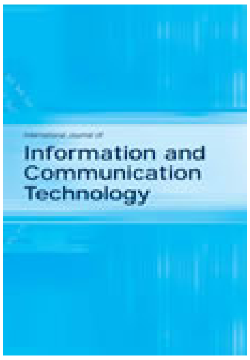
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Effectiveness analysis of speech visualisation technology applied to English speech teaching

Zhumin Huang

DOI: 10.1504/IJICT.2025.10071259

Article History:

Received:	09 October 2024
Last revised:	20 February 2025
Accepted:	20 February 2025
Published online:	11 June 2025

Effectiveness analysis of speech visualisation technology applied to English speech teaching

Zhumin Huang

School of Foreign Languages,
Nanchang Institute of Technology,
330044, China
Email: huangzhumin2024@163.com

Abstract: At present, English speech teaching has developed into an intelligent form, and speech recognition function is combined in speech teaching to perform spoken English correction, but the accuracy of speech recognition needs to be improved. In order to improve the effect of speech intelligent recognition in English speech teaching, this paper combines speech recognition technology with visualisation technology to propose a speech recognition visualisation technology, designs and implements ASR algorithm based on Conformer encoder and CTC decoder, and realises VITS speech synthesis model. At the same time, this paper uses knowledge distillation method to obtain a lightweight speech synthesis model, uses MobileNetV3 network to realise the lightweight YOLOv5s model, and combines DeepSORT tracking algorithm and statistical function to realise the target statistical function. According to the comprehensive test results, it can be seen that the model proposed in this paper has high speech recognition accuracy and speed. In addition, it can be seen from the comparative test results that the model proposed in this paper has certain advantages in speech recognition compared with the existing research, and can meet the needs of intelligent English pronunciation teaching.

Keywords: speech; visualisation; English; speech; teaching effectiveness.

Reference to this paper should be made as follows: Huang, Z. (2025) ‘Effectiveness analysis of speech visualisation technology applied to English speech teaching’, *Int. J. Information and Communication Technology*, Vol. 26, No. 17, pp.16–40.

Biographical notes: Zhumin Huang is currently a Lecturer at Nanchang Institute of Technology, specialising in applied linguistics research and English language teaching. Her research focuses on second language acquisition, language teaching methodology, and corpus linguistics. In recent years, she has conducted in-depth research in technology-enhanced language learning and academic English writing. She has led and participated in several research projects and teaches courses including College English, Academic English Writing, and Introduction to Linguistics, emphasising task-based teaching and blended learning.

1 Introduction

Educational informatisation has laid an environmental foundation for multi-modal learning data analysis, and learners' online learning behaviour based on the rich resources and diverse functions of the platform contains a large amount of multi-modal data. Each information acquisition method can be called a 'modal', and the data of two or more different representations or source channels is called multi-modal data (Alabsi, 2020), such as the recorded audio, expressions, EEG, gestures and other data of the learning process. Multi-modal data in online learning environment mainly comes from two aspects. One is online learning system, learning management system, information service system and other platforms, and the other is digital tools such as intelligent recording and broadcasting system, wearable devices, EEG, and eye trackers. The data of teacher-student and student interaction behaviour in online learning environment can be obtained through questionnaires and platform log files, but the emotional and cognitive changes of learners in the learning process are difficult to be reflected through questionnaires and other data. During the learning process, learners' emotional and cognitive input fluctuates, which makes it difficult to comprehensively evaluate learners' learning process by questionnaire pre-test and post-test. By analysing and fusing multi-modal data, learning behaviour, cognition, belief and emotion can be deeply excavated and explored (Al-Jarf, 2021), so as to grasp the input of cognition, behaviour and emotion. In addition, emotional, cognitive and behavioural involvement is the key factors affecting learning effect and satisfaction. Therefore, it is of great significance for students, teachers and educators to properly intervene and improve the teaching process by collecting and analysing multi-modal data, so as to improve the learning effect (Dizon and Thanyawatpokin, 2021).

After experiencing the stage driven by economic development, it has changed from the growth mechanism mainly relying on factor input to the improvement of market environment and the accumulation of human capital. Higher English pronunciation education institutions have become institutions that store and disseminate knowledge. The high-level development of data-driven higher English pronunciation education needs to meet the urgent needs of national and social development as soon as possible. At the same time, a series of information generated in education, teaching and campus activities, from students' learning behaviour to data monitoring of colleges and universities forms educational big data, which can clearly construct every teaching carrier. Moreover, how to deeply integrate it into the development of English pronunciation education and lead the innovation of educational informatisation concept has become a hot topic of attention from all walks of life and scholars. Through in-depth mining of English pronunciation education data, we can stimulate the vitality of internal elements of colleges and universities, further optimise the allocation of internal resources of higher education, and comprehensively improve the core quality of higher education. In addition, adopting related artificial intelligence technologies such as big data analytics and trend forecasting is changing many fields. Data-driven decision-making has been shown to be effective in increasing productivity. In the field of higher education, the management basis of macro-strategy of colleges and universities is to systematically collect and analyse a large amount of data materials of specific activities and measure the 'output' and 'income' of these activities. After a large amount of human and financial investment, the data collected and analysed has become the most important asset of colleges and universities. Therefore, the demand for de-isolation of data-driven higher education decision-making in the organisational process is gradually increasing. Simultaneously, comprehensive research and judgement of various types of data and attention to more reliable data in the process of reasonable expectation planning and efficient resource utilisation undoubtedly puts forward higher requirements for the evaluation and analysis based on comprehensive data of higher education.

In order to improve the effect of speech intelligent recognition in English speech teaching, this paper combines speech recognition technology with visualisation technology to propose a speech recognition visualisation technology, designs and implements ASR algorithm based on Conformer encoder and CTC decoder, and realises VITS speech synthesis model. At the same time, this paper uses knowledge distillation method to obtain a lightweight speech synthesis model, uses MobileNetV3 network to realise the lightweight YOLOv5s model, and combines DeepSORT tracking algorithm and statistical function to realise the target statistical function.

2 Related work

2.1 Speech recognition algorithm based on phone recogniser (PR)

Phoneme is an acoustic unit divided according to the pronunciation characteristics of speech. Generally speaking, phones correspond to people's pronunciation actions one-to-one. For different language types, even if the phone composition is similar, the statistical information reflected by the phone composition is different. By using the difference information between different speeches, different speeches can be distinguished, that is, speech recognition using phone information. In this kind of speech recognition algorithm, the speech signal needs to be further processed. The processing process is to use the trained PR to extract the mono-phone or tri-phone state of the speech signal, and then combine it with the language model to discriminate the speech. Therefore, this speech recognition algorithm is usually called phone recogniser algorithm followed by language model (PRLM) (Du et al., 2022). Because these phone states corresponding to different speeches are not the same, when the number of speeches to be recognised is not large, the speech recognition algorithm can be divided into two parts: front-end and back-end, and the acoustic model of each speech is taken as the front-end, and then the language model is used to score and discriminate the speech in the back-end part of the speech recognition algorithm. This method uses multiple parallel PRs at the front end of the algorithm, so it is also called the parallel phone recogniser algorithm followed by language model (PPRLM) (Ekayati, 2020). Although the performance of the speech recognition algorithm is improved to a certain extent after using the phone information of multiple languages, the features used by training the PR are relatively single, the speech recognition performance is easily affected by the PR, and the training of the PR requires a large number of labelled samples, which poses great challenges to PR-based speech recognition algorithms (Evers and Chen, 2022).

