Wireless Personal Communications
https://doi.org/10.1007/s11277-021-08773-w

Design of Voice Recognition Acoustic Compression System Based on Neural Network

Yuan Xiwen[1]

Accepted: 4 August 2021

Abstract
The deep neural network halts its application to mobile devices because of its high complexity. So, it motivates us to compress and accelerate the deep neural network model. In order to improve the operating effect of the voice recognition acoustic compression system, this paper improves on the traditional neural network, and stitches the transfer features of multiple deep convolutional neural networks together to obtain a more discriminative feature representation than the transfer learning features in a single convolutional neural network. Also to construct a voice recognition acoustic compression system based on deep convolutional neural networks, this paper combines the actual needs. After the system framework construction, the performance and recognition accuracy of the voice acoustic system are studied from two perspectives. The experimental findings show that the voice recognition acoustic compression system constructed in this paper has excellent performance. The voice data processing speed of the voice recognition acoustic compression model is 111(s) and the average accuracy is 94.1%.

Keywords Deep convolutional neural network · Voice recognition · Acoustic compression · Machine learning · Compressed sensing theory · Nyquist sampling theorem · GoogleNet

1 Introduction

According to the theorem adopted by Nyquist, it is known: "In order to restore the signal without distortion after sampling, the sampling frequency must be greater than twice the highest frequency of the signal" [1]. The Nyquist sampling theorem mention the signals that are periodically sampled two or three time to get the highest frequency component, the signal data will be lost if there is no frequency component. Also, the sampling frequency is always greater than analog signal frequency in Nyquist sampling theorem. It is used to convert analog frequency signal into digital frequency signal through sampling. Therefore, more bandwidth resources and data storage space are required. With the advent of the digital and information age, people's ever-increasing demand for information puts forward higher requirements on many aspects such as signal sampling, transmission and storage.

✉ Yuan Xiwen
yuanxiwen9703@163.com

[1] National University of Defense Technology, Hefei 230071, Anhui, China

It can be known from the compressed sensing theory that if the signal itself is sparse or has sparseness after some transformation, it can be observed and sampled at a standard far lower than the traditional Nyquist sampling theorem. At this time, the under-determined sampling observation sequence value can be obtained, and the original signal before processing can be restored almost accurately through the sampling sequence value [2]. The under-determined sampling observation sequence value is predicted when the compressed sensing theory signal is lesser than the traditional Nyquist theorem signal. It is obtained with the help of compressed sensing theory. Compressed sensing theory efficiently compresses the signal data while observing and sampling the signal, which greatly saves the data storage space and the amount of data sent [3].

Compressed sensing theory changes the traditional signal processing method, and does not need to meet the signal sampling frequency requirement of Nyquist theorem. In a specific transformation domain, the signal has a sparse representation. As a result, the compressed sensing principle is applied to recreate it with a limited number of acquire samples. The compressed sensing theory address the high scan time issue in measuring Fourier co-efficient. Moreover, the compressed sensing theory does not have the same limitations as the Nyquist sampling theorem. As long as the signal itself is sparse or has sparseness after some transformation, the signal can be compressed at the same time of observation and sampling and the information required in signal compression can be collected, and the observation sampling sequence with all the information details can be directly obtained. Using nonlinear optimization theory to restore and reconstruct the collected sampling sequence, the original signal can be restored and reconstructed approximately accurately. When compressing signals through the compressed sensing theory, there is no need for high-speed measurement rate, high-speed equipment support, and large-capacity storage equipment. Due to the above advantages, the compressed sensing theory has attracted much attention after it was proposed. The proposal of compressed sensing theory has produced breakthrough changes in many existing fields. The significance of voice signal compression is to remove redundancy in voice information and reduce storage [4]. The neural network is the algorithm which identify the relationship between data set through processing. It is used to for various applications such as risk management, sales forecasting and data validation, etc. By using neural networks, the computers are trained to perform task by analyzing datasets. Also, the voice recognition is nothing but the voice signal is taken as an input for computer program. Also, Countless patterns and elements are identified using voice recognition.

2 Related Work

The only prerequisite for applying compressed sensing theory is the sparsity of the signal. If the signal contain large number of non-zero elements is said to be sparse signal. Then the compressed sensing theory is used to solve the issues regarding this problem. Compressed sensing is one of the signal processing technique which reconstruct the signal in order to predict the underdetermined linear systems, but it recovers the data under two conditions i.e., sparsity and incoherence. Orthogonal transformation is the first to use signal sparse representation of the transformation, but not all signals after orthogonal transformation are sparse [5]. The choice of basis function will directly affect the sparseness of the signal. In response to this problem, the literature [6] proposed a signal transformation theory based on an orthogonal basis dictionary to make the transformed signal sparse effect better.

When the over-complete dictionary library is used as the signal sparse transformation base and the frame theory are successively applied to the sparse representation of the signal, the sparse effect of the signal is better [7]. In the selection of observation matrix, the initial choice is random matrix, such as the common random Gaussian matrix, which is widely used in the measurement sampling of compressed sensing [8]. The literature [9] proposed that the measurement matrix should have a certain degree of linear independence and characteristic random properties.

The literature [10] proposed the characteristic of finite equidistance constraint and used it as a criterion for the selection of measurement matrix. In addition, other matrices satisfying the properties of RIP, such as random Bernoulli matrices, partial Fourier matrices, and partial Hadamard matrices, can also be used in compressed sensing observation sampling. The Restricted isometry property (RIP) is used to characterize the matrices which are operating in the sparse vectors and used to prove many theorems in the compressed sensing field. Also, RIP having certain features such as the networks are updated periodically. The literature [11] proposed an adaptive observation matrix to fully improve the utilization of the number of observation points. Under the condition that the reconstruction error remains unchanged, it can greatly reduce the total number of observation points required and improve the efficiency of signal compression. In terms of signal reconstruction and recovery, it is necessary to reconstruct the signal before compression from the under-determined sampling sequence. Literature [12] pointed out that signal reconstruction is a linear optimization problem for solving minimum $l0$ norm. Therefore, its sub-optimal solution is often sought in applications to approximate and accurately reconstruct the signal. At present, commonly used reconstruction algorithms include greedy pursuit algorithm, convex relaxation method, lp optimization algorithm, combination algorithm, etc. [13].

