Neural Computing and Applications (2024) 36:2371–2383
https://doi.org/10.1007/s00521-023-08959-2

S.I.: MACHINE LEARNING AND BIG DATA ANALYTICS FOR IOT SECURITY AND PRIVACY (SPIOT 2022)

Intelligent speech recognition algorithm in multimedia visual interaction via BiLSTM and attention mechanism

Yican Feng[1]

Received: 17 February 2023 / Accepted: 15 August 2023 / Published online: 29 November 2023

Abstract
With the rapid development of information technology in modern society, the application of multimedia integration platform is more and more extensive. Speech recognition has become an important subject in the process of multimedia visual interaction. The accuracy of speech recognition is dependent on a number of elements, two of which are the acoustic characteristics of speech and the speech recognition model. Speech data is complex and changeable. Most methods only extract a single type of feature of the signal to represent the speech signal. This single feature cannot express the hidden information. And, the excellent speech recognition model can also better learn the characteristic speech information to improve performance. This work proposes a new method for speech recognition in multimedia visual interaction. First of all, this work considers the problem that a single feature cannot fully represent complex speech information. This paper proposes three kinds of feature fusion structures to extract speech information from different angles. This extracts three different fusion features based on the low-level features and higher-level sparse representation. Secondly, this work relies on the strong learning ability of neural network and the weight distribution mechanism of attention model. In this paper, the fusion feature is combined with the bidirectional long and short memory network with attention. The extracted fusion features contain more speech information with strong discrimination. When the weight increases, it can further improve the influence of features on the predicted value and improve the performance. Finally, this paper has carried out systematic experiments on the proposed method, and the results verify the feasibility.

Keywords Speech recognition · Multimedia visual interaction · BiLSTM · Attention

1 Introduction

Speech is the most natural way of interaction for human communication. People get a lot of information from the outside world through speech interaction every day. According to statistics, about 75% of communication in human daily life is completed by voice [1, 2]. At the beginning of the invention of computer, human–computer interaction can only be realized through keyboard, mouse and buttons. With the development of smart tablet, smart phone and other devices, human–computer interaction has developed from keyboard input to touch input [3, 4]. However, these human–computer interaction methods have gradually failed to meet the needs, and the development of speech recognition technology can make people and machines interact better [5].

Let the machine understand the world is the initial purpose of studying speech recognition. In the process of human and machine interaction, the first step is to let the machine hear human voice. The second step is to let the machine translate the sound into words to achieve the purpose of accurate listening. The third step is to let the machine accurately understand the emotion expressed by human voice to achieve the purpose of understanding [6, 7]. Speech recognition technology includes near-field speech technology and far-field speech technology, which are bounded by 30 cm. Due to the limitations of near-field

✉ Yican Feng
fengyican95@163.com

[1] Advanced Film School, Chung-Ang University, Seoul 156756, Korea

speech, the research focus of speech recognition is on the far-field speech interaction technology. This technology includes voice wake-up, recognition and understanding [8, 9]. With the development of artificial intelligence, speech recognition technology is also improving. Deep learning has been used for modeling tasks in speech recognition and achieved good results. These technologies are also applied to intelligent products [10].

Using voice to realize multimedia human–computer interaction is the key research direction of artificial intelligence. This mainly includes automatic speech recognition, natural language processing, speech synthesis and other technologies. Automatic speech recognition technology, also known as speech recognition technology, is responsible for the conversion of speech to text [11, 12]. This is an interdisciplinary task, involving multiple disciplines such as signal processing, pattern recognition, physiology, psychology, computer science and linguistics. Automatic speech recognition technology is the key first step to realize multimedia human–computer interaction using speech and is the core of this paper. Speech recognition technology is divided into two main development directions according to different computing platforms. One is a large vocabulary continuous speech recognition system based on computer platform. The other is the specific control instruction recognition system based on the special speech recognition chip [13, 14]. The processor of special speech recognition chip is a low power consumption, low cost and small size intelligent chip. Compared with computers, its operation processing speed and storage capacity are very limited. This is mainly used to identify short voice control commands and some fixed voice data of specific people. The large vocabulary continuous speech recognition system via computer platform can adjust model scale according to the use scenario, which has more powerful functions [15].

Although automatic speech recognition has developed rapidly for more than half a century, it still has many application problems in real life [16]. On the one hand, this is caused by the defects of speech recognition itself. Yet, academics have great aspirations for voice recognition, despite the fact that the research in this area is inconsistent. The use of deep learning in automatic voice recognition has made significant strides as we have entered the third stage of human evolution [17, 18]. The recognition ability of machine systems has been substantially improved under ideal settings as a result of the rapid development of deep learning technology, which has sparked a technological revolution in the field of voice recognition. The recognition rate of many advanced technologies can even exceed 95% [19]. Under this technical background, various kinds of smart voice products such as mobile phone voice assistant and smart speaker have rapidly opened the market. At present, intelligent voice has become one of the most important multimedia human–computer interaction modes in the consumer market [20].

This work proposes a new method for speech recognition in multimedia visual interaction. The contributions are:

- First, this work proposes three kinds of feature fusion structures to extract speech information from different angles. This extracts three different fusion features based on the low-level features and higher-level sparse representation.
- Secondly, the fusion feature is combined with the BiLSTM with attention mechanism to extract more discriminative feature.
- Finally, this work has carried out systematic experiments on the proposed method, and the experimental results verify the feasibility of this work.

The second section of this paper will carry out the corresponding literature analysis. The third section will elaborate the proposed method. The fourth section is related experiments and analysis. The fifth section is the conclusion.

2 Related work

Reference [21] applied GMM-HMM to speech recognition model, and speech recognition model based on neural network also originated at this time. However, due to the influence of factors such as computing power and data at that time, this did not achieve particularly good results. Literature [22] proposed a connection timing classification model. This used a single model to directly model the speech frame sequence, and the one-way model could be applied to the field of streaming speech recognition. Literature [23] proposed LAS based on attention mechanism. This translated speech recognition task into Seq2Seq modeling task, which directly mapped audio sequence to corresponding text sequence. With the continuous improvement of computing performance, the neural network framework was also iteratively optimized. In the aspect of speech recognition based on DNN-HMM framework, there were various structural optimizations for acoustic models. Literature [24] proposed CTC extended structure, which was a kind of cyclic neural network converter. This integrated the input sequence and the historical output sequence, which could optimize the acoustic and language models at the same time. This solved the defects of the CTC output independence assumption, and the use of one-way coding structure could be applied to streaming speech recognition. CTC, LAS and RNN-T were the three main end-to-end speech recognition frameworks. Its basic

assumptions, training objectives, training process and reasoning process were different. Literature [25] proposed time-delay neural network, which would apply convolutional neural network, which was brilliant in the field of image recognition, to speech recognition. It used hole convolution for selective input, which reduced the amount of computation and increased the field of view of the top input. Literature [26] proposed LSTMP, which was based on the traditional LSTM by introducing an affine layer between the memory block and the output layer. This effectively reduced the number of model parameters while retaining the advantages of LSTM for serialization modeling. Literature [27] proposed the maximum mutual information training criterion to take into account the correct path and the confused path at the same time. This increased the scoring difference between the correct path and other paths. This could improve the correct path score and more match the recognition target.

