Computing (2020) 102:663–682
https://doi.org/10.1007/s00607-019-00753-0

English speech recognition based on deep learning with multiple features

Zhaojuan Song[1]

Received: 30 May 2019 / Accepted: 17 August 2019 / Published online: 26 August 2019

Abstract
English is one of the widely used languages, with the shrinking of the global village, the smart home, the in-vehicle voice system and voice recognition software with English as the recognition language have gradually entered people's field of vision, and have obtained the majority of users' love by the practical accuracy. And deep learning technology in many tasks with its hierarchical feature learning ability and data modeling capabilities has achieved more than the performance of shallow learning technology. Therefore, this paper takes English speech as the research object, and proposes a deep learning speech recognition algorithm that combines speech features and speech attributes. Firstly, the deep neural network supervised learning method is used to extract the high-level features of the speech, select the output of the fixed hidden layer as the new speech feature for the newly generated network, and train the GMM–HMM acoustic model with the new speech features; secondly, the speech attribute extractor based on deep neural network is trained for multiple speech attributes, and the extracted speech attributes are classified into phoneme by deep neural network; finally, speech features and speech attribute features are merged into the same CNN framework by the neural network based on the linear feature fusion algorithm. The experimental results show that the proposed English speech recognition algorithm based on deep neural network with multiple features can directly and effectively combine the two methods by combining the speech features and the speech attributes of the speaker in the input layer of the deep neural network, and it can improve the performance of the English speech recognition system significantly.

Keywords Deep neural network · Fusion · Speech recognition · Multiple features

Mathematics Subject Classification 68T10 · 68T35 · 68T50

✉ Zhaojuan Song
Zhaojuan55@yeah.net

[1] School of Translation Studies of Qufu Normal University, Rizhao 276826, Shandong, China

1 Introduction

Speech is the most natural, convenient, basic and most effective means and method for human beings to obtain information. Voice contains rich inner feelings in the process of transmitting information [1]. With the continuous development of science and technology, the global village is shrinking, and the use of English is becoming more and more common. The emergence of artificial intelligence computers that can understand English speech will inevitably greatly promote the new experience and comprehensive intelligence of human life and work in the future. Intelligent English speech recognition and language interaction technology can not only affect our work and study life, but also have extensive application and significant promotion significance in important strategic fields such as military, education, language promotion, etc. It is the focus of research at home and abroad [2].

The research of speech recognition technology began in the 1950s. In 1952, Bell Labs developed 10 isolated digital identification systems [3]. Since the 1960s, Reddy et al. of Carnegie Mellon University in the United States have conducted continuous speech recognition research, but this period has been slow to develop. Beginning in the 1980s, Coutrot et al. [4] proposed the HMM (Hidden Markov Model, HMM). The statistical model based on this method [5] gradually occupied a dominant position in speech recognition research. The HMM model can well describe the short-term stationary characteristics of speech signals, and integrates knowledge of acoustics, linguistics, and syntax into a unified framework. The research and application of HMM has gradually become the mainstream [6, 7]. The first “non-specific continuous speech recognition system” is the SPHINX [8] system developed by Kai-Fu Lee. The core framework is the GMM–HMM framework (Gaussian mixture model–Hidden Markov Model, GMM–HMM), in which the GMM (Gaussian mixture model, GMM) is used to observe the probability of speech. Modeling, HMM models the timing of speech. In the late 1980s, the ANN (Artificial Neural Network, ANN), the predecessor of DNN (Deep Neural Network, DNN), became a direction of speech recognition research [9, 10]. However, this kind of shallow neural network has a general effect on speech recognition tasks, and its performance is not as good as the GMM–HMM model. Yan et al. [11] proposed a discriminative training criterion and a model adaptive method based on the GMM–HMM acoustic model. Sailor et al. [12] proposed a DBN (Deep Belief Network, DBN) for initializing nodes of a neural network using a RBM (restricted boltzmann machine, RBM). DBN solves the problem that it is easy to fall into local optimum during deep neural network training, and gradually becomes the mainstream of research [13, 14]. Ali et al. [15] proposed a speech recognition method that combines learning features with MFCC (Mel-Frequency Cepstral Coefficients, MFCC) features and can be used for audio scripts of different lengths.

The hybrid Gaussian model in the traditional method is simpler, and has insufficient modeling ability for massive data. Most of the methods in noise reduction work on feature processing. The acoustic model is not robust enough to noise, and the system is Performance is good under quiet conditions, but performance is much lower in noisy environments. The generation of the shallow neural network model slightly relieved the pressure in this area, but the effect was not significant. Compared with hybrid Gaussian models and shallow neural networks, a large number of parameters of deep networks can effectively model massive data, and high-level networks can extract more distinguishing features of speech features. Because this paper uses the deep neural network supervised learning method to extract the high-level features of speech and a variety of speech attribute features, train a deep neural network, select the output of the fixed hidden layer as the new speech feature for the newly generated network, with new The speech features train the acoustic model and ultimately fuse the speech features and speech attribute features into a CNN (convolutional neural network, CNN) framework. The algorithm can not only use a large amount of training data to train deep network with stronger robustness and generalization ability, but also improve the recognition accuracy of English speech recognition system and reduce the influence of noise on English speech recognition system.

The contributions of this article are as follows:

1. Using the deep neural network supervised learning method to extract the high-level features of the speech, training a deep neural network, selecting the output of the fixed hidden layer as the new speech feature for the newly generated network, and training the GMM with the new speech features. HMM acoustic model.
2. Training the speech attribute extractor based on deep neural network for multiple speech attributes, and classifying the extracted speech attributes through deep neural network.
3. The linear feature fusion algorithm is used to fuse the speech features and speech attribute features based on neural network into the same framework, which significantly improves the performance of the English speech recognition system.

The rest of this paper is organized as follows. Section 2 discusses methodology, including speech recognition and deep learning, deep neural network based speech feature extraction and deep neural network model for blending speech features and speech attribute features. Section 3 shows the simulation experimental results, and Sect. 4 concludes the paper with summary and future research directions.