At present, voice compression coding technology has been rapidly progressed and widely used, and it plays an indispensable role in mobile phone communication, satellite phone communication and network communication. Voice compression coding technology has a history of more than 60 years. In particular, with the recent rise of computer technology and microelectronics technology, the progress of voice compression coding technology has become more eye-catching [14]. In 1972, the International Telephone and Telegraph Consultative Committee (CCITT) identified 64 kb/s PCM as the first voice compression coding standard (G.711). This kind of encoding obtains a relatively high voice synthesis effect at the expense of a large amount of bandwidth resources. CCITT began to explore lower-rate encoding methods after 1980 and achieved success in four years. It determined the 32 kb/sADPCM voice compression coding standard (G.721), which not only has higher voice synthesis quality but also lower bit error rate [15]. After that, the 16 kb/s low-delay code excitation linear prediction (LD-CELP) standard (G.728) came out. It has low speed and high performance, so it has been widely used. Then, the 8 kb/s voice compression coding standard (G.729) of the conjugate structure-algebraic code excited linear prediction (CS-ACELP) was proposed. It has low latency, less bandwidth, and high synthesized voice quality [16].

With the rise of compressed sensing, the research on voice compression coding has a new goal. Combining the advantages of compressed sensing theory and the characteristics of voice signals, a lot of research has been done in the direction of combining voice compression and compressed sensing at home and abroad and a series of results have been achieved. The literature [17] successfully applied CS theory to voice compression coding technology. The literature [18] conducted in-depth research on the sparseness of voice signals based on the excitation vocal tract model and combined with the theory of compressed sensing to make a breakthrough in the processing of sparse excitation signals. The

literature [19] combined the theoretical basis of compressed sensing to put forward the selection criteria for the observation matrix suitable for modeling. It is a long-term and arduous task to achieve high-efficiency compression of voice signals by combining compressed sensing theory with voice compression.

3 Deep Convolutional Neural Network

Convolutional Neural Network (CNN) is a deep learning method mainly used for image recognition and classification developed on the basis of multilayer neural network. It is one of the well-known neural networks. Convolution neural network is one the type of deep neural network, which is mostly used for computer vision applications and image classification. It extracts the features from the input image. In that, the classifier is used to classify different image contents from the set of image contents. The sampling layer is also known to be pooling layer, in which the size of the feature map is reduced without significant image information. And, the convolution layer is used as a filter to extract the features from the image for activation function. Moreover, its structure is shown in Fig. 1 [20].

It has achieved great success in image recognition through supervised training and backpropagation algorithms. The Back propagation is a supervised learning technique for multilayer feed-forward networks in the field of Artificial Neural Networks. Feed-forward neural networks are motivated by the information transmission of one or more neural cells called neurons. It is used to calculate the error function of the network with respect to neural network weights. Then, the supervised training algorithm examine the trained data for mapping. The accuracy of the algorithm is measured through error and loss function.

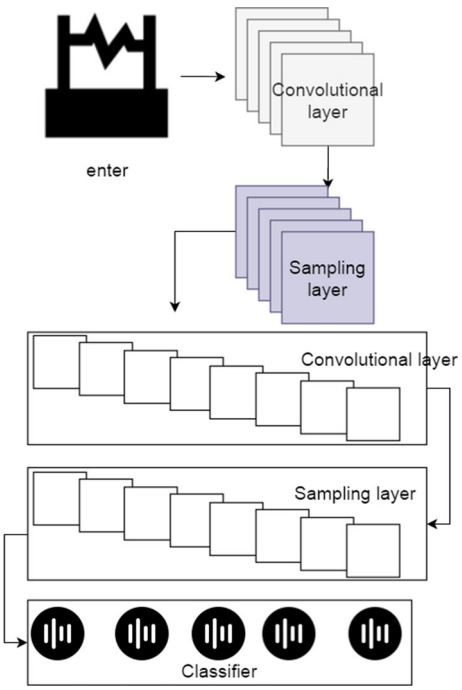


Fig. 1 Structure diagram of convolutional neural network

Therefore, many researchers have applied convolutional neural networks to fields such as voice recognition and text classification, and have made certain breakthroughs.

3.1 Inception Network

The Inception series of networks originated from Google's GoogleNet, which adopted the design idea of Network in Network and proposed a modular Inception network structure. The GoogleNet is one of the convolutional neural network model with deep 22 layers. It is used to perform new task in transfer learning. In GoogleNet the pertained network is loaded in different dataset such as Places365 and ImageNet. Inception v3 is used to achieve the accuracy greater than 78.1% of ImageNet dataset and it is one of the most widely used image recognition model. But the Inception v4 model is used to simplify the network architecture using more inception modules.

As shown in Fig. 2, it is the most basic Inception structure. In order to make the dense parameters approximate the optimal sparse structure, it usually contains 3 convolutional layers of different sizes and a maximum pooling layer, which are stitched at the end. Due to such a special structure, the network can perceive local image areas of different sizes in one layer, and integrate multi-scale features. The size of the convolution kernel is $1 \times 1, 3 \times 3$ and 5×5 respectively. However, since the convolution of 5×5 will still generate a lot of calculations, 1×1 convolution kernels are added to the Inception structure to reduce dimensionality, as shown in Fig. 3. The convolution kernel of 1×1 can not only merge the features extracted from convolutions of different sizes, increase the nonlinearity of the network, but also control the channel size of the feature map and reduce the amount of network calculations. Design modules in Convolutional Neural Networks are used to make more powerful processing and deeper networks feasible by reducing dimension, with a 1-to-1 stack. The modules were designed to address the computer costs dilemma and to override other problems. Therefore, such a structure can

