ITC 3/55
Information Technology and Control
Vol. 55 / No. 3/ 2026
pp. 1509-1528
DOI 10.5755/j01.itc.55.3.44268

Implementation of Synchronized and Coordinated Motion in 3D Facial Animation Based on Weighted Dirichlet Algorithm

Received 2026/02/09 | Accepted after revision 2026/06/02

HOW TO CITE: Zhang, H. (2026) Implementation of Synchronized and Coordinated Motion in 3D Facial Animation Based on Weighted Dirichlet Algorithm. *Information Technology and Control*, 55(3), 1509-1528. https://doi.org/10.5755/j01.itc.55.3.44268

Implementation of Synchronized and Coordinated Motion in 3D Facial Animation Based on Weighted Dirichlet Algorithm

Huidong Zhang
Arts College, Changchun University, Changchun, 130000, China; e-mail: HuidongZhangz@outlook.com

Corresponding author: HuidongZhangz@outlook.com

With the continuous advancement of Virtual Reality (VR) technology, people's requirements for human-computer interaction are becoming increasingly high. To achieve realistic facial expression animation in VR human-computer interaction, a method based on weighted Dirichlet algorithm is proposed to effectively synchronize and coordinate complex facial expressions and mouth shape changes in 3D facial animation. This method achieves natural transitions between different facial expressions and mouth shapes by applying weighted processing to key facial points and combining Dirichlet energy minimization technology. It also combines a mouth shape prediction model based on a Transformer deep learning architecture to achieve speech-driven lip animation synchronization and coordination. The findings denoted that the predicted animation trajectory of the weighted Dirichlet algorithm highly overlapped with the actual motion trajectory on the corners of the mouth and middle of the lips, with maximum errors reduced to 3.0mm and 1.8mm, respectively. In addition, the errors in the lip and nasolabial groove areas were reduced by 46.49% and 54.12%, respectively. In summary, the 3D facial animation synchronized and coordinated motion based on the weighted Dirichlet algorithm effectively achieves precise simulation and coordinated motion of facial animation actions, significantly improving the performance of 3D facial animation.

KEYWORDS: Weighted Dirichlet Algorithm, 3D Facial Animation, Synchronized and Coordinated Motion, Facial Expressions, Virtual Reality

1. Introduction

With the rapid advancement of technologies such as Virtual Reality (VR), Augmented Reality (AR), and digital human-computer interaction, 3D Facial Animation (FA), has become a key indicator for measuring interactive experience in terms of realism and naturalness, serving as the core carrier of human-computer emotional communication. 3D FA has a wide range of applications in film and television production, game development, VR, and human-computer interaction. For example, in scenarios such as virtual anchors, film and television special effects, and remote medical care, 3D faces not only need to accurately reproduce the detailed changes in facial expressions, but also need to achieve synchronous coordination between expressions and speech, head posture, and body movements, to create immersive emotional resonance for users [23]. However, due to the complex anatomical structure of the face, which includes over 60 facial muscles, and the diverse motion patterns of the face, existing 3D FA generation often suffers from inconsistent timing of facial region movements [24]. These factors lead to inconsistent animation and unnatural shaking of expressions that violate anatomical rules. Therefore, how to achieve synchronous and coordinated movement of facial expressions and mouth shapes under dynamic conditions remains an urgent problem to be solved. Traditional methods often rely on a large amount of manual adjustments and complex algorithms, but these methods are difficult to balance efficiency and effectiveness [6]. In recent years, researchers have attempted to optimize facial motion coordination through physical simulation and data-driven methods. However, physical simulation methods are based on muscle dynamics models to construct motion constraints. Although they can simulate muscle activity realistically, they have high computational complexity and are difficult to meet real-time interaction requirements [15]. However, data-driven methods often rely on large-scale 3D facial motion capture data. Although they can ensure local authenticity, they are prone to motion breakage when transferring across identities and scenes [14]. In this context, Partial Differential Equations (PDEs) have gradually become a new direction for solving synchronization and coordination problems due to their efficient ability to describe continuous field distributions. Among them, the Dirichlet algorithm, as a classic elliptic PDE solving method, has gradually attracted the attention of researchers due to its advantages in handling complex geometric shapes and dynamic deformations. Chauhan et al. [30] found that the Dirichlet algorithm can provide a mathematical framework for global coordination of facial motion by constraining the function distribution within the field through boundary conditions [30]. However, the traditional Dirichlet algorithm uses uniform weights for all facial regions, which cannot reflect the motion priority of different parts [21]. Therefore, an innovative 3D FA generation method based on weighted Dirichlet algorithm was designed. This method constructs a regional differentiation weight matrix and sets fixed weights based on anatomical priors of facial motion amplitudes in advance, distinguishing high, medium, and low dynamic regions spatially. The innovation of this study lies in: (1) changing the unified facial weights in the traditional Dirichlet algorithm to differentiated weight allocation based on regional motion priority, and adopting an architecture of weighted Dirichlet geometric deformation and Transformer temporal prediction running in series to achieve phased collaboration between spatial geometric deformation and temporal lip motion prediction. (2) Introducing a cyclic weighting mechanism during the decoding process reduces computational complexity to linear while maintaining the ability to model long sequences. The main contribution of this study lies in: (1) transforming the facial motion coordination problem into a weighted elliptical PDE constraint, breaking through the limitation of traditional Dirichlet methods that use uniform weights for all facial regions. (2) constructing a dynamic weight matrix based on regional motion priority to achieve anatomical differentiation control of high, medium, and low dynamic regions. (3) establishing a cross-modal mapping mechanism from speech features to lip motion coefficients, and use Transformer architecture to decouple long-range temporal dependencies in audio-visual synchronization.

2. Related Works

With the continuous advancement of VR technology, computer 3D FA has gradually received attention from many experts and scholars. Stan et al. [25] designed a new method based on diffusion technique to generate non-deterministic and more realistic speech-driven 3D FAs. This method utilized a pre-trained large speech representation model, Hubert, to encode audio input. The findings denoted that this method could generate FAs [25]. Liu et al. [11] denoted a new method called Emoface to create audio-driven 3D FAs with vivid emotional dynamics. This method used independent speech encoders and emotion encoders to learn audio and map it to a sequence of controller values to generate facial expressions with multiple emotions. The findings denoted that this method was feasible [11]. Fan et al. [5] proposed a joint audio text model to capture contextual information of expressive-speech driven 3D FAs. This method combines acoustic and text modalities to maintain audio stream lip synchronization while integrating real-life facial expressions. The findings denoted that the model performed well compared to existing state-of-the-art methods [5].

Data-driven and deep learning methods have made significant progress in the generation of 3D FAs. Pan et al. [18] introduced a new equipment parameter based emotional dialogue facial dataset and proposed the Latents2Rig model, which helps to convert 2DFA into 3D FA. The results indicate that this method is significantly superior to existing state-of-the-art methods in terms of authenticity and accuracy [18]. Romeo et al. [20] proposed a data-driven facial simulation workflow to enhance the realism of facial performances. This method used an extended position-based dynamic solver to add physical simulations to FAs, and controlled the behavior of the simulations through activation maps. The findings denoted that this method was effective [20]. Wang et al. [29] designed a deep learning-based expression dynamic capture and 3D animation generation method to achieve the goal of generating more vivid FA expressions. The findings denoted that this method could achieve real-time detection of facial regions in video images [29].

The Dirichlet algorithm plays an important role in coordinating facial movements. Ding et al. [3] proposed a new global local interference decoupling method for federated facial expression recognition, and used the Dirichlet algorithm to control the distribution of expression categories among different samples, thereby solving the unique challenge of heterogeneous expression data and heterogeneous interference in federated facial expression recognition [3]. To provide detection performance for micro expressions, Jain Rakesh Kumar et al. introduced a new weighted edge node locally constrained graph attention network and automatically applied Dirichlet to node features and global edge features in the graph attention network to ensure the uniqueness of each class feature [8]. Tsurumaki et al. [28] constructed a probability generation model based on a multi-layer multimodal latent Dirichlet algorithm, and characterized the dynamic correlation between facial expressions and emotional stimuli through a probability generation framework.

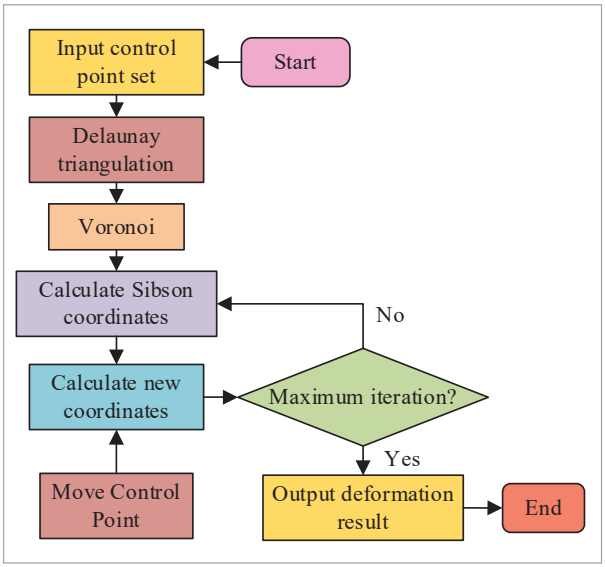
In summary, although many achievements have been made in the technology of 3D FA, current research still has certain limitations in the synchronization and coordination of complex expressions and mouth shape changes. In addition, robust feature representation and distributed privacy protection learning strategies in complex environments have been explored in related perception tasks. To this end, a method based on weighted Dirichlet algorithm is proposed for synchronously coordinating complex expressions and mouth shape changes in 3D FA, combined with Transformer model, to achieve speech-driven lip animation synchronization and coordination.

