3D Face Animation Generation Method Based on Self-Supervised Speech Coding and Lattice Convolutional Architecture

Yunying Wang
Academy of Art Design, Henan Institute of Technology, Henan, 453003, China
Email: wang11358@outlook.com

Keywords: self-supervised learning, coding, convolution, 3D face animation, expression, speech features, pose

Received: July 9, 2025

Existing 3D face animation generation methods focus on lip movement and audio synchronization, ignoring the ability to synchronize expressions and poses. To address this problem, the study proposes a Self-Supervised Speech-Driven 3D Face Animation via Lattice Convolution Networks. The study first selected students from a certain school to read aloud the same corpus and record audio and video as the dataset. Through self-supervised learning and encoder-decoder structure, the speech features were extracted and mapped, and the obtained facial parameters were applied to the Face Latent Animated Mesh Estimator model to achieve lip-sync. Then, by combining the optical flow information in the video stream with the changes of facial key points, the grid convolutional network is used to model the expression dynamics and head postures, achieving multimodal feature fusion. In the experiment of analyzing the naturalness and accuracy of the generated animation, the lip shape vertex error, naturalness score, lip reading character error rate, and pixel error rate of the proposed method were only 2.54mm ^{2}, 9.45, 2.34%, and 2.53% respectively. In the performance analysis experiment of the emotion and posture recognition model, the accuracy rate of expression recognition and the posture offset error were 92.34% and 1.2° respectively. The lighting sensitivity, micro-expression fidelity and rendering frame rate of the generated face animation were 5.01, 93.48% and 63.74FPS respectively. The proposed 3D face animation generation method can effectively improve the realism and synchronization of the animation and achieve more accurate face animation generation.

Povzetek: Raziskava predstavi metodo, ki iz govora ne animira le premikanja ustnic, ampak tudi obrazno mimiko in položaj glave, zato so 3D animacije bolj naravne in bolje usklajene z govorom.

1 Introduction

As animation technology continues to advance, 3D animation has progressively emerged as a new medium that can depict settings and characters that are more colorful and lifelike. 3D face animation (3DFA) is a significant subset of it that is becoming more and more significant in the creation of video games, virtual reality, and movies and television shows [1]. The face expression and lip synchronization of 3D animation are the key factors affecting the fineness of animation. To improve the quality of animation production, more and more scholars have begun to study the face generation of 3D animation [2]. Sun Z et al. proposed a diffusion model-based generation framework in order to realize face generation for 3D animation. The framework extracted style encoders embedded with styles from short reference videos and built classifiers based on speech and style to further guide face generation. The results indicated that the 3D animation generated by this method had high accuracy [3]. To conduct an in-depth study on 3D animation task generation, Sha T et al. summarized the survey on the scope of research, recent advances, and technological trends in animation generation. The findings indicated that virtual fitting, digital human body etc. were the application areas of the study [4]. Wang B et al. proposed a 3D animation generation method for more convenient and faster human-computer interaction. The method employed support vector machine to extract the facial features of the face and used C++ and OpenGL for rendering simulation. The results indicated that the method was able to achieve real-time detection of face regions in video images [5]. Hou Z D et al. proposed a new process in order to achieve personalized virtual face generation. The process performed topology on a real human face model using R3ds Wrap, after which the model was deformed using a mesh deformation algorithm to generate a virtual face. Experimental results indicated that the virtual face generated by this method had a high degree of realism and personalization [6].

Self-supervised learning is an emerging deep learning paradigm. It achieves efficient learning by mining the inherent structural information of the data itself, and has a wide range of applications in various fields [7]. Yang Z et al. proposed a lightweight self-supervised model in order to achieve high-precision surgical navigation in endoscopic scenarios. The model combined lightweight convolutional neural network (CNN) and Transformer to

extract texture features and shape features respectively. The method realized attitude prediction through global information sensing. According to the findings, the method's prediction accuracy was 92.36% [8]. Wang T et al. proposed a method based on self-supervised learning in order to simulate human muscle movement and provide the animation industry with simulable joint movement sequences. This method introduces the loss function based on physical principles to achieve skeletal nerve simulation and combines the deformation effect to simulate muscle movement. The results show that the simulation authenticity of this method exceeds that of the traditional methods [9]. There are several uses for CNNs, which are shift-invariant or spatially invariant artificial neural networks, in domains like signal and image processing [10]. Ghazal T M et al. proposed a CNN based recognition system in order to achieve effective recognition of handwritten documents. The system achieved high accuracy recognition of handwritten documents through multi-level feature extraction and adaptive learning mechanism. Experimental results indicated that the system showed excellent recognition performance on a variety of handwritten samples, with a recognition accuracy rate of more than 90% [11]. Jain N et al. proposed an emotion recognition method based on convolutional neural networks in order to simulate the emotions of real human faces into animations. This method uses Mask R-CNN for character detection and combines a deep learning model for sentiment classification. The results show that the recognition accuracy of this method has reached over 90% [12].

In summary, existing 3DFA generation methods suffer from the problems of insufficiently fine speech feature extraction and poor expression synchronization. To address this problem, the study proposes a 3DFA generation model based on self-supervised speech coding and Lattice Convolution Networks. The model uses voice video to achieve the expression pose synchronization and lip-sound synchronization of 3D face. The study aims to achieve accurate voice-expression synchronized animation through the established 3DFA generation method, which provides technical support for real-time interactive applications. The innovation of the study is to apply the waveform vector self-supervised learning framework Wav2vec 2.0 to the speech feature extraction task. Furthermore, a new self-supervised learning model is designed to achieve more accurate audio-visual synchronization of 3DFA by combining the encoder-decoder structure. In addition, the study adopts a convolutional structure to extract facial expression features, geometric feature parameters, and combines them with the self-supervised speech coding model. The fineness of 3D animation feature capture is further improved.

