Liu *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:40
https://doi.org/10.1186/s13636-021-00226-3

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

U^2-VC: one-shot voice conversion using two-level nested U-structure

Fangkun Liu[1,2†], Hui Wang[1†], Renhua Peng[2,3*], Chengshi Zheng[2,3] and Xiaodong Li[2,3]

Abstract

Voice conversion is to transform a source speaker to the target one, while keeping the linguistic content unchanged. Recently, one-shot voice conversion gradually becomes a hot topic for its potentially wide range of applications, where it has the capability to convert the voice from any source speaker to any other target speaker even when both the source speaker and the target speaker are unseen during training. Although a great progress has been made in one-shot voice conversion, the naturalness of the converted speech remains a challenging problem. To further improve the naturalness of the converted speech, this paper proposes a two-level nested U-structure (U^2-Net) voice conversion algorithm called U^2-VC. The U^2-Net can extract both local feature and multi-scale feature of log-mel spectrogram, which can help to learn the time-frequency structures of the source speech and the target speech. Moreover, we adopt sandwich adaptive instance normalization (SaAdaIN) in decoder for speaker identity transformation to retain more content information of the source speech while maintaining the speaker similarity between the converted speech and the target speech. Experiments on VCTK dataset show that U^2-VC outperforms many SOTA approaches including AGAIN-VC and AdaIN-VC in terms of both objective and subjective measurements.

Keywords: Voice conversion, U^2-Net, Sandwich adaptive instance normalization

1 Introduction

It is well-known that speech information is composed of four components: timbre, rhythm, pitch, and content. As stated in [1], the content represents the linguistic part of the speech, and the timbre represents the speaker identity. Voice conversion (VC) aims to convert the timbre while maintaining the linguistic content unchanged. This technique can be applied in many fields, such as pronunciation assistance [2, 3], personalized speech synthesis [4, 5], and even dubbing.

The existing voice conversion systems can be roughly divided into parallel voice conversion [6–9] and non-parallel voice conversion [10–17], which depend on whether the model of this system is trained on paired utterances of the same linguistic content spoken by source and target speakers or not [18]. Due to the difficulty of collecting parallel data in reality, non-parallel VC has gained more attention, and many researchers have proposed lots of effective methods, such as generative adversarial network (GAN) [10–13], variational auto-encoder (VAE) [14], and attention-mechanism [15]. However, many conventional non-parallel VC algorithms cannot work well when converting a voice of one unseen speaker in the training data without being retrained.

Recently, one-shot voice conversion has been proposed to solve the problem of the conventional non-parallel voice conversion. For one-shot voice conversion, either or both of the source speaker and the target speaker can be unseen during training. Several works [19–22] have been done for this challenging task, while the naturalness of the converted speech still remains an unsolved problem.

In this paper, we propose a one-shot voice conversion algorithm based on U^2-Net [23] called U^2-VC, where U^2-Net was first used in salient object detection (SOD).

*Correspondence: pengrenhua@mail.ioa.ac.cn
†Fangkun Liu and Hui Wang contributed equally to this work and should be considered co-first authors.
[2]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190 Beijing, China
[3]University of Chinese Academy of Sciences, 100049 Beijing, China
Full list of author information is available at the end of the article

Different from U-Net [24], U^2-Net is consisted of many residual U-blocks (RSU) with different layers [23]. Each residual U-block can extract both local feature and multi-scale feature from input image to preserve more detailed information about salient object. For voice conversion, if one regards the log-mel spectrogram as a 2D image, the speech spectrogram can be regarded as the salient object of this "2D image," and the residual U-block can be adopted to extract local feature and multi-scale feature from the input speech to preserve more detailed information, which can be expected to improve the naturalness of the converted speech. This motivates us to design a voice conversion algorithm based on U^2-Net to improve the naturalness of the converted speech. Inspired by AGAIN-VC [22], the sigmoid function is adopted as the activation guidance at the last of encoder to guide the content embedding, so that we can learn more content information. Moreover, sandwich adaptive instance normalization (SaAdaIN) [25], which is first proposed for neural style transformation to reduce content loss during transformation process, is also adopted for speaker identity transformation in decoder to maintain more content information of the source speech and keep the speaker similarity between the converted speech and the target speech simultaneously. To the best of our knowledge, it is the first one-shot voice conversion algorithm with U^2-Net and SaAdaIN. Objective evaluation results such as mel-cepstral distortion (MCD) [26], NISQA model [27], and subjective listening tests with mean opinion score (MOS) showed that the proposed approach outperforms many state-of-the-art (SOTA) approaches such as AdaIN-VC [21] and AGAIN-VC [22]. To validate the robustness of the proposed approach, we also perform experiments in cross-lingual scenarios, where the results also verify the better performance of the proposed approach.

2 Related work

2.1 One-shot voice conversion

One-shot voice conversion can be achieved by decoupling content and speaker identity with a content encoder and a speaker encoder, respectively. AUTOVC [19] uses a pre-trained speaker encoder with a generalized end-to-end loss [28] and a content encoder with a well-designed information bottleneck to limit the leakage of speaker information of the source speaker. However, the pre-trained speaker encoder might affect the robustness of the system because it is just trained for speaker verification and the structure of information bottleneck with hard sampling may cause the content information loss, which makes the converted speech sound unnatural. VQVC+ [20] jointly uses vector quantization (VQ) [29] and U-Net [24] for extracting content information and improving the reconstruction simultaneously. Although VQVC+ performs well on speaker conversion, the naturalness still cannot meet the specification because of the content information loss. AdaIN-VC [21] is the first one-shot voice conversion to perform speaker transformation through adaptive instance normalization (AdaIN) [30]. Although AdaIN can separate the content information and the speaker information very well, one can see that the perceptual quality of the converted speech is still unsatisfied, e.g., the naturalness. Very recently, AGAIN-VC [22] was proposed with a single encoder for encoding both content and speaker identity, where the sigmoid function was added to the end of the encoder as an information bottleneck to prevent the content embedding from leaking speaker information. The quality of the converted speech is better than many existing algorithms, and the naturalness is still a big problem because the harmonic distortion still often occurs in the converted speech.

As mentioned above, both the content information loss and the harmonic distortion lead to the degradation of the converted speech. Recently, some methods have been proposed to improve the naturalness of the converted speech [13, 31–33]. Kwon et al. [31] use the attention mechanism to modify the information bottleneck structure to preserve more linguistic information, which can prevent the content loss of the converted speech. CycleGAN-VC3 [13] uses time-frequency adaptive normalization (TFAN) to reduce the harmonic distortion of the converted speech in order to make it sound more natural. Text-to-speech (TTS) [32, 33] and automatic speech recognition (ASR) [33] techniques also have been introduced to overcome the problem of mispronunciation in the converted speech. In this paper, the proposed U^2-VC is introduced to improve the naturalness of the converted speech by extracting multi-scale features through newly designed 1-2-1 residual U-blocks and the usage of SaAdaIN to maintain more content information during speaker identity transformation

2.2 U^2-Net

U^2-Net is a kind of two-level nested U-structure [23], which is originally used for SOD. Residual U-blocks(RSU) with different layer (L) are the main components of U^2-Net to extract different scale features. Each RSU has the same structure: an input convolution layer to extract the local features, a U-Net like encoder-decoder block with layer of L to extract the multi-scale features from the local features, and a residual connection to fuse the local features and the multi-scale features by summation. The characteristic of U^2-Net makes it able to extract more details from input features. As demonstrated in [23], U^2-Net can perform well on SOD. Due to the importance of preserving more harmonic components for the converted speech in improving speech quality, we introduce the U^2-Net to the one-shot VC task, where we verify that

the naturalness of the converted speech can be improved significantly by doing so.

3 Proposed approach

For voice conversion, the goal is to design a system that can convert a source speaker to the target one, while keeping the source linguistic content unchanged, which can be represented as:

$$\widehat{X}_{1\to 2} = \mathcal{C}\left(X_1, X_2\right) \tag{1}$$

where X_1 denotes the log-mel spectrogram of the source speech, and X_2 denotes that of the target speech. $\mathcal{C}$ represents the nonlinear mapping function. $\widehat{X}_{1\to 2}$ denotes the log-mel spectrogram of the converted speech.

