Noisy-to-Noisy Voice Conversion Under Variations of Noisy Condition

Chao Xie and Tomoki Toda, *Senior Member, IEEE*

***Abstract*—Voiceconversion (VC) refers to the transformation of the speaker identity of a speech to the target one without altering the linguistic content. As recent VC techniques have made significant progress, implementing them in real-world scenarios is also considered, where speech data have some inevitable interferences, the most common of which are background sounds. On the other hand, background sounds are informative and need to be retained in some applications, such as VC in movies/videos. To address these issues, we have proposed a noisy-to-noisy (N2N) VC framework that does not rely on clean VC data and models the noisy speech directly by using noise as conditions. Previous experimental results have proven its effectiveness. In this article, we further improve its performance by introducing the pre-trained noise-conditioned VC model. Moreover, to further explore the impacts of introducing noise conditions, the performance in more realistic situations is evaluated in which the training set possesses speaker-dependent noisy conditions. The experimental results demonstrate the effectiveness of the pre-training strategy and the degradation of its performance under strict noisy conditions. We then proposed a noise augmentation method to overcome the limitation. Further experiments showed the effectiveness of the augmentation method.**

***Index Terms*—Voice conversion (VC), noisy-to-noisy VC, noise robustness, noisy speech modeling, data augmentation.**

I. Introduction

VOICE conversion (VC) is a technique of converting the vocal timbre of a speech from the source speaker to the target one without changing its linguistic content. VC has been extensively studied before the advent of deep learning. The early works were mainly based on statistical modeling of speech signals with parallel training data and then extended to non-parallel scenarios. Many approaches, such as exemplar-based sparse representation [1], vector quantization (VQ) [2], and Gaussian mixture modeling [3] have been proposed, establishing the foundation for the present deep-learning-based methods. With the emergence of deep learning, the components of voice conversion (VC) have undergone significant advancements, encompassing speech analysis, feature mapping, and vocoders. Neural network-based VC methods consistently push the boundaries of naturalness and similarity, as evidenced by the notable progress showcased in the recent Voice Conversion Challenge (VCC) 2020 [4]. Many approaches, such as those based on generative adversarial networks [5], [6], [7], [8], variational autoencoders (VAEs) [9], [10], [11], [12], automatic speech recognition (ASR), and text-to-speech (TTS) [13], [14], [15], are frequently focused on in the latest VC studies. Recent developments have witnessed the application of VC in novel domains, including dubbing [16], audio data augmentation [17], and other demanding real-world scenarios.

However, in contrast to controlled experimental environments where high-quality data is prepared in advance for both training and evaluation, real-world speeches often encounter various interferences, with background noise being the most prevalent. In certain scenarios where training data consists solely of noisy samples, the performance of VC experiences noticeable degradation in terms of naturalness and similarity due to the entangled factors of linguistic content, speaker identity, and background noise, as demonstrated in [18], [19]. On the other hand, background sounds are informative and may need to be preserved in some cases. For instance, VC for singing focuses on converting the vocals with the accompaniments being removed beforehand to ensure the quality of the conversion. However, it is crucial to keep the accompaniments intact during inference, for they can be optionally layered over the converted vocals. Similar considerations are required in VC-based data augmentation [17], where the inherent background noise is also a resource to enhance a system's robustness. Nevertheless, most of the related studies [20], [21], [22], [23], [24], [25] primarily focus on noise-robust VC, in which the background sounds are typically discarded as undesirable interference. Furthermore, most of these studies rely on clean source/target speech data for training.

To address the above issues, we have proposed a noisy-to-noisy (N2N) VC framework [18]. For the first "noisy", our method does not require clean data: all the source/target training data can be noisy. For the second "noisy", the background sound in the converted sample can either be preserved or removed, depending on specific scenarios. The proposed framework comprises off-the-shelf denoising and VC modules. The denoising module is utilized to separate the speech and noise signals in the time domain. The VC module is trained using denoised speech. During inference, only the estimated speech signal is converted. The separated noise can then optionally be superimposed, depending on the specific scenario. However, using the denoising module introduces undesirable distortion that can significantly impact the performance of the VC downstream. Therefore, the

Manuscript received 31 January 2023; revised 27 June 2023 and 13 August 2023; accepted 14 August 2023. Date of publication 20 September 2023; date of current version 20 October 2023. This work was supported by JST CREST under Grant JPMJCR19A3, Japan. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jong Won Shin. *(Corresponding author: Chao Xie.)*

Chao Xie is with the Graduate School of Informatics, Nagoya University, Aichi 464-8601, Japan (e-mail: xie.chao@g.sp.m.is.nagoya-u.ac.jp).

Tomoki Toda is with Information Technology Center, Nagoya University, Aichi 464-8601, Japan (e-mail: tomoki@icts.nagoya-u.ac.jp).