2 Methodology

2.1 English speech recognition and deep learning

English is one of the widely used languages. The speech recognition technology with English as the main research object has become a research hotspot [16, 17]. Figure 1 shows the application of English speech recognition technology. English speech recognition technology converts a piece of English speech signal into corresponding text information. The system is mainly composed of acoustic feature

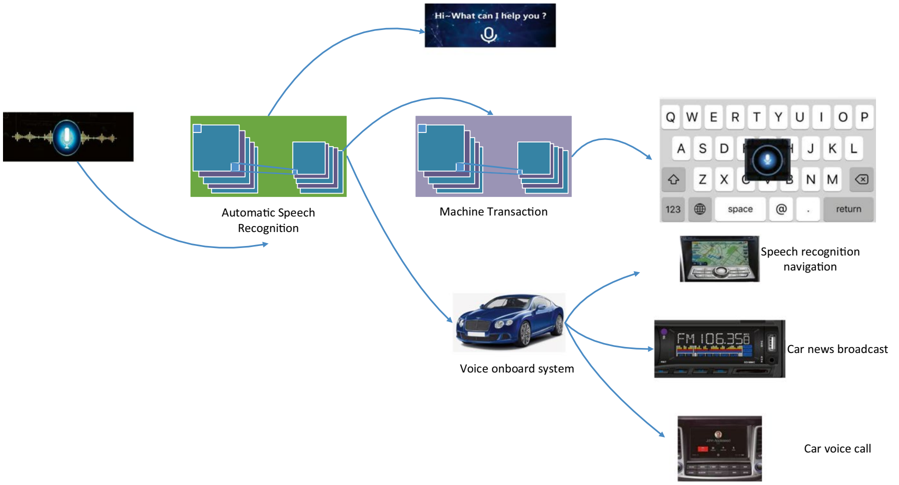

Fig. 1 Application of English speech recognition technology

extraction, language model, acoustic model and decoder [18]. The process of training recognition is that the acoustic features extracted from the original waveform speech data are trained to obtain an acoustic model, and the vocal dictionary and the language model form a network, and the new speech extraction features are represented by an acoustic model and identified by Viterbi decoding result.

The current speech recognition systems are almost always based on HMM. Given the acoustic feature vector sequence $O_T^1 = \{o_1, o_2, \ldots, o_T\}$ of the speech signal, the acoustic model and the language model are combined to perform the decoding search. The most probable sequence of words is $W^* = \{w_1, w_2, \ldots, w_N\}$. Then the process of speech recognition can be regarded as the known feature O_T^1, and the word sequence W* which maximizes the posterior probability $P\left(W|O_T^1\right)$ is found according to the maximum posterior probability criterion, namely:

$$W^* = \text{argmax}\{P\left(O_T^1|W\right)P(W)/P(O_T^1)\} \tag{1}$$

In the above formula, $P(W)$ refers to the probability of language model, which refers to the probability of occurrence of the word sequence W to be identified, which is a prior probability independent of the O_T^1 sequence, which can be statistically obtained through a large number of corpora; $P\left(O_T^1|W\right)$ is the acoustic model probability, which represents the probability of matching the given word sequence W with the acoustic feature vector O_T^1; $P(O_T^1)$ is the probability of occurrence of the acoustic O_T^1, for a certain observation sequence, its size is not It will change, so it can be ignored, so Eq. (1) can be changed to:

$$W^* = \text{argmax}\{P\left(O_T^1|W\right)P(W)\} \tag{2}$$

The above formula is the basic formula of speech recognition. Its purpose is to find the optimal word sequence W*, so that the calculation result is the largest, then the word sequence is the final result of speech recognition. The logarithm of the right part of the above formula is simplified as follows:

$$W^{*} = \operatorname{argmax}\left\{\log\left(P\left(O_{T}^{1}|W\right)\right) + \lambda * \log(P(W))\right\} \tag{3}$$

In the above formula (3), $\log(P\left(O_{T}^{1}|W\right))$ is the acoustic model score, and $\log(P(W)$ is the language model score, where λ is the scaling factor, which is used to weigh the acoustics respectively trained by the acoustic feature training library and the text corpus. The key variables of the contribution of the model and language model to the selection of the word sequence W.

In English speech recognition, the research on continuous vocabulary and non-specific continuous speech recognition is still very challenging, a large vocabulary continuous speech recognition system (LVCSR, Large Vocabulary Continuous Speech Recognition) [19] block diagram, such as Fig. 2.

It can be seen from Fig. 2 that the entire identification system can be roughly divided into the following two parts: the model training phase and the speech recognition phase. Among them, the training of acoustic model is the hub of the whole identification system. How to train an acoustic model with high precision and robustness quickly and effectively is also a key content of this paper. The dictionary in the figure contains a collection of words that the entire system needs to process, indicating the pronunciation of the words, which can form a mapping relationship between the acoustic model and the language model modeling unit, so that the two models can be connected. A searched state space is formed such that the decoder can efficiently perform decoding recognition.

The acoustic feature is an important parameter that can represent the acoustic signal by extracting the important information of the voice file. In the process of identification, it is necessary to select reasonable and effective acoustic features in order to use it to train a better acoustic model. The acoustic model can be seen as the most central part of the entire speech recognition system, which

