thick solid line is the value of the loss function after smoothing. It can be seen that with the increase of the number of iterations, the loss function value gradually decreases from the overall point of view.

The recognition rate of the disguised voice speaker recognition system based on the DBN model for different disguised voice tags is shown in Table 3.

In order to clearly see the recognition rate of different acoustic models for different disguised voice tags, the form of a histogram is used, as shown in Fig. 18.

From the experimental data of Table 3 and Fig. 18, we can see that the classification effect of the DBN algorithm model on fast, bass and slow disguised voice tags on the disguised voice database is better, and the recognition rate is more than 90%, while the recognition effect on nose pinch and whisper disguised voice tags is relat-

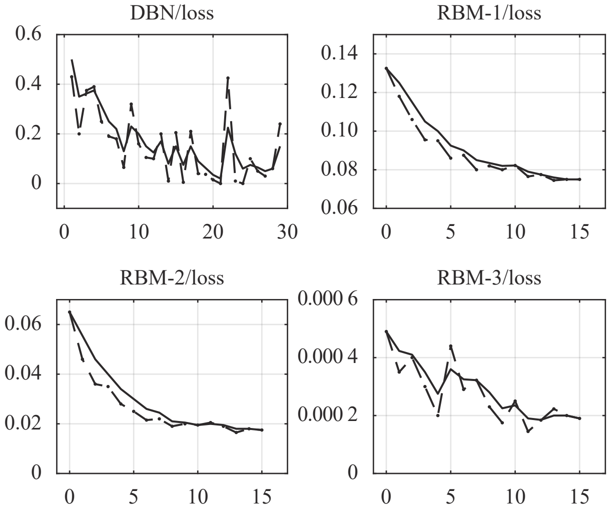


Fig. 17 Graph of loss function values

Table 3 System recognition rate of different disguised voices based on DBN

Disguised voice	GMM	DBN	DBN of adding dropout
Fast	95.92	96.87	97.01
Slow	93.88	95.82	95.92
Treble	67.35	76.02	76.84
Bass	89.80	90.71	91.22
Whisper	40.82	52.85	53.67
Pinch nose	55.10	61.33	61.84
Bite pencil	65.31	70.61	70.92

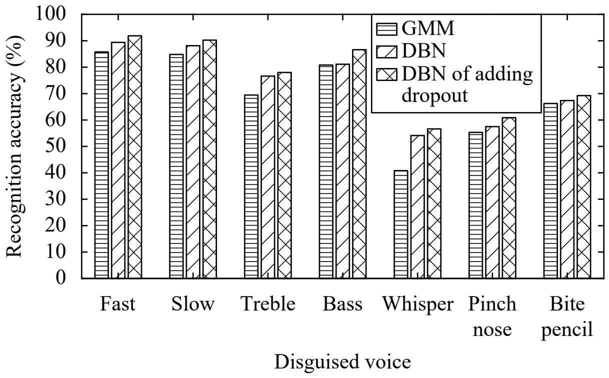


Fig. 18 Recognition rate of acoustic model

ively poor, but compared with the traditional GMM model[18,19], the recognition rate is improved.

Through the comparison of the experimental results, it can be seen that the system recognition rate of the method of using DBN with dropout strategy to establish acoustic model is the highest, and the method of using GMM to establish an acoustic model is the most unsatisfactory.

Compared with the shallow network such as GMM, the deep network such as DBN can describe the feature data in detail, mine the useful information, and its nonlinear modeling ability can express the original voice signal better. The special structure of DBN makes its modeling ability very outstanding. The parameters obtained by RBM pre-training in an unsupervised way can provide a good initial value for the model, and then the parameters of the network can be fine-tuned by the supervised back-propagation algorithm, so as to effectively solve the problem of local optimum.

DBN with dropout strategy has a better recognition rate, because it is equivalent to the role of regularization, Criminal Investigation Police University of China can prevent overfitting, and effectively increase the robustness of the neural network.

4 Conclusions

The proposed voiceprint recognition system based on mixed features can learn more representative voiceprint features from voice data, and the recognition rate is higher than the traditional GFCC features. Therefore, it is a very effective method to extract mixed features to solve the disguised voice speaker recognition. Based on the powerful nonlinear modeling function of the DBN, higher expression level features can be mined for classification. Compared with the traditional GMM model, the recognition rate of the model based on the DBN network is improved. And the dropout strategy can further improve the recognition accuracy. It is a very effective method to solve the problem of disguised voice speaker recognition by using deep learning from the level of speaker model building.