3.1 Structure overview

Figure 1 plots the overall structure of the proposed U^2-VC, where one can see that U^2-VC is mainly consisted of three parts: an encoder to disentangle content information and speaker identity information through instance normalization (IN), a decoder to achieve speaker identity transformation through sandwich adaptive instance normalization (SaAdaIN), and an output module to generate the final converted log-mel spectrogram of the converted speech. For encoder, the source log-mel spectrogram X_1 is passed through IN layers following the 1-2-1 residual U-blocks and the sigmoid function to eliminate its speaker information, which is for generating content embedding. Meanwhile, the skip-connection structure passes the speaker embedding μ and σ, which are calculated from target log-mel spectrogram X_2, of each instance normalization (IN) layer to the paired SaAdaIN layer in decoder for speaker identity transformation. The output module has the same purpose as the saliency map fusion module of U^2-Net [23]. Firstly, side out log-mel spectrograms of the converted speech $\widehat{X}^i_{1\to 2}$, with the index $i = 1, \cdots, 6$, are generated by the generation blocks following the 1-2-1 RSU blocks in decoder. The generation block consists of two GRU layers and a linear layer, and the output module reshapes the side out log-mel spectrograms from 2D to 1D and then fuses them with a concatenation operation followed by a 1×1 convolution layer. Finally, a reshaping operation is needed to generate the final log-mel spectrogram of the converted speech. The 1-2-1 residual U-blocks and SaAdaIN are the most important components of U^2-VC to improve the naturalness of the converted speech, and we will describe them in the next two sections separately. The loss function of U^2-VC will also be introduced finally, and the configuration of the proposed U^2-VC is summarized in Table 1.

3.2 1-2-1 residual U-block

RSU was first developed to extract the multi-scale features from image for salient object detection. More detailed information can be preserved with this approach. To make residual U-block suitable for voice conversion task, we redesigned the residual U-block named 1-2-1 residual U-block (1-2-1 RSU). The detailed structure of 1-2-1 residual U-block is shown in Fig. 2b. In Fig. 2, the ReBNConv1D

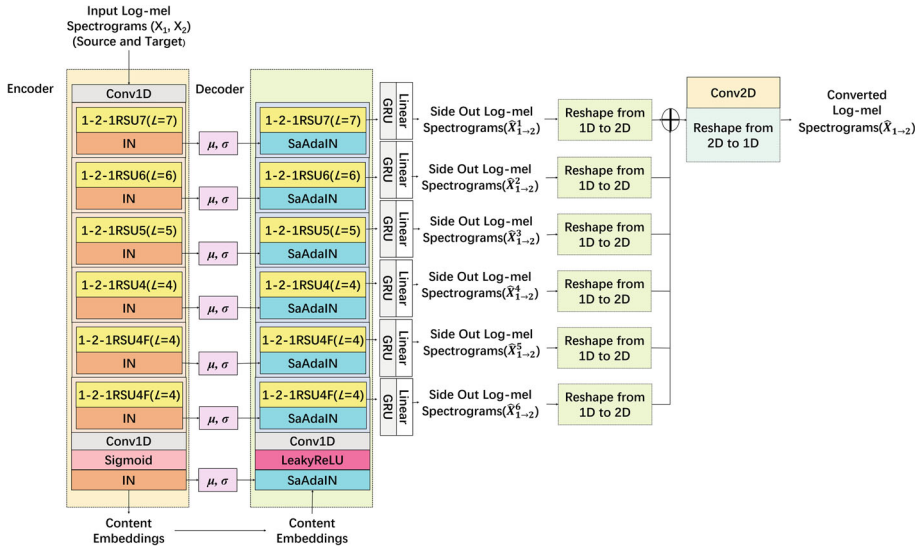

Fig. 1 The architecture of U^2-VC. The 1-2-1 residual U-blocks (RSU) with different layers (1-2-1RSU7,$\cdots$,1-2-1RSU4F) consist of the U-Net like encoder-decoder structure. "7", "6", "5," and "4" represent the layers (L) of 1-2-1 residual U-blocks. Greater L means the1-2-1 residual U-block could capture more large-scale information . In this network, we set the L from large to small in order to extract the features from the global to the detail. This process preserves more fine details of input features which could be better for the naturalness of converted speech. Inspired by AGAIN-VC, sigmoid function is used at the end of encoder. Sandwich adaptive instance normalization (SaAdaIN) is adopted in decoder for speaker identity transformation

Table 1 Bold words represent the three parts of our U^2-VC as noted in section 3. "I" denotes the input size. "O" denotes the output size. "K" denotes the kernel size. "S" denotes the stride. "D" denotes the dilation. "H" denotes the hidden size. "L" denotes the layer. As stated before, residual U-block has the same structure as original residual U-block except the input layer and the reshaping operation

	Encoder	**Decoder**
	Conv1D_{IN} I:80 O:256 K:1×1	Conv1D_{IN} I:4 O:256 K:3×3
	Conv1D_{OUT} I:256 O:4 K:1×1	
1-2-1RSU7	Conv1D_{IN} I:256 O:256 K:3×3	Conv1D_{IN} I:256 O:256 K:3×3
	Conv2D_{EN1} I:1 O:16 K:3×3	Conv2D_{EN1} I:1 O:16 K:3×3
	$\text{Conv2D}_{EN2\sim6}$ I:16 O:16 K:3×3	$\text{Conv2DEN}_{EN2\sim6}$ I:16 O:16 K:3×3
	MaxPool2D K:3×3 S:2	MaxPool2D K:3×3 S:2
	Conv2D_{EN7} I:16 O:16 K:3×3 D:2	Conv2D_{EN7} I:16 O:16 K:3×3 D:2
	$\text{Conv2D}_{DE6\sim2}$ I:32 O:32 K:3×3	$\text{Conv2D}_{DE6\sim2}$ I:32 O:32 K:3×3
	Conv2D_{DE1} I:32 O:1 K:3×3	Conv2D_{DE1} I:32 O:1 K:3×3
1-2-1RSU6	Conv1D_{IN} I:256 O:256 K:3×3	Conv1D_{IN} I:256 O:256 K:3×3
	Conv2D_{EN1} I:1 O:16 K:3×3	Conv2D_{EN1} I:1 O:16 K:3×3
	MaxPool2D K:3×3 S:2	MaxPool2D K:3×3 S:2
	$\text{Conv2DEN}_{EN2\sim5}$ I:16 O:16 K:3×3	$\text{Conv2DEN}_{EN2\sim5}$ I:16 O:16 K:3×3
	Conv2D_{EN6} I:16 O:16 K:3×3 D:2	Conv2D_{EN6} I:16 O:16 K:3×3 D:2
	$\text{Conv2D}_{DE5\sim2}$ I:32 O:32 K:3×3	$\text{Conv2D}_{DE5\sim2}$ I:32 O:32 K:3×3
	Conv2D_{DE1} I:32 O:1 K:3×3	Conv2D_{DE1} I:32 O:1 K:3×3
1-2-1RSU5	Conv1D_{IN} I:256 O:256 K:3×3	Conv1D_{IN} I:256 O:256 K:3×3
	Conv2D_{EN1} I:1 O:16 K:3×3	Conv2D_{EN1} I:1 O:16 K:3×3
	MaxPool2D K:3×3 S:2	MaxPool2D K:3×3 S:2
	$\text{Conv2D}_{EN2\sim4}$ I:16 O:16 K:3×3	$\text{Conv2D}_{EN2\sim4}$ I:16 O:16 K:3×3
	Conv2D_{EN5} I:16 O:16 K:3×3 D:2	Conv2D_{EN5} I:16 O:16 K:3×3 D:2
	$\text{Conv2D}_{DE4\sim2}$ I:32 O:32 K:3×3	$\text{Conv2D}_{DE4\sim2}$ I:32 O:32 K:3×3
	Conv2D_{DE1} I:32 O:1 K:3×3	Conv2D_{DE1} I:32 O:1 K:3×3
1-2-1RSU4	Conv1D_{IN} I:256 O:256 K:3×3	Conv1D_{IN} I:256 O:256 K:3×3
	Conv2D_{EN1} I:1 O:16 K:3×3	Conv2D_{EN1} I:1 O:16 K:3×3
	MaxPool2D K:3×3 S:2	MaxPool2D K:3×3 S:2
	$\text{Conv2D}_{EN2\sim3}$ I:16 O:16 K:3×3	$\text{Conv2D}_{EN2\sim3}$ I:16 O:16 K:3×3
	Conv2D_{EN4} I:16 O:16 K:3×3 D:2	Conv2D_{EN4} I:16 O:16 K:3×3 D:2
	$\text{Conv2D}_{DE3\sim2}$ I:32 O:32 K:3×3	$\text{Conv2D}_{DE3\sim2}$ I:32 O:32 K:3×3
	Conv2D_{DE1} I:32 O:1 K:3×3	Conv2D_{DE1} I:32 O:1 K:3×3
1-2-1RSU4F	Conv1D_{IN} I:256 O:256 K:3×3	Conv1D_{IN} I:256 O:256 K:3×3
	Conv2D_{EN1} I:1 O:16 K:3×3	Conv2D_{EN1} I:1 O:16 K:3×3
	Conv2D_{EN2} I:16 O:16 K:3×3 D:2	Conv2D_{EN2} I:16 O:16 K:3×3 D:2
	Conv2D_{EN3} I:16 O:16 K:3×3 D:4	Conv2D_{EN3} I:16 O:16 K:3×3 D:4
	Conv2D_{EN4} I:16 O:16 K:3×3 D:8	Conv2D_{EN4} I:16 O:16 K:3×3 D:8
	Conv2D_{DE3} I:32 O:32 K:3×3 D:4	Conv2D_{DE3} I:32 O:32 K:3×3 D:4
	Conv2D_{DE2} I:32 O:32 K:3×3 D:2	Conv2D_{DE2} I:32 O:32 K:3×3 D:2
	Conv2D_{DE1} I:32 O:1 K:3×3	Conv2D_{DE1} I:32 O:1 K:3×3

Output Module: GRU I:256 H:256 L:2; linear I:256 O:80; Conv2DFuse I:6 O:1 K:1×1

denotes the block consists of a 1D CNN, a batch normalization layer, and a leaky Relu activation function, while the ReBNConv2D has the similar architecture and the only difference is that the 1D CNN in the ReBNConv1D is replaced by a 2D CNN.