The novelty of the proposed method compared with the existing models lies in the deep integration of self-supervised speech representation learning and 3D facial dynamic modeling. It uses Wav2vec 2.0 to extract high-level semantic features from the original audio, avoiding the reliance of traditional methods on manually labeled phonemes. Meanwhile, the grid convolutional network is introduced to jointly model the local and global features of the facial topological structure, significantly enhancing the fineness and synchronization accuracy of expression changes, and achieving an end-to-end audio and video collaborative generation framework that can be trained without pairing data.

Its key contributions are as follows:

(1) A self-supervised pre-training strategy without lip-reading annotation is proposed for voice-driven 3D facial animation generation, effectively reducing the reliance on large-scale labeled data and enhancing the model's generalization ability in real scenes.

(2) A facial dynamic decoder based on grid convolution was designed, which can accurately capture micro-expression changes and maintain long-term temporal consistency.

(3) An end-to-end audio and video collaborative training framework was constructed, achieving high-precision synchronization of voice and facial movements without paired data.

2 Methods and materials

To achieve accurate synchronization of face animation, a 3DFA generation method based on self-supervised speech coding and Lattice Convolution Networks is proposed. The method combines speech features and video features to finely capture face expression and lip shape changes to ensure synchronization with real-life audio and video.

2.1 Synchronized lip sync generation for face animation based on self-supervised speech coding

One hundred and twenty students from School A, consisting of 60 males and 60 females, are selected as subjects for the study. These subjects read the same spoken English sentences while recording audiovisual videos, totaling 425 utterances. The length of each utterance sequence is about 3.0s~4.0s. These utterances come from Corpus of Contemporary American English (COCA), which covers a wide range of life scenarios and topics in the corpus. The study uses data collected from 80 subjects as a training set and data from another 40 subjects as a test set. The study's volunteers are evenly distributed in terms of age, gender, and accent to guarantee the samples' diversity and representativeness. The video frames in the dataset are sampled at 30 frames per second, and the audio signals are resampled to 16kHz. The audio preprocessing adopts short-time Fourier transform to extract MEL spectrum features and normalizes them to eliminate individual differences. The MEL spectrogram is used as the speech input mode and sent into the Wav2vec 2.0 model for feature encoding to extract high-dimensional speech latent variables. In the video modality, the FLAME network is adopted to locate the key points of the face, and the corresponding expression coefficients and lip movement parameters are generated through the three-dimensional regresser. The grid convolution module models the topological structure of the human face and captures the fine-grained changes in

local muscle movements. The self-supervised learning framework optimizes the speech-to-expression mapping relationship without paired annotations. To extract the speech features in the audiovisual video, the study uses the waveform vector self-supervised learning framework Wav2vec 2.0 to process the speech signals. Wav2vec 2.0 extracts high-quality features from the original speech waveforms by self-supervised learning, which can effectively capture the subtle changes of speech [13-14]. Fig. 1 depicts the model's construction.

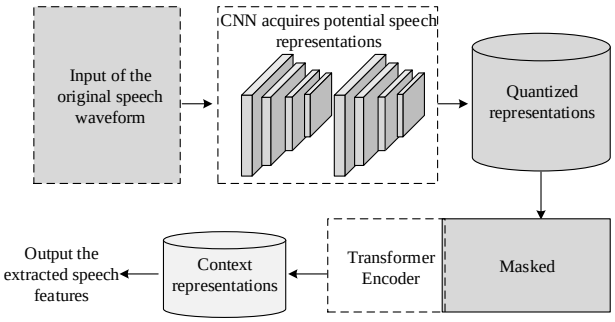


Figure 1: Wav2vec 2.0 model structure

As shown in Figure 1, the Wav2vec 2.0 model consists of multiple convolutional layers, quantization modules, and Transformer encoders. The convolutional layer of the Wav2vec 2.0 model is used to process the original speech waveform and extract local features. The quantization module discretizes continuous features into latent representations. The Transformer encoder captures long-distance dependencies. The Transformer has 12 layers, 768 hidden dimensions, and 12 attention heads. The learning rate is 5e-4, the batch size is set to 16, and the number of training rounds is 100. The cosine annealing strategy is adopted to dynamically adjust the learning rate to enhance the model's convergence. During the training process, the input speech waveform is divided into 20ms frames with a sampling frequency of 16kHz. By comparing the prediction tasks, the model is pre-trained on unlabeled data to enhance the ability to express speech features. The feature output dimension is 768, achieving seamless integration with the subsequent 3D face mesh convolutional network. The encoder extracts the context embedding vector, the convolutional layer processes the local features, and the quantization module converts the continuous speech into a discrete representation. 3D face modeling, as a popular and effective face representation, has the ability to express expressive motion and understand audio-visual synchronization better than 2D images [15-16]. To realize realistic 3D animation effects, deep learning techniques are investigated to generate offset vertex sequences of 3D face templates by taking raw speech features as input. Finally, the offset vertex sequences are mapped onto the 3D face model, and the 3DFA is generated by fine-tuning. To this end, the research designs a self-supervised speech coding network. It learns the geometric structure of the 3D face model through the encoder and decoder, and outputs the animation. The structure of the self-supervised speech coding network designed in the study is shown in Fig. 2.