3. Synchronized and Coordinated Motion of 3D Facial Animation Based on Weighted Dirichlet Algorithm

3.1. 3D Facial Motion Based on Weighted Dirichlet Algorithm

With the rapid advancement of artificial intelligence technology, people's demand for the quality of VR realism is becoming increasingly high, and the generation and control technology of 3D FA has therefore become a current research hotspot [12]. Traditional

Figure 1
3D facial motion deformation process based on Dirichlet algorithm.



3D FA methods often rely on keyframe interpolation or physical model-based methods, but these methods often have certain limitations when dealing with complex facial expression changes and fine motion coordination [19, 26]. The core idea of Dirichlet algorithm, as a deformation method based on Laplace operator, is to construct an energy function about the vertex position and achieve the deformation of the model by solving the minimum value of the energy function [1]. This algorithm has the advantages of high computational efficiency, high degree of freedom deformation, and strong adaptability. Therefore, research is conducted on the motion deformation design of 3D faces based on the Dirichlet algorithm. The process is shown in Figure 1.

In Figure 1, the purpose of the three-dimensional face motion deformation process based on the Dirichlet algorithm is to establish spatial topology and initial geometric weight measurement. Firstly, the control point set is subjected to Delaunay tetrahedralization to construct spatial adjacency relationships. Based on this partition, the dual can be used to construct a three-dimensional Voronoi diagram (consisting of convex polyhedral cells surrounding each control point and their common faces, edges, and vertices). Subsequently, further the Sibson local coordinates are calculated through the intersection of Voronoi cell volumes, which are used to quantify the initial spatial weights of the passive points relative to the nearby control points. The above Delaunay triangulation, Voronoi diagram, and Sibson coordinate calculation provide the geometric weight measurement and local neighborhood information required for deformation, while the final vertex displacement is determined by the subsequent weighted Dirichlet energy minimization as the core solver. The mathematical expression of Voronoi diagram is shown in Equation (1).

$$V = D\left(p, p_i\right) \le D\left(p, p_j\right), \forall i \ne j\,. \tag{1}$$

In Equation (1), V represents the Voronoi diagram of the set of control points, $D(p, p_i)$ represents the distance from point p to point p_i, and $D(p, p_j)$ denotes the distance from point p to point p_j. The Voronoi diagram region contains all p_i points that are closer in distance than any other point p_j. The Voronoi diagram of the set of active control points is shown in Equation (2).

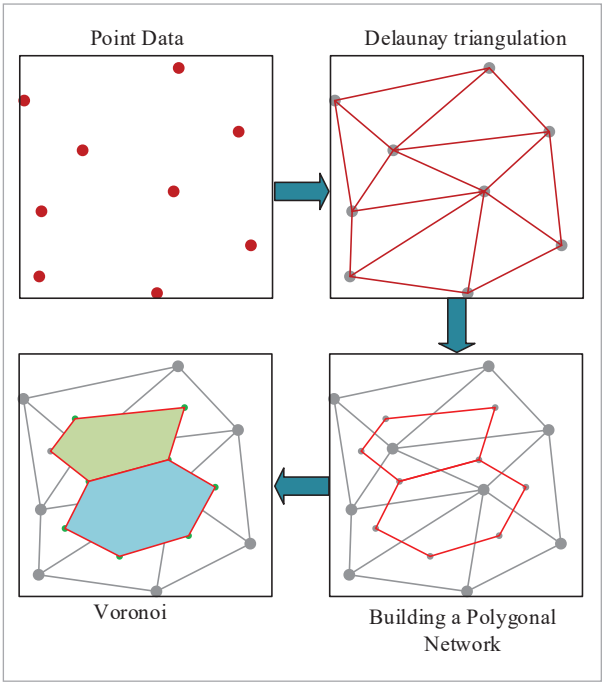
$$V_P = \left\{V_{p_1}, V_{p2}, ..., V_{pn}\right\}. \tag{2}$$

In Equation (2), V_P represents the Voronoi diagram of the active control point set P, and n represents the total number of control points. The schematic diagram of the Voronoi diagram generation process is denoted in Figure 2.

In Figure 2, firstly, during the generation, irregularly distributed point data is used to construct the Delaunay triangulation. Then, the center of the circumcircle of each triangle is determined. Finally, by connecting the centers of the circumcircles of adjacent triangles, a polygon network is formed with each triangle vertex as the generating element, known as the Voronoi diagram. Sibson coordinates, as a coordinate system based on Voronoi diagrams, can more accurately represent the geometric relationship between points and their surrounding environment. The calculation method of Sibson coordinates is denoted in Equation (3).

$$S_i = \frac{V_{p_i} \cap V_q}{V_q}, P \in Q\,. \tag{3}$$

Figure 2
Schematic diagram of Voronoi diagram generation process.



In Equation (3), S_i means the Sibson coordinate value of point p_i to point q, V_q represents the Voronoi diagram of the passive control point q set, and Q represents the set of control points q. Research applies determinant-based methods to calculate the volume of any polyhedron, to determine the relative positions and weights of points within the polyhedron. The determinant expression of a triangle is shown in Equation (4).

$$H(ABC)=\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{vmatrix}. \tag{4}$$

In Equation (4), $H(ABC)$ represents the determinant of a triangle, where $A(x_1,x_2,x_3)$, $B(y_1,y_2,y_3)$, and $C(z_1,z_2,z_3)$ respectively represent the vertices in the three directions of the triangle. To facilitate geometric transformations and calculations, homogeneous coordinates are used when calculating volume, so the rightmost column of the determinant is 1. Thus, the determinant expression of the polygon can be obtained as shown in Equation (5).

$$H(O_1,O_2,...O_m)=\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ x_m & y_m & z_m & 1 \end{vmatrix}. \tag{5}$$

In Equation (5), $H(O_1,O_2,...,O_m)$ means the determinant of the polygon, where $O_m(z_m,z_m,z_m)$ means the coordinates of the m-th vertex. Equation (5) can be regarded as a direct extension of equation (4) from a triangle to any polygon vertex, and its determinant structure remains in the same form as in the case of a triangle. Based on the above determinant expression, for any polyhedron, its volume can be expressed as one sixth of the sum of the positive determinants on each face of the polyhedron [22]. The equation for calculating the volume of polyhedra is shown in Equation (6).

$$V(O_1,O_2,...O_m)=\frac{\sum \vartheta(O_1,O_2,...O_m)}{6}. \tag{6}$$

In Equation (6), $V(O_1,O_2,...,O_m)$ represents the volume of the polyhedron, and $\Sigma\vartheta(O_1,O_2,...,O_m)$ represents the sum of determinants on all faces of the polyhedron. In the traditional Dirichlet algorithm, the influence of active control points on surrounding passive control points is unbalanced [32]. Therefore, to make the deformation effect more flexible and accurate, the study introduces weights to adjust the change values of control points, thereby controlling the deformation strength in different regions. The calculation of weighted Dirichlet energy $E(\Delta V)$ is shown in Equation (7).

$$E(\Delta V)=\frac{1}{2}(L\Delta V)^T W(L\Delta V). \tag{7}$$

In Equation (7), ΔV represents the displacement vectors of all vertices in the grid, L represents the Turapuras matrix, and W represents the diagonal weight matrix. The diagonal elements of W are pre-determined by the anatomical priors of the motion amplitudes of each facial region, and they vary differently in space while remaining constant in time. Minimize the energy and apply a fixed displacement or target position of the active control point as the bound-

ary condition to obtain a linear equation system as shown in Equation (8).

$$P' = \dot{P} + \sum_{i=0}^{n} P_i' \Delta P_i . \tag{8}$$

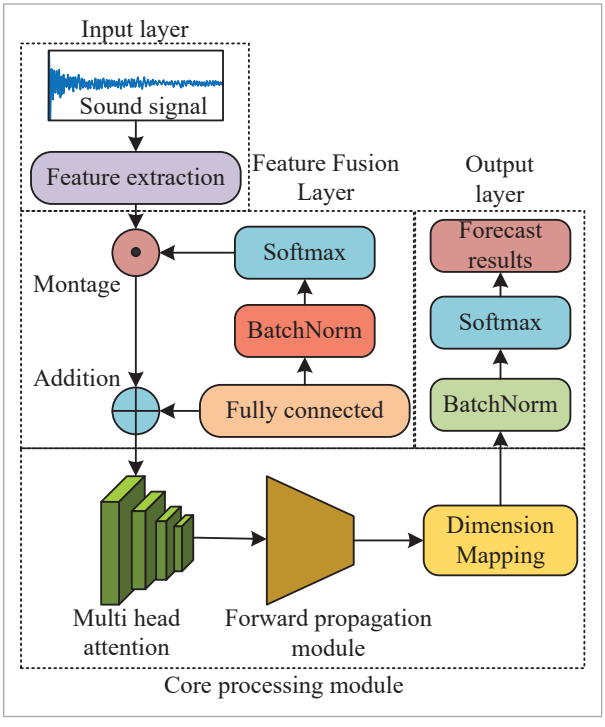
In Equation (8), b represents the right-hand term generated by boundary conditions. The coefficient matrix is symmetric positive definite, so the solution is unique and can be efficiently calculated using sparse solvers such as conjugate gradient method. This weighting mechanism does not change the essence of the Dirichlet method's partial differential equations, only adjusting the metric of the Laplacian operator through the weight matrix W, so that high dynamic regions are given priority in energy minimization. The differentiated weights are applied in a segmented linear approximation method for non-linear deformation areas such as the docking area, meeting the accuracy requirements for moderate deformation.

