# Animation Character Mouth Matching Model Considering Reinforcement Learning and Feature Extraction

Hang Zhao
College of Culture and Communication, Liming Vocational University, Quanzhou, 362000, China
E-mail: tarotworld@126.com

**Keywords:** animation production, mouth matching, reinforcement learning, feature extraction, actor-critic method

**Received:** May 14, 2024

*With the development of the times, animation production has become increasingly sophisticated. Mouth matching is one of the key points to ensure the vividness and realism of animated characters. Therefore, this study proposed an animation character mouth matching model, which is based on Actor-Critic method of reinforcement learning. This model was combined with audio feature extraction and facial action coding to predict and generate animation character mouth matching synchronized with speech. This model was optimized by strategy gradient algorithm, aiming to achieve realistic and emotion-rich animated mouth matching. RAVDESS, VIDTIMIT, and SAVEE were selected as experimental datasets. Accuracy, F1 value, and peak signal-to-noise ratio were selected as performance indicators. GRU, CNN+GRU, and CNN+LSTM were selected as experimental comparison algorithms. The experimental results showed that the proposed model had an average accuracy of 95.61% and an F1 value of 97.13% on three databases. Meanwhile, the peak signal-to-noise ratio and structural similarity index of the proposed model were 41.77 and 0.93, respectively, which were better than the comparative methods. In addition, the study tested the error of mouth shape under different emotions, and the results showed an average mean square error of only 6.639. Finally, the user survey results showed that the animated characters generated by the proposed model received more recognition in mouth shape matching and realism, with a highest selection rate of 98.64%. The successful application of the proposed model provides new ideas and methods for research in related fields, laying the foundation for further promotion and innovation of animation production technology*

*Povzetek: Raziskava predstavlja model za usklajevanje ust animiranih likov, ki temelji na metodi Actor-Critic v okviru stojnega učenja. Model uporablja ekstrakcijo značilnosti zvoka in obraznih potez za sinhronizirano in realistično gibanje ust, kar omogoča izražanje čustev.*

## 1 Introduction

In today’s digital age, animation production plays an increasingly important role in the entertainment industry. Animation production not only provides viewers with rich and colorful visual enjoyment, but also becomes an important medium for people to express creativity and emotions [1, 2]. With the continuous advancement of technology, animation production is no longer limited to simple visual presentation, but requires deeper shaping of animated characters, especially in terms of expression and communication [3, 4]. In addition, the charm of animation is far beyond just the visuals. Dubbing and music, as the emotional guide, deepen the emotional resonance of animation works [5, 6]. In this context, matching the mouth shape of animated characters has become a challenging task, which is directly related to the expression and emotional communication of animated characters [7, 8]. However, it is extremely complex to generate synchronized mouth movements, involving multiple challenges such as video generation, precise matching from audio to video generation, and training the model's generalization ability. Traditional mouth shape matching methods should meet the needs of dynamic scenes and complex emotional expression [9, 10]. Generating synchronized mouth movements with emotional features is always a challenge in computer animation research. Therefore, this study proposes an innovative animation character mouth matching model that takes into account reinforcement learning and feature extraction. This study focuses on the mouth design model of 3D character animation and speech emotion synchronization algorithm. The Actor-Critic method in reinforcement learning is introduced to continuously optimize and improve the performance of the model in learning mouth generation. Meanwhile, the feature extraction of audio and facial Action Units (AUs) ensures that the generated mouth shape is not only realistic, but also able to express rich emotional features. The goal of the research is not only to achieve the production of animated characters with real emotions and mouth shape matching, but also to promote the development of the entire animation production technology. This innovative research not only provides audiences with a richer audio-visual experience, but also promotes the digital entertainment industry. This leads animation production into a more fascinating era.

The study is divided into five parts. The first part introduces the current research on different methods of animation production and mouth matching worldwide. The different applications of the Actor-Critic method are introduced too. The second part mainly introduces animation production, animated characters, and other content. The third part provides an in-depth introduction to the methods of mouth shape matching models for animated characters. The fourth part conducts experiment on the performance of the mouth shape matching model proposed in the study to verify its feasibility. The last part is a summary and discussion of the article.

## 2 Related works

The advancement of technology has provided strong support for the development of the animation industry. With the continuous advancement of computer and network technology, the cost and cycle of animation production have gradually decreased, while also improving the efficiency and quality of animation production [11]. There are also many studies on animation production around the world. Ye et al. proposed a new method based on support vector machine and Augmented Reality (AR) animator system to address the challenge of creating virtual AR animated characters that closely interacted with the real environment. This allowed users to easily create in situ character animations that closely interacted with different real-world environments [12]. Arshad et al. proposed a method to explore the basic process of character assembly systems in 3D animation production, addressing the flexibility limitations in the character assembly process. This study suggested that chain assembly systems were the best choice for the animated characters. Therefore, the assembly process in the animation directly affected the actions and poses of the character in the final animation [13]. Paier et al. proposed a new hybrid animation framework to tackle the challenging task of creating realistic facial animations in computer graphics. This framework utilized the latest advances in deep learning to provide an interactive animation engine for facial expression editing through simple and intuitive visualization [14]. Facial areas, such as eyes or teeth, could not be stably synthesized in the generation of animated characters. Then, K. Gu et al. proposed a landmark-driven dual-flow network to obtain data from multi-source images to achieve facial region synthesis and learn more details, thus effectively improving the fidelity of animated characters [15].

The advantage of the Actor-Critic method is that it has high efficiency and scalability when dealing with large state spaces [16]. In practice, the Actor-Critic method has been widely applied in fields such as robot control and game intelligence, which has achieved significant results. Hong et al. proposed a dual time scale stochastic approximation algorithm for a two-layer optimization problem, which included an external objective function and an internal strongly convex problem. The convergence speed of the Actor-Critic method was fast, providing an effective method for solving global optimal strategies and double-layer optimization problems [17]. Xi et al. proposed a new control strategy to address the shortcomings of traditional automatic power generation control in dealing with strong stochastic disturbances caused by renewable energy infiltration. The Actor-Critic method and incentive heuristic mechanism were introduced to improve the system control performance, improve the dynamic performance of the power system, and achieve regional optimal coordinated control [18]. Han et al. proposed a control method based on the Actor-Critic method for a reinforcement learning framework to address the stability of model free reinforcement learning in robot control tasks. Through empirical evaluation, the learned strategies could to some extent restore the system to a balanced state or path point when the system was disturbed by uncertain factors. This demonstrated its advantages in improving system robustness and stability [19]. Zhong et al. proposed a dynamic multi-channel access framework based on deep Actor-Critic reinforcement learning to effectively utilize limited spectrum resources. This considered the situation where both single and multiple users attempted to access the channel simultaneously, effectively improving the resource utilization [20].

The summary of related works related to the research content is shown in Table 1.

Table 1: Summary of related works

| Method | Result | Disadvantage | Reference |
|---|---|---|---|
| Animation character creation based on support vector machine and AR animated system | Users can easily create in-situ character animations that interact closely with different real-world environments | Not paying attention to animation character details | Ye et al. |
| Using chain assembly system to personify 3D animated characters | The animated characters created are more realistic and the creation efficiency is improved. | Not paying attention to animation character details | Arshad et al. |

| | | | |
|---|---|---|---|
| Building a facial expression editing method using deep learning | It is easy and intuitive visualization for facial expression editing. | The neural texture model used does not consider the view dependence of texture. | Paier et al. |
| Landmarks-driven two-stream networks that obtain data from multiple source images. | The fidelity of animated characters is improved. | Large computing resources and long time consuming. | Gu et al. |
| A dual time scale stochastic approximation algorithm framework | The Actor-Critic method has faster convergence rate, which solves the dual optimization and provides an efficient method | The Actor-Critic method is not applied to a specific domain. | Hong et al. |
| Introduce Actor-Critic method and incentive mechanism | The system control performance is improved, the dynamic performance of the power system is improved, and the regional optimal coordinated control is realized. | High computational complexity. | Xi et al. |
| Control method of reinforcement learning framework based on Actor-Critic method | When the system is disturbed by uncertain factors, it can restore the system to the equilibrium state or path point to a certain extent | The method is very sensitive to the choice of hyperparameters. | Han et al. |
| Dynamic multi-channel access framework based on deep Actor-Critic reinforcement learning | The resource utilization is improved. | The stability of the algorithm is poor. | Zhong et al. |

In summary, the development of animation production is becoming increasingly technological, and the Actor-Critic method is also applied in many fields. However, there is limited research on mouth shape generation. Some methods generate animated videos by selecting frames from specific character databases and combining them. However, these methods highly rely on specific roles and incur significant costs when transitioning to new speakers. Therefore, this study proposes an animation character mouth matching model based on the Actor-Critic method. Then, the generation of synchronized mouth movements is regarded as a sequence generation, achieving the goal of character independence by matching speech with corresponding features. The innovation of the research includes the generalization of the model, which can be generalized across different languages and speaker roles. Next is to expand to 2D to 3D, introducing 2D video datasets to compensate for the shortcomings of 3D character datasets. Finally, emotional feature extraction is performed by analyzing audio emotional features to recreate facial micro changes. This is to reflect the contextual information and emotional intensity in the input speech.

