Speech-to-Face Conversion Using Denoising Diffusion Probabilistic Models

Shuhei Kato¹, Taiichi Hashimoto¹
¹RevComm Inc., Japan
{kato, taiichi.hashimoto}@revcomm.co.jp

Abstract

Speech-to-face conversion is the task of generating face images from speech signals. Many studies have been conducted to address this task, and achieved good performances. In this paper, we introduce denoising diffusion probabilistic models (DDPMs) to generate face images instead of generative adversarial networks (GANs) or autoencoders, which are used in most of the prior studies. Moreover, unlike prior studies, several components of our system are designed to use high-resolution face image datasets instead of audio-visual paired data. As a result, our system can generate high-resolution face images from speech signals with an architecture that is simpler and more flexible than the ones used in prior studies. In addition, introducing DDPMs enables us to utilize techniques that control outputs of DDPMs or improve performance of them in succeeding studies.

Index Terms: speech-to-face, denoising diffusion probabilistic model, classifier-free guidance

1. Introduction

Humans have the ability to imagine how a speaker looks when they hear a voice but cannot see the person [1, 2]. This relationship between speech and appearance is partially obvious, considering a speech signal conveys various speaker attributes such as age and gender [3]. Recent developments in machine learning have enabled researchers to take on the challenge in generating a human’s face solely from his or her speech while keeping the speaker’s attributes [4, 5, 6, 7, 8, 9, 10, 11]. Such a speech-to-face conversion technique can be applied, for example, to generating virtual face images reflecting speakers’ attributes on the phone apps, or generating avatars for entertainment purposes or privacy protection.

Most of prior studies use generative adversarial networks (GANs) [12] or autoencoders, and have achieved good performance. Despite their success, because audio-visual paired data are needed to train the whole system, the output image resolution is limited.

In the field of image generation, denoising diffusion probabilistic models (DDPMs) [13, 14] have achieved state-of-the-art performance in recent years [15, 16, 17, 18, 19]. DDPMs tend to have the ability to generate samples that are more diverse than those of GANs and autoencoders while maintaining or improving quality, and can be stably trained with stationary loss functions. In addition, we can use simple and flexible methods to train conditional models [17, 20].

In this paper, we propose a novel method of speech-to-face conversion using DDPMs and investigated how well generated face images reflect the speakers’ attributes. Generated image samples by our system and prior studies are shown in Fig. 1. The key contributions of this paper are as follows:

• We introduce DDPMs to speech-to-face conversion to replace GANs or autoencoders. DDPMs enable our system to generate high-resolution face images with a simpler and more flexible architecture. Introducing DDPMs also enables us to apply techniques for controlling outputs of DDPMs or improving performance of them in succeeding studies.

• In contrast to prior studies, audio-visual paired data is only required to train the speech encoder, not the whole system. To train the face decoder and super-resolution model, we can use high-resolution face image datasets, which are more effortlessly available and easier to build.

Figure 1: Samples of generated face images generated from speech. (a): Generated by a GAN-based decoder (128 × 128). (b): Generated by an autoencoder-based decoder (128 × 128). (c): Ours (DDPM-based, each image is 512 × 512). Samples (a) and (b) are referenced from original papers.
2. Related work

2.1. Speech-to-face conversion

Speech-to-face conversion is the task of generating face images from speech signals while keeping the speakers’ attributes. Oh et al. [5] use a voice encoder that predicts an embedding that is decoded to a normalized face image by an autoencoder-based pre-trained face decoder [21]. The voice encoder is trained to predict embeddings as close to the corresponding face embeddings produced by a pre-trained face recognizer using a paired dataset of speech and face images. Instead of generating normalized face images, as in Oh et al. [5], most subsequent studies generate (non-normalized) face images using GAN-based [6, 7, 9, 11] or autoencoder-based [8, 10] decoders that are trained simultaneously with the other components.

