Speaking Digital Person Video Generation Methods Review Report Talking Head

Qinghua Yu[a]
Elite Engineers College (Innovation Entrepreneurship College), Dongguan University of Technology, Dongguan, Guangdong, China

Keywords: Speak Digital Person, Video Generation, Generate Confrontation Model, Diffusion Model, Neural Radiation Field.

Abstract: In recent years, the rapid development of deep learning technology has significantly promoted the progress of virtual digital human technology, especially in the field of speaking digital human video generation has made a significant breakthrough. The research on this technology has shown great potential and value in many application scenarios such as video translation, film production and virtual assistant. This paper systematically summarizes and summarizes the main methods and research progress of voice-driven speech, and discusses the key technologies, data set construction and evaluation strategies. At the key technical level, advanced artificial intelligence technologies such as Generative Adversarial Network (GAN), Diffusion Model (DM) and Neural Radiance Field (NeRF) play a central role. At the same time, the size and diversity of the dataset have a decisive impact on the effect of the model training, while the optimization of the evaluation strategy contributes to a more objective and comprehensive measurement of the quality of the generated results. Although the technology has made significant progress, there are still many challenges and opportunities. In the future, this field is expected to further promote technological development through continuous innovation and breakthrough, and bring more convenience and value to human society.

1 INTRODUCTION

With the rapid progress of artificial intelligence technology, the Digital Human Generation (DHG) video generation and its application have gradually become the focus of academic attention. As a kind of virtual entity built by computer technology, digital human not only simulates human behavior and interaction in the virtual environment but also shows a wide range of application potential in the real world. Voice-driven speech digital human is a frontier branch of virtual digital human research, the core of which is to use advanced artificial intelligence technology to build virtual characters that can express speech content naturally (Song et al., 2023； Zhen et al., 2023; chen et al., 2020). Specifically, through the input audio information, combined with the image or video clips containing the characteristics of the target person, the digital speaker, through the steps of information extraction, semantic expansion, data fusion and alignment, generates a video of the target person naturally. The core challenge of this technology is the fusion and presentation of multimodal data, which aims to visually display the speech content of the target person through the visual form. In the generation process of speech, digital speaker, generation methods such as Generative Adversarial Network (GAN), Diffusion Model (DM) and Neural Radiance Field (NeRF) play a key role.

Owing to its distinctive adversarial training architecture, the Generative Adversarial Network (GAN) has attained remarkable accomplishments within the domain of digital human generation. Notably, it has demonstrated significant prowess in enhancing the fidelity and diversity of generated images. In the context of digital human creation, the implementation of GAN not only substantially elevates the visual authenticity of virtual images but also imbues them with a rich repertoire of facial expressions and the ability to exhibit diverse postural variations. As GAN technology continues to undergo progressive evolution, it holds the promise of yielding

[a] https://orcid.org/0009-0003-6062-0519

Yu, Q.
Speaking Digital Person Video Generation Methods Review Report Talking Head.
DOI: 10.5220/0013700200004670

In *Proceedings of the 2nd International Conference on Data Science and Engineering (ICDSE 2025)*, pages 502-508
ISBN: 978-989-758-765-8

even more exquisitely detailed and vivid digital human images, thereby facilitating a more seamless and natural human - computer interaction experience.

The diffusion model (DM), as a new generation method, has opened up a new research direction for the field of digital human generation by simulating the diffusion and reverse diffusion process to generate data. The model can generate high-quality and diversified digital human images, with strong robustness and controllability. In terms of digital human customization, the diffusion model can realize the precise regulation of the virtual image characteristics and meet the personalized needs of users. In addition, its application potential in the generation of digital human animation also provides more possibilities for the dynamic performance of virtual characters.

The NeRF method encodes and enders the three-dimensional scene through the neural network, which provides a three-dimensional technical path for digital human generation. This method can realize the high precision reconstruction and rendering of digital human 3D scenes, so as to generate a virtual image with real three-dimensional sense. In the field of Virtual Reality (VR) and Augmented Reality (AR), the neural radiation field NeRF method enables digital people to better integrate into the virtual environment and provide users with an immersive interactive experience. At the same time, its application in the fashion industry such as virtual fitting and virtual makeup also provides a new idea for business model innovation.