The main differences between the original residual U-block and the proposed 1-2-1 residual U-block are the input convolution layer and the reshaping operation. As mentioned before, speech signal is a kind of sequence while its spectrogram can be regarded as an image with one channel. Accordingly, we adopt a 1-2-1D CNN structure in newly designed residual U-block according to the pros and cons of 1D CNN and 2D CNN, as pointed out in [12]: a 1D CNN is good at capturing dynamic change,

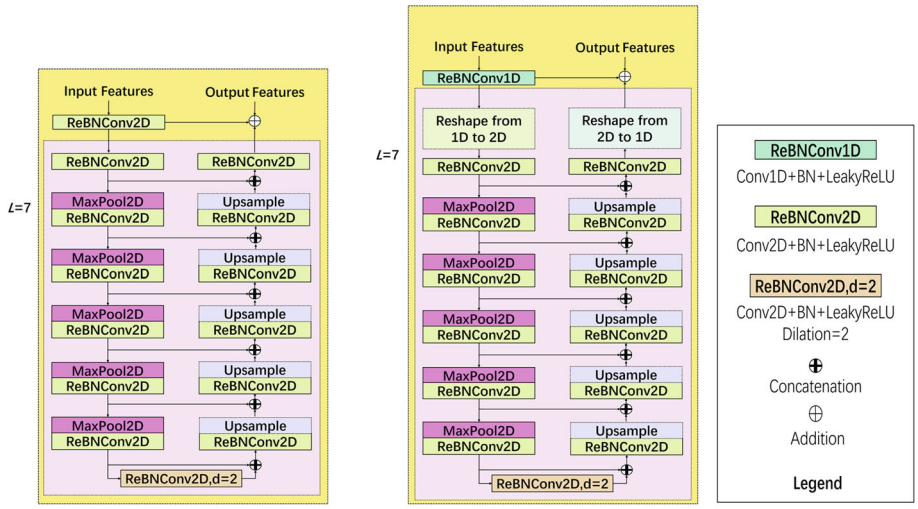

Fig. 2 The comparison of residual U-block (RSU) and 1-2-1 residual U-block (1-2-1 RSU). We set the number of U-block layers (L) equal to 7 as an example in both residual U-block and 1-2-1 residual U-block. 1-2-1 RSU with other number of layers have the same structures as 1-2-1 residual U-block (L=7)

while a 2D CNN is good at converting features while preserving the original structures. Therefore, a 1D CNN with batch normalization(BN) followed LeakyReLU is used as the input layer to extract the local feature, and many different layers of 2D CNN with batch normalization and LeakyReLU are used for downsampling and upsampling in U-block to extract multi-scale feature. The output feature of U-block has to be reshaped from 2D to 1D because we only use 1D IN to extract the speaker identity information along the time axis. The main reason is that the speaker identity information is a kind of time-invariant feature.

Figure 3 compares the 1-2-1 residual U-block (RSU) used in U^2-VC and the plain residual block used in AGAIN-VC. The operation of the plain residual block can be described as:

$$\mathcal{F}(x) = x + \mathcal{H}_2(\mathcal{H}_1(x)), \tag{2}$$

where $\mathcal{F}(x)$ denotes the output feature of the plain residual block when given the input feature x, which is extracted from log-mel spectrograms of the source speech and the target speech; $\mathcal{H}_1$ represents the operation of ReBNConv1D and $\mathcal{H}_2$ represents the operation of plain 1D convolution block. Both operations aim to extract the local feature. Instead of using the plain convolution block to extract the local feature only, multi-layer U-block is proposed in RSU to extract different scale features. The output feature of RSU can be represented as:

$$\mathcal{F}_{\mathrm{RSU}}(x) = \mathcal{H}_1(x) + \mathcal{H}_U(\mathcal{H}_1(x)), \tag{3}$$

where $\mathcal{F}_{\mathrm{RSU}}(x)$ denotes the output feature of RSU; $\mathcal{H}_1$ stands for the convolution operation to extract the local feature, and $\mathcal{H}_U$ represents the operation of U-block to extract the multi-scale feature from the local feature $\mathcal{H}_1(x)$. The benefit of using RSU is that it helps U^2-VC preserve more detailed features when compared to AGAIN-VC, which is helpful to improve the naturalness of the converted speech.

3.3 Sandwich adaptive instance normalization

Sandwich adaptive instance normalization (SaAdaIN) is the extended application of sandwich batch normalization for style transformation [25]. AdaIN firstly preforms instance normalization on content feature, and then affine transformation is performed on normalized content feature with the statistic of style feature. AdaIN can be formulated as:

$$\mathrm{AdaIN}(C, S) = \sigma(S)\frac{C - \mu(C)}{\sigma(C)} + \mu(S), \tag{4}$$

where C is the content input, and S is the style input. Different from batch normalization, $\mu(\cdot)$ and $\sigma(\cdot)$ represent the channel-wise average and standard deviation of the input, respectively. As discussed in [25], AdaIN would lead to the content loss in the output because the style-dependent re-scale might further amplify the intrinsic data heterogeneity brought by the variety of the input content images. To reduce the content loss problem in AdaIN, SaAdaIN is proposed with shared sandwich affine layer after the instance normalization of content feature to reduce data heterogeneity, which can make the output preserve more content information. The proposed SaAdaIN can be formulated as:

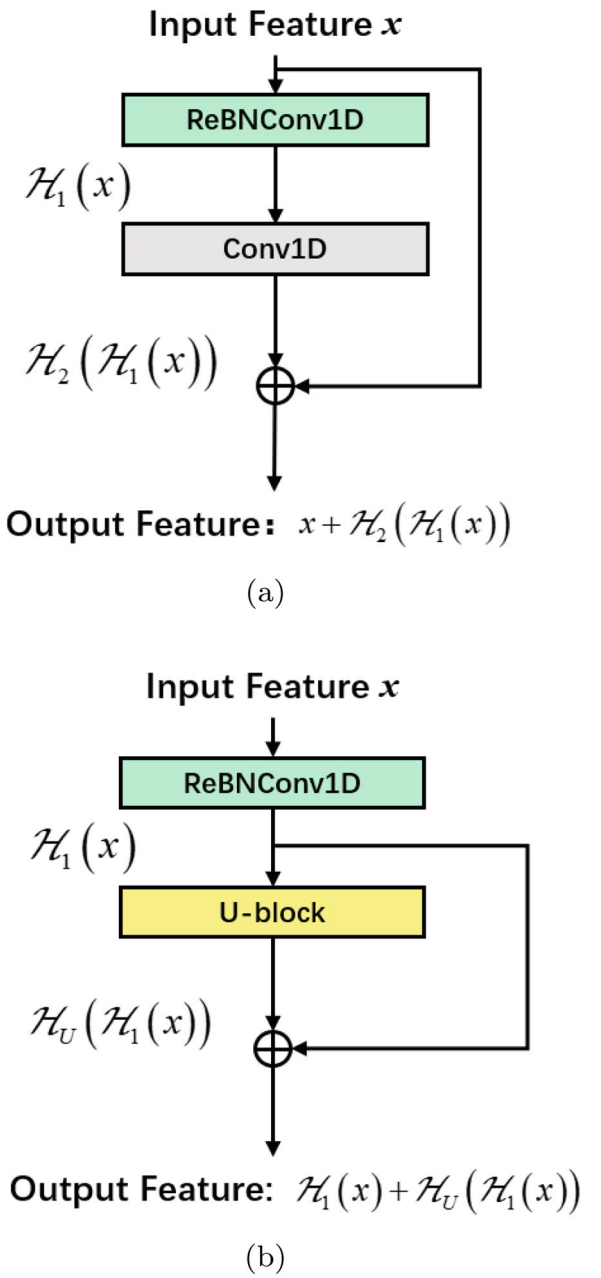

Fig. 3 The comparison of 1-2-1 residual U-block and plain residual block. **a** The residual block. **b** The 1-2-1 residual U-block (1-2-1RSU). The constitution of the output feature from each block illustrates why the 1-2-1 residual U-block can preserve more details than plain residual block

$$\mathrm{SaAdaIN}(C,S)=\sigma(S)\left(\gamma_{sa}\left(\frac{C-\mu(C)}{\sigma(C)}\right)+\beta_{sa}\right)+\mu(S), \tag{5}$$

where γ_{sa} and β_{sa} are the parameters that can be learned from the added affine layer, and their shapes are both the same as the number of channels of input feature. Experimental results on image style transformation have verified the superiority of SaAdaIN compared to AdaIN [25]. And the ablation experimental results in this paper also show the importance of introducing SaAdaIN for voice conversion in improving the naturalness of the converted speech.

