computers

MDPI

Article

The Generation of Articulatory Animations Based on Keypoint Detection and Motion Transfer Combined with Image Style Transfer

Xufeng Ling [1], Yu Zhu [2], Wei Liu [1], Jingxin Liang [1] and Jie Yang [3,*]

[1] AI School, Tianhua College, Shanghai Normal University, No. 1661 North Sheng Xin Road, Shanghai 201815, China; lxf1131@sthu.edu.cn (X.L.); lw2091@sthu.edu.cn (W.L.); ljx2665@sthu.edu.cn (J.L.)
[2] Shanghai Library, Institute of Scientific and Technical Information of Shanghai, 1555 West Huaihai Road, Shanghai 200030, China; yuzhu@libnet.sh.cn
[3] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai 200240, China
* Correspondence: jieyang@sjtu.edu.cn

Abstract: Knowing the correct positioning of the tongue and mouth for pronunciation is crucial for learning English pronunciation correctly. Articulatory animation is an effective way to address the above task and helpful to English learners. However, articulatory animations are all traditionally hand-drawn. Different situations require varying animation styles, so a comprehensive redraw of all the articulatory animations is necessary. To address this issue, we developed a method for the automatic generation of articulatory animations using a deep learning system. Our method leverages an automatic keypoint-based detection network, a motion transfer network, and a style transfer network to generate a series of articulatory animations that adhere to the desired style. By inputting a target-style articulation image, our system is capable of producing animations with the desired characteristics. We created a dataset of articulation images and animations from public sources, including the International Phonetic Association (IPA), to establish our articulation image animation dataset. We performed preprocessing on the articulation images by segmenting them into distinct areas each corresponding to a specific articulatory part, such as the tongue, upper jaw, lower jaw, soft palate, and vocal cords. We trained a deep neural network model capable of automatically detecting the keypoints in typical articulation images. Also, we trained a generative adversarial network (GAN) model that can generate end-to-end animation of different styles automatically from the characteristics of keypoints and the learned image style. To train a relatively robust model, we used four different style videos: one magnetic resonance imaging (MRI) articulatory video and three hand-drawn videos. For further applications, we combined the consonant and vowel animations together to generate a syllable animation and the animation of a word consisting of many syllables. Experiments show that this system can auto-generate articulatory animations according to input phonetic symbols and should be helpful to people for English articulation correction.

Keywords: computing methodologies; artificial intelligence; machine learning

Citation: Ling, X.; Zhu, Y.; Liu, W.; Liang, J.; Yang, J. The Generation of Articulatory Animations Based on Keypoint Detection and Motion Transfer Combined with Image Style Transfer. *Computers* **2023**, *12*, 150. https://doi.org/10.3390/computers12080150

Academic Editors: Zhitao Xiao and Guangxu Li

Received: 24 June 2023
Revised: 20 July 2023
Accepted: 25 July 2023
Published: 28 July 2023

1. Introduction

For language learners, how to learn accurate pronunciation is a very important task that even directly determines the success of language learning. When Chinese people learn English, many of them struggle with accurate pronunciation due to the influence of their native language. For example, they often confuse the pronunciation of "θ" and "s" or have difficulty distinguishing between the sounds of "i" and "i:". At times like these, simply relying on audio recordings and teacher guidance is not enough. However, if they have pronunciation videos, they can visually observe the typical positions of the articulatory organ, such as the teeth, lips, tongue, and jaw. This allows them to understand

the characteristics of the aforementioned sounds and consequently learn to pronounce them accurately. In order to solve the above problems, some researchers have proposed a pronunciation visualization method, which uses deep learning, graphics, voice processing, and other technologies to display the position and motion characteristics of the tongue and other important parts in the process of articulation in an animated way. Researchers have undertaken much work in the area of articulatory visualization and achieved success [1].

1.1. Related Research

Some researchers have created pronunciation animations through anatomical methods and physical modeling methods, while others have generated pronunciation animations using 3D computer graphics and deep learning methods.

Li et al. [2] proposed a 3D visualization method for tongue movement in Chinese pronunciation. By extracting the electromagnetic articulometer (EMA) data from three points on the surface of the back of the tongue as the driving source, the tongue movement in Chinese pronunciation can be truly reproduced using spring-mesh technology. Furthermore, a computer graphics method is used to simulate the detailed effect of tongue movement, and its X-ray images are compared. The experiment showed that the 3D tongue movement realized by the method conformed to the real tongue movement.

Mi et al. [3] proposed a lip animation method synchronized with speech based on muscle models and co-articulation modeling according to the pronunciation habits for Chinese and the requirements for natural and continuous lip animation in speech visualization technology. The co-articulation modeling method was designed based on a differential geometry description. Experiments and analysis showed that the lip animations generated by this method were more realistic and conformed to the pronunciation habits for Chinese.

