Zero-Shot Face-Based Voice Conversion: Bottleneck-Free Speech Disentanglement in the Real-World Scenario

Shao-En Weng, Hong-Han Shuai, Wen-Huang Cheng
National Yang Ming Chiao Tung University
anita4213.ee09@nycu.edu.tw, hhshuai@nycu.edu.tw, whcheng@nycu.edu.tw

Abstract

Often a face has a voice. Appearance sometimes has a strong relationship with one's voice. In this work, we study how a face can be converted to a voice, which is a face-based voice conversion. Since there is no clean dataset that contains face and speech, voice conversion faces difficult learning and low-quality problems caused by background noise or echo. Too much redundant information for face-to-voice also causes synthesis of a general style of speech. Furthermore, previous work tried to disentangle speech with bottleneck adjustment. However, it is hard to decide on the size of the bottleneck. Therefore, we propose a bottleneck-free strategy for speech disentanglement. To avoid synthesizing the general style of speech, we utilize framewise facial embedding. It applied adversarial learning with a multi-scale discriminator for the model to achieve better quality. In addition, the self-attention module is added to focus on content-related features for in-the-wild data. Quantitative experiments show that our method outperforms previous work.

Introduction

Voice conversion (VC) (Qian et al. 2019; Kaneko and Kameoka 2018; Chen et al. 2021; Lin et al. 2021) changes the voice characteristics of a source speaker to a target speaker while conserving linguistic information. To do the conversion, how to disentangle the acoustic and linguistic information is a thorny problem. Once having a good disentangle strategy, the model can generate a high quality of speech from the given utterance and style. A successful VC can be applied to various fields, such as personal electrical support as an audio assistant (Lu et al. 2021), entertainment usage for dubbing (Mukhneri, Wijayanto, and Hadiyoso 2020), and industrial applications for voice changers, etc.

With the advancement of deep learning, a recent line of study focuses on solving the VC task using data-driven approaches (Sisman et al. 2021). Some works use the bottleneck strategy (Qian et al. 2019; Lu et al. 2021), vector quantization (VQ) (Wu, Chen, and Lee 2020), cycle-consistent generative adversarial network (Kaneko and Kameoka 2018), instance normalization (Chen et al. 2021), etc. They are eager to find a good solution for speech disentanglement. With the bottleneck method, the size of the bottleneck should be chosen carefully, as should the VQ-based method. With cycle-consistent training, style diversity is bounded by the training domain. However, most research studies are based on laboratory data, which is quite different from speech in the real-world scenario. There is still a gap between research and real-world application.

In addition, we argue that there should be other ways, in addition to voice, to control the style of speech, such as the face. Therefore, in this work, we focus on the synthesis of speech from a given facial image. That is, a face-based voice conversion (Lu et al. 2021). Face-based VC can be applied in numerous applications. For example, for movie or animation dubbing, the face of the character can help to generate a more suitable voice matching audiences' minds; for personal audio assistance, one can change the voice by giving a facial image without collecting the vocal records. Meanwhile, not only is it an interesting topic, but it is also an important issue for how to deal with the cross-modal problem.

To perform face-based voice conversion, there are two challenges: face-speech transformation and learning how to use the in-the-wild data. First of all, since the face-speech transformation is cross-domain learning, there is some redundant facial information for voice, i.e., winking or face angle, etc. This information perturbs the mapping process from the facial to the acoustic domain, so the model tends to synthesize a general voice for each utterance to give the optimal result. Second, existing datasets that contain face and speech are in-the-wild datasets, that is, not recorded in a laboratory environment and not with expert record equipment. These data include some background noise or echo and thus are of lower quality than the clean dataset. Moreover, unlike the clean data recorded in the specific corpus, in-the-wild data do not contain duplicate sentences and have various accents. These also make it difficult to train voice conversion with in-the-wild data or for real-world applications.

To solve these problems, FaceVC (Lu et al. 2021) proposed a three-stage model. Models are based on the backbone of AutoVC (Qian et al. 2019), which is trained by providing a suitable size of content embedding for the decoder to limit the received information; therefore, it can disentangle speech. This technique is called bottleneck adjustment. With this skill, a good disentanglement of the speech can convert the voice. For high-quality data synthesis, FaceVC trains its model on clean and in-the-wild datasets, respectively. The model trained with in-the-wild data is used to teach the face encoder to produce a speaker-related embedding; the model trained with clean data is to be used as a reference generator. Then, they used a fully connected layer to warp the facial distribution to the acoustic distribution.

However, it takes at least twice as long to make the bottleneck adjustment to learn the facial and acoustic distribution. Furthermore, since the distribution of in-the-wild and lab-collected data are varied from each other, it cannot reflect the actual mapping space for face and voice. Moreover, due to the fact that FaceVC trains the content encoder on clean data, in regard to speech in the real-world scenario, the quality degrades as much as unseen words and noise.