3.4 Loss

It has to be noted that log-mel spectrogram of the source and that of the target are both extracted from the speech signals with the same speaker during the training phase in order to calculate all the side self-reconstruction losses and final self-reconstruction loss. We adopt deep supervision similar to U^2-Net, and the loss is defined as:

$$\mathrm{Loss}=\omega_f l_f+\sum_{i=1}^{M}\omega_i l_i \tag{6}$$

$$l_i=\left\|X_1-\widehat{X}^i_{1\to 1}\right\|_1^1 \tag{7}$$

$$l_f=\left\|X_1-\widehat{X}_{1\to 1}\right\|_1^1 \tag{8}$$

where X_1 represents the source log-mel spectrogram; $\widehat{X}^i_{1\to 1}$, with $i=1,\cdots,M$, represents the ith side output reconstructed log-mel spectrogram, and $\widehat{X}_{1\to 1}$ represents the final reconstructed log-mel spectrogram. l_i denotes the ith side self-reconstruction loss between the side output reconstucted log-mel spectrogram and the source log-mel spectrogram with $M=6$ as shown in Fig. 4, while l_f is the self-reconstruction loss between the final reconstructed log-mel spectrogram and the source log-mel spectrogram. ω_i and ω_f are the weights of the two loss terms. For all l_i and l_f, we use L1 loss as self-reconstruction loss.

4 Experimental setup

We implement three experiments, including the ablation study, mono-lingual voice conversion evaluation, and cross-lingual voice conversion evaluation, to verify the effectiveness of our proposed algorithm in improving the naturalness of the converted speech. AdaIN-VC and AGAIN-VC are chosen as the baselines for comparison in both mono-lingual voice conversion evaluation and cross-lingual voice conversion evaluation. Details will be given in the following parts.

4.1 Dataset

VCTK dataset [34] is chosen for training, ablation study, and comparing the proposed approach with other approaches in mono-ilingual scenario. VCTK is an English dataset consisted of 46-h speech data with 109 speakers. In addition, the dataset for cross-lingual VC task taken from Voice Conversion Challenge (VCC) 2020 [18] is also used to evaluate the robustness of proposed approach in cross-lingual scenario. This dataset includes 6 speakers consisting both male and female speakers. Each speaker

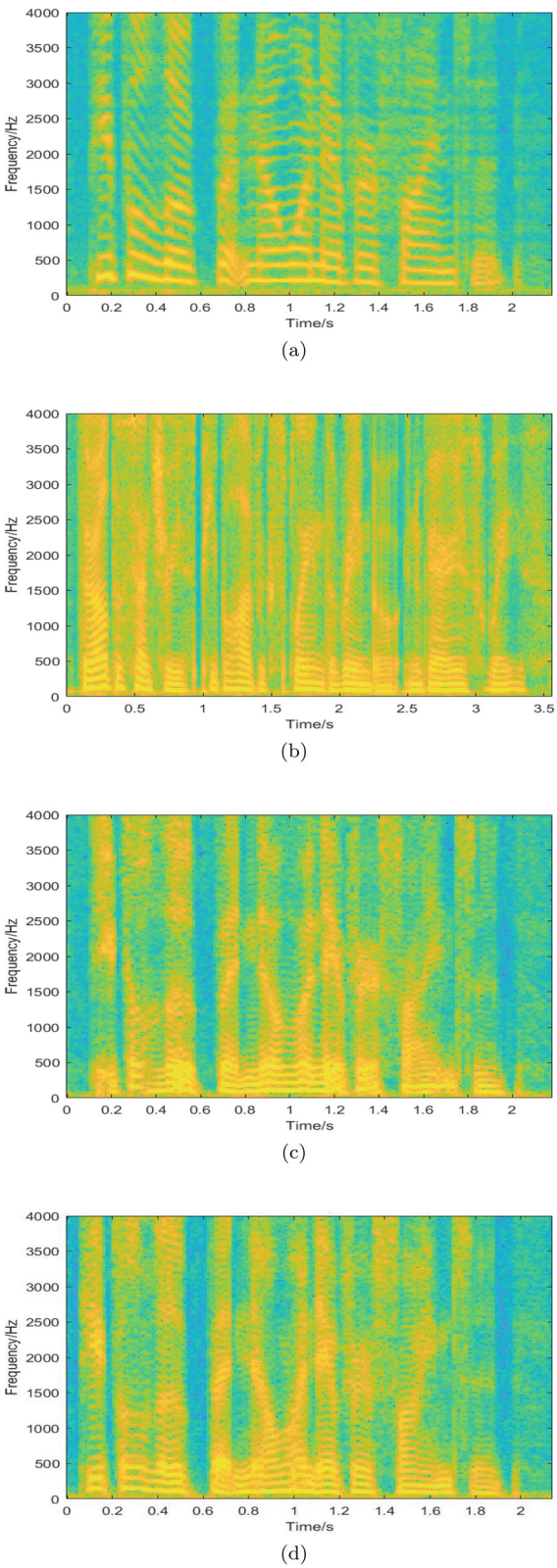

Fig. 4 Comparison among source, target and the converted spectrograms. **a** Source speech. **b** Target speech. **c** The converted speech of AGAIN-VC. **d** The converted speech of U^2-VC

has around 70 utterances. There are 3 languages in this dataset including Finnish, German, and Mandarin.

4.2 Vocoder

Because the output of the proposed U^2-VC is log-mel sectrogram of the converted speech, we need a vocoder to convert the log-mel spectrogram to time-domain waveform. The pretrained MelGAN [35] is chosen as the vocoder when considering its high inference speed of waveform generation and high quality of generated speech. Note that MelGAN is used for all the baselines and the proposed approach as the vocoder to give a fair comparison when implementing all the evaluations.

4.3 Training details

During the training phase, we randomly select 80 speakers from VCTK corpus, and then 200 utterances are randomly chosen for each speaker. Meanwhile, the remaining speakers are randomly chosen for evaluation in unseen-to-unseen conversion scenario. All the raw speech signals are downsampled to 22.05 kHz and transform it into log-mel spectrogram with 1024 STFT window length, 256 hop length, and 80 mel-frequency bins for training and evaluation according to the configuration of MelGAN [35]. During the training phase, AdamW optimizer [36] is used to train our network with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and its initial learning rate set to be $5e^{-4}$. The proposed U^2-VC is implemented based on Pytorch 1.5.1. The training is conducted on NVIDIA Tesla V100 with 32 GB memory, and the number of training step is 170k.

4.4 Evaluation metrics

We evaluate the proposed U^2-VC using the naturalness of the converted speech and the speaker similarity between the converted speech and the target one. Both naturalness and the speaker similarity need objective evaluation and subjective evaluation. Three evaluation metrics are used for objective evaluation and subjective evaluation, which include (1) Mel-cepstral distortion (MCD), (2) NISQA model, and (3) mean opinion score (MOS). The reason for choosing these three evaluation metrics can be summarized as follows:

1. Mel-cepstral distortion: Mel-cepstral distortion (MCD) measures the difference between the target and converted spectral features. It can be calculated between the converted and target Mel-cepstral coeffificients or MCEPs [37]. The lower scores, the better performance.
2. NISQA model: NISQA [27] is a speech quality prediction model. The model can not only well predict the overall MOS, but also measure the four speech quality dimensions including noisiness, coloration, discontinuity, and loudness. We use

NISQA as an objective measurement on the naturalness of the converted speech through the predicted overall MOS. The higher scores, the better performance.

3. Mean opinion score (MOS): mean opinion score (MOS) is used for subjective evaluation on both naturalness and the similarity to the target speaker of the converted speech. For similarity, annotators were asked to rate the score from 1 to 5 depending on how confident they considered these two speech signals were uttered by the same speaker for subjective evaluation on similarity, where 1 represents being totally different and 5 represents being absolutely same after listening the target speech and the converted one. For naturalness, annotators were asked to rate the score from 1 to 5 depending on the naturalness of the converted speech, where 1 represents being completely unnatural and 5 represents being completely natural. The higher score, the better performance.

4.5 Statistical testing method

We use analysis of variance (ANOVA) as the statistical testing method to verify that the proposed approach outperforms the baselines in a statistically significant manner. In ANOVA, we set the confidence level to be 0.95. If the significance between two approaches is less than 0.05, it means that there is a significant difference between the two approaches.