Zheng et al. [4] proposed an improved cooperative articulatory model and used this method to synthesize the articulation track of Chinese characters. The experiment showed that the synthetic articulation track obtained by the improved method was closer to the real articulation track. Zhi et al. [5] analyzed and discussed the physiological characteristics of vowel articulation by using an electromagnetic articulation instrument and carried out visual voice training research based on the 3D articulation physiological model. The experiment showed that the scores of subjects using the 3D articulation model for vowels, consonants, and tones were higher than those using audio, indicating that the visual 3D articulation model was more helpful for learners' pronunciation learning than audio.

Yu [6] proposed a facial animation system for visual singing synthesis. With a reconstructed 3D head mesh model, both the finite element method and an anatomical model were used to simulate the articulatory deformation corresponding to each phoneme with a musical note. Based on an articulatory song corpus, articulatory movements, phonemes, and musical notes were trained simultaneously to obtain the visual co-articulation model using a context-dependent hidden Markov model (HMM). Articulatory animations corresponding to all phonemes were concatenated by a visual co-articulation model to produce the song-synchronized articulatory animation. Experimental results demonstrated that the system could synthesize realistic song-synchronized articulatory animations and increase the human–computer interaction capability objectively and subjectively.

Jiang et al. [7] proposed an accurate 3D tongue model based on anatomy and biomechanics and calculated the dynamic deformation of the tongue model through anatomical modeling, biomechanical modeling, and the finite element method. The experiments showed that the model had high accuracy and could generate realistic tongue movements according to muscle excitation and synthesize highly realistic tongue animations. Chen et al. [8] introduced visualization for articulatory animation. On the basis of the study of the anatomical structure and movement patterns of the tongue, a 3D tongue muscle model was established to simulate the common movements and shapes in articulation by combining the X-ray images of the articulation. Then, this visualization technology was used to help the hearing-disabled and speech-disabled achieve correction and rehabilitation.

Generative adversarial networks (GANs) are currently the most popular generative deep learning model. GANs mainly consist of a generator (G) and a discriminator (D). The goal of the generator is to generate data that are as close to real as possible, while the discriminator aims to judge whether the data are real or fake. During training, the generator tries its best to fool the discriminator, and the discriminator keeps honing its investigation skills. Through continuous confrontation, the model reaches a converged result. GANs have different variant structures, like the DCGAN [9], WGAN [10], and StyleGAN [11], which can be applied to different generation tasks.

1.2. Key Technologies in Articulatory Animation Generation

The key technologies for articulatory animation generation in our method included keypoint detection, motion detection, motion estimation, and image generation.

The most commonly used method for keypoint detection is the heatmap method [12]. Compared with the regression method, the heatmap method has the characteristics of stability and easy training, especially in human pose detection where the body movements vary a lot, making it difficult for the network to learn. Thus, researchers use the heatmap prediction method to detect keypoints, which can achieve good training results.

First, a convolutional neural network (CNN) was used to extract features; then, a fully connected network output heatmaps with N channels. Each channel in the heatmaps represented one keypoint. The number of channels equaled the number of keypoints. In each channel, the keypoint location was modeled as a 2D Gaussian distribution centered at that point. The most straightforward approach is to find the point with the maximum response above a certain threshold in each channel and take its coordinates as the location of that keypoint category. Classical heatmap neural networks include Hourglass [13], Openpose [14], and HRNet [15].

StyleGAN [11] is a new type of generative adversarial network (GAN) that has made significant progress in image generation. Its main advantages include the following:

1. It employs a stylized generator architecture that can explicitly control the style and content of images;
2. It uses a mapping network to match the distribution of random noise, resulting in more diverse generation results;
3. It introduces a feature mapping technique that can synthesize high-quality, high-resolution images;
4. It utilizes noise input, making the image quality clearer and the style more coherent;
5. It can generate highly realistic scene images.

StyleGAN has broad application prospects in many generative model applications. The StyleGAN series models are continuously improving and developing and are an important technology in the current field of image generation.

1.3. Main Innovation of Our Method

Based on the advancements in the self-supervised training [16–18] and the image generation mode of GANs [11,19], we developed a model that can generate a target articulatory animation using a driving articulatory animation and a typical target-style articulation image. The main innovations of this paper are as follows:

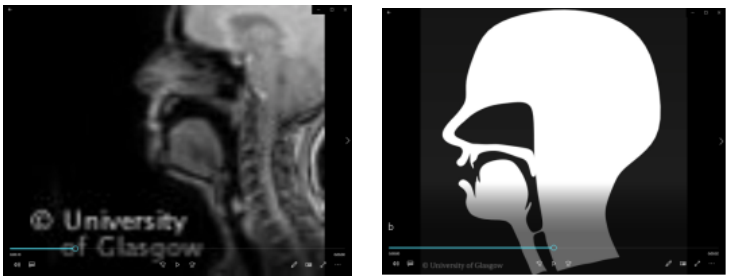
1. Different occasions require pronunciation animations of different styles, and the workload of manual drawing is very high. With our system, given a specific style image and a driving articulation animation, the system can automatically generate an articulation animation of the target style based on the image, thus greatly reducing the workload of manual drawing;
2. Our method creatively uses the techniques of keypoint extraction and keypoint registration to realize the automatic generation of articulation animations. Firstly, the keypoint extraction network is used to extract the keypoints of the target image. Then, the keypoints of the keyframe of the driving video are extracted. Next, there is a

registration match between the keypoints of the two images. This is followed by the prediction of pixel motion in the target animation through a dense motion field and animation generation is realized through a GAN network;

3. In order to improve the accuracy and visual effect of the generated animation, we also provide a manual configuration method. If the keypoints extracted from the original target image and the keyframe of the driving video do not match perfectly or have errors, we can calibrate them through manual methods. This can greatly improve the accuracy and visual effect of our video generation and only requires a very small increase in manual workload;
4. With the articulatory animation for each vowel and consonant, we provide a method to combine them into phonemes and to combine animations of different phonemes to form word animations to help English learners to improve their pronunciation.