3.2. Synchronization and Coordination of Facial Animation Based on Transformer Model

The study introduces weight adjustment control point influence through weighted Dirichlet algorithm to achieve more accurate and flexible facial deformation effects. To further enhance the synchronization and coordination of 3D FA, it is necessary to introduce more complex temporal modeling methods to achieve synchronization and coordination of FA. The difficulty of synchronized and coordinated motion in FA modeling lies in the coordination between speech and lip shape. Conventional expressions such as smiling, sad, etc. can be achieved through the deformation of control points such as the eyes, eyebrows, and corners of the mouth [7, 17]. However, different speech pronunciations involve different changes in the mouth, so to achieve synchronous coordination in FA, research needs to design 3d facial coordinated movements driven by speech. Due to the susceptibility of Recurrent Neural Networks (RNNs) to vanishing gradients when processing long sequences, it is difficult to capture temporal dependencies. Due to the sensitivity of RNNs to vanishing gradients when processing long sequences, it is difficult to capture temporal dependencies. Therefore, the Transformer architecture is chosen to generate 3D FAs [31]. The deformation parameters output by the weighted Dirichlet algorithm are used as static spatial geometric constraints and run in series with the lip motion coefficients predicted by the Transformer in the pipeline. The former predefines the deformation weights of each facial region, while the latter independently trains to generate temporal coefficients corresponding to speech. There is no feedback mechanism or shared loss function between the two, and the weights of different regions determine the driving amplitude of the lip motion coefficient. The mouth shape prediction model based on Transformer architecture is shown in Figure 3.

In Figure 3, the model is mainly broken into an input layer, a feature fusion layer, a core processing module, and an output layer. Among them, the input layer receives audio signals as the initial input of the model. The feature fusion layer fuses the input temporal features through concatenation and addition operations to form the initial input features of the model. The core processing module is based on the

Figure 3
Mouth shape prediction model based on Transformer architecture.



Transformer architecture and utilizes multi-head attention mechanism to capture long-range dependencies between features. Subsequently, feature transformation and dimension mapping are performed through the forward propagation module and feature extraction module. In the output layer, batch normalization is performed through the BatchNorm layer to alleviate the distribution offset problem during the training process. Finally, the coefficient prediction layer composed of a fully connected layer and a Softmax activation function outputs the predicted mouth action coefficients related to audio. The mathematical expression for the mouth movement coefficient is shown in Equation (9).

$$C_{1,2,\ldots,a} = E_C\left(F_{1,2,\ldots,a}, C_{1,2,\ldots,a-1}\right). \tag{9}$$

In Equation (9), $C_{1,2,\ldots,a}$ represents the mouth action coefficient corresponding to the audio features of the previous a frame, E_C denotes the function for estimating the mouth action coefficient, and $F_{1,2,\ldots,a}$ represents the audio features of the previous a frame. The mathematical expression of the input feature is shown in Equation (10).

$$G_{C_{1,2,\ldots,a}} = F_{1,2,\ldots,a} + E_a + FC\left(C_{1,2,\ldots,a-1}\right). \tag{10}$$

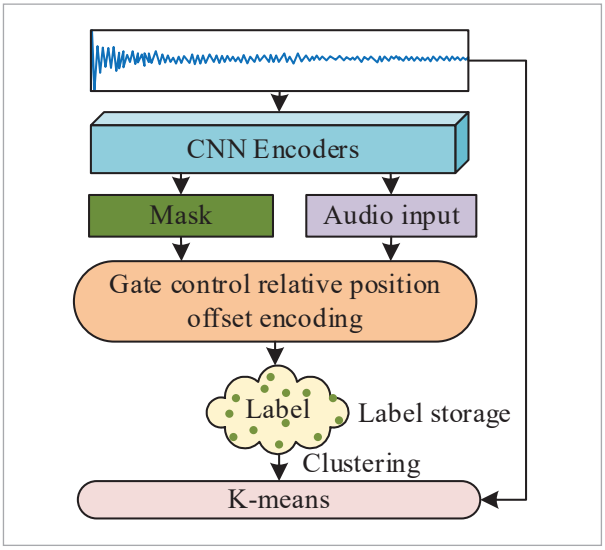
In Equation (10), $G_{1,2,\ldots,a}$ represents the final input feature generated for predicting the mouth motion coefficient in frame a, E_a represents the mouth motion feature embedding in frame a, and $FC(C_{1,2,\ldots,a-1})$ represents the output after feature extraction of the mouth shape coefficient in the previous a–1 frames. The predicted input features of the mouth motion coefficient in the first frame are shown in Equation (11).

$$g_{C_1} = f_1 + E_a. \tag{11}$$

In Equation (11), g_{C_1} represents the input feature used to predict the mouth shape coefficient of the first frame, and f_1 represents the audio feature of the first frame. To better extract features from speech and improve the effectiveness of facial expression synthesis, the latest Transformer-based audio pre-training model, the Wave Language Model (WavLM), is used for self-supervised pre-training

Figure 4

Network architecture based on WavLM.



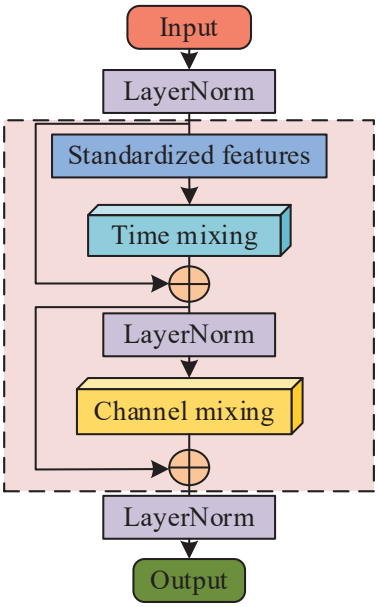
on a large unlabeled speech dataset [2]. The network structure based on WavLM is shown in Figure 4.

In Figure 4, in this network structure, acoustic units are first modeled and corresponding labels are generated through unsupervised K-means clustering. Next, gate relative position offset encoding is introduced, and the input audio is transformed to simulate mixed background sound. Then, the Transformer encoder uses the labels generated from the first iteration to predict the mask region. The feature output after self-attention encoding is shown in Equation (12).

$$h_a = \sum_{b=1}^{T} \omega' v_b. \tag{12}$$

In Equation (12), h_a represents the feature output of the a-th frame after self-attention encoding, ω' represents the attention weight in the self-attention mechanism, v_b represents the input vector of the b-th frame, and T represents the length of the input sequence [16]. To generate realistic and coordinated facial action expression sequences while reducing computational complexity and memory requirements, a deep learning architecture called Recurrent Weighted Key-Value (RWKV) that combines the advantages of RNN and Transformer is proposed for decoding. The network structure based on RWKV is shown in Figure 5.

Figure 5
Network architecture based on RWKV.



In Figure 5, the structure mainly contains a time mixing module and a channel mixing module, which are used for time series dependency modeling and feature channel interaction, respectively. Firstly, the input sequence data is normalized by LayerNorm to enhance the stability of subsequent modules. The standardized features enter the time mixing module, where weighted temporal responses are calculated using the Weighted Key-Value (WKV) unit in the RWKV architecture. While maintaining the ability to model long sequence dependencies, the time complexity is reduced from $O(n^2)$ to $O(n)$. The time mixed features are passed through LayerNorm again and enter the channel mixing module. The channel mixing module achieves information fusion between feature channels through nonlinear transformation and linear recombination of channel dimensions. Finally, the output of the channel mixing module is used as a feature of the current timing step to generate predictions for the final task. The weighted temporal response formula of the time mixing module is shown in Equation (13).

$$r_t = W_r \times \left(\mu_r \alpha_t + \left(1-\mu_r\right)\alpha_{t-1}\right). \tag{13}$$

In Equation (13), r_t represents the feature representation of the current time t in the time mixing module, W_r and μ_r represent the weight matrix and smoothing coefficient in the time mixing module, respectively, and α_t denotes the input feature of the current time t. The feature fusion formula of the channel mixing module is shown in Equation (14).

$$k_t = W_k \times \left(\mu_k \alpha_t + \left(1-\mu_k\right)\alpha_{t-1}\right). \tag{14}$$

