Noisy-to-Noisy Voice Conversion Framework with Denoising Model

Chao Xie*, Yi-Chiao Wu†, Patrick Lumban Tobing†, Wen-Chin Huang* and Tomoki Toda†
* Graduate School of Information Science, Nagoya University, Nagoya, Japan
E-mail: {xie.chao, wen.chinhuang}@g.sp.m.is.nagoya-u.ac.jp
† Information Technology Center, Nagoya University, Nagoya, Japan
E-mail: {yichiao.wu, patrick.lumbantobing}@g.sp.m.is.nagoya-u.ac.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—In a conventional voice conversion (VC) framework, a VC model is often trained with a clean dataset consisting of speech data carefully recorded and selected by minimizing background interference. However, collecting such a high-quality dataset is expensive and time-consuming. Leveraging crowd-sourced speech data in training is more economical. Moreover, for some real-world VC scenarios such as VC in video and VC-based data augmentation for speech recognition systems, the background sounds themselves are also informative and need to be maintained. In this paper, to explore VC with the flexibility of handling background sounds, we propose a noisy-to-noisy (N2N) VC framework composed of a denoising module and a VC module. With the proposed framework, we can convert the speaker's identity while preserving the background sounds. Both objective and subjective evaluations are conducted, and the results reveal the effectiveness of the proposed framework.**

***Index Terms*: noisy-to-noisy voice conversion, denoise, background sounds separation, deep learning.**

I. Introduction

Voice conversion (VC) is a technique to convert the voice characteristics of a source speaker into that of a target speaker while preserving the linguistic contents. With the advent of deep learning, VC also enters a new era by dramatically improving the naturalness and similarity of the converted speech. According to the latest Voice Conversion Challenge (VCC) [1] held in 2020, the state-of-the-art method [2] shows that the similarity is comparable to natural target speech with slight disparity for naturalness.

However, in real-world scenarios, we can not always get a large amount of high-quality VC data as it is very costly to collect them. Although background noise sounds usually interfere with the input speech signal, it would be much appreciated to leverage such mega data to train a VC model in a data-driven technique. Therefore, it is essential to suppress the background noise to achieve better VC performance. However, speaking aside from conventional VC, as in VCC, we do not always filter out the background noise obtained from real-world speech signals. Consider VC usage in a video or a movie; it is essential to only convert the speech segments and preserve the background sounds. In other cases, such as VC-based speech data augmentation [3] for automatic speech recognition (ASR), the background noise is a valuable resource that further improves the robustness of the downstream system. Therefore, flexibly dealing with the background sounds in VC is more beneficial in general.

The majority of previous research works, such as [4], [5], [6] focus on noise-robust VC, in which the background sounds are considered as interference to be discarded. These works employ the use of noisy input speech and clean target speech. On the other hand, there has been proposed a text-to-speech method [7] that can convert noisy speech while controlling the noise. To disentangle the speaker identity and the noise attributes, the method augments the clean training set with a copy that mixes with the noise clips but reuses the same transcript and speaker label. By doing so, two latent variables can be used to represent speaker identity and noise attributes, respectively. They are modeled by the variational autoencoder (VAE) and introduced to condition the generative process so that both the speaker identity and the background noise are controllable.

In this paper, we propose a noisy-to-noisy (N2N) VC as a new VC framework, where speaker conversion is achieved while maintaining input background noise without linguistic input/supervision. The noisy-to-noisy VC signifies that the available dataset for VC training contains only noisy speech signals; therefore, we cannot simply train a noisy-to-clean speech model. To handle such a noisy speech dataset in VC training, as a first step, we propose to utilize a denoising module for separating speech signal and noise signal, where this denoising module is developed beforehand with publicly available datasets. Then, we propose to employ a conversion network that is developed with the use of the denoised VC dataset. Finally, in the conversion phase, the separated background noise is added back to the converted speech to maintain the input background sounds. We will show that the utilization of the denoising module can enhance the N2N VC system compared to the usage of purely noisy signals in the development of the conversion network. The main contributions of our work are as follows:

- Unlike previous works focusing on noisy-to-clean VC (noise-robust VC), we aim at noisy-to-noisy VC in our work. The first "noisy" means only noisy data are available for the VC task. The second "noisy" indicates that the background sounds are maintained, and we can add the background sounds back or suppress them based on different scenarios.
- In this work, we integrate the state-of-the-art denoising

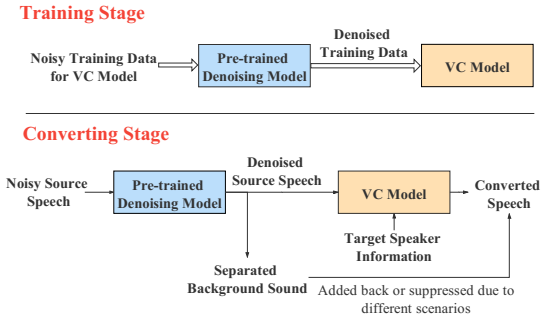

Fig. 1. The overall workflow of our proposed N2N VC framework.

model, Deep Complex Convolution Recurrent Network (DCCRN) [8], and speaker VC model, vector-quantized variational autoencoder (VQ-VAE), into the proposed N2N VC framework. We adopt the DCCRN to separate speech and background sounds, and VQ-VAE receives the denoised data as training/conversion input.

- To further investigate how the denoising model would influence the downstream VC performance, we insert another famous denoising model Conv-TasNet [9] into our framework to compare with DCCRN.
- We conduct objective and subjective evaluation, and the experimental results indicate that our method achieves an acceptable conversion performance with well-preserved background sounds.

II. Noisy-to-Noisy Voice Conversion Framework

Our framework is composed of a denoising module and a VC module. Fig. 1 illustrates the overall workflow. In our framework, the denoising module is pre-trained on the mega dataset to guarantee ideal denoising performance, and it is utilized as a separation model to separate the speech and the background sounds:

$$b(t) = x_n(t) - x_e(t), \tag{1}$$

where $b(t)$ denotes the estimated background sounds signal in the time-domain. x_n and x_e represent the time-domain noisy speech signal and the estimated speech signal, respectively. As mentioned previously, N2N VC is proposed to address the situation that only a noisy VC dataset is available, and the background noise is required to be preserved. In the training stage, the noisy VC training data pass through the denoising module, and only the denoised data are sent to train the VC module. In the conversion stage, the noisy source speech is separated by the denoising module, and only the estimated speech signal is delivered to the VC module. After the conversion, the separated background sounds can be either added back or dropped out, based on individual scenarios.

III. Framework Implementation

The motivation of our method comes from the encouraging results of the speech enhancement (SE) domain, wherein the latest Deep Noise Suppression (DNS) Challenge 2020 [10], DCCRN [8] has demonstrated the state-of-the-art performance. In line with our hypotheses, we expect that a reliably tested denoising module could bring reasonable improvements in developing a conversion network for N2N VC, where we do not have clean signals for the VC dataset. On the other hand, for the conversion network, we propose to utilize a non-parallel and linguistically unsupervised module based on VQ-VAE, which has been shown to be capable of performing the disentanglement of content and speaker information better compared to conventional variational autoencoder and autoencoder [11].

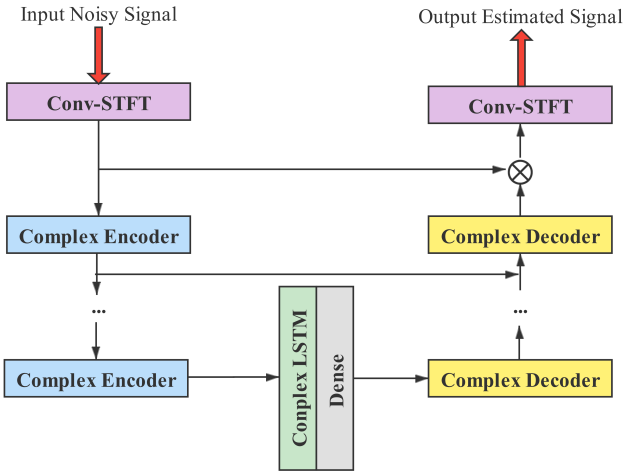

Fig. 2. The overall structure of DCCRN.