To better address the challenges, we propose a novel method, called SP-FaceVC, that is bottleneck-free by changing the data preprocessing to avoid the difficult bottleneck adjustment. Specifically, since speech is composed of the frequency response of the vocal track and the glottal pulse, it can be passed by a lowpass liftering [1] on the cepstrum [2] to obtain content-related features (in this work, we call it SP). We adopted SP as input data instead of the entire Mel spectrogram, so the speech can be easily disentangled without bottleneck adjustment.

For the second challenge of face-speech transformation, we train our model directly on the in-the-wild data to prevent the facial distribution from fitting into the many-to-one acoustic distribution. Here, we obtain the facial embedding by taking the arithmetic mean of all frames instead of choosing only one frame. This can eliminate the bias of one facial photo from making the general style. Moreover, we leverage the reparameterization trick, which transfers the data distribution into a Gaussian distribution by sampling a hyperparameter from a Normal distribution, to operate for unseen speakers. However, sampling for reparameterization may make the generated speech with unstable timbre or low quality, i.e., the speech may sound like a woman at the beginning and a man afterward. Therefore, we propose a multi-scale discriminator to teach the generator by differentiating the good- and bad-quality speech, which makes the speech more natural for zero-shot voice conversion without multi-stage learning. Moreover, for adoption in the real-world scenario with in-the-wild data, we add a self-attention module on the content encoder to aid it in paying attention to the content-only features to improve the speech quality. Since noise is not related to the context, the attention module can learn which part is highly related to the content through the training process.

Summarize our contributions:

1. We propose a novel bottleneck-free VC for easy disentanglement.
2. We eliminate the bias that causes general style by averaging frame-level facial features and achieve style for unseen speakers with the reparameterization trick and a multi-scale discriminator.
3. We adopt a self-attention module for the content encoder to apply in the real-world scenario with noisy data.

Related Work

Voice Conversion

A recent line for the voice conversion of non-parallel training data can be categorized mainly into autoencoder-based approaches (AE-based) (Qian et al. 2019; Chen et al. 2021; Lin et al. 2021) and generative-adversarial-based approaches (GAN-based) (Kaneko and Kameoka 2018; Kameoka et al. 2018). AE-based methods usually use the bottleneck to disentangle speech. AgainVC (Chen et al. 2021) further uses instance normalization to separate global and temporal information. However, how to select a suitable size of the bottleneck still relies on experiments. Moreover, the quality of speech for using adaptive instance normalization is relatively low, i.e., the converted speech might contain the original speaker sound, especially when the styles of two utterances are far from each other. This is because the temporal information might still contain parts of acoustic information with instance normalization. For GAN-based methods and their subsequent works, such as CycleganVC (Kaneko and Kameoka 2018), the discriminator judges the quality of the synthesized speech only within the training domain style. The scalability and robustness of handling more than two domains using cycle consistency learning are limited.

Cross Modality Learning

In addition to FaceVC (Lu et al. 2021) mentioned in the introduction, several works also leverage the high correlation between voices and faces for different applications. For example, Speech2Face (Oh et al. 2019) discusses this idea and proposes a cross-modal learning method to synthesize faces from voices. The work mainly focuses on how to make the voice encoder learn the multimodal relationship with knowledge distillation loss. Additionally, Face2Speech (Goto et al. 2020) inputs facial images and text to generate speech. It first trains a general text-to-speech (TTS) model and then trains a face encoder, trying to map the facial style to that of the speech. Taking a similar thought, HearingFace (Plüster et al. 2021) trains a face-based TTS with style transfer learning from the facial to acoustic domain. Taking as input the speech, text transcript, and facial characteristic, FR-PSS (Wang et al. 2022) applies the prior information and takes advantage of the idea of residual to remove the main similar part of the speech, as a result of highlighting the subtle changes that could be caused by the facial characteristics. However, except for FaceVC, all other works study the text as an input feature to avoid disentanglement.

Methodology

To achieve the one-to-one face-based voice conversion in the real world scenario, we introduce the proposed model and the training strategy in the following. The input data preprocessing is first presented to alleviate the difficult disentanglement problem for voice conversion and associate the cross-modality. Then, we present the detailed model structure that is trained on in-the-wild dataset.

[1] Just like the filter, but by adopting on cepstrum.

[2] Cepstrum is the thing that takes the inverse discrete Fourier transform from the Mel spectrum.

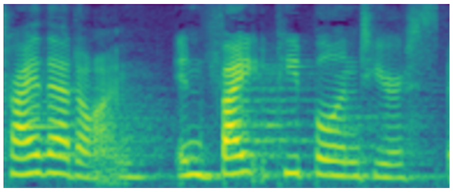

Figure 1: The demonstration of original Mel spectrogram.

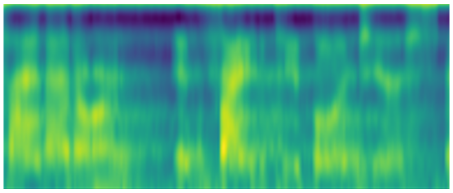

Figure 2: The demonstration of a Mel spectrogram being passed by a low-pass lifter (SP).

