Audio-video synchronous generation algorithm based on Deepfake

Yongyu Sun

June 17, 2024

Abstract

With the rapid development of deep learning technology in audio-video processing, audio-video synthesis technology has shown broad application prospects. Based on Deepfake technology, this study proposes a new audio-video synchronization method by combining SimSwap and VITS models, aiming to generate highly realistic audio-video content both visually and auditorily. We utilize the identity information injection module and weak feature matching algorithm of the SimSwap framework to achieve high-fidelity face swapping. Simultaneously, we introduce the GPT-SoVITS framework based on the VITS model to enable end-to-end text-to-speech conversion. By comparing with existing face-swapping and voice cloning techniques, this study verifies the effectiveness of the proposed algorithm in enhancing the realism of synthetic media. Finally, we summarize the practical application effects and limitations of this technology, and propose future research directions, particularly focusing on how to further enhance the expressive capability of the model and improve the naturalness of synthetic media.

Keywords: Deepfake Simswap VITS GAN audio-video Synchronization

1 Introduction

In recent years, with the vigorous development of deep learning technology, Deepfake technology has particularly attracted attention. As a major breakthrough in image processing and video synthesis, Deepfake technology has garnered widespread attention. Specifically, Deepfake is a technology that utilizes deep learning techniques, especially Generative Adversarial Networks (GANs), to synthesize highly realistic artificial audio and video.

By applying deep learning models to audio and video separately, we hope to achieve more refined and lifelike Deepfake audio-video content. The rise of this technology has sparked extensive discussion, attention, and profound reflection. It not only has great entertainment and creative potential but may also be used for malicious acts such as fake news and online fraud.

The audio-video synchronization generation algorithm is an important branch of Deepfake technology, and its deep technical foundation brings together the essence of deep learning, image processing, signal processing, and other aspects. In particular, it benefits from deep learning image generation techniques such as Generative Adversarial Networks (GANs) and Autoencoders[2]. These technologies provide powerful tools and theoretical support for audio-video synchronization generation. Through deep learning models' in-depth study of massive amounts of real audio-video data, we can generate highly realistic audio-video content. Additionally, image processing and signal processing techniques also play a crucial role in audio-video synchronization generation, including face detection and recognition, speech synthesis, and motion tracking. The combined power of these advanced technologies not only improves the accuracy of audio-video synthesis but also further enhances its realism, making the generated audio-video content increasingly lifelike.

In this context, the audio-video synchronization generation algorithm based on Deepfake has become one of the research hotspots. Through in-depth study of related theories and methods, we hope to further improve the quality and efficiency of audio-video synchronization generation, expanding its application in various scenes such as film and television production, advertising marketing, e-commerce, social entertainment, etc.[3]. At the same time, it also provides technical support and solutions to address the risks that Deepfake technology may bring. Therefore, this article aims to explore the audio-video synchronization generation algorithm based on Deepfake, deeply analyze its

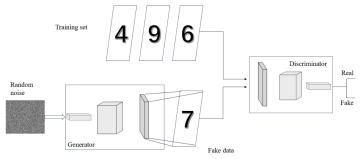

Figure 1: Training process diagram.

principles and methods, discuss its challenges and opportunities in practical applications, and propose corresponding technical improvements and coping strategies. Thus, it provides theoretical and practical support for research and application in related fields.

2 Relevant methodological foundations

2.1 deep learning model

Generative Adversarial Networks (GANs) are composed of two main components: the Generator and the Discriminator. The Generator acts as a data generator, dedicated to generating fake data (or "fake samples"), aiming to simulate the original data as realistically as possible in order to deceive the discriminator; while the Discriminator plays the role of a data discriminator, whose mission is to accurately differentiate between these fake data and real data. Through continuous adversarial training and mutual optimisation, the models gradually learn the intrinsic distribution patterns of the real data until the discriminator finds it difficult to distinguish the fake data from the real data. This process enables the generator to eventually output highly realistic simulated data. Specifically, the generator's task is to take a random vector (often referred to as noise or samples from the latent space) and transform it into an image similar to the training data. In short, the generator endeavours to create data that is sufficiently realistic with a view to deceiving the discriminator. The task of the discriminator is to evaluate the received set of images to determine whether they are real (i.e., derived from the training dataset) or made by the generator, which acts as a binary classifier for evaluating whether the images belong to the real data distribution.