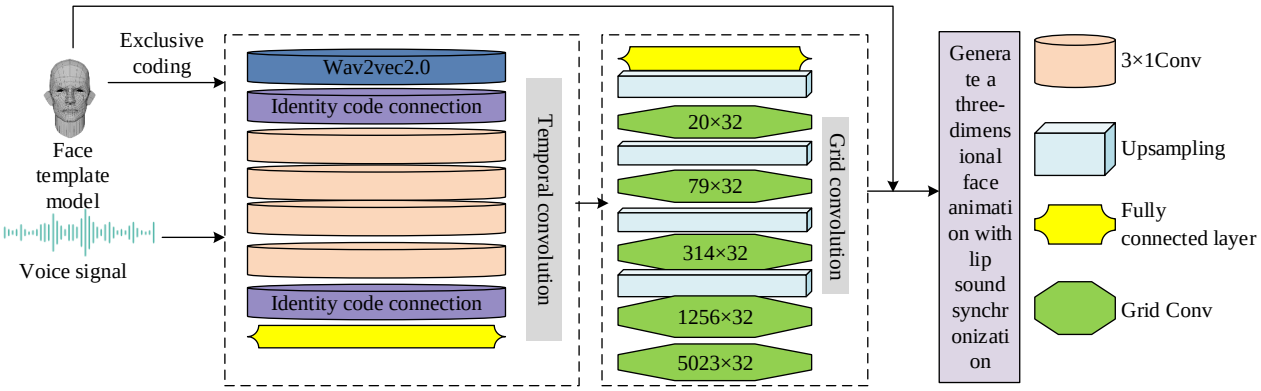


Figure 2: The self-supervised speech coding network structure designed by the research

In Fig. 2, the encoder includes a Wav2vec 2.0 module and a 3D face feature extraction module. The module includes two identity coded connectivity layers, four CLs, and one fully connected layer (FCL). To ensure that the model can recognize and understand individual differences, the study uses one-hot unique hot coding vectors to encode various training objects in the dataset in an attempt to learn the speaking styles of various individuals. The study then fuses the coding vectors with the extracted speech features and the CL outputs in a FCL to control the network to output 3D animated faces with different speaking styles. The step size of all convolution layers is 2×1 and the convolution kernel size is set to 3×1. The study uses spectral convolution in the grid convolution layer to capture finer geometric details and enhance the realism of the animation. The spectral convolution is shown in Equation (1).

$$y = \sum_{i=1}^{C} g_{w_{ij}} \phi x_i \tag{1}$$

In Equation (1), y is the convolution output. $g_{w_{ij}}$ is the lattice kernel filter. w is the Chebyshev coefficient vector. ϕ is the Laplace real symmetric matrix. x_i is the mesh vertices of the 3D model. C is the number of

features. To ensure that the generated 3DFA sequence is consistent with the real situation, the study introduces a joint loss function (LF) consisting of reconstruction loss and velocity loss. The reconstruction loss ensures the accuracy of the geometric structure, while the velocity loss smoothes the animation transition and avoids abrupt changes. The LF is specifically shown in Equation (2).

$$Loss = L_1 + \xi \cdot L_2 \quad (2)$$

In Equation (2), ξ is the hyperparameter. The study sets it to 0.1 to balance reconstruction and velocity loss. L_1 and L_2 are the reconstruction loss and velocity loss values, respectively. $Loss$ is the loss value of the model training process. The study uses mean square error to calculate the reconstruction loss. The study uses face latent animated mesh estimator (FLAME) to generate the base 3D face model. Moreover, it is combined with the established self-supervised speech coding network to realize the lip sync of 3D animation. Before that, the study performs 3D mesh refinement of the face model. Fast spectral convolution is used to learn the nonlinear representation of the face, and the hierarchical mesh representation is realized by mesh sampling operations. The filter in fast spectral convolution is expressed in Equation (3).

$$g(l) = \sum_{i=0}^{Q-1} w \cdot T_i(\tilde{l}) \quad (3)$$

In Equation (3), $g(l)$ is the filter kernel function. l and $\tilde{l}$ are the Laplace operators before and after scaling. T_i is the recursive computation result. Q is the order of the polynomial. This defines the spectral convolution as shown in Equation (4).

$$y_k = \sum_{i=1}^{M} g(l) \cdot x_i \quad (4)$$

In Equation (4), x_i is the input feature. y_k is the spectral convolution output result. k is the output feature number. M is the number of input features. The specific flow of speech-driven 3DFA generation is shown in Fig. 3.

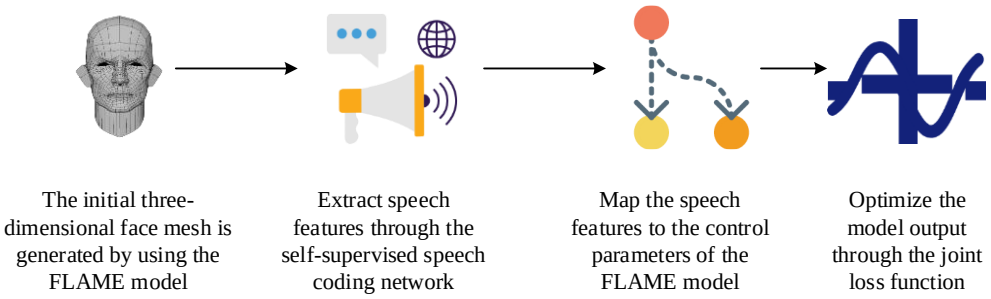


Figure 3: The specific process of generating 3D face animations driven by voice

In Fig. 3, the study first generates an initial 3D face mesh using the FLAME model. Then, speech features are extracted by self-supervised speech coding network and mapped to the control parameters of the FLAME model. Then, the mapped parameters are refined using grid convolution layer to ensure that the lip shape is highly matched with the speech. Finally, the model output is optimized by the joint LF to generate realistic 3D animated face sequences to achieve accurate synchronization between speech and lip shape.

