International Journal of Advanced Engineering Research and Science (IJAERS)
Peer-Reviewed Journal
ISSN: 2349-6495(P) | 2456-1908(O)
Vol-12, Issue-5; May, 2025
Journal Home Page Available: https://ijaers.com/
Article DOI: https://dx.doi.org/10.22161/ijaers.125.3

DreamTalk-DMT: A Lightweight Sparse Mechanism Model with Dynamic Thresholds

Jia Zhang[1,*], Lin Po Shang[2]

[1]Department of Electronic Information Engineering, Electronic Information Science and Technology major, Guangdong University of Petrochemical Technology, China
[2]Department of Energy and Power Engineering, Process Equipment and Control Engineering major, Guangdong University of Petrochemical Technology, China
*Corresponding author:19578022107@163.com

Received: 11 Apr 2025,

Receive in revised form: 09 May 2025,

Accepted: 15 May 2025,

Available online: 20 May 2025

***Keywords*— Diffusion model; Dynamic threshold sparsification; Mutual information constrained optimization; Decoupled decoder; Cross-modal feature fusion module**

***Abstract*— Aiming at the shortcomings of the DreamTalk 2D digital human synthesis model in computational efficiency and expression generation fineness, this paper proposes an optimization method combining adaptive sparsity and cross-modal feature enhancement. By introducing a dynamic threshold sparsity mechanism into the diffusion model, the sparsity ratio was dynamically adjusted based on the learnable threshold and Exponential Moving Average (EMA), and the Mutual information Constraint (MI Constraint) was combined to minimize the information loss, which reduced the calculation amount of the model while retaining key features. The model architecture is improved, and the decoupled decoder is designed to decompose the facial expression into the upper and lower regions for independent processing. The dynamic linear layer is combined to realize parameter adaptation under the style condition, and the detail expression of expression generation is improved. In addition, Tacotron speech features and Wav2Vec acoustic features are fused to enhance the synchronization of speech and expression, and skip connections are used to optimize the information transmission efficiency.*

I. INTRODUCTION

From the perspective of technology evolution, digital human synthesis technology has experienced a significant transformation from traditional methods based on physical models to data-driven deep learning methods. Initially, DaViT regress 3DMM parameters from the input image to roughly scout the shape and texture of the face. Although 3DMM provides valuable information, its linear nature limits its realism.[1] Subsequently, an innovative approach developed by Buhari et al.[2] combined graph theory and FACS to extract useful features (68 landmark points) that can distinguish between various microexpressions.[3] The development of deep learning technology, methods based on Generative Adversarial networks (GAN) have made breakthroughs in the field of image generation, such as the StarGAN-VC model, which has attracted people's attention because it can solve this problem using only a single generator. However, there is still a gap between real and converted speech.[4] Diffusion Model has aroused a new upsurge of research in the field of digital human synthesis due to its theoretical completeness and generation quality advantages. Among them, DreamTalk model, as the landmark achievement in the field of speech-driven expression synthesis, is an audio-driven framework based on two-stage diffusion, which uses emotional conditional diffusion model and lip refinement network[5] to improve facial emotional expression while maintaining high video quality. DREAM-Talk represents a major leap forward in the field of emotional conversational face generation, enabling the creation of realistic and emotionally engaged digital human representations in a wide range of applications.[5]

Many scientific research institutions and enterprises continue to make efforts in digital generation technology and other related fields. In the direction of expression generation, VASA-1, a diffusion-based global facial dynamics and head motion generation model proposed by Microsoft Research Asia, can not only generate lip movements perfectly synchronized with audio, but also generate a large number of facial nuances and natural head movements, providing high video quality through realistic facial and head dynamics IC. Online generation of 512×512 videos at up to 40 FPS with negligible startup latency is also supported.[6]; OTAvatar[7] proposed by Ma et al., OTAvatar invert the portrait image into a motion-free identity code, and then use the identity code and motion code to modulate an efficient CNN to generate a three-plane formula volume. Finally, the image is generated by volume rendering, and the identity and motion in the latent code are decoupled by a novel anti-phase decoupling strategy. The face image is constructed based on generalized controllable three-plane rendering. In addition, the Make-A-Video model[8] launched by Meta AI tries to model the multi-modal generation of text-speech-image in a unified way. Although it shows strong potential in creative content generation, there are still technical bottlenecks in the accurate synchronization of voice and expression.