The end-to-end speech recognition framework was relatively simple. The unified optimization goal avoided the cumulative error problem caused by multiple modules and did not require forced alignment. At the same time, single words could be used as the modeling unit, without additional pronunciation dictionary, and the model had good extensibility. Literature [28] proposed a code-decode structure, which had been widely used in the field of machine translation for the first time and achieved good experimental results. Literature [29] extended the application of code-decode structure to the field of speech recognition. Therefore, in addition to CTC, sequence-to-sequence end-to-end speech recognition framework had also become one of the mainstream methods. It was not necessary to assume the alignment of input and output sequences in advance, and this could learn encoding, decoding and how to align at the same time. Cyclic neural network needed to calculate the response of each time frame in sequence at the training time. Its main defect was that it was easy to lose the historical information of long words. This had disadvantages for speech recognition tasks with monotonic relationship between audio and text. In addition, the training should to calculate each time frame in time sequence. Literature [30] proposed the Transformer model, which replaced RNN with self-attention mechanism to model the global timing signal. At the same time, it could train in parallel and improve the calculation speed of the model. Literature [31] applied Transformer model to specific mandarin speech recognition tasks and achieved results beyond LAS model. Literature [32] combined CTC with LSTM model and inserted the vector output from the top of LSTM into the CTC model. LSTM-CTC model was constructed using CTC decoding method. The model not only reduced the word error rate, but also shortened the training time of the model. Reference [33] proposed LFR-DFSMN, which added jump connections between memory layers. This alleviated the problem of gradient disappearance when building a deep model structure. Literature [34] proposed a streaming multi-level truncated attention model, which introduced a special multi-level attention mechanism.

3 Speech recognition via BiLSTM and attention mechanism

First, this work proposes three kinds of feature fusion structures to extract speech information from different angles. This extracts three different fusion features based on the low-level features and higher-level sparse representation. Secondly, the fusion feature is combined with BiLSTM with attention mechanism. The extracted fusion features contain more speech information with strong discrimination.

3.1 Speech sparse representation extraction

When extracting fusion features, it is necessary to extract corresponding sparse representation features from the underlying features. As a high-level feature, the sparse representation feature has good uniqueness and distinctiveness. In this paper, KSVD dictionary learning algorithm and BPDN sparse decomposition algorithm are used to extract sparse features. KSVD algorithm is improved to solve the problem of slow learning of dictionary learning algorithm under a large amount of voice data.

In this paper, KSVD algorithm is used to learn the dictionary. The dictionary learned by this algorithm from the training data can be regarded as a reduced-dimension representation of the training set. Therefore, this can well represent the characteristics of training data. In this algorithm, first of all, data are randomly selected from the training data as the initial dictionary atom. Then, the OMP algorithm is used to solve the sparse representation while the dictionary is left unmodified. Then, the SVD algorithm is used to update the dictionary atoms while maintaining the sparsity. As long as there is significant discrepancy between the sample data and the data rebuilt using the created dictionary, iterate until a minimal error is achieved. The speech training data chosen in this research at the frame level is substantial, and the system requires frequent iteration. Due to the sluggish acquisition of sparse representation from the learning dictionary, this paper randomly selects the training dictionary from each voice training sample. This study refines the method of training a dictionary in an effort to address this issue. The enhanced KSVD chooses a specified amount of data from each type of voice data to create a short training dataset that is then

used to train the dictionary's first sample. The dictionary is then used to obtain the sparse representation of all the voice data.

The basic idea of the algorithm is to update all atoms until the overall reconstruction error is minimal. The target function of KSVD is as follows:

$$J_K = \min_{D,A} \|X - DA\|_F^2 \tag{1}$$

where X is the training data, D is the initial dictionary, A is the sparse sample.

KSVD algorithm is divided into three stages. First is the initialization stage, which generates initial MFCC feature training samples according to the improved method proposed in this paper. These samples contain different phoneme categories. Each category takes several samples from the training data to form a small training sample. This initial training data covers all phoneme types. Then, select some samples from the initial samples as the atoms of the initial dictionary. Next is the sparse coding stage, which fixes the dictionary obtained in the previous step and normalizes each column of atoms in the dictionary. This uses the sparse decomposition algorithm OMP to solve sparse data. Finally, dictionary update stage, in which the dictionary atoms and sparse features need to be updated at the same time. Generally, it is assumed that one optimization variable is fixed, another variable is optimized, and the variable is updated alternately.

For solving the problem of sparse representation, the OMP algorithm in the greedy algorithm adopted in document [35]. The flaw of this algorithm type is that during the iterative process, only the local optimal solution is guaranteed. At the same time, this sparsity is pre-fixed by the matching pursuit algorithm. Complex speech data is represented with varying degrees of sparsity across distinct speech areas. Consequently, it is not appropriate to apply the same sparsity to the sparse representation of each voice frame when dealing with voice data. In light of these issues, this article employs the algorithm of base tracking noise reduction in order to address the sparse representation. The optimization problems of conventional sparse representation are as follows:

$$A = \arg\min_A a_{i0} \tag{2}$$

where a_i is sparse parameter.

In the actual solution process, the above problems are usually converted into the problem of solving the L1 norm. When noise is considered, the sample signal becomes a superimposed signal. If BP algorithm, namely, BPDN algorithm [36], is used for noise suppression, the optimization problem is:

$$A = \arg\min_A a_{i1} \tag{3}$$

where a_i is sparse parameter.

BPDN algorithm can find the global optimal solution by minimizing the reconstruction error while obtaining the rarest solution of the signal. After a learning dictionary is obtained by using KSVD algorithm, all training samples are processed by BPDN algorithm to obtain corresponding sparse representation. Finally, the sparse representation features are sent into the recognition model for classification.

3.2 Structural design of multi-level feature fusion

This paper proposes several structures of feature fusion, which can extract the features of speech signals from different angles. When one mode feature is damaged, other mode features can help fill in the missing information to enhance the robustness of the signal. Finally, the interaction between the characteristics of different patterns can be considered as an enhancement factor.