A. Denoising module: DCCRN

DCCRN is a convolution recurrent network (CRN) based single-channel denoising model. Fig. 2 shows the overall structure of DCCRN. Two-dimensional convolution (Conv2D) blocks are stacked to constitute the encoder/decoder. Each Conv2D block consists of a convolution/deconvolution layer along with batch normalization and activation function. The DCCRN has been shown to outperform conventional CRN [12] by a large margin thanks to the handling of the problems of complex calculation that are observed in the CRN. Specifically, complex convolution neural network, complex batch normalization layer, and complex long short-term memory (LSTM) are implemented for encoder/decoder, guaranteeing that the DCCRN can model the correlation between magnitude and phase. More details can be found in [8].

In our work, since we utilize DCCRN as a separation model, the power of the estimated speech should be matched to the clean target speech. Hence, the original scale-invariant signal-to-noise ratio (SI-SNR) loss [9] is replaced by scale-dependent signal-to-distortion (SD-SDR) loss [13], which is formulated as:

$$\text{SD-SDR} = 10\log_{10}\left(\frac{\|\alpha s\|^2}{\|s-\hat{s}\|^2}\right), \tag{2}$$

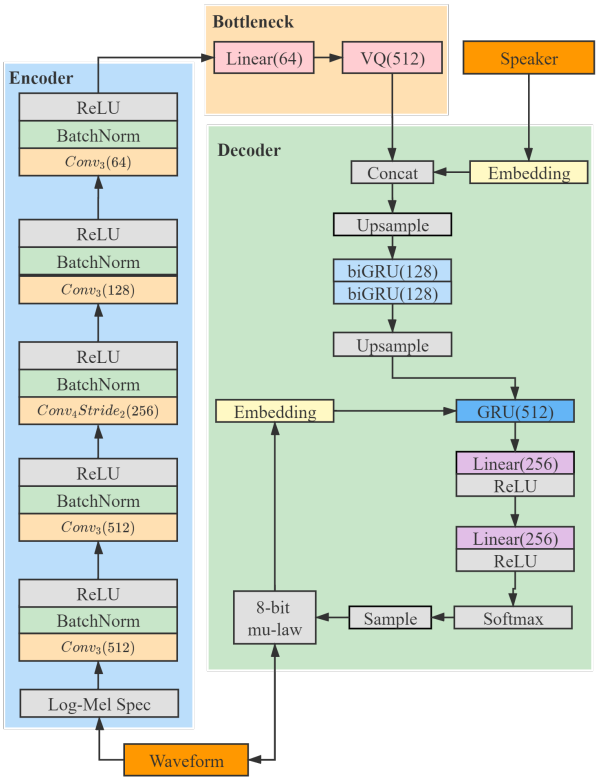

Fig. 3. The model structure of the VQ-VAE: An encoder (blue) encodes the log mel-spectrogram into latent representation and passes to the VQ bottleneck (orange). The decoder (green) then reconstructs the waveform from the discrete representation using an autoregressive stream and a speaker embedding as condition. The subscript number of *Conv* represents its convolutional kernel size, and that of *Stride* means the length of stride in the convolutional computation.

where s and $\hat{s}$ indicate the target signal and the estimate of the target respectively, and α denotes an optimal scaling factor defined as:

$$\alpha = \hat{s}^T s / \|s\|^2. \tag{3}$$

B. VC module: VQ-VAE

In our work, as illustrated in Fig. 3, we implement a VQ-VAE-based VC module [14], which has three main components: encoder, bottleneck layer, and decoder.