In Equation (14), k_t represents the feature representation of the current time t in the channel mixing module, while W_k and μ_k respectively represent the weight matrix and smoothing coefficient in the channel mixing module [27]. The weighting and formula for temporal step features is shown in equation (15).

$$wkv_t = \frac{\sum_{t=1}^{t-1} e^{-(t-1)\psi + k_t} v_t + e^{u+kt}}{\sum_{t=1}^{t-1} e^{-(t-1)\psi + k_t} + e^{u+kt}}. \tag{15}$$

In Equation (15), wkv_t represents the feature representation of the weighted combination of t at the current time, ψ represents the smoothing parameter, and v_t represents the feature value of t at wthe current time [9]. By normalizing and weighting the temporal step features, long sequence dependency modeling can be effectively carried out, improving the realism and coordination of generation, while significantly reducing computational complexity and memory requirements. The model training adopts a multi-task joint loss function. For the lip movement coefficient prediction of the main task, the mean absolute error (MAE) loss is used to supervise the regression of frame-by-frame coefficients. The spatial position of the three-dimensional vertices is constrained by the mean square error loss for geometric consistency, and a temporal smoothness regularization term is introduced to suppress frame-to-frame jitter. To address the imbalance problem of emotion categories, the focal loss is introduced in the auxiliary emotion classification branch that shares the encoded features. By down-weighting the majority class and up-weighting the minority class, the model focuses on the lip shape features under rare emotions, thereby enhancing the prediction robustness in extreme expressions.

4. Synchronized and Coordinated Motion Verification of 3D Facial Animation Based on Weighted Dirichlet Algorithm

4.1. Experimental Environment Setup

To assess the performance of synchronized and coordinated motion in 3D FA based on weighted Dirichlet algorithm, an experimental environment was built in Ubuntu 16.04.7 LTS environment. The experimental setup included a CPU with Intel Core i7-9700, a GPU with NVIDIA GeForce GTX 1080Ti, and 32GB of memory. In addition, the experiment was equipped with a facial expression capture system from Natural Point OptiTrack and a deep learning framework from PyTorch. The hyperparameter configuration for model training is as follows: the AdamW optimizer was used in the study, with an initial learning rate set to 1×10^{-4}, a batch size of 32, and 100 training iterations. In addition, the Dropout rate was set to 0.3, the number of multi-head attention heads was 8, the hidden layer dimension was 512, and both the encoder and decoder layers were set to 6. The datasets used were the publicly available BIWI 3D facial dataset and VOCASET dataset. To further enrich the dataset, the study also utilized a facial expression capture system for data collection at a speed of 100 frames per second. During the data collection process, the study selected 10 healthy subjects aged between 22 and 35, including 5 males and 5 females, all of whom had no history of facial deformities or movement disorders. After marking 37 key landmarks on the subjects' faces, they were asked to complete an action sequence that included 6 basic emotions and common pronunciations in Mandarin Chinese and English. This self-built dataset contained a total of 12000 sample sequences, each with a length of 100 frames. The sample distribution is as follows: in the emotion category, neutral accounts for 25%, joy 20%, sadness 18%, anger 17%, surprise 12%, and fear 8%. In the distribution of phonemes, the Chinese initial consonants were evenly distributed, accounting for approximately 4.76% each. The vowels consisted of 10% monophthongs, 15% diphthongs, and 10% nasal vowels. Vowels accounted for 20% and consonants accounted for 24% of English phonemes. To address the imbalance between minority classes such as fear and surprise and neutral majority classes in the dataset, a triple optimization strategy was adopted. Firstly, the minority class samples were oversampled to augment the representation of underrepresented categories. Second, the majority class samples were undersampled to mitigate the overwhelming dominance of neutral samples while retaining their key distribution characteristics. Finally, focus loss is adopted in the auxiliary emotion classification branch to alleviate the impact of class imbalance. The specific implementation details are as follows: In the stage of minority class oversampling, the study did not adopt a simple sample replication method, but instead used SMOTE synthetic minority class oversampling technology. By calculating the feature similarity between minority class samples, synthetic samples that conform to the distribution law of sample features were generated, achieving a reasonable expansion of the number of minority class samples. In the majority undersampling stage, a hierarchical undersampling method was used to first divide neutral majority class samples into multiple subsets based on dimensions such as inter frame variation amplitude and feature similarity of the action sequence. Then, a certain proportion of samples were randomly selected from each subset to ensure the preservation of the core feature information of the majority class samples and avoid the loss of effective information caused by excessive deletion of samples. In the optimization stage of the loss function, an improved Focal Loss function was introduced. By dynamically adjusting the focus parameter, a lower loss weight was set for neutral majority samples, and a higher loss weight was assigned to minority samples such as fear and surprise. This forced the model to focus on feature learning of minority samples during training, ultimately achieving a balanced contribution of each category sample to the model training, ensuring the recognition performance and generalization ability of the model for different category samples. Finally, after mixing and organizing all the data, they were broken into a training set and a validation set in a 3:7 ratio. The specific experimental environment configuration is shown in Table 1.

Table 1
Experimental environment configuration.

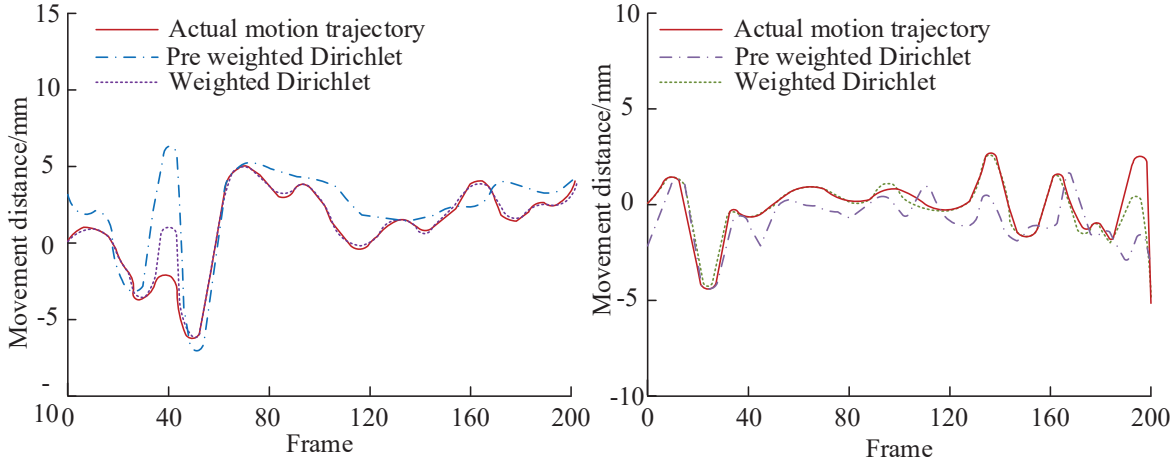
Environment	Configuration
Operating system	Ubuntu 16.04.7 LTS
Processor	Intel Core i7-9700
Graphics card	NVIDIA GeForce GTX 1080Ti
Memory	32GB
Facial expression capture system	Natural Point OptiTrack
Deep Learning Framework	PyTorch

This study used the following indicators to evaluate the animation performance. The marker point distance is the mean of the Euclidean distances between the predicted positions and the actual positions of the facial marker points. The lip vertex error is the root mean square error of the displacements of all vertices in the lip area from the predicted to the actual displacements. The synchronization consistency is quantified by the Pearson correlation coefficient between the audio and the lip movement curve. The synchronization accuracy is defined as the proportion of frames where the time difference between the peak of the lip movement and the peak of the phoneme is less than a threshold (50 ms). The naturalness score is obtained from the subjective ratings (1 to 5 points) given by the participants on the authenticity of the animation. Synchronous and coordinated movement refers to the alignment of the lip movement and the speech in time sequence and the natural coordination of the movements in each facial area.

4.2. Validation of the Deformation Effect of Weighted Dirichlet Algorithm

To study the deformation effect of the weighted Dirichlet algorithm, a comparison was made between the predicted FA trajectories of the Dirichlet algorithm before and after weighting, as shown in Figure 6. From Figure 6(a), the predicted animation trajectory generated by the weighted Dirichlet algorithm highly coincided with the actual motion trajectory on the corner of the mouth. When approaching 40 frames, the error distance was the largest, and the actual motion trajectory was -2.0mm. However, the predicted motion trajectories of the weighted pre - and post Dirichlet algorithms were 6.3mm and 1.0mm, respectively, resulting in errors of 8.3mm and 3.0mm. From Figure 6(b), the weighted Dir-

Figure 6
Comparison of pre - and post weighted Dirichlet algorithms for predicting facial animation trajectory.



(a) The movement trajectory of the corners of the mouth

(b) The movement trajectory of the middle of the lips

ichlet algorithm also exhibited a highly overlapping state on the motion trajectory of the middle lip. When approaching 200 frames, the error reached its maximum value, and the actual trajectory was 2.3mm. The predicted motion trajectories of the weighted pre - and post Dirichlet algorithms were -1.4mm and 0.5mm, respectively, resulting in errors of 3.7mm and 1.8mm. In summary, the weighted Dirichlet algorithm significantly reduces errors in predicting FA trajectories, making the generated animation trajectories more consistent with the actual motion trajectories.