For the same segment of voice data, feature extraction is performed from frequency domain and cepstrum domain, respectively. First, the CQT spectrum corresponding to the voice data can be obtained by CQT transformation of the initial audio data. CQT transform is different from Fourier transform in that its spectrum is nonlinear and the length of filtering window is variable. This can be changed according to the change of spectral line frequency to obtain better performance. In addition, this paper calculates Mel cepstrum coefficients of speech to obtain MFCC features, so as to achieve the extraction of underlying features from different angles. The two extracted low-level features are fused to form a low-level fusion feature. Then, the fused features are extracted from the sparse representation by dictionary learning to obtain the fused sparse features. The fusion structure is shown in Fig. 1:

The specific fusion expression can be expressed as:

$$F_{ll} = \min_{D_{mc}, a_{ll}^{(i)}} \left\| x_{mc}^{(i)} - D_{mc} a_{ll}^{(i)} \right\|_2^2 + \mu' \left\| a_{ll}^{(i)} \right\|_1 \tag{4}$$

where x_{mc} is the fusion sample, D_{mc} is the fusion dictionary after dictionary learning after the fusion of low-level and low-level features.

This method fuses two kinds of low-level features extracted from different angles, and then extracts the high-level sparse representation. The advantage of this method is that it combines the information of the two underlying modes before learning together. The sparse representation after learning can capture the relationship between the two modal features. This can also play a complementary role and make up for the lack of single type features.

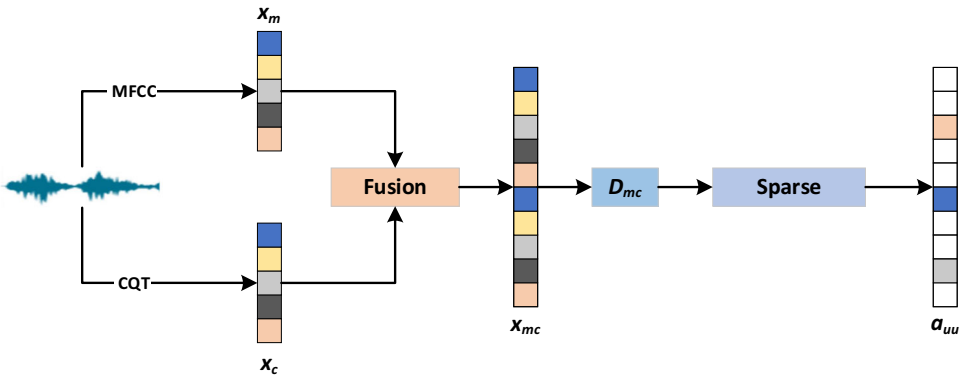

Fig. 1 Fusion of low-level feature and low-level feature

Fusion of high-level feature and high-level feature adopts the method of directly sparse coding the underlying features. First of all, after a series of preprocessing of the original data of the same voice segment, two different underlying features, MFCC and CQT, are extracted. The method of implementing parallel sparse coding after extracting the underlying features. This is to learn a dictionary for the low-level features of each mode and then conduct sparse coding to obtain two different high-level feature representations. Finally, the high-level sparse representation obtained separately will be fused in parallel. Figure 2 depicts this modal parallel sparse coding scheme. The scheme learns two dictionaries for the two underlying modal features in parallel. Finally, the two high-level sparse feature representations are fused according to the following structure to obtain new fusion features.

Finally, the sparse representation obtained in parallel is fused according to the above structure to obtain the fused feature representation. The specific expression is as follows:

$$F_{hhm} = \min_{D_m, a_m^{(i)}} \left\| x_m^{(i)} - D_m a_m^{(i)} \right\|_2^2 + \mu' \left\| a_m^{(i)} \right\|_1 \tag{5}$$

$$F_{hhc} = \min_{D_c, a_c^{(i)}} \left\| x_c^{(i)} - D_c a_c^{(i)} \right\|_2^2 + \mu' \left\| a_c^{(i)} \right\|_1 \tag{6}$$

$$a_{hh} = [a_m, a_c] \tag{7}$$

where a_m and a_c are high-level sparse representations extracted from the two features, respectively.

The method of fusing two high-level features is a simple method of encapsulating features from two modes. This increases the effective information contained in the feature to a certain extent, but completely discards the underlying effective information. This can't conduct joint training on voice data like the fusion between the bottom layer and the bottom layer. Therefore, it cannot capture the correlation between the two features that may be beneficial to the recognition task. And, the same as the fusion between the bottom layers, the extracted features are the fused higher-level features, which will cause some beneficial information at the bottom layer to be completely discarded.

For the high-level fusion and the low-level fusion, the two feature fusion structures are finally transformed to obtain a fused sparse representation feature, which will be sent into the classification model. However, this completely discards the underlying features, which also discards a lot of important information contained in the underlying features. In view of the problems of the above two structures, this paper has made further improvements on the structure designed previously. This paper designs and adopts a structure that integrates the low-level features and high-level features and extracts features, as shown in Fig. 3.