Data Processing

Preprocessing of Facial Feature We leverage FaceNet (Schroff, Kalenichenko, and Philbin 2015) to generate speaker embeddings from the images. To make the speaker embedding more representative, frame-wise features obtained from an utterance are taken as the arithmetic mean. The means of speaker embedding serve as the speaker style in an utterance. This way, redundant characteristics can be removed, and it reduces the possible bias caused by sampling from one facial image.

Preprocessing of Acoustic Feature If we perform the inverse discrete Fourier transform (F^{-1}) to convert the Mel spectrogram into the cepstrum, the data can be divided into low-quefrency and high-quefrency[3] as follows.

$$Cepstrum = Low(F^{-1}[Mel]) + High(F^{-1}[Mel]), \quad (1)$$

where Low(-) and High(-) are the functions of low-pass liftering and high-pass liftering, respectively. The theoretical basis is from audio signal processing (Murphy and Akande 2007). Low-quefrency components of a cepstrum represent content-related features. On the other hand, the high-quefrency components of a cepstrum represent style features, such as timbre, formants, relative phonemes, etc. Therefore, we take the low-quefrency part and transform it back into a speaker-independent Mel spectrogram SP as follows.

$$SP = F[Low(F^{-1}[Mel])] \quad (2)$$

SP replaces the Mel spectrogram as the input for the model. Figures 1 and 2 demonstrate the Mel spectrogram and the parts after processing by a low-pass lifter, respectively. As we can see, only the energy part and the contour of the linguistic information remain, and the frequency about the pitch disappears after processing. Equipped with this preprocessing, the model can better learn the synthesis process without learning to disentangle the speech, which simplifies the training process.

Model Architecture

Our goal is to learn a generator that can synthesize natural, cross-modal, and style-diverse speech from in-the-wild data. Since the target is not an audio, but an image, unlike a normal speech synthesis technique, in our case, simplifying data distribution by providing more conditional information (Ren et al. 2022; Choi et al. 2021; Lee et al. 2021), such as pitch or energy, cannot be used. We introduce a discriminator-guided network to teach the decoder how to learn cross-modal information to generate high-quality speech. The general structure of the model is shown in Figure 3. The following describes the structure in detail.

Content Encoder Previous work (Lu et al. 2021; Qian et al. 2019) used content encoders to extract the content by providing a speaker embedding and a whole Mel spectrogram, which might confuse the encoder since the acoustic information is encoded. Since the proposed SP-FaceVC takes SP as input, which has already removed the speaker-related signal, it is easier for the content encoder E_{cont} to obtain the embedding of the content $e_{cont.}$ as follows.

$$e_{cont.} = E_{cont}(SP). \quad (3)$$

However, the preprocessed SP could still contain some undesired signals, e.g., background noise or some channel disturbance in the original audio caused by the record device. Therefore, through the training process, a self-attention module is used to learn context-related characteristics from thousands of utterances. The module enhances the linguistic information and minimizes potential noise from the given in-the-wild data. We also add a mask before the self-attention module to help train the model about how to find the relationship between context and context.

Speaker Encoder Once the facial embedding is obtained, it serves as input to the speaker encoder. The speaker encoder is made up of fully connected layers and an activation function as ReLU, which are followed by the reparameterization trick. The reparameterization trick can turn the embedding into a Gaussian distribution. This way, when we input an unseen facial image, the model will find a suitable distribution to represent the speaker style instead of just a point.

Decoder and Postnet The goal of the decoder is to integrate the content embedding from the source speaker and the style embedding from the facial image for face-based voice conversion. For the speech synthesis task, since the embeddings of speakers are often closely related to their utterances, traditional speaker extraction approaches (Dehak et al. 2010; Snyder et al. 2017) take the characteristics of every frame of speech to accurately represent the style of speech. However, a facial image cannot represent this kind of information. It is difficult for the model to map a visual domain distribution to an acoustic domain distribution.

To solve this problem, the SP-based content embedding $e_{cont.}$ and the reparameterized style embedding e_{style} are

[3] Quefrency is the unit of the cepstrum.