2.2 Speech recognition algorithm based on underlying acoustic features

Speech recognition algorithm based on underlying acoustic features is an algorithm that uses the underlying acoustic features of speech signals to distinguish which speech an audio signal belongs to. Because the underlying acoustic features of different speeches reflect different statistical characteristics, different speeches can be distinguished according to this feature (Hsu et al., 2023). Because the features used by speech recognition algorithms based on underlying acoustic features can be extracted from the speech spectrum of speech signals, and there is no need for a lot of manual labelling of audio samples, this method has been focused on by relevant scholars for many years. The most commonly used parameterised underlying feature in speech recognition algorithms is the Mel frequency cepstrum coefficient (MFCC). In the research process of speech recognition algorithms, the proposed shifted delta cepstra (SDC) has also played a very good role in promoting the development of speech recognition algorithms based on underlying acoustic features. On the basis of shifted differential cepstrum features, the speech recognition method of Gaussian mixture model-universal background model (GMM-UBM) has promoted the rapid development of speech recognition algorithms based on underlying acoustic features (Jia and Hew, 2022). At that time, a serious problem in the field of speech recognition was the low performance of the algorithm caused by insufficient speech data. Therefore, the proposed model made the speech recognition algorithm overcome the problem caused by insufficient data to a certain extent. However, as a generative statistical modelling method, GMM-UBM cannot effectively distinguish the speech features of mixed speech, so discriminative modelling algorithms based on GMM-UBM are gradually proposed. Although these proposed methods improve the performance of speech recognition algorithms to a certain extent, they do not deal with the differences between speech and channel space well. On the other hand, the covariance matrix of UBM model needs to be estimated in GMM-UBM modelling, and the demand for speech data in the parameter estimation process is still large. If enough speech data cannot be provided, the performance of speech recognition algorithm based on GMM-UBM still cannot reach the optimal (Kumar et al., 2022). Milliner and Dimoski (2024) proposed the use of factor analysis (FA) method to process the underlying acoustic features and reduce the negative impact of noise on the speech recognition algorithm. Moreover, the GMM mean super-vector dimension of speech is reduced. This method makes the original estimation of UBM covariance matrix become the estimation of low-rank matrix covariance, which reduces the calculation amount of the algorithm to some extent. In addition, joint factor analysis (JFA) in speech print recognition algorithm has also been gradually applied to speech recognition algorithm, and has achieved good recognition results (Mukhamadiyev et al., 2022). The total variability (TV) method greatly improves the recognition performance of speech recognition algorithms based on underlying acoustic features. The TV method is based on the total difference space, which transforms the high-dimensional mixed Gaussian model super-vector of each audio data into a lower-dimensional vector representation without losing effective speech information as much as possible, thus ensuring the performance of the speech recognition algorithm. This low-dimensional vector representation is the identity authentication vector i-vextor. Up to now, the i-vextor-based speech recognition algorithm still has important research value and application value in the field of speech recognition (Nguyen and Pham, 2022).

2.3 Speech recognition algorithm based on deep learning

In recent years, the continuous advent of intelligent speech technology, deep learning technology, artificial intelligence and other related technical means has also injected new vitality into the development of speech recognition algorithms. Researchers' research ideas on speech recognition algorithms are mainly to divide the algorithms into front-end and back-end, and then study the speech recognition methods based on deep learning from three aspects: front-end frame-level feature extraction, back-end sentence-level full difference space modelling and end-to-end recognition directly targeting speech (Ran et al., 2021). Among the improvements in acoustic features and TV modelling, the contribution of senone-based deep neural network (DNN) to speech recognition algorithms is the most prominent. In the acoustic feature extraction part of the front end of the speech recognition system, Shadiev et al. (2020) used the output layer data of the DNN to calculate the posterior probability of each senone on each frame of the speech signal to obtain the acoustic features of the audio signal. Then, using the DNN middle layer data, an additional DNN indicating phoneme information is trained using data unrelated to the training set. The bottleneck layer data of DNN is a new speech recognition feature, that is, deep bottleneck characteristics (DBF). This neural network model is naturally called deep bottleneck network (DBN) (Tai and Chen, 2021). This speech recognition method is called DBF-TV method, and there are many speech recognition algorithms based on this method in recent years. In terms of back-end modelling improvement, it is mainly the improvement of TV algorithm. The reason for this is to obtain i-vextor (Tsai, 2023), which represents speech information more accurately.

3 Algorithm design and verification

3.1 Function overview

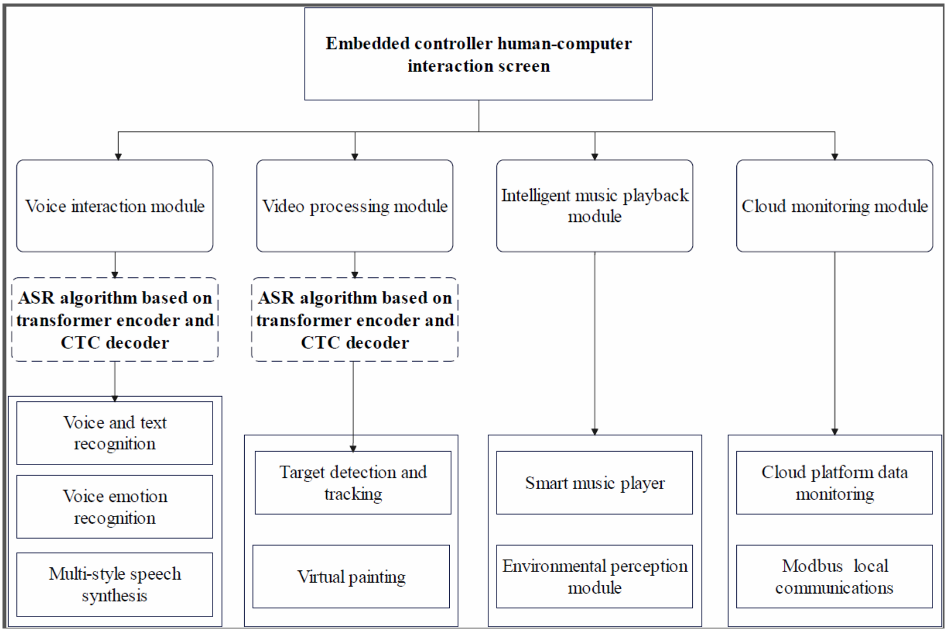
The research content of this paper is summarised, and the research ideas and model construction process are obtained. An overview of the functions of intelligent interactive devices based on speech and video processing is shown in Figure 1. At the software level, there are four parts: speech interaction module, video processing module, intelligent music playback module and cloud monitoring module, among which speech interaction and video processing modules are the key functions, and all of them rely on deep learning algorithms. Therefore, this chapter will design the corresponding algorithm according to the functional needs of the above modules, and carry out experimental verification on the PC side. The main contribution points are:

1 In speech interaction: ASR module based on conformer and CTC is designed, which achieves high recognition accuracy and small number of parameters. Based on the idea of knowledge distillation, the lightweight of VITS speech synthesis model is completed, and the real-time performance of speech interaction on embedded hardware platform is further improved. Then, the SER function is realised based on emotion2vec speech emotion feature extraction model, which improves the user experience of speech interaction.

2 In terms of video processing, the lightweight of YOLOv5s model is realised by using MobileNetV3 lightweight network, and the target statistical function is realised by combining DeepSORT tracking algorithm and statistical function. Based on the COCO dataset, the scene label is obtained, and the scene application of the target detection module is realised. In order to further improve the interactive experience,

the virtual painting function is implemented based on MediaPipe gesture recognition and OpenCV video processing library (Van et al., 2021).

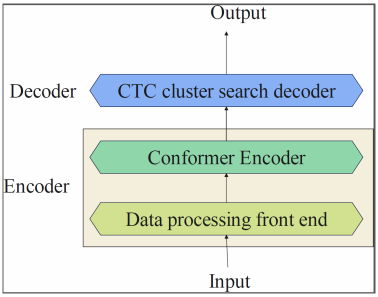
Figure 1 Function overview of intelligent interactive device based on speech and video processing


Embedded controller human-computer interaction screen
Voice interaction module
Video processing module
Intelligent music playback module
Cloud monitoring module
ASR algorithm based on transformer encoder and CTC decoder
ASR algorithm based on transformer encoder and CTC decoder
Voice and text recognition
Voice emotion recognition
Multi-style speech synthesis
Target detection and tracking
Virtual painting
Smart music player
Environmental perception module
Cloud platform data monitoring
Modbus local communications

3.2 *Design and verification of speech and text recognition algorithm*

In this section, a complete end-to-end ASR model based on Conformer and CTC is designed for ASR tasks. The model structure is shown in Figure 2, which is an encoder-decoder structure. Among them, Encoder consists of conformer module and data processing front-end. The data processing front-end consists of three parts: data pre-processing, feature extraction and data enhancement. Decoder is implemented by a CTC module decoded using a bundled search algorithm. At the same time, the model is trained and tested on the PC side.