## 3 The development history of animation production and animated characters

Animation is an art and technical form that creates motion effects by continuously playing a series of still images or graphics [21]. Animation has become an indispensable part of the global cultural industry, providing audiences with endless creativity and entertainment [22]. Animation production is a complex process involving multiple steps, from concept and storyboard design, to character modeling and scene layout, and then to the creation of keyframe animation and interval frames [23]. This process also includes audio design, animation rendering, and post-processing. The final animated work can be published and distributed on media platforms such as movies, television, and the internet. The origin of animation can be traced back to the late 19th century, and its development history is summarized as shown in Figure 1.

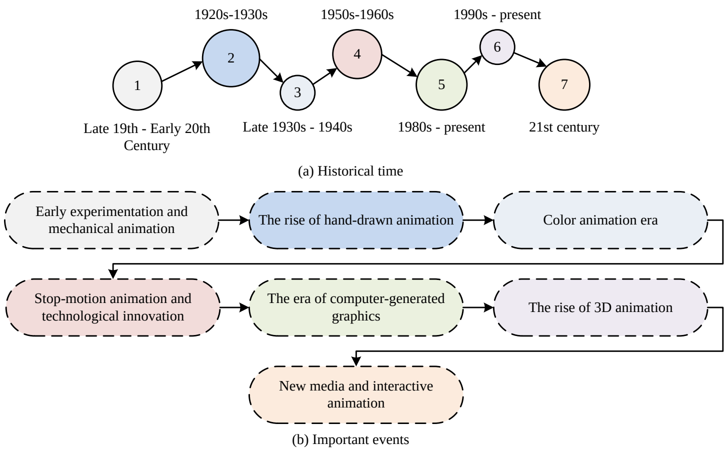

1920s-1930s
1950s-1960s
1990s - present
1
2
3
4
5
6
7
Late 19th - Early 20th Century
Late 1930s - 1940s
1980s - present
21st century
(a) Historical time
Early experimentation and mechanical animation
The rise of hand-drawn animation
Color animation era
Stop-motion animation and technological innovation
The era of computer-generated graphics
The rise of 3D animation
New media and interactive animation
(b) Important events

Figure 1: Development history of animation production

In Figure 1 (a), animation production has roughly gone through 7 historical stages and key developments. Figure 1 (b) shows seven key events in the development of animation production. Specifically, experimenters such as Émile Reynaud from France and J. from the United Kingdom Stuart Blackton attempted to conduct animation experiments using mechanical means such as rotating discs and slides during the early experimental and mechanical animation period (late 19th to early 20th century). The rise of hand drawn animation occurred from the 1920s to the 1930s. In 1928, Disney's Mickey Mouse became the first animation to synchronize sound and image, driving the mainstream development of hand drawn animation. The era of color animation emerged from the late 1930s to the 1940s, with Snow White in 1937 becoming the first full-length color animated film. The period of stop motion animation and technological innovation occurred from the 1950s to the 1960s, with 1955's Candy House being the first film to use stop motion animation technology. During this period, special effects animation films such as Disney's Cinderella (1950) and Mary Popins (1964) achieved great success.

The 1980s ushered in the era of computer-generated graphics. In 1982, Pixar's short film “Andre's Dream” became the first animation to use computer-generated graphics. Since the 1990s, the flourishing of 3D animation has become mainstream. Toy Story in 1995 was the first fully computer-generated full-length animated film, marking the dominant position of computer technology in animation production. During this period, many classic works emerged, including “The Lion King” (1994), “Frozen” (2013), and so on. In the 21st century, new media and interactive animation have emerged. With the popularization of the Internet and mobile devices, animation has entered a multi-platform era. New media forms such as interactive animation and virtual reality animation have gradually emerged. This evolution has witnessed the rapid progress of animation production technology and the innovative spirit of creators at different times. It is worth noting that animated characters play a crucial role in animation production, as shown in Figure 2.

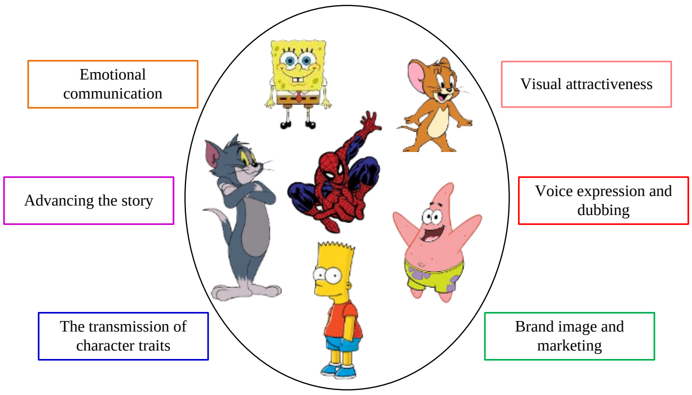

Emotional communication
Visual attractiveness
Advancing the story
Voice expression and dubbing
The transmission of character traits
Brand image and marketing

Figure 2: The key role of animated characters in animation

In Figure 2, animated characters are not only the main body of the story, but also key elements for emotional communication and visual attraction. Firstly, animated characters convey rich emotions through exquisite expressions and actions, allowing the audience to be more deeply immersed in the story plot. Secondly, as the driving force of the story, the conflicts, goals, and developmental trajectories of animated characters drive the entire story forward, increasing the depth and appeal of the story [26]. Each animated character is unique, showcasing various values and cultural characteristics through their words and actions, enriching the connotation of the entire work. Visually, the design of animated characters directly determines the attractiveness of a work. Clever images can attract the audience's attention, making the animation more engaging. In terms of voice, dubbing can express the personality and emotions of characters, giving them a unique timbre [27, 28]. In addition, some successful animated characters have become powerful brand images, surpassing the animation itself and driving the sales and successful marketing of related products [29]. Therefore, animated characters are not only characters, but also bridge the emotional connection between the audience and the work, contributing irreplaceable elements to the success and charm of animated works.

# 4 Establishment of mouth shape matching model for animated characters in action production

Animated characters play a crucial role in storytelling and visual appeal as key elements. Therefore, a mouth shape matching model is proposed for animated characters that makes them more realistic and vivid. The study aims to establish models from three aspects, namely audio feature extraction, facial feature extraction of animated characters, and the application of Actor-Critic methods.

## 4.1 Audio feature extraction and processing in animation character mouth matching model

When performing mouth shape matching in animation production, extracting key acoustic features from the input audio signal can enable the mouth shape matching model of animated characters. As a result, information, such as tone, speed, and emotion of speech, can be understood and captured. Therefore, the first step in constructing a model is to analyze the audio data. Then, the model can better adapt to the speech styles of different speakers and provide a foundation for subsequent mouth shape matching. The study selects Mel Frequency Cepstral Coefficients (MFCC) as the key method for audio feature extraction. The MFCC design is inspired by the way the human ear perceives sound, which can better simulate the perceptual characteristics of the human ear, making it more suitable for processing human speech signals. The MFCC extraction in the proposed model is shown in Figure 3.