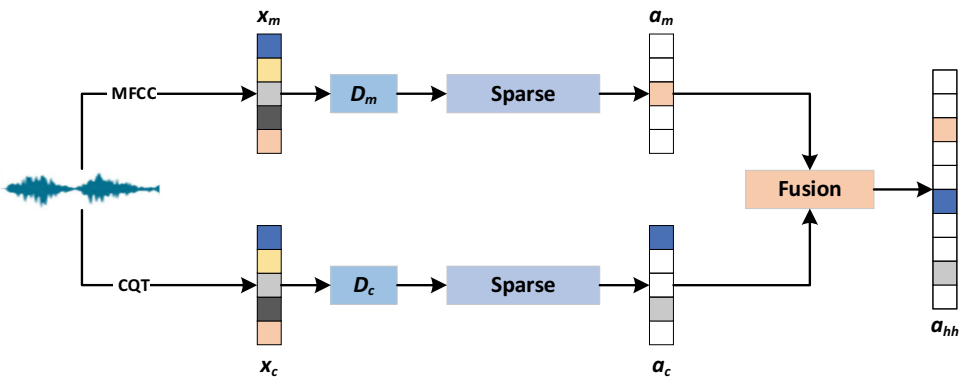

Fig. 2 Fusion of high-level feature and high-level feature

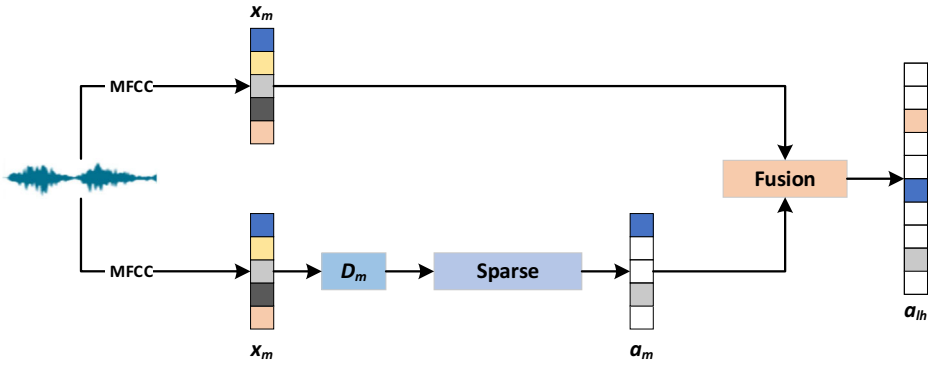

Fig. 3 Fusion of low-level feature and high-level feature

This structure first extracts MFCC features directly from voice data. On the other hand, a sparse high-level feature is obtained by sparse coding after extracting the low-level feature. Finally, two different modal features are fused through the above fusion structure to obtain a new fusion feature. The sparse representation extraction formula is as follows:

$$F_{lh} = \min_{D_m, a_m^{(i)}} \left\| x_m^{(i)} - D_m a_m^{(i)} \right\|_2^2 + \mu \left\| a_m^{(i)} \right\|_1 \tag{8}$$

$$a_{lh} = [x_m, a_m] \tag{9}$$

where x_m is MFCC feature, a_m is sparse feature.

As a low-level feature, MFCC contains a lot of original information. The more original information, the more complete the information. It accurately describes the nonlinear characteristics of human ear frequency. As the underlying features, these features usually contain redundant information, which will interfere with the expression of effective information. The features after sparse representation have good distinguishability, and will be supplemented when the underlying information is ambiguous. This makes it easier for the classifier to obtain the effective information contained in the signal. Therefore, compared with the previous two structures, the fusion structure of the bottom layer and the top layer will be easier to accurately mine the effective information in the voice signal.

3.3 Speech recognition based on BiLSTM and feature fusion

Speech recognition is a typical signal processing problem with time characteristics. The hidden layer neurons in DNN and CNN are only affected by the current voice signal. RNN is a neural network structure that allows hidden layer neurons to have self-feedback loops. Its hidden layer input not only comes from the voice signal at the current moment, but also receives the memory information from the hidden layer at the previous moment. This structure mimics the speech perception ability of human brain, that is, to predict the speech information of the next moment through historical speech information. In theory, RNN model can model speech signals of any time sequence length. However, in the process of long sequence training, the RNN network will produce a high power of the matrix, which will lead to gradient disappearance or explosion. If the gap between the previous voice information and the current frame becomes very large, RNN will lose the ability to learn to connect such a long-distance information.

Unit state, forgetting gate, input gate, and output gate are all components of LSTM, a unique subclass of RNNs. The unit state is where all of your long-term memories from before this moment are kept. The forgetting gate eliminates unwanted or unnecessary information from long-term memory. Short-term memory is used by the input gate to update permanent storage. The output gate combines the information stored in both the short- and long-term memories to produce a final output. The issue of long-term dependency of RNN has been resolved by LSTM, which can handle both long- and short-term memory. The traditional LSTM model is a one-way transmission structure, while speech recognition is a context-dependent phoneme distribution problem, so this paper uses BiLSTM to process speech. Two independent forward and backward LSTM networks process the front and back information, respectively, and then splice the two outputs to get the acoustic model at that time.

In this paper, attention mechanism is added to the model, which searches for some inputs related to the current prediction output from the input through the calculation of attention weight in the network. The closer the relationship, the greater the weight value assigned to the input vector. When the relevant input features contain more effective information, the greater the impact of the relevant input features given greater weight on the current output. This can improve the recognition accuracy. Therefore, this chapter sends the fusion features with different levels of information extracted from the previous chapter into the BiLSTM network via attention mechanism. This makes use of the characteristics of attention to strengthen the influence of relevant input characteristics on the current

prediction output, so as to achieve the effect of reducing the recognition error rate. The feature fusion speech recognition model based on attention and BiLSTM (FF-ATT-BiLSTM) is shown in Fig. 4.

The structure of BiLSTM model based on attention mechanism is composed of coding layer, attention mechanism end and decoding layer. In terms of model building, BiLSTM is selected for classification. Compared with the traditional GMM-HMM and support vector machine models, the specific structure of BiLSTM can connect the information far away from the previous to the current task. This can make full use of the voice information at the front and back ends and also avoid the problem of gradient disappearance. And, this kind of serialized network structure is suitable for processing voice data.

When forecasting the current time, the hidden layer features of all the current input times are matched with the predicted state of the previous time to calculate the score. The matching score can be calculated by a small neural network. A small neural network is trained to learn the relationship between the higher-order features and the predicted state at the last time. The key step is to use the softmax function, which passes the scores at all times through the softmax function to ensure that the sum of all weight values is 1. Finally, the weighted sum of the hidden layer features and corresponding weights output by the coding end is the context vector. The specific relationship of the context vector at the first moment can be expressed as follows:

$$c_1 = \sum_{i=1}^{T} x_{1t} h_t \tag{10}$$

where x is input.

Attention mechanism will give greater weight to the input sequence most relevant to the current prediction output. According to this principle, when the input feature contains more effective information, the proportion of information in the input feature that is more relevant to the current output will increase as the corresponding weight increases. The greater the impact of the relevant input characteristics after being given greater weight on the current output. Therefore, this paper puts the multimodal features into the model based on attention mechanism. This can increase the effective input information and also strengthen the effect of effective input on output through the model.

In order to alleviate the over-fitting phenomenon in the training process, this paper uses the Dropout method and a certain probability to randomly delete the number of hidden layer units in the model to improve the generalization ability of the model. In addition, this paper also sets up the Early Stopping mechanism during the training process. This is stopped in time when the validation accuracy on a