This paper focuses on the technology of speaking digital human generation and mainly conducts systematic research on key technologies and related data sets. First, the paper discusses the latest progress in the Talking Head generation method from the perspective of GAN, DM and NeRF. Then, the current research results of relevant data sets in the Talking Head field are summarized. By systematically combing the existing research, this paper aims to provide theoretical support for the further development of this field and provide a valuable reference for technology selection and optimization in practical application.

2 GENERATIVE APPROACH

The mainstream generation methods of Talking Head mainly include generation methods based on GAN, diffusion model and neural radiation field. Each of these three methods has its own advantages and disadvantages, as shown in Table 1.

Table 1: Comprehensive comparison of voice-driven speech.

model	scene	superiority	inferior strength or position
GAN	General people	The picture quality is high	Training is unstable
DM	General people	Rich in details	High demand for computing resources
NERF	persona certa	A strong sense of reality	High demand for computing resources

2.1 Methods Based on Generative Adversarial Models

The Generative Adversarial Network (GAN) is composed of a generator and a discriminator. The generator is designed to produce realistic samples, while the discriminator's function is to differentiate between the generated samples and the genuine real - world samples. The generator undertakes the task of learning the distribution of real - world data, with the aim of generating samples that closely resemble the real data. Conversely, the discriminator assesses the authenticity of the samples presented to it. Throughout the training phase, the generator and the discriminator are optimized in an alternating manner. The generator endeavors to deceive the discriminator by generating samples that are increasingly difficult to distinguish from real ones. Simultaneously, the discriminator refines its discriminatory capabilities to better discern the true nature of the samples. This iterative process continues until an equilibrium is reached between the two components. The detailed operational process is graphically depicted in Figure 1.

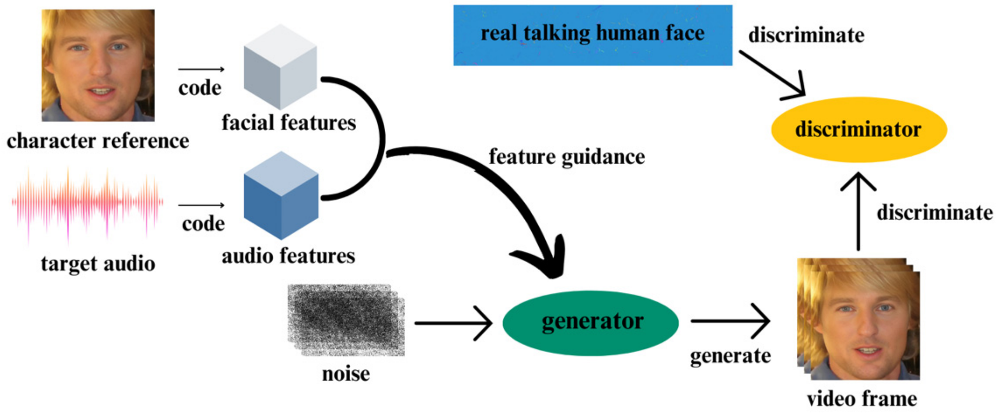

Figure 1: Voice-driven speech digital person based on the generative adversarial model. (Picture credit: Original)

In the literature (Zakharov et al., 2019), a system endowed with few - sample capabilities is put forward. This system conducts extensive meta - learning on a vast video dataset. It formulates the few - sample and single - sample learning of hitherto unseen human neural speech head models as adversarial training problems, leveraging high - capacity generators and discriminators. The system has the ability to initialize the parameters of the generator and the discriminator in a manner specific to each individual. Moreover, the training process can be efficiently accomplished relying solely on a limited number of images. By virtue of this approach, highly realistic and personalized models of novel characters, and even portrait talking heads, can be learned.

Literature (Tang et al., 2018) uses deep convolution generative adversarial network (DCGAN) to introduce Convolutional neural networks (CNN) for unsupervised training based on traditional generating adversarial network, and adds conditional extension to conditional model. Combined with deep convolution generated against the network and conditions against the advantages of the network, establish Conditional-DCGAN (C-DCGAN), using the convolution neural network powerful feature extraction ability, on the basis of the auxiliary samples, the structure is optimized and used for image recognition, the experimental results show that the method can effectively improve the image recognition accuracy.