2.2 3D face animation generation based on Lattice Convolution Networks and voice video

The study implements synchronized lip sync generation for face animation based on self-supervised speech coding. However, in 3DFA, only implementing lip sync cannot achieve a more ideal animation generation effect. Therefore, the study considers combining expression and pose parameters to further improve the face model, and first preprocesses the captured video frames. The processing flow is shown in Fig. 4.

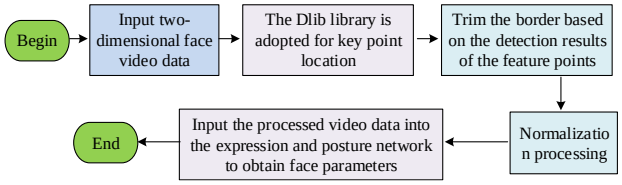


Figure 4: Video frame data processing flow

In Fig. 4, the study begins with the detection of feature points in the face video. The detection process uses the Dlib library for key point localization. Previous video frame preprocessing methods mainly use the edges of the face feature point detection in the image for individual cropping. In this way, although the faces of all frame images are accurately cropped, it is more troublesome and easier to lose the background information [17-18]. Therefore, it is investigated to compute the maximum cropped face borders based on the edge positions of the face feature points of all frames, covering the faces of all frame sequences and cropping them uniformly. After processing the 2D video data, the study applies it to the expression as well as the pose of the network to obtain the expression parameters and pose parameters. The method works by combining the expression parameters, the pose parameters, and the results of the lip offset points of the

3D face obtained from the speech drive. It is also input into the FLAME model for comprehensive optimization to generate a more natural and dynamic 3DFA. The specific structure of the 3DFA generation method based on Lattice Convolution Networks and voice video designed in the study is shown in Fig. 5.

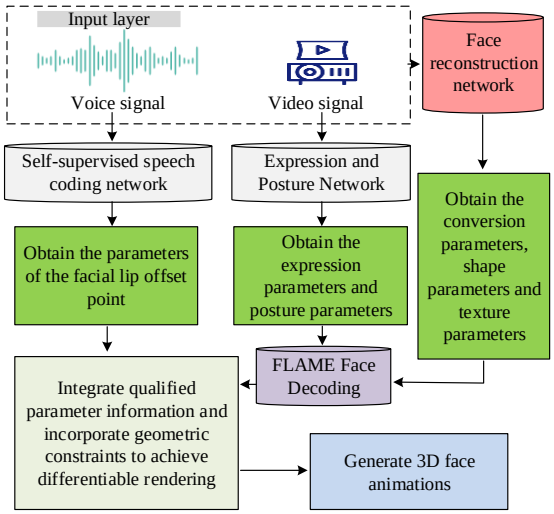


Figure 5: The specific structure of the 3D face animation generation method based on the Lattice Convolution Networks and voice video

In Fig. 5, the study inputs the video data into the expression pose network (EPN) to extract the expression and pose parameters. In addition, the study introduces a 3D face reconstruction network to predict the transformation parameters, shape parameters, and texture parameters to further optimize the 3D face model. The study uses a temporal CNN in the EPN to capture the temporal information and enhance the dynamic expression. In the process of expression recognition, the study uses MobileNet model to extract the visual features of each frame in the video, which is combined with the timing information for comprehensive analysis. The video data is first fed into the MobileNet model to extract visual features. After that, it is inputted into the temporal convolution layer to further extract the temporal features. The size of the temporal CL designed by the study is 5×2 with a step size of 1. When performing the fusion of the two modal features of pose and expression, the study first combines them by splicing, and then analyzes the visually relevant expression parameters and pose parameters by regressing them through two FCLs. Moreover, these parameters are input into the FLAME model for face decoding. Finally, it is combined with the speech-driven module to generate the final face animation. MobileNet model is a CNN architecture. It uses an inverted residual block and a linear bottleneck structure, which effectively reduces the amount of computation and improves the model efficiency. Its structure is shown in Fig. 6.

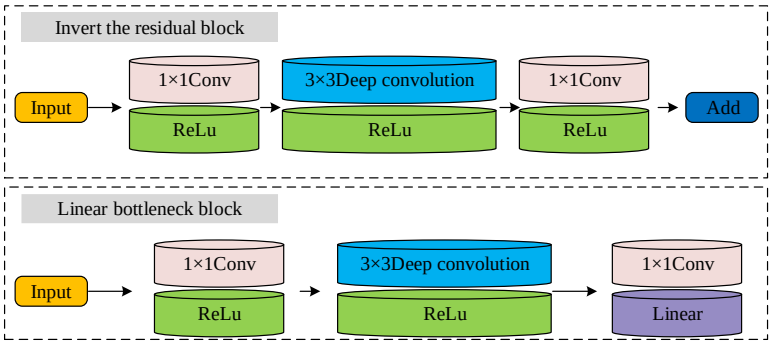


Figure 6: The specific structure of the MobileNet model

In Fig. 6, the inverted residual block contains a 1×1 CL for feature dimensionality reduction, followed by a deep CL. A separate convolutional filter is used to extract localized features, and then the dimensionality is restored by a 1×1 CL. The linear bottleneck block reduces the information loss and improves the accuracy of feature extraction by reducing the use of activation functions. To train the EPN, the study uses a LF to constrain it. The study combines emotion and lip-reading loss, pose consistency loss and geometric constraint loss to construct a multi-task learning framework. All three consistency losses are computed using mean square error. The geometric constraint loss is calculated using the Euclidean distance between the actual feature offsets and the projected feature offsets. The calculation process is shown in Equation (5).