The encoder consists of five one-dimensional convolution (Conv1D) blocks, and each block is composed of a convolution layer along with batch normalization and activation function. The input log mel-spectrogram sequence $\{x_t, t = 1, \ldots, T\}$ is computed as a stream of latent vectors $\{z_j, j = 1, \ldots, N\}$ by the encoder and sent to a vector-quantized bottleneck with a 64-dimensional trainable codebook $\{e_i, i = 1, \ldots, 512\}$ to discard speaker information. In the forward pass, the latent vectors of the encoder $\{z_j, j = 1, \ldots, N\}$ are mapped into the nearest vectors in the codebook by

$$k = \arg\min_i \|z_j - e_i\|^2, \tag{4}$$

and z_j is replaced with e_k as a discrete latent representation $\{\hat{z}_j, j = 1, \ldots, N\}$. The decoder adopts a lightweight recurrent network to reconstruct the waveform based on the embedded speaker identity information and the discrete representation $\{\hat{z}_j, j = 1, \ldots, N\}$ from the VQ bottleneck in an autoregressive manner that predicts the current sample based on the past ones.

In the backward pass, the gradient of the loss through the codebook is approximated via the straight-through estimator [15], due to that the $argmin$ is not differentiable. The values of the codebokk are updated by exponential moving averages [16]. Additionally, a commitment loss [16] is introduced to encourage the output vector of the encoder z_j to be close to its selected vector e_k of the codebook. The VQ-VAE is trained to minimize a sum of two loss terms: the negative log-likelihood of the reconstruction loss and the commitment loss as follows:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T} \log p\left(x_t \mid \hat{x}_t\right) + \beta \frac{1}{N}\sum_{j=1}^{N} \|z_j - \mathrm{sg}\left(\hat{z}_j\right)\|^2, \tag{5}$$

where $\{\hat{x}_t, t = 1, \ldots, T\}$ is the output sequence of the decoder. β is the commitment weight and set to 0.25 according to [16], and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation.

IV. Experimental Evaluations

We conducted experimental evaluations to investigate the effectiveness of the proposed N2N VC framework. Since we focus on VC application for telecommunication, such as telephone speech conversion or data augmentation for speaker recognition of telephone speech, as one of our target applications of N2N VC, we used 8 kHz sampled speech data in the experimental evaluations.

A. Dataset

1) Dataset for denoising model: For the training of the denoising model, we used DNS Challenge 2020 dataset [10], which is a vast and high-quality dataset for the SE task. The dataset consisted of two sub-datasets: the clean speech dataset and the noise dataset. The clean speech dataset was derived from a dataset of public audiobooks, Librivox [17]. The organizers of the DNS Challenge had already cherry-picked the speech files via subjective quality evaluation. The resulting clean speech dataset had 500 hours of speech from 2,150 speakers in various languages, most of which were in English. 6,000 speech clips were randomly sampled as the validation data.

The noise dataset was collected from Audioset [18] and Freesound [19]. Preprocessed by the organizers, the selected dataset had about 150 audio classes and a total of 65,000 audio clips. 500 clips were randomly picked into the validation set. We built up the noisy dataset by uniformly sampling a noise clip and adding it to a clean speech. The SNR levels were also sampled from a uniform distribution between 5 and 20 dB.

TABLE I
THE OBJECTIVE EVALUATION RESULTS OF DCCRN AND CONV-TASNET, HIGHER IS BETTER

(a) SI-SDR and SAR on the noisy VCC training dataset

Model	Eval. Target	SI-SDR (dB)					SAR (dB)				
		7 dB	11 dB	15 dB	19 dB	Avg.	7 dB	11 dB	15 dB	19 dB	Avg.
DCCRN	Speech	17.86	20.17	22.73	25.19	**21.49**	18.55	20.82	23.32	25.76	**22.11**
	Background sounds	10.48	8.55	6.87	4.86	**7.69**	11.27	9.32	7.65	5.68	**8.48**
Conv-TasNet	Speech	15.39	18.01	20.46	22.77	19.16	16.04	18.66	21.05	23.44	19.80
	Background sounds	7.45	5.64	3.42	1.06	4.39	8.27	6.53	4.29	1.84	5.23

(b) PESQ and STOI on the noisy VCC training dataset

Model	PESQ					STOI				
	7 dB	11 dB	15 dB	19 dB	Avg.	7 dB	11 dB	15 dB	19 dB	Avg.
DCCRN	3.20	3.41	3.57	3.72	**3.47**	0.96	0.97	0.98	0.99	**0.98**
Conv-TasNet	2.84	3.05	3.23	3.39	3.13	0.94	0.96	0.97	0.98	0.96

2) Dataset for VC model: The dataset for the VC model ought to be unseen for the denoising model. We chose VCC 2018 dataset [20] as the clean speech dataset and PNL 100 Nonspeech Sounds [21] as the noise dataset to simulate the real-world situation.