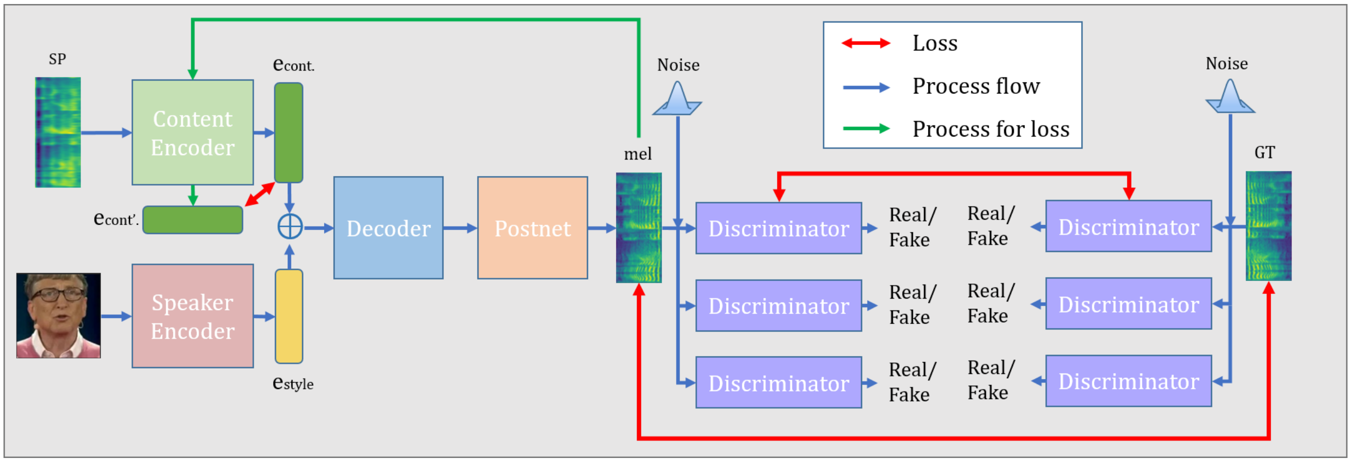

Figure 3: The model structure and training process. For simplify the loss flow in the figure, only the reconstruction losses and feature matching loss are shown. The losses for adversarial training are not shown in the figure.

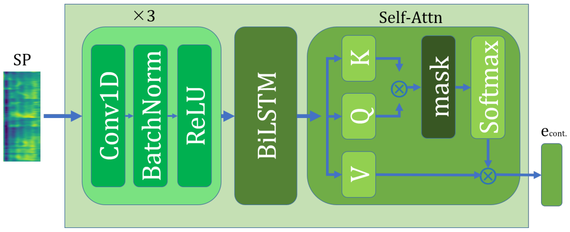

Figure 4: The architecture of the content encoder. ⊗: matrix product.

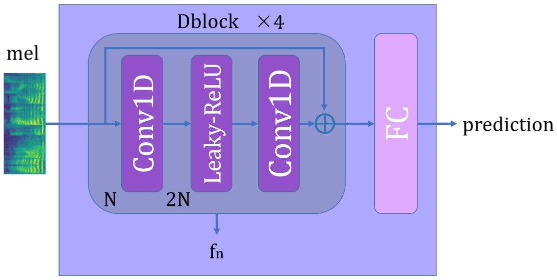

Figure 5: The architecture of the sub-discriminator.

first concatenated as input to the decoder Dec. Since the embedding of SP-based content is independent of acoustic information, the decoder is forced to learn the style elements from the reparameterized style embedding. Therefore, it can transform the style distribution to the acoustic distribution. Afterward, a postnet P is used to improve speech quality.

Multi-Scale Discriminator Synthesized speech sometimes has distortion and blurriness because it tends to be over-smooth (Ren et al. 2022) with a simple autoencoder architecture. To solve this issue, we propose a multi-scale discriminator to help the decoder generate better results. The discriminator judges whether the synthesized quality of speech is good or not. Since audio has a different structure at each level (Kumar et al. 2019), the discriminator consists of three sub-discriminators. The architecture of the sub-discriminator is shown in Figure 5. Each sub-discriminator aims to distinguish different scale data. Therefore, we use average pooling with kernel size 3 to downsample the Mel spectrogram. Moreover, we add noise to the input data to increase the data complexity and stabilize the discriminator. A residual connection and spectral normalization are used in each Dblock $DB(\cdot)$ for optimization.

Vocoder Since the Mel spectrogram is used as input data, we need a vocoder to generate a time-domain waveform from the given spectrograms. We use a flow-based model, Waveglow (Prenger, Valle, and Catanzaro 2019) as our vocoder due to the high quality result and efficiency as WaveNet (Oord et al. 2016), the vocoder used in AutoVC and FaceVC.

Training Strategy

Reconstruction Loss To avoid the decoder from generating high-quality speech but ignoring the conditions (that is, given a low-pitched face style but generating a high-pitched voice style), two reconstruction losses are used as in 5, 6:

$$\hat{y} = P[Dec(e_{cont.}, e_{style})] \tag{4}$$

$$L_{cont} = \|E_{cont}(y) - E_{cont}(\hat{y})\|_1 \tag{5}$$

$$L_{recons} = \|y - \hat{y}\|_1 \tag{6}$$

y: the ground truth Mel spectrogram

Adversarial Loss We use the least squares formulation in LSGAN (Mao et al. 2017) for adversarial training, since it gets better performance than hinge loss for audio generation according to MelGAN (Kumar et al. 2019). To discourage sub-discriminators D_k from being too confident, label smoothing is performed. Instead of setting the real label to 1, 0.9 is set. Discriminators and generator ($G = E_{cont.}, E_{style}, Dec, P$) are trained by the following loss:

$$\min_{D_k} \mathbb{E}\left[\|D_k(y) - 0.9\|_2 + \|D_k(\hat{y})\|_2\right], \forall k = 1, 2, 3 \tag{7}$$

$$L_{adv} = \mathbb{E}\left[\sum_{k=1}^{3} \|D_k(\hat{y}) - 0.9\|_2\right] \tag{8}$$

Feature Matching Loss To combat mode collapse, a loss of matching of characteristics for the output feature of the Dblock is used (Lee et al. 2021) to improve the discriminator by learning more representative characteristics.

$$L_{fm} = \mathbb{E}\left[\sum_{k=1}^{3}\sum_{i=1}^{4} \|DB_k^i(y) - DB_k^i(\hat{y})\|_1\right], \tag{9}$$

where DB_k^i is the i-th Dblock for the k-th sub-discriminator.