$$\text{Loss}_v = \left\| \left(K_j - K_{j-1} \right) - \left(\mu\left(M_i \right) - \mu\left(M_{i-1} \right) \right) \right\| \quad (5)$$

In Equation (5), Loss_v is the geometric loss. K_j is the current frame face feature point coordinates. K_{j-1} is the previous frame face feature point coordinates. $\mu(\cdot)$ is the value to which the face model feature point is projected by the predictive camera model. M_i and M_{i-1} are both 3D face feature points in the FLAME model. The consistency loss is calculated as shown in Equation (6).

$$\begin{cases} \text{Loss}_e = \left\| \psi_v - \psi_R \right\|^2 \\ \text{Loss}_c = \left\| \gamma_v - \gamma_R \right\|^2 \end{cases} \quad (6)$$

In Equation (6), Loss_e and Loss_c are the loss values of

emotion and lip-reading consistency, respectively. ψ_v and γ_v are the original emotion features and lip features, respectively. ψ_R and γ_R are the emotion features and lip features obtained from rendering. In summary, the study fine-tunes the 3D face model by combining face expression, pose, and speech features. This ensures synergistic consistency among the features, resulting in more natural and realistic expression and pose reproduction.

The research adopts a self-supervised speech encoder to extract high-dimensional semantic features from the input speech and reduces the reliance on labeled data through a pre-trained model. The grid convolutional structure is constructed based on the topological connection relationship of the FLAME model. The graph convolutional network is utilized to perform local neighborhood aggregation on the vertices of the 3D face, enhancing the spatial perception ability. The model takes geometric loss and consistency loss as the joint optimization goals. Through end-to-end training, it gradually adjusts the mapping relationship between expression parameters, pose parameters and speech-driven features to improve the accuracy and fluency of dynamic expression generation. During the training process, the Adam optimizer is used for parameter updates. The initial learning rate is set to 0.001, and a learning rate decay strategy is adopted based on the loss changes on the validation set. The learning rate is multiplied by 0.9 every 10 epochs. The training period was set to a total of 50 epochs, with a batch size of 32. All experiments were conducted on NVIDIA A100 Gpus. To prevent overfitting, random discard and data augmentation strategies are introduced to perform time-domain masking and noise addition processing on the input speech features.

3 Results

To analyze the generation effect as well as the model performance of the 3DFA generation method proposed in the study, a series of experiments are carried out. The performance is discussed based on the experimental results.

3.1 Lip sync effect based on self-supervised speech coding

To examine the performance of the lip sync face generation method based on self-supervised speech coding designed in the study, the study compares it with existing lip sync techniques. The comparison methods include Real3D-Portrait and GeneFace. The comparison metrics include the mean error of lip vertex generation, the mean square error of all vertices of a 3D face, and the naturalness score of lip sync. The naturalness score is evaluated by five experts in the specialized field of 3D animation production and has a total score of 10. A higher score indicates a better generation of lip sync. Table 1 displays the comparing results.

Table 1: Performance comparison of three 3D face animation lip sync methods

Project	Face vertex error			Lip shape vertex error			Naturalness score (0-10)
	Maximum value	Minimum value	Average value	Maximum value	Minimum value	Average value	
Research method	2.83	3.54	3.26	2.25	2.69	2.54	9.45
Real3D-Portrait	2.85	3.47	3.24	2.67	3.11	3.02	8.23
GeneFace	2.76	3.44	3.36	2.78	3.15	2.97	8.54

In Table 1, there is basically no difference in the mean square error values of the overall vertices of the face for the three methods, and the error value of the proposed method is slightly higher in the study. This is due to the fact that the current method only combines speech features. This further justifies the subsequent combination of video expression and pose features. In addition, the lip shape vertex error value of the study's proposed method is significantly lower than the other methods, and the naturalness score is higher than the other two methods. The mean values of the two metrics are 2.54 mm^2 and 9.45 points, respectively. This indicates that the proposed method performs better in terms of lip synchronization accuracy and naturalness. The self-supervised speech coding network is able to capture the association between speech and lip shape effectively.

To further validate the effects of lip sync, the study tests a multilingual sample. The comparison languages include English, Chinese, Japanese, and Spanish. The changes of average lip vertex error value of the three methods with the increase of sample number under four languages are shown in Fig. 7.

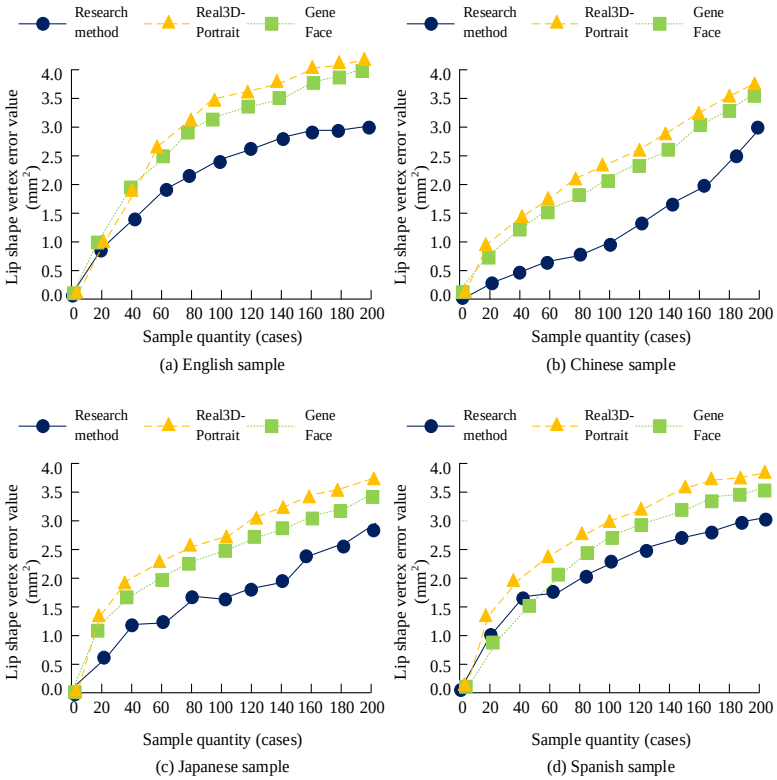


Figure 7: The variation of the average lip vertex error values of the three methods with the increase of the sample size