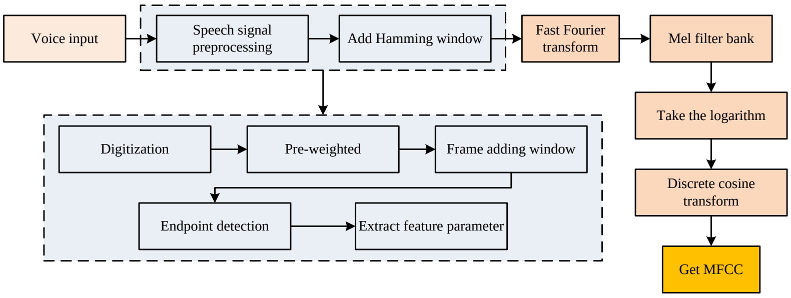


Figure 3: Flow chart of extracting MFCC

In Figure 3, to extract MFCC, the first step is to obtain a speech signal and input it. The preprocessing and Hamming window operation of the input speech audio signal is studied to distinguish between high-frequency signals and low-frequency signals. Then, the signal is converted to the frequency domain through fast Fourier transform. A Mel filter bank is used to uniformly distribute it on the Mel frequency scale, simulating human ear perception. The logarithm is taken to simulate nonlinear volume perception. Subsequently, the spectral signal is transformed into cepstral coefficients through discrete cosine transform, and the key MFCC is ultimately extracted. One advantage of MFCC is that it has relatively less interference with noise. Meanwhile, it is easier to display the non-uniform structure of frequency in feature representation through the nonlinear characteristics of Mel scale. There is a mapping relationship between the Mel frequency and the actual frequency of the signal. This is approximately linear in the frequency range below 1kHz and logarithmic in the high frequency range. The research establishes a relationship model between the two as shown in equation (1).

$$F_{mel} = 2595\log_{10}\left(1+\frac{F_t}{700}\right) \tag{1}$$

In equation (1), $F_{mel}$ and $F_t$ represent the Mel frequency and actual frequency, respectively. In the preprocessing process, the audio signal needs to be processed in frames. The simplification effect of longer frames will be compromised due to the difficulty in extracting effective features from shorter frames.

Therefore, the entire audio is divided into a set of 20-40 millisecond frame segments. A pre-emphasis operation is adopted to counteract the suppression effect of the system on the high-frequency part of the speech signal. The processed effect is shown in Figure 4.

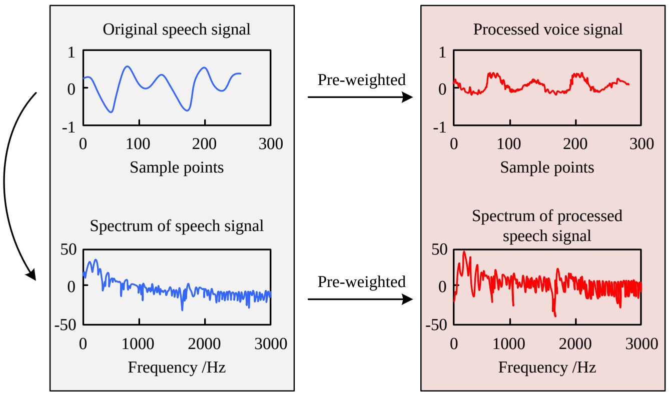


Figure 4: Contrast of signal and spectrum before and after pre-weighted processing

In Figure 4, the purpose of the pre-emphasis operation is to emphasize the high-frequency part and compensate for the high-frequency components in the speech signal that are suppressed by the system due to vocal fold and lip effects. A first-order limited length impulse response feedback filter is introduced to weight the speech signal, making the amplitude of the high-frequency part relatively enhanced. This can better capture high-frequency features in subsequent processing, improve the clarity and distinguishability of speech signals. Speech signals exhibit non-translational characteristics under the influence of changes in motion state. For the convenience of processing, this study divides longer speech signals into shorter speech signal segments and introduces frame shift to ensure a smooth transition in information processing. Frame shifting refers to a certain overlap between adjacent speech frames, rather than tightly connecting the beginning and end of the frames. This can ensure a smooth transition between two frames during the framing stage, avoiding the loss of key information. The combination of finite terms may cause Gibbs effect when using Fourier transform to analyze the spectral characteristics of frame signals. The Gibbs effect may deviate from the original speech signal after signal processing at the beginning and end of the frame. Therefore, a Hamming window is added. The Hamming window function is shown in equation (2).

$$\omega(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \qquad (2)$$

In equation (2), $\omega(n)$ is the value of the window function at the discrete sequence $n$, and $N$ is the length of the window. Window operation can minimize the amplification deviation caused by discontinuous parts in the application of fast Fourier transform, obtain a more continuous wave output, and reduce the influence of the Gibbs effect. This helps improve the accuracy of spectrum analysis. In addition, the characteristics of speech signals change over time in actual speech. Therefore, for each time $t$, a delta coefficient is calculated by weighting and summing the MFCC features of a certain number of frames before and after to reflect the dynamic characteristics of that time. The delta coefficient is shown in equation (3).

$$delta(t) = \frac{\sum_{i=1}^{I} n \cdot \left[MFCC(t+i) - MFCC(t-i)\right]}{2 \cdot \sum_{i=1}^{I} i^2} \quad (3)$$

In equation (3), $delta(t)$ represents the dynamic feature, $i$ represents the cycle in the calculation, and $MFCC(\ )$ represents the MFCC feature vector at a certain time. $I$ represents the size of the calculation window, which specifies how many frames of dynamic information to consider when calculating the delta coefficient. The processing of research helps to capture the temporal changes in speech signals, providing more comprehensive information for audio analysis of mouth movements.

## 4.2 Facial feature extraction and expression processing of animated characters

In addition to considering the matching of audio and mouth shape, this study focuses on the emotional features conveyed in speech to make the facial features of animated characters more realistic. Then, it reflects these emotional characteristics in the facial animation performance of the character. Therefore, the study introduces the Facial Action Coding System (FACS). FACS systematically encodes facial expressions by dividing different facial movements into basic elements called AU. The study first names and distinguishes various regions of the face, as shown in Figure 5.

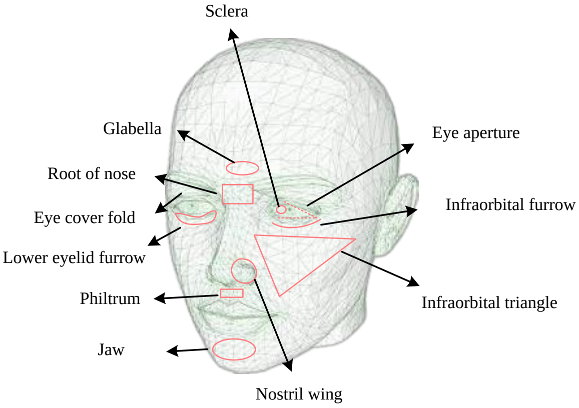


Figure 5: The naming and location of each area of the human face

Figure 5 shows that the human face is divided into 11 regions, namely sclera, glabella, root of nose, eye cover fold, lower eyelid furrow, jaw, nostril wing, infraorbital triangle, infraorbital furrow, and eye aperture. FACS contains multiple AUs, each representing the movement of muscles or muscle groups in different parts of the face. The combination of these AUs can form various facial expressions, making FACS a detailed and comprehensive system for recording and analyzing small changes in facial movements. The study lists some major AU codes and their meanings, as shown in Table 2.