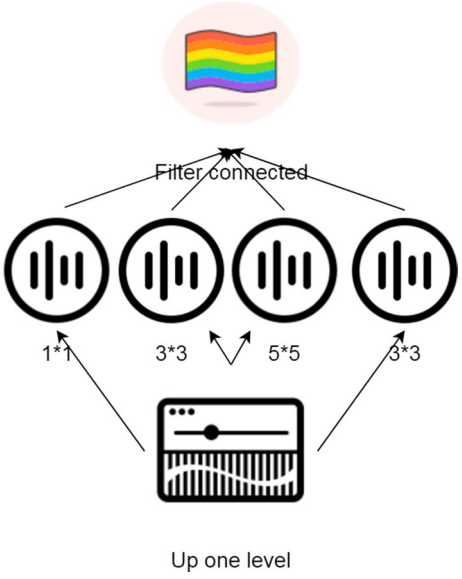


Fig. 2 Basic structure of inception

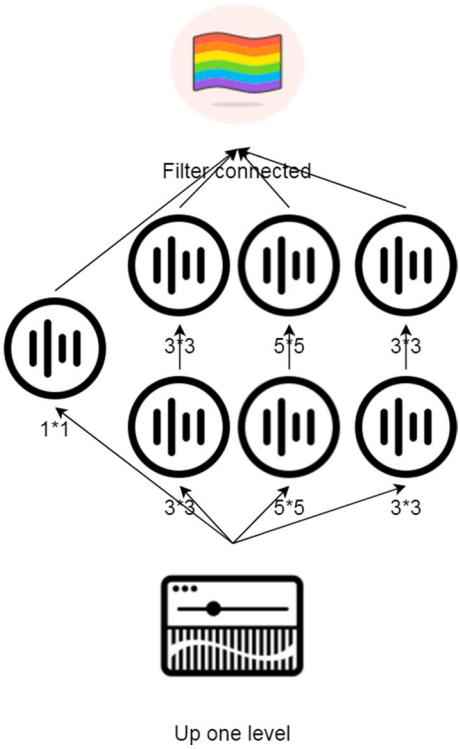


Fig. 3 Inception structure with 1×1 convolution kernel

greatly improve the expressive ability of the network. The experimental results prove that it cannot only expand the width and depth of the network, but also prevent over fitting and improve the performance of the network [21].

Inception-v3 proposes an asymmetric split large convolution method based on the above. That is, 3×3 convolution is split into 3×1 convolution and 1×3 convolution instead of stacking two 2×2 convolutions, as shown in Fig. 4. Such a structure can further reduce the amount of parameters, and the parameter of the former is $(3 \times 1 + 1 \times 3)/3 \times 3 = 66.7\%$ of the latter. It also reduces the classification layer introduced in Inception-v1, and adds a more complex pooling layer. The 1*1 convolution kernel is used as a filter to extract the features from the Inception-v3 structure. The kernel is a matrix that converts the data, to run the product using the input data subregions in order to collects the output as the dot product matrix. In addition, a method of smoothing labels is also proposed, that is, to constrain the classification labels, so that the Inception-v3 network has good generalization ability [22].

Finally, Inception-v3 optimized the second half of the network and adopted a parallel dimensionality reduction structure. The dimensionality reduction structure of the feature maps between 35/17/8 is used to improve the basis of Inception-v3 and v4. Specifically, the two structures shown in Fig. 5 are used between 35/17/8 to reduce the size of the feature map.

Inception-v4 is an improvement on the basis of Inception-v3, and adopts a short-circuit connection method, which further increases the width and depth of the network

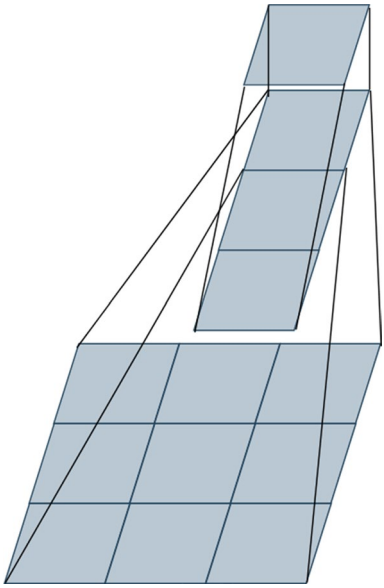
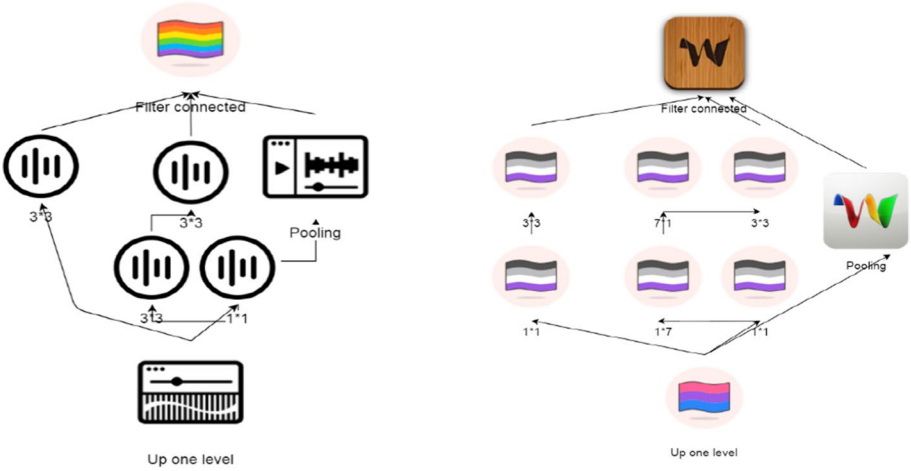


Fig. 4 3×3 convolution is split into 3×1 convolution and 1×3 convolution



(a) The dimensionality reduction structure of the feature maps between 35/17

(b) The dimensionality reduction structure of feature maps between 17/8

Fig. 5 The dimensionality reduction structure of feature map

to achieve better learning accuracy [23]. The short-circuit connection method are used to improve the gradient of the deep learning neural network.

3.2 ResNet Network

As the depth of the network increases, the characterization capabilities of the network will increase. However, in experiments, it is found that this is not necessarily the case [24].

When the network reaches a certain depth, continuing to increase the number of layers of the network will not improve the performance of the network, but will lead to degradation. The Residual Network (ResNet) is one of the standard neural network. It is mostly used as a backbone for many computer networks. Also, the ResNet is allowed to train deep neural networks with 150 layers.