The training of GANs is a typical zero-sum game in which generators and discriminators compete with each other and take turns to optimise. In the initial phase, the generator and discriminator start training by randomly initialising the weights. In each training step, the discriminator processes a portion of the real samples and a portion of the samples generated by the generator, learning to distinguish between the two. The goal is to identify the real vs. generated samples as accurately as possible. Meanwhile, the generator works to create samples that are as realistic as possible in order to maximally deceive the discriminator. This process is carried out by continuously tuning the parameters of the generator and the discriminator, aiming to minimise the difference between real and generated samples while maximising the classification accuracy of the discriminator. This competitive iteration continues until some predefined end point is reached, such as a fixed number of training rounds or the quality of generated samples meets the requirements. The training process is shown in Figure 2.1.

The mathematics of GANs involves achieving a dynamic balance between generative and discriminative models through adversarial training. At its core is a max-minimax problem (minimax game), where the generative model G receives a random noise z as input and tries to generate samples similar to real data; its goal is to learn the distribution of real data from a latent space.

Networks of GANs consisting of generators and discriminators compete with each other through adversarial training as a means of generating realistic data, which has a wide range of applications in fields such as image generation.

2.2 audio processing

Mel-spectrogram is a widely used technique in the field of speech processing and audio analysis, which is based on a nonlinear mel scale and logarithmic spectral representation combined with a linear cosine transform. This method not only correlates the energy distribution of audio signals in the frequency domain, but also more closely matches human auditory perception.

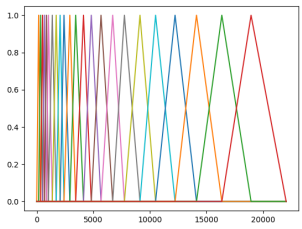

Figure 2: Mel filters within Hz frequencies (24).

The Mel scale is a nonlinear frequency metric that simulates human auditory perception. Because of the non-linear nature of human frequency perception - i.e., the difference in perception between different frequencies is not equal - the Mel Scale provides an accurate simulation by experimentally measuring frequencies. This scale converts conventional linear frequencies to Mel frequencies, and the conversion equation is shown in equation (2-1):

$$M(f) = 2595 \times \log_{10}(1 + \frac{f}{700}) \qquad (2-1) \tag{1}$$

where f is the linear frequency and M(f) is the corresponding Mel frequency. A Meier filter bank is a set of filters that are uniformly distributed on the Meier scale. Each filter corresponds to a frequency range, and these frequency ranges are usually triangular or trapezoidal. A Mel filter bank typically analyses a signal in the frequency domain, with each filter corresponding to a range of frequencies and with different weighting factors for frequencies within that range. These filters are designed to mimic the frequency perception properties of human hearing. A Mel filter bank within the Hertz frequency is shown in Figure 2.3, with a number of 24.

3 Design of Deepfake-based audio-video synchronous generation method

3.1 Deepfake Technology Overview

Deepfake technology is an advanced synthetic media technology that combines deep learning and artificial intelligence to generate virtual audio and video content with such a high degree of realism that viewers find it difficult to recognise its authenticity. The development of the technology began in 2014 when researchers started applying deep learning algorithms to create synthetic face images. By 2017, with significant advances in deep learning algorithms and computing power, Deepfake technology attracted widespread attention, especially the realistic face-swapping video posted by a user named Deepfake on the Reddit platform became the centre of attention. Since then, as the technology has rapidly gained popularity, the quality and realism of the content it generates has continued to improve with technological advances, making the line between real and fake increasingly blurred. Initially, Deepfake was mainly used for special effects in the film industry, but as the technology matured, it was recognised that it could also be used for fraud and the dissemination of false information.At its core, Deepfake relies on deep learning models, in particular Generative Adversarial Networks (GANs), which play a key role in the field of face counterfeiting, generating high-quality data, learning data distributions, and creating diversity. data, learn data distributions, and create diversity.The GANs framework consists of two main neural networks: a generator network, which is responsible for generating fake media, and a discriminator network, which is responsible for recognising the difference between real and fake media. Through continuous adversarial training, the generator continuously learns how to create more realistic content, while the discriminator improves its ability to recognise fake content, and together they drive the Deepfake technology towards the goal of high-quality generation.