4.6 Experiment implementation details

As mentioned above, we implement three experiments including the ablation study, mono-lingual voice conversion evaluation, and cross-lingual voice conversion evaluation. The experiment implementation details are summarized as follows:

1. Ablation study: Ablation study is performed on U^2-VC to verify the effectiveness of the proposed structure and SaAdaIN. Both seen-to-seen conversion and unseen-to-unseen conversion scenarios are included in this ablation study, where seen-to-seen conversion means both the source speaker and the target one are included in the training set, while unseen-to-unseen conversion indicates that both of them are not included in the training phase. Both objective and subjective metrics are used for evaluation. For subjective evaluation, raters can speak authentic English fluently.
2. Mono-lingual voice conversion evaluation: Mono-lingual conversion performance comparison is conducted to verify the advantage of the proposed U^2-VC. AGAIN-VC and AdaIN-VC are chosen as baselines for validation. Both seen-to-seen conversion and unseen-to-unseen conversion cases are included in this experiment. Both objective and subjective metrics are selected for evaluation. For subjective evaluation, raters can speak authentic English fluently. ANOVA is used here to show the advantage of the proposed approach compared with the two baselines in a statistically significant manner.
3. Cross-lingual voice conversion evaluation: We conduct cross-lingual conversion performance comparison in order to evaluate the robustness of the proposed U^2-VC in cross-lingual scenario inspired by [38], which can also test the ability of the proposed algorithm in disentangling the content information and the speaker identity of the input speech. In this experiment, AGAIN-VC and AdaIN-VC are also chosen as baselines. All the models are trained with VCTK dataset to give a fair comparison; meanwhile, we choose the speech signals of Mandarin speakers from VCC dataset as the target and the unseen speakers from VCTK dataset as the source because the raters of subjective evaluation can speak both Mandarin and English fluently. Both objective and subjective metrics are used for this evaluation. For objective evaluation, we only use NISQA model to evaluate the naturalness because the MCD requires that the converted speech and the target speech have the same content. MOS is chosen as subjective evaluation metric. ANOVA is also used here to further show the advantage of the proposed approach when compared with the baselines in a statistically significant manner.

5 Experiment results

5.1 Ablation study

In this ablation study, we use "S," "T," "F," and "M" to represent the source speech, the target speech, female, and male, respectively. As an example, "SF2TF" represents the conversion from a source female speech to a target female speech.

Tables 2 and 3 present the ablation study results of the converted speech of AGAIN-VC, voice conversion based on only U^2-Net, voice conversion with only SaAdIN and our U^2-VC in speaker similarity, and naturalness through objective evaluation metrics. Table 2 shows the results in seen-to-seen scenario. One can see that all of these approaches nearly have the same MCD score. When focusing on the predicted MOS, U^2-VC shows much better performance compared to AGAIN-VC, and the predicted MOS of the proposed U^2-VC always get the best performance. The maximum difference is 0.3 compared to AGAIN-VC, which is a significant improvement. The same trend can be observed in Table 3, which measures the unseen-to-unseen conversion scenario.

Table 2 Objective evaluation results of the ablation study on architecture in seen-to-seen conversion scenario. "AGAIN-VC" represents the network has neither U^2-Net structure nor SaAdaIN. "U^2-VC" represents the network has both U^2-Net structure and SaAdaIN

	MCD (dB)					Predicted MOS by NISQA				
	SF2TF	SF2TM	SM2TF	SM2TM	Average	SF2TF	SF2TM	SM2TF	SM2TM	Average
AGAIN-VC	6.33	6.07	6.32	6.33	6.26	3.87	3.63	3.93	4.02	3.86
w/o SaAdaIN, with U^2-Net	6.35	6.13	6.36	6.42	6.32	3.97	3.88	3.96	4.02	3.96
w/o U^2-Net, with SaAdaIN	6.34	6.04	6.23	6.31	6.23	4.01	3.83	3.99	3.99	3.96
U^2-VC	6.36	6.11	6.32	6.39	6.29	4.13	3.93	4.14	4.05	4.06

Tables 4 and 5 present the subjective evaluation results of the ablation study to make it more convincing. From these results, one can see that both U^2-Net structure and SaAdaIN can improve the naturalness of the converted speech in both seen-to-seen and unseen-to-unseen conversion scenarios, which is more beneficial by introducing the U^2-Net structure. The integration of U^2-Net structure and SaAdaIN can achieve the highest improvement according to the subjective results. This is because the U^2-Net structure and the SaAdaIN are complementary to each other. Instead of learning the learnable parameters of the SaAdaIN directly from the local features without U^2-Net structure, the proposed approach learns these parameters from the multi-scale features generated by the U^2-Net structure, which can improve the performance of the SaAdaIN. Meanwhile, the proposed approach with the SaAdaIN makes the U^2-Net structure generate better multi-scale features compared with the conventional approach without the SaAdaIN.

In summary, the ablation study results demonstrate the effectiveness of U^2-Net structure and SaAdaIN in improving the naturalness of the converted speech in both seen-to-seen and unseen-to-unseen scenarios.

5.2 Comparison of mono-lingual conversion performance

Tables 6 and 7 present the comparison results of objective evaluation and subjective evaluation with the standard deviation of mono-lingual conversion in seen-to-seen scenario, respectively. Meanwhile, Tables 8 and 9 present the mentioned comparison results in unseen-to-unseen scenario.

From Table 6, one can find that the proposed U^2-VC and AGAIN-VC nearly have the same MCD scores, and the MCD scores of the proposed U^2-VC improves a lot compared to those of AdaIN-VC. When focusing on the predicted MOS, the proposed U^2-VC always gets the best performance compared to AGAIN-VC and AdaIN-VC. The difference of average predicted MOS is 0.2 compared to that of AGAIN-VC, up to 0.92 compared to that of AdaIN-VC, which is an obvious improvement. As stated in [18], MCD score is not always related to the human perception. The subjective evaluation is more important because the results represent the authentic naturalness and similarity of a voice conversion system. From Table 7, one can see that the proposed U^2-VC always shows the best performance compared to AGAIN-VC and AdaIN-VC in both similarity and naturalness, which means the converted speech's perceptual quality of the proposed approach is much higher. Tables 8 and 9 present the experimental results of the unseen-to-unseen scenario, which has the similar trend with the seen-to-seen scenario as presented in Tables 6 and 7.

We perform the statistical significance evaluation of MOS through ANOVA with the confidence level of 0.95 to further confirm the better performance of the proposed approach in perceptual speech quality. Tables 10 and 11 present the results in seen-to-seen scenario and unseen-to-unseen scenario, respectively. From the similarity test results, one can see that there is statistical significance between the proposed approach and AdaIN-VC in the four separate cases, but there is no obvious significance between the proposed approach and AGAIN-VC. For all

Table 3 Objective evaluation results of the ablation study on architecture in unseen-to-unseen conversion scenario. "AGAIN-VC" represents the network has neither U^2-Net structure nor SaAdaIN. "U^2-VC" represents the network has both U^2-Net structure and SaAdaIN

	MCD (dB)					Predicted MOS by NISQA				
	SF2TF	SF2TM	SM2TF	SM2TM	Average	SF2TF	SF2TM	SM2TF	SM2TM	Average
AGAIN-VC	5.95	6.03	5.96	6.02	5.99	3.71	3.75	3.82	3.93	3.80
w/o SaAdaIN, with U^2-Net	6.11	6.16	6.19	6.20	6.17	3.85	3.91	3.81	3.97	3.89
w/o U^2-Net, with SaAdaIN	6.01	6.02	5.96	6.01	6.05	3.88	3.74	3.83	3.89	3.84
U^2-VC	6.01	6.09	6.02	6.03	6.04	4.00	3.95	3.85	3.97	3.94

Table 4 Subjective evaluation results of the ablation study on architecture in seen-to-seen conversion scenario. "AGAIN-VC" represents the network has neither U^2-Net structure nor SaAdaIN. "U^2-VC" represents the network has both U^2-Net structure and SaAdaIN

	MOS (similarity)					MOS (naturalness)				
	SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Average**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Average**
AGAIN-VC	3.50	3.38	3.00	3.30	3.30	3.25	3.38	3.00	3.63	3.32
w/o SaAdaIN, with U^2-Net	3.13	3.15	3.13	3.13	3.14	3.25	3.75	3.13	3.70	3.44
w/o U^2-Net, with SaAdaIN	3.25	3.25	3.00	3.25	3.19	3.50	3.37	3.10	3.67	3.41
U^2-VC	3.63	3.38	3.39	3.69	3.53	4.00	4.13	3.69	3.91	3.93

Table 5 Subjective evaluation results of the ablation study architecture in unseen-to-unseen conversion scenario. "AGAIN-VC" represents the network has neither U^2-Net structure nor SaAdaIN. "U^2-VC" represents the network has both U^2-Net structure and SaAdaIN

	MOS (similarity)					MOS (naturalness)				
	SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Average**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Average**
AGAIN-VC	3.13	2.89	2.75	3.00	2.94	3.00	3.00	3.13	3.50	3.16
w/o SaAdaIN, with U^2-Net	3.25	3.00	2.63	3.13	3.00	3.50	3.13	3.50	3.62	3.44
w/o U^2-Net, with SaAdaIN	3.13	3.13	3.00	3.10	3.09	3.25	3.00	3.15	3.62	3.26
U^2-VC	3.25	3.25	3.23	3.13	3.22	3.75	3.88	3.80	3.88	3.83