At present, in the aspect of film and television special effects, the application of digital human is more and more widely, and the fidelity of image and motion has been improved. The continuous expansion of digital human application scenarios to strong interaction fields such as real-time broadcast, virtual idol interaction, and intelligent education, the limitations of existing technologies have become increasingly prominent. Aided by the diffusion model mechanism, the DreamTalk model represents a major leap forward in the field of emotional talk face generation, enabling the creation of realistic and emotionally engaging digital human representations in a wide range of applications[9]. However, with the expansion of application scenarios and the improvement of requirements, its defects gradually appear. In terms of computational efficiency, the model parameters are dense, and in real-time interaction scenarios, the memory footprint is high and the reasoning time is long, which seriously affects the interaction fluency. For expression generation, it is difficult for a single decoder to accurately simulate the differentiated motion of the eyebrow, mouth and other regions, and synthesize expression detail distortion. In cross-modal fusion, the simple feature concatenation method cannot deeply explore the complex relationship between speech prosody and expression dynamics, and the matching degree of expression and speech emotion is not good.

To address the above technical challenges, this study proposes a dynamic threshold sparsification and decoupling generation framework based on information theory and dynamic system theory. By introducing the learnable sparse threshold and Exponential Moving Average (EMA) mechanism[10], combined with the mutual information loss function[11], the framework reduced the floating-point operation efficiency while ensuring that the key information was not lost. The decoupled decoder was designed, the facial expression space was divided into the upper and lower halves, and the dynamic linear layer was used to realize the adaptive adjustment of parameters to improve the naturalness of expression. The gated fusion module of Tacotron acoustic features and Wav2Vec speech representation is constructed, and the gradient transfer path is optimized by combining jump connection, which greatly improves the accuracy of speech-expression synchronization, and provides an innovative solution for the practical development of digital human technology.

II. DREAMTALK

In the field of speech-driven expression synthesis of digital human, DreamTalk model uses the diffusion mechanism[5], uses the Transformer-based EmoDiff network, and performs temporal denoising learning of 3D expression under the conditions of audio, portrait and emotional style, and realizes the end-to-end generation of speech to expression. Excellent results are achieved on the VoxCeleb dataset, which alleviates the mode collapse problem of traditional GAN. The diffusion mechanism adopted by the method is derived from the denoising Diffusion Probability Model (DDPM)[12], which is based on the Markov chain[13]. The data generation is realized through the process of adding Gaussian noise forward and reverse iterative denoising. Compared with traditional generative models, diffusion models have more solid theoretical foundation and stronger conditional generation ability, and have shown significant advantages in the field of multimodal generation, which provides important technical support for models such as DreamTalk. denoising diffusion probabilistic model (DDPM) is a class of generative models based on probabilistic diffusion process. In recent years, remarkable progress has been made in the field of deep learning and generative models. The core idea of diffusion model is to treat the process of data generation as a random process that gradually changes from simple distribution (e.g., Gaussian distribution) to complex data distribution. Diffusion models usually include two processes: the Forward Diffusion Process and the Reverse Process. However, both of them are a parameterized Markov chain in nature, which has stationary property. That is, if a probability changes with time, it will tend to a stationary distribution under the action of the Markov chain, and the longer the time, the more stable the distribution will be. It was this stationarity that allowed him to gradually restore the image, given a neural network that predicted the noise.

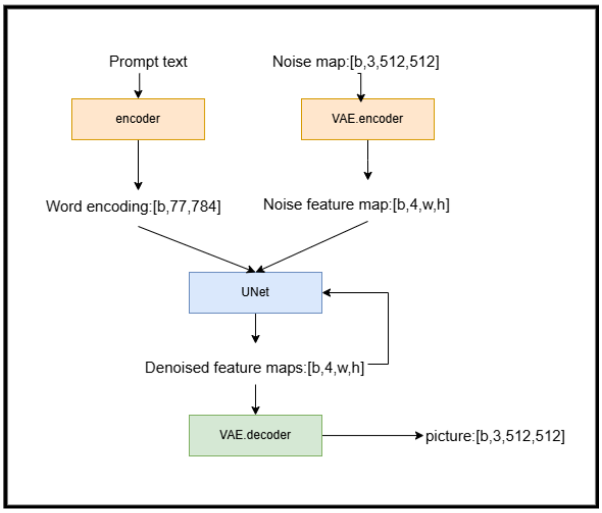

Fig.1 Diffusion Model generation process

The Forward Diffusion Process is a process that continuously adds noise to the data to be trained. The process usually starts from a simple distribution (e.g., Gaussian distribution, etc.), and through multiple rounds of small cardinality noise, the image data to be trained is closer to a complex data distribution. Meanwhile, at each step, the model predicts the noise at the next step based on the current data state and noise level, thus gradually pushing the data into a high-dimensional and complex distribution space.

In the forward process, given the initial data distribution x_0~q(x), the noise with standard deviation β_t is gradually added to the initial data according to the schedule to obtain the noise data.

$$q(x_t \mid x_{t-1}) = N(x_i; \sqrt{1-\beta_t} x_{t-1}, \beta_t I) \quad (1)$$

Where t represents the final time, as t continues to increase, the noise data gradually approaches the Gaussian distribution.