Table 2: Main action units in FACS

| Action unit | Description | Action unit | Description | Action unit | Description |
|---|---|---|---|---|---|
| AU1 | Inner eyebrow raised | AU13 | Pull the corners of the mouth up | AU25 | Parted lips reveal teeth |
| AU2 | Outer eyebrow raised | AU14 | The corners of the mouth shrink towards the teeth | AU26 | Parted lips see tongue |
| AU4 | The eyebrows are generally lowered | AU15 | Pull the corners of the mouth straight down | AU27 | Parted lips see throat |
| AU5 | Raise the eyelid | AU16 | Pull the lower lip down | AU28 | Suck the lips over the teeth |
| AU6 | Lift the cheek | AU17 | Squeeze the lower lip up to the top | AU41 | Slightly lowered the lid |
| AU7 | Eye contraction | AU18 | Tuck the mouth in the middle | AU42 | Lower the eyelids |
| AU9 | Retraction and lifting nose | AU20 | Pull the lips back | AU43 | Close the eyes |
| AU10 | Lift the upper lip | AU22 | Curl the lips into a funnel | AU44 | Lower eyelid up |
| AU11 | Deepen the middle nasal lip | AU23 | Clench the lips into one word | AU45 | Blink the eyes |
| AU12 | Turn up the corners of the mouth | AU24 | Squeeze the lips together | AU46 | Monocular blink |

In Table 2, FACS encodes many subtle facial expressions, which is crucial for accurately interpreting and simulating subtle differences in facial expressions. Analyzing each expression yields a combination of AU. For example, the combination of happy expressions is AU6+AU12, which means lifting the cheeks and lifting the corners of the mouth. The combination of sad expressions is AU1+AU4+AU15, which means lifting the internal eyebrows, lowering the eyebrows as a whole, and pulling the corners of the mouth vertically downwards. The combination of surprise expressions is AU1+AU2+AU5+AU26, which means raising the inner eyebrows, raising the outer eyebrows, lifting the upper eyelids, and separating the lips to see the tongue. The combination of fear expressions is quite complex, consisting of AU1+AU2+AU4+AU5+AU7+AU20+AU26. It includes a combination of lifting the internal eyebrows, raising the external eyebrows, lowering the eyebrows as a whole, lifting the upper eyelids, contracting the eyes, pulling the lips backwards, and separating the lips to see the facial features of the tongue. Principal Component Analysis (PCA) is used to generate face grid after obtaining facial features when expressions change through FACS. Then, key features can be extracted from face data and a low-dimensional representation of faces can be generated. The goal of PCA is to find a transformation that projects the original data into a low-dimensional space while preserving the variance of the original data as much as possible. Firstly, facial feature points are preprocessed, as shown in equation (4).

$$x_c = x - \mu \tag{4}$$

In equation (4), $x_c$ represents the data after centralization, $x$ represents the feature vector of the original dataset, and $\mu$ represents the mean vector. The covariance matrix $C$ for the centralized data is calculated, as shown in equation (5).

$$C = \frac{1}{n}\sum\nolimits_{i=1}^{n} x_{c,i} x_{c,i}^T \tag{5}$$

In equation (5), $n$ represents the number of samples, and $x_{c,i}$ represents the $i$ -th centralized sample. The eigenvalues of covariance matrix are decomposed to obtain the eigenvalues and corresponding eigenvectors. The eigenvector corresponding to the largest eigenvalue defining the main direction of the PCA space can be selected according to the size of the eigenvalues. The transformation matrix $W$ is constructed using the selected eigenvectors. The original data are projected into the PCA space, as shown in equation (6).

$$x_{PCA} = W^T x_c \tag{6}$$

It is worth noting that the study constructs several assumptions when applying PCA. The assumptions include that the data changes are primarily linear, that the centralized feature points are independent of each other, that the eigenvalue distribution allows most of the variance to be preserved by selecting a small number of principal components, that the data need to be normalized to eliminate scaling effects, and that a certain reconstruction error is accepted. Therefore, the face $Q$ is approximately represented as shown in equation (7).

$$Q = \overline{E} + \sum_{i=1}^{l} \alpha_i E_i \tag{7}$$

In equation (7), $\overline{E}$ represents the average face, $E_i$ represents the $i$ -th PCA face vector, and $\alpha_i$ is the coefficient. The study first needs to obtain the closest grid of neutral expressions in PCA space. The definition of feature point matching energy is shown in equation (8).

$$E_1 = \sum_{j=1}^{m_i} \left\| v_{i_j} - c_j \right\|^2 + \sum_{k=1}^{m_c} \left\| M v_{c_k} - s_k \right\|^2 \tag{8}$$

In equation (8), $c_j$ represents the 3D position of feature point $j$, and $v_{i_j}$ represents its mesh vertex. $s_k$ represents the two-dimensional spatial feature points of the face, $v_{c_k}$ represents the corresponding three-dimensional spatial feature points of the face, and $M$ represents the projection matrix from two-dimensional to three-dimensional. The index of contour feature points on the grid can be determined through equation (8). The energy term for matching depth maps is defined as equation (9).

$$E_2 = \sum_{j=1}^{n_d} \left\| v_{d_j} - p_j \right\|^2 \tag{9}$$

In equation (9), $v_{d_j}$ represents the vertex of the mesh, and $p_j$ is the point closest to $v_{d_j}$ in the depth map. The study aims to minimize energy through a series of least squares optimizations, which are iterated 5 to 8 times. Between consecutive iterations, update the mesh vertices corresponding to the contour feature points in equation (8) and the nearest point of each mesh vertex in equation (9). After obtaining the grid of neutral expressions, the grid of other expressions can be calculated based on the AU changes in FACD. Meanwhile, the grid deformation can use all facial feature points on color images as additional positional constraints.

## 4.3 The actor-critic method in animation character mouth matching model

The study regards the task of generating mouth animation as generating a sequence. Each time step corresponds to a specific facial expression or mouth state. The model uses the speech audio features and facial features of the character from the previous text as input states for the current time step, thereby predicting the facial features for the next time step.

This study uses the Actor-Critic method, which allows the model to gradually unfold the sequence of mouth animation generation. At each time step, the Actor generates a prediction of facial features, and the Critic evaluates the quality of this prediction and provides feedback. The schematic diagram of the Actor-Critic method is shown in Figure 6.

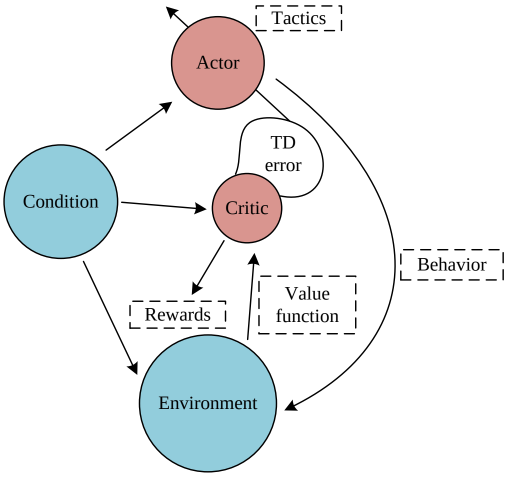


Figure 6: Actor-Critic process diagram

In Figure 6, firstly, the method begins training by initializing the parameters of the Actor network and the Critic network. Then, based on the current state, the Actor network is used to generate actions and execute them, observing rewards and new states. Next, a Critic network is used to calculate the value function or state action value function of the current state. It updates the parameters of the Actor network based on the policy gradient algorithm to maximize the expected reward. Meanwhile, the algorithm updates the parameters of the Critic network based on the value function to reduce estimation errors. This process will continue to repeat until convergence or termination conditions are met. Through this approach, the Actor-Critic method can gradually learn the optimal strategy and achieve maximum expected return. The strategy gradient formula is shown in equation (10).

$$\nabla_\theta J = \sum\nolimits_{s,a} \pi(a \mid s) \nabla_\theta \log \pi(a \mid s) Q(s,a) \quad (10)$$

In equation (10), $J$ is a function of expected return, $\theta$ is a policy parameter, $\pi(a \mid s)$ is a policy, and $Q(s,a)$ is a state action value function. The policy gradient theorem is the basis of this formula, which states that the policy gradient is equal to the gradient of the product of the state action value function $Q(s,a)$ and the policy $\pi(a \mid s)$ with respect to the policy parameter $\theta$. The gradient of expected returns with respect to policy parameters can be obtained by summing up the state action space. The formula for updating the value function is shown in equation (11).

$$\Delta V(s) = \sum\nolimits_a \pi(a \mid s)\left[Q(s,a) - V(s)\right] \quad (11)$$

In equation (11), $V(s)$ represents the state value function. This formula is derived based on the Bellman equation. The Bellman equation states that the state value function is equal to the weighted sum of the expected returns obtained by taking all possible actions in that state. The update amount $\Delta V(s)$ of the state value function can be obtained by calculating the difference between the state action value function $Q(s,a)$ and the state value function $V(s)$ and performing a weighted sum. The strategy update formula is shown in equation (12).

$$\theta \leftarrow \theta + \beta \nabla_\theta \log \pi(a \mid s) Q(s,a) \quad (12)$$

In equation (12), $\beta$ represents the learning rate. This formula is derived based on policy gradients. The update direction of policy parameter $\theta$ can be obtained by calculating the policy gradient $\nabla_\theta \log \pi(a \mid s) Q(s,a)$ through research. Then, the learning rate $\beta$ is used to control the update step size and updates the policy parameters along the gradient direction. In this way, the study can gradually adjust strategies to maximize expected returns. In summary, the overall process of the animation character mouth shape matching model based on the Actor-Critic method is shown in Figure 7.