Table 6 Objective comparison results of mono-lingual conversion in seen-to-seen scenario

	MCD (dB)					Predicted MOS by NISQA				
	SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Average**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Average**
AdaIN-VC	7.11	6.62	7.09	6.97	6.95	3.01	2.94	3.07	3.53	3.14
AGAIN-VC	6.33	6.07	6.32	6.33	6.26	3.87	3.63	3.93	4.02	3.86
U^2-VC	6.36	6.11	6.32	6.39	6.29	4.13	3.93	4.14	4.05	4.06

Table 7 Subjective comparison results with standard deviation (std dev) of mono-lingual conversion in seen-to-seen scenario. The results are listd as MOS/std dev

	MOS (similarity)/std dev					MOS (naturalness)/std dev				
	SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Average**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Average**
AdaIN-VC	2.00/0.30	2.04/0.37	2.04/0.30	2.19/0.40	2.07	2.07/0.77	2.01/0.35	2.10/0.32	2.21/0.46	2.10
AGAIN-VC	2.92/0.46	2.76/0.40	2.87/0.45	3.40/0.59	2.99	3.38/0.65	3.18/0.49	3.12/0.41	3.62/0.43	3.33
U^2-VC	3.30/0.41	3.24/0.40	3.28/0.38	4.02/0.42	3.46	3.91/0.56	3.92/0.47	3.78/0.31	4.16/0.32	3.94

Table 8 Objective comparison results of mono-lingual conversion in unseen-to-unseen scenario

	MCD (dB)					Predicted MOS by NISQA				
	SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Average**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Average**
AdaIN-VC	6.53	6.57	6.59	6.84	6.63	2.97	2.54	2.81	2.98	2.83
AGAIN-VC	5.95	6.03	5.96	6.02	5.99	3.71	3.75	3.82	3.93	3.80
U^2-VC	6.01	6.09	6.02	6.03	6.04	4.00	3.95	3.85	3.97	3.94

Table 9 Subjective comparison results with standard deviation (std dev) of mono-lingual conversion in unseen-to-unseen scenario. The results are listd as MOS/std dev

	MOS (similarity)/std dev					**MOS (naturalness)/std dev**				
	SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Average**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Average**
AdaIN-VC	2.01/0.31	2.02/0.48	2.04/0.29	2.10/0.29	2.04	2.07/0.61	2.08/0.67	2.05/0.58	2.06/0.68	2.07
AGAIN-VC	2.68/0.49	2.88/0.57	2.94/0.47	3.31/0.35	2.95	3.14/0.82	3.24/0.65	3.02/0.82	3.50/0.51	3.23
U^2-VC	3.14/0.43	3.40/0.49	3.34/0.43	3.62/0.37	3.37	3.94/0.77	4.04/0.47	3.74/0.78	4.05/0.56	3.94

Table 10 Statistical significance of the MOS results of mono-lingual conversion in seen-to-seen scenario

		Statistical significance of MOS (similarity)					**Statistical significance of MOS (naturalness)**				
		SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Overall**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Overall**
AdaIN-VC	AGAIN-VC	0.000	0.015	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000
	U^2-VC	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AGAIN-VC	AdaIN-VC	0.000	0.015	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000
	U^2-VC	0.380	0.094	0.083	0.007	0.003	0.086	0.014	0.018	0.038	0.000
U^2-VC	AdaIN-VC	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AGAIN-VC	0.380	0.094	0.083	0.007	0.003	0.086	0.014	0.018	0.038	0.000

Table 11 Statistical significance of the MOS results of mono-lingual conversion in unseen-to-unseen scenario. "Overall" represents the overall statistical analysis of all the four conversion cases

		Statistical significance of similarity					**Statistical significance of naturalness**				
		SF2TF	**SF2TM**	**SM2TF**	**SM2TM**	**Overall**	**SF2TF**	**SF2TM**	**SM2TF**	**SM2TM**	**Overall**
AdaIN-VC	AGAIN-VC	0.005	0.003	0.000	0.002	0.000	0.003	0.000	0.009	0.000	0.000
	U^2-VC	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AGAIN-VC	AdaIN-VC	0.005	0.003	0.000	0.002	0.000	0.003	0.000	0.009	0.000	0.000
	U^2-VC	0.215	0.063	0.051	0.604	0.009	0.023	0.007	0.045	0.037	0.000
U^2-VC	AdaIN-VC	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AGAIN-VC	0.215	0.063	0.051	0.604	0.009	0.023	0.007	0.045	0.037	0.000

Table 12 Objective evaluation results of voice conversion in cross-lingual scenario

	Predicted MOS by NISQA			
	VCTK2VCC	**VCC2VCTK**	**VCC2VCC**	**Average**
AdaIN-VC	2.83	2.72	2.81	2.79
AGAIN-VC	3.56	3.40	3.64	3.53
U^2-VC	3.60	3.82	3.72	3.71

Table 13 Subjective evaluation results with standard deviation (std dev) of voice conversion in cross-lingual scenario. The results are listd as MOS/std dev

	MOS (similarity)/std dev				**MOS (naturalness)/std dev**			
	VCTK2VCC	**VCC2VCTK**	**VCC2VCC**	**Average**	**VCTK2VCC**	**VCC2VCTK**	**VCC2VCC**	**Average**
AdaIN-VC	2.04/0.38	1.81/0.60	1.90/0.38	1.92	2.24/0.25	1.91/0.33	2.00/0.34	2.05
AGAIN-VC	2.94/0.45	2.47/0.58	2.68/0.44	2.70	3.19/0.50	2.74/0.48	2.58/0.60	2.84
U^2-VC	3.33/0.37	3.10/0.48	3.14/0.49	3.19	3.78/0.48	3.44/0.46	3.26/0.40	3.49

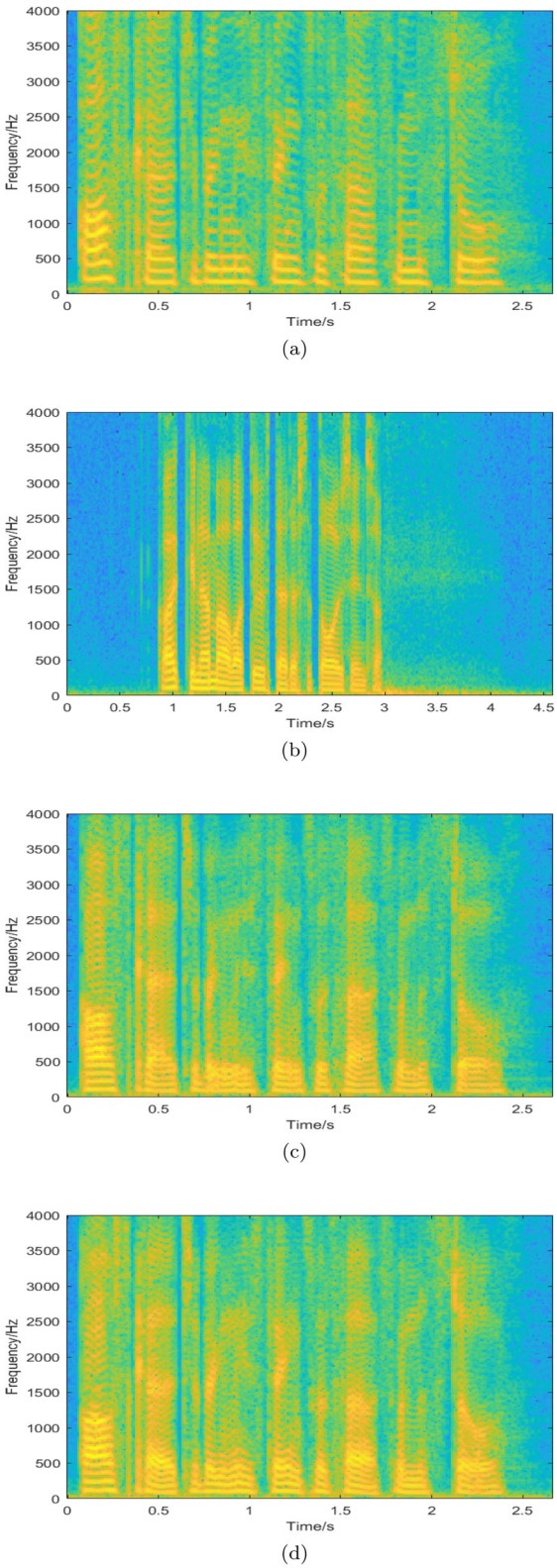

Fig. 5 Comparison among source, target, and the converted spectrograms in cross-lingual scenario(VCTK2VCC). **a** Source speech. **b** Target speech. **c** The converted speech of AGAIN-VC. **d** The converted speech of U^2-VC