In Fig. 7(a), in the English samples, Real3D-Portrait error is the largest, GeneFace is the second largest, and the average error of the research-proposed method is only 2.54 mm^2. In Fig. 7(b), in the Chinese samples, the research-proposed method has the smallest lip-vertex error value. Its average value is only 2.31 mm^2, which is significantly better than the other two methods. In Fig. 7(c), for the Japanese sample, the research method still performs best with an error of 2.45 mm^2. As shown in Fig. 7(d) for the Spanish sample, the error is 2.38 mm^2, further validating the superiority of the method. This indicates that the research method exhibits high lip sync accuracy in different languages.

3.2 Effect of expression pose generation based on voice video and Lattice Convolution Networks

To realize more accurate and natural 3DFA generation, this study introduces expression and pose networks of video data on the basis of speech-driven. To test the effect of the proposed emotion, pose generation, the study introduces emotion recognition accuracy rate, head pose, offset error of actual pose, and face vertex error value as evaluation indexes. It is also compared and experimented with FaceAnime model and 3DMM-Blendshape proposed by Tsinghua University. The performance of the three models is compared through multiple datasets. The comparison datasets include BU-3DFE and Tufts face dataset. Table 2 displays the comparing results.

Table 2: Comparison of the expression and posture generation effects of three models under two datasets

Project	BU-3DFE			Tufts face dataset		
	Accuracy rate (%)	Attitude offset error (°)	Vertex error value (mm^2)	Accuracy rate (%)	Attitude offset error (°)	Vertex error value (mm^2)
Research method	94.10	0.92	2.49	90.58	1.38	2.67
FaceAnime	87.14	1.50	2.94	82.47	1.83	3.15
3DMM-Blend shape	90.33	1.32	2.79	85.44	1.69	3.00

In Table 2, in both datasets, the emotion recognition accuracy rate, head pose offset error, and face vertex error of the study's proposed method are significantly better than the other two methods. The mean values of these three metrics are 92.34%, 1.2°, and 2.58 mm^2, respectively. Especially on the BU-3DFE dataset, the emotion

recognition accuracy rate is as high as 94.10%, the head pose offset error is only 0.92°, and the face vertex error is only 2.49 mm^2. This indicates that the proposed method can effectively capture the expression and pose details of the subject.

To further test the effect of expression, pose generation, the study tests the dynamic video sequences, which are selected as the video sequences recorded by the students of School A. With the change of time, the recognition accuracy rate, head pose offset error, and face vertex error results of the three methods are shown in Fig. 8.

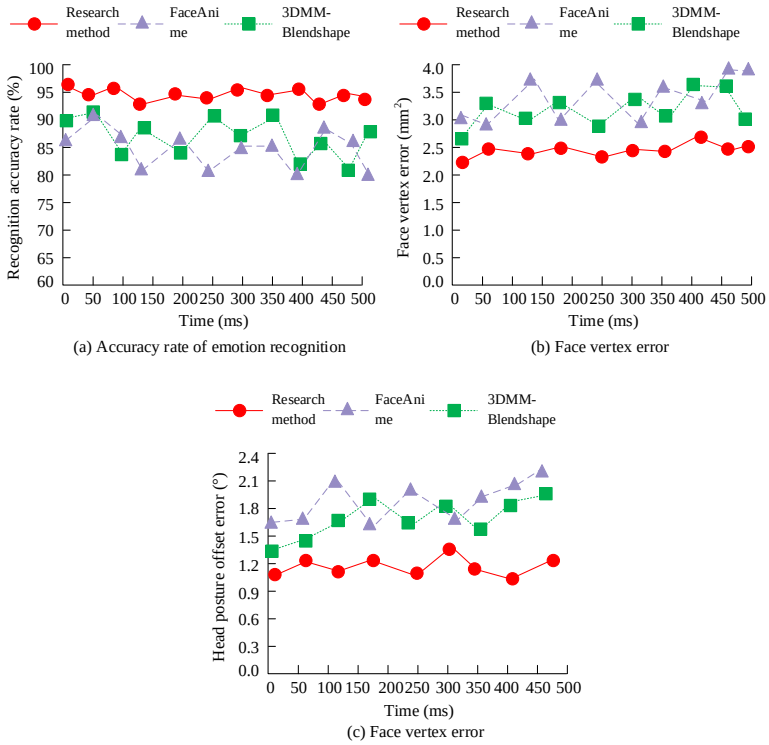
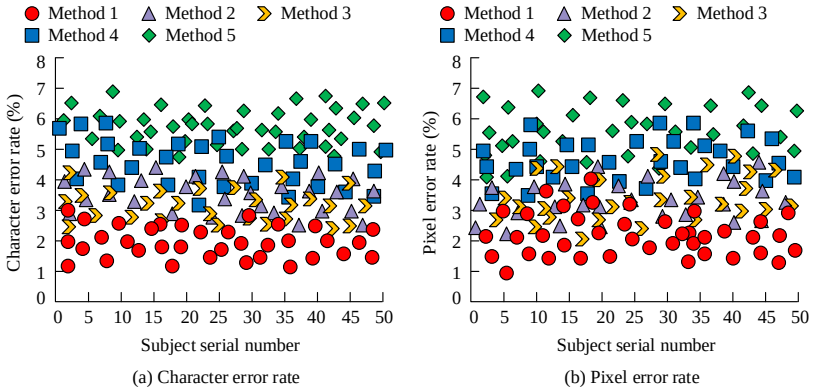