2.2. Denoising diffusion probabilistic models

DDPMs [13, 14] are a class of latent variable models that convert Gaussian noise into samples matching a data distribution via a finite iterative denoising process. Conditional models are possible, for example on class labels, text, low-resolution images, or spectrograms [15, 17, 22, 23, 18, 24, 25, 26]. Let $x_1, \ldots, x_T$ be a sequence of latents with the same dimension as the data $x_0 \sim q(x_0)$. The forward diffusion process is defined by a variance schedule $\beta_1, \ldots, \beta_T$ and condition $c$ as

$$ q(x_{1:T} \mid x_0, c) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}, c) \tag{1} $$

where

$$ q(x_t \mid x_{t-1}, c) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I) \tag{2} $$

The reverse backward denoising process is defined as

$$ q(x_{0:T-1} \mid x_0, c) = \prod_{t=1}^{T} q(x_{t-1} \mid x_t, c) \tag{3} $$

and the models are trained to minimize the objective

$$ \mathbb{E}_{x_0, c, \epsilon \sim \mathcal{N}(0,1), \epsilon \sim \mathcal{U}(1 \ldots, T)}[\epsilon - \epsilon_\theta(x_t, c)]^2 \tag{4} $$

$$ = \mathbb{E}_{x_0, c, \epsilon, \ell}[\epsilon - \epsilon_\theta(\sqrt{\alpha_t}x_0 + \sqrt{1 - \alpha_t}\epsilon, c)]^2 \tag{5} $$

where $\epsilon_\theta$ is an estimator to predict $\epsilon$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Classifier-free guidance [20] is a simple and flexible method for training conditional DDPMs. In this method, DDPMs are trained on conditional and unconditional objectives by randomly (e.g., with a 10% probability) dropping $c$ during training. When sampling, we use the following linear combination of conditional and unconditional $c$-estimators

$$ \tilde{\epsilon}_\theta(x_t, c) = (1 + w)\epsilon_\theta(x_t, c) - w\epsilon_\theta(x_t) \tag{6} $$

where $\epsilon_\theta(x_t, c) = \epsilon_\theta(x_t, c = 0)$ and $w$ is the guidance weight. Setting $w = 0$ means no guidance, and larger $w > 0$ intensify the effect of guidance, resulting in improving sample quality while reducing diversity.

DDPMs have outperformed other state-of-the-art methods recently on image generation tasks [15, 16, 17, 18, 19]. In particular, DDPMs have achieved success on conversion tasks, such as text-to-image [18, 27], text-to-video [28, 29], image-to-image [30], text-to-audio [24], spectrogram-to-waveform [25, 26], and text-to-waveform [22, 23] conversions. In this paper, we apply a text-to-image conversion technique in our speech-to-face conversion system. Specifically, our system generates face images with the guidance of speech embeddings.

3. Model architecture and training

3.1. Model architecture

The proposed model consists of three sequentially connected components: a speech encoder, face decoder, and super-resolution model (Fig. 2). Through this pipeline, an input spectrogram of speech data is converted into a face image.
First, the speech encoder converts an input spectrogram into a speech embedding. We use a convolutional-neural-network-based architecture almost the same to the one used in Oh et al. [5]. Second, the face decoder generates a 64 × 64 face image from Gaussian noise with the guidance of a speech embedding predicted by the speech encoder. The decoder is based on a DDPM, the backbone of which is a time-conditional U-Net [33], conditioned by speech embeddings using the classifier-free guidance. Finally, the super-resolution model upsamples a 64 × 64 face image generated by the face decoder to a 512 × 512 one. We take the same super-resolution architecture used in Saharia et al. [18].

This pipeline architecture enables us to use a face image dataset instead of an audio-visual dataset to train the face decoder and super-resolution model, resulting in generate images in higher resolution than those of prior studies.

3.2. Training

We train the speech encoder, face decoder, and super-resolution model separately.