However, the efficiency of stepwise iteration based on Equation (1) is very low, and the training process consumes a lot of time. To improve the efficiency of computing, introducing the α_t=1-β_t, $\overline{\alpha_t} = \prod_{s=0}^{t} \alpha_s$, type (1) can be converted to:

$$q(x_t \mid x_0) = N\left(x_i; \sqrt{\overline{\alpha}_t} x_{01}, (1-\overline{\alpha}_t) I\right) \quad (2)$$

The noisy data x_t at any time t can be obtained。

Reverse Diffusion Process is a process that gradually recovers useful information from noisy data.

The goal is to gradually recover the distribution of the original data from the pure noise state (the final result of the forward diffusion process). It is the opposite of the forward diffusion process and tries to learn how to remove the noise added at each time step so as to recover the original data.

The backward diffusion process takes advantage of the fact that the way noise is added in the forward diffusion process is known, and gradually restores the noisy data to the original data by training a neural network to predict how much noise should be subtracted at each step.In the backward diffusion process, the neural network is constructed to fit $p_\theta(x_{t-1}|x_t)$, and the original data is gradually recovered from the noise, which can be expressed as follows.

$$p_\theta(x_{t-1}|x_t) = \mathrm{N}(x_{t-1}; \mu_\theta(x_t, \mathrm{t}), \Sigma_\theta(x_t, \mathrm{t})) \quad (3)$$

Where θ is the neural network parameter, $\mu_\theta(x_t, \mathrm{t})$and $\Sigma_\theta(x_t, \mathrm{t})$ are the mean and variance, respectively.

The training process in the diffusion model is achieved by optimizing the variational lower bound of the negative log-likelihood with p_θ. To simplify the training process, the variance of the model is set to a constant and the coefficients of the loss function are removed, so the loss function is:

$$L(\theta) = \mathbb{E}_{t, x_t, \epsilon 1}\left[|| \epsilon - \epsilon_\theta (x_t, t) ||^2 \right] \quad (4)$$

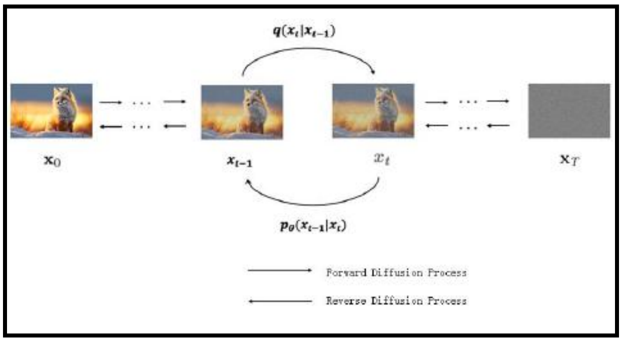

Fig.2 Diffusion Process

III. IMPROVED MODEL ARCHITECTURE AND KEY TECHNOLOGIES

Aiming at the technical bottlenecks of DreamTalk model in terms of computational efficiency, expression generation accuracy and cross-modal fusion, this study proposes a dynamic threshold sparsifation-decoupling generation framework (DTS-DG). The framework realized systematic optimization through four core modules. At the level of efficiency optimization, the dynamic sparse threshold and EMA dynamic adjustment mechanism were used, and the mutual information loss function was combined to reduce the amount of calculation while ensuring the loss of information. In the cross-modal fusion dimension, the gated fusion module of Tacotron[14] and Wav2Vec features[15] is constructed, supplemented by skip connection to optimize the gradient transfer path and enhance the depth correlation between speech and expression. In the aspect of expression generation, the upper and lower half decoupling decoder is designed, and the parameters are adaptively adjusted by the dynamic linear layer, which significantly improves the accuracy of expression detail description and emotion synchronization, and provides a new solution for speech-driven digital human synthesis technology.

Through the four-layer optimization system, the improved model achieves a significant improvement in computational efficiency and generalization ability while maintaining the naturalness of speech synthesis, which provides a new technical path for the lightweight of end-to-end speech synthesis models.

3.1 ynamic threshold sparsification mechanism

When the original dreamtalk model deals with high-dimensional features, there are problems such as large consumption of computing resources and slow inference speed. A large number of redundant parameters not only increase the computational burden, but also may lead to overfitting. In view of this, this study introduces a dynamic threshold sparsification mechanism, which dynamically screens features based on a dynamic sparse mask. By setting a learnable threshold, the feature dimensions that contribute less to the model are automatically identified and eliminated.

The computational efficiency of the improved model is significantly improved, and the number of parameters and reasoning time are reduced compared with the original model, which effectively alleviates the bottleneck of computing resources. At the same time, the generalization ability of the model is enhanced because the redundant information interference is reduced. In addition, the dynamic threshold sparsification mechanism ensures that the model can still maintain a high level of performance while being lightweight by retaining key features, which facilitates the deployment in practical applications.