VCC 2018 dataset is a high-quality and publicly available dataset specialized for VC tasks. The speech data was recorded by professional US English speakers in a professional studio without significant noise effects. There were a total number of 972 utterances for training and 420 utterances for evaluation, involving 12 male/female speakers: 8 source speakers denoted as (VCC2SM1, VCC2SM2, VCC2SM3, VCC2SM4, VCC2SF1, VCC2SF2, VCC2SF3, VCC2SF4) and four target speakers denoted as (VCC2TM1, VCC2TM2, VCC2TF1, VCC2TF2). Each speaker uttered 81 and 35 sentences for training and evaluation, respectively, resulting in a total of around 13 minutes of audio.

The PNL 100 Nonspeech sounds consisted of 100 clips and 20 categories of environmental records, such as crowd noise, cry, tooth brushing, and so on. We uniformly sampled the noise clips to mix with the VCC 2018 train/evaluation dataset at four certain SNR levels: 7 dB, 11 dB, 15 dB, and 19 dB.

For VCC evaluation data, to guarantee that the participants of the subsequent subjective evaluation could concentrate on marking appropriate scores, the number of evaluating utterances was limited to a proper amount. Four speakers (VCC2SM3, VCC2SM4, VCC2SF3, VCC2SF4) were selected as source speakers, which resembled the non-parallel (SPOKE) task of VCC 2018 [20], and two speakers (VCC2TF2, VCC2TM2) as target speakers. Hence there were 8 conversion pairs, and each pair had 35 utterances. Since the evaluation dataset aimed to compare the conversion performance of different systems equally, the same utterances in different speakers shared the same pattern of background sounds.

B. Model training details

1) DCCRN training: We trained the DCCRN model implemented by Asteroid [22]. The type of DCCRN was "DCCRN-CL." The window length, hop size, and FFT length were set to 50 ms, 12.5 ms, and 512, respectively. We observed degradation of the denoising performance with the original settings [8], which were for a sampling rate of 16 kHz; hence, we used our own 8 kHz optimized settings. The batch size was 64. Adam was used as the optimizer and set the initial learning rate to $1 \cdot 10^{-4}$. The learning rate would decay 0.5 if the validation loss did not go down within 4 epochs. Additionally, an early stopping mechanism was introduced to choose an optimized model. It took about 22 days of training on a single RTX 3080 to get the best model.

2) VQ-VAE training: We used a PyTorch-based implementation for the VQ-VAE model [14]. Log mel-spectrogram was extracted as the input and 8 bits mu-law decoded waveform as ground-truth. The window length, the hop size, and the FFT length were set to 20 ms, 5 ms, 1024, respectively. The batch size was 64, and the optimizer was Adam with an initial learning rate of $2 \cdot 10^{-4}$. The learning rate would be halved after 300k steps. The total training steps were 600k steps, which cost around two days on a single RTX 3080.

C. Experimental setup

To demonstrate the performance of our proposed N2N VC framework, we set two models as our baseline. The first baseline was a VQ-VAE trained on the clean VCC dataset, denoted as Clean-VC. The other was the VQ-VAE directly trained on the noisy VCC dataset, denoted as Noisy-VC. In simpler terms, the Clean-VC and the Noisy-VC respectively represented the upper and lower bound of our framework.