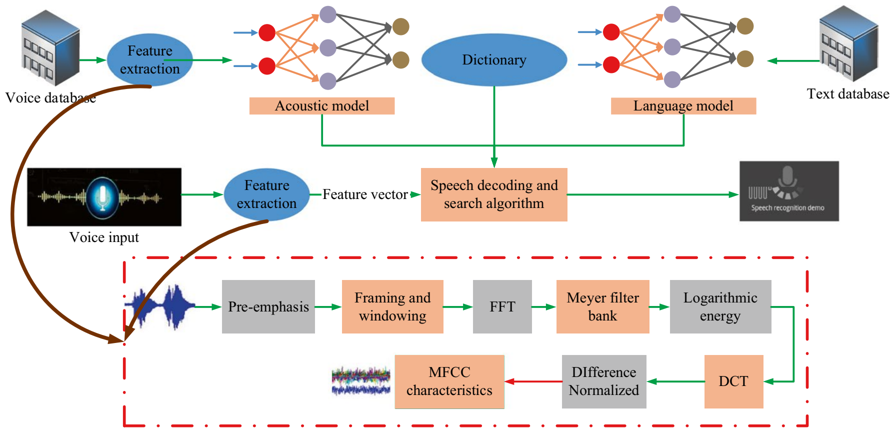

Fig. 2 Block diagram of continuous English speech recognition system

can describe how the feature sequence is generated by acoustic primitives. In the process of speech recognition and decoding, the LM (Language Model, LM) is an important part, which can effectively improve the recognition performance of the system. The acoustic primitives for acoustic modeling in English are the phonemes. There are many kinds of phonemes. There are many kinds of phonemes that can be obtained by constructing phonemes. It is difficult to accurately obtain the words corresponding to the phonemes by relying on acoustic models. Since the language model is derived from some prior linguistic knowledge, the connection structure between sequences is described by probability and statistics. In the recognition process, the search space of words can be effectively constrained, thus effectively eliminating those impossible. Some situations that arise can effectively improve search efficiency and accuracy of recognition results. When the feature extraction is completed and the model is trained, the work to be completed by the speech recognition is to decode and recognize the given speech. The principle of recognition is the process of finding a set of optimal word sequences in the recognition network using a search algorithm in a network of given acoustics, language models, and pronunciation vocabularies.

For a long time, most of the techniques of machine learning and information processing are in a shallow architecture of a single-layer nonlinear transformation, although shallow models [20–22] can achieve good results in some limited problems. However, it does not perform well for the processing of some natural voices, images, videos, etc. Deep learning can learn data by simulating neurons in the human brain. It is found that deep learning has the superior ability to judge while extracting appropriate features [23–25]. Deep learning currently uses more models: DA (De-noising Auto-encoders, DA), DNN (Deep neural networks, DNN), RNN (recurrent neural network, RNN), and CNN. Figure 3 shows the noise reduction auto-encoder and DNN model.

2.2 Speech feature extraction based on deep neural network

The essence of speech recognition is the classification problem, and the feature plays an important role in it. It can be said that to some extent, the pros and cons of the speech feature determine the performance improvement of the speech recognition system. The acoustic features used in previous speech recognition have achieved good results in pure speech corpus, but the use of this feature still does not achieve the effect of low intra-class discrimination and high discrimination between classes. Moreover, some irrelevant information contained in the feature has no meaning for classification, but will affect the classification effect. Focusing on the basic principles of deep learning, this paper analyzes the application of deep neural networks in deep speech feature extraction, builds deep neural networks, and extracts deep features.

In this paper, the network is trained with a 0–1 target value, and the output unit approximates the probability. When an infinite number of training data can be given, the output value represents the probability. When there is only a finite number of

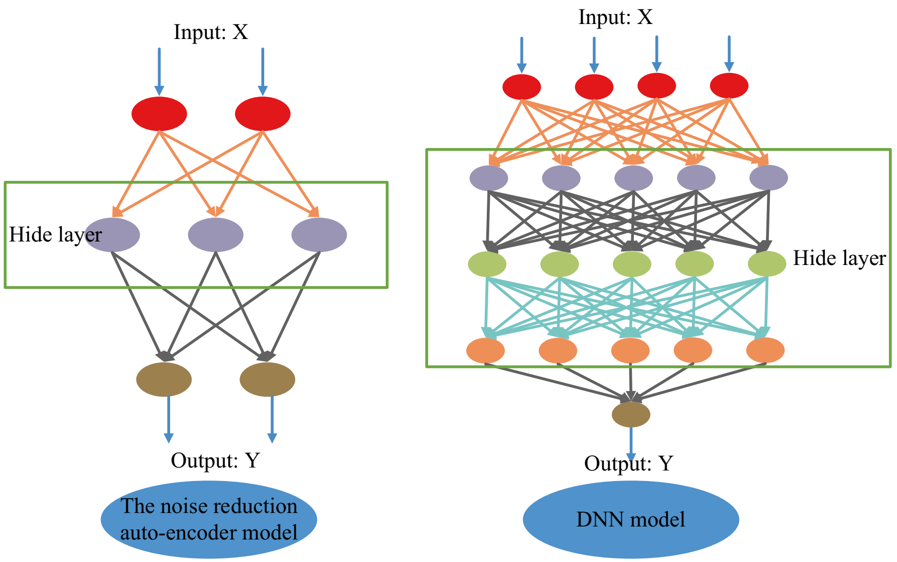

Fig. 3 Noise reduction automatic encoder and DNN model

training data, the output will not represent the probability. Only the exponential output unit nonlinear function can be selected to approximate the output probability, and the output is normalized to 1.0 for each mode. The resulting formula is as follows:

$$Y_k^n = e^{\Theta^k X^n} \Bigg/ \left(\sum_t e^{\Theta^t X^n} \right) \tag{4}$$

When it is a multi-classification problem, the target random variable can be written into the following form in accordance with the multinomial distribution likelihood function:

$$P(t|X,\Theta) = \coprod_{n=1}^{N} \coprod_{k=1}^{K} (Y_k^n)^{t_k^n} \tag{5}$$

Then take the opposite of the likelihood function, and the final error criterion function is:

$$E(\Theta) = -\ln(P(t|X,\Theta)) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_k^n \ln(Y_k^n) \tag{6}$$

Because the network inputs Gaussian features, the input layer and the first hidden layer are modeled using Gaussian–Bernoulli RBM. The first layer of the generated

RBM model is a fixed-dimensional Gaussian-type node layer, and the upper layer is a 720-dimensional Bernoulli node layer. The training algorithm is as follows:

1. Given a sample v of training data, the activation probability of the hidden layer node hj can be expressed as:

$$P(h_j = 1|v) = \sigma(b_j + \sum v_i w_{ij}) \tag{7}$$

   In the above formula, σ(x) is a sigmoid function, specifically $\sigma(x) = 1/(1 + exp(-x))$.
2. For (1) obtaining the hidden layer node value to be randomized to generate an activation state of 0, 1, the visible layer input v′ may be derived according to the state of the hidden layer node. For linear visible layer units, the reconstruction formula is expressed as:

$$v_i' = N(b_i + \sum h_j w_{ij}, 1) \tag{8}$$

   In the above formula, $N(\mu,\ V)$ belongs to a Gaussian distribution with a mean of μ and a variance of V.
3. The reconstructed visible layer state value v′ is used as the input of the RBM structure, and the hidden layer probability h′ is calculated according to step (1).
4. Update the weight parameters as follows, where <> is expressed as an average of all samples in each mini-batch, and ε is the learning rate.

$$\Delta w_{ij} = \varepsilon(< v_i h_j > - < v_i' h_j' >) \tag{9}$$

The initialization parameters of the Gaussian–Bernoulli RBM model are as follows: the connection weight is randomly small initially and the node offset is zero. Each mini-batch contains 256 samples, the learning rate is 0.01, and the training stops after 40 rounds. The activation probability value of each node of the last training hidden layer h1 is retained as the visible layer input data of the upper layer RBM in the superposition structure.