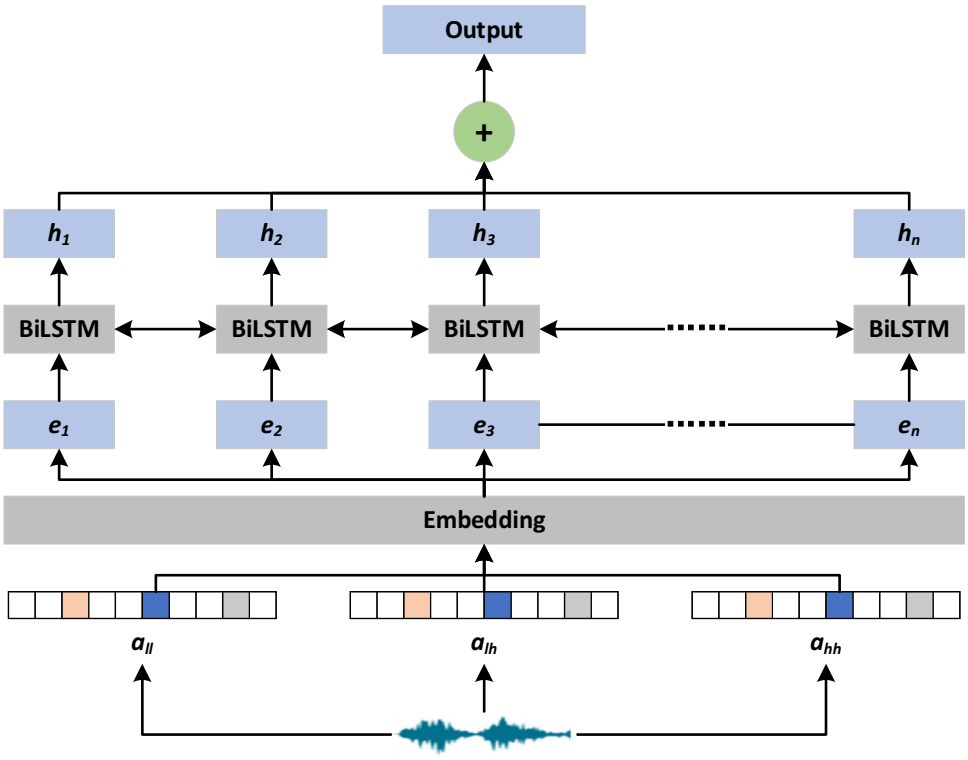

Fig. 4 FF-ATT-BiLSTM pipeline

certain number of validation sets has not improved, so as to prevent the training process from over-fitting.

4 Experimental result

4.1 Dataset and experimental environment

This work collects voice data during multimedia visual interaction to build two datasets SRA and SRB. The two data sets contain different amounts of data. SRA has 20,318 training samples and 8026 test samples. SRB has 52,189 training samples and 19,053 test samples. The model is a deep learning model. The experimental environment is in Table 1.

The speech recognition task is essentially a classification task. The performance evaluation indicators selected for this work are the accuracy and F1 score.

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$F1 = \frac{2 \times R \times P}{R + P} \tag{12}$$

4.2 Analysis on FF-ATT-BiLSTM training

FF-ATT-BiLSTM is a deep learning model, so it must be trained and optimized first. This work first analyzes the training process of FF-ATT-BiLSTM, and the main analysis object is the loss in the training process as Fig. 5.

With the increase in training times, the loss of FF-ATT-BiLSTM on both data sets gradually decreased. When the training reaches 60 epochs, the loss will not change significantly, which indicates that the model has converged at this time. In addition, when the model converges, the loss of FF-ATT-BiLSTM on SRB is smaller. This is because the dataset is larger and the model training is more sufficient.

Table 1 Experimental environment

Option	Configuration
Operating system	Windows10
CPU	Intel(R) Core(TM) i7-9700
GPU	NVIDIA RTX3070
Deep learning framework	PyTorch 1.6
Development language	Python

4.3 Method comparison

To further verify the progressiveness of FF-ATT-BiLSTM in the field of speech recognition, this paper compares it with others. The methods contain RNN, LSTM, BiLSTM and Transformer. To ensure the consistency of the experiment, keep the experimental parameters unchanged as much as possible, the result is in Table 2.

FF-ATT-BiLSTM proposed in this paper can obtain the best performance. Specifically, FF-ATT-BiLSTM achieves 92.1% accuracy and 90.5% F1 score on SRA and 95.5% accuracy and 92.8% F1 score on SRB. Compared with the hot Transformer algorithm this year, F-ATT-BilSTM achieves 2.2% accuracy improvement and 3.2% F1 score improvement in SRA. In SRB, the corresponding increases are 2.9 and 2.3%, respectively. These performances and improvements verify the feasibility and superiority of FF-ATT-BiLSTM in the field of speech recognition in the process of multimedia visual interaction.

To further verify the advantages of FF-ATT-BiLSTM, this paper compares the performance of different training stages, as shown in Fig. 6.

At different stages of training, FF-ATT-BiLSTM can achieve different degrees of performance improvement compared with other methods. This further verifies the superiority of this method.

4.4 Analysis on feature fusion

FF-ATT-BiLSTM proposed three different feature combinations and fused the three sets of features. To verify the feasibility of this feature fusion, this paper carries out comparative experiments for different features. To ensure the comparability of the experiment, this work tries to keep the experimental parameters unchanged. The experimental results are shown in Table 3.

From the data shown in the table, we can see that the performance corresponding to the fusion of low-level features and low-level features is the lowest. If high-level features are combined with high-level features, the performance will be improved to a certain extent. The fusion of low-level features and high-level features can further improve the accuracy and F1 score. However, a single feature combination cannot achieve the best performance. Only after these three groups of features are fused can FF-ATT-BiLSTM achieve the highest speech recognition performance.

4.5 Analysis on improved KSVD

This work uses the improved KSVD algorithm to sparsely represent speech features. In order to verify the feasibility