The speech encoder is trained with an audio-visual dataset such as videos of people talking (Fig. 3). The speaker’s face is cropped from a single frame of each video, and the corresponding 512-dimensional face embedding, \( \mathbf{v}_F \), is extracted via the ArcFace [31] network. The speech encoder receives an input spectrogram and converts it into a speech embedding \( \mathbf{v}_S \), which is expected to approximate \( \mathbf{v}_F \) through the training. We choose the cosine distance between \( \mathbf{v}_F \) and \( \mathbf{v}_S \) as the loss function.

The face decoder and super-resolution model are trained with a high-resolution face image dataset. For the training of the face decoder (Fig. 4), the ArcFace embedding is extracted from each face image. Then the face decoder learns the data distribution of 64 × 64 face images with the guidance of the corresponding face embeddings. For the training of the super-resolution model, each face image is resampled at a 64 × 64 and 512 × 512 image pair, and the model is trained to predict 512 × 512 images from the corresponding 64 × 64 images.

3.3. Implementation details

We select the AVSpeech dataset [34], a large-scale audio-visual dataset from YouTube videos, as the training dataset for the speech encoder. In the manner of Oh et al. [5], up to 6 seconds of audio taken from the beginning of each video in the AVSpeech is transformed into a spectrogram and fed into the speech encoder. If the duration of the video is less than 6 s, the audio is repeated so that it becomes at least 6 s long. All the audio samples are resampled at 16 kHz and converted to single channel samples. Spectrograms are calculated using a short-time Fourier transform with a 25-ms Hann window, 10-ms hop length, and 512 frequency bands. Then, both real and imaginary parts of each spectrogram \( \mathbf{s} \) are independently compressed as \( \text{sgn}(\mathbf{s}) |\mathbf{s}|^{0.2} \), where \( \text{sgn}(\cdot) \) denotes the sign function. We use MediaPipe\(^1\) to detect and crop faces from the frames of the AVSpeech videos. Only the frames containing a single face are used for training and testing. As a result, the training, validation, and test sets for the speech encoder have 102 k, 11 k, and 8.2 k spectral-face embedding pairs, respectively. The speech encoder is optimized by LAMB [35], the initial learning rate is \( 1 \times 10^{-3} \) with an exponential decay of 0.95 at every 10,000 steps, the batch size is 128, and the speech encoder is trained for 430 k steps.

For the training of the face decoder and super-resolution model, we choose the Flickr-Faces-HQ Dataset [36], a large-scale 1024 × 1024 face image dataset of Flickr photos, as the training dataset. We also use the MediaPipe to detect and crop faces. As a result, the training and validation sets have 60 k and 6.6 k face images, respectively. The hyperparameters for the face decoder and super-resolution model are described in Table 1.

4. Evaluation

4.1. Facial attribute evaluation

To assess how well our models capture facial attributes, we compare the age and gender estimated by inaFaceAnalyzer\(^4\) of the generated images and ground-truth ones. As shown in Table 2, larger values of the classifier-free guidance weight \( w \) tend to increase similarity, as is expected considering the function of the classifier-free guidance. Confusion matrices for each of the attributes when the value of \( w \) is 4.0 are shown in Fig. 5. As can be seen, the classification results for both age and gender have a certain correlation.

\(^1\)https://google.github.io/mediapipe/

\(^4\)https://github.com/ina-foss/inaFaceAnalyzer
Table 1: Hyperparameters for the face decoder and super-resolution model.