2.3 Deep neural network model combining speech features and speech attribute features

Through in-depth study of the process of human recognition of speech, it is found that words in a person's memory unit are stored in segments as a basic unit, and are distinguished from each other by a series of feature sets, which are called distinguishing features. Distinctive features are characteristic parameters used to characterize phonetic pronunciation knowledge and distinguish speech segments from each other. Considering the broad concept of speech distinguishing features, it can be defined from different aspects of speech such as: acoustics, rhythm, linguistics, pronunciation position, pronunciation method. These distinguishing speech features are called speech attributes.

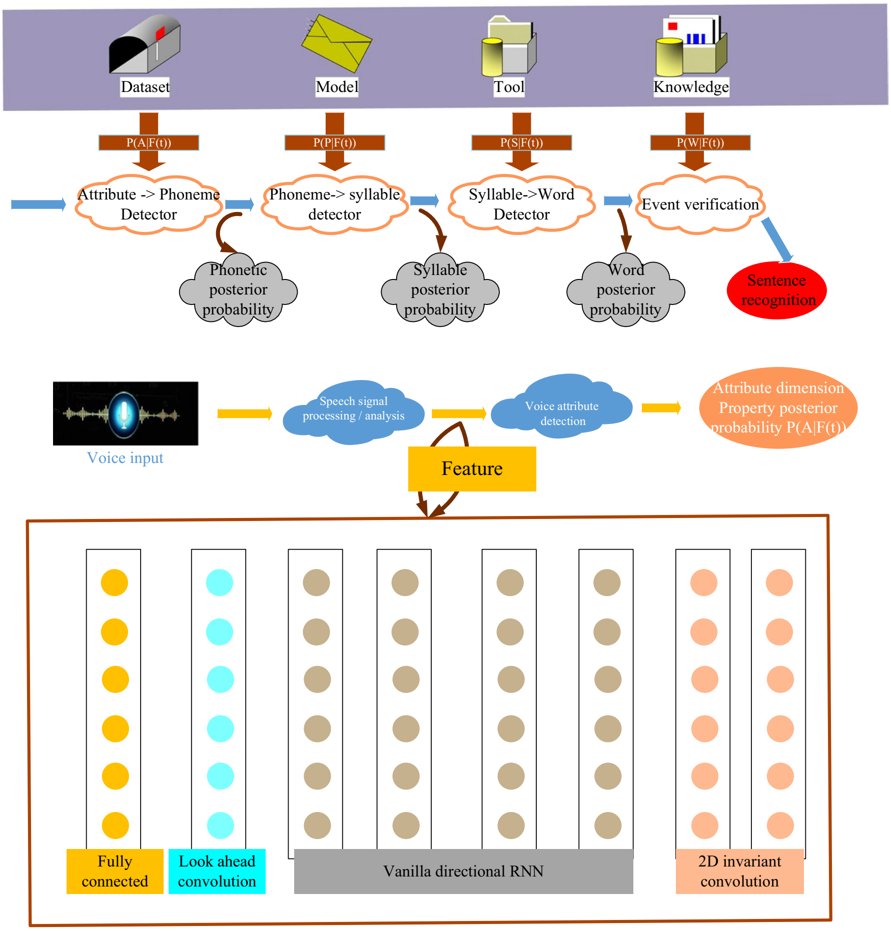
Dataset
Model
Tool
Knowledge
P(A|F(t))
P(P|F(t))
P(S|F(t))
P(W|F(t))
Attribute -> Phoneme Detector
Phoneme-> syllable detector
Syllable->Word Detector
Event verification
Phonetic posterior probability
Syllable posterior probability
Word posterior probability
Sentence recognition
Voice input
Speech signal processing / analysis
Voice attribute detection
Attribute dimension Property posterior probability P(A|F(t))
Feature
Fully connected
Look ahead convolution
Vanilla directional RNN
2D invariant convolution

Fig. 4 CNN–RBM–ASAT hybrid system structure

Since the input requirement of CNN is a fixed-size speech attribute feature, and the speech signal is a signal with a continuous timing and variable length, CNN cannot be used to directly model the speech signal. Therefore, this paper combines the RBM model of Sect. 2.2. Based on the extracted speech features and the features of speech attributes extracted by CNN model, combined with ASAT model, an English speech recognition algorithm based on deep learning is proposed. The model hybrid structure of CNN–RBM–ASAT is shown in Fig. 4.

In the CNN–RBM–ASAT system, for all state sets $s \in [1, S]$, the training network is used to estimate the posterior probability $p\left(q_t = s|x_t\right)$ of the state (where x_t is the acoustic observation at time t), and then to model for different states. The selected input feature is not a single frame, but a window $x_t = [o_{\max(0,t-w)}, \dots, o_t, \dots, o_{\max(0,t+w)}]$ of 2w + 1 frame size, so that the information of adjacent frames can be fully utilized effectively. The CNN–RBM–ASAT network is labeled with the Viterbi algorithm for forced alignment, and its performance is closely related to the quality of the annotation. After training the CNN–RBM–ASAT network, a posterior probability $P(s|o)$ corresponding to the data of each frame on each bound state can be calculated based on the Bayesian formula (10). Among them, o is the observed vector sequence.

$$P(o|s) = P(s|o)P(o)/P(s) \tag{10}$$

In Eq. (10), $P(s)$ is a prior probability of states, and $P(o)$ represents the prior probability of observation vector o, usually replaced by uniform distribution, when all the values of binding probability are obtained. The ASAT model can then be used for decoding recognition.