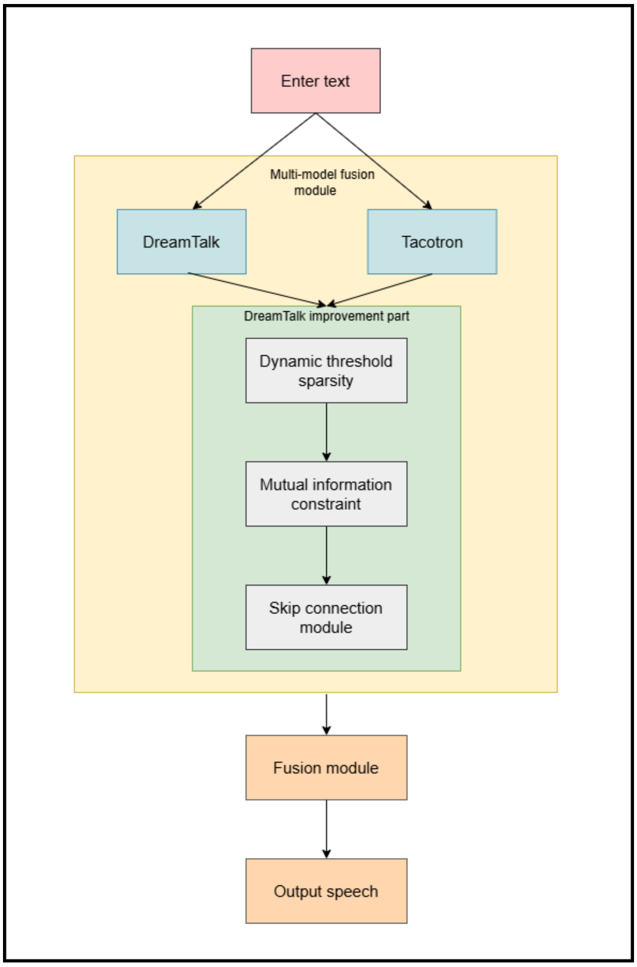
Enter text
Multi-model fusion module
DreamTalk
Tacotron
DreamTalk improvement part
Dynamic threshold sparsity
Mutual information constraint
Skip connection module
Fusion module
Output speech

Fig.3 Model framework

Under the key requirements of model computational efficiency optimization, the dynamic threshold sparsification mechanism becomes one of the core innovations of this research. The mechanism aims to solve the problem that the traditional fixed sparsity method cannot adapt to the dynamic changes of features in the process of model training. By introducing a learnable threshold and combining with the Exponential Moving Average (EMA) technology, the dynamic adjustment of the sparsity ratio of model parameters is realized, and the calculation amount is reduced while the key information is retained to the maximum extent, ensuring that the model performance is not significantly affected.

In the training process of the diffusion model, the data characteristics show a complex change trend with the advancement of time steps. To effectively capture these changes and adjust the sparsification strategy accordingly, we design a dynamic threshold calculation method based on learnable threshold and EMA. First, we define a learnable threshold parameter θ, which is optimized through backpropagation during model training. To map the values of θ to a reasonable range, we use the sigmoid function [16] to convert it to θ', i.e.

$$\theta' = \frac{1}{1+e^{-\theta}} \quad (5)$$

The value range of θ' is limited to the interval of (0,1), which enables the threshold to be adjusted in a reasonable dynamic range.

At the same time, in order to track the dynamic changes of features, we introduce EMA[10] to calculate the mean value μ_t of the absolute values of features. EMA is a commonly used time series smoothing technique, which is able to dynamically update statistics based on historical information and current data. In this study, μ_t is calculated as follows.

$$\mu_t = \alpha\mu_{t-1} + (1-\alpha)\mathbb{E}(|x_t|) \quad (6)$$

α is the smoothing coefficient of EMA, which is usually set to a value close to 1, and α=0.9 was taken in this study. This means that the calculation of μ_t is more dependent on the historical mean μ_{t-1}, but at the same time, it is also adjusted according to the expectation $\mathbb{E}(|x_t|)$ of the absolute value of the feature at the current time. In this way, μ_t can better reflect the overall trend of the absolute value of the feature, and it is somewhat robust to sudden outliers.