This structure is designed to fit the residual through several stacked convolutions [25]:

$$F(x) := H(x) - x \tag{1}$$

That is, for input x, the original target output is $H(x)$. However, when the input x is re-added to the output through a shortcut connection, the learning goal becomes:

$$F(x) = H(x) - x \tag{2}$$

Therefore, the network no longer needs to learn the entire output but only needs to learn the difference between the output and the input, the "residual". This makes it easier to train the network, approximates the original mapping more realistically, and makes it possible to train ultra-deep networks.

4 Voice Recognition Model

4.1 Linear Prediction Cepstrum Coefficient

Linear prediction cepstral coefficient (LPCC) is based on linear prediction coefficient (LPC) to obtain its representation in the cepstrum domain. Linear predictive coding (LPC) is the method, which is used to represent the digital data in a compressed form with the help of speech processing and audio signal processing. It is one of the powerful speech analysis techniques with low bit rate and good quality. Its advantage is that it can reduce the influence of irrelevant excitation information generated during the sound signal generation process, and its calculation process is very simple, and a set of accurate parameters can be obtained to characterize the sound signal through a small amount of calculation [26]. The Linear prediction cepstral coefficient (LPCC) are determined from the linear prediction co-efficient (LPC). Cepstral analysis is widely used in speech synthesis because of its ability to symbolize speech waveforms perfectly and functions with a small size of features.

When LPCC is extracted, the LPC parameters need to be calculated first. The basic idea is to predict the acoustic signal samples at this moment by calculating the p acoustic signal sample points before a certain moment, and then the p samples are linearly combined to represent. The impulse response having impulse function with same time of continuity but the step response having discontinuity. Also, in impulse response the output of the system is generated when the input is impulse and the input in step response is taken as step. The uttered voice of the acoustic signal can be described by all-pole voice, and its transfer function $H(z)$ is:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3}$$

where$c(n)$—impulse response of the system,$\hat{H}(n)$—complex cepstrum.

If we assume that the impulse response of the system is $c(n)$ and the complex cepstrum is $\hat{H}(n)$, then the following formula exists:

$$\hat{H}(z) = \lg H(z) = \sum_{n=0}^{\infty} c(n) z^{-n} \tag{4}$$

where $H(z)$—transfer function expression, z^{-n}—partial derivative.

When the transfer function expression $H(z)$ is substituted and the partial derivative is solved for z^{-n}, we can get [27]:

$$\left(1 - \sum_{k=1}^{p} a_k z^{-k}\right) \sum_{n=1}^{\infty} n c(n) z^{-n+1} = \sum_{k=1}^{p} k a_k z^{-k+1} \tag{5}$$

When the above formula is expanded and calculated, the recurrence relationship between c_n and a_n can be obtained, namely:

$$\begin{cases} c_1 = a_1 \\ c_n = a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k}, 1 < n \leqslant p \\ c_n = a_n + \sum_{k=1}^{p} \frac{k}{n} c_k a_{n-k}, n > p \end{cases} \tag{6}$$

Among them, $c_1, c_2, \ldots, c_n$ is the cepstrum coefficient.

4.2 Mel Frequency Cepstrum Coefficient

Mel Frequency Cepstral Coefficient (MFCC) is a feature extraction algorithm designed based on the auditory characteristics of the human ear and the sounding principle of voice. The human ear has different sensitivity to sounds of different frequencies and presents a non-linear relationship. Generally, the human ear can better distinguish low-frequency sound signals, and the higher the frequency of the signal, the worse its perception ability. According to this characteristic, the MFCC characteristic is proposed, and the Mel filter is used to approximate the hearing characteristics of the human ear. Like LPCC, the Mel Frequency Cepstral Coefficient (MFCC) is derived from MFC (Mel-frequency cepstrum). It is also derived from a type of cepstral representation of audio clip. The bidirectional recurrent neural network (BRNN) with self-organizing map (SOM)-based classification scheme is proposed for automatic Tamil speech recognition. In which the input speech signal is pre-processed to improve the signal strength with the help of savitzky-Golay filter and the classification accuracy are enhanced with the help of perceptual linear predictive co-efficient (LPC) [28]. The conversion relationship between the Mel frequency and the time domain frequency is as follows:

$$Mel(f) = 2595 \lg (1 + f/700) \tag{7}$$

Among them, f is the actual frequency and $Mel(f)$ is the Mel frequency.

The MFCC feature is obtained by designing a set of triangular filters from dense to sparse, high to low frequency to filter the acoustic signal and through logarithm and DCT transformation. The energy power spectrum is the graph which is used to examine the co-efficients of each and every frequency by FFT power values.

For acoustic signal $x(n)$, the calculation process of extracting MFCC is as follows:

(1) First, the acoustic signal is preprocessed;
(2) Then, the fast Fourier transform (FFT) is performed to obtain the frequency domain information of the acoustic signal, and the calculation formula is as follows:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad 0 < k \leqslant N-1 \tag{8}$$

Among them, N is the number of FFT transform points. Then, the short-term energy spectrum $P(f)$ is:

$$P(f) = |X(k)|^2 = \left\{ [ReX(k)]^2 + [ImX(k)]^2 \right\} \tag{9}$$

(3) $P(f)$ passes through the Mel triangle filter bank, and the short-term energy spectrum is converted to the energy spectrum of the Mel frequency. The transfer function of the m-th filter is set to:

$$H_m(k), m = 1, 2, \ldots, M \tag{10}$$

Among them, M is the number of Mel filters, and the transfer function of each Mel triangle filter is as follows:

$$H_m(k) = \begin{cases} 0, k < f[m-1] \\ \dfrac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])}, \\ f[m-1] \leqslant k \leqslant f[m] \\ \dfrac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m] - f[m-1])}, \\ k > f[m+1] \end{cases} \tag{11}$$

Among them, $f[m]$ is the center frequency of the Mel filter.
(4) The Mel frequency spectrum is subjected to logarithmic operation to obtain a set of logarithmic frequency spectrum coefficients$S(m)$:

$$S(m) = In\left[\sum_{k=0}^{N-1} |X(k)|^2 H_m(k)\right] \quad m = 1, 2, \ldots, M \tag{12}$$

(5) Finally, the logarithmic spectrum coefficients obtained above are subjected to discrete cosine transform (DCT), that is, the MFCC coefficients can be obtained:

$$C(n) = \sqrt{\frac{2}{N}} \sum_{m=0}^{M-1} S(m) \cos\left[\frac{\pi n(m+0.5)}{M}\right] \quad n = 1, 2, \ldots, M \tag{13}$$

5 Common Classification Algorithms

The classification algorithm is used to predict and extract the important data for future trends in machine learning. Support vector machine (SVM) is one of the learning algorithm, which can be used for both classification and regression. And it used to determine the datasets with high plane with two classes. The SVM are effective in great dimensional spaces and its dimensions are also superior to samples.