In traditional speech recognition, the HMM-based recognition system is mainly used. It is a top-down recognition process. This system has achieved great success in commercial applications because it is simple and effective in training recognition. There is still a big gap in human recognition. ASAT adds the knowledge of phonetics based on the traditional probability model to simulate a bottom-up speech attribute detection and cognitive recognition architecture for human speech recognition. The ASAT framework handles voice mainly in three aspects: attribute event detection of voice, voice event integration, and voice event confirmation. The speech attribute detection module realizes the analysis and processing of the original speech signal, and then extracts the speech attributes required by the system. Finally, the combined analysis of these features determines the semantic hypotheses related to the speech, and finally passes the language model, the vocal dictionary, etc. The identification network is composed for decoding. It can be seen from the above figure that the framework is mainly divided into two parts: the front end and the back end. The front end mainly performs speech signal analysis, firstly generates the speech feature parameter $F(t)$, and then combines these features and the attribute detection model to obtain the speech attribute A. The posterior probability $P(A|F(t))$, and finally the attribute features are constructed with the generated posterior probability. The back-end processing is to integrate the attribute features and then use it for the extraction of higher-level voice information. The CNN–RBM–ASAT algorithm proposed in this paper adopts the ASAT framework, which combines deep learning algorithms that combine speech features and speech attribute features. It can not only utilize multiple features of speech information, but also utilize the bottom-up of the ASAT framework. The architecture system improves the efficiency of English speech recognition.

3 Results and discussion

3.1 Performance evaluation index

After setting up a speech recognizer, it is necessary to construct a test standard to evaluate the experimental results. A good evaluation standard can often analyze the problems more in a targeted manner, avoid blind research, and train for parameters and models. And the optimization of algorithms and other aspects provide a good guiding role. In the evaluation of the recognition results, the commonly used performance evaluation indicators are as follows: word recognition accuracy, word error rate, word correct rate and sentence error rate. The formula is expressed as follows:

$$Word\ recognition\ accuracy:\ Acc = (N - D - S - I)/N \tag{11}$$

$$Word\ error\ rate:\ Wer = (S + D + I)/N \tag{12}$$

$$Word\ correct\ rate:\ Wcr = (N - D - S)/N \tag{13}$$

$$Sentence\ error\ rate:\ Ser = (the\ number\ of\ right\ sentences)/(the\ number\ of\ sentences) \tag{14}$$

In the above formula, N represents the number of words to be recognized, D indicates that the number of words are deleted in the recognition result, and S indicates the number of correct words replaced by other words, and I indicates that inserted the number of extra word in the result of recognition. It is not difficult to find that the word correct rate ignores the insertion error and its accuracy is higher than the word recognition accuracy. Among them, the word error rate indicator is the most used, and the sentence error rate indicator requires that each word in the sentence be correctly identified. For the phoneme level modeling, the result is often not high.

For the English speech recognition system, in addition to the examination of the recognition rate, the complexity of the task needs to be investigated. The language model not only provides grammatical knowledge for system identification, but also constrains the search space. The degree of complexity directly reflects the degree to which the recognition system is grammatically constrained. The higher the complexity, the smaller the grammatical constraints, indicates that the identification is more difficult. Entropy and the degree of branching of identification are often used as a criterion for the complexity of system identification. The definition of entropy is as follows:

$$g(L) = -\lim_{m->\infty} \frac{\sum_{w_1, w_2, \dots w_m} P(w_1, w_2, \dots, w_m) \log_2(w_1, w_2, \dots, w_m)}{m} \tag{15}$$

In the above formula, w is a sequence of words, m is its length, and $P(w_1, w_2, \dots, w_m)$ represents the probability of a sequence of words in a language model. The above formula contains all possible sequences of words. If the language is treated as a random process, then the formula can be:

$$g(L) = \frac{-\lim_{m->\infty}(\log_2(P(w_1, w_2, \dots, w_m)))}{m} \tag{16}$$

When m in the above formula becomes large enough, then a random identification unit can be expressed as follows:

$$g_0(L) = \frac{-\log_2(P(w_1, w_2, \dots, w_m))}{m} \tag{17}$$

If the next word is predicted from the current word, it is necessary to select one of the $2^{g_0(L)}$ words with the same probability. Then the branching degree of the recognition task can be defined as follows:

$$\mathrm{FF(L)} = 2^{g_0(L)} \tag{18}$$

Analysis of the above formula shows that the larger the FF(L), the larger the language model PPL is, and the more difficult it is to identify it accordingly. Conversely, the smaller the FF(L), the easier and more accurate the identification.

3.2 Experimental parameter setting

The experiment used 863 English speech database as the training data of deep neural network. The entire speech library consists of two major blocks: training set data and test set data. There are 166 speakers in total, and the ratio of male to female is 1:1. We considered age and selected men and women in six age groups, such as 10, 20, 30, 40, 50 and 60. In this paper, the training set data is 42,638 pieces of all the training data in the original data. The test set data is randomly selected from the test data in the original data. The training corpus of the language model is the sum of some 863 texts and news corpora.

Data normalization operation for all features is beneficial to reduce the difference of features caused by different speakers. Gaussian distribution is used in the input layer of deep neural network. After normalization, the input data distribution satisfies $a_i = 0$ and $\sigma_i = 1$. This can directly avoid the re-estimation of the distribution of training samples.

The speech attribute extraction is based on deep neural network-based attribute feature extraction network for each kind of speech attribute. The attribute feature extraction network is used to extract the required attributes. According to the knowledge of phonetics, the speech attributes are divided into 21 categories, and each attribute feature is extracted. The network structure is $39 \times 256 \times 720 \times 720 \times 256 \times 2$, the network input is a 39-dimensional feature, the output is a two-dimensional representation attribute category, and then a sample is extracted through 21 attribute feature extraction networks to form a 42-dimensional speech attribute, which is a 42-dimensional The feature is used as an attribute feature to extract attributes extracted by the network for a certain sample of data.