Acknowledgements

This work was supported by Natural Science Foundation of Liaoning Province (Nos. 2019-ZD-0168 and 2020-KF-12-11), Major Training Program of Criminal Investigation Police University of China (No. 3242019010), Key Research and Development Projects of Ministry of Science and Technology (No. 2017YFC0821005), and Second Batch of New Engineering Research and Practice Projects (No. E-AQGABQ20202710).

References

[1] Y. H. Zheng. Development and application strategy of voiceprint recognition technology. *Technology Wind*, no. 21, pp. 9–10, 2017. DOI: 10.19392/j.cnki.1671-7341.201721007. (in Chinese)

[2] Z. Lian, Y. Li, J. H. Tao, J. Huang, M. Y. Niu. Expression analysis based on face regions in real-world conditions. *International Journal of Automation and Computing*, vol. 17, no. 1, pp. 96–107, 2020. DOI: 10.1007/s11633-019-1176-9.

[3] T. Kinnunen, H. Z. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010. DOI: 10.1016/j.specom.2009.08.009.

[4] J. H. Tao, J. Huang, Y. Li, Z. Lian, M. Y. Niu. Semi-supervised ladder networks for speech emotion recognition. *International Journal of Automation and Computing*, vol. 16, no. 4, pp. 437–448, 2019. DOI: 10.1007/s11633-019-1175-x.

[5] C. L. Zhang. Acoustic Study of Disguised Voice, Ph. D. dissertation, Nankai University, China, 2005. (in Chinese)

[6] L. L. Stoll. Finding Difficult Speakers in Automatic Speaker Recognition, Ph. D. dissertation, University of California, USA, 2011.

[7] A. R. Reich. Detecting the presence of vocal disguise in the male voice. *The Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1458–1461, 1981. DOI: 10.1121/1.385778.

[8] H. Hollien, W. Majewski. Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 975–980, 1977. DOI: 10.1121/1.381592.

[9] X. H. Shen, T. Jin, C. Z. Zhang, R. C. Wan. Feasibility analysis on identification of disguised falsetto. *Journal of Criminal Investigation Police University of China*, no. 2, pp. 124–128, 2018. DOI: 10.14060/j.issn.2095-7939.2018.02.024. (in Chinese)

[10] Y. Matveev. The problem of voice template aging in speaker recognition systems. In *Proceedings of the 15th International Conference on Speech and Computer*, Springer, Pilsen, Czech Republic, pp. 169-175, 2013. DOI: 10.1007/978-3-319-01931-4_46.

[11] H. J. Wu, Y. Wang, J. W. Huang. Identification of electronic disguised voices. *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 489–500, 2014. DOI: 10.1109/TIFS.2014.2301912.

[12] Z. Z. Wu, A. Khodabakhsh, C. Demiroglu, J. Yamagishi, D. Saito, T. Toda, S. King. SAS: A speaker verification spoofing database containing diverse attacks. In *Proceed-*

ings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, South Brisbane, Australia, pp. 4440-4444, 2015. DOI: 10.1109/ICASSP.2015.7178810.

[13] Y. Wang, H. J. Wu, J. W. Huang. Verification of hidden speaker behind transformation disguised voices. *Digital Signal Processing*, vol. 45, pp. 84–95, 2015. DOI: 10.1016/j.dsp.2015.06.010.

[14] W. Zhang. Auditory recognition of disguised speech. *Science & Technology Vision*, no. 13, pp. 10–12, 2016. DOI: 10.3969/j.issn.2095-2457.2016.13.005. (in Chinese)

[15] Y. P. Li, L. Lin, D. Y. Tao. Research on identification of electronic disguised voice based on GMM statistical parameters. *Computer Technology and Development*, vol. 27, no. 1, pp. 103–106, 2017. (in Chinese)

[16] P. Zhou, H. Shen, K. P. Zheng. Speaker recognition based on combination of MFCC and GFCC feature parameters. *Journal of Applied Sciences*, vol. 37, no. 1, pp. 24–32, 2019. DOI: 10.3969/j.issn.0255-8297.2019.01.003. (in Chinese)

[17] K. P. Zheng. The Research of Voiceprint Recognition Method Based on MFCC and GFCC Mixed Cepstrum, Master dissertation, Guilin University of Electronic Technology, China, 2017. (in Chinese)

[18] J. Cao, P. Pan. Research on GMM based speaker recognition technology. *Computer Engineering and Applications*, vol. 47, no. 11, pp. 114–117, 2011. DOI: 10.3778/j.issn.1002-8331.2011.11.033. (in Chinese)