GAN models have powerful generation capabilities, generating high-quality, realistic speaker face images and videos. Through training, you can learn various features of the face, including expressions, posture and lip shape, to generate face images that match the audio content. It also has the flexibility to generate different styles of speaker face images based on the input audio content and face features, and can also cope with complex face scenes, including different angles, different gestures and different expressions.

However, the training of GAN is prone to problems such as mode collapse, that is, the generated images lack diversity and only contain a few modes. The GAN model may overfit the training data during training, and may not be well adapted to the unseen audio content and face features.

2.2 Speech-driven Digital Speaker Based on the Diffusion Model

The diffusion model is mainly based on two core processes: the Forward Diffusion Process (FDP) and the Reverse Diffusion Process (RDP). The forward diffusion process is a parameterized Markov chain that gradually adds noise to the raw data until the data eventually becomes pure noise. This process can be seen as a process of gradually "blurring" or "destroying" the data, so that the data gradually loses its original characteristics. The reverse diffusion process is the inverse process of the forward process, which starts from pure noise, gradually removes noise and recovers the original data. This process is implemented by a deep learning model (usually a convolutional neural network) which is how the original image is progressively recovered from a noisy image. The specific process is shown in Figure 2. The ability of the neural network to predict the noise is trained by the process of adding the speaker video frame, and the corresponding digital human video frame is generated through the process of gradually denoising the noise.

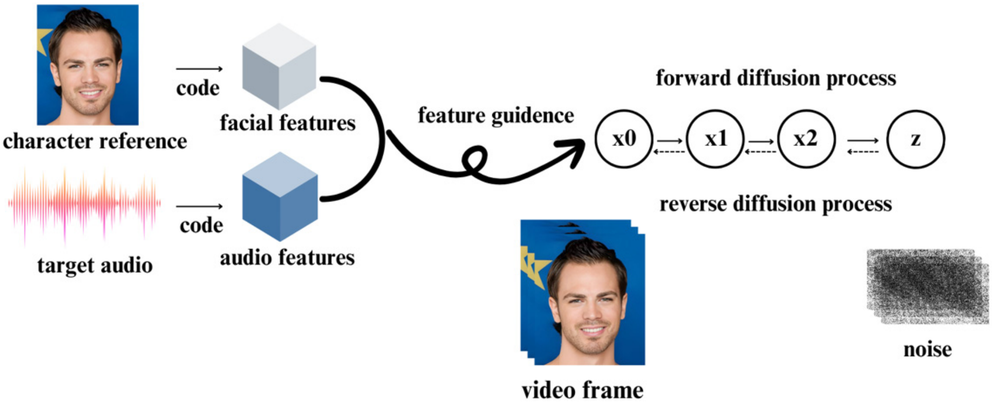

Figure 2: A Speech-Driven Speech Digital Person Based on the Diffusion Model. (Picture credit: Original)

Literature (Xiao, 2024) presents a high-quality and easy-to-operate digital portrait editing framework. This framework only uses monocular portrait video and text instructions as input to generate dynamic driving portrait models with temporal consistency and 3D consistency. By combining the semantic prior of the human body and the expression of facial prior to method improvement, according to the corresponding text instructions for accurate local editing, and keeping the accurate reconstruction of expression, thus according to the other information input stable control editing result expression and attitude, shows that the method in the field of dynamic digital portrait editing effectiveness and superiority.

In the literature (Liu et al., 2024), the diffusion model and the conditional model structured by U - Net are employed for the extraction of features. This approach aims to incorporate the interaction between noise and semantic features. Additionally, multi - scale images are integrated to fortify the subtle structural and boundary texture features present in low - contrast images. When confronted with intricate, low - contrast, and boundary - blurred images, in comparison to U - Net and SOLOv2, the image segmentation method founded on the Transformer - based diffusion model demonstrates enhanced stability and robustness.