To further assess the deformation effect of the weighted Dirichlet algorithm, it was compared with other advanced algorithms. Comparative algorithms included traditional Dirichlet algorithm, Laplacian deformation algorithm, and As-Rigid-As-Possible Deformation (ARAP) deformation algorithm. The comparison of vertex errors in key facial regions using different deformation algorithms is shown in Table 2. From Table 2, the weighted Dirichlet algorithm had the lowest vertex error in both high and medium dynamic regions. In low-dynamic areas such as cheeks and forehead, the ARAP deformation algorithm had the lowest vertex error. The error of ARAP deformation algorithm in the forehead area was 1.62 ± 0.24, and the error of weighted Dirichlet algorithm was 1.74 ± 0.25, with an increase of only 7.40% in error. For the lip region, the vertex error of the weighted Dirichlet algorithm was 1.83 ± 0.29. Compared with the traditional Dirichlet algorithm, the error was reduced by 46.49%. Compared with the Laplacian deformation and ARAP deformation, the error was reduced by 56.53% and 37.96%, respectively. In the nasolabial groove area, the error of the weighted Dirichlet algorithm was 1.78 ± 0.27, which was 54.12%, 32.06%, and 41.63% lower than the Laplacian deformation, ARAP deformation, and traditional Dirichlet algorithm, respectively. In summary, the weighted Dirichlet algorithm has superior overall performance in 3D FA deformation.

To test the performance of the weighted Dirichlet algorithm in practical applications, a heatmap of the weight distribution of control points was studied and analyzed, as shown in Figure 7. From Figure 7, the weight values of the mouth area were all above 4.0, while the weights of the eyebrow and cheek areas were concentrated around 1.5. In contrast, the weight of low dynamic areas such as the forehead was the lowest, only around 0.5. In summary, the weighted Dirichlet algorithm can flexibly and accurately allocate weights in the design of synchronized and coordinated motion in 3D FA, enhancing the realism and coordination of 3D FA. From a mechanistic perspective, the phenomenon of significantly higher weights in the perioral region was related to the higher muscle distribution density and greater range of motion during pronunciation. Weighted processing essentially set the constraint stiffness coefficients of each region in the Dirichlet energy functional through differentiation, so that high dynamic regions are given priority in the minimization process, thereby improving the accuracy of local deformation. In Figure 7, the weight around the mouth (>4.0) was approximately 8 times that of the low dynamic area of the forehead (~0.5). This difference was consistent with the physiological characteristics of high muscle density and large movement amplitude around the mouth, indicating that the differentiated weight matrix effectively maps anatomical differences in movement priority, making energy allocation more in line with facial physiological driving laws.

Figure 7
Thermogram of weight distribution of control points.

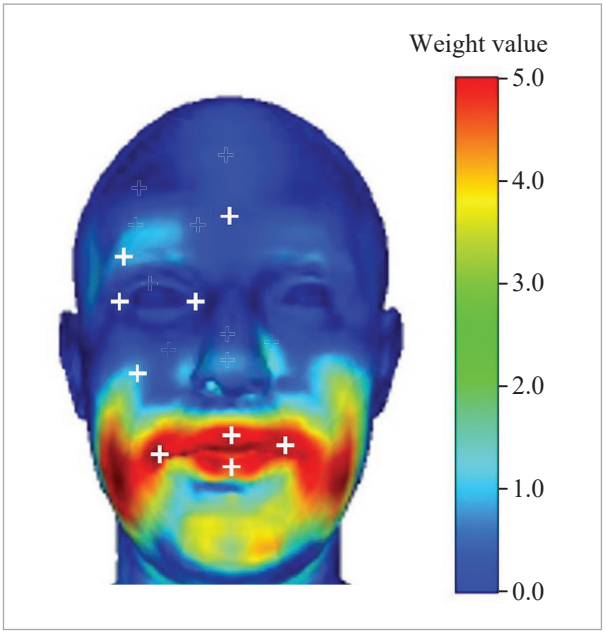


Table 2
Comparison of vertex errors in key facial regions using different deformation algorithms.

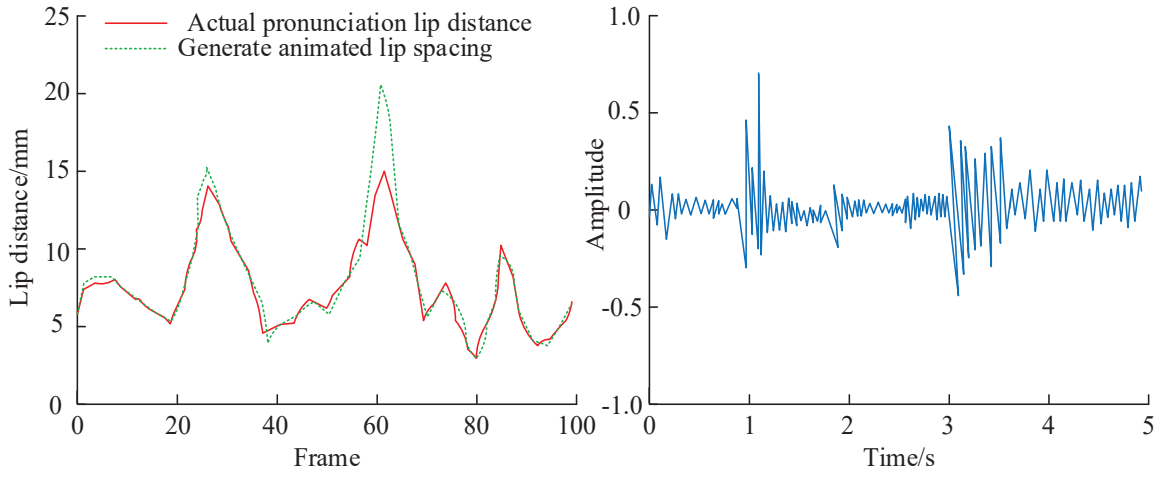
Facial area		Laplacian	ARAP	Traditional Dirichlet	Weighted Dirichlet
High dynamic	Lip	4.21±0.63	2.95±0.44	3.42±0.51	1.83±0.29
	Nasolabial fold	3.88±0.59	2.62±0.39	3.05±0.48	1.78±0.27
Middle dynamic	Eyebrow	2.97±0.42	1.89±0.30	2.15±0.33	1.67±0.24
	Eye Corners	3.12±0.47	1.95±0.31	2.40±0.36	1.71±0.25
Low-dynamic	Cheeks	2.15±0.33	1.75±0.26	1.98±0.28	1.81±0.26
	Forehead	1.82±0.27	1.62±0.24	1.76±0.25	1.74±0.25

4.3. Synchronized and Coordinated Motion Coordination Verification

To verify the synchronization and coordination between FA and speech generated based on the Transformer model, an analysis was carried out through the degree of lip closure, as shown in Figure 8. From Figure 8(a), the actual lip distance changes in pronunciation highly overlapped with the lip distance changes in the generated animation. Although the maximum error occurred at frame 60, with an actual lip distance of 15mm and an animation generated lip distance of 21mm, resulting in an error of 6mm, the overall lip movement trend remained highly consistent with the real data. Figure 8(b) shows the speech waveform generated for this animation segment. Usually, higher syllables require a larger lip distance for pronunciation. The peak trends in Figure 8(a)-(b) are basically consistent, indicating that the mouth movements in the animation are consistent with the actual pronunciation of speech. Taking the 6mm peak deviation around frame 60 in Figure 8(a) as an example, the time propagation characteristics of errors were further studied. In the following 20 frames, the error quickly fell back to below 2.5mm and remained stable, without any trend of error accumulation and amplification along the time axis. In the entire 100 frame sequence, the sliding window mean of lip distance error remained in the range of 1.8-2.3mm, indicating that the self-attention mechanism of Transformer can effectively suppress cross frame error propagation and avoid drift phenomena at the end of long-term sequences. This temporal stability benefited from the explicit modeling ability of the model for long-range dependencies.

To further validate the synchronization performance of FA based on the Transformer model, a comparative analysis was conducted with other advanced models. Other models included 3D Convolutional Neural Network (3D CNN), Bidirectional Long Short Term Memory Network (Bi-LSTM), FaceFormer, and Diffusion based Talking Head (DiffTalk) [4, 10, 13, 33]. Among them, the 3D CNN architecture selected for the study adopted a 3-layer spatiotemporal convolution kernel with a kernel size of 3×3×3, which achieves lip shape prediction by capturing audio and visual cross modal temporal features. The Bi-LSTM model adopted a two-layer bidirectional structure with a hidden layer dimension of 256. It introduced layer normalization and dropout regularization to optimize the cross-frame feature transfer mechanism based on the infrastructure, to meet the real-time requirements of speech-driven applications. The synchronization performance comparison of different speech-driven FA models is shown in Table 3. From Table 3, the Transformer model performed well in all evaluation metrics. Among them, the landmark distance of this model was only 2.58 ± 0.39 mm, which was 38.72%, 35.18%, 10.73%, and 14.57% lower than 3D CNN, Bi-LSTM, FaceFormer, and DiffTalk models, respectively. The lip vertex error of the Transformer model was also lower than all other models, with reductions of 37.74%, 33.97%, 10.07%, and 14.23%, respectively. In terms of synchronized and coordinated motion, the Transformer model scored 0.91 ± 0.02, which was 24.66%, 19.74%, 4.60%, and 7.06% higher than other models, respectively. In summary, FA based on the Transformer model exhibits significant advantages in synchronization performance, accuracy, and consistency.