5.1 Support Vector Machine

As shown in Fig. 6, it is a simple two-classification problem. The figure contains two types of sample data. We assume that the data sample is $(x_i, y_i), i = 1, 2, \ldots, N$, N is the number of samples, and the feature vector is $x_i \in R^n$. The category label is $y_i, y_i \in \{-1, 1\}$, representing two different categories. Then, it is necessary to find a classification surface $w^T x + b = 0$ that can distinguish the data of the two types of samples and maximize the interval between the two types of samples, namely:

$$\begin{aligned} & y_i\left(w^T x_i + b\right) \geqslant 1 \\ & i = 1, 2, \ldots, N \end{aligned} \tag{14}$$

Among them, w is the weight vector and b is the offset. Then, the above conditions can be combined into Eq. 14:

The classification interval of the two types of samples is $d = 2/\|w\|$. To make the maximum classification interval distance d is equivalent to finding the minimum $\|w\|$, that is, finding the minimum value of $\|w\|^2$ or $\|w\|^2/2$. Then, the solution equation of the optimal hyperplane can be transformed into an optimal problem solution with constraints.

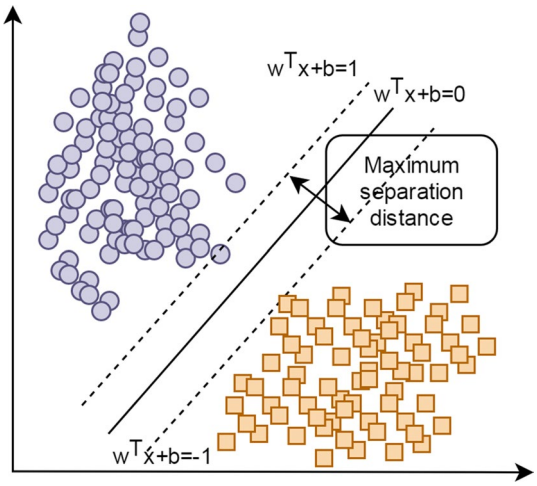


Fig. 6 Simple SVM two classification problem

$$\begin{cases} \min \frac{1}{2}\|w\|^2 \\ s.t. y_i\left(w^T x_i + b\right) \geqslant 1, \\ i = 1, 2, \ldots, N \end{cases} \tag{15}$$

Lagrangian is used to solve the minimum value:

$$L(w, a, b) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} a_i \left[y_i\left(wx_i + b\right) - 1\right] \tag{16}$$

Among them, $a_i \geqslant 0$ is the Lagrangian coefficient. We respectively find the partial derivatives of w and b for $L(w, a, b)$ and make them equal to 0, and then according to the KKT complementary condition, we obtain:

$$\begin{cases} w^* = \sum_{i=1}^{n} a_i^* y_i x_i \\ b^* = y_j - \sum_{i=1}^{n} a_i^* y_i x_i x_j \end{cases} \tag{17}$$

Among them, w^* is the optimal weight and b^* is the optimal bias.

5.2 Over-limit Learning Machine

An over-limit learning machine algorithms are also known as extreme learning machine (ELM) algorithm. The extreme machine learning is used for wide range of applications such as incremental learning, sequential learning and batch learning due to its fast convergence and effortlessness implementation. The over-limit learning machine (ELM) is a generalized single hidden layer feedforward neural network structure composed of input layer, hidden layer and output layer. The voice structure of ELM is shown in Fig. 7. The number of neurons in the input layer is N, and the number of neurons in the hidden layer is L. For ELM, the weights and biases of the neurons in the input layer are randomly initialized, and then the parameters of the output layer are obtained through the idea of least squares, and finally the optimal value of the network weight is solved through the generalized inverse matrix.

For any N data samples $\left(x_j, o_j\right)$, $x_j = \left[x_{j1}, x_{j2}, \ldots, x_{jn}\right]^T \in R^n$, the expected output is $y_j = \left[y_{j1}, y_{j2}, \ldots, y_{jn}\right]$, and the actual output is $o_j = \left[o_{j1}, o_{j2}, \ldots, o_{jn}\right]$. Then, the goal of the neural network is to minimize the difference between the expected output and the actual output:

$$\sum_{i=1}^{N} \left\|o_i - y_i\right\| = 0 \tag{18}$$

That is, there are α_i, b_i and β_i to make the following formula hold:

$$\begin{aligned} y_j &= \sum_{i=1}^{L} \beta_i g\left(\alpha_i, b_i, x_j\right) \\ j &= 1, 2, \ldots, N \end{aligned} \tag{19}$$