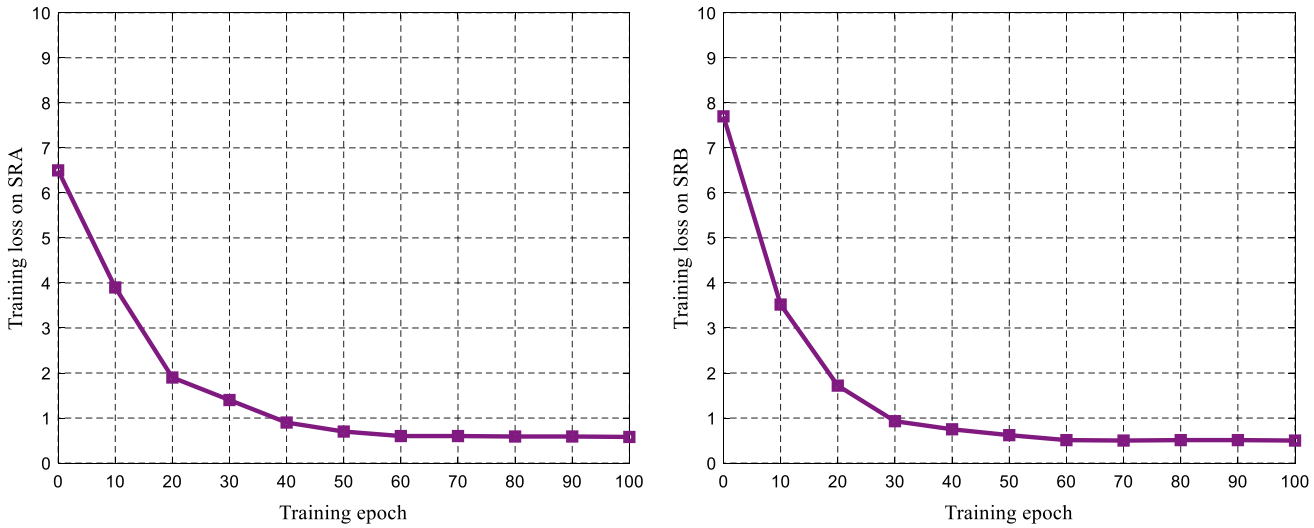

Fig. 5 Training loss of FF-ATT-BiLSTM

Table 2 Method comparison

Method	SRA		SRB	
	ACC	F1	ACC	F1
RNN	81.2%	80.1%	85.1%	83.3%
LSTM	84.1%	82.2%	86.9%	84.9%
BiLSTM	87.5%	85.9%	90.3%	88.7%
Transformer	89.9%	87.3%	92.6%	90.5%
FF-ATT-BiLSTM	92.1%	90.5%	95.5%	92.8%

Table 3 Analysis on feature fusion

Feature	SRA		SRB	
	ACC	F1	ACC	F1
a_{ll}	88.2%	86.2%	90.3%	88.5%
a_{hh}	89.3%	86.7%	92.0%	89.1%
a_{lh}	90.5%	87.8%	93.6%	91.1%
$a_{ll} + a_{hh} + a_{lh}$	92.1%	90.5%	95.5%	92.8%

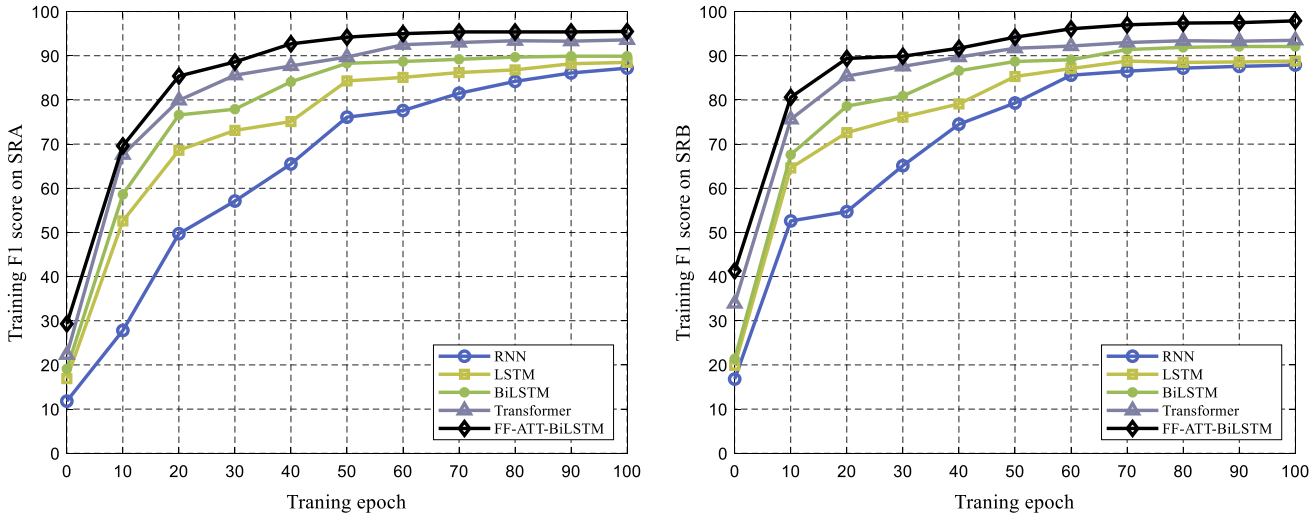

Fig. 6 Training F1 score on different stage

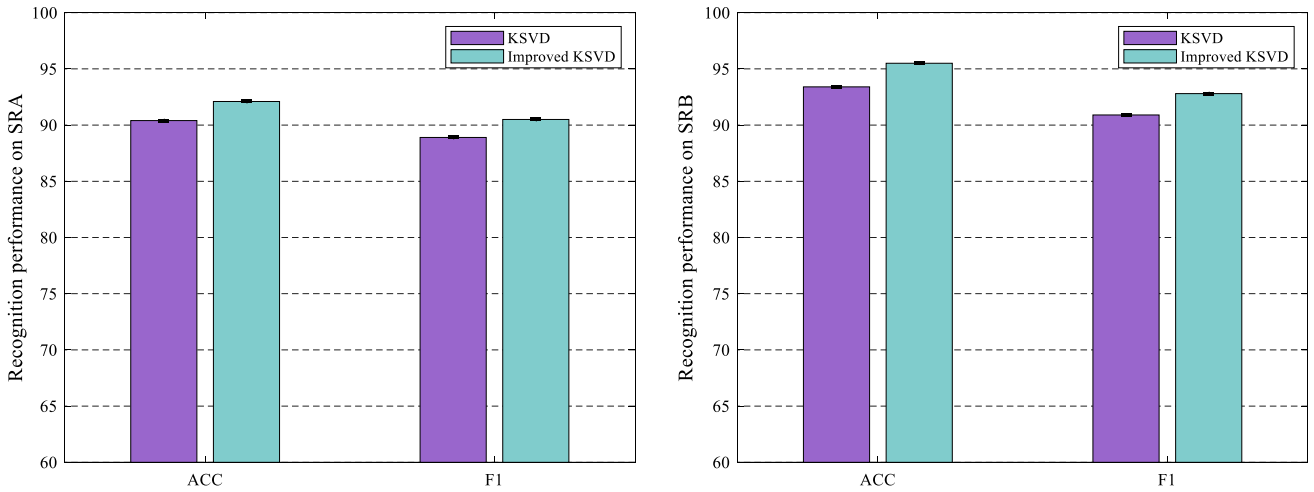

Fig. 7 Analysis on improved KSVD

of this improvement measure, this paper compares the model recognition performance when the improvement measure is not used and the improved KSVD algorithm is used. The experimental result is shown in Fig. 7.

After the corresponding improvement of KSVD, the speech recognition performance of FF-ATT-BiLSTM can be improved to a certain extent. Specifically, the accuracy and F1 score on SRA are improved by 1.7 and 1.6%, respectively. Similarly, the accuracy rate and F1 score on SRB are improved by 2.1 and 1.9%, respectively.