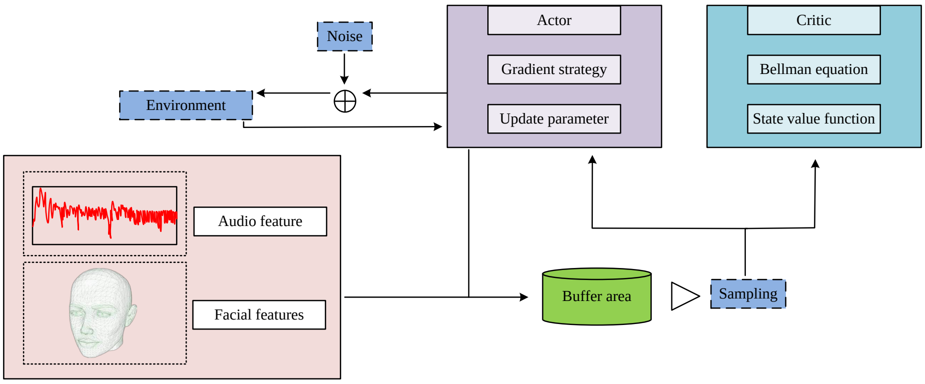


Figure 7: Animation character mouth matching model

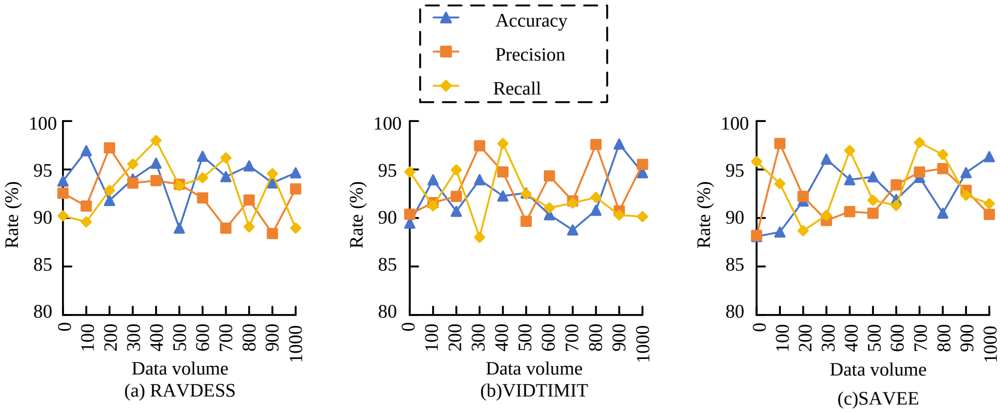
In Figure 7, a parameterized Actor model is trained to generate a sequence. The state includes audio features and previously generated expression features, while the action is the expression feature for the next time step. In addition, a parameterized Critic model was introduced to guide the Actor model in generating more realistic mouth movements through rewards. During the training process, real sequence data and synthetic sequence data are used to optimize the Critic model. To improve performance, a replay buffer is introduced and a deep deterministic policy gradient algorithm is used for training to minimize the error of the Critical network.

# 5 Performance verification and actual effect analysis of animation character mouth matching model

To explore the effectiveness and superiority of the research content, this study conducted experimental verification on the mouth shape matching model of animated characters. Firstly, it compared the performance of the model on different datasets, and then compared the different methods with the generated renderings of the model.

## 5.1 Performance verification of mouth shape matching model for animated characters

To verify the effectiveness and superiority of the mouth shape matching model for animated characters, this study selected three databases: Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS), Video Limit (VIDTIMIT), and Surrey Audio Visual Express Emotion (SAVEE) from the University of Surrey in the UK. RAVDESS is a database used for studying emotion recognition, containing audio and video data from different actors. The voice and song samples in this database cover different emotional states, such as joy, sadness, anger, etc. VIDTIMIT is a database used for studying audio and video synthesis. It is an extension of the TIMIT database, providing video data on facial and lip movements corresponding to speech. SAVEE contains speech samples from different emotional states in English, used for emotion recognition and related research. The study randomly selected 3/5 of the data in the database as the training set, and the remaining 2/5 of the data as the testing set. The training results are shown in Figure 8.



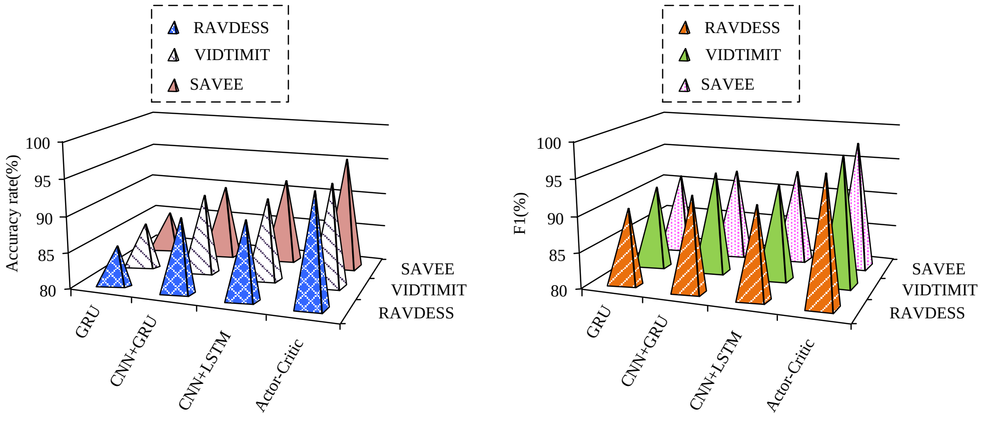
(a) RAVDESS

(b)VIDTIMIT

(c)SAVEE

Figure 8: Comparison of training results

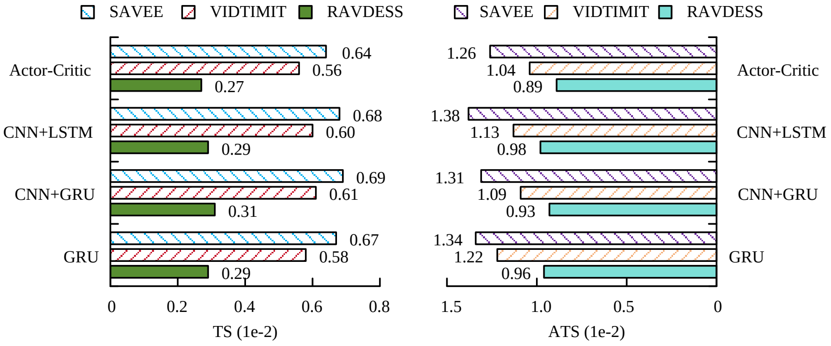
Figure 8 (a) shows that the training accuracy, precision, and recall of the RAVDESS database are all above 90%. Figure 8 (b) shows that the training accuracy and recall of the VIDTIMIT database are both above 90%. The accuracy is 89.56% when the data volume is 900, and the rest are above 90%. Figure 8 (c) shows that the training results of the SAVEE database are also excellent, with accuracy, precision, and recall basically above 90%. Therefore, this study further examined the performance indicators on the test set and compared them with mouth shape matching models supported by Gated Recurrent Unit (GRU), Convolutional Neural Network+Gated Recurrent Unit (CNN+GRU), and CNN+Long Short-Term Memory (CNN+LSTM) models. GRU is suitable for processing sequential data and is able to capture long-term dependencies in time series. CNN+GRU combines convolutional neural networks (for extracting visual features) and GRU (for processing time series features) to improve the accuracy of mouth matching. CNN+LSTM replaces the GRU with a Long Short-Term Memory (LSTM), which is equally good at processing sequence data and is able to learn long-term dependencies on information. In the experiment, the key hyperparameter settings are as follows. The evaluation frequency is to evaluate the model performance every 1000 training steps. The learning rate of the Adam optimization algorithm is adjusted every 3×10^4 steps. The batch size is 256. The discount factor is used to determine the current value of future rewards, which is set to 0.99. The temperature parameter is used to control the updating amplitude of the policy gradient, which is set to 0.005. The comparison between accuracy and F1 value is shown in Figure 9.