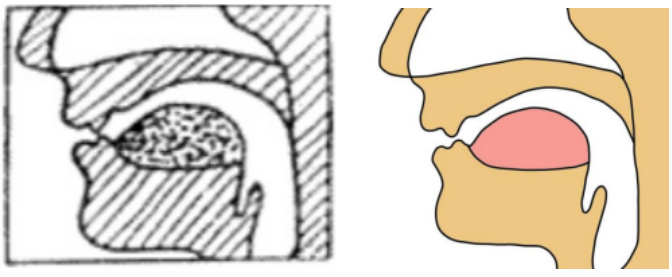
2. Method

Some articulatory animations currently exist. We mainly collected three datasets: the animation dataset from Glasgow University in the UK [20], shown as Figure 1; the dataset from the University in Canada (UBC) [21]; and the dataset from SpeechTrainer in Germany. Each of the above animation datasets has its own animation style. However, we found that, in practical applications, people's needs for animation styles vary. Some textbooks need cartoon animation styles, some videos need medical-style animation, and some need 3D style animation. To meet these different requirements, our method involves creating a picture of the target style and then selecting the existing driving animation so as to obtain the articulation animation in a new style. Of course, the auto-generated articulatory animation will not be perfect and will need to be manually fine-tuned, but it can greatly reduce the human workload. In addition, our system can also form syllable animations with vowels and consonants and then form word animations and sentence animations. Vowels and consonants drive the acquisition of animations, and ultrasonic images of vowels and consonants were obtained through various resources, as well as some animation images made by previous workers. Here, we need to see that the MRI image is complete and the manually drawn image is incomplete. Figure 2 shows the publicly used standard articulatory image and the decorated articulatory image drawn by our team.



(**a**) (**b**)

Figure 1. (**a**) Original MRI video, (**b**) hand-drawn articulation animation.



(**a**) (**b**)

Figure 2. (**a**) Original low-quality standard articulation image from textbook, (**b**) well-drawn articulation image from the original image.

2.1. Selection of Driving Video and Target-Style Image

We use the deep-learning motion transfer model to drive the target images to generate articulatory animations of different styles. There are five main implementation steps in the method:

1. First, we collect the publicly used articulatory animations to form our dataset. These animation datasets can be used as driving animations and also as training datasets;
2. In the second step, we preprocess the dataset, detect the articulatory region, and clip the region of interest, which can optimize the training data and greatly improve the follow-up training effect;
3. The third step is to design and train a deep learning network that includes keypoint detection, movement transfer, and style transfer. This network can transfer the style of the target image to the driving animation, thus forming a new style of animation. In the training process, the network collects a series of video samples, which contain different samples of the same object, and learns the potential expression of video motion features. With the characteristics of a single image frame, the model can reconstruct the video;
4. The fourth step is animation cascade by cascading consonants and vowels into a syllable animation. When multiple syllables are cascaded into a word animation, multiple connecting words can be cascaded into a sentence animation;
5. The fifth step is to optimize and perfect the animation manually according to the situation such that the articulation is accented, toned, and connected and, finally, a relatively complete system can be generated. The system framework is shown in Figure 3.

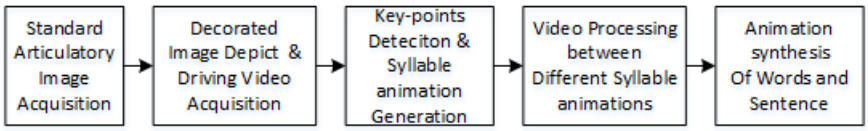


Figure 3. Flowchart of the implementation of the articulation animation system.

2.2. Keypoint Detection in Feature Space

Inspired by former research [22–24], we designed a motion migration model using artificial intelligence to realize style migration for articulation animations. There are two main tasks. First, the driving video is an MRI image, and the target image is a beautiful, manually drawn standard tongue image. The second step is to form a similar animation with the beautiful tongue image driven by MRI animation and then to form a clear and beautiful animation of the causal consonant pronunciation through a small amount of manual processing by artists.

In the training process, the model collects a series of video samples, including different samples of the same types of objects, such as different facial expressions and postures and different running horse videos. Then, the model learns the potential expression of the video motion features. With the features of a single image frame, the model can reconstruct the video. By observing a pair of different image frames extracted from the same video, the model can learn the motion coding, which mainly includes the specific keypoint displacements of the motion, and the affine transformation of the local image.