Figure 8
Lipstick closure degree result.



(a) Comparison between actual lip distance changes and animated lip distance changes

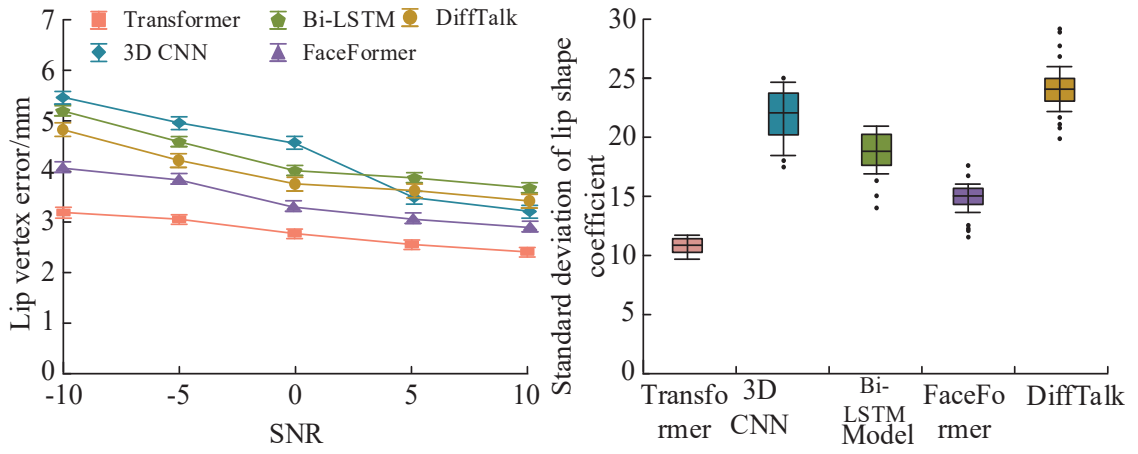
(b) Speech waveform of animation

Table 3
Comparison of synchronization performance of different speech-driven facial animation models.

Model	Distance of marker points/mm	Lip vertex error/mm	synchronized and coordinated motion
3D CNN	4.21±0.62	3.87±0.58	0.73±0.05
Bi-LSTM	3.98±0.59	3.65±0.54	0.76±0.06
FaceFormer	2.89±0.43	2.68±0.39	0.87±0.03
DiffTalk	3.02±0.45	2.81±0.41	0.85±0.04
Transformer	2.58±0.39	2.41±0.35	0.91±0.02

To verify the performance in noisy environments, the stability of mouth shape prediction under different models in noisy environments was compared and studied, as shown in Figure 9. From Figure 9(a), the lip vertex error was maximum at a Signal-to-Noise Ratio (SNR) of -10dB. At this point, the Transformer model had an error of 3.2 ± 0.5mm, while 3D CNN, Bi-LSTM, FaceFormer, and DiffTalk models had errors of 5.6 ± 0.8mm, 5.2 ± 0.7mm, 4.1 ± 0.6mm, and 4.8 ± 0.9mm, respectively. The research model reduced the error by 42.86%, 38.46%, 21.95%, and 33.33%, respectively. When the SNR was 10dB, the lip vertex error was minimized, and the error of the research model was 2.4 ± 0.3mm, while the errors of other models were all above 2.70mm. In the comparison of the standard deviation of lip shape coefficient fluctuations under extreme noise in Figure 9(b), the mean standard deviation of the Transformer model was 0.11, while 3D CNN, Bi-LSTM, FaceFormer, and DiffTalk were 0.22, 0.19, 0.15, and 0.24, respectively. The research model reduced the standard deviation by 50%, 42.11%, 26.67%, and 54.17%, respectively. In summary, the lip prediction performance of the Transformer model is significantly better than other models in noisy interference environments, demonstrating higher stability and accuracy.

Figure 9
Comparison of mouth shape prediction stability under different model noise environments.



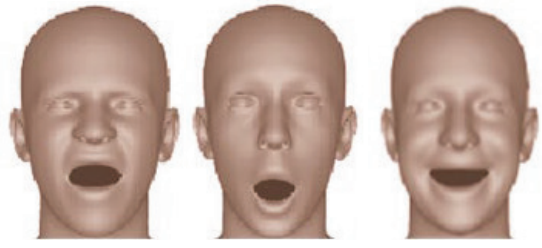
(a) Comparison of lip vertex errors under different noise intensities

(b) Comparison of standard deviation of lip shape coefficient fluctuation under extreme noise

4.4. Comprehensive Performance Verification

To visually verify the performance of the model, FAs were generated under different intonations, as shown in Figure 10. The animation sequentially generated FAs pronounced as "a", "o", and "e". The pronunciation "a" belongs to the open vowel, at which point the lower jaw in the animation dropped significantly and the lips naturally unfold. The pronunciation of 'o' is a rounded vowel sound, and the mouth shape in the animation appears as two lips that are round and protruding forward, with a smaller mouth shape that accurately captured the features of pronunciation. The pronunciation of 'e' is an open lip vowel, and at this time, the mouth shape in the animation was manifested as the corners of the mouth stretching to both sides, and the animated expression was highly consistent with the actual pronunciation of the mouth shape. The research model can accurately capture and reproduce the mouth shape changes during pronunciation when generating FAs with different intonations.

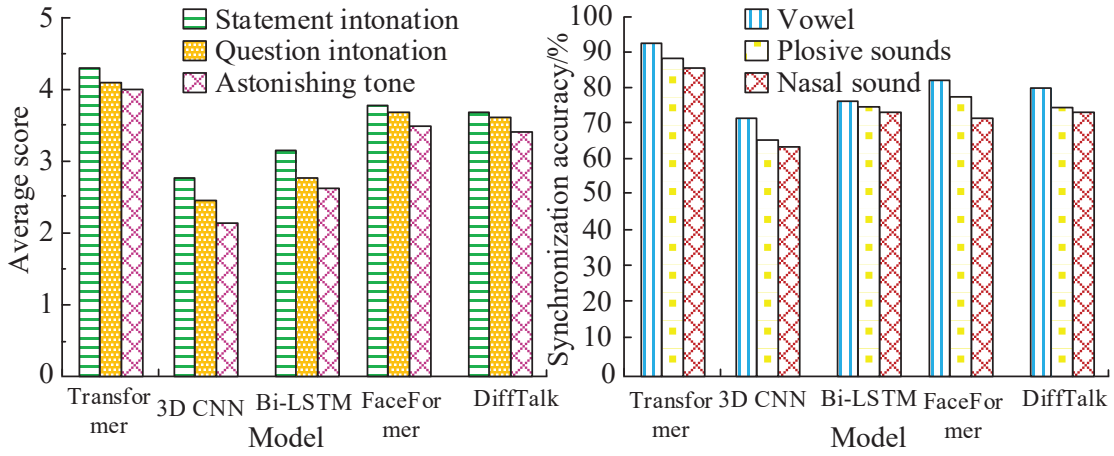
Figure 10
Facial animations with different tones.



(a) Pronunciation a (b) Pronunciation o (c) Pronunciation e

To further verify the coordination of generated FAs, subjective experimental evaluations were conducted. A study was conducted using 10 mixed Chinese English speech segments containing three tones: question, statement, and exclamation. Thirty random participants were selected to assess the naturalness and audio-visual synchronization of the generated FAs. The subjective experimental evaluation results are denoted in Figure 11. From Figure 11(a), the naturalness score of the Transformer model was significantly higher than other models under different intonations. The average rating of this model under declarative tone was 4.3, interrogative tone was 4.1, and exclamation tone was 4.0, while the ratings of other models were all below 4.0. From Figure 11(b), in terms of synchronicity score, the Transformer model had the highest synchronicity accuracy in vowel pronunciation, reaching 92.54%. In contrast, the synchronization accuracies of 3D CNN, Bi-LSTM,

Figure 11
Subjective experimental evaluation results.



(a) Comparison of naturalness ratings for different models

(b) Comparison of synchronicity ratings for different models

FaceFormer, and DiffTalk models were 71.35%, 76.40%, 82.46%, and 80.21%, respectively. Even on nasal sounds with the lowest synchronicity score, the accuracy of the Transformer model reached 85.39%, while the accuracy of other models was below 75%. In summary, the Transformer model exhibits significant advantages in generating FAs in terms of naturalness and audio-visual synchronization.