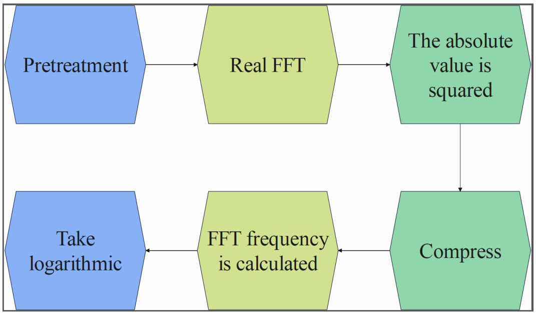
Figure 2 Architecture diagram of speech and text recognition module (see online version for colours)



3.2.1 Data processing front-end

1 Mandarin feature extraction: in the ASR task, the acoustic features obtained by processing the Mel spectrum can be used as the input of the acoustic model. In this paper, the linear energy map method is used for audio pre-processing, and its processing flow is shown in Figure 3.

Figure 3 Linear audio pre-processing (see online version for colours)



In this paper, it is assumed that the real FFT result is $X(f)$, and its dimension is $N \times M$, where N represents the number of frequency points, M represents the number of time windows, and the window function $w(t)$ is a Hamming window with length L. Then, the linear pre-processing process is shown in the following formula:

$$Y(f) = |X(f)|^2 \tag{1}$$

$$S = \sum_{i=0}^{L-1} w(i)^2 \tag{2}$$

$$Y'(f) = \frac{2}{S} Y(f) \tag{3}$$

Among them, S is the total energy of the window function, and the compression factor $2/S$ is the result of removing the frequency of 0 and the highest frequency. Finally, the linear energy map of speech spectrum is obtained, that is, the power distribution of speech in frequency domain. The linear energy map contains most of the speech information and can be used as input for ASR tasks. Since the length of audio files may be different, which results in an inconsistent number of features extracted from each audio file, a uniform audio length is required. As shown in Figure 4, this paper uses tail filling to unify the audio length. After reading a batch of data, the data is sorted to find the longest audio label, create a zero tensor according to the longest audio length, and then fill the original data into the zero tensor to obtain audio data with uniform length.

2 Data enhancement

The intelligent interactive device is placed in the complex outdoor place of users, so the speech interactive module should have good robustness. In this paper, three data enhancement methods, noise enhancement, speech speed enhancement and volume enhancement, are designed to improve the robustness of ASR model.

As a completely end-to-end ASR model, conformer can realise the mapping of speech signals to text sequences without relying on language models, greatly simplifying the recognition process. Moreover, conformer uses feature coupling units to fuse CNN and transformer together, making full use of the advantages of CNN in local features and transformer in global planning, so it has certain advantages in model size and training speed (Zhang and Zhang, 2022).

CER is the ASR general performance index character error rate (CER), and it is as follows:

$$CER = \frac{(insert + delete + replace)}{Total\ number\ of\ words} \tag{4}$$

Among them, the insertion, deletion and replacement are parts that need to be adjusted after comparing the predicted text with the correct text.

The conformer encoder architecture is shown in Figure 5. The acoustic features of speech data are obtained through the data processing front end. The acoustic features and text data are input to the Conformer module at the same time, and the probability distribution and context information of the text corresponding to the acoustic features can be obtained.

Figure 4 Schematic diagram of unified audio length processing (see online version for colours)

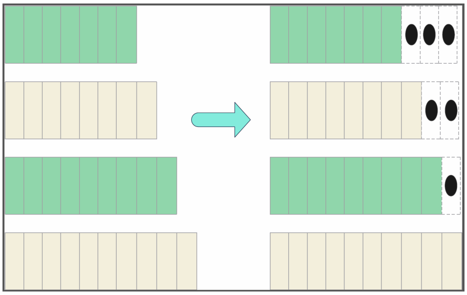
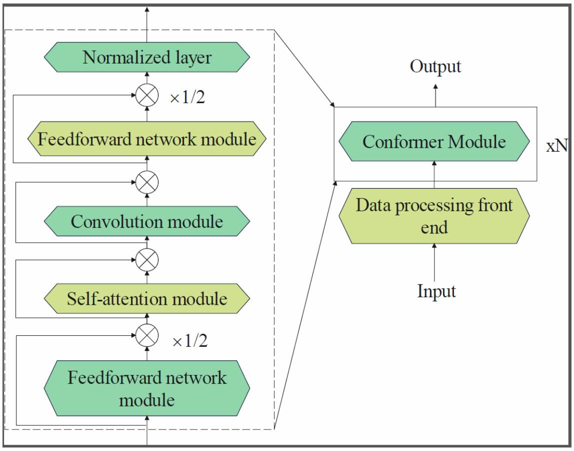
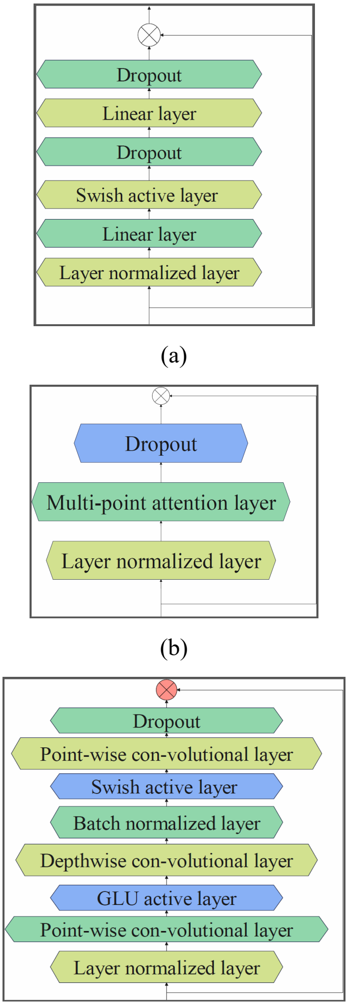


Figure 5 Conformer encoder architecture (see online version for colours)



The conformer encoder is composed of N Conformer modules, and each conformer module is composed of a feedforward network module, a self-attention module, a convolution module, and another feedforward network module in turn. The architecture of each module is shown in Figure 6.

Figure 6 Conformer module architecture, (a) feedforward network module, (b) self-attention module, (c) convolution module (see online version for colours)



(a)

(b)

(c)

The feedforward network module [Figure 6(a)] is usually composed of one or more fully connected layers, which can carry out nonlinear transformation on the input features to extract higher-level feature representation. These feature representations are then used in the subsequent decoding process to generate the final speech recognition results.

The self attention module [Figure 6(b)] constructs an attention matrix by calculating the attention weight of each element in the sequence to other elements. This matrix reflects the relative importance or correlation between elements in the sequence. Then, the model uses this matrix to calculate the weighted sum of the input sequence to generate a new feature representation. These feature representations fuse the global information in the sequence and help the model to make more accurate predictions or classifications.

The deep separable convolution layer in the convolution module [Figure 6(c)] can effectively reduce the number of model parameters, which not only helps to accelerate the model training and reasoning process, but also prevents over-fitting to a certain extent. At the same time, the use of residual connection can better alleviate the problem of model degradation, accelerate the convergence speed of the model, and further improve the stability and performance of the model.

In the encoder part, X_i represents the input of the i^{th} conformer module, and Y_i represents its corresponding output. The calculation process in the conformer module is shown in formulas (5) to (8):

$$X_i = X_i + \frac{1}{2} FFN\left(X_i\right) \tag{5}$$

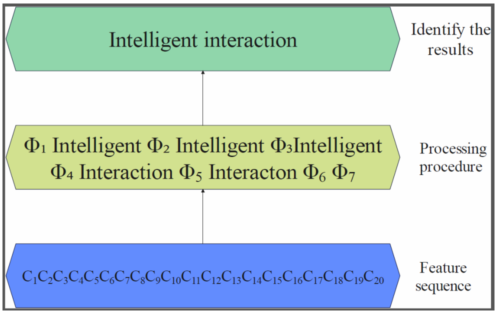
$$X_i' = X_i + MHA\left(X_i\right) \tag{6}$$

$$X_i'' = X_i' + Conv\left(X_i'\right) \tag{7}$$

$$Y_i = Layernorm\left(X_i'' + \frac{1}{2} FFN\left(X_i''\right)\right) \tag{8}$$

Among them, *FFN* represents a feed-forward neural network module, *MHA* represents multi-head attention module, *Conv* represents a convolution module, and *Layernorm* represents the normalisation operation.