(a) Accuracy comparison (b) F1 value comparison

Figure 9: Comparison results of accuracy and F1 value of four algorithms

Figure 9 (a) shows that for accuracy, GRU is below 85%. CNN+GRU and CNN+LSTM are basically around 90%. Among them, CNN+GRU has the highest accuracy on the VIDTIMIT database, reaching 90.67%, and CNN+LSTM has the highest accuracy on the SAVEE database, reaching 91.34%. The accuracy of the Actor-Critic method proposed in the study is superior to the first three, reaching a maximum of 96.48%, with an average of 95.61% on the three databases. In Figure 9 (b), the F1 value of the Actor-Critic method is also higher than that of the three comparison algorithms, with an average of 97.13%. Furthermore, the study compared the time series analysis performance of the four algorithms using Temporal Smoothness (TS) and Adaptive Temporal Smoothness (ATS) as indicators. The results are shown in Figure 10.



(a) TS (b) ATS

Figure 10: Comparison of TS and ATS

In Figure 10 (a), the Actor-Critic method has the smallest TS in all three databases, with values of 0.64, 0.56, and 0.27, respectively. In Figure 10 (b), the ATS of the Actor-Critic method is also the smallest, with values of 1.26, 1.04, and 0.89, respectively. This means that the Actor-Critic method has a smoother temporal variation, which helps to reduce noise and abrupt changes, making the mouth matching results more stable and coherent. The study continued to compare the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) of the four algorithms, as shown in Figure 11.

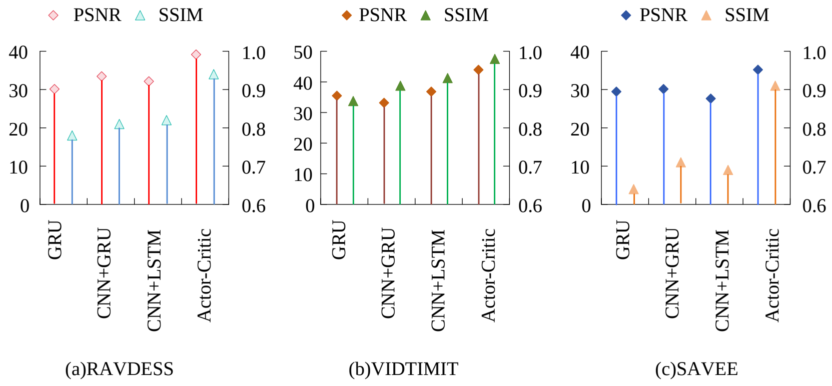


Figure 11: Comparison of PSNR and SSIM

Figure 11 (a) shows that on the RAVDESS database, the Actor-Critic method has a PSNR of 39.67 and a SSIM of 0.93. Figure 11 (b) shows that on the VIDTIMIT database, the Actor-Critic method has a PSNR of 46.58 and a SSIM of 0.96. According to Figure 11 (c), on the SAVEE database, the PSNR of the Actor-Critic method is 37.06 and the SSIM is 0.91. The results indicate that the Actor-Critic method has better image quality and more accurate preservation of structure and details compared to the comparative algorithms. Finally, the accuracy of the four methods is compared, and the results are shown in Table 3.

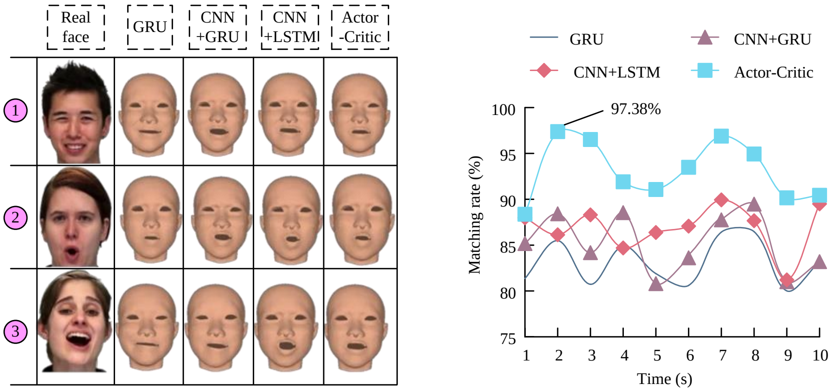
Table 3: Accuracy comparison of the four methods

| - | Accuracy (%) | | | |
|---|---|---|---|---|
| | GRU | CNN+GRU | CNN+LSTM | Actor-Critic |
| RAVDESS | 89.46 | 91.24 | 92.43 | 96.85 |
| VIDTIMIT | 88.52 | 91.94 | 93.02 | 95.97 |
| SAVEE | 87.46 | 90.61 | 92.99 | 97.05 |
| Mean value | 88.48 | 91.26 | 92.81 | 96.62 |

From Table 3, the Actor-Critic method has the highest accuracy on the three datasets, reaching 96.62% on average. The average accuracy of GRU, CNN+GRU, and CNN+LSTM is 88.48%, 91.26% and 92.81%, respectively. Therefore, the mouth matching model based on Actor-Critic method has high precision, F1 value, and accuracy rate, which has important practical application significance in the animation industry. It can generate highly synchronized and emotion-rich animation mouth shapes with speech, significantly improving the animation quality and enhancing the expressiveness of characters.

## 5.2 Analysis of Actual Effects of Animation Character Mouth Matching Model

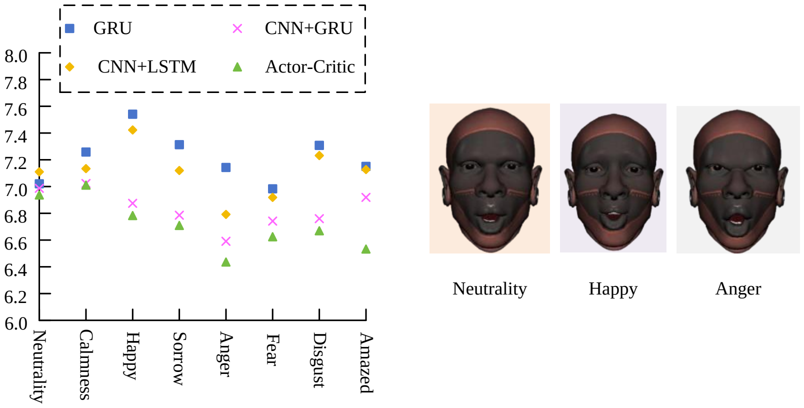
To further demonstrate the superiority of the animation character mouth matching model supported by the Actor-Critic method proposed in the study, the effect images generated by various comparison algorithms were compared. The matching effect and matching rate of the animation character's mouth shape matching effect are shown in Figure 12.