The overall loss of updating generator is with the following objective:

$$\min_G L = \alpha L_{adv} + \beta L_{fm} + \gamma L_{recons} + \delta L_{cont} \tag{10}$$

where α, β, γ and δ are the hyperparameters controlling the importance of each term. Empirically, we set $\alpha = 1$, $\beta = 0.1$, $\gamma = 100$, and $\delta = 0.1$. Due to a good speech disentanglement and framewise facial embedding, the model with self-attention and reparameterization modules can be adversarially trained end-to-end. The generator and the discriminator are trained one-by-one. Following the previous work, the small batch size is used to generate a good speech. Here, we set the batch size to 2.

Experiments

Dataset

LRS3 LRS3 (Afouras, Chung, and Zisserman 2018) dataset is collected from TED and TEDx videos downloaded from YouTube. To extract facial images, a face alignment (Bulat and Tzimiropoulos 2017) is first used to detect whether there is a face in the given frame. If so, the image would be cropped and sent through MTCNN (Zhang et al. 2016), a face detection network. Here, the margin parameter is set to 50 in MTCNN.

Evaluation Criteria

NISQA The non-intrusive speech quality assessment method (Mittag et al. 2021) mimics the mean option score (MOS) evaluation.

MCD Mel-cepstral distortion (Kubichek 1993) is a measurement to tell the difference of two mel-frequency cepstrums. The utterance-wise speeches are taken to evaluate the result. The unit is dB.

Cosine Similarity The speaker embedding of the speech generated from the same speaker are calculated.

Except for MCD, the higher the values are, the better.

Experiment Setup

The training speakers in the LRS3 dataset are randomly selected. We choose 100, 200, and 400 speakers to evaluate the performance of the style stability of our model. For fast convergence, the training set of 100 speakers is used for ablation studies. All audios were first sampled at 22050HZ. For the generator and discriminator, we use ADAM (Kingma and Ba 2015) as optimizers. The learning rate for the generator is set to 0.0001. For the discriminator, it is set to 0.0004 and with $\beta_1 = 0.9$, $\beta_2 = 0.999$.

Model	F2F	M2M	M2F	F2M	Avg.
GT	-	-	-	-	3.294
FaceVC	2.431	2.359	2.317	2.366	2.368
Ours	2.474	2.326	2.427	2.470	2.424
Ours w. M.	**2.483**	**2.636**	**2.635**	**2.572**	**2.582**

Table 1: MOS quality of unseen speech evaluated by NISQA. GT: ground truth speech. F2F: female-to-female; M2M: male-to-male; M2F: male-to-female; F2M: female-to-male; Avg.: average scores for the four terms; M.: mask

Model Comparison To demonstrate the power of our model for real-world data, we compare it with a baseline model and ground-truth speech. Since FaceVC is the first and only work for face-based VC, it is adapted as our baseline model. We use the pre-trained model provided by the authors. As shown in Table 1, since ground truth speech comes from the in-the-wild dataset, the MOS quality provided by the NISQA model is only 3.294. The overall quality of FaceVC is only about 2.368. Our speech quality, which is 2.424, is beyond FaceVC. If we add the mask before self-attention module, the MOS score achieves to 2.582.

Furthermore, the other important factor for face-based VC is the stability of styles for cross-modal learning. To demonstrate style diversity and stability, we randomly sample 14 unseen utterances from 7 female and 7 male speakers to be the source utterances and randomly select 4 unseen female faces and 4 unseen male faces to be the style providers, as shown in Figure 6. In Figure 6a, the embedded data from the unseen target extracted for FaceVC scatter on the projected two-dimensional surface. On the other hand, in Figure 6b, our embeddings are well clustered for each unseen speaker.

We also demonstrate the visualization results for the Mel spectrogram and pitch contour to show the style conversion. An unseen woman's speech and a male face are chosen. As in Figures 7a and 7b, the pitch contour for the woman is approximately 300 HZ, but it is approximately 200 HZ for the man. The pitch contour of the FaceVC result in Figure 7c is much higher than the target pitch in Figure 7b, and the contour is not similar to the original utterance. This means that the content information is distorted and that the style cannot reflect the speaker's style. However, our model can generate a more similar contour for the source utterance and the target style of about 200 HZ, as in Figure 7d.

It can be concluded that our model achieves a higher performance in synthesized speech quality and outperforms in style representation.

Style Stability for Cross-Modality For most VC works, the number of training speakers is usually less than 100. These studies extract acoustic information from the Mel spectrogram. However, for face-based voice conversion, it is a cross-modal work. Without a large number of styles provided, the model cannot learn a moderate relationship between the facial and acoustic domains. For example, FaceVC learns only utterances from 20 speakers, so speaker diversity is low and cannot synthesize a moderate speech style for the unseen speaker, as in Figure 6a.