We adopted the method from [22–24] and introduced three constraints without the supervision of handmade labels. The detected keypoints must be able to reflect the visual concept consistently with human perception. Specifically, each landmark has a corresponding detector. The detector convolution outputs the detection score heatmap, and the detected landmark will be at the maximum value. In this framework, we use the deep neural network to convert the image $\mathbf{I}$ into a $(K+1)$ channel detection confidence map $\mathbf{D}$ in $[0,1]^{W\times H\times(K+1)}$. This confidence map detects K landmarks, and the $(K+1)$ channel represents the background. The resolution of $\mathbf{D}^{W\times H}$ can be equal to or less than $\mathbf{I}$, but they

should have the same aspect ratio. A lightweight hourglass network was developed to obtain the original detection score chart, as Equation (1) shows.

$$\mathbf{R} = hourglass_I(\mathbf{I}; \theta_I) \in \boldsymbol{R}^{W \times H \times (K+1)} \tag{1}$$

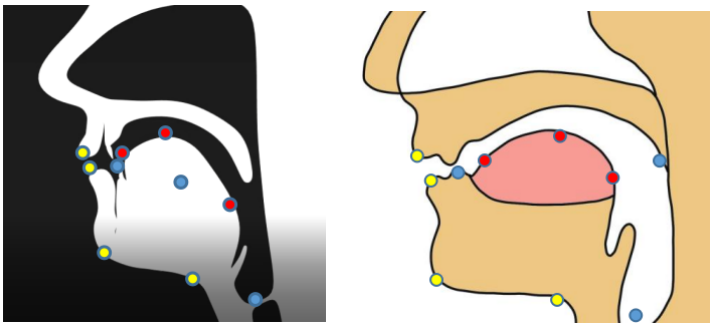
Here, θ_I represents a parameter. The hourglass architecture allows the detector to focus on the key local mode of the landmark location while taking advantage of the higher-level context. Then, we convert the unbounded original score into a probability and encourage each channel to detect different patterns. To this end, we use softmax to normalize the $\mathbf{R}$ across channels (including the background) to obtain the detection confidence diagram, as Equation (2) shows.

$$\mathbf{D}_{(u,v)} = \frac{exp(\mathbf{R}_k(u,v))}{\sum_{k'=1}^{K+1} exp(\mathbf{R}_{k'}(u,v))} \tag{2}$$

where matrix $\mathbf{D}_K$ is the kth channel of $\mathbf{D}$, scalar $\mathbf{D}_{K(u,v)}$ is the value of K at pixel (u,v), and the vector $\mathbf{D}(u,v)$ in $[0,1]^{K+1}$ can also be used to represent the multi-channel value of $\mathbf{D}$ at (u,v). The same representation is also applicable to other triaxial tensors. In $\mathbf{D}_K$, there is a weighted map. We take the weighted average coordinates as the location of the k landmark, as Equation (3) shows.

$$(x_k, y_k) = \frac{1}{\zeta_k} \sum_{v-1}^{H} \sum_{u=1}^{W} (u,v) \cdot \mathbf{D}_k(u,v) \tag{3}$$

By observing a pair of different image frames extracted from the same video, the model can learn the motion coding, which mainly includes the specific keypoint displacement of the motion, and the affine transformation of the local image, as shown in Figure 4.



(**a**) (**b**)

Figure 4. The keypoint detection model located the left 10 landmarks in the source image (**a**) and the right 10 landmarks in the target image (**b**).

2.3. Style Transfer Model for Articulation Image

In our approach, inspired by the method from [25,26], we trained a motion transfer model with a small sample. We set the source image S and the drive video frame D as the input. By combining a single frame and the potential expression of motion in the learned video, we can train a model to reconstruct the training video. We extract two different image frames from the same video. The model encodes the keypoint displacement motion and obtains the local affine transformation.

Our model consists of two modules: a motion estimation module and an image generation module. The purpose of the motion estimation module is to predict a dense motion field from an image $\mathbf{D}$ in $\mathbb{R}^{3 \times H \times W}$with dimensions of $H \times W$ in the drive video $\mathcal{D}$ combined with the source image frame $\mathbf{S}$ in $\mathbb{R}^{3 \times H \times W}$.

Assuming an abstract reference frame $\mathbf{R}$, we independently estimate two changes: from $\mathbf{R}$ to $\mathbf{S}(\mathcal{T}_{\mathbf{S} \leftarrow \mathbf{R}})$ and from $\mathbf{R}$ to $\mathbf{D}(\mathcal{T}_{\mathbf{D} \leftarrow \mathbf{R}})$. Therefore, we can process $\mathbf{D}$ and $\mathbf{S}$, respectively. The source image and drive image of the model input may be quite different. The

motion estimation module does not directly predict $\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}$ and $\mathcal{T}_{\mathbf{SleftarrowR}}$ and instead the prediction is carried out in two steps.