(a) Effect drawing (b) Matching rate

Figure 12: Animated characters' mouth matching effect

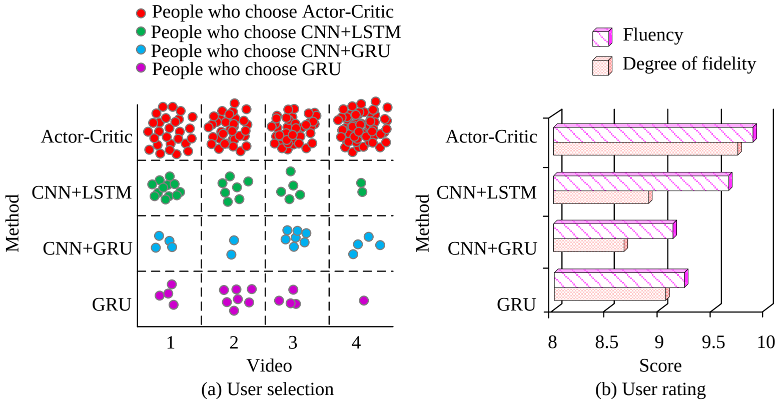
Figure 12 (a) shows that the animated character facial model generated by the Actor-Critic method is more realistic and more similar to the expressions of real faces. Figure 12 (b) calculates and compares the matching rate of the matching effect map within 10 seconds. Figure 12 (b) shows that the matching rate of the Actor-Critic method is consistently better than the four comparison algorithms. The highest matching rate is achieved in the second, reaching 97.38%, with an average matching rate of 94.29%. The average matching rate of GRU renderings is 82.54%. The average matching rates of CNN+GRU and CNN+LSTM renderings are 84.97% and 83.93%, respectively. To further understand the performance of each model on the test set, this study calculated the mean square error of parameters corresponding to different emotions and facial data. This study provided facial models that said the same sentence under three different emotions. This is to better observe the changes in facial expressions, as shown in Figure 13.



(a) Mean square error (b) Different emotional mouth movements

Figure 13: Animated characters' mouth matching effect

Figure 13 (a) shows that, overall, the mean square error of the Actor-Critic method is consistently lower than other comparison methods. When the emotion is anger, the mean square error is the smallest, at 6.436. This is because angry expressions contain more AUs and facial features are more prominent. The average mean square error of the Actor-Critic method is only 6.639. Figure 13 (b) shows the differences in facial expressions of the same sentence under different emotions. This indicates that the Actor-Critic method not only has a higher mouth shape matching rate, but also a more realistic expression. Finally, the study presented users with four animated videos generated by each of the four methods and asked them to choose videos that were more natural and realistic. Meanwhile, each video was rated. The results are shown in Figure 14.



(a) User selection

(b) User rating

Figure 14: User selection and rating results

Figure 14 (a) shows that users believe that the mouth shape matching and facial expressions of animated characters in videos generated by the Actor-Critic method are more vivid and realistic. After careful statistics, more than 90% of users choose the Actor-Critic method, and the selection rate of video number 4 even reaches 98.64%. Figure 14 (b) shows that for the videos generated by the Actor-Critic method, the fluency and realism ratings reached 9.87 and 9.72, respectively. However, in the comparison algorithm, the highest flow smoothness is only 9.64. For videos from CNN+LSTN, the highest fidelity is only 9.04, while for videos from GRU. This indicates that the Actor-Critic method has achieved significant success in mouth shape matching and character expression generation, which has been highly recognized by users. This can demonstrate the effectiveness and superiority of the proposed animation character mouth shape matching model.

# 6 Discussion

The model of mouth matching was put forward. First of all, from the perspective of performance indicators, the proposed mouth matching model was superior to the methods listed in Table 1 in terms of accuracy, F1 value, PSNR, and SSIM. Specifically, the average accuracy of the research model on the three databases reached 95.61%, the F1 value was 97.13%, the PSNR was about 41.77, and the SSIM was about 0.93. In contrast, other methods in Table 1 had lower performance on these indicators, for example, the method using support vector machine and AR animator system [12], and the method using chain assembly system [13]. Although they improve the authenticity of animated characters, they do not pay special attention to the accuracy of mouth matching and the richness of emotional expression. The reasons for the difference in performance can be attributed to several key factors. The proposed model adopted Actor-Critic method and combined audio feature extraction and facial action coding, which not only improved the accuracy of mouth matching, but also enhanced the richness of emotional expression. In addition, 2D video datasets were introduced to make up for the shortcomings of 3D character datasets, helping to improve the generalization ability of the model. The advantages of the proposed model are that it can optimize mouth shape generation by Actor-Critic method, improve the learning and prediction ability of the model, and capture the emotional features in audio and reflect them in the facial animation of animated characters. However, the model mainly focuses on mouth matching, which does not deeply analyze other details such as blinking and turning the head. In summary, the proposed model provides a new solution for animation production by introducing reinforcement learning and feature extraction techniques. It not only improves the accuracy of mouth matching and the emotional expression ability of animated characters, but also lays a foundation for the further promotion and innovation of animation production technology.

In addition, MFCC was used as the key method for audio feature extraction. However, MFCC may be sensitive to noise, whose sensitivity to tone change may lead to feature distortion in speech processing with different intonation or pitch. At the same time, MFCC has limited ability to capture nonlinear features and has a certain dependence on audio quality. Therefore, MFCC can be replaced by Perceptual Linear Prediction (PLP), another feature extraction method simulating auditory characteristics of human ears. PLP takes the perceptual nonlinearity into account in the extraction. Therefore, PLP may have better ability to capture the nonlinear features of speech signals.

When using the Actor-Critic method, the research did not deeply consider the setting of hyperparameters in the Actor-Critic method. The parameter setting in the research referred to existing literature. Therefore, the subsequent research can further carry out a detailed analysis of the hyperparameter setting scheme. In addition, as the size of the dataset increases, the calculation requirements for the Actor-Critic method are also the same. The scalability of the model needs to be considered to handle larger datasets. For example, the model can be trained on multiple Gpus simultaneously through data parallelization techniques, or the model parallelization method can be used to distribute different parts of the model to different computing devices. In addition, incremental learning allows models to be updated gradually as new data arrive, rather than being trained from scratch each time, which not only increases efficiency but also reduces the waste of computational resources. Distributed training frameworks can also enable models to be trained on multiple compute nodes to handle larger datasets. The performance of the model can be maintained or improved through these methods. The computational complexity can be effectively managed and reduced to ensure that the model can adapt to the growing volume of data.

As automation advances, technologies must ensure that innovation is accompanied by appropriate ethical guidelines and regulatory measures to prevent misuse of technology, protect personal privacy, clarify liability, and promote the healthy development of technology. In addition, the use of automated methods to create realistic animation requires consideration of how to maintain the artistic and creative nature of animation. The automated methods should ensure that the spread of technology does not exacerbate social inequalities and update education and training systems to adapt to technological change. With these measures, future research can make better use of automated lip matching technology, while ensuring that it is applied ethically and responsibly.

# 7 Conclusion

A mouth shape matching model based on reinforcement learning and feature extraction is proposed to address the issue of mouth shape matching in animation production. The study first explores the development history of animation production and illustrates the importance of animated characters. Then, in this context, audio feature extraction, facial feature extraction, and Actor-Critic method are combined to complete the mouth shape matching of animated characters. Meanwhile, its final experimental verification is carried out. The experimental results showed that the model training results were good, with accuracy, precision, and recall basically above 90%. On the test set, the average accuracy and F1 value of the Actor-Critic method were better than GRU, CNN+GRU, and CNN+LSTM, reaching 95.61% and 97.13%, respectively. In addition, the Actor-Critic method had the smallest TS in all three databases, with values of 0.64, 0.56, and 0.27, respectively. ATS was also the smallest, with values of 1.26, 1.04, and 0.89, respectively. The average PSNR and SSIM of the Actor-Critic method on the three databases were approximately 41.77 and 0.93, respectively. Finally, a user survey was conducted, and over 90% of users chose the Actor-Critic method. Overall, this study provides an advanced mouth shape matching model for the animation production, making significant contributions to improving the expressiveness and realism of animated characters. Although the research delves into the emotional changes in mouth matching, the simulation of key details such as blinking and turning of the head has not been addressed. Future research can focus on simulating the natural rhythm of blinking and the fluency of head movements to enhance the realism of animated characters. In addition, there are important directions to improve the naturalness of character animation, including capturing facial micro-expressions, developing context-aware action generation algorithms, and realizing personalized and real-time action generation. At the same time, multi-modal emotion analysis, user interaction customization, and cross-cultural emotion expression research will further enrich the character's emotional level and expression.

# Funding statement

The research is supported by: This work was supported (in part) by the Scientific research start-up fund project for introducing high-level talents from Liming Vocational University, under Grant (No. 24YG01S).

# Conflict of interest

There is no conflict of interest between the authors.