To assess the comprehensive performance of the model, the efficiency performance of different models was compared and studied, as denoted in Table 4. All performance data of the models were measured uniformly in the same experimental environment (see Table 1 for details) to ensure consistency of the comparison benchmark. From Table 4, the Transformer model only required 18.7 hours of training time, while the training time of other models was all over 20 hours. The single frame prediction delay of the Transformer model was only 4.22ms, which was 27.61%, 54.87%, 66.93%, and 87.68% lower than 3D CNN, Bi-LSTM, FaceFormer, and DiffTalk models, respectively. Although the proposed model integrates multiple modules, its efficiency advantage mainly comes from the linear attention mechanism of the RWKV decoder and the pipeline design of weighted Dirichlet geometric deformation decoupling, rather than relying solely on architecture simplification. In terms of model size, the size of the research model was 52.8MB, only slightly larger than the 48.2MB of the 3D CNN model. But compared with Bi-LSTM, FaceFormer, and DiffTalk models, they reduced by 21.78%, 58.22%, and 83.44%, respectively. In addition, the memory usage of the research model was only 1.9GB, while other models were all above 2GB. In terms of energy consumption, the research model was only 3.6 kWh, which was also greatly better than other models. In addition, comparing the traditional Dirichlet algorithm with the weighted Dirichlet algorithm, it can be found that the weight matrix storage and update introduced by the weighting mechanism does indeed bring slight computational overhead. In a single deformation calculation, the average time consumption of the weighted Dirichlet algorithm was 0.25ms, which was about 39% higher than the traditional method's 0.18ms. The memory usage of the weighted Dirichlet algorithm was 0.78GB, which was higher than the traditional Dirichlet algorithm's 0.72GB. However, this increase was not due to iterative calculations, but rather to a one-time mapping operation of the weight matrix. For the computational cost of the weighted Dirichlet module itself, due to its deformation solution still being reduced to a sparse linear equation system of the same order as the number of

Table 4
Comparison of efficiency and performance of different models.

Model	Training time/h	Single frame prediction delay/ms	Model size/MB	Memory usage/GB	Energy consumption/kWh
3D CNN	23.5	5.83	48.2	2.1	4.7
Bi-LSTM	36.2	9.35	67.5	3.5	7.2
FaceFormer	58.3	12.76	126.4	4.9	11.5
DiffTalk	142.6	34.25	318.7	9.3	29.8
Transformer	18.7	4.22	52.8	1.9	3.6
Traditional Dirichlet	-	0.18	-	0.72	-
Weighted Dirichlet	-	0.25	-	0.78	-

Note: The data for the traditional Dirichlet algorithm and the weighted Dirichlet algorithm are the average time and memory usage per frame when the de formation module runs separately on a facial mesh containing 1200 vertices.

Table 5
Results of ablation experiment.

Model Variant	Distance of marker points/mm	Lip vertex error/mm	Synchronization consistency
Full Model	2.58	2.41	0.91
w/o RWKV	2.99	2.76	0.86
w/o WavLM	2.91	2.70	0.87
w/o Weighted Dirichlet	3.68	3.33	0.83

vertices, introducing a weight matrix only increases the O (n) operations of scalar multiplication per vertex, theoretically without changing the asymptotic complexity of the algorithm.

To quantify the independent contributions of each technical module, this study designed ablation experiments. On the basis of the complete model, the RWKV decoder (replaced with a conventional Transformer decoder), WavLM pre trained features (replaced with MFCC acoustic features), and weighted Dirichlet mechanism (replaced with uniformly weighted Dirichlet) were sequentially removed. The results of the ablation experiment are shown in Table 5. From Table 5, the performance of the complete model was the best, and the degradation was most significant after removing the weighted Dirichlet. The boundary distance and lip vertex error increased by 42.6% and 38.2% respectively, and the synchronization consistency decreased by 8.8%, confirming the core contribution of differentiated region weights to deformation accuracy.

4.5. Error Case Analysis

To gain a deeper understanding of the limitations of the proposed model, a representative misclassification case of extreme emotion expression was selected for detailed analysis, and the root cause of the error was revealed through feature visualization techniques. The comparison of attention weights under extreme anger emotions is shown in Figure 12. From Figure 12, under extreme anger, there was a confusion in the attention weights between high dynamic areas of the face such as the lips and nasolabial folds, and low dynamic areas such as the forehead and cheeks. At this point, the model overly focused on the muscle movement characteristics of the brow area, with a weight value of 3.2, while the attention weight of the key control points around the lips was

Figure 12
Comparison of attention weights under extreme anger emotions.

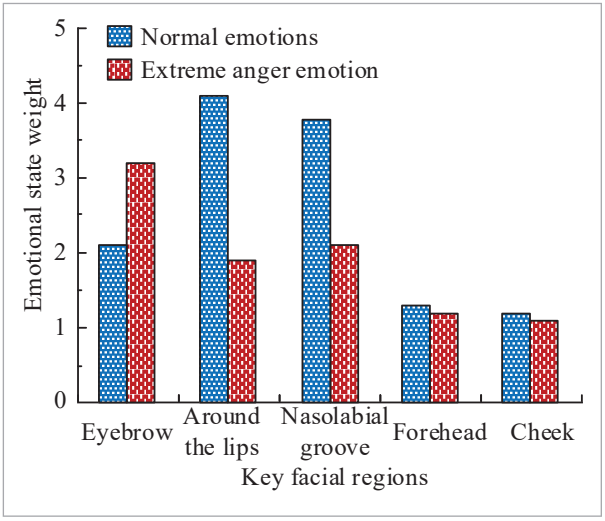


Figure 13
Generate animation results for 6 basic facial expressions.



only 1.9, compared to 4.1 under normal emotions. This is because extreme emotions cause delays in updating the dynamic weight matrix, resulting in insufficient deformation weights in the lip area and ultimately leading to prediction bias. This phenomenon indicates that the model lacks sufficient modeling of the strong coupling relationship between emotions and lip shape, especially in extreme scenarios of multi muscle coordinated movement, and the dynamic adaptability of weight allocation needs to be improved.

To further validate the natural transition effect of the model between different facial expressions, basic expression generation experiments were supplemented in the study. The experiment selected six basic facial expressions, including neutral, joy, sadness, anger, surprise, and fear, from the dataset. Participants were asked to complete a sequence of facial expression switching actions, and real facial expression data was recorded using a facial capture system. Then, the corresponding facial expression animations were generated using our model. The animation results of the six basic facial expressions are shown in Figure 13. From Figure 13, the various expressions generated by the model highly matched the real expressions in key features. For example, in the joyful expression, the animation accurately reproduced the core features of upward corners of the mouth and bulging zygomaticus major muscle. The key movements of the fearful expression, such as widening the eyes and retracting the corners of the mouth, were accurately reproduced by the model. During the transition of facial expressions, the animation generated by the model had no stuttering or abrupt deformation. When switching from neutral to joy, sadness to anger, and other cross expression transitions, facial muscle movements followed anatomical laws and the transition was natural and smooth.

5. Conclusion

With the quick advancement of digital entertainment and VR technology, the application of 3D FA in film production, game development, and human-computer interaction has become increasingly widespread. To improve the authenticity and coordination of facial expressions in FA, a three-dimensional FA synchronization and coordination motion implementation method based on weighted Dirichlet algorithm was proposed. The results indicated that in the verification of deformation effects, the weighted Dirichlet algorithm significantly reduced the error in predicting FA trajectories. On the motion trajectories of the corners of the mouth and the middle of the lips, the predicted animation trajec-

tory of the weighted Dirichlet algorithm coincided with the actual motion trajectory, and the maximum error was reduced to 3.0mm and 1.8mm, respectively. Compared with other deformation algorithms, the weighted Dirichlet algorithm had the lowest vertex error in the high dynamic and medium dynamic regions, and the error in the lip and nasolabial groove regions was reduced by 46.49% and 54.12%, respectively. Through weight distribution thermogram analysis, the weighted Dirichlet algorithm had weight values above 4.0 in the mouth area, while the weights in areas such as eyebrows and cheeks were concentrated around 1.5, and the weights in low dynamic areas such as forehead were the lowest, only around 0.5. This indicates that the weighted Dirichlet algorithm could flexibly and accurately allocate weights, significantly enhancing the realism and coordination of 3D FA. In subjective experimental evaluation, the naturalness score of the animation was above 4.0, with a vowel synchronization accuracy of 92.54% and a nasal synchronization accuracy of 85.39%. In terms of efficiency, the single frame prediction delay of the model was only 4.22ms, and the training time was only 18.7h, demonstrating its high efficiency and practicality in practical applications. In summary, the 3D FA synchronization and coordinated motion method based on weighted Dirichlet algorithm effectively improves the accuracy and realism of FA. However, research has mainly focused on key facial regions and has not fully covered all detailed areas of the entire face. Therefore, future research can further expand the application scope of the algorithm, covering more facial detail areas to strengthen the realism and detail representation of the overall animation.

Acknowledgement

Not received funding.

Data Availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Author's Contribution

H.D.Z. processed the numerical attribute linear programming of communication big data, and the mutual information feature quantity of communication big data numerical attribute was extracted by the cloud extended distributed feature fitting method. H.D.Z. combined with fuzzy C-means clustering and linear regression analysis, the statistical analysis of big data numerical attribute feature information was carried out, and the associated attribute sample set of communication big data numerical attribute cloud grid distribution was constructed. H.D.Z. did the experiments, recorded data, and created manuscripts. All authors read and approved the final manuscript.