Literature (Liu & Huang, 2024) proposes a model of face image repair based on the probability model of denoised diffusion. We improve the denoised diffusion probability model by using the U-Net network structure in Guided diffusion and introducing fast Fourier convolution into the network, and finally perform the model training and result evaluation on the CelebA-HQ HD face image dataset. Experimental results show that the improved denoising diffusion probability model in repairing random mask face image, repair results and the original peak signal to noise ratio (PSNR) can reach 25.01, SSIM (structural similarity) can reach 0.886, better than the improved before noise diffusion probability model and the existing generated based on the network face image repair model.

Diffusion models are often able to generate high-quality images and videos because they are based on probabilistic diffusion processes that can progressively recover target images from noise, thus retaining more detail and texture, while having relatively high flexibility to accommodate different datasets and task requirements.

However, the training and inference process of the diffusion model usually requires high computational cost, because multiple iterations are needed to gradually recover the target image from the noise, which also leads to a slow generation speed. In the application scenarios that require rapid generation of results, the performance is inferior to other generation models.

2.3 Speech-driven Digital Speaker Based on the Neural Radiation Field

The neural radiance field (NeRF) is an advanced 3D scene rendering approach rooted in deep - learning paradigms. Its fundamental principle revolves around leveraging neural networks to represent and render intricate 3D scenes. The input parameters for the neural radiance field encompass the camera pose (which precisely defines the position and orientation of the camera) and real - world images. The output is an implicit representation of the scene. Specifically, for every spatial point within the scene, the neural radiance field takes into account both its 3D coordinate position and the viewing direction. Subsequently, it computes and outputs the color and density values of that point. The rendering process of the neural radiance field is accomplished through volume rendering. Volume rendering is a sophisticated technique that generates images by meticulously calculating the color and brightness of light as it interacts with matter while traversing a three - dimensional scene. In the context of the neural radiance field, for each ray projected from the camera, the neural radiance field calculates the color and density of each point along the ray. These values are then subjected to a weighted summation process to determine the final color of that ray. The detailed process is illustrated in Figure 3.

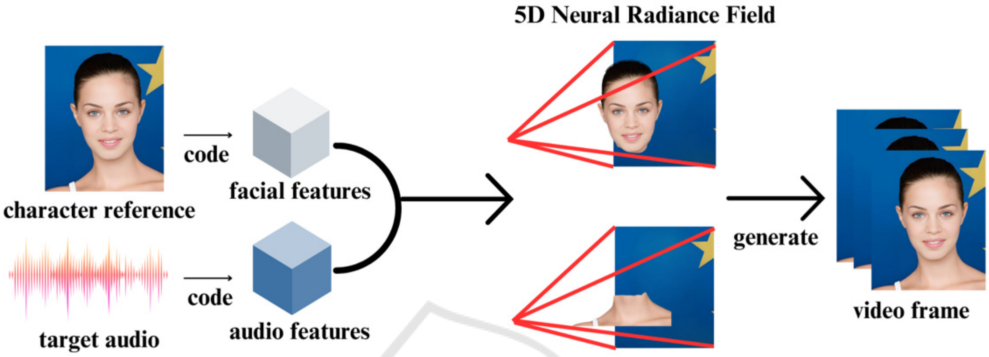

Figure 3: Speech-driven speech digital person based on the neural radiation field. (Picture credit: Original)

Literature (Sheng et al., 2024) presents a new method to rapidly generate high-precision face models based on monocular RGB videos, while constructing a new framework focusing on accurate modeling of face and neck regions. This framework integrates neural elements into parametric models of face and neck, and uses the Head-And-neCK (HACK) model to replace the commonly used Face Latent Animated Mesh Estimator (FLAME) model, which significantly improves the accuracy and efficiency of 3D face reconstruction. In addition, a real-time adaptive neural radiation field is specifically designed to effectively speed up the training and reconstruction process. By introducing a multi-resolution hash grid to compute deformation gradients using nearest triangle searches within the deformation space, the method enables rapid reconstruction of high-fidelity face and neck models within minutes. After extensive quantitative and qualitative evaluation, the experimental results show that the model exhibits significant advantages in terms of rendering quality and training time over the existing state-of-the-art methods.