Based on the computed θ' and μ_t, we generate the dynamic threshold $\theta'\cdot\mu_t$ and construct the sparse mask M_t accordingly. For each element $x_t[i]$ in the feature vector x_t, the element $M_t[i]$ of the sparse mask M_t is generated according to the following rules:

$$M_t[i] = \begin{cases} 1, & if\ |x_t[i]| > \theta' \cdot \mu_t \\ 0, & otherwise \end{cases} \quad (7)$$

When $|x_t[i]|$ is greater than the dynamic threshold, the value of $M_t[i]$ is 1, and the corresponding element is retained in the sparsification process. Otherwise, $M_t[i]$ is 0, and the corresponding element is set to zero, thus sparsifying the feature vector x_t. This dynamic threshold setting allows the sparsification process to be dynamically adjusted according to the importance and distribution of features. The key features that have larger absolute values and contribute more to the model output are more likely to be retained; However, the relatively unimportant features are sparsified to reduce the amount of calculation.

During backpropagation, to ensure that the sparsified model can still learn effectively, we only perform gradient updates on the corresponding parameters with a value of 1 in the sparse mask M_t. This not only ensures that the model can continue to be optimized in the case of parameter compression, but also avoids the invalid calculation of the sparsified (zeroed) parameters, which further improves the computational efficiency.

Through the above dynamic threshold sparsifying mechanism, the model can dynamically adjust the sparsity ratio of the parameters during the training process, and flexibly balance the computational efficiency and model performance under different training stages and data feature distributions. This mechanism not only effectively reduces the computational burden of the model and improves the inference speed, but also ensures the accuracy and stability of the model in tasks such as expression generation by retaining key information. In practical applications, this mechanism enables the model to maintain good performance under limited computing resources when dealing with large-scale data and complex tasks.

3.2 Mutual information constrained optimization mechanism

In the process of dynamic threshold sparsification, the original dreamtalk model is easy to cause the loss of feature information, which affects the model's ability to capture key semantic and emotional information, and leads to the decline of the accuracy and integrity of the generated results.

In order to solve this problem, based on Mutual Information[17] and Kullback-Leibler Divergence theory[18] in information theory, this study constructs a mutual information constrained optimization mechanism. Mutual information was proposed by Shannon in 1948 to quantify the dependence between two random variables. KL divergence was defined by Kullback and Leibler in 1951 as a measure of how different two probability distributions are.

The basic definition of KL divergence is as follows.

$$D_{KL}(p||q) = \sum_x \quad p(x)\log\frac{p(x)}{q(x)} \quad (8)$$

Here, p(x) and q(x) represent two probability distributions, and the formula measures the difference between p(x) and q(x) by calculating the weighted sum of log ratios over all values x.

The basic definition of mutual information is based on joint distribution and marginal distribution, which is expressed as follows.

$$I(X;Y) = D_{KL}(p(X,Y)||p(X)p(Y)) \quad (9)$$

That is, the mutual information is equal to the KL divergence between the joint probability distribution p(X,Y) and the product p(X)p(Y) of the marginal probability distributions, which reflects the amount of information shared between two random variables X and Y.

In this study, the original feature distribution is denoted as $q(x_t)$, and the feature distribution under the action of sparse mask Mt is denoted as $p(x_t|M_t)$. Based on the above theory, the mutual information loss function is constructed as follows.

$$\mathcal{L}_{\mathrm{MI}} = \mathbb{E}[\mathrm{KL}(p(x_t|M_t) \parallel q(x_t))] = \mathbb{E}[\textstyle\sum_{i=1}^{D} \quad p(i)\log \frac{p(i)}{q(i)}]$$

(10)

This formula quantifies the information loss during dynamic threshold sparsification by calculating the KL divergence of the feature distribution before and after sparsification. In the actual calculation, because it is difficult to estimate the probability distribution directly, the feature mean and variance are used to approximate the distribution. In the training process, $\mathcal{L}_{\mathrm{MI}}$ is incorporated into the total loss function, and the model parameters and sparse threshold are optimized through back propagation, which effectively retains key information while reducing the amount of calculation and maintaining the performance of the model.

After introducing this mechanism, the model performs well in information retention, the retention rate of key features is improved, and the performance degradation caused by information loss is effectively avoided. At the same time, the mutual information constraint optimization mechanism makes the model more accurately balance the computational efficiency and information retention in the sparsification process, which provides a guarantee for the stable training and efficient operation of the model.

3.3 (Multi-model Fusion network) Cross-modal feature fusion module

The original dreamtalk model has the problems of insufficient synchronization and insufficient feature fusion when processing speech and facial expression features, which leads to the inability to accurately match the generated facial expression and speech, and poor expression naturalness and dynamic correlation. The cross-modal feature fusion module constructed in this study strengthens the dynamic association between speech and expression by deeply fusing Tacotron speech features and Wav2Vec acoustic features.

After the introduction of this module, the model achieves a significant improvement in speech-expression synchronization, and the time deviation between lip movements and speech phonemes is reduced, which greatly improves the phenomenon of phonetic and painting synchronization. In addition, the cross-modal feature fusion module effectively enhances the network's ability to express multimodal information through the gate mechanism and skip connection[19], so that the model can better capture the complex mapping relationship between speech and expression.