Figure 7 Schematic diagram of CTC decoding (see online version for colours)


Intelligent interaction
Identify the results
Φ1 Intelligent Φ2 Intelligent Φ3Intelligent
Φ4 Interaction Φ5 Interacton Φ6 Φ7
Processing procedure
C1C2C3C4C5C6C7C8C9C10C11C12C13C14C15C16C17C18C19C20
Feature sequence

3.2.2 CTC decoder

The CTC algorithm is a loss function for sequence labelling problems, which is particularly suitable for dealing with the alignment problem of input and output labels by expanding the set of labels and introducing the concept of blank. As a standard sequence labelling task, ASR usually requires the input data to be completely aligned with the output labels at every moment. However, the introduction of CTC allows training when the input and output sequences are inaccurately aligned, and blank is inserted between repeated labels as labels, effectively reducing the error rate when merging labels, thereby completing ASR tasks more efficiently.

As shown in Figure 7, it assumes that the input feature sequence is 20 frames, and the correct result is 4 characters.

CTC decoding is the process of finding probability maximisation, as shown in formula (9). If $C = \{c_1, c_2, \cdots, c_T\}$ is assumed to be the input feature sequence, the goal of CTC decoding is to find the output label sequence Y^* with the highest probability. Therefore, the decoding algorithm based on CTC is shown in formula (10). CTC will calculate the output unit with the highest probability at each moment $t(1 \leq t \leq T)$, and then delete repeated characters and blank tags to obtain the output sequence Y^*.

$$Y^* = \arg\max_{Y} \{P(Y \mid C)\} \tag{9}$$

$$Y^* = \arg\max_{Y} \prod_{t=1}^{T} P\left(y_t \middle| c_t\right) \tag{10}$$

In this paper, beam search (BS) is used as the prefix algorithm of CTC decoding process to complete the decoding task. BS algorithm is an efficient heuristic graph search algorithm. By introducing the beam width (BW), only W possible solutions with the highest probability of each path can be retained by clipping among multiple 'bundles' output by decoding task, and then the optimal solution can be selected by comparing all 'bundles'. The decoding process of the BS algorithm is shown in formula (11):

$$Y_t = \underset{y_{1,[t]}, \cdots y_{w,[t]} \in y_y}{\arg\max} \; score\left(y_{w,[t-1]}, y_{w,[t]} \middle| x\right) a \tag{11}$$

Among them, x represents the input feature, Y_t refers to the optimal output sequence at time t, y_t and y_{t-1} refer to the output sequences at time t and $t - 1$, respectively, and u$Y_{w,[t]}$ refers to the specific sequence considered after adding the bundle width w,[t] constraint. The whole calculation process is based on the scoring function score, and the process of finding the solution with the maximum probability among all the best solutions constrained by w,[t]. The score scoring function is shown in formula (12), and its meaning is to obtain the logarithm of the output probability of the current time step given the output and input conditions of the previous time step.

$$score\left(y_t, y_{t-1} \middle| x\right) = \sum \log p\left(y_t \middle| y_{t-1}, x\right) \tag{12}$$

After introducing the BS search algorithm, the whole decoding process is shown in Figure 8. The decoder can traverse all possible solutions and automatically tailor the possibilities with low probability, which improves the decoding accuracy and reduces the computational difficulty.

Combined with the CTC + BS decoder, the process of character recognition of the input speech by the ASR module is as follows:

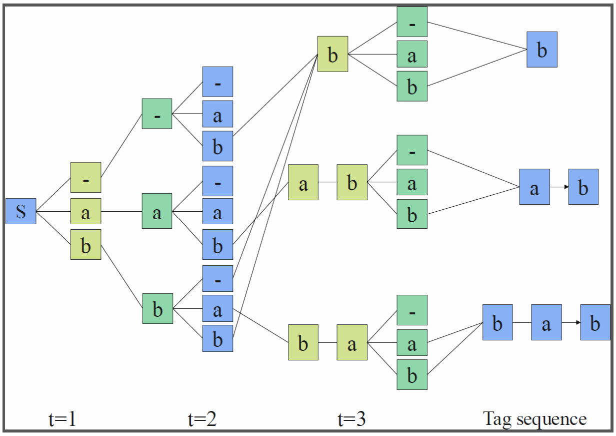
1 It performs feature extraction on the input speech signal and obtains an acoustic feature sequence S.

2 It inputs the acoustic feature sequence S into the conformer encoder, and obtains the probability distribution of the audio signal in the time step through the decoder, that is, the probability distribution O of the characters and phonemes that may correspond

to each time step. In the conformer encoder, this includes not only the probability distribution of phonemes to text, but also the probability distribution of text context information.

3 BS algorithm is used to decode CTC Loss, so as to solve the optimal sequence distribution according to the probability distribution O, that is, the final recognition result.

Figure 8 CTC + BS decoding process (see online version for colours)



3.3 *Design and verification of speech emotion recognition (SER) algorithm*

The diversity and uncertainty of speech emotion expression make speech emotion features complex and diverse. Therefore, for SER task, traditional feature extraction methods using MFCC or Fbank cannot meet the requirements in semantic information. At present, the mainstream method is to extract emotional features from pre-trained self-supervised learning models based on speech, but there is no good general model. In this paper, based on the emotion2vec basic model of general emotion representation, the emotion features of CA-ES dataset are extracted, and the SER model is built by using bi-directional long short-term memory (Bi-LSTM) neural network.

Knowledge distillation is a model compression and information extraction technology, which takes a complex teacher model and training data as the training input of the student model at the same time. In this way, the student model can learn the features of the teacher model and reduce the number of parameters at the same time. The model includes a CNN-based feature extractor F and a transformer-based backbone network B. When the original audio $X = [x_1, \cdots, x_{Nx}]$ is given, the teacher model T and the student model S use their respective feature extractors F^T and S^T to obtain the downsampled feature $Z = [z_1, \cdots, z_{Nz}]$, as shown in the following formula:

$$Z^T = F^T(X) \tag{13}$$

$$Z^S = F^S(X) \tag{14}$$

In the teacher model T, the downsampled feature Z^T is directly input into the backbone network B^T. However, in the student model S, random masking and utterance-level embedding operations are performed on the downsampled feature Z^S, and the next i-frame information is continuously masked from a certain frame with probability P, then the learnable utterance embedding $E = [e_1, \cdots, e_{Ne}]$ is added and finally input to the backbone network B^S of the student model, as shown in formulas (15) and (16). Through random masking and utterance-level embedding operations, the model's understanding of input features is enhanced, and the overall semantic and local features of speech are better captured, thus improving the learning efficiency and generalisation ability of the model.

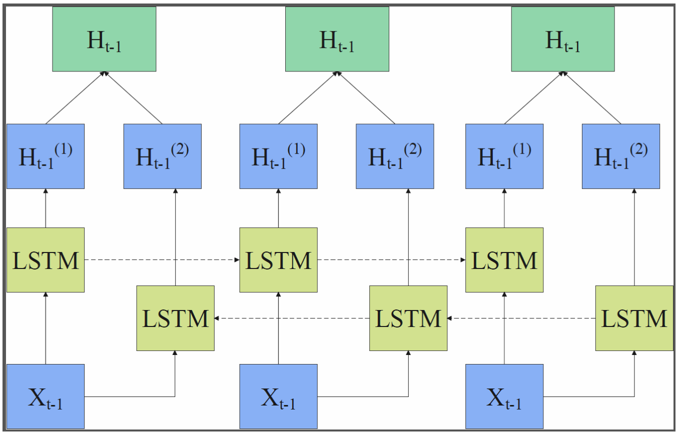
$$Y^T = \frac{1}{k}\sum_{i=1}^{k} B_T^i \left(Z^T\right) \tag{15}$$

$$U^S, Y^S = B^S\left(U, Mask\left(Z^S\right)\right) \tag{16}$$

Among them, Y^T is the average output embedding value of the k^{th} layer (usually the top layer) of the multi-layer transformer module B_T^i in the teacher model backbone network B^T, U^S is the utterance-level output embedding, Y^S is the frame-level output embedding, U^S and Y^S together serve as the output of the student model backbone network B^S, and mask is the mask operation. The dimensions on the hidden layers of Y^T, Y^S, and U^S are the same, so the speech emotion feature dimensions output by the teacher model and the student model and the feature dimensions embedded at the utterance level are consistent. Y^T and Y^S are consistent in the temporal dimension, which also illustrates that they represent different levels of output, namely, utterance level and frame level, respectively.