Table 1 Structural combination of CNN–RBM–ASAT

Number of offset frames	The model
1	117*720*720*720*1399
2	195*720*720*720*1399
3	273*720*720*720*1399
4	351*720*720*720*1399

Table 2 Speech recognition error rate

Number of offset frames	1	2	3	4
The model	117*720*720*720*1399	195*720*720*720*1399	273*720*720*720*1399	351*720*720*720*1399
Word error rate	0.335	0.297	0.271	0.207
Sentence error rate	0.457	0.381	0.367	0.256

The structural parameters of the deep neural network are mainly the number of hidden layers, the number of units included in each hidden layer, and the choice of hidden layer unit types. In this paper, the deep neural network is mainly composed of one input layer, five layers of hidden layer and one layer of output layer. In the experiment, many combinations of layer number combinations are tried. See the following table for details.

3.3 Performance analysis of speech features and speech attribute features

The features extracted from the model described in Table 1 were compared from the word error rate and the sentence error rate, and the results are shown in Table 2.

It can be seen from Table 2 that as the number of left and right frames of the input data of the deep neural network increases, the new features extracted gradually become better, and it can be seen that the five frames are currently the optimal values. The best effects of the features extracted by the deep neural network are reduced in word error rate and sentence error rate, respectively. It can be seen that the deeper neural network has a certain advantage in continuous speech recognition by extracting relevant information between speech frames effectively in speech recognition.

In order to verify the speech attribute-based speech recognition algorithm proposed in this paper, a speech attribute extraction network is built. The attribute extracted by the attribute feature extraction network is recorded as Attr1, and the

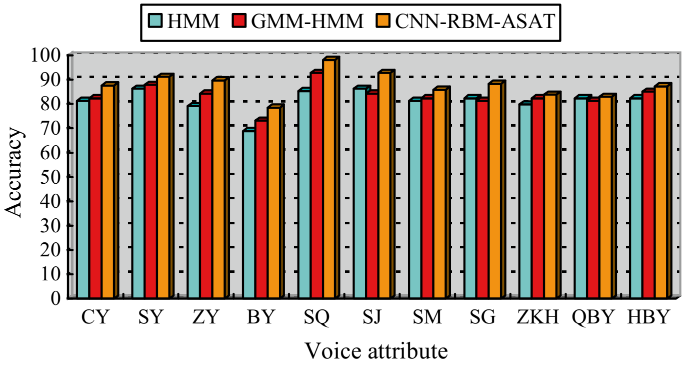

Fig. 5 Speech attribute recognition correct rate

Table 3 Speech attribute feature word recognition error rate

The model with Arr1	HMM	GMM–HMM	CNN–RBM–ASAT
Word error rate	0.185	0.129	0.119
Sentence error rate	0.231	0.207	0.181

GMM–HMM-based speech attribute identifier and HMM-based training are respectively trained by Attr1. The recognition rate of the speech attribute recognizer is shown in Fig. 5. The recognition rate of the speech attribute through the trained word recognizer is shown in Table 3. In the recognition of speech attributes, compared with the methods of GMM–HMM and HMM, the recognition rate based on CNN–RBM–ASAT speech attribute extraction network is improved. This shows that deep neural networks have advantages over data modeling of speech attributes compared to GMM–HMM and HMM algorithms.

3.4 Performance analysis of deep learning algorithms combining speech features and speech attributes

The performance of the English speech recognition model based on deep neural network proposed in this paper is shown in Fig. 6. The abscissa indicates the number of iterations of deep neural network model training, and the ordinate indicates that the model is on the training set and test set corresponding accuracy performance. Figure 7 shows a graph of the average penalty value of the network as a function of the number of iterations. The abscissa represents the number of iterations of the deep neural network model training, and the ordinate represents the average penalty value of the model on the training set and the test set.

A recognition system based on CNN–HMM acoustic model, DNN–HMM acoustic model, CNN–RBM acoustic model and CNN–RBM–ASAT acoustic model is built. The error rates for the four acoustic models are shown in Table 4.

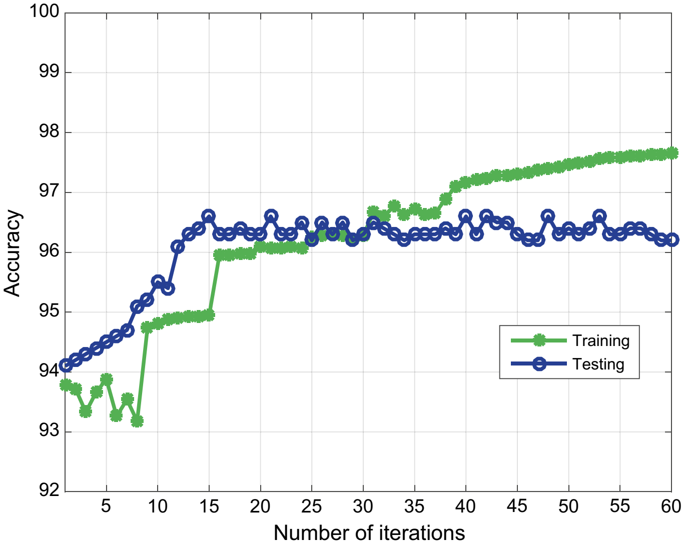

Fig. 6 Deep neural network training performance

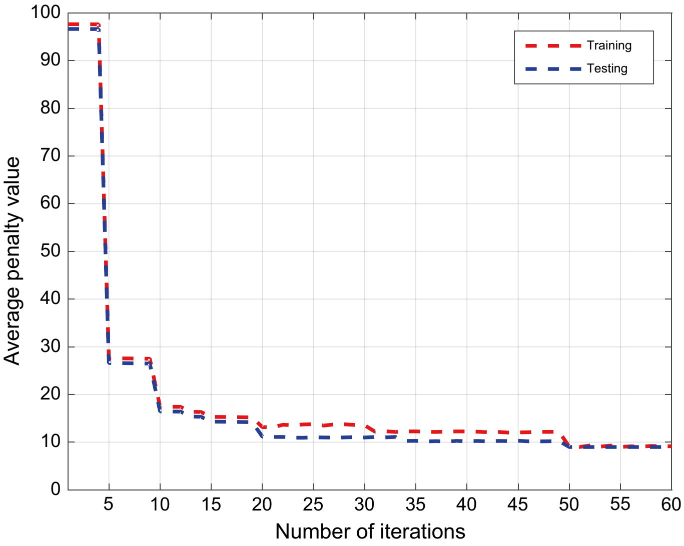

Fig. 7 Average penalty value of CNN–RBM–ASAT network with iterations

Table 4 Error rates for four acoustic models

The model	CNN–HMM	DNN–HMM	CNN–RBM	CNN–RBM–ASAT
Word error rate	0.185	0.105	0.095	0.082
Sentence error rate	0.231	0.157	0.138	0.121