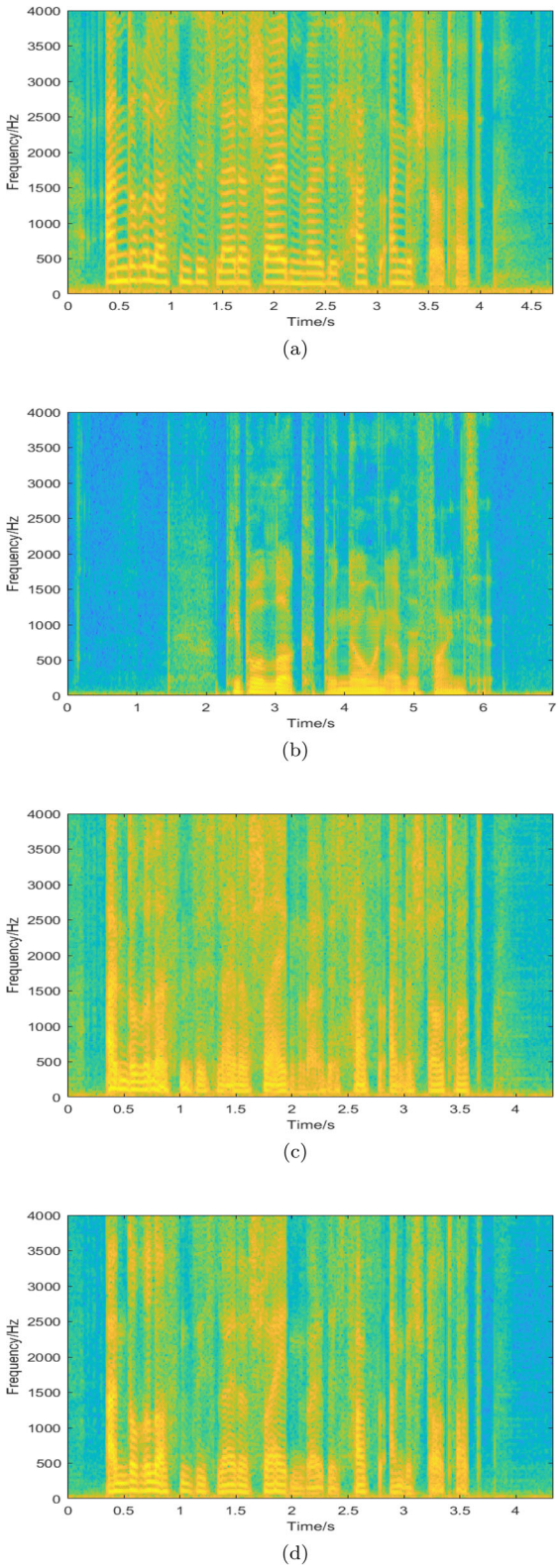

Fig. 6 Comparison among source, target, and the converted spectrograms in cross-lingual scenario(VCC2VCTK). **a** Source speech. **b** Target speech. **c** The converted speech of AGAIN-VC. **d** The converted speech of U^2-VC

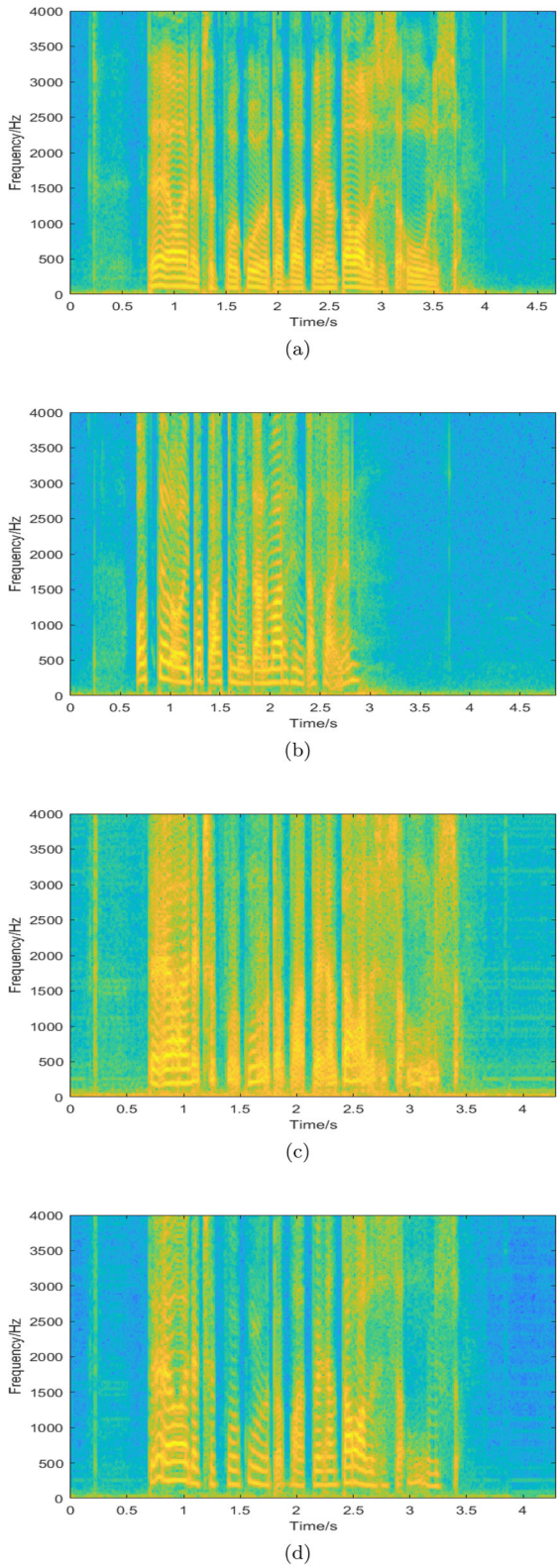

Fig. 7 Comparison among source, target, and the converted spectrograms in cross-lingual scenario(VCC2VCC). **a** Source speech. **b** Target speech. **c** The converted speech of AGAIN-VC. **d** The converted speech of U^2-VC

of the four cases, there is statistical significance between the proposed approach and the two baselines, which indicates that the proposed approach shows better performance on the similarity compared with the baselines. From the naturalness test results, there is statistical significance between the proposed approach and the baselines in both separate and overall cases. Focusing on the subjective evaluation, it is shown that the proposed approach does improve the naturalness of converted speech without degrading the speaker similarity compared with the baselines in both seen-to-seen and unseen-to-unseen scenarios.

In summary, the comparison results, especially the subjective evaluation results, show the advantage of the proposed U^2-VC in improving the quality of the converted speech.

5.3 Comparison of cross-lingual conversion performance

For cross-lingual conversion evaluation, we set 3 conversion cases which are VCTK2VCC, VCC2VCTK, and VCC2VCC. The language of source speech has to be English and Mandarin, while the target speech can be any language in these two corpuses. The results of objective evaluation and subjective evaluation are shown in Tables 12 and 13, respectively.

From Tables 12 and 13, one can see that the proposed U^2-VC always gets the best performance compared to AdaIN-VC and AGAIN-VC in both objective evaluation and subjective evaluation with the standard deviation. The evaluation results demonstrate that even though the proposed approach somewhat degrades its performance in cross-lingual scenario, it is still much better than the competing approaches. The spectrograms of the converted speech signals in Figs. 5, 6, and 7 show that the proposed approach can solve the problems of harmonic distortion and content loss, which makes the converted speech sound better in both similarity and naturalness.

Statistical significance evaluation of the MOS results is also performed to verify the advantage of the proposed approach in cross-lingual scenario. Table 14 presents the evaluation result, and one can also see that the proposed approach actually improves the naturalness of converted speech without sacrificing the speaker similarity, which has the same conclusion as that of mono-lingual conversion.

In summary, the proposed U^2-VC has the potential to improve the system robustness in cross-lingual scenario, which is proved by the evaluation results compared to AdaIN-VC and AGAIN-VC.

6 Conclusion

In this paper, we propose U^2-VC, which is a new one-shot voice conversion system with U^2-Net and SaAdaIN. The proposed approach has the capability of extracting

Table 14 Statistical significance of the MOS results in cross-lingual conversion scenario. "Overall" represents the overall statistical analysis of all the three conversion cases

		Statistical significance of similarity				Statistical significance of naturalness			
		VCTK2VCC	VCC2VCTK	VCC2VCC	Overall	VCTK2VCC	VCC2VCTK	VCC2VCC	Overall
AdaIN-VC	AGAIN-VC	0.001	0.041	0.004	0.000	0.001	0.002	0.029	0.000
	U^2-VC	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AGAIN-VC	AdaIN-VC	0.001	0.041	0.004	0.000	0.001	0.002	0.029	0.000
	U^2-VC	0.090	0.049	0.070	0.001	0.017	0.007	0.013	0.001
U^2-VC	AdaIN-VC	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	AGAIN-VC	0.090	0.049	0.070	0.001	0.017	0.007	0.013	0.001

detailed features of input log-mel spectrogram, which can improve the quality of the converted speech, especially its naturalness. In the near future, we will focus on improving the robustness of the proposed approach in more challenging scenarios, such as noisy and reverberant environments.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13636-021-00226-3.

Additional file 1: Demo samples of the proposed approach. The conversion samples are available on the demo page: https://tjulfk.github.io/U2VC/.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61631016 and Grant 61501410, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132018XNG1805.