When mixing the noisy VCC dataset, the whole PNL 100 Nonspeech was sampled for both the training and the evaluation set, which indicated that the Noisy-VC had already seen all the patterns of background sounds during training. Hence, Noisy-VC should have its own optimal performance

TABLE II
MCD ON THE CLEAN REFERENCE OF VCC EVALUATION DATASET, LOWER IS BETTER

Systems	MCD (dB)				
	SF-TF	SM-TM	SF-TM	SM-TF	Avg.
Clean-VC	7.17	7.26	7.35	7.67	7.36
DCCRN-VC	7.55	7.78	7.80	8.38	7.88
ConvTas-VC	7.55	7.86	8.0	8.26	7.92

TABLE III
NATURALNESS SCORES (MOS) WITH 95% CONFIDENCE INTERVALS ON NOISY VCC EVALUATION DATASET, HIGHER IS BETTER

Systems	MOS [1, 5]		
	7 dB	15 dB	Avg.
Clean-VC	3.46 ± 0.12	3.52 ± 0.11	3.49 ± 0.08
DCCRN-VC	3.07 ± 0.13	3.08 ± 0.12	3.08 ± 0.09
ConvTas-VC	3.0 ± 0.13	3.14 ± 0.12	3.07 ± 0.09
Noisy-VC	1.99 ± 0.11	2.15 ± 0.11	2.07 ± 0.08

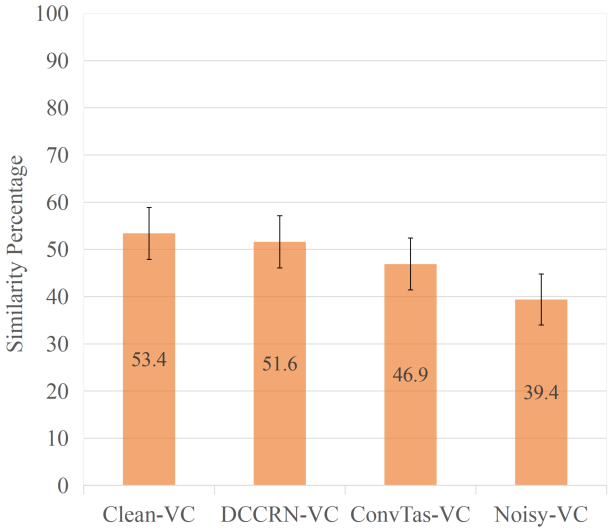

Fig. 4. Similarity percentage with 95% confidence intervals on noisy VCC evaluation dataset. The similarity percentage is defined as the added percentage of *Definitely the same* and *Maybe the same*.

on the evaluation dataset. It needs to be emphasized that for our framework, both the noise dataset and speech dataset were unseen for the denoising model.

To further probe into how the denoising model would affect the VC performance in our framework, another well-known denoising model Conv-TasNet [9] was selected as a comparison. Due to the training of the denoising model was very time-consuming, we used the pre-trained Conv-TasNet model provided by Asteroid. It was trained on the single-speaker enhancement task of the Libri3Mix dataset [23]. To differentiate the use of the denoising models, i.e., DCCRN and Conv-TasNet, we denoted our method as DCCRN-VC and that of ConvTas-VC. Overall, there were four systems to compare and evaluate:

- Clean-VC trained on clean VCC dataset.
- Noisy-VC trained on noisy VCC dataset.
- DCCRN-VC trained on DCCRN-denoised noisy VCC dataset.
- ConvTas-VC trained on ConvTasNet-denoised noisy VCC dataset.

D. Evaluation results

1) Objective Evaluation: First, the relative performance of the two denoising models, DCCRN and Conv-TasNet, was assessed with several measurements as follows: scale-invariant signal-to-distortion ratio (SI-SDR) [13], signal-to-artifact ratio (SAR), PESQ [24], and STOI [25]. In this objective evaluation, the noisy VCC training dataset was used instead of the VCC evaluation dataset, owing to the former that covered the whole of PNL 100 Nonspeech Sounds dataset compared to the latter that was consisted of only 35 different background sound clips at the most.

The results are demonstrated in Table I. It is evident that DCCRN outperforms Conv-TasNet on all metrics among all SNR levels for both speech and separated background sounds, on which we infer that DCCRN-VC would also provide better performance compared to ConvTas-VC in the following VC evaluations. We can also observe that as the SNR level increases, the SI-SDR and the SAR of the clean speech increase, while those of the separated background sounds decrease; which is reasonable considering that clean speech could be extracted more easily from a signal with higher signal-to-noise powers (higher SNR) condition rather than from a signal with lower SNR condition, and vice versa for extracting the noise signal.