In this paper, Bi-LSTM is used to design an emotion classification model. For speech data, LSTM can capture the above information and input it into the following, but it cannot transmit it in the reverse direction. Bi LSTM can realise the two-way transmission of information by capturing the context information in the two directions. As shown in Figure 9, the Bi LSTM is composed of two layers of reversed LSTMs. Moreover, the inputs of every two LSTMs at time steps are the same and they are both x_t, but the context direction is opposite, and the total output of each time step is $H_t = [h^{(1)}, h^{(2)}]$.

Figure 9 Schematic diagram of Bi LSTM structure (see online version for colours)



The SER model network based on Bi LSTM is shown in Figure 10, and the functions of each network layer are as follows:

1 The six-dimensional vector of speech emotion features is linearly transformed into 512 dimensions through the fully connected layer 1, which increases the representation ability of the model.

2 In the Bi LSTM layer, two LSTM networks learn the forward time series feature sequence and the reverse feature sequence respectively to obtain 512-dimensional hidden states respectively, and after splicing the hidden states, 1,024-dimensional hidden information with bidirectional context relationship is obtained, and output to the Tanh activation layer.

3 Tanh activation function can increase the nonlinear ability of the model and help the model learn more complex feature representations.

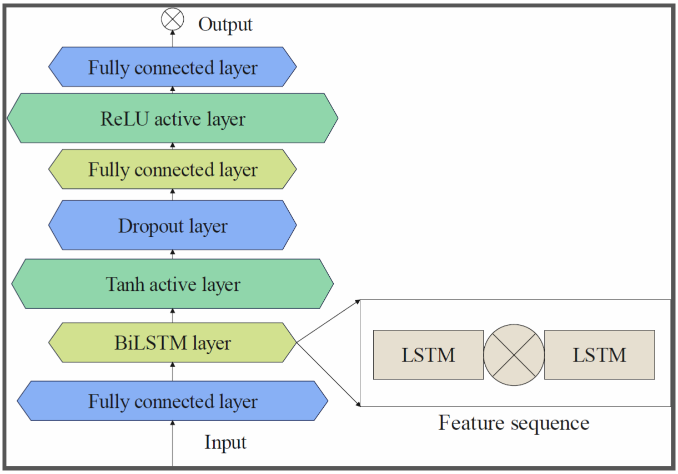
4 Dropout discard layer randomly discards some parameters to prevent overfitting and improve the generalisation ability of the model.

5 The fully connected layer 2 performs dimensionality reduction operation to reduce the number of parameters of the model, thereby reducing the amount of calculation and storage requirements, improving the efficiency of the model and reducing the risk of overfitting.

6 Re LU activation function, which sets the negative value of the feature to zero and retains the positive value, provides the sparsity of the network model and improves the training efficiency.

7 Fully connected layer 3 reduces the dimensions to 6 dimensions, and each dimension represents the score of an emotion category, and it finally obtains the optimal classification situation.

Figure 10 Architecture diagram of SER model (see online version for colours)



3.4 Lightweight and verification of speech synthesis algorithm

In this paper, the BZNSYP and AISHELL-3 datasets will be used to implement single-TTS with high quality and multi-TTS with diverse styles, respectively.

1 Loss function: the joint loss function is used in training to optimise the reconstruction loss and KL (Kullback-Leibler) divergence loss caused by VAE and the adversarial loss of GAN. The joint loss function L_s is as follows:

$$L_s = L_r + L_{kl} + L_{dur} + L_{adv} + L_{fin}(G) \tag{18}$$

Among them, L_r is the reconstruction loss function, which is used to measure the difference between synthesised speech and original speech, as shown in formula (19), and x_{mel} and $\hat{x}_{mel}$ are the Mel spectrum of original audio and synthesised speech, respectively.

$$L_r = \left\| x_{mel} - \hat{x}_{mel} \right\|_1 \tag{19}$$

L_{kl} is a KL divergence loss function, which can be used to constrain the regularisation of potential vector distribution. The KL divergence generalisation of glow-TTS is used for reference in VITS, and its expression is shown in formula (20):

$$L_{kl} = \log q_\varphi\left(z \middle| x_{lin}\right) - \log p_\theta\left(z \middle| c_{text}, A\right) \tag{20}$$

Among them, $q_\varphi(z| \ x_{lin})$ is the posterior distribution of the linear spectrum x output implicit variable z, $p_\theta(z| \ c_{text}, A)$ is the prior distribution of the implicit variable z given condition c output, and the expression of the implicit variable z is shown in formula (21):

$$z \sim q_\varphi\left(z \middle| x_{lin}\right) = N\left(Z; \mu_\varphi\left(x_{lin}\right), \sigma_\varphi\left(x_{lin}\right)\right) \tag{21}$$

The role of $L_{dur} + L_{adv} + L_{fin}$ is to optimise the adversarial loss caused by adversarial training, which is consistent with the loss function used by Hi Fi-GAN.

4 System test analysis

4.1 System and test environment

The system designed and implemented in this paper is shown in Figure 11.

From the figure, it can be seen that the model structure of this article is a hierarchical model, which constructs the overall system structure through five levels. This structure can not only improve the efficiency of speech recognition data processing, but also effectively enhance the internal aggregation effect of the system.

The system structure hierarchical design divides the system structure at the logical level. Based on this logical hierarchical structure, the overall architecture design of this system is shown in Figure 12.

There are two main types of data transmission between terminal modules and cloud background services: speech data and unified signalling. Among them, speech data is collected and generated in real time by the microphone array at the terminal, and system signalling can be divided into two categories: one is actively generated at the terminal, which is the terminal's report of its own status, and the other type is actively generated in the cloud, which is the state control or message push of the terminal by the cloud.

In the virtual interaction module based on video processing, this paper mainly uses Mozilla Common Voice dataset to train and verify the object detection model, while the virtual painting module uses the pre-trained model based on Media Pipe.

Figure 11 System structure hierarchy diagram (see online version for colours)

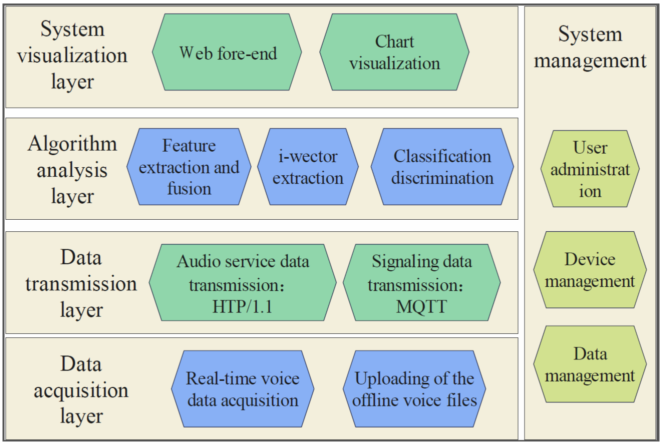

System visualization layer
Web fore-end
Chart visualization
System management
Algorithm analysis layer
Feature extraction and fusion
i-wector extraction
Classification discrimination
User administrat ion
Data transmission layer
Audio service data transmission: HTP/1.1
Signaling data transmission: MQTT
Device management
Data acquisition layer
Real-time voice data acquisition
Uploading of the offline voice files
Data management

Figure 12 System architecture design (see online version for colours)

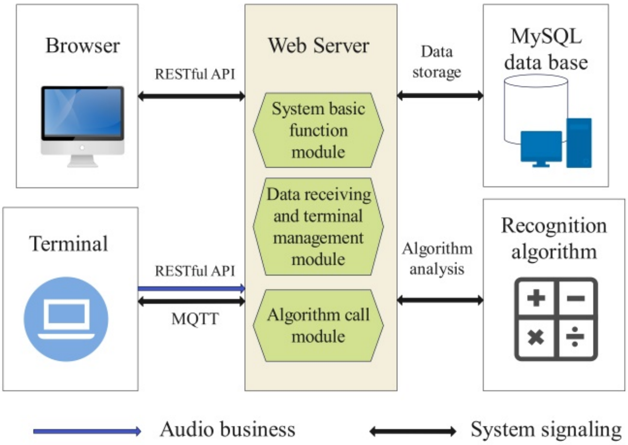


Table 1 Test hardware environment

Equipment	*Configuration*
Operating system	Windows11 Ultimate (64-bit)
Memory	64G
CPU	Intel i9-12900K
GPU	NVIDIAGeForceRTX3090 (24G)
Python	3.8.12
Cuda	cuda11.2 + cudnn8.2. 1

In the accuracy of speech recognition, several databases, such as Mozilla common voice, ljspeech, ryanspeech, Hi-Fi multi speaker English TTS dataset, were used to carry out experiments and comparisons.