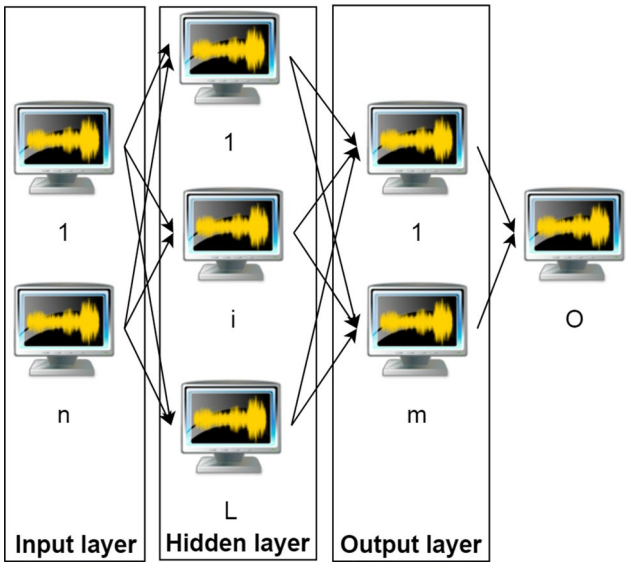


Fig. 7 Three-layer structure of the over-limit learning machine

Among them, $\alpha_i = \left[\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{in}\right]^T$ is the connection weight between the i-th hidden layer node and the input layer, b_i is the bias of the i-th hidden layer node, $\beta_i = \left[\beta_{i1}, \beta_{i2}, \ldots, \beta_{iL}\right]^T$ is the weight between the i-th hidden layer node and the output layer, and $g(x)$ is the activation function of the network. After that, we introduce a matrix for representation, namely:

$$H\beta = T \tag{20}$$

Finally, we obtain the following results by using the least square method:

$$\hat{\beta} = H^{+}T \tag{21}$$

Among them, H^{+} is the generalized inverse matrix of H.

6 Voice Recognition Acoustic Compression System Test

First, we define the K of the high-frequency coefficients of each level, and then adaptively select the number of decomposition stages and the number of observations in each level according to the sparsity of the high-frequency coefficients of the specific level, so as to ensure good reconstruction performance under the premise of sparse signals. We propose an adaptive multi-level wavelet compressed sensing algorithm AMCS, which performs adaptive multi-level decomposition and compression at the originating end (Fig. 8), and the receiving end is exactly the same as MCS. An adaptive multi-level wavelet compressed sensing algorithm (AMCS) is used to perform compression and multi-level decomposition and compression of the signals.

Voice recognition acoustic compression system model are shown in Fig. 8. First the voice signal is taken as an input, then the input voice signal is processed to the windowed

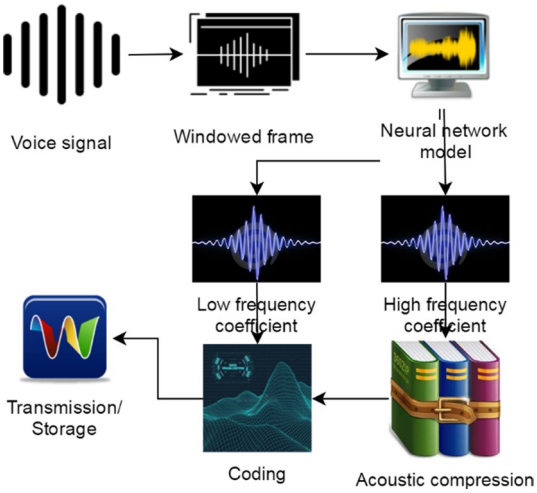


Fig. 8 Voice recognition acoustic compression system model

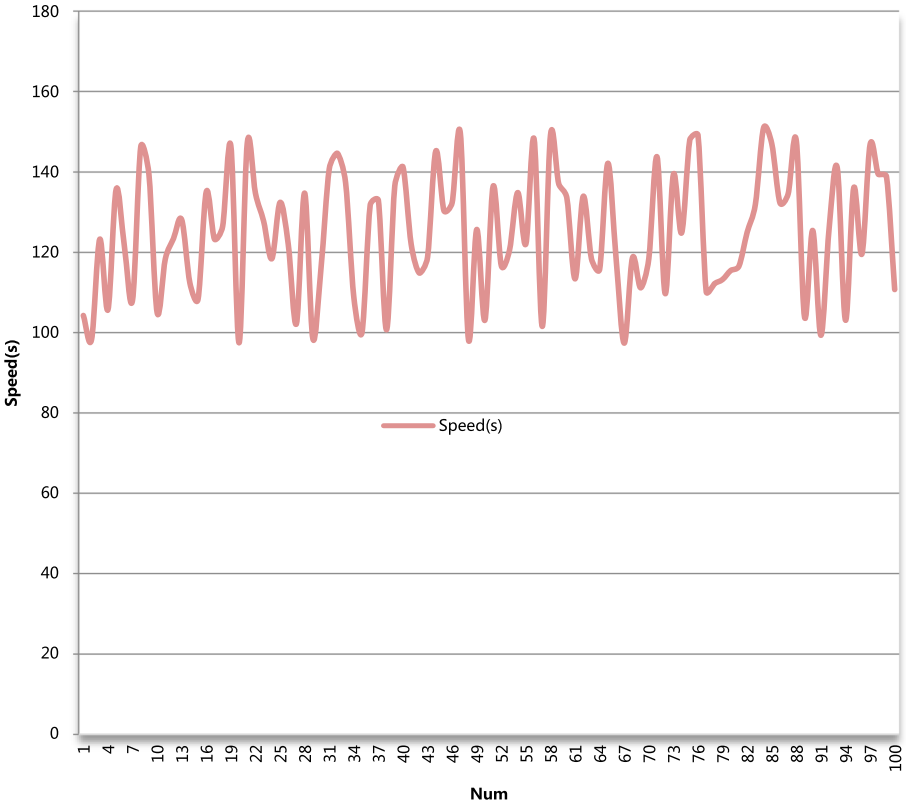


Fig. 9 Statistical table of the accuracy of voice recognition of the voice recognition acoustic compression model

Table 1 Statistical table of the voice data processing speed of the voice recognition acoustic compression model

Num	Speed (s)	Num	Speed (s)	Num	Speed (s)	Num	Speed (s)
1	104	26	122	51	136	76	149
2	98	27	102	52	117	77	110
3	123	28	135	53	121	78	112
4	106	29	99	54	135	79	113
5	136	30	116	55	122	80	116
6	122	31	141	56	148	81	117
7	108	32	145	57	102	82	125
8	146	33	137	58	149	83	132
9	139	34	109	59	137	84	151
10	105	35	100	60	134	85	147
11	118	36	132	61	113	86	132
12	123	37	133	62	134	87	135
13	128	38	101	63	118	88	148
14	112	39	137	64	116	89	104
15	108	40	141	65	142	90	125
16	135	41	122	66	119	91	99
17	123	42	115	67	97	92	126
18	127	43	119	68	119	93	141
19	147	44	145	69	111	94	103
20	97	45	130	70	118	95	136
21	147	46	132	71	144	96	120
22	135	47	150	72	110	97	147
23	128	48	98	73	139	98	139
24	118	49	126	74	125	99	139
25	132	50	103	75	148	100	111

frame for feature extraction and the signal are transmitted to neural network model in which the input voice signal is recognized into two types i.e., low frequency coefficient and high frequency coefficient. Then the low frequency signals are directly processed for coding but the high frequency co-efficient are processed in Acoustic compression to compress and manipulate the signal to generate output. The compressed signal is transmitted to coding and finally the output is stored to process transmission.