In the first step, we approximate these two transformations from the sparse track set and obtain them by using the keypoints from the self-supervised learning. The positions of keypoints in D and S are predicted by the coder–decoder network. This sparse motion representation is very suitable for animation. During the test, the keypoints of the source image can be moved using the keypoint track in the driving video. We use local affine transformation to build a motion model near each keypoint. Local affine transformations allow us to model a larger family of transformations than if we use only key displacements. We use Taylor expansion to express $\mathcal{T}_{\mathbf{D}\leftarrow\mathbf{R}}$ through a set of keypoint positions and affine transformations. To this end, the key detector network outputs the key position and the parameters of each affine transformation.

In the second step, the dense motion network is combined with local approximation to obtain the dense motion field $\hat{\mathcal{T}}_{\mathbf{S}\leftarrow\mathbf{D}}$. In addition to dense motion fields, the network also outputs an occlusion mask $\hat{\mathcal{O}}_{\mathbf{S}\leftarrow\mathbf{D}}$ that indicates which image parts of $\mathbf{D}$ can be reconstructed by distorting the source image and which parts should be redrawn; that is, inferred from the context. Finally, the generation module renders the moving image provided by the source object in the drive video. Here, we use a generator network G, which is based on $\hat{\mathcal{T}}_{\mathbf{S}\leftarrow\mathbf{D}}$, and it distorts the source image and fills the occluded part in the source image. The framework is shown in Figure 5.

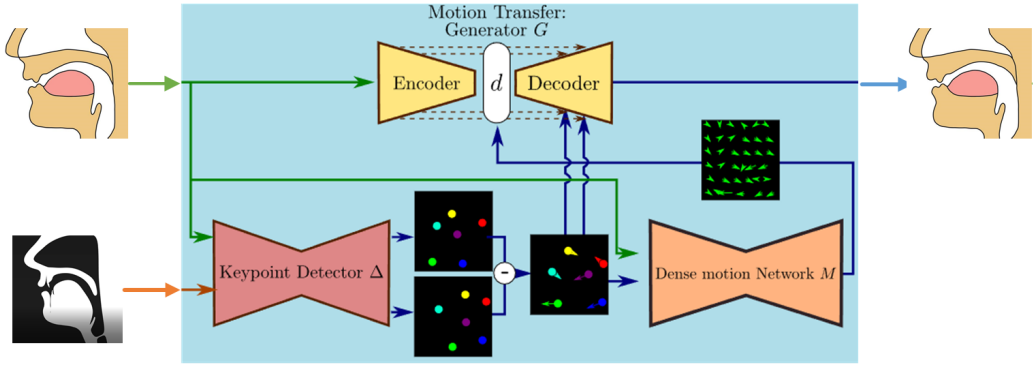


Figure 5. The system architecture, The lower graph demonstrates the model used to detect the keypoints and estimate the dense motion field, while the upper graph demonstrates the animating generator combined with the motion transformation.

2.4. Animation Generation for Syllables and Words

In English pronunciation, stressed and unstressed sounds correspond to the intensity and duration of the pronunciation. In terms of generating the syllable animation, in accordance with the characteristics of English pronunciation, we combine consonants and vowels into syllables and generate an animation. We can then easily control the duration of the animation generation, primarily through the control of the duration of consonants and vowels. Our statistics show that, for stressed sounds, consonants are approximately 1.1 times longer in duration and vowels are approximately 1.5 times longer. For unstressed sounds, consonants are about 0.9 times longer and vowels are about 0.67 times longer in duration. we prepare each word into a series of syllable combinations.

3. Experiments and Results

The experimental platform utilized Ubuntu 20.04 and Python 3.8. Four NVIDIA GEFORCE 3090 GPUs were used to train and tune the deep learning network. In the experiment, we chose the Pytorch 1.12 deep learning framework and created an independent virtual environment with miniconda3.

3.1. Dataset Preparation

The image dataset in this study included four kinds of videos, with 50 videos for each type and 200 videos in total. The first type were MRI videos of vowel and consonant pronunciation, and the other three types were animation videos with different styles. The videos mainly showed the animation of the tongue, lips, and upper and lower jaw on the side of the mouth during pronunciation. As the experimental process was unsupervised, all samples were test samples. Video samples from the dataset are shown in the Figure 6.