Literature (Hu et al., 2024) proposed a dynamic neural radiation field coding and transmission method, through multi-resolution residual radiation field, using the bit rate and quality of multi-resolution feature grid to represent the dynamic neural radiation field construction volume video, the complete dynamic neural radiation field as characteristic voxel grid, and decomposed into multiple different resolution characteristic voxel grid, can be combined according to user requirements. Combined with the end-to-end joint optimization coding method, reduces time redundancy and ensures efficient multi-resolution characterization, improving compression efficiency and reconstruction quality. In addition, an adaptive bit rate scheme based on user experience is designed to realize flexible dynamic detail transmission.

NeRF can generate high-quality, realistic 3D models, suitable for complex objects such as faces, and can truly present the surface and texture details of objects. By using two independent neural radiation fields, it has powerful video editing capabilities, such as changing posture and replacing the background, and can generate models from any number of input images without specific processing. However, NeRF requires substantial computational resources and time, and may have problems when handling large-scale scenarios and complex illumination, affecting the generation quality. When the perspective is insufficient, it may lead to occlusion and emptiness. Moreover, existing methods have limitations in the fusion of audio and visual features, making it difficult to accurately map audio to facial areas.

3 DATA SET

Datasets play an important role in the speaking digit person generation task. In the speech digital human generation task, deep learning, as a typical data-driven technology, requires sufficient data to train the model to enable the model to learn the features of the task. On the one hand, the data set can serve as a common platform to evaluate the performance of different speech digital human generation algorithms; on the other hand, the size and diversity of the datasets also pose increasingly complex challenges to this task. With the size of the data set, the deep learning model can learn more complex features in the data; the high-resolution data set enables the deep learning model to generate detailed and more realistic digital speakers; In addition, the more diverse data set (including expressions, head movements and body posture) can help the model to better generalize, making the generated digital figures can better meet the needs of real life. The overall situation is shown in Table 2.

Table 2: Common speech-driven speech.

data set	Time / year	number of people	Without obvious head movements	Are there any emotional labels	class
GRID	2006	33	not have	not have	be in common use
CREMA-D	2014	91	not have	have	be in common use
LRW	2017	1000+	not have	not have	be in common use
MEAD	2020	60	have	have	be in common use
HDTF	2021	362	have	not have	be in common use
ObamaSet	2017	1	have	not have	specially appointed

In the QMUL Underground Re-Identification Dataset (GRID) dataset, there were 34 speaking individuals (18 men and 16 women), each facing the camera and reading 1,000 short sentences, consisting of six words randomly selected from a small dictionary containing only 51 words. All of the individuals had no significant mood fluctuations and head movements while speaking. The dataset was published by the University of Surrey, UK in 2006.

In the Crowds-Sourced Emotional Multimodal Actors Dataset (CREMA-D) data set, including 91 actors aged 20 to 74 (48 men and 43 women), different from other data sets, each actor with different categories of emotion (a total of 6 types of emotion) and emotional intensity (contains a total of 4 kinds of emotional intensity) to repeat the same sentence, each say a total of 12 sentences and there is no obvious head movements when talking.

The Lip Reading in the WildDataset (LRW) data set contains 500-word video clips read by hundreds of speaking individuals. One of them is that some videos contain only one speaking individual facing the camera, while others contain a group debate with more than one speaking individual. The dataset was collected from BBC TV broadcasts, where each video segment was short and there were no obvious head movements when individuals spoke. The data set was created by Imperial College London, UK.

The Multi - view Emotional Audio - Visual Dataset (MEAD) is a comprehensive, large - scale collection of audio - visual data centered on emotionally expressive speaker faces. This dataset has been meticulously designed to facilitate the generation of speaker faces imbued with specific emotions. Comprising data from 60 actors (an equal distribution of 30 male and 30 female participants), each actor has been recorded under strictly controlled conditions. They have been captured expressing different emotion categories (encompassing a total of eight distinct emotions) and emotional intensities (three levels in total). The recordings were made from multiple angles, with the aim of comprehensively capturing the minute details of the actors' facial expressions during emotional manifestations. This dataset was jointly introduced by SenseTime and other collaborating organizations.