The voice recognition acoustic compression model constructed in this paper needs to have high data processing efficiency and has certain requirements on the recognition accuracy of voice acoustics. Therefore, the model in this paper is mainly studied from the two perspectives of voice acoustic system performance and recognition accuracy. First, the data processing speed of the acoustic compression model for voice recognition is calculated. The recognition speed statistics are performed by inputting 100 sets of data, and the results are shown in Table 1 and Fig. 9.

From the research results shown in Table 1 and Fig. 10, the model constructed in this paper performs well in the speed of data processing. Next, the accuracy of the model's voice acoustic recognition is calculated. The results are shown in Table 2 and Fig. 9.

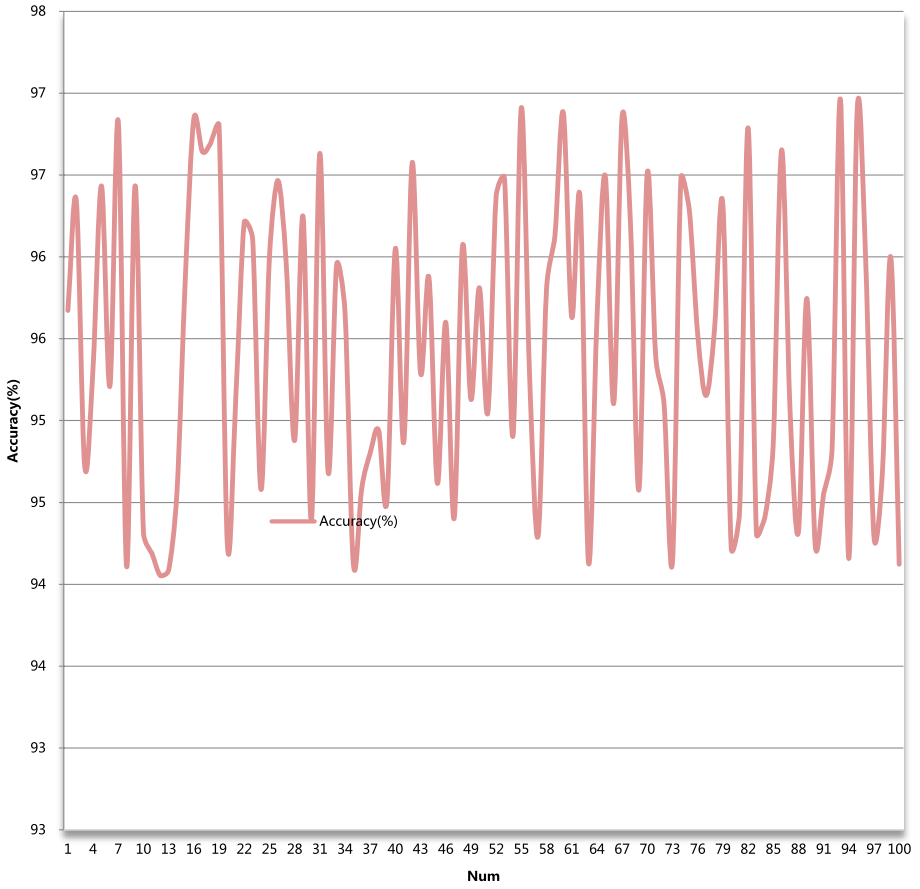


Fig. 10 Statistical table of the voice data processing speed of the voice recognition acoustic compression model

7 Conclusions

The acoustic model based on the deep neural network contains a large number of model parameters and extremely high computational complexity, which has caused a huge obstacle to the application of the acoustic model to mobile devices with limited resources. In addition, the ever-increasing amount of data requires an increase in the scale of acoustic models to absorb to ensure that voice recognition performance is improved. This article combines deep convolutional neural networks to improve voice recognition algorithms, and builds a voice acoustic compression system based on actual needs. Different convolutional neural networks can extract different information in the spectrogram. Therefore, the transfer features of multiple deep convolutional neural networks are spliced together to obtain a more discriminative feature representation than the transfer learning features in a single convolutional neural network. In addition, the above-mentioned classifier is used to verify it through experiments, and the

Table 2 Statistical table of the accuracy of voice recognition of the voice recognition acoustic compression model

Num	Accuracy (%)	Num	Accuracy (%)	Num	Accuracy (%)	Num	Accuracy (%)
1	95.7	26	96.5	51	95.0	76	95.5
2	96.3	27	96.0	52	96.4	77	95.2
3	94.7	28	94.9	53	96.5	78	95.6
4	95.3	29	96.2	54	94.9	79	96.3
5	96.4	30	94.4	55	96.9	80	94.2
6	95.2	31	96.6	56	95.3	81	94.5
7	96.8	32	94.7	57	94.3	82	96.8
8	94.1	33	96.0	58	95.8	83	94.3
9	96.4	34	95.7	59	96.1	84	94.4
10	94.3	35	94.1	60	96.9	85	94.9
11	94.2	36	94.6	61	95.6	86	96.7
12	94.1	37	94.8	62	96.4	87	95.1
13	94.1	38	94.9	63	94.1	88	94.3
14	94.6	39	94.5	64	95.6	89	95.7
15	95.9	40	96.1	65	96.5	90	94.2
16	96.8	41	94.9	66	95.1	91	94.6
17	96.6	42	96.6	67	96.9	92	94.8
18	96.7	43	95.3	68	96.2	93	97.0
19	96.8	44	95.9	69	94.6	94	94.2
20	94.2	45	94.6	70	96.5	95	96.9
21	95.1	46	95.6	71	95.4	96	96.0
22	96.2	47	94.4	72	95.1	97	94.3
23	96.1	48	96.1	73	94.1	98	94.7
24	94.6	49	95.1	74	96.5	99	96.0
25	96.0	50	95.8	75	96.3	100	94.1

experimental research also proves that the algorithm model in this paper has certain effects with an accuracy of 94.1% and speed of 111 s per 100 item sets. Further, in future an improved voice recognition algorithm is used to construct voice acoustic compression system to achieve better accuracy.