Authors' contributions

Fangkun Liu and Hui Wang: software and writing—original draft. Renhua Peng: platform and writing—review and editing. Chengshi Zheng and Xiaodong Li: supervision and writing-review and editing. All authors read and approved the final manuscript.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author details

[1]State Key Laboratory of Media Convergence and Communication, Communication University of China, 100024 Beijing, China. [2]Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, 100190 Beijing, China. [3]University of Chinese Academy of Sciences, 100049 Beijing, China.

Received: 20 June 2021 Accepted: 11 October 2021
Published online: 24 November 2021

References

1. K. Qian, Y. Zhang, S. Chang, D. Cox, M. Hasegawa-Johnson, Unsupervised speech decomposition via triple information bottleneck (2021). http://arxiv.org/abs/2004.11284. Accessed 2020
2. L. W. Chen, H. Y. Lee, T. Yu, in *Interspeech 2019*. Generative adversarial networks for unpaired voice transformation on impaired speech (ISCA, 2019). https://doi.org/10.21437/interspeech.2019-1265
3. K. Nakamura, T. Toda, H. Saruwatari, K. Shikano, Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech. Speech Commun. **54**(1), 134–146 (2012)
4. M. Zhang, X. Wang, F. Fang, H. Li, J. Yamagishi, in *Interspeech 2019*. Joint training framework for text-to-speech and voice conversion using multi-source tacotron and wavenet (ISCA, 2019). https://doi.org/10.21437/interspeech.2019-1357
5. S. Zhao, T. H. Nguyen, H. Wang, B. Ma, in *Proc. Interspeech 2020*. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion, (2020), pp. 2927–2931. https://doi.org/10.21437/Interspeech.2020-1163
6. L. Sun, S. Kang, K. Li, H. Meng, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Voice conversion using deep bidirectional long short-term memory based recurrent neural networks (IEEE, 2015). https://doi.org/10.1109/icassp.2015.7178896
7. T. Kaneko, H. Kameoka, K. Hiramatsu, K. Kashino, in *Interspeech 2017*. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks (ISCA, 2017). https://doi.org/10.21437/interspeech.2017-970
8. T. Kou, H. Kameoka, T. Kaneko, N. Hojo, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Atts2s-vc: sequence-to-sequence voice conversion with attention and context preservation mechanisms (IEEE, 2019). https://doi.org/10.1109/icassp.2019.8683282
9. J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, L.-R. Dai, Sequence-to-sequence acoustic modeling for voice conversion. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(3), 631–644 (2019). https://doi.org/10.1109/TASLP.2019.2892235
10. H. Kameoka, T. Kaneko, K. Tanaka, N. Hojo, in *2018 IEEE Spoken Language Technology Workshop (SLT)*. Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks, (2018), pp. 266–273. https://doi.org/10.1109/SLT.2018.8639535
11. T. Kaneko, H. Kameoka, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Cyclegan-vc: non-parallel voice conversion using cycle-consistent adversarial networks, (2018), pp. 2100–2104. https://doi.org/10.23919/EUSIPCO.2018.8553236
12. T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Cyclegan-vc2: improved cyclegan-based non-parallel voice conversion, (2019), pp. 6820–6824. https://doi.org/10.1109/ICASSP.2019.8682897
13. T. Kaneko, H. Kameoka, K. Tanaka, N. Hojo, in *Proc. Interspeech 2020*. CycleGAN-VC3: examining and improving CycleGAN-VCs for mel-spectrogram conversion, (2020), pp. 2017–2021. https://doi.org/10.21437/Interspeech.2020-2280
14. W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, H.-M. Wang, in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. Voice conversion based on cross-domain features using variational auto encoders, (2018), pp. 51–55. https://doi.org/10.1109/ISCSLP.2018.8706604
15. R. Levy-Leshem, R. Giryes, in *2020 28th European Signal Processing Conference (EUSIPCO)*. Taco-vc: a single speaker tacotron based voice conversion with limited data, (2021), pp. 391–395. https://doi.org/10.23919/Eusipco47968.2020.9287448
16. T. Li, Y. Liu, C. Hu, H. Zhao, Cvc: contrastive learning for non-parallel voice conversion (2021). https://doi.org/10.21437/interspeech.2021-137

17. H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, T. Toda, Many-to-many voice transformer network. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 656–670 (2021). https://doi.org/10.1109/TASLP.2020.3047262
18. B. Sisman, J. Yamagishi, S. King, H. Li, An overview of voice conversion and its challenges: from statistical modeling to deep learning. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 132–157 (2021). https://doi.org/10.1109/TASLP.2020.3038524
19. K. Qian, Y. Zhang, S. Chang, X. Yang, M. Hasegawa-Johnson, AUTOVC: zero-shot voice style transfer with only autoencoder loss (2019). http://arxiv.org/abs/1905.05879. Accessed 2019
20. D.-Y. Wu, Y.-H. Chen, H.-Y. Lee, VQVC+: one-shot voice conversion by vector quantization and U-Net architecture (2020). http://arxiv.org/abs/2006.04154. Accessed 2020
21. J.-C. Chou, H.-Y. Lee, in *Proc. Interspeech 2019*. One-shot voice conversion by separating speaker and content representations with instance normalization, (2019), pp. 664–668. https://doi.org/10.21437/Interspeech.2019-2663
22. Y.-H. Chen, D.-Y. Wu, T.-H. Wu, H.-Y. Lee, in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Again-vc: a one-shot voice conversion using activation guidance and adaptive instance normalization, (2021), pp. 5954–5958. https://doi.org/10.1109/ICASSP39728.2021.9414257
23. X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, M. Jagersand, U^2-net: going deeper with nested u-structure for salient object detection. Patt. Recog. **106**, 107404 (2020). https://doi.org/10.1016/j.patcog.2020.107404
24. O. Ronneberger, P. Fischer, T. Brox, *U-net: convolutional networks for biomedical image segmentation*. (Springer, Cham, 2015)
25. X. Gong, W. Chen, T. Chen, Z. Wang, Sandwich batch normalization (2021). http://arxiv.org/abs/2102.11382. Accessed 2021
26. R. Kubichek, in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. Mel-cepstral distance measure for objective speech quality assessment, vol. 1, (1993), pp. 125–1281. https://doi.org/10.1109/PACRIM.1993.407206
27. G. Mittag, B. Naderi, A. Chehadi, S. Möller, NISQA: a deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets (2021). http://arxiv.org/abs/2104.09494. Accessed 2021
28. L. Wan, Q. Wang, A. Papir, I. L. Moreno, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Generalized end-to-end loss for speaker verification, (2018), pp. 4879–4883. https://doi.org/10.1109/ICASSP.2018.8462665
29. A. van den Oord, O. Vinyals, K. Kavukcuoglu, Neural discrete representation learning (2017). http://arxiv.org/abs/1711.00937. Accessed 2018
30. X. Huang, S. Belongie, in *2017 IEEE International Conference on Computer Vision (ICCV)*. Arbitrary style transfer in real-time with adaptive instance normalization, (2017), pp. 1510–1519. https://doi.org/10.1109/ICCV.2017.167
31. Y. Kwon, S.-W. Chung, H.-S. Heo, H.-G. Kang, Learning in your voice: non-parallel voice conversion based on speaker consistency loss (2020). http://arxiv.org/abs/2011.02168. Accessed 2020
32. K.-W. Kim, S.-W. Park, M.-C. Joe, Assem-VC: realistic voice conversion by assembling modern speech synthesis techniques (2021). http://arxiv.org/abs/2104.00931. Accessed 2021
33. W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, T. Toda, Pretraining techniques for sequence-to-sequence voice conversion. IEEE/ACM Trans. Audio Speech Lang. Process. **29**, 745–755 (2021). https://doi.org/10.1109/TASLP.2021.3049336
34. C. Veaux, J. Yamagishi, K. MacDonald, CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. University of Edinburgh. The Centre for Speech Technology Research (CSTR) (2017). https://doi.org/10.7488/ds/1994
35. K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, A. Courville, MelGAN: generative adversarial networks for conditional waveform synthesis (2019). http://arxiv.org/abs/1910.06711. Accessed 2019
36. I. Loshchilov, F. Hutter, Decoupled weight decay regularization (2017). http://arxiv.org/abs/1711.05101. Accessed 2019
37. A. W. Black, H. T. Bunnell, Y. Dou, P. Kumar Muthukumar, F. Metze, D. Perry, T. Polzehl, K. Prahallad, S. Steidl, C. Vaughn, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Articulatory features for expressive speech synthesis, (2012), pp. 4005–4008. https://doi.org/10.1109/ICASSP.2012.6288796
38. T.-H. Huang, J.-H. Lin, H.-Y. Lee, in *2021 IEEE Spoken Language Technology Workshop (SLT)*. How far are we from robust voice conversion: a survey, (2021), pp. 514–521. https://doi.org/10.1109/SLT48900.2021.9383498

Publisher's Note