On the basis of computational efficiency optimization, this study constructs a cross-modal feature fusion module to solve the problem of speech and expression synchronization. The Tacotron model is used to extract the 512-dimensional speech feature $f_{\mathtt{taco}}$ containing prosodic and semantic information, while the Wav2Vec model is used to obtain the 1024-dimensional feature $f_{\mathtt{w2v}}$ focusing on acoustic details, providing multi-dimensional speech information for fusion.

The module adopts the gating mechanism to realize feature fusion, and learns from the idea that the LSTM gating unit[20] controls the information flow through the Sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$). Firstly, the two features are concatenated and linear transformed, and then the gating signal g is generated by the Sigmoid function: $g = \sigma(\mathtt{Linear}([f_{\mathtt{w2v}};f_{\mathtt{taco}}]))$. Based on this, the fusion feature $f_{\mathtt{fusion}} = g \cdot f_{\mathtt{w2v}} + (1-g) \cdot f_{\mathtt{taco}}$ is obtained by weighted summation, so that the model can adaptively adjust the feature weight according to the speech characteristics. In addition, the jump connection $f_{\mathtt{final}} = f_{\mathtt{fusion}} + f_{\mathtt{input}}$ of ResNet is introduced to ensure the effective transmission of key information, improve the expression ability of the network, and realize the deep correlation between speech features and expression generation.

3.4 Decouple the decoder

The original dreamtalk model uses a single decoder to process facial expression generation, which is difficult to accurately model the movement of different facial regions. It is easy to interfere with emotional expression and mouth movement, resulting in unnatural local expressions and loss of details. In this study, based on the decoupled generation network designed by FACS theory[21], the facial expression space is divided into the upper and lower half regions, which are processed independently and modeled by the dual-branch structure respectively.

The introduction of decoupled decoder effectively solves the defects of the original model. In the generation of eye expressions, the movements of eyebrows and eyelids are more consistent with emotional semantics, and the emotional expression is more accurate. In terms of mouth movement generation, the synchronization between mouth shape and speech is further enhanced, and the speech-expression synchronization error is reduced. At the same time, the structure avoids the interference between the actions of different regions, which greatly improves the naturalness and accuracy of local expressions. The generated facial expressions are more vivid and realistic, and have more advantages in detail processing.

Considering that a single decoder is difficult to accurately simulate the movement of different facial regions, this study designs a decoupled generation network based on FACS theory, and divides the facial expression space into the upper and lower half regions for independent processing. The upper half is responsible for emotional expression, while the lower half is closely related to speech articulation.

The decoupled decoder adopts a dual-branch structure, each branch is equipped with a dynamic linear layer, and its design refers to the idea of conditional normalization. The eyebrow decoder uses $W_{\mathtt{eye - brow}} = \mathtt{Softmax}(\frac{\mathrm{MLP}(s_{\mathtt{emo}})}{T}) \cdot \mathbf{W}_{\mathtt{shared - emo}}$ to generate a weight matrix based on s_{emo}. According to the acoustic features $f_{\mathtt{acoustic}}$, the mouth decoder uses $W_{\mathtt{mouth}} = \mathtt{Softmax}(\frac{\mathrm{MLP}(f_{\mathtt{acoustic}})}{T}) \cdot \mathbf{W}_{\mathtt{shared - mouth}}$

Determine the parameters, including $\mathtt{Softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} \quad e^{x_j}}$. Finally, the output of the upper and lower halves is concatenated to avoid the mutual interference between emotional expression and mouth movement, realize the accurate control of eye emotional transmission and mouth speech synchronization, and significantly improve the naturalness and detail accuracy of expression generation.

IV. EXPERIMENTAL ANALYSIS

4.1 Experimental Environment and experimental data set

In this study, an end-to-end training approach is used to jointly optimize modules such as speech feature extraction, cross-modal feature fusion, and expression generation. In the early stage of training, the parameters of pre-trained models such as Tacotron and Wav2Vec are fine-tuned with a small learning rate to adapt them to the speech feature extraction task of this study. Then, the cross-modal feature fusion module was gradually introduced to decouple the decoder, and the alternating training strategy was adopted. The parameters of the expression generation network were fixed, and the feature fusion module was optimized to enhance the correlation between speech and expression features. Then the feature fusion module is fixed, and the decoupled decoder is trained to improve the quality of expression generation. In the training process, the Early Stopping method is used to avoid overfitting, and the training rounds are dynamically adjusted according to the expression naturalness index on the validation set.