[19] X. Yu, S. He, Y. X. Peng, W. Zhou. Pattern matching of voiceprint recognition based on GMM. *Communications Technology*, vol. 48, no. 1, pp. 97–101, 2015. DOI: 10.3969/j.issn.1002-0802.2015.01.020. (in Chinese)

[20] L. Lv. Research on Speaker Recognition Based on Deep Learning, Master dissertation, Southeast University, China, 2016. (in Chinese)

[21] H. Pan. Design and Implementation of Speaker Recognition System Based on Deep Learning, Master dissertation, Heilongjiang University, China, 2016. (in Chinese)

[22] N. Srivastava, G. Hinton, A. Krizhevsky, A. Sutskever, R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014. DOI: 10.5555/2627435.2670313.

[23] Y. B. Xing, X. W. Zhang, C. Y. Zheng, T. Y. Cao. Establishment of bone-conducted speech database and mutual information analysis between bone and airconducted speeches. *Technical Acoustics*, vol. 38, no. 3, pp. 312–316, 2019. DOI: 10.16300/j.cnki.1000-3630.2019.03.013. (in Chinese)

[24] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle. Greedy layer-wise training of deep networks. In *Proceedings of the 19th International Conference on Neural Information Processing Systems*, Vancouver, Canada, pp. 153-160, 2006. DOI: 10.5555/2976456.2976476.

Nan Jiang received the Ph. D. degree in engineering from Northeastern University, China in 2007, and completed her postdoctoral research in computer science and technology at Northeastern University, China in 2012. She is currently an associate professor in Department of Acoustic and Imaging Information Inspection Technology, Criminal Investigation Police University of China. She is the young talents of criminal science and technology in China, and has been selected into the Millions of Talents Project in Liaoning Province. During her school time, she has been engaged in teaching, scientific research and case handling of judicial voice examination and voice recognition. She has published more than 40 academic papers.

Her research interests include pattern recognition, speech recognition, speech emotion recognition, and multimodal emotion recognition.

E-mail: zgxj_jiangnan@126.com

ORCID iD: 0000-0002-5497-7007

Ting Liu received the B. Sc. degree in automation from Northeastern University, China in 2007 and the M. Sc. and Ph. D. degrees in control theory and control engineering from Northeastern University, China in 2009 and 2014, respectively. She is now a lecturer in electrical engineering and automation at Liaoning University, China, and has been engaged in the research of nonlinear algorithms and speech recognition algorithms.

Her research interests include pattern recognition, feedback control systems, and control theory.

E-mail: 195361952@qq.com (Corresponding author)

ORCID iD: 0000-0001-9520-5081

IJAC International Journal of Automation & Computing

Citation: N. Jiang, T. Liu. Research on voiceprint recognition of camouflage voice based on deep belief network. *International Journal of Automation and Computing*, vol.18, no.6, pp.947–962, 2021. https://doi.org/10.1007/s11633–021–1283–2

Articles may interest you

Developing soft sensors for polymer melt index in an industrial polymerization process using deep belief networks. *International Journal of Automation and Computing*, vol.17, no.1, pp.44-54, 2020.
DOI: 10.1007/s11633-019-1203-x

Multi-layer contribution propagation analysis for fault diagnosis. *International Journal of Automation and Computing*, vol.16, no.1, pp.40-51, 2019.
DOI: 10.1007/s11633-018-1142-y

Deep learning based hand gesture recognition and uav flight controls. *International Journal of Automation and Computing*, vol.17, no.1, pp.17-29, 2020.
DOI: 10.1007/s11633-019-1194-7

Deep learning based single image super-resolution: a survey. *International Journal of Automation and Computing*, vol.16, no.4, pp.413-426, 2019.
DOI: 10.1007/s11633-019-1183-x

An advanced analysis system for identifying alcoholic brain state through eeg signals. *International Journal of Automation and Computing*, vol.16, no.6, pp.737-747, 2019.
DOI: 10.1007/s11633-019-1178-7

Dual-modal physiological feature fusion-based sleep recognition using cfs and rf algorithm. *International Journal of Automation and Computing*, vol.16, no.3, pp.286-296, 2019.
DOI: 10.1007/s11633-019-1171-1

A performance evaluation of classic convolutional neural networks for 2d and 3d palmprint and palm vein recognition. *International Journal of Automation and Computing*, vol.18, no.1, pp.18-44, 2021.
DOI: 10.1007/s11633-020-1257-9