High - resolution Audio - visual Dataset (HDTF) represents a high - resolution digital speaker dataset. This dataset is sourced from YouTube and subsequently undergoes processing and annotation procedures to facilitate the generation of high - resolution digital speakers. Comprising approximately 362 distinct videos, the HDTF dataset has an aggregate duration of 15.8 hours. All of the videos within this dataset possess a video resolution of either 720P or 1080P.

The ObamaSet represents a specialized audiovisual dataset. It is centered around the in - depth analysis of the speeches delivered by former US President Barack Obama. Functioning as a dedicated database, it serves a particular speaking digital human generation task. All video materials within this dataset are sourced from Obama's weekly addresses.

4 CONCLUSIONS

This research undertakes an in - depth and comprehensive assessment of the latest developments in the generation techniques of speaking digital humans. It approaches this from two critical perspectives: the fundamental technical components and the datasets involved. Broadly speaking, propelled by the rapid and remarkable progress of artificial intelligence technologies, which are firmly rooted in deep - learning algorithms, the current video generation technology for speaking digital humans has attained significant headway. However, it still contends with a multitude of complex and arduous challenges that impede its seamless and widespread implementation.

This academic treatise delves into the contemporary status quo and prospective evolution of voice - driven speech systems. In the wake of unceasing technological advancements, remarkable enhancements in the fidelity and naturalness of digital humans have been observed, attributed to generative methodologies including the generative adversarial network model, diffusion model, and neural radiance field. The progressive augmentation and diversification of datasets have furnished a more comprehensive and copious resource for digital human creation.

Within sectors such as entertainment, healthcare, and education, digital avatars are anticipated to emerge as pivotal instruments, endowing users with a more lifelike, intuitive, and immersive experience. Concurrently, with the relentless progression of technology and the continuous expansion of application scenarios, the technology for digital human generation is confronted with a plethora of challenges and concomitant opportunities.

REFERENCES

Chen, L., Cui, G., Kou, Z., Zheng, H., & Xu, C. (2020). What comprises a good talking-head video generation?: A survey and benchmark. arXiv preprint arXiv:2005.03201.

Hu, Q., Zhong, H. Q., Wang, W. S., et al. (2024). Efficient encoding and transmission method of volumetric video based on neural radiation field. The Radio and Television Network, 2024(S2), 41-45. https://doi.org/10.16045/j.cnki.catvtec.2024.s2.019

Liu, J. H., & Huang, X. X. (2024). Face inpainting model based on denoising diffusion probability models. Journal of Northeastern University (Natural Science), 45(9), 1227-1234.

Liu, Y., Wu, M. Y., Hu, Y., Qi, K., Wang, Y. B., Zhao, Y., & Song, J. L. (2024). Preliminary application of a cervical vertebra segmentation method based on Transformer and diffusion model for lateral cephalometric radiographs in orthodontic clinical practice. Journal of Shanghai Jiao Tong University (Medical Science), 44(12), 1579-1586.

Song, Y., Zhang, W., Chen, Z., & Jiang, Y. (2023). A survey on talking head generation. Journal of Computer-Aided Design & Computer Graphics, 35(10), 1457-1468.

Tang, X. L., Du, Y. M., Liu, Y. W., et al. (2018). Image recognition method for generating adversarial networks based on conditional deep convolution. Journal of Automation, 44(5), 855-864.

Xiao, H. Y. (2024). Digital human portrait editing based on radiation field and generative diffusion model (Unpublished doctoral dissertation). University of Science and Technology of China, Hefei, China.

Zhen, R., Song, W., He, Q., Cao, J., Shi, L., & Luo, J. (2023). Human-computer interaction system: A survey of talking-head generation. Electronics, 12(1), 218.

Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 9459-9468). IEEE.

Sheng, X. M., Zhao, J. L., Wang, G. D., et al. (2024). High-fidelity face generation algorithm based on neural radiation field. Computer Science, 1-15. Retrieved from http://kns.cnki.net/kcms/detail/50.1075.TP.20241225.1825.004.html