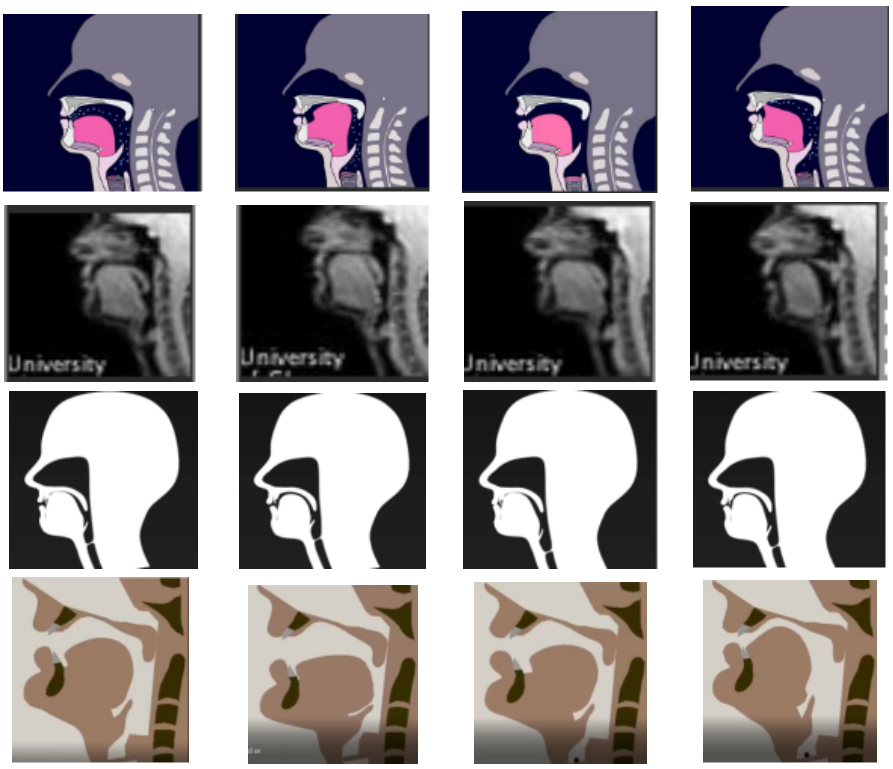


Figure 6. Samples of the articulation images and animations created from public sources.

We standardized and normalized the video mainly via the following three steps.

1. The articulation videos were extracted into frame-by-frame images in RGB format. All image frames for the same video were saved in the same directory. Each image contained different numbers of frames: the minimum was 10 frames and the maximum was 120 frames;
2. For the original images that were large, most of the areas were still and did not move, and the moving parts were mainly in the lower right corner. We used software to automatically intercept the lower right corner area of each image. The width and height of the intercepted area were equal, and it was a square area;
3. We normalized the intercepted area into RGB images 256 pixels in height and width with three channels in image format and finally obtained 256 × 256 × 3 RGB images.

3.2. Evaluation Method

It was difficult to evaluate the effect of the image animation generation model. Referring to past experience, we used two tasks to evaluate the model. The first was to evaluate the video reconstruction task performed by the model by extracting the keypoints of each frame of the driving video and, at the same time, extracting the first frame of the target image to regenerate the driving video. The second was to manually evaluate the accuracy of keypoint extraction. In the experiment, we selected 10 key points, including 3 points on the tongue, 2 points on the lips, 1 point on the lower teeth, 1 point on the vocal cords, 1 point on the small part of the tongue, and 2 points on the lower jaw.

3.3. Training

Before the formal training of the model, we downloaded the pre-trained model. As the number of samples was small and the results of training images from scratch are poor, our main training work was in optimizing the pre-trained model. In the process of model optimization, it was necessary to reasonably set the initial values of the parameters. The model was easier to train and tended to converge quickly. The batch size for network training was set to 24, and the total number of training epochs was set to 100. The learning rate of the model decreased with the increase in the number of training epochs. The model training curve is shown in Figure 7.

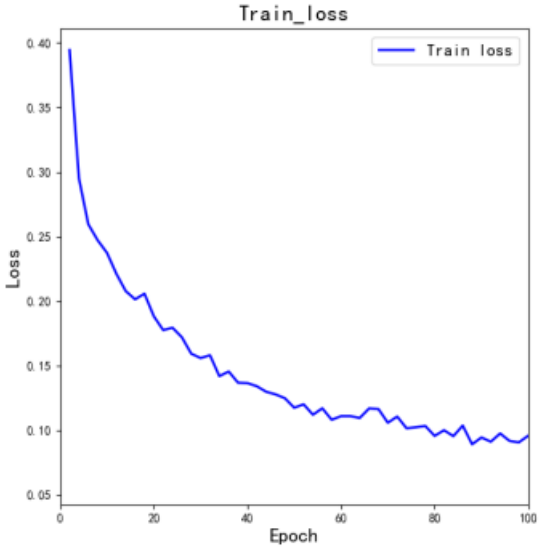


Figure 7. Training loss with the model.

3.4. The Results of Keypoint Detection

The trained network could automatically extract keypoints, as shown in Figure 8. We tested the test pictures and estimated the average error to be 12.3%. The error is the relative error. The average error D represents the error for the corresponding point between the model output result and the annotation result. As shown in Equation (4), p_i indicates that 10 keypoints were detected, q_i refers to the 10 pre-labeled keypoints, and W refers to the image width. We calculated the square sum of relative errors for keypoint extraction from a single photo.