The dataset used in this experiment is VoxCeleb. The VoxCeleb dataset is an open source dataset maintained by the Visual Geometry Group at the University of Oxford. The dataset is derived from speech clips in YouTube videos related to celebrities. It is split into VoxCeleb1, which has more than 100,000 voice clips of 1,251 celebrities, and VoxCeleb2, which is much larger, with more than 1 million voice clips of 6,112 celebrities and each clip is at least 3 seconds. It is characterized by a high diversity of speech, including different races, accents, ages, and complex backgrounds, while being of high quality and carefully screened. It has a wide range of applications in speech recognition, speaker verification, speech sentiment analysis, speech synthesis and other fields, which provides rich and high-quality data resources for speech-related research and application.

The experimental platform environment configuration used in this experiment is shown in *Table 1*

Table 1 Experimental environment

Name	version informatio
Operating system	Microsoft Windows11
CPU	12th Gen Intel(R) Core(TM) i7-12700
GPU	NVIDIA GeForce RTX 4060 Ti
Memory capacity	16GB
Deep Learning FrameworkPython	PyTorch 3.10.14
CUDA	11.8
PyTorch	2.1.2
TorchVision	0.16.2

4.2 Comparative analysis of data

We use a variety of evaluation metrics to evaluate the experimental results, and the experimental results are shown in Table 3, which show the experimental results of the four methods SadTalker, Wav2lip, TANGO and Ours, respectively. In this paper, SSIM, SIFT and PSNR are selected as the performance evaluation metrics, which measure the quality of the 2D digital human video generated based on the diffusion model from different perspectives, thus providing a comprehensive evaluation of the performance of the method.

SSIM is a full-reference image quality assessment index, which measures the similarity of images from three aspects: brightness, contrast and structure. SSIM values range from [0,1], with higher values indicating lower image distortion. Therefore, for the similarity curve of video frames, a higher SSIM value is better, and a flatter curve is better, because it means that the similarity between video frames does not change much and the video quality is stable.

Table 2 Comparison of experimental results of different methods

Methods	SSIM↑	LPIPS↓	PSNR↑
DMT	0.7970	0.1093	28.2298
DreamTalk	0.6973	0.4582	20.3429
SadTalker	0.6693	0.5348	12.8915
Wav2lip	0.8470	0.1277	34.6643
TANGO	0.8758	0.1359	29.0019

LPIPS is a deep learning-based image similarity evaluation metric, which evaluates image similarity by comparing perceptual differences between image patches. The smaller the LPIPS value, the more similar the images. For the similarity curve of video frames, a lower LPIPS value is better, and a flatter curve is better, which indicates that the perceptual difference between video frames is small and the video quality is high.

PSNR is a commonly used metric to evaluate video and image quality, which is calculated by comparing the original signal with the distorted signal. A higher PSNR value indicates less distortion of the video frame. For the similarity curve of video frames, the higher the PSNR value, the better, the upward of the curve indicates that the video quality is improving, and the downward of the curve indicates that the video quality is decreasing.

The experimental results show that the model proposed in this study outperforms the previous methods in many aspects. The cross-modal feature fusion module realized the deep fusion of speech features through the gate mechanism and skip connection, which significantly improved the synchronization. The decoupled decoder separated the upper and lower half of the facial motion based on FACS theory, and combined with the dynamic linear layer to enhance the expression detail generation ability. The dynamic threshold sparsification and mutual information constrained optimization mechanism greatly reduce the computational complexity under the premise of controllable information loss. When mutual information constrained optimization is disabled, the inference time of the model decreases but the performance index deteriorates significantly. These results prove that the collaborative design of model components is the key to achieve efficient and natural expression generation.

V. CONCLUSION

In this study, the DreamTalk speech synthesis model is optimized, and the performance of the model is significantly improved by introducing techniques such as adaptive threshold sparsification method, mutual information constraint, multi-model fusion and skip connection. In terms of speech synthesis quality, computational efficiency and generalization ability, the improved model is significantly better than the traditional DreamTalk model and other comparison models.

However, there are still some shortcomings in this study. In the process of multi-model fusion, the current simple average fusion method essentially treats the output of each model with equal weight, which fails to fully consider the differences in the advantages of different models in processing specific speech features or scenes, and it is difficult to maximize the effectiveness of each model in complex speech synthesis tasks. In the field of cross-modal applications, although the speech-image matching has been improved, there is still a large room for improvement in the quality and diversity of image generation. There is a gap between the generated image and the real image and user expectation in detail texture, color richness and creative expression. When the adaptive sparszization method faces extreme data distribution, such as a small number of abnormal speech samples or a serious imbalance of data feature distribution, the stability of the model will be affected, and problems such as fluctuations in the quality of synthesized speech and abnormal parameter update may occur.