4.6 Analysis on attention

FF-ATT-BiLSTM uses attention to enhance features. To verify the role of attention in speech recognition, this work analyzes the model performance when attention is not used and when attention is used. The experimental results are shown in Table 4.

After embedding attention module, the speech recognition performance of FF-ATT-BiLSTM can be improved to a certain extent. Specifically, the accuracy and F1 score on SRA are improved by 1.3 and 1.4%, respectively. Similarly, the accuracy and F1 score on SRB are improved by 1.3 and 1.2%, respectively.

Table 4 Analysis on attention

Method	SRA		SRB	
	ACC	F1	ACC	F1
FF-BiLSTM	90.8%	89.1%	94.2%	91.6%
FF-ATT-BiLSTM	92.1%	90.5%	95.5%	92.8%

4.7 Analysis on BiLSTM

FF-ATT-BiLSTM uses BiLSTM to extract speech features. To verify the high-performance extraction effect of BiLSTM on speech features, this work analyzes the model performance when using LSTM and BiLSTM. The experimental results are shown in Fig. 8.

After embedding BiLSTM module, the speech recognition performance of FF-ATT-BiLSTM can be improved to a certain extent. Specifically, the accuracy and F1 score on SRA are improved by 1.2 and 1.0%, respectively. Similarly, the accuracy and F1 score on SRB are improved by 1.3 and 0.9%, respectively.

4.8 Analysis on dropout

FF-ATT-BiLSTM uses Dropout to enhance neural network. To verify the role of Dropout in speech recognition, this work analyzes the model performance when Dropout is not used and when Dropout is used. The experimental results are shown in Table 5.

After embedding Dropout strategy, the speech recognition performance of FF-ATT-BiLSTM can be improved to a certain extent. Specifically, the accuracy and F1 score on SRA are improved by 1.0 and 0.7%, respectively. Similarly, the accuracy and F1 score on SRB are improved by 0.7 and 0.6%, respectively.

4.9 Analysis on early stopping

FF-ATT-BiLSTM uses Early Stopping to constrain neural network. To verify the role of Early Stopping in speech recognition, this work analyzes the model performance when Early Stopping is not used and when Early Stopping is used. The experimental results are shown in Fig. 9.

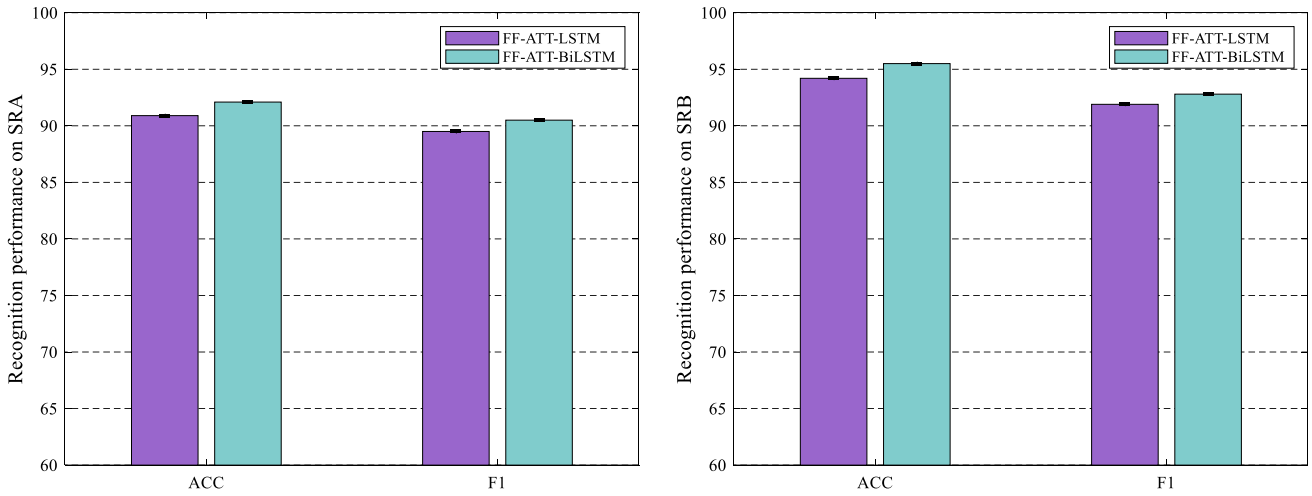

Fig. 8 Analysis on BiLSTM

Table 5 Analysis on Dropout

Method	SRA		SRB	
	ACC	F1	ACC	F1
Without Dropout	91.1%	89.8%	94.8%	92.2%
With Dropout	92.1%	90.5%	95.5%	92.8%

After embedding Early Stopping strategy, the speech recognition performance of FF-ATT-BiLSTM can be improved to a certain extent. Specifically, the accuracy and F1 score on SRA are improved by 1.8 and 2.0%, respectively. Similarly, the accuracy and F1 score on SRB are improved by 2.2 and 1.6%, respectively.

5 Conclusion

With the continuous popularity of multimedia intelligent devices, the realization of multimedia visual interaction has gradually become the focus of research. Therefore, human–computer interaction has become a key research field. The accuracy of speech recognition technology has far-reaching significance in this field. The accuracy of a speech recognition system depends heavily on the accuracy of the classification model and the quality of the derived acoustic data. Therefore, it is required to include as much information as possible in the sound content. At the same time, this should also minimize the interference of information irrelevant to classification. Only the distinguishable features extracted can make the features more conducive to recognition. This work proposes a new method for speech recognition in multimedia visual interaction. First of all,