$$D = \sum_{i}^{n} \frac{(p_i - q_i)^2}{W^2} \tag{4}$$

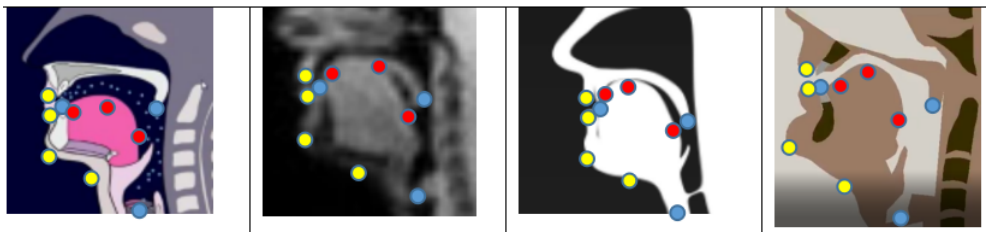
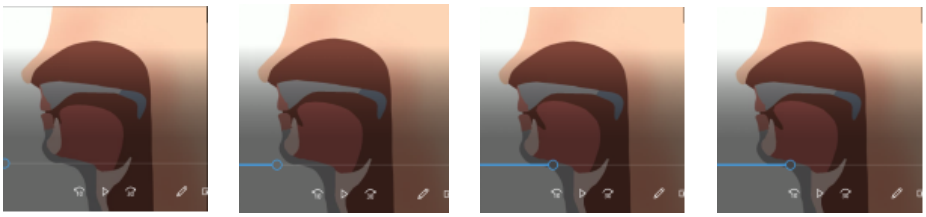


Figure 8. The result of keypoint detection.

3.5. Animation Generation

The sample frames in a generated animation is shown in Figure 9. Assessing the quality of the images generated by the method was a rather challenging task due to the subjective nature of evaluation criteria and the absence of annotated datasets. Therefore, we employed a manual evaluation method, comparing the animations generated by the model with animations manually crafted in MAYA software. In the first step, we randomly selected

108 pairs of videos with similar object poses in the first frame. We presented three videos to the evaluators—one as the driving video, the second as the video automatically generated by the model, and the third as the video hand-drawn in MAYA—but the evaluator did not know which one was generated from the network and which was drawn by hand. The evaluators were asked to select the video that best resembled the driving video. We engaged 25 evaluators to assess the videos, and the evaluation results are presented in Table 1. For vowel animations, the videos generated by our model were superior to the hand-drawn MAYA videos. Conversely, for consonant videos, the model-generated videos were slightly inferior to the hand-drawn MAYA videos.



(a) (b) (c) (d)

Figure 9. The generated animation with slight manual fine-tuning. (**a**–**d**) are the sample frames at different time from the animation.

Table 1. User appraisal: animation generated via model generation vs. animation hand-drawn in MAYA.

Types	Video Automatically Generated by the Model	Video Hand-Drawn in MAYA Software
Consonants	56.22%	43.78%
Vowels	48.17%	51.83%

3.6. Discussion

Advantages. The accuracy and quality of our method mainly depend on three factors. First, it is critical that the extraction of the 10 keypoints is accurate and that they can be registered correctly. If the extraction and registration of the 10 keypoints are not precise, it will significantly affect the accuracy of subsequent work. Second, after extracting the 10 keypoints, the prediction of the dense motion field for each pixel based on the keypoints must be accurate. Third, an excellent style-based generative model is needed to produce the desired style for the articulation animations. Our method is continuously improvable and optimizable. Improvements in these three steps can significantly enhance the output of our method. If the keypoint extraction becomes more accurate, the system will be able to generate animations fully automatically without manual intervention. If the prediction of the dense motion field is accurate, it will result in better image quality.

Disadvantages. The extraction of the 10 keypoints was sometimes not very accurate. At these times, manual intervention was required through interactive operation to adjust the keypoints. We hope that future improvements will significantly enhance the accuracy of the extraction of the 10 keypoints. At present, we lack a baseline model and evaluation formula to determine the effectiveness of our method. Currently, we can only evaluate the outputs through manual methods. Our method involves an animation generation model, which is different from image generation. The baseline model and evaluation methods for image generation are not directly applicable in evaluating the quality of the generated articulation animations. We hope to have such a baseline model in the future.

Future research. As we continue to improve our style transfer and generative models, such as by using the currently more effective diffusion model [27,28], we might achieve better results. We plan to generate videos based on the diffusion model in our future work. Going beyond GANs, diffusion models [27] have emerged as powerful new deep generative models, achieving state-of-the-art results in image synthesis and other generation tasks. By

gradually blurring the original image by continuously adding noise and then gradually restoring a clear image through a reverse process, this process of progressively adding Gaussian noise and then gradually denoising can generate high-quality synthetic images. The main advantages are:

1. The process can generate high-resolution and high-quality images. The images are rich in detail and the quality is close to that of real pictures;
2. The training process is more stable and can handle complex, high-resolution images;
3. The generation process is deterministic, and the same image can be generated each time the same noise is input;
4. The model framework is simple and easy to implement.

There are a vary of diffusion models, including DDPM [29], DDIM [30], etc. As the models continue to be optimized, diffusion models are becoming one of the important methods in the field of image generation.

4. Conclusions

During the process of learning English, displaying dynamic mouth and tongue movements can assist individuals in practicing pronunciation accurately. However, the creation of articulation animations requires significant resources. To address these challenges, this paper proposes a method that utilizes keypoint detection to locate crucial points during oral articulation. Subsequently, a motion transfer model is trained to track the essential motion trajectories, enabling the transfer of articulatory images. By combining different image styles, the method generates articulation animations that can meet specific requirements. The experimental results demonstrate the ability of this method to simulate realistic articulation animations based on input phonetic symbols. Nevertheless, due to the limited size of the driving animation dataset, the automatically generated animations may lack clarity, necessitating further manual adjustments. Nonetheless, this approach significantly reduces manual labor by over 90% and greatly improves the efficiency of animation generation. Initial usage indicates that the method is beneficial for practicing English articulation, particularly in correcting issues related to incorrect mouth and tongue positions that lead to inaccurate articulation. Future work will involve acquiring additional driving video data, further enhancing the motion transfer model to enable high-definition animation output, and applying the developed model to synthesize voice-synchronized tongue animations.