To address these shortcomings, future research will be carried out in several directions. In the aspect of multi-model fusion, the fusion strategy based on attention mechanism and dynamic weight allocation will be deeply explored. By constructing an intelligent evaluation system, the model can automatically allocate the weight of each sub-model according to the characteristics of the input speech, and give full play to the advantages of different models. In the field of cross-modal research, we plan to combine generative adversarial networks and self-supervised learning technology to further explore the potential correlation between speech and image, build a more powerful cross-modal mapping model, improve the quality and diversity of image generation, and realize more creative and realistic image generation driven by speech. For the adaptive sparsification method, a dynamic adjustment threshold strategy and an abnormal data detection mechanism are introduced. By real-time monitoring of data distribution characteristics, the sparsification process is dynamically optimized, and the stability of the model in extreme data environments is enhanced, so as to further improve the overall performance and application range of the model, which provides more powerful support for the development of speech synthesis technology and cross-modal research.

ACKNOWLEDGEMENTS

This paper received funding from "Guangdong University of Petrochemical Technology College Student Innovation and Entrepreneurship Project No. 24A014".

REFERENCES

[1] Priyadharshini, A. R., & Annamalai, R. (2024). Identification and Reconstruction of Human Faces into 3D Models Using SSD-Based and Attention Mesh Models in Real-Time. *SN Computer Science*, *5*(8), 1-9.

[2] Buhari, A. M., Ooi, C. P., Baskaran, V. M., Phan, R. C., Wong, K., & Tan, W. H. (2020). FACS-based graph features for real-time micro-expression recognition. *Journal of Imaging*, *6*(12), 130.

[3] Chauhan, A., & Jain, S. (2024). FMeAR: FACS Driven Ensemble Model for Micro-Expression Action Unit Recognition. *SN Computer Science*, *5*(5), 598.

[4] Kaneko, T., Kameoka, H., Tanaka, K., & Hojo, N. (2019). Stargan-vc2: Rethinking conditional methods for stargan-based voice conversion. *arXiv preprint arXiv:1907.12279*.

[5] Zhang, C., Wang, C., Zhang, J., Xu, H., Song, G., Xie, Y., ... & Feng, J. (2023). Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*.

[6] Xu, S., Chen, G., Guo, Y. X., Yang, J., Li, C., Zang, Z., ... & Guo, B. (2024). Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, *37*, 660-684.

[7] Ma, Z., Zhu, X., Qi, G. J., Lei, Z., & Zhang, L. (2023). Otavatar: One-shot talking face avatar with controllable tri-plane rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 16901-16910).

[8] Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., ... & Taigman, Y. (2022). Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

[9] Ma, Y., Zhang, S., Wang, J., Wang, X., Zhang, Y., & Deng, Z. (2023). DreamTalk: When Emotional Talking Head Generation Meets Diffusion Probabilistic Models. *arXiv preprint arXiv:2312.09767*.

[10] Chen, R., Pang, K., Wang, Z., Liu, Q., Tang, C., Chang, Y., & Huang, M. (2025). A self-supervised graph convolutional model for recommendation with exponential moving average. *Neural Computing and Applications*, 1-17.

[11] Zhao, S., Wang, Y., Yang, Z., & Cai, D. (2019). Region mutual information loss for semantic segmentation. *Advances in Neural Information Processing Systems*, *32*.

[12] Nair, N. G., Mei, K., & Patel, V. M. (2023). At-ddpm: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 3434-3443).

[13] Hou, J., Lu, Y., Wang, M., Ouyang, W., Yang, Y., Zou, F., ... & Liu, Z. (2024). A Markov Chain approach for video-based virtual try-on with denoising diffusion generative adversarial network. *Knowledge-Based Systems*, *300*, 112233.

[14] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

[15] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

[16] Zhao, Y., & Li, X. (2024, September). Better approximation of sigmoid function for privacy-preserving neural networks. In *Journal of Physics: Conference Series* (Vol. 2852, No. 1, p. 012007). IOP Publishing.

[17] A Mutual Information Based Approach for Feature Subset Selection and Image ClassificationA Mutual Information Based Approach for Feature Subset Selection and Image Classification.

[18] Xu, H., Li, Y., Zhang, M., & Tong, P. (2024). Sonar image segmentation using a multi-spatial information constraint fuzzy C-means clustering algorithm based on KL divergence. *International Journal of Machine Learning and Cybernetics*, 1-18.

[19] Chen, H., Lu, X., Li, S., & He, L. (2025). Improving aluminum surface defect super-resolution with diffusion models and skip connections. *Materials Today Communications*, *42*, 111297.

[20] Wang, W., Han, D., Duan, X., Yong, Y., Wu, Z., Ma, X., ... & Dai, K. (2024). Fast-Activated Minimal Gated Unit: Lightweight Processing and Feature Recognition for Multiple Mechanical Impact Signals. *Sensors*, *24*(16), 5245.

[21] Gilbert, M., Demarchi, S., & Urdapilleta, I. (2021). FACSHuman, a software program for creating experimental material by modeling 3D facial expressions. *Behavior Research Methods*, *53*(5), 2252-2272.