Figure 8: The recognition accuracy rates, head pose offset errors, and face vertex error results of the three methods

In Fig. 8(a), the recognition accuracy rate of the research design method remains stable with minimal fluctuations over time. Its average recognition accuracy rate value remains above 92%. The other two methods both produced different levels of fluctuations over the time series. In Fig. 8(b), the face vertex error has less variation over time, with an average value of 2.53 mm^2. This indicates that the research method has higher stability and reliability in dynamic expression recognition. In Fig. 8(c), the head pose offset error remains the smallest in the dynamic sequence with a mean value of 1.1°.

3.3 Application effect of face animation generation method based on speech coding and convolutional architecture

To test the effectiveness of the proposed face animation generation method (Method 1) based on self-supervised speech coding and Lattice Convolution Networks, the study conducts comparative experiments with the generation method (Method 2) in the literature [18], the generation method (Method 3) in the literature [19], the generation method (Method 4) in the literature [20], and the generation method (Method 5) in the literature [21]. The experiment is applied to the generation of 3DFA for 50 students in School A, and lip-reading loss is introduced as an evaluation index. Among them, it includes the character error rate and the optogenetic error rate representing the error rate of the corresponding visual mouth unit during pronunciation. The comparison results are shown in Fig. 9.



(a) Character error rate (b) Pixel error rate

Figure 9: The lip-reading loss results of 3D face animations for 50 students

In Fig. 9(a), the fluctuation range of the character error rate for Method 1 is from 2.1% to 3.5%, which is significantly smaller than that of the other four methods in the 3DFA generation for 50 students. In Fig. 9(b), the fluctuation range of the optic error rate for Method 1 is 1.8% to 2.9%. The fluctuation range for Method 2 is 3.2% to 4.7%. For Method 3, it is 4.1% to 5.6%. Methods 4 and 5 achieve an average error rate of 4% or more. In summary, Method 1 proposed in the study exhibits higher stability and accuracy in both character and reticle error rates, verifying its superiority in face animation generation.

To further validate the effectiveness of the face animation generated by the proposed method, the study introduces vertex position error (VPE), root-mean-square error (RMSE) of expression movement, illumination sensitivity (IS), micro-expression fidelity (MEF), and rendering frame rate (RFR) metrics to further analyze its performance. Among them, the MEF metrics are obtained by utilizing facial action coding system (FACS) assessment. Table 3 displays the comparing results.

Table 3: Comparison of the generation effects of several 3D face animation generation methods

Project	VPE (mm)	RMSE	IS	MEF (%)	RFR (FPS)
Method 1	1.22	0.15	5.01	93.48	63.74
Method 2	1.75	0.25	5.71	89.97	58.46
Method 3	1.36	0.19	5.48	90.22	60.00
Method 4	1.92	0.28	6.00	85.36	54.23
Method 5	2.03	0.30	6.03	85.12	51.04

In Table 3, Method 1 performs optimally in all performance metrics, with VPE and RMSE significantly lower than the other methods. It has a moderate IS with a high MEF of 93.48% and the highest RFR, which further confirms its comprehensive advantage in generating face animation. In addition, Method 1 shows greater adaptability in complex lighting environments and dynamic expression capture. The VPE value and the MSE value of expression movement of Method 1 are 1.22 mm and 0.15, respectively. Its RFP reaches 63.74 FPS and its IS value is only 5.01, which enables it to effectively cope with a variety of lighting changes. In conclusion, the proposed 3DFA generation method is leading in terms of accuracy, smoothness and environment adaptability. It is capable of generating realistic and detailed dynamic expressions, and accurately capturing the changes in the character's mouth shape when he/she is speaking.

4 Discussion and conclusion

To ensure that the face generation animation has natural facial expressions and movements, and can realize high-precision lip sync, the study proposed a 3DFA generation method based on self-supervised speech coding and Lattice Convolution Networks. The method realized lip sync by self-supervised speech coding network and FLAME model. At the same time, the research combined the video data, extracted the expression and pose parameters, and fused them with speech features to generate the final face animation. The experimental results indicated that the mean values of lip shape vertex error value and naturalness score of the research proposed method were 2.54 mm^2 and 9.45, respectively. The mean error value was less than 2.6 mm^2 in different language samples. The mean values of emotional recognition accuracy rate, head pose offset error, and face vertex error of the research proposed method in different datasets were 92.34%, 1.2°, and 2.58 mm^2, respectively. Moreover, the expression pose feature capture of the proposed method was significantly better and more stable in both dynamic testing situations. After the application of Method 1, the average values of lip-reading character error rate and optic pixel error rate of the generated animation were only 2.34% and 2.53%, respectively. Its VPE, RMSE, MEF, RFP, and IS metrics were 1.22 mm, 0.15, 93.48%, 63.74 FPS, and 5.01, respectively. Compared with other methods, it was superior in detail performance. The method designed in

the current study focuses on three modal data: expression, movement, and speech. In the future, details such as teeth and hairs of the face can be refined to further improve the realism of the animation.