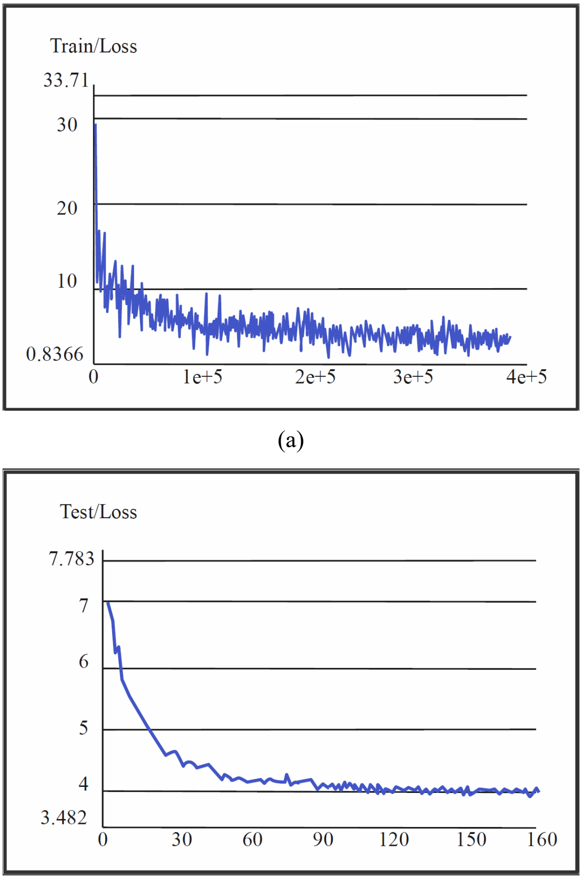
In this paper, the experiment is carried out in a model training environment built on a high-performance host. Table 1 shows the test hardware environment of this paper.

4.2 Results

The model is trained with batch_size = 128, num_workers = 24, and the AISHELL-1 dataset is used. Moreover, considering the size of the model, this test only train for 186 rounds, and the training time is about 60 hours. The results are shown in Figure 13.

The training results of the SER model are shown in Figure 14. After 200 rounds of training, the best loss is 0.22. Figures 14(b) and 15(c) show the change of the accuracy rate of the general performance index of the classification function, and the best accuracy rate can reach 90%.

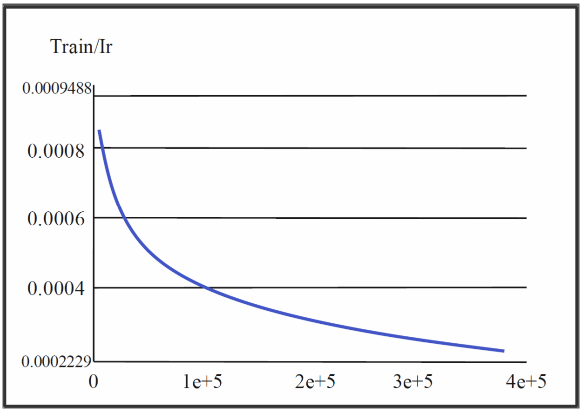
Figure 13 Training results of speech and text recognition model, (a) training set loss, (b) testing set loss, (c) training learning rate, (d) test set word error rate (see online version for colours)



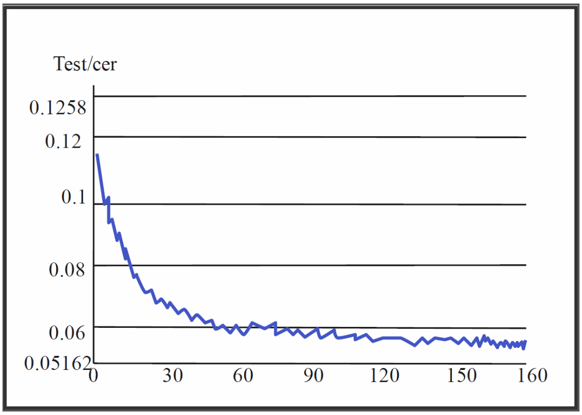
(a)

(b)

Figure 13 Training results of speech and text recognition model, (a) training set loss, (b) testing set loss, (c) training learning rate, (d) test set word error rate (continued) (see online version for colours)



(c)



(d)

VITS speech synthesis model is implemented using Pytorch framework, we text processing is used to normalise text, and bidirectional encoder representations from transformers (BERT) pre-training model is added to improve the naturalness of synthesised speech. The speaker vector matrix is embedded in VITS to improve the learning ability of the model to speaker labels, and at the same time, the initial weight is obtained, which saves training time.

In this speech synthesis test, a total of 180 k generations are iterated, and the training results are shown in Figure 15.

The comparison parameters in this paper include recognition time, accuracy, recall rate and F1 value. The models of Dizon and Thanyawatpokin (2021), Du et al. (2022), Ekayati (2020) and Kumar et al. (2022) are taken as controls, and the specific test results are shown in Table 2. The results of further comparison of speech recognition accuracy through multiple datasets are shown in Table 3.

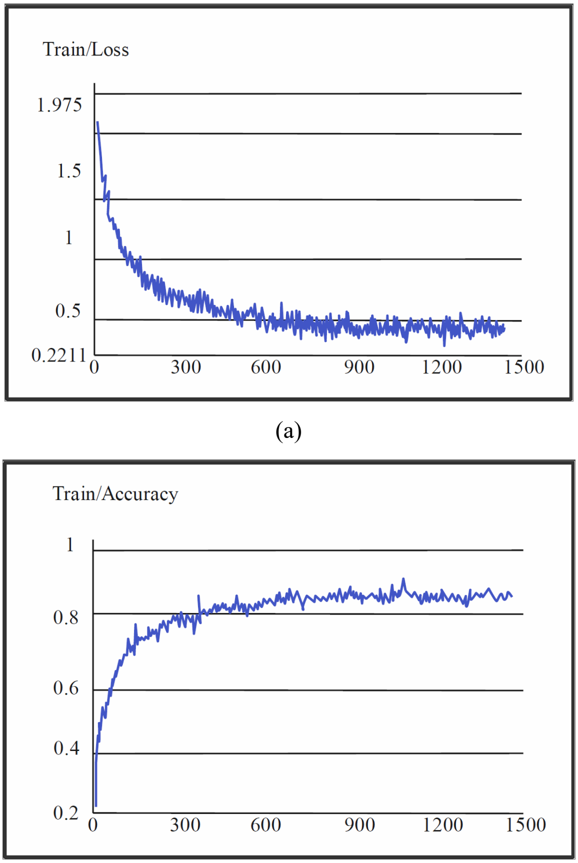
$$Rwecall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{22}$$

Among them, true positive refers to the case where the model prediction result is positive and the actual result is also positive, and false negative refers to the case where the prediction result is negative but the actual result is positive.