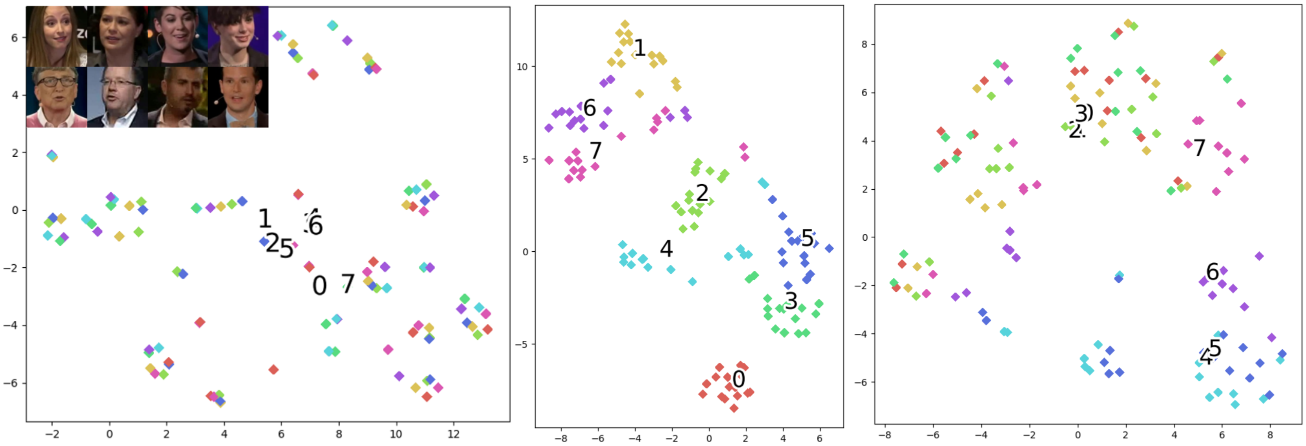

Figure 6: Visualization result for the speaker embedding distribution derived from the conversion speech with t-SNE. We adopt unseen utterances for FaceVC and our model with 100 and 400 speakers. The same color points stand for the same speakers. Numbers 0-3 are females, 4-7 are males. Their facial images are shown on the left top corner. Numbers from left to right, top to down are 0-8, sequentially. Figures from left to right: the result from (a) FaceVC (b) Our model with 100 speakers (c) Our model with 400 speakers.

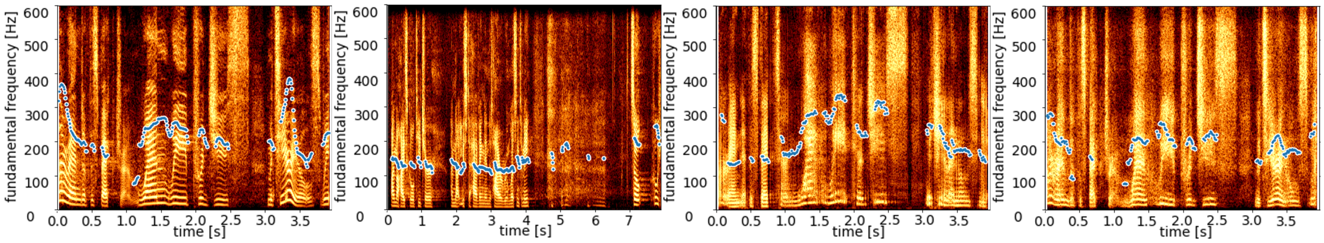

Figure 7: The comparision results of Mel Spectrogram. The blue dot line indicates the pitch contour. We use the Parselmouth (Jadoul, Thompson, and De Boer 2018) library. From left to right:(a) Mel Spectrogram of source utterance; (b) Mel Spectrogram of target utterance; (c) The conversion result with FaceVC; (d) The conversion result with our model.

Therefore, we conduct experiments with different numbers of training speakers to show how the synthesized speech quality of our proposed work will be affected by the training data. Training in a 100-speaker model gives the highest overall MOS results among the three models, as shown in Table 2. When the number of training speakers increases, the quality starts to degrade. The reason is that when the number of training samples increases, a large complication makes the model degrade no matter on the content or style in Figure 6c. For linguistic information, since various background noises and ascents appear in the training set, the model learns hard to distinguish them from the words spoken by the speakers. For acoustic information, the increasing uncertainty for face and voice mapping along with the growing number of speakers also makes the speaker similarity decrease. However, our models, which are for 200 and 400 training speakers, still outperform FaceVC in all evaluation metrics. This shows that our method works better for many-to-many and zero-shot voice conversion.

#Spk.	F2F	M2M	M2F	F2M	Avg.
100	2.483	**2.636**	**2.635**	**2.572**	**2.582**
200	2.345	2.564	2.562	2.500	2.493
400	**2.531**	2.294	2.478	2.446	2.437

Table 2: Overall MOS quality for different number of training speakers evaluated by NISQA.