Author Contributions: X.L. proposed the innovation and wrote part of the manuscript. J.Y. guided the research. Y.Z., W.L. and J.L. provided the data, wrote the programs, and finished the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China, ID. 42075134.

Data Availability Statement: No permissions are required.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Li, X.; Zhang, Z. Review of Speech Driven Facial Animation. *Comput. Eng. Appl.* **2017**, *22*, 142–149.
2. Li, R.; Yu, J.; Luo, C.; Wang, Z. 3D Visualization Method for Tongue Movements in Pronunciation. *PR AI* **2016**, *5*, 142–149.
3. Mi, H.; Hou, J.; Li, K.; Gan, L. Chinese Speech Synchronized 3D Lip Animation. *Appl. Res. Comput.* **2015**, *4*, 142–149.
4. Zheng, H.; Bai, J.; Wang, L.; Zhu, Y. Visual Speech Synthesis Based on Articulatory Trajectory. *Comput. Appl. Softw.* **2013**, *6*, 142–149.
5. Zhi, N.; Li, A. Phonetic Training Based on Visualized Articulatory Model. *J. Foreign Lang.* **2020**, *1*, 142–149.
6. Tang, Z.; Hou, J. Speech-driven Articulator Motion Synthesis with Deep Neural Networks. *Acta Autom. Sin.* **2016**, *6*, 142–149.
7. Jiang, C.; Yu, J.; Luo, C.; Wang, Z. Physiology Based Tongue Modeling and Simulation. *J. Comput.-Aided Des. Comput. Graph.* **2015**, *12*, 142–149.
8. Chen, Z.; Xin, Q.; Zhu, Y.; Lin, Q.; Wang, L. Visualization Study of Virtual Human Tongue in Speech Production. *Chin. J. Rehabil. Theory Pract.* **2013**, *10*, 142–149.

9. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2016**, arXiv:1511.06434v2.
10. Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; Courville, A.C. Improved Training of Wasserstein GANs. *arXiv* **2017**, arXiv:1704.00028v3.
11. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. *arXiv* **2020**, arXiv:1912.04958v2.
12. Nibali, A.; He, Z.; Morgan, S.; Prendergast, L. Numerical Coordinate Regression with Convolutional Neural Networks. *arXiv* **2018**, arXiv:1801.07372v2.
13. Newell, A.; Yang, K.; Deng, J. Stacked Hourglass Networks for Human Pose Estimation. *arXiv* **2016**, arXiv:1603.06937v2.
14. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.-E.; Sheikh, Y. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv* **2019**, arXiv:1812.08008v2.
15. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep High-Resolution Representation Learning for Human Pose EstimationKe. *arXiv* **2019**, arXiv:1902.09212v1.
16. Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O.K.; Singhal, S.; Som, S.; et al. Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasksar. *arXiv* **2022**, arXiv:2208.10442v1.
17. Cheng, B.; Schwing, A.; Kirillov, A. Per-Pixel Classification is Not All You Need for Semantic Segmentation. *arXiv* **2021**, arXiv:2107.06278v2.
18. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030v2.
19. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv* **2019**, arXiv:1812.04948v3.
20. Available online: https://www.gla.ac.uk/ (accessed on 1 January 2012).
21. Available online: https://enunciate.arts.ubc.ca/ (accessed on 1 January 2018).
22. Zhang, Y.; Guo, Y.; Jin, Y.; Luo, Y.; He, Z.; Lee, H. Unsupervised Discovery of Object Landmarks as Structural Representations. *arXiv* **2018**, arXiv:1804.04412v1.
23. Jakab, T.; Gupta, A.; Bilen, H.; Vedaldi, A. Conditional Image Generation for Learning the Structure of Visual Objects. *arXiv* **2018**, arXiv:1806.07823v1.
24. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. Animating Arbitrary Objects via Deep Motion Transfer. *arXiv* **2019**, arXiv:1812.08861v3.
25. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First Order Motion Model for Image Animation. *arXiv* **2020**, arXiv:2003.00196v3.
26. Siarohin, A.; Woodford, O.J.; Ren, J.; Chai, M.; Tulyakov, S. Motion Representations for Articulated Animation. *arXiv* **2021**, arXiv:2104.11280v1.
27. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. *arXiv* **2020**, arXiv:2006.11239v2.
28. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv* **2022**, arXiv:2112.10741v3.
29. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. *arXiv* **2022**, arXiv:2010.02502v4.
30. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.H. Diffusion Models: A Comprehensive Survey of Methods and Applications. *arXiv* **2023**, arXiv:2209.00796v10.