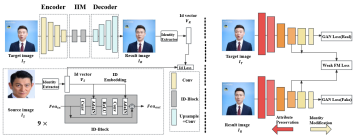

Figure 3: Simswap's framework diagram.

3.2 Simswap basic process and related algorithms

SimSwap distinguishes itself from previous technologies that only target specific identities or require complex attribute adjustments by introducing an innovative ID Injection Module that enables identity transfer between arbitrary source and target faces. This module grafts the identity information of the source face to the target face at the feature level. Meanwhile, in order to preserve the original attributes of the target face, such as expression, posture and lighting conditions, Simswap effectively maintains the identity of the target face by employing the Weak Feature Matching Loss (WFML) technique. The framework about Simswap is shown in Figure 3.1.

4 A concrete implementation of Deepfake-based audio and video synchronous generation

4.1 overall design

In this chapter, we detail the overall architecture of the Deepfake-based audio-video synchronisation generation algorithm, aiming at multidimensional deepfaking at the visual and auditory levels. For this purpose we mainly use SimSwap and GPT-SoVITS as the main frameworks, aiming to improve the realism and usefulness of the generated content.

In the Simswap framework, we use insightface [33] for image preprocessing such as face detection, feature point localisation, face alignment, etc. for video frames. Among them, RetinaFace [34] is used to accurately locate the face in the image. Feature point localisation helps us to obtain the positions of key feature points (e.g., eyes, nose, mouth, etc.) of a face, and then achieve alignment processing to ensure that the feature points can be correctly matched with each other when a face is exchanged, so as to improve the naturalness and coherence of the exchanged face. In addition, insightface can also be used for feature extraction and identity preservation, which not only extracts the basic features for neural network learning and generating new faces, but also preserves the core identity features of the original face. Subsequently on the detected faces we apply face-parsing for deep parsing to segment the face image into multiple specific facial regions (e.g., eyes, eyebrows, mouth, hair, etc.) and annotate each region. In the feature fusion stage, we use a model trained by Identity GAN (ID-GAN) to apply the parsing results obtained by face-parsing to insightface , align the feature points of the source face with the corresponding feature points of the target face and fuse the corresponding features of the two, and after fusion is completed, the model generates a face mask that contains the target's features. After the fusion, the model will generate a face mask containing the target facial features, which will be combined with the source image to replace the original facial features. Finally, the visual effect is further enhanced by post-processing such as colour correction.

In the GPT-SoVITS framework, we first used UVR5 for vocal accompaniment separation due to the presence of background music and other noises in most of the audio. For the needs of this experiment, we selected the audio without harmony for vocal extraction, so we used the HP2 model for vocal separation. Before cutting the audio we adjusted the maximum volume of the audio through FFmpeg to ensure that it did not exceed -9dB to -6dB. audio-slicer was then used to perform the cuts, in which we used RMS (Root Mean Score) to measure the quietness of the audio and to detect muted parts. The RMS value is calculated for each frame (frame length is set to jump size) and all frames with RMS below the threshold are considered as silent frames. And once the valid (sound) part is detected to reach the minimum length since the last slicing and the silent part is detected beyond the minimum interval, the audio will be separated from the frame with the lowest RMS value in the silent region. This speech data is then automatically transcribed with the help of an offline batch Automatic Speech Recognition (ASR) tool. Based on the ASR results, SubFix is used for data cleaning of the speech cor-

development tool	Configuration
operating system	Linux
programming language	Python
Deep learning frameworks	PyTorch=1.8
GPU acceleration	NVIDIA GeForce GTX 2080
CPU	Intel(R) Xeon(R) CPU E5-2678 v3 @ 2.50GHz

Table 1: Simswap Experimental Environment.

development tool	Configuration
operating system	Windows 10
programming language	Python
Deep learning frameworks	PyTorch=1.8
GPU acceleration	NVIDIA GeForce GTX 1650
CPU	Intel(R) Core(TM) i7-9750H CPU @ 2.60GHz 2.59 GHz