It can be seen from Table 4 that the acoustic modeling method based on CNN–RBM–ASAT model is better than CNN–HMM acoustic model, DNN–HMM acoustic model and CNN–RBM acoustic model. This shows that the CNN model has stronger modeling ability than the DNN model on complex data. CNN's data modeling ability is more advantageous than that of the deep network. The huge parameters contained in the deep network can describe the feature data in great detail. The mining of useful information is more than enough. It can be seen from Figs. 6 and 7 that the performance of the model on the training set is significantly better than the test set, because the model is prone to overfitting on the training set, and the test set can prevent the system from overfitting and improve the system. The role of generalization ability; after multiple iterations of the model, the performance of the model on the training set and the test set tends to be stable, indicating that the model is gradually converging. The unique structure of CNN network makes its modeling ability extremely prominent. The convolutional layer in the network can fully filter out the noise contained in the feature and transfer the useful information to the higher layer. The pooling layer samples the feature and reduces the data.

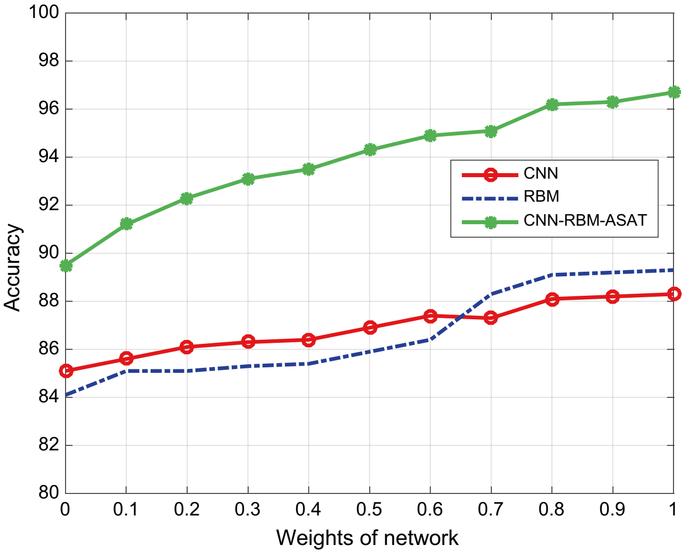

Fig. 8 Performance comparison of linear fusion of CNN network and RBM network

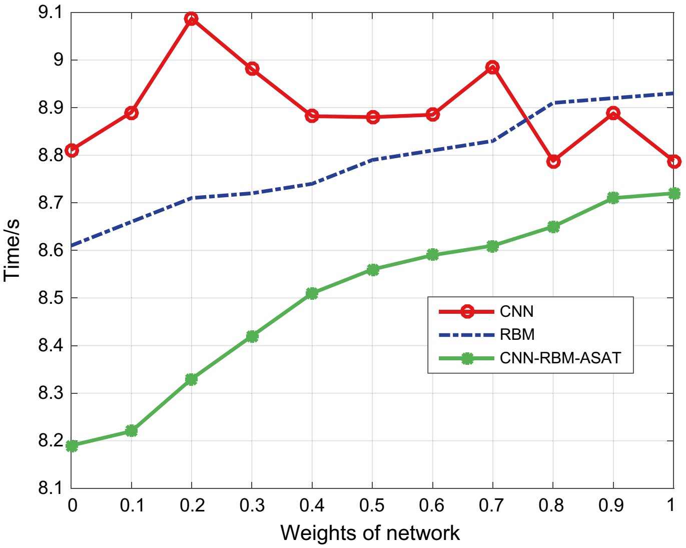

Fig. 9 Comparison of time energy consumption between linear integration of CNN network and RBM network

Dimensions and retained useful information. Deep network training has undergone a large number of supervised global parameter adjustments. This method of differentiated training has an advantage in speech recognition.

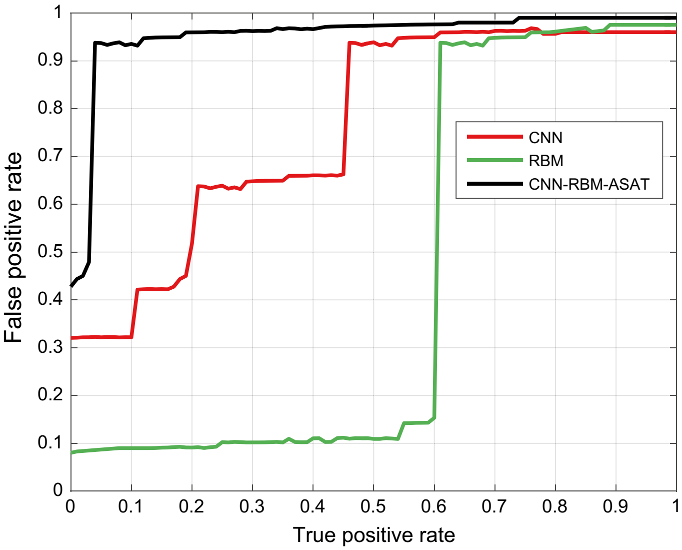

Fig. 10 The ROC curve of CNN network and RBM network

In order to further verify the effectiveness of the proposed feature network, the speech recognition algorithm based on CNN network, the speech recognition algorithm based on RBM network and the speech recognition of CNN–RBM–ASAT network with linear fusion speech features and speech attributes are built. Figure 8 is a performance comparison diagram, and Fig. 9 is a time energy consumption diagram corresponding to each algorithm.