# References

[1] Y. Deng, “Fluid equation-based and data-driven simulation of special effects animation,” Advances in Mathematical Physics, vol. 2021, no. 4, pp. 1-11, 2021. https://doi.org/10.1155/2021/7480422

[2] J. Zhang, G. Liao, and N. Li, “Combining active learning and local patch alignment for data-driven facial animation with fine-grained local detail,” Neurocomputing, vol. 398, no. 7, pp. 431-441, 2020. https://doi.org/10.1016/j.neucom.2020

[3] W. Paier, A. Hilsmann, and P. Eisert, “Interactive facial animation with deep neural networks,” IET Computer Vision, vol. 14, no. 6, pp. 359-369, 2020. https://doi.org/10.1049/iet-cvi.2019.0790

[4] S. Long, S. Andreopoulos, S. Patterson, J. Enkinson, and D. P. Ng, “Metabolism in motion: Engaging biochemistry students with animation,” Journal of Chemical Education, vol. 98, no. 5, pp. 1795-1800, 2021. https://doi.org/10.1021/acs.jchemed.0c01498

[5] L. Wang, and J. Kim, “Exploring the caricature style identification and classification using convolutional neural network and human-machine interaction under artificial intelligence,” International Journal of Humanoid Robotics, vol. 19, no. 3, pp. 26-38, 2022. https://doi.org/10.1142/S0219843622400096

[6] W. Paier, A. Hilsmann, and P. Eisert, “Example-based facial animation of virtual reality avatars using auto-regressive neural networks,” IEEE Computer Graphics and Applications, vol. 41, no. 4, pp. 52-63, 2021. https://doi.org/10.1109/MCG.2021.3068035

[7] R. Ploetzner, S. Berney, and M. Bétrancourt, “A review of learning demands in instructional animations: The educational effectiveness of

animations unfolds if the features of change need to be learned," Journal of Computer Assisted Learning, vol. 36 , no. 6, pp. 838-860, 2020. https://doi.org/10.1111/jcal.12476

[8] C. A. Sanchez, and K. Weber, "Using relevant animations to counter stereotype threat when learning science," Journal of Applied Research in Memory and Cognition, vol. 8，no. 4, pp. 463-470, 2019. https://doi.org/10.1016/j.jarmac.2019.08.003

[9] X. Xu, "Multimedia VR image improvement and simulation analysis based on visual VR restructuring algorithm," Informatica, vol. 48, no. 1, 2024. https://doi.org/10.31449/inf.v48i1.5368

[10] A. Berg, D. Orraryd, A. J. Pettersson, and M. Hulténl, "Representational challenges in animated chemistry: self-generated animations as a means to encourage students' reflections on sub-micro processes in laboratory exercises," Chemistry Education Research and Practice, vol. 20, no. 6, pp. 710-737, 2019. https://doi.org/10.1039/C8RP00288F

[11] M. Hanif, "The development and effectiveness of motion graphic animation videos to improve primary school students' sciences learning outcomes," International Journal of Instruction, vol. 13, no. 3, pp. 247-266, 2020. https://doi.org/10.29333/iji.2020.13416a

[12] H. Ye, K. C. Kwan, W. Su, and H. Fu, "ARAnimator: In-situ character animation in mobile AR with user-defined motion gestures," ACM Transactions on Graphics, vol. 39, no. 4, pp. 1-83, 2020. https://doi.org/10.1145/3386569.3392404

[13] M. R. Arshad, K. H. Yoon, A. A. A. Manaf, and M. A. Ghazali, "Physical rigging procedures based on character type and design in 3D animation," International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 3, pp. 4138-4147, 2019. https://doi.org/10.35940/ijrte.C5484.098319

[14] W. Paier, A. Hilsmann, and P. Eisert, "Interactive facial animation with deep neural networks," IET Computer Vision, vol. 14, no. 6, pp. 359-369, 2020. https://doi.org/10.1049/iet-cvi.2019.0790

[15] K. Gu, Y. Zhou, and T. Huang, "Flnet: landmark driven fetching and learning network for faithful talking facial animation synthesis," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 10861-10868, 2020. https://doi.org/10.1609/aaai.v34i07.6717

[16] J. Li, R. Wang, and K. Wang, "Service function chaining in industrial Internet of things with edge intelligence: A natural actor-critic approach." IEEE Transactions on Industrial Informatics, vol. 19, no. 1, pp. 491-502, 2022. https://doi.org/10.1109/TII.2022.3177415

[17] M. Hong, H. T. Wai, Z. Wang, and Z. Yang, "A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic," SIAM Journal on Optimization, vol. 33, no. 1, pp. 147-180, 2023. https://doi.org/10.1137/20M1387341

[18] L. Xi, J. Wu, Y. Xu, and H. Sun, "Automatic generation control based on multiple neural networks with actor-critic strategy," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 6, pp. 2483-2493, 2020. https://doi.org/10.1109/TNNLS.2020.3006080

[19] M. Han, L. Zhang, J. Wang, and W. Pan, "Actor-critic reinforcement learning for control with stability guarantee," IEEE Robotics and Automation Letters, vol. 5, no. 4, pp. 6217-6224, 2020. https://doi.org/10.1109/LRA.2020.3011351

[20] C. Zhong, Z. Lu, M. C. Gursoy, and S. Velipasalar, "A deep actor-critic reinforcement learning framework for dynamic multichannel access," IEEE Transactions on Cognitive Communications and Networking, vol. 5, no. 4, pp. 1125-1139, 2019. https://doi.org/10.1109/TCCN.2019.2952909.

[21] S. Starke, Y. Zhao, F. Zinno, and T. Komura, "Neural animation layering for synthesizing martial arts movements," ACM Transactions on Graphics, vol. 40, no. 4, pp. 1-16, 2021. https://doi.org/10.1145/3450626.3459881

[22] P. Preethi, and H. R. Mamatha, "Region-based convolutional neural network for segmenting text in epigraphical images," Artificial Intelligence and Applications, vol. 1, no. 2, pp. 119-127, 2023. https://doi.org/10.47852/bonviewAIA2202293

[23] H. Chen, D. Zhao, and J. Barbič, "Capturing animation-ready isotropic materials using systematic poking," ACM Transactions on Graphics, vol. 42, no. 6, pp. 1-27, 2023. https://doi.org/10.1145/3618406

[24] J. Wolper, Y. Fang, M. Li, J. Lu, M. Gao and C. Jiang, "CD-MPM: continuum damage material point methods for dynamic fracture animation," ACM Transactions on Graphics, vol. 38, no. 4, pp. 1-15, 2019. https://doi.org/10.1145/3306346.3322949

[25] H. Eom, D. Han, J. S. Shin, and J. Noh, "Model predictive control with a visuomotor system for physics-based character animation," ACM Transactions on Graphics, vol. 39, no. 1, pp. 1-11, 2019. https://doi.org/10.1145/3360905

[26] B. John, S. Jörg, S. Koppal, and E. Jain, "The security-utility trade-off for iris authentication and eye animation for social virtual avatars," IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 5, pp. 1880-1890, 2020. https://doi.org/10.1109/TVCG.2020.2973052

[27] M. Guo, F. Xu, S. Wang, Z. Wang, M. Lu, X. Cui, and X. Ling, "Synthesis, style editing, and animation of 3D cartoon face," Tsinghua Science and Technology, vol. 29, no. 2, pp. 506-516, 2024. https://doi.org/10.26599/TST.2023.9010028

[28] W. Carlson, R. Hackathorn, and R. Parent, "Computer graphics and animation at the ohio state university," IEEE Computer Graphics and

Applications, vol. 41, no. 3, pp. 8-17, 2021. https://doi.org/10.1109/MCG.2021.3070624

[29] M. Brehmer, B. Lee, P. Isenberg, and E. K. Choe, “A comparative evaluation of animation and small multiples for trend visualization on mobile phones,” IEEE Transactions on Visualization and Computer Graphics, vol. 26, no. 1, pp. 364-374, 2020. https://doi.org/10.1109/TVCG.2019.2934397

[30] Z. Ma, C. Li, X. Liu, H. Wu, and Z. Wen, “Separating shading and reflectance from cartoon illustrations,” IEEE Transactions on Visualization and Computer Graphics, vol. 30, no. 7, pp. 3664-3679, 2024. https://doi.org/10.1109/TVCG.2023.3239364