To evaluate the performance of our VC model combined with the denoising process, we leveraged clean evaluation reference to assess Clean-VC, DCCRN-VC, and ConvTas-VC via measuring the mel-cepstral distortion (MCD) [26]. Table II presents the results of MCD on the VCC evaluation dataset. It can be observed that all three systems achieve better performance for the intra-gender conversions (SF-TF and SM-TM) compared to the cross-gender conversions (SF-TM and SM-TF), where the SM-TF conversion pair is the worst, and the SF-TF conversion pair is the best. The best average MCD is reasonably achieved by the Clean-VC system with a value of 7.36, and our proposed DCCRN-VC method shows a considerable gap of 7.88 with respect to the Clean-VC. Although DCCRN outperforms Conv-TasNet much in denoising tasks, the ConvTas-VC with an average MCD value of 7.92 shows only a slightly worse average MCD than the DCCRN-VC.

Additionally, we also investigated the perceptual quality of the denoised speech and the converted speech. We observed that the denoised samples of DCCRN sounded clean but with a bit of distortion, while Conv-TasNet remained residual noise throughout the audio. As for the converted speech, samples from DCCRN-VC and ConvTas-VC sounded with comparable

quality. The distortion by DCCRN leads to further distortion by the VQ-VAE downstream, and the residual noise by Conv-TasNet also degrades the performance of its VQ-VAE. As we have observed in Table II, the resulting MCD values are very close to each other, but as the MCD calculation was performed on only speech segments while the non-speech regions that could still contain the residual noise for the ConvTas-VC were discarded, it is possible that the residual noise in the ConvTas-VC causes adverse effects in a scenario when only clean converted speech needs to be presented.

2) Subjective Evaluation: As our final goal is to achieve high-quality VC while preserving the background sounds, eventually, the overall performance was evaluated through subjective evaluations. The mean opinion score (MOS) by an opinion test was applied to measure the naturalness of the converted samples. The participants were asked to give a naturalness score from 1 to 5 (higher is better). The four systems were evaluated on the same noisy VCC evaluation dataset mentioned in Section IV-A2. To guarantee that the complete subjective evaluation would not take too long so that the participants can submit high-quality answers, we further limited the amount of the evaluation data. From the evaluation dataset, six utterances were randomly selected for each conversion pair, where three utterances were set for the 7 dB SNR, and the other three were set for the 15 dB SNR. This resulted in a total number of 204 audio samples: 48 audio samples per system and 12 samples from noisy ground-truth target speech. As our goal is N2N VC, converted samples from the DCCRN-VC and the ConvTas-VC were superimposed with the respective separated background sounds. For Clean-VC, we superimposed the original record of background sounds for a fair comparison. Furthermore, the participants were required to give their scores based on the overall naturalness of both the speech and the background sounds. To assist the participants in making judgment, the category of the superimposed background sounds was given in the evaluation.

Lastly, we conducted the similarity (SIM) evaluation proposed in [20]. In the SIM test, each of the participants was presented with two audio samples at a time, consisting of a converted speech and a reference speech of the target speaker, and asked to determine whether these samples came from the same speaker. In judging each of the audio pairs, four options were given: *1. Definitely the same; 2. Maybe the same; 3. Maybe different; 4. Definitely different*. We asked the participants to ignore the quality of the speech and the background sounds and focus on the speaker similarity. From the evaluation dataset, four utterances were randomly selected for each conversion pair, where two utterances were set for the 7 dB SNR, and the other two were set for the 15 dB SNR. This resulted in a total of 128 converted samples to be evaluated by each participant and 32 audio samples per system.