It can be seen from Fig. 8 that the accuracy of the CNN–RBM–ASAT algorithm proposed in this paper is much higher than that of CNN and RBM algorithms, so combining the two can bring out the advantages of these two algorithms and further improve the accuracy. It can be further seen from Figs. 9 and 10 that the linear fusion method based on the deep learning-based speech feature model and the speech attribute model can make full use of the effective information of the speech feature and the attribute feature than the single speech feature model and the single speech attribute model. Although the linear combination of the speech feature model and the speech attribute model is linearly used, the time used is the same as the time when the speech feature model is used alone and the speech attribute model is used. This verifies the deep neural network-based English based on the fusion of multiple features the performance of the speech recognition model.

4 Conclusion

As the hottest research of artificial intelligence, deep learning is widely used in the recognition of speech, image and text and has achieved amazing results. As the main interface of future human–machine interface, English speech recognition directly affects the user experience of intelligent systems. This paper focuses on the application of deep learning in acoustic features and speech attributes, and combines the two features to propose a deep speech-based English speech recognition algorithm that combines multiple features. On the one hand, the large amount of training data collected by the English speech recognition system helps to train deeper networks with stronger robustness and generalization ability. On the other hand, better and stronger deep networks can effectively improve the recognition accuracy of English speech recognition systems and reduce the impact of noise on English speech recognition systems. Our next work considers introducing a clustering algorithm before feature extraction, and first screening the features, which can provide a new idea for feature extraction of the network.

References

1. Nassif AB, Shahin I, Attili I et al (2019) Speech recognition using deep neural networks: a systematic review. IEEE Access 7(99):19143–19165
2. Toth L, Hoffmann I, Gosztolya G et al (2018) A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. Curr Alzheimer Res 15(2):130–138
3. Schillingmann L, Ernst J, Keite V et al (2018) AlignTool: the automatic temporal alignment of spoken utterances in German, Dutch, and British English for psycholinguistic purposes. Behav Res Methods 50(2):466–489
4. Coutrot A, Hsiao JH, Chan AB (2018) Scanpath modeling and classification with hidden Markov models. Behav Res Methods 50(1):362–379
5. Ali Z, Abbas AW, Thasleema TM, Uddin B, Raaz T, Abid SAR (2015) Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN. Int J Speech Technol 18(2):271–275
6. Satori H, Zealouk O, Satori K et al (2017) Voice comparison between smokers and non-smokers using HMM speech recognition system. Int J Speech Technol 20(12):1–7
7. Bocchieri E (2017) System and method for speech recognition modeling for mobile voice search. Jersey Citynj Usphiladelphiapa Uschathamnj Us 47(10):4888–4891
8. Telmem M, Ghanou Y (2018) Estimation of the optimal HMM parameters for amazigh speech recognition system using CMU-Sphinx. Procedia Comput Sci 127:92–101
9. Siniscalchi SM, Salerno VM (2017) Adaptation to new microphones using artificial neural networks with trainable activation functions. IEEE Trans Neural Netw Learn Syst 28(8):1959–1965
10. Enarvi S, Smit P, Virpioja S et al (2017) Automatic speech recognition with very large conversational finnish and estonian vocabularies. IEEE/ACM Trans Audio Speech Lang Process 25(11):2085–2097
11. Yan Z, Qiang H, Jian X (2013) A scalable approach to using DNN-derived features in GMM–HMM based acoustic modeling for LVCSR. Math Comput 44(170):519–521
12. Sailor HB, Patil HA, Sailor HB et al (2016) Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition. IEEE/ACM Trans Audio Speech Lang Process 24(12):2341–2353
13. Cairong Z, Xinran Z, Cheng Z et al (2016) A novel DBN feature fusion model for cross-corpus speech emotion recognition. J Electr Comput Eng 2016(4):1–11
14. Affonso ET, Rosa RL, Rodríguez DZ (2017) Speech quality assessment over lossy transmission channels using deep belief networks. IEEE Signal Process Lett 25(1):70–74
15. Ali H, Tran SN, Benetos E et al (2018) Speaker recognition with hybrid features from a deep belief network. Neural Comput Appl 29(6):13–19
16. Jian L, Li Z, Yang X et al (2019) Combining unmanned aerial vehicles with artificial-intelligence technology for traffic-congestion recognition: electronic eyes in the skies to spot clogged roads. IEEE Consum Electron Mag 8(3):81–86

17. Toshitatsu T, Masumura R, Sakauchi S et al (2018) New report preparation system for endoscopic procedures using speech recognition technology. Endosc Int Open 6(6):E676–E687
18. Ishimitsu S (2018) Speech recognition method and speech recognition apparatus. J Acoust Soc Am 94(109):3538
19. Abdelaziz AH (2018) Comparing fusion models for DNN-based audiovisual continuous speech recognition. IEEE/ACM Trans Audio Speech Lang Process 26(3):475–484
20. Fadlullah ZM, Tang F, Mao B et al (2017) State-of-the-art deep learning: evolving machine intelligence toward tomorrow's intelligent network traffic control systems. IEEE Commun Surv Tutor 19(4):2432–2455
21. Tang D, Bing Q, Liu T (2015) Deep learning for sentiment analysis: successful approaches and future challenges. Wiley Interdiscip Rev Data Min Knowl Discov 5(6):292–303
22. Chen Miaochao, Shengqi Lu, Liu Qilin (2018) Global regularity for a 2D model of electrokinetic fluid in a bounded domain. Acta Math Appl Sin Engl Ser 34(2):398–403
23. Tomczak JM, Gonczarek A (2017) Learning invariant features using subspace restricted boltzmann machine. Neural Process Lett 45(1):173–182
24. Zhang F, Mao Q, Shen X et al (2018) Spatially coherent feature learning for pose-invariant facial expression recognition. ACM Trans Multimed Comput Commun Appl 14(1s):1–19
25. Yin J (2019) Study on the progress of neural mechanism of positive emotions. Transl Neurosci 10(1):93–98. https://doi.org/10.1515/tnsci-2019-0016

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.