<table>
<thead>
<tr>
<th>Hyperparameter</th>
<th>Face decoder</th>
<th>Super-resolution model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusion steps</td>
<td>2,000</td>
<td>2,000</td>
</tr>
<tr>
<td>Noise schedule</td>
<td>Cosine</td>
<td>Cosine</td>
</tr>
<tr>
<td>Input channels</td>
<td>256</td>
<td>64</td>
</tr>
<tr>
<td>Number of ResNet [37] blocks</td>
<td>3, 3, 3, 3</td>
<td>1, 1, 2, 4, 8</td>
</tr>
<tr>
<td>Channel multiples</td>
<td>1, 2, 3, 4</td>
<td>1, 2, 4, 8, 16</td>
</tr>
<tr>
<td>Cross-attention heads</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Self-attention heads</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Self-attention resolution</td>
<td>2, 4, 8</td>
<td>16</td>
</tr>
<tr>
<td>Drop rate for classifier-free guidance</td>
<td>0.1</td>
<td>-</td>
</tr>
<tr>
<td>Number of groups for group normalization [38]</td>
<td>32</td>
<td>16</td>
</tr>
<tr>
<td>Optimizer</td>
<td>LAMB</td>
<td>LAMB</td>
</tr>
<tr>
<td>Batch size</td>
<td>64</td>
<td>128</td>
</tr>
<tr>
<td>Learning rate scheduling</td>
<td>Linear</td>
<td></td>
</tr>
<tr>
<td>Peak learning rate</td>
<td>$2 \times 10^{-4}$</td>
<td>$5 \times 10^{-5}$</td>
</tr>
<tr>
<td>Iteration steps</td>
<td>1.9 M</td>
<td>7.4 M</td>
</tr>
</tbody>
</table>

Table 2: Correlation coefficients and accuracy for the estimated age and gender of generated images and ground-truth ones for various values of classifier-free guidance weight $w$.

<table>
<thead>
<tr>
<th>$w$</th>
<th>Corr. coeff. (age)</th>
<th>Accuracy (gender)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.22</td>
<td>0.83</td>
</tr>
<tr>
<td>1.0</td>
<td>0.22</td>
<td>0.85</td>
</tr>
<tr>
<td>2.0</td>
<td>0.24</td>
<td>0.85</td>
</tr>
<tr>
<td>3.0</td>
<td>0.24</td>
<td><strong>0.86</strong></td>
</tr>
<tr>
<td>4.0</td>
<td><strong>0.25</strong></td>
<td>0.85</td>
</tr>
</tbody>
</table>

4.2. Craniofacial attribute evaluation

We also compare craniofacial attributes. The selected measurements are commonly used ones in the literature to capture ratios and distances in the face\(^{3}\), and are used in Oh et al. [5]. Because both the generated images and ground-truth ones have various facial angles and sizes, we normalize them using Hsu et al.’s method [39] to obtain the frontal face images, and extract attributes from them using the dlib library\(^{4}\). The correlation coefficients of the attributes of the generated images and ground-truth ones are listed in Table 3. As found in Oh et al. [5], the “nasal index” has the largest correlation, albeit our value is smaller. In contrast to the face attribute results, there is no tendency in the results when the value of $w$ is changed.

4.3. Limitations of our system

From the perspective of similarity of generated face images to the ground-truth ones, our system has a limitation compared to several existing studies [5, 6, 7, 9, 11]. This is partially because it is practically impossible to use the information from generated face images or the face decoder to train the speech encoder, as do most of these prior studies, as the inference speeds of DDPMs are much slower than the speed of other models such as GANs and autoencoders. Recent improvements in accelerating inference speeds of DDPMs [40, 41] might help alleviate this problem. Although the similarity is not the main topic of this paper, we could select a more sophisticated audio representations and loss function that does not use any information from other components, such as the ones used in Hong et al. [9] to improve performance. We could of course train the face decoder to use ground-truth or any value of age, gender, or any other attributes such as expression, clothes, as guidance and control generated images if we desired to do so during inference. Such flexibility is an advantage of our approach over the prior studies.

5. Conclusions

We proposed a novel method for speech-to-face conversion using DDPMs. Our system comprises a speech encoder, face decoder DDPM, and super-resolution DDPM, and they are separately trained using an audio-visual dataset (speech encoder) or a high-resolution face image dataset (face decoder and super-resolution model) while prior studies need to be train the whole system using an audio-visual dataset. Although there is a limitation on the similarity of the generated images, we succeeded in building a simple and flexible speech-to-face conversion system based on DDPMs that can generate high-resolution face images. We expect our system to be an stepping stone to apply techniques for controlling outputs or improving performance in succeeding speech-to-face conversion studies.

---

6. References