Ablation Study To demonstrate the contribution of each component, we first show how the bottleneck will affect our model. We input the Mel spectrogram and SP into the content encoder and keep the remaining part the same. We set a fine bottleneck size and try to make the conversion result stay speaker diversity as high as our model, and evaluate the results with MCD. As shown in Figure 8, the model with SP as input data (bottleneck-free) gets lower values than with the Mel spectrogram (bottleneck-dependent) regardless of whether utterances or facial embedding are seen or unseen. Since deciding to make a better diverse style or clear content information is a trade-off for the bottleneck-dependent

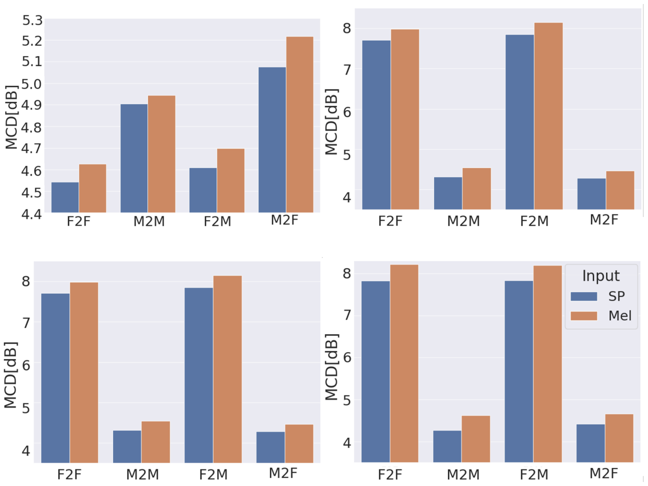

Figure 8: MCD values of synthesized speech. We compare the model result with SP and with Mel spectrogram as input to see the effect of bottleneck-free and bottleneck-dependent on the linguistic information. The left bar is for bottleneck-free, the right bar is for bottleneck-dependent. From top left to bottom right:(a) seen-to-seen; (b) unseen-to-unseen; (c) seen-to-unseen; (d) unseen-to-seen. The values are lower, the better.

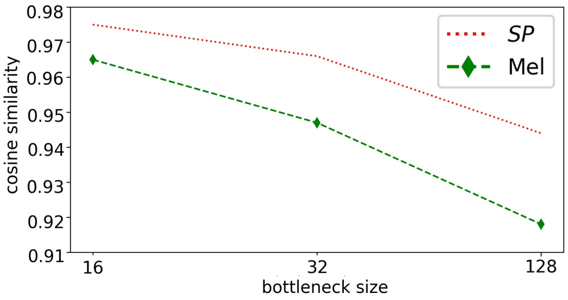

Figure 9: The speaker similarity affected by bottleneck size.

technique, the MCD values are therefore high. Some audio samples can be found on the demo website [4] to show the difference between these two methods. In Figure 9, we increase the size of bottleneck from 32 to 128 and decrease the size from 32 to 16 for bottleneck-free and bottleneck-dependent models. Since the size of the bottleneck is a hyperparameter for the model, the performance will have changed slightly when the bottleneck size is adjusted. But our method should be more stable than the bottleneck disentanglement method. As a result, the speaker similarity has a significant drop when the model is bottleneck dependent, since the acoustic information is affected by the bottleneck. On the other hand, our method changes little. That is, our model can learn to disentangle speech without bottleneck adjustment .

For the overall quality shown in Table 3, the MOS scores decrease when there is at least one component unavailable.

A.	R.	D.	F2F	M2M	M2F	F2M	Avg.
v	v	v	2.483	**2.636**	**2.635**	**2.572**	**2.582**
-	v	v	2.415	2.398	2.410	2.381	2.401
v	v	-	**2.501**	2.479	2.480	2.487	2.487
v	-	v	2.429	2.512	2.417	2.423	2.445
v	-	-	2.270	2.506	2.198	2.545	2.380
-	-	v	2.415	2.327	2.452	2.298	2.373
-	v	-	2.382	2.282	2.390	2.337	2.347
-	-	-	2.461	2.305	2.528	2.291	2.397

Table 3: Ablation study with MOS quality evaluated by NISQA. The first row with all components is our model. A.: masked self-attention ;R.: reparameterization trick; D.: multiscale discriminator

For example, if the model only contains the reparameterization module, the quality becomes worse compared to a good-bottleneck-adjusted plain autoencoder architecture. This is because the reparameterization module strives to map the cross-modal distribution into a proper distribution; however, only a reconstruction loss cannot take care of the speech quality and the cross-modal distribution at the same time. Just in reverse, by adding the self-attention modules, it achieves the highest score within the cases of addition with a single component. It removes some background echo or noise for content embedding. In conclusion, it can be said that simply adding one of the proposed modules might assist in style learning or noise removal, but cannot improve the overall quality. Both the self-attention module and the reparameterization trick need to be guided by the discriminator to achieve a better style or quality of speech. Three modules need to work together to obtain the best style diversity and speech quality from the wild data.