References

1. Aimi, A., Dallospedale, S., Desiderio, L., Guardasoni, C. A Space-Time Energetic BIE Method for 3D Elastodynamics: The Dirichlet Case. Computational Mechanics, 2023, 72(5), 885-905. https://doi.org/10.1007/s00466-023-02312-z
2. Chandran, P., Zoss, G., Gross, M., Gotardo, P., Bradley, D. Facial Animation with Disentangled Identity and Motion Using Transformers. Computer Graphics Forum, 2022, 41(8), 267-277. https://doi.org/10.1111/cgf.14641
3. Ding, H., Yan, Y., Lu, Y., Xue, J., Wang, H. Global-Local Disturbance Decoupling for Federated Facial Expression Recognition. IEEE Transactions on Affective Computing, 2026, 17(2), 1985-1998. https://doi.org/10.1109/TAFFC.2026.3666221
4. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T. FaceFormer: Speech-Driven 3D Facial Animation with Transformers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, 18770-18780. https://doi.org/10.1109/CVPR52688.2022.01821
5. Fan, Y., Lin, Z., Saito, J., Wang, W., Komura, T. Joint Audio-Text Model for Expressive Speech-Driven 3D Facial Animation. Proceedings of the ACM on Computer Graphics and Interactive Techniques, 2022, 5(1), 1-15. https://doi.org/10.1145/3522615
6. Gan, W., Xue, J., Lu, K., Yan, Y., Gao, P., Lyu, J. FEA-FA+: An Extended Well-Annotated Dataset for Facial Expression Analysis and 3D Facial Animation. Proceedings of the Fourteenth International Conference

on Digital Image Processing (ICDIP 2022), 2022, 12342(2), 307-316. https://doi.org/10.1117/12.2643588

7. Hu, Z., Ding, Y., Wu, R., Li, L., Zhang, R., Hu, Y., Fan, C. Deep Learning Applications in Games: A Survey from a Data Perspective. Applied Intelligence, 2023, 53(24), 31129-31164. https://doi.org/10.1007/s10489-023-05094-2
8. Jain Rakesh Kumar, A., Bhanu, B. Micro-Expression Classification with Weighted Locally Constrained Graph Network. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2025, 7(4), 695-709. https://doi.org/10.1109/TBIOM.2025.3550430
9. Kadambi, A., de Melo, C., Hsieh, C. J., Srivastava, M., Soatto, S. Incorporating Physics into Data-Driven Computer Vision. Nature Machine Intelligence, 2023, 5(6), 572-580. https://doi.org/10.1038/s42256-023-00662-0
10. Kim, S., Jin, S., Park, J., Kim, K., Kim, J., Nam, J., Kim, S. ModiTalker: Motion-Disentangled Diffusion Model for High-Fidelity Talking Head Generation. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(4), 4302-4310. https://doi.org/10.1609/aaai.v39i4.32452
11. Liu, C., Lin, Q., Zeng, Z., Pan, Y. EmoFace: Audio-Driven Emotional 3D Face Animation. 2024 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2024, 387-397. https://doi.org/10.1109/VR58804.2024.00060
12. Liu, J., Hui, B., Li, K., Liu, Y., Lai, Y. K., Zhang, Y., Yang, J. Geometry-Guided Dense Perspective Network for Speech-Driven Facial Animation. IEEE Transactions on Visualization and Computer Graphics, 2022, 28(12), 4873-4886. https://doi.org/10.1109/TVCG.2021.3107669
13. Maddu, R., Murugappan, S. Students' Engagement Prediction in Online Learning Context via Face Emotion and Data Features with Improved LinkNet and Bi-LSTM Architecture. Journal of Ambient Intelligence and Humanized Computing, 2025, 16(8), 993-1009. https://doi.org/10.1007/s12652-025-04997-x
14. Moser, L., Chien, C., Williams, M., Serra, J., Hendler, D., Roble, D. Semi-Supervised Video-Driven Facial Animation Transfer for Production. ACM Transactions on Graphics, 2021, 40(6), 1-18. https://doi.org/10.1145/3478513.3480515
15. Mourot, L., Hoyet, L., Le Clerc, F., Schnitzler, F., Hellier, P. A Survey on Deep Learning for Skeleton-Based Human Animation. Computer Graphics Forum, 2022, 41(1), 122-157. https://doi.org/10.1111/cgf.14426
16. Nguyen, D. P., Ho Ba Tho, M. C., Dao, T. T. Enhanced Facial Expression Recognition Using 3D Point Sets and Geometric Deep Learning. Medical & Biological Engineering & Computing, 2021, 59(6), 1235-1244. https://doi.org/10.1007/s11517-021-02383-1
17. Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G. E., Neff, M. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. Computer Graphics Forum, 2023, 42(2), 569-596. https://doi.org/10.1111/cgf.14776
18. Pan, Y., Liu, C., Xu, S., Tan, S., Yang, J. VASA-Rig: Audio-Driven 3D Facial Animation with 'Live' Mood Dynamics in Virtual Reality. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(5), 2416-2425. https://doi.org/10.1109/TVCG.2025.3549168
19. Pang, L., Zeng, Z., Li, Y., Zhang, G., Zhang, S. Speech-2Face3D: A Two-Stage Transfer-Learning Framework for Speech-Driven 3D Facial Animation. IET Image Processing, 2025, 19(1), 70155-70159. https://doi.org/10.1049/ipr2.70155
20. Romeo, M., Schvartzman, S. C. Data-Driven Facial Simulation. Computer Graphics Forum, 2020, 39(6), 513-526. https://doi.org/10.1111/cgf.14089
21. Sharma, S., Kumar, V. 3D Face Reconstruction in Deep Learning Era: A Survey. Archives of Computational Methods in Engineering, 2022, 29(5), 3475-3507. https://doi.org/10.1007/s11831-021-09705-4
22. Sharma, S., Kumar, V. Voxel-Based 3D Face Reconstruction and Its Application to Face Recognition Using Sequential Deep Learning. Multimedia Tools and Applications, 2020, 79(25), 17303-17330. https://doi.org/10.1007/s11042-020-08688-x
23. Song, W., Wang, X., Jiang, Y., Li, S., Hao, A., Hou, X., Qin, H. Expressive 3D Facial Animation Generation Based on Local-to-Global Latent Diffusion. IEEE Transactions on Visualization and Computer Graphics, 2024, 30(11), 7397-7407. https://doi.org/10.1109/TVCG.2024.3456213
24. Song, W., Wang, X., Zheng, S., Li, S., Hao, A., Hou, X. TalkingStyle: Personalized Speech-Driven 3D Facial Animation with Style Preservation. IEEE Transactions on Visualization and Computer Graphics, 2025, 31(9), 4682-4694. https://doi.org/10.1109/TVCG.2024.3409568

25. Stan, S., Haque, K. I., Yumak, Z. FaceDiffuser: Speech-Driven 3D Facial Animation Synthesis Using Diffusion. Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games, 2023, 13(1), 1-11. https://doi.org/10.1145/3623264.3624447

26. Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y., Yu, M., Liu, Y. DiffPoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models. ACM Transactions on Graphics, 2024, 43(4), 1-9. https://doi.org/10.1145/3658221

27. Tian, S., Zheng, M., Liang, X. TransFusion: A Practical and Effective Transformer-Based Diffusion Model for 3D Human Motion Prediction. IEEE Robotics and Automation Letters, 2024, 9(7), 6232-6239. https://doi.org/10.1109/LRA.2024.3401116

28. Tsurumaki, K., Hieida, C., Miyazawa, K. Study of Emotion Concept Formation by Integrating Vision, Physiology, and Word Information Using Multilayered Multimodal Latent Dirichlet Allocation. IEEE Transactions on Affective Computing, 2025, 16(4), 2975-2987. https://doi.org/10.1109/TAFFC.2025.3585882

29. Wang, B., Shi, Y. Expression Dynamic Capture and 3D Animation Generation Method Based on Deep Learning. Neural Computing and Applications, 2023, 35(12), 8797-8808. https://doi.org/10.1007/s00521-022-07644-0

30. Wenting, Z., Mandan, L., Baoxu, Z., Lizhi, X. 3D Facial Expression Animation Generation Based on Interactive Evolutionary Algorithm. Journal of East China University of Science and Technology, 2025, 51(4), 538-551. DOI: 10.14135/j.cnki.1006-3080.20240831001.

31. Yang, Y., Guo, Y., Xiong, J., Liu, Y., Pan, H., Wang, P., Tong, X., Guo, B. Swin3D: A Pretrained Transformer Backbone for 3D Indoor Scene Understanding. Computational Visual Media, 2025, 11(1), 83-101. https://doi.org/10.26599/CVM.2025.9450383

32. Zhang, L., Zeng, C., Zhang, Q., Lin, H., Cao, R., Yang, W., Xu, L., Yu, J. Video-Driven Neural Physically-Based Facial Asset for Production. ACM Transactions on Graphics, 2022, 41(6), 1-16. https://doi.org/10.1145/3550454.3555445

33. Zhao, J., Mao, X., Zhang, J. Learning Deep Facial Expression Features from Image and Optical Flow Sequences Using 3D CNN. The Visual Computer, 2018, 34(10), 1461-1475. https://doi.org/10.1007/s00371-018-1477-y

This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (http://creativecommons.org/licenses/by/4.0/).