At present, the research still has certain limitations, which are specifically manifested in the limited adaptability to low-resolution input videos and the decline in the estimation accuracy of expression parameters in extreme occlusion scenarios. In addition, model training relies on a large amount of labeled data, and there is still room for improvement in generalization ability across racial and age groups. Future work will introduce lightweight network structures to optimize reasoning efficiency and combine physical driving mechanisms to enhance dynamic simulation of facial details, in order to further improve the realism and robustness of generated animations. Meanwhile, explore unsupervised domain adaptive strategies to alleviate the reliance on data annotation and enhance the applicability of the model among people of different cultures and age groups.

References

[1] Melnik A, Miasayedzenkau M, Makaravets D, Pirshtuk D, Akbulut E, Holzmann D, Ritter H(2024). Face generation and editing with stylegan: A survey. IEEE Transactions on pattern analysis and machine intelligence, 46(5): 3557-3576. https://doi.org/10.1109/TPAMI.2024.3350004

[2] Bora N P, Jain D C(2023). A web authentication biometric 3D animated CAPTCHA system using artificial intelligence and machine learning approach. Journal of Artificial Intelligence and Technology, 3(3): 126-133. https://doi.org/10.37965/jait.2023.0216

[3] Sun Z, Lv T, Ye S, Lin M, Sheng J, Wen Y H, Liu Y J(2024). Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. ACM Transactions on Graphics (TOG), 43(4): 1-9. https://doi.org/10.1145/3658221

[4] Sha T, Zhang W, Shen T, Li Z, Mei T(2023). Deep person generation: A survey from the perspective of face, pose, and cloth synthesis. ACM Computing Surveys, 55(12): 1-37. https://doi.org/10.1145/3575656

[5] Wang B, Shi Y(2023). Expression dynamic capture and 3D animation generation method based on deep learning. Neural Computing and Applications, 35(12): 8797-8808. https://doi.org/10.1007/s00521-022-07644-0

[6] Hou Z D, Kim K H, Lee D J, Zhang G H(2022). Real-time markerless facial motion capture of personalized 3D real human research. International Journal of Internet, Broadcasting and Communication, 14(1): 129-135.

[7] Zhang Y, Nie R, Cao J, Ma C(2023). Self-supervised fusion for multi-modal medical images via contrastive auto-encoding and convolutional information exchange. IEEE Computational Intelligence Magazine, 18(1): 68-80. https://doi.org/10.1109/MCI.2022.3223487

[8] Yang Z, Pan J, Dai J, Sun Z, Xiao Y(2024). Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer. IEEE Transactions on Medical Imaging, 43(5): 1934-1944. https://doi.org/10.1109/TMI.2024.3352390

[9] Wang T, Liu S(2025). Hierarchical Neural Skinning Deformation with Self-supervised Training for Character Animation. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 8(1): 1-20. https://doi.org/10.1145/3728300

[10] Zhou Z, Zhang J, Gong C(2023). Hybrid semantic segmentation for tunnel lining cracks based on Swin Transformer and convolutional neural network. Computer-Aided Civil and Infrastructure Engineering, 38(17): 2491-2510. https://doi.org/10.1111/mice.13003

[11] Ghazal T M(2022). Convolutional neural network based intelligent handwritten document recognition. Computers, Materials & Continua, 70(3): 4563-4581. https://doi.org/10.32604/cmc.2022.021102

[12] Jain N, Gupta V, Shubham S, Madan A, Chaudhary A, Santosh K C(2022). Understanding cartoon emotion using integrated deep neural network on large dataset. Neural Computing and Applications, 34(24): 21481-21501. https://doi.org/10.1007/s00521-021-06003-9

[13] Wu Y, Daoudi M, Amad A(2023). Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. IEEE Transactions on Affective Computing, 15(1): 157-172. https://doi.org/10.1109/TAFFC.2023.3263907

[14] Towfek S K, Khodadadi N(2023). Deep convolutional neural network and metaheuristic optimization for disease detection in plant leaves. Journal of Intelligent Systems and Internet of Things, 10(1): 66-75. https://doi.org/10.54216/JISIoT.100105

[15] Jung G, Jung S G, Cole J M(2023). Automatic materials characterization from infrared spectra using convolutional neural networks. Chemical Science, 14(13): 3600-3609. https://doi.org/10.1039/D2SC05892H

[16] Singh S A, Desai K A(2023). Automated surface defect detection framework using machine vision and convolutional neural networks. Journal of Intelligent Manufacturing, 34(4): 1995-2011. https://doi.org/10.1007/s10845-021-01878-w

[17] Nsugbe E(2023). Toward a Self-Supervised Architecture for Semen Quality Prediction Using Environmental and Lifestyle Factors[C]//Artificial Intelligence and Applications. 1(1): 35-42. https://doi.org/10.47852/bonviewAIA2202303

[18] Niu L, Xie W, Wang D, Cao Z, Liu X(2024). Audio2AB: Audio-driven collaborative generation of virtual character animation. Virtual Reality &

Intelligent Hardware, 6(1): 56-70. https://doi.org/10.1016/j.vrih.2023.08.006

[19] Xu P, Zhu Y, Cai S(2022). Innovative research on the visual performance of image two-dimensional animation film based on deep neural network. Neural Computing and Applications, 34(4):2719-2728. https://doi.org/10.1007/s00521-021-06140-1

[20] Yuan Z, Lee J H, Zhang S(2023). Correction to: Research on simulation of 3D human animation vision technology based on an enhanced machine learning algorithm. Neural Computing & Applications, 35(7): 5589-5589. https://doi.org/10.1007/s00521-022-07267-5

[21] Shi Y, Wang B(2023). Optimization algorithm of an artificial neural network-based controller and simulation method for animated virtual idol characters. Neural Computing and Applications, 35(12): 8873-8882. https://doi.org/10.1007/s00521-022-07697-1