Table 2 Comparison of speech recognition effects

	Recognition time	*Accuracy*	*Recall*	*F1*
The models of Dizon and Thanyawatpokin (2021)	1.260	0.709	0.238	0.128
The models of Du et al. (2022)	1.397	0.710	0.168	0.133
The models of Ekayati (2020)	1.292	0.692	0.182	0.189
The models of Evers and Chen (2022)	0.964	0.763	0.177	0.107
The models of this paper	0.664	0.848	0.435	0.372

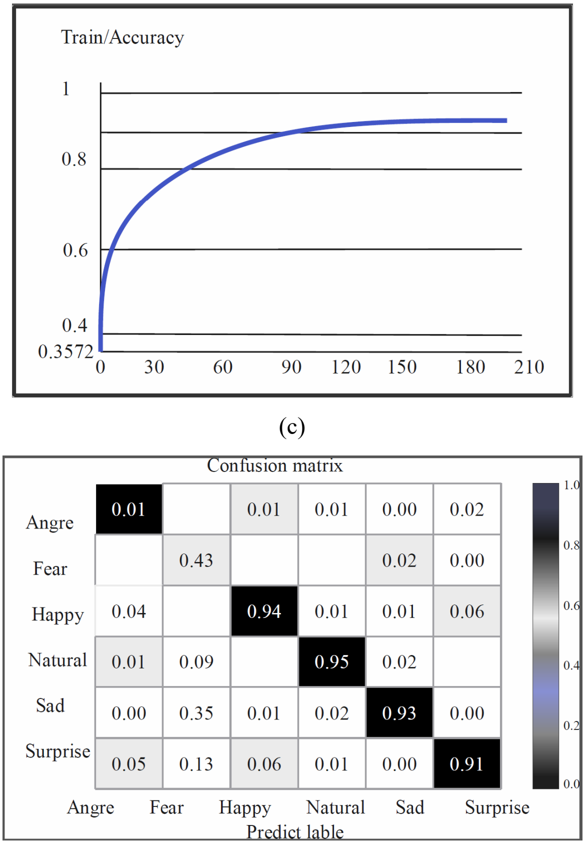
Figure 14 Training results of SER model, (a) training set loss, (b) training-set model classification accuracy, (c) test-set model classification accuracy, (d) confusion matrix (see online version for colours)



(a)

(b)

Figure 14 Training results of SER model, (a) training set loss, (b) training-set model classification accuracy, (c) test-set model classification accuracy, (d) confusion matrix (continued) (see online version for colours)

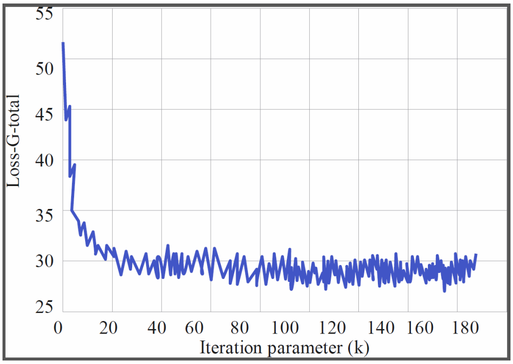


	Angre	Fear	Happy	Natural	Sad	Surprise
Angre	0.01		0.01	0.01	0.00	0.02
Fear		0.43			0.02	0.00
Happy	0.04		0.94	0.01	0.01	0.06
Natural	0.01	0.09		0.95	0.02	
Sad	0.00	0.35	0.01	0.02	0.93	0.00
Surprise	0.05	0.13	0.06	0.01	0.00	0.91

(c)

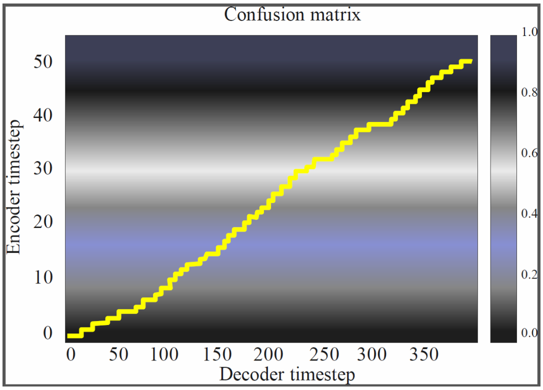
(d)

Figure 15 VITS model training results, (a) generator training loss, (b) attention matrix visualisation (see online version for colours)



(a)

Figure 15 VITS model training results, (a) generator training loss, (b) attention matrix visualisation (continued) (see online version for colours)



(b)

Table 3 Comparison of speech recognition accuracy of multiple datasets

	ljspeech	*ryanspeech*	*Hi-Fi multi speaker English TTS dataset*
The models of Dizon and Thanyawatpokin (2021)	0.770	0.733	0.756
The models of Du et al. (2022)	0.711	0.687	0.732
The models of Ekayati (2020)	0.742	0.770	0.729
The models of Evers and Chen (2022)	0.811	0.839	0.827
The models of this paper	0.847	0.881	0.881

4.3 Analysis and discussion

As shown in Figure 14, Figure 14(a) illustrates that in GAN training, the generator (G) training loss converges stably, and the hidden information of the data can be learned, which is helpful to generate samples similar to the original audio, and the model training fits better. Figure 14(b) is a t-SNE-based attention matrix visualisation, and it can be observed that at each time step of the generation sequence, the decoder mainly focuses on the output of the encoder at the same time step (yellow represents a high degree of attention). This shows that in the process of generating speech features, the decoder can accurately align with the corresponding time step of the input speech signal, which can show that the model establishes an effective correspondence relationship when processing the input and output sequences.

The speech recognition algorithm of this system is based on script call to access the speech recognition service background, and the coupling between the two is low, so it is convenient for the system to upgrade the speech recognition algorithm, access other speech recognition algorithms or access other platforms.

System performance testing generally refers to the response time of the system, which is manifested in two aspects in the intelligent speech recognition system studied in this article. One is the response speed of the speech recognition process, and the other is the response time of the system after receiving instructions, that is, the response time of the system from the silent state to the awake state. If these two times are shorter in the test, it indicates that the system performance is higher, and the speed of response to the user terminal is also faster, which can provide users with a very good experience. The system in this article collects speech signals in real time through hardware devices at the terminal and processes them through internal algorithms. Therefore, the system in this article does not need to be connected for a long time to perform data processing, Moreover, the stability of the system in this article is ensured. The system in this article can simultaneously perform data transmission and data processing during speech signal recognition. The system effectively improves data processing efficiency and reduces system redundancy through a layered processing mode. Stable audio signal processing can be performed with one request. Compared with traditional audio processing modes, the model in this paper does not require multiple uploads of audio signals, thus effectively improving the processing efficiency of speech signals

Since the system model of this paper uses speech visual recognition technology, in the actual recognition process, it can not only perform recognition through speech recognition methods, but also perform speech recognition in combination with visual recognition schemes. From the existing research, it can be seen that a single speech recognition method will inevitably have certain drawbacks. Therefore, this paper combines speech recognition technology and visual recognition to effectively integrate their advantages and overcome the shortcomings of a single recognition method.

By comparing the performance of the methods in different reference articles with that of this method in recognition time, accuracy, recall rate and F1 score, the following analysis can be carried out:

1 Recognition time:

   The recognition time of this method is the shortest, only 0.664, which is significantly lower than that of the methods in other references. The recognition time of reference 1 is 0.964, which is shorter than other references, but still longer than the method in this paper. The recognition time of reference 2 is 1.292, that of reference 3 is 1.397, and that of reference 4 is 1.260, all of which are longer than that of the method in this paper. In this paper, the lightweight emotion recognition method effectively reduces the system redundancy and promotes the efficiency of data processing. However, other methods need to process multiple voice data conversion, and then output the recognition results. Therefore, the existing algorithms take more recognition time than the method in this paper.

2 Accuracy:

   The accuracy of this method is the highest, reaching 0.848. The accuracy of reference 1 is also high, 0.763. The accuracy of reference 2 is 0.692, which is lower than that of this paper and reference 1. The accuracy rates of reference 4 and reference 3 are similar, which are 0.709 and 0.710, respectively. Their accuracies are relatively good, but still lower than that of the method in this paper. By using the emotion2vec emotion feature extraction scheme, the SER algorithm based on Bi-LSTM is designed. According to the dialogue history and context, the algorithm can capture the complex patterns of speech, enhance the model's ability to understand context, and improve the recognition accuracy.