Table 2: Table 4.2 GPT-SoVITS experimental environment.

responding to the text, so that each sentence of speech can be better tailored to the text. The training set is then formatted, which includes converting the audio into semantic tokens containing timbre using $\mathrm{CN}_H ubert$[35], $a speech feature encoder with a pre-training model, and RVQ. Then comes the fine-tuning training. Due to the autoregressive nature of the GPT-like model, the tokens that follow in the inference will always in$
$like model to complete the tokens that contain the reference timbre, and use the VITS decoder to reconstruct the audio for the rest$
$tuned trained GPT-like, SoVITS model to perform TTS inference processing on the preprocessed text.$

After generating DeepFake audio and video respectively, the source video and audio are preprocessed by tools such as FFmpeg, including the separation of video and audio, the slicing of audio, and the acquisition of break intervals. In GPT-SoVITS, the corresponding target audio is generated based on the sliced audio and merged with the break interval to generate the final audio. Finally, the target video and audio are synchronised and post-processed with image enhancement to ensure optimal output quality.

In summary, our overall design framework covers the steps of face detection, feature point localisation and alignment, face segmentation and feature fusion, face swapping, cutting through audio, automatic speech recognition and fine-tuning training. Through these processes, we are able to generate high-quality target video and audio, and finally achieve their simultaneous output with image enhancement processing.

4.2 experimental environments

In order to conduct Deepfake-based audio/video synchronisation experiments, we set up two separate experimental environments suitable for the task. The following are the detailed configurations of these two environments:

5 Experimental results and analysis

5.1 Video face-swapping results and quantitative assessment

In this experiment, in order to evaluate the effect of video face-swapping, we randomly selected 10 frames from the source video and the generated video at the same time point for comparison and analysis. As shown in Fig. 5.1, from the comparison of the source and target frames, it can be observed that the action and expression features maintain a high degree of consistency between the two, and also show good consistency in terms of lighting and contrast.

To quantify this visual similarity, we assessed it using the Structural Similarity Index (SSIM), a measure of the visual similarity of two images that consists of the following three main factors of comparison:

1. Brightness comparison: average brightness of two images.

Figure 4: Comparison of random source and target frames.

2. Contrast comparison: standard deviation of two images.

3. Structural comparison: covariance of two images.

The SSIM value analysis shows that the SSIM value of the target face after face-swapping is 0.50565, which indicates that the generated face is moderately similar to the target face structurally, although there may be problems with loss of detail, lighting variations and contrast adjustment. The SSIM value of the source and target faces after face-swapping is 0.89086, which indicates that the source and target faces are very similar in structure, verifying the high success of the face-swapping effect, and suggesting that the source facial features can be accurately adjusted to match the target facial features.

By comparing the random frame images of the source video and the generated video, it was observed that the movement and expression features maintained a high degree of consistency between the two, and also showed good agreement in terms of lighting and contrast. Quantitative evaluation using the Structural Similarity Index (SSIM) shows that the structural similarity between the target face and the source face is moderate, while the structural similarity between the source and the target face is very high. We can see that both the face-swapping results and quantitative assessment of the video verify the success of the face-swapping effect.

5.2 synthesis

At the technical level, we evaluated the project through a combination of synchronisation checks and analysis of the quality of the audio and video editing. In the synchronisation check, although the audio waveforms were largely aligned with the keyframes of the character's mouth movements in the video, we found that the generated audio speech was slightly slower in the end part of the video, resulting in incomplete alignment of the sound and picture. We corrected this using FFmpeg, which resulted in some degree of synchronisation correction. In terms of audio and video editing quality, the face-swapping effect was more consistent in terms of skin colour, lighting, shadows and edge backgrounds, and the resolution of the face-swapped portion matched the rest of the video, with no over-pixelated or blurred areas, and generally appeared relatively natural. On a perceptual level, we assessed the realism of the face-swapped video and audio through blind testing and collecting viewer feedback. According to audience feedback, most people were able to quickly recognise the replaced target characters and gave high ratings on the similarity of the audio. However, the emotional expression of the audio was slightly undermatched with the facial expressions of the characters in the video, and there is still room for improvement in the integration with the real situation.