Conclusion

We propose a face-based bottleneck-free voice conversion, with the aim of mapping a facial acoustic distribution and synthesizing speech in the real-world scenario. The preprocessing of the cepstrum with low-pass liftering can simply lead the model without the bottleneck strategy. With the assistance of the self-attention module, the content encoder can pay attention to preserving linguistic-related information without perturbation by background noise. In addition, we use an average facial embedding with reparameterization trick, and an adversarial guide decoder on how to find the relationship between two modalities. Furthermore, the overall quality of speech is under the control of the multi-scale discriminator. Extensive experiments demonstrate that our model is superior to previous work, regardless of the cross-modal diversity or the overall quality of speech. We believe that this work illuminates more possibilities for cross-modal research for face and speech. In the future, how to improve the quality to reach excellent performance on the in-the-wild data and how to make the model more suitable for extra training samples are the further work to be done.

[4] https://sites.google.com/view/spfacevc-demo/

Acknowledgments

This work is supported in part by the National Science and Technology Council (NSTC) of Taiwan under the grants NSTC-109-2221-E-009-114-MY3 and NSTC-109-2223-E-009-002-MY3.

References

Afouras, T.; Chung, J. S.; and Zisserman, A. 2018. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*.

Bulat, A.; and Tzimiropoulos, G. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision*.

Chen, Y.-H.; Wu, D.-Y.; Wu, T.-H.; and Lee, H.-y. 2021. Again-vc: A one-shot voice conversion using activation guidance and adaptive instance normalization. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5954–5958. IEEE.

Choi, H.-S.; Lee, J.; Kim, W.; Lee, J.; Heo, H.; and Lee, K. 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34: 16251–16265.

Dehak, N.; Kenny, P. J.; Dehak, R.; Dumouchel, P.; and Ouellet, P. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4): 788–798.

Goto, S.; Onishi, K.; Saito, Y.; Tachibana, K.; and Mori, K. 2020. Face2Speech: Towards Multi-Speaker Text-to-Speech Synthesis Using an Embedding Vector Predicted from a Face Image. In *INTERSPEECH*, 1321–1325.

Jadoul, Y.; Thompson, B.; and De Boer, B. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71: 1–15.

Kameoka, H.; Kaneko, T.; Tanaka, K.; and Hojo, N. 2018. Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, 266–273. IEEE.

Kaneko, T.; and Kameoka, H. 2018. Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, 2100–2104. IEEE.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, 125–128. IEEE.

Kumar, K.; Kumar, R.; de Boissiere, T.; Gestin, L.; Teoh, W. Z.; Sotelo, J.; de Brébisson, A.; Bengio, Y.; and Courville, A. C. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Lee, S.-H.; Yoon, H.-W.; Noh, H.-R.; Kim, J.-H.; and Lee, S.-W. 2021. Multi-spectrogan: High-diversity and high-fidelity spectrogram generation with adversarial style combination for speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 13198–13206.

Lin, Y. Y.; Chien, C.-M.; Lin, J.-H.; Lee, H.-y.; and Lee, L.-s. 2021. Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5939–5943. IEEE.

Lu, H.-H.; Weng, S.-E.; Yen, Y.-F.; Shuai, H.-H.; and Cheng, W.-H. 2021. Face-based Voice Conversion: Learning the Voice behind a Face. In *Proceedings of the 29th ACM International Conference on Multimedia*, 496–505.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mittag, G.; Naderi, B.; Chehadi, A.; and Möller, S. 2021. NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. *Proc. Interspeech 2021*.

Mukhneri, F. M.; Wijayanto, I.; and Hadiyoso, S. 2020. Voice conversion for dubbing using linear predictive coding and hidden markov model. *Journal of Southwest Jiaotong University*, 55(4).

Murphy, P. J.; and Akande, O. O. 2007. Noise estimation in voice signals using short-term cepstral analysis. *The Journal of the Acoustical Society of America*, 121(3): 1679–1690.

Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7539–7548.

Oord, A. v. d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; and Kavukcuoglu, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Plüster, B.; Weber, C.; Qu, L.; and Wermter, S. 2021. Hearing Faces: Target Speaker Text-to-Speech Synthesis from a Face. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 757–764.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3617–3621. IEEE.

Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; and Hasegawa-Johnson, M. 2019. Autovc: Zero-shot voice style transfer with only autoencoder loss. In *International Conference on Machine Learning*, 5210–5219. PMLR.

Ren, Y.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2022. Revisiting Over-Smoothness in Text to Speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8197–8213.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Sisman, B.; Yamagishi, J.; King, S.; and Li, H. 2021. An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 132–157.

Snyder, D.; Garcia-Romero, D.; Povey, D.; and Khudanpur, S. 2017. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, volume 2017, 999–1003.

Wang, J.; Wang, Z.; Hu, X.; Li, X.; Fang, Q.; and Liu, L. 2022. Residual-Guided Personalized Speech Synthesis based on Face Image. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4743–4747.

Wu, D.-Y.; Chen, Y.-H.; and Lee, H.-y. 2020. VQVC+: One-Shot Voice Conversion by Vector Quantization and U-Net Architecture. *Proc. Interspeech 2020*, 4691–4695.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10): 1499–1503.