3 Recall:

The recall rate of this method is 0.435, which belongs to the medium level among all methods. The recall rate of reference 3 is the lowest, only 0.168. The recall rate of reference 2 is 0.182, slightly higher than that of reference 3. The recall rate of reference 4 is 0.238, slightly higher than that of reference 3 and reference 2. The recall rate of reference 1 was 0.177, which was in the lower middle level. The reason for the low recall rate of this model may be related to the insufficient number of samples used for training this model. Therefore, this model has the function of intelligent learning and progress. When there are enough samples, it can further improve the recognition performance of the model.

4 F1-score (F1):

The F1 score of this method is 0.372, which is medium in all indicators. Reference 2 has the highest F1 score of 0.189.

The F1 score of reference 1 is 0.107, which is relatively low. The F1 scores of reference 4 and reference 3 are 0.128 and 0.133, respectively, slightly lower than reference 2, but higher than reference 1.

The lightweight speech synthesis model kd-vits is obtained by using knowledge distillation method; in the target detection module, MobileNetV3 network is used to realise the lightweight of yolov5s model. All these make the model have higher recognition efficiency and accuracy than the existing models.

To sum up, although this method is not optimal in terms of recall rate and F1 score, it has the shortest recognition time and the highest accuracy, showing good overall performance. The methods in other references are different in different indicators, but the overall performance is not as balanced as the method in this paper.

From the recognition results of multiple speech datasets in Table 3, the method in this paper has the highest recognition accuracy in each speech dataset, which also verifies the above results.

This verifies that the model proposed in this paper has the fastest response speed in speech recognition and the fastest feedback from users. However, several other methods have a certain delay, which affects the user experience. From the perspective of recall rate and F1 value, the model proposed in this paper also has the best effect in speech recognition. Therefore, on the whole, the model in this paper has certain advantages in speech recognition compared with existing studies. At the same time, it also verifies that the application effect of speech visualisation technology in speech recognition results is very obvious, providing a model reference for the subsequent construction of English speech intelligent teaching system.

The main limitations of Vits speech synthesis model are model complexity, training difficulty and dependence on specific datasets and hardware resources.

First, although Vits model can generate high-quality voice, its model structure is relatively complex, which may limit the deployment and reasoning efficiency of the model in some application scenarios. The complex model structure means a higher demand for computing resources, which may not be conducive to running on resource constrained devices.

Secondly, the training process of Vits model is relatively cumbersome. Developers need to prepare data in specific formats for the model, including unmarked audio and text, and optimise the model through complex pre-processing and training processes. In addition, the code of the multi person training model may need to be modified by developers, and the training time is long, which increases the difficulty and cost of model application.

Finally, Vits model has a certain dependence on specific datasets and hardware resources. For example, in the training process, a large amount of voice data is needed to ensure the generalisation ability of the model. At the same time, high-performance GPU resources are also the key factors to accelerate model training and improve model performance. This dependence may limit the application of the model in some scenarios, especially when data is scarce or hardware resources are limited.

5 Conclusions

Aiming at the current problem that it is difficult to improve the correction effect of real-time speech recognition in English pronunciation teaching, this paper applies speech visualisation technology to pronunciation teaching. This paper conducts in-depth research on multi-dimensional visualisation and proposes a new multi-dimensional visualisation method. Moreover, the algorithm design and implementation of each functional module of intelligent interactive device in the process of pronunciation teaching are analysed. In addition, this paper uses emotion2vec emotion feature extraction scheme to design a SER algorithm based on Bi LSTM. In the task of speech synthesis, the VITS speech synthesis model is implemented, and the lightweight speech synthesis model KD-VITS is obtained by using knowledge distillation method. In the object detection module, the MobileNetV3 network is used to realise the lightweight of the YOLOv5s model. According to comprehensive experimental research, this paper has high speech recognition accuracy and recognition speed. Finally, from the results of comparative tests, it can be seen that the model proposed in this paper has certain advantages compared with the existing studies in speech recognition, and can meet the needs of intelligent English pronunciation teaching.

Because the interaction methods based on speech and video processing are universal, this paper focuses on speech interaction and video virtual interaction. However, in some usage scenarios, there may be more suitable interaction methods. Therefore, in the actual use of intelligent interactive devices, other interactive modules can be combined to further improve the user's interactive experience, which is also the follow-up research direction of this paper.

Declarations

All authors declare that they have no conflicts of interest.

References

Alabsi, T. (2020) 'Effects of adding subtitles to video via apps on develo** EFL students' listening comprehension', *Theory and Practice in Language Studies*, Vol. 10, No. 10, pp.1191–1199.

Al-Jarf, R. (2021) 'TED Talks as a listening resource in the EFL college classroom', *International Journal of Language and Literary Studies (IJLLS)*, Vol. 2, No. 3, pp.256–267.

Dizon, G. and Thanyawatpokin, B. (2021) 'Language learning with Netflix: exploring the effects of dual subtitles on vocabulary learning and listening comprehension', *Computer-Assisted Language Learning Electronic Journal*, Vol. 22, No. 3, pp.52–65.

Du, G., Hasim, Z. and Chew, F.P. (2022) 'Contribution of English aural vocabulary size levels to L2 listening comprehension', *International Review of Applied Linguistics in Language Teaching*, Vol. 60, No. 4, pp.937–956.

Ekayati, R. (2020) 'Shadowing technique on students' listening word recognition', *IJEMS: Indonesian Journal of Education and Mathematical Science*, Vol. 1, No. 2, pp.31–42.

Evers, K. and Chen, S. (2022) 'Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults', *Computer Assisted Language Learning*, Vol. 35, No. 8, pp.1869–1889.

Hsu, H.L., Chen, H.H.J. and Todd, A.G. (2023) 'Investigating the impact of the Amazon Alexa on the development of L2 listening and speaking skills', *Interactive Learning Environments*, Vol. 31, No. 9, pp.5732–5745.

Jia, C. and Hew, K.F.T. (2022) 'Supporting lower-level processes in EFL listening: the effect on learners' listening proficiency of a dictation program supported by a mobile instant messaging app', *Computer Assisted Language Learning*, Vol. 35, Nos. 1–2, pp.141–168.

Kumar, L.A., Renuka, D.K., Rose, S.L. and Wartana, I.M. (2022) 'Deep learning based assistive technology on audio visual speech recognition for hearing impaired', *International Journal of Cognitive Computing in Engineering*, Vol. 3, No. 2, pp.24–30.

Milliner, B. and Dimoski, B. (2024) 'The effects of a metacognitive intervention on lower-proficiency EFL learners' listening comprehension and listening self-efficacy', *Language Teaching Research*, Vol. 28, No. 2, pp.679–713.

Mukhamadiyev, A., Khujayarov, I., Djuraev, O. and Cho, J. (2022) 'Automatic speech recognition method based on deep learning approaches for Uzbek language', *Sensors*, Vol. 22, No. 10, pp.3683–3692.

Nguyen, T.D.T. and Pham, V.P.H. (2022) 'Effects of using technology to support students in develo** speaking skills', *International Journal of Language Instruction*, Vol. 1, No. 1, pp.1–8.

Ran, D., Yingli, W. and Haoxin, Q. (2021) 'Artificial intelligence speech recognition model for correcting spoken English teaching', *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 2, pp.3513–3524.

Shadiev, R., Wu, T.T. and Huang, Y.M. (2020) 'Using image-to-text recognition technology to facilitate vocabulary acquisition in authentic contexts', *ReCALL*, Vol. 32, No. 2, pp.195–212.

Tai, T.Y. and Chen, H.H.J. (2021) 'The impact of immersive virtual reality on EFL learners' listening comprehension', *Journal of Educational Computing Research*, Vol. 59, No. 7, pp.1272–1293.

Tsai, S.C. (2023) 'Learning with mobile augmented reality-and automatic speech recognition-based materials for English listening and speaking skills: effectiveness and perceptions of non-English major English as a foreign language students', *Journal of Educational Computing Research*, Vol. 61, No. 2, pp.444–465.

Van, L.K., Dang, T.A., Pham, D.B.T., Vo, T.T.N. and Pham, V.P.H. (2021) 'The effectiveness of using technology in learning English', *AsiaCALL Online Journal*, Vol. 12, No. 2, pp.24–40.

Zhang, S. and Zhang, X. (2022) 'The relationship between vocabulary knowledge and L2 reading/listening comprehension: a meta-analysis', *Language Teaching Research*, Vol. 26, No. 4, pp.696–725.