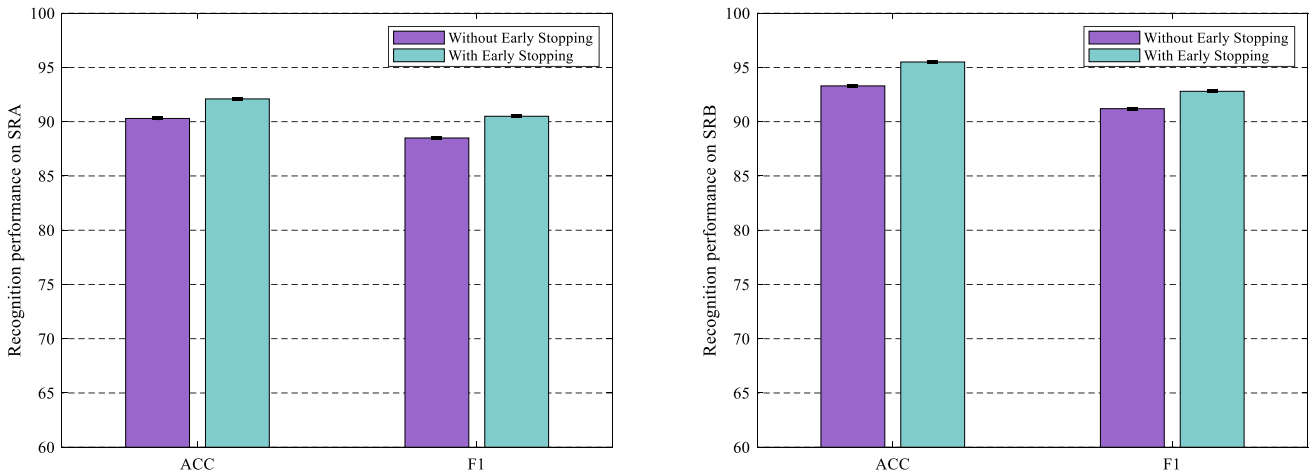

Fig. 9 Analysis on Early Stopping

this work considers the problem that a single feature cannot fully represent complex speech information. This paper proposes three kinds of feature fusion structures to extract speech information from different angles. This extracts three different fusion features based on the low-level features and higher-level sparse representation. Secondly, this work relies on the strong learning ability of neural network and the weight distribution mechanism of attention model. In this paper, the fusion feature is combined with BiLSTM and attention. The extracted fusion features contain more speech information with strong discrimination. When the weight increases, it can further improve the influence of features on the predicted value and improve the performance. Finally, this paper has carried out systematic experiments on the proposed method, and the results verify the feasibility.

Data availability The datasets used during the current study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare no conflict of interest exists.

References

1. Zgank A (2022) Influence of highly inflected word forms and acoustic background on the robustness of automatic speech recognition for human-computer interaction. Mathematics 10(5):711
2. Liu M (2022) English speech emotion recognition method based on speech recognition. Int J Speech Technol 25(2):391–398
3. Šumak B, Brdnik S, Pušnik M (2022) Sensors and artificial intelligence methods and algorithms for human–computer intelligent interaction: a systematic mapping study. Sensors 22(1):20
4. Liu Y, Sivaparthipan CB, Shankar A (2022) Human–computer interaction based visual feedback system for augmentative and alternative communication. Int J Speech Technol 1:1–10
5. Sang Y, Chen X (2022) Human-computer interactive physical education teaching method based on speech recognition engine technology. Front Public Health 10:941083–941097
6. Markl N, Lai C(2021) Context-sensitive evaluation of automatic speech recognition: considering user experience & language variation[C]. In: Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pp 34–40
7. Oh EY, Song D (2021) Developmental research on an interactive application for language speaking practice using speech recognition technology. Educ Tech Res Dev 69(2):861–884
8. Ran D, Yingli W, Haoxin Q (2021) Artificial intelligence speech recognition model for correcting spoken English teaching. J Intell Fuzzy Syst 40(2):3513–3524
9. Fu Q, Fu J, Zhang S et al (2021) Design of intelligent human-computer interaction system for hard of hearing and non-disabled people. IEEE Sens J 21(20):23471–23479
10. Pei J, Yu Z, Li J, et al (2022) TKAGFL: a federated communication framework under data heterogeneity. IEEE Trans Netw Sci Eng 1:1–11
11. Weng Z, Qin Z, Tao X, et al 2023 () Deep learning enabled semantic communications with speech recognition and synthesis. IEEE Trans Wirel Commun 1:6227–6240
12. Subramanian AS, Weng C, Watanabe S et al (2022) Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. Comput Speech Lang 75:101360
13. Oruh J, Viriri S, Adegun A (2022) Long short-term Memory Recurrent neural network for Automatic speech recognition. IEEE Access 10:30069–30079
14. Fendji JLKE, Tala DCM, Yenke BO et al (2022) Automatic speech recognition using limited vocabulary: a survey. Appl Artif Intell 36(1):2095039
15. Bhangale KB, Kothandaraman M (2022) Survey of deep learning paradigms for speech processing. Wirel Pers Commun 125(2):1913–1949
16. Dua S, Kumar SS, Albagory Y et al (2022) Developing a speech recognition system for recognizing tonal speech signals using a convolutional neural network. Appl Sci 12(12):6223
17. Gupta AK, Gupta P, Rahtu E (2022) FATALRead-fooling visual speech recognition models: put words on lips. Appl Intell 1:1–16
18. Lu Y J, Chang X, Li C, et al (2022) ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding. arXiv preprint arXiv:2207.09514
19. Agarwal P, Kumar S (2022) Electroencephalography-based imagined speech recognition using deep long short-term memory network. ETRI J 44(4):672–685
20. Shashidhar R, Patilkulkarni S, Puneeth SB (2022) Combining audio and visual speech recognition using LSTM and deep convolutional neural network. Int J Inf Technol 14(7):3425–3436
21. Baum LE (1972) An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. Inequalities 3(1):1–8
22. Graves A, Graves A (2012) Connectionist temporal classification. Supervised Seq Labell Recurr Neural Netw 1:61–93
23. Chan W, Jaitly N, Le Q, et al (2016) Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4960–4964
24. Graves A (2012) Sequence transduction with recurrent neural networks. arXiv preprint arXiv:1211.3711
25. Waibel A, Hanazawa T, Hinton G et al (1989) Phoneme recognition using time-delay neural networks. IEEE Trans Acoust Speech Signal Process 37(3):328–339
26. Liu H, Zhao L (2019) A speaker verification method based on TDNN–LSTMP. Circ Syst Signal Process 38:4840–4854
27. Normandin Y (1996) Maximum mutual information estimation of hidden Markov models. Autom Speech Speak Recog: Adv Top 1:57–81

28. Cho K, Van Merriënboer B, Gulcehre C, et al (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
29. Bahdanau D, Chorowski J, Serdyuk D, et al (2016) End-to-end attention-based large vocabulary speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4945–4949
30. Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Adv in Neural Information Processing Systems 30:1–11
31. Zhou S, Dong L, Xu S, et al (2018) Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin Chinese. arXiv preprint arXiv:1804.10752
32. Zhang Y, Lu X (2018) A speech recognition acoustic model based on LSTM-CTC[C]. In: IEEE 18th International Conference on Communication Technology (ICCT). IEEE, pp 1052–1055
33. Zhang S, Lei M, Yan Z, et al (2018) Deep-FSMN for large vocabulary continuous speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5869–5873
34. Cheng X, Xu M, Zheng TF (2020) A multi-branch ResNet with discriminative features for detection of replay speech signals. APSIPA Trans Signal Inform Process 9:28
35. Sivaram G, Nemala S K, Elhilali M, et al (2010) Sparse Coding for Speech Recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp 4346–4349
36. Chen S, Saunders D (2001) Atomic decomposition by basis pursuit. SIAM Rev 43(1):129–159