Funding None.

Declarations

Conflict of interest The author declares that there is no conflict of interest.

References

1. Orlandi, S., Garcia, C. A. R., Bandini, A., et al. (2015). Application of pattern recognition techniques to the classification of full-term and preterm infant cry. *Journal of Voice, 30*(6), 656–663.
2. Hsu, C. C., Cheong, K. M., Chi, T. S., et al. (2015). Robust voice activity detection algorithm based on feature of frequency modulation of harmonics and its DSP implementation. *IEICE Transactions on Information and Systems, E98.D*(10), 1808–1817.
3. Kumar, P. H., & Mohanty, M. N. (2016). Efficient feature extraction for fear state analysis from human voice. *Indian Journal of Science & Technology, 9*(38), 1–11.
4. Rhodes, R. (2017). Aging effects on voice features used in forensic speaker comparison. *International Journal of Speech Language & the Law, 24*(2), 177–199.
5. Ngoc, Q. K., & Duong, H. T. (2015). A review of audio features and statistical models exploited for voice pattern design. *Computer Science, 03*(2), 36–39.
6. Sarria-Paja, M., Senoussaoui, M., & Falk, T. H. (2015). The effects of whispered speech on state-of-the-art voice based biometrics systems. *Canadian Conference on Electrical and Computer Engineering, 2015*(1), 1254–1259.
7. Leeman, A., Mixdorff, H., O'Reilly, M., et al. (2015). Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison. *International Journal of Speech Language and the Law, 21*(2), 343–370.
8. Hill, A. K., Rodrigo, A., Cárdenas, Wheatley, J. R., et al. (2017). Are there vocal cues to human developmental stability? Relationships between facial fluctuating asymmetry and voice attractiveness. *Evolution & Human Behavior, 38*(2), 249–258.
9. Woźniak, M., & Połap, D. (2017). Voice recognition through the use of Gabor transform and heuristic algorithm. *Nephron Clinical Practice, 63*(2), 159–164.
10. Haderlein, T., Döllinger, M., Matoušek, V., et al. (2015). Objective voice and speech analysis of persons with chronic hoarseness by prosodic analysis of speech samples. *Logopedics Phoniatrics Vocology, 41*(3), 106–116.
11. Nidhyananthan, S. S., Muthugeetha, K., & Vallimayil, V. (2018). Human recognition using voice print in LabVIEW. *International Journal of Applied Engineering Research, 13*(10), 8126–8130.
12. Malallah, F. L., Saeed , K. N. Y. M. G., Abdulameer, S. D., et al. (2018). Vision-based control by hand-directional gestures converting to voice. *International Journal of Scientific & Technology Research, 7*(7), 185–190.
13. Sleeper, M. (2016). Contact effects on voice-onset time in Patagonian Welsh. *Acoustical Society of America Journal, 140*(4), 3111–3111.
14. Mohan, G., Hamilton, K., Grasberger, A., et al. (2015). Realtime voice activity and pitch modulation for laryngectomy transducers using head and facial gestures. *Journal of the Acoustical Society of America, 137*(4), 2302–2302.
15. Kang, T. G., & Kim, N. S. (2016). DNN-based voice activity detection with multi-task learning. *Ieice Transactions on Information & Systems, E99.D*(2), 550–553.
16. Choi, H. N., Byun, S. W., & Lee, S. P. (2015). Discriminative feature vector selection for emotion classification based on speech. *Transactions of the Korean Institute of Electrical Engineers, 64*(9), 1363–1368.
17. Herbst, C. T., Hertegard, S., Zangger-Borch, D., et al. (2016). Freddie Mercury—Acoustic analysis of speaking fundamental frequency, vibrato, and subharmonics. *Logopedics Phoniatrics Vocology, 42*(1), 1–10.
18. Al-Tamimi, J. (2017). Revisiting acoustic correlates of pharyngealization in Jordanian and Moroccan Arabic: Implications for formal representations. *Laboratory Phonology, 8*(1), 1–40.
19. Abdel-Hamid, O., Mohamed, A., Jiang, H., et al. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(10), 1533–1545.
20. Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(7), 1315–1329.
21. Noda, K., Yamaguchi, Y., Nakadai, K., et al. (2015). Audio-visual speech recognition using deep learning. *Applied Intelligence, 42*(4), 722–737.
22. Qian, Y., Bi, M., Tan, T., et al. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(12), 2263–2276.
23. Li, J., Deng, L., Gong, Y., et al. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22*(4), 745–777.

24. Besacier, L., Barnard, E., Karpov, A., et al. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication, 56*(3), 85–100.
25. Watanabe, S., Hori, T., Kim, S., et al. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing, 11*(8), 1240–1253.
26. Vincent, E., Watanabe, S., Nugraha, A. A., et al. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language, 46*(3), 535–557.
27. Swietojanski, P., Ghoshal, A., & Renals, S. (2014). Convolutional neural networks for distant speech recognition. *IEEE Signal Processing Letters, 21*(9), 1120–1124.
28. Lokesh, S., Priyan, M. K., Ramya Devi, M., Parthasarathy, P., & Gokulnath, C. (2019). An automatic Tamil speech recognition system by using bidirectional recurrent neural network with self-organizing map. *Neural Computing and Applications, 31*(5), 1521–1531.

Yuan Xiwen is presently working in National University of Defense Technology. He has a master's degree from the National University of Defense Technology. His research direction is electronic and communication engineering.