The results of the MOS and of the SIM tests are shown in Table III and Fig 4, respectively, where the SIM score is defined as the sum of the percentages from *Definitely the same* and *Maybe the same* decisions. Undoubtedly, Clean-VC acquires the best performance with the MOS score of 3.49 and SIM score of 53.4 on average. Our proposed framework DCCRN-VC reached the MOS score of 3.08 and the SIM score of 51.6 on average, which is far more beyond Noisy-VC that gets 2.07 and 39.4 but still has a margin from Clean-VC. Thanks to the powerful denoising model DCCRN, our method achieves approximate scores under different SNR levels. While for Noisy-VC, which is sensitive to noise powers, it reaches better performance under higher SNR level because a higher SNR level means less noise interference. A similar situation is observed for the ConvTas-VC.

It is worth noting that ConvTas-VC reaches similar naturalness scores on average to DCCRN-VC's. DCCRN-VC only leads ConvTas-VC with a slight margin, which is consistent with the trend in MCD. As for ConvTas-VC, due to the inability of Conv-TasNet to completely remove the background noise, the residual noise also exists in the converted samples of the VQ-VAE. However, after the separated background sounds are superimposed, it is difficult to perceptually notice the interference, which allows ConvTas-VC to obtain a tolerable score in the overall naturalness evaluation. However, in another scenario, when clean converted speech is required, such residual noise will bring in adverse effects. It is worthwhile to conduct the subjective evaluation in such a scenario, which would be in our future work.

As for DCCRN-VC, although we can observe in Table I that the performance of the noise removal of the DCCRN is better than that of the Conv-TasNet, as has also been mentioned, the DCCRN introduces some artifacts that are, in turn, propagated to the VQ-VAE VC module. We believe that this is the reason that the naturalness score of the DCCRN-VC to be in the same range as that of the ConvTas-VC, which would imply that compared to the residual background noise, the unwanted artifacts produced by the denoising module cause more adverse effects to the downstream VC model in terms of audio quality.

V. Conclusions

In this paper, we have presented a noisy-to-noisy VC framework that relies on only noisy VC training data and is capable of preserving the background sounds for the converted speech waveform. Our framework consists of a state-of-the-art denoising model DCCRN and a VC model based on VQ-VAE. In the training stage, the noisy VC dataset is denoised by the denoising model, on which the VC model is trained. In the conversion stage, the noisy source speech is separated by the denoising model to get the estimated speech signal and background sounds, and the speech signal is sent to the VC model for conversion. The background sounds can be superimposed or suppressed flexibly according to a specific application. The experimental results show that our framework outperforms the conventional noisy-to-noisy VC that is directly trained on the noisy VC dataset and achieves acceptable noisy-to-noisy VC performance with room for improvement. In future work, we aim to bridge the gap between our framework and Clean-VC.

ACKNOWLEDGMENT

This work was partly supported by JST CREST Grant Number JPMJCR19A3, Japan.

REFERENCES

[1] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z.-H. Ling, and T. Toda, "Voice Conversion Challenge 2020 — Intra-lingual semi-parallel and cross-lingual voice conversion —," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[2] L.-J. Liu, Y.-N. Chen12, J.-X. Zhang, Y. Jiang12, Y.-J. Hu, Z.-H. Ling, and L.-R. Dai, "Non-parallel voice conversion with autoregressive conversion model and duration adjustment," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 126–130.

[3] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario," *Proc. Interspeech 2020*, pp. 4382–4386, 2020.

[4] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, p. 151, 2020.

[5] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," in *Eighth ISCA Workshop on Speech Synthesis*, 2013.

[6] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *2012 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 313–317.

[7] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5901–5905.

[8] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Proc. Interspeech 2020*, 2020, pp. 2472–2476.

[9] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "Icassp 2021 deep noise suppression challenge," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[11] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[12] K. Tan and D. Wang, "Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6865–6869.

[13] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[14] B. van Niekerk, L. Nortje, and H. Kamper, "Vector-Quantized Neural Networks for Acoustic Unit Discovery in the ZeroSpeech 2020 Challenge," in *Proc. Interspeech 2020*, 2020, pp. 4836–4840.

[15] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[16] A. van den Oord, O. Vinyals, and k. kavukcuoglu, "Neural discrete representation learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[17] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Reviews*, 2014.

[18] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[19] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[20] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 195–202.

[21] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.

[22] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[23] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[24] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.

[26] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.