Digital Object Identifier 10.1109/TASLP.2023.3313426

framework is improved in [26] by leveraging the noise signal as a condition within the VC module to model the noisy speech to alleviate the distortion introduced by the denoising module. The experimental results show that this modification shortened the gap in the mean opinion score (MOS) of the naturalness from the upper bound by up to 65% [26].

In this article, we describe more details of the proposed N2N VC framework. As the noise-conditioned VC method is novel, our initial evaluation focuses on its performance in generating clean converted samples, assessing whether introducing the noise condition within the VC model affects its VC efficacy and whether it is comparable to the performance of the original VC model. To further enhance the robustness of our method in noisy environments, we adopt the pre-training strategy which has been proven beneficial in numerous natural language processing (NLP) and VC studies. In our previous work [26], the training set has an 8 kHz sampling rate and speaker-independent (SI) noisy conditions, where the clean corpus is disrupted by various types of noise clip at several signal-to-noise ratios (SNRs). In this article, we extend the experimental conditions to a 16 kHz sampling rate and a more realistic speaker-dependent (SD) noise environment, where each speaker's utterances have a fixed type of background noise and SNR so that the noisy conditions and speaker information are correlated. However, results show that the performance of the N2N framework degrades. We further proposed three data augmentation methods to address the disentanglement of speaker identity and noisy conditions, and one of the methods successfully mitigates the performance degradation, as substantiated by our experimental results.

II. Related Work

In general, most VC methods rely on high-quality training data and may become fragile when exposed to noisy environments, which can limit their practical applications. Compared to conventional VC which has been extensively studied, relatively fewer research efforts have been dedicated to noise-robust VC.

A. Noise-Robust VC With Statistical Methods

In the realm of statistic-based methods, Takashima et al. [20] proposed an exemplar-based noise-robust VC method, in which a noisy speech is represented as a weighted combination of both speech and noise exemplars using non-negative matrix factorization. During inference, the clean target speech is generated with the target speech exemplars and the weighted source exemplars, but without the noise ones. Both clean source and target speech data are needed for training this method.

B. Noise-Robust VC With Speech Enhancement Methods

As a pioneer, Valentini-Botinhao et al. [21] proposed a cascading method for TTS in noisy environments. A recursive neural network (RNN)-based speech enhancement (SE) model is used to denoise a noisy waveform before it is passed onto the TTS module. Clean references are necessary as the SE and TTS models are jointly trained. In a similar vein, Chan et al. [22] adopted a lightweight SE component to preprocess noisy data for VC downstream, necessitating clean speech data because the SE and VC components are jointly trained. Miao et al. [23] realized noise-robust VC by implementing two filtering methods in the pre- and post-processing stages. Furthermore, the dimensions of the input feature Mel-cepstral coefficients (MCEPs) are extended, and only the sub-band cepstrum is converted to alleviate the interference in high-quefrency components. Choi et al. [27] proposed a three-module cascading VC framework consisting of one VC and two SE modules to deal with background noise and reverberation in real-world scenarios separately. Each SE module's preprocessing is designed to control noise and reverberation independently in the converted samples. Given that the two SE modules are pre-trained and fixed, the framework does not necessitate clean training data for VC.

C. Noise-Robust VC With Training Strategies

Studies on ways to mitigate the impact of noise on VC have also been conducted in terms of training strategies. Du et al. [24] presented a noise-robust framework adopting domain adversarial training (DAT). The framework is based on a zero-shot VC method AdaIN-VC [28] consisting of a speaker encoder, content encoder, and decoder. It receives clean and noisy data as the input while only predicting the clean reconstructed data during training. A gradient reversal layer and a domain classifier are added after each encoder, and a DAT loss term is introduced, enabling each encoder to extract noise-invariant speaker and content representations from both clean and noisy speech. In another study, Huang et al. [25] explored general degradation-robust VC via denoising and adversarial training. They considered three major types of degradation: background noise, reverberation, and band rejection. Additionally, adversarial samples generated by embedding attack [29] were adopted to enhance robustness further. During training, 60% of clean data in a mini-batch are augmented with randomly chosen degradation, then 50% of the data in the batch are further applied to generate adversarial examples. Consequently, the VC model processes clean data, augmented data with distortion, and adversarial examples all in one mini-batch, whereas the loss is calculated using the corresponding clean data, effectively integrating denoising and adversarial training.

Generally, most related studies concentrate on noise-robust VC, where the background noise is suppressed, generating only clean converted speech. Moreover, in the majority of cases, clean utterances from source/target speakers are essential. Significantly, Hsu et al. [30] proposed a TTS-based VC method capable of controlling the background noise within the converted speech. The training dataset is augmented by duplicating the clean speech and mixing it with additive noise while maintaining the transcript and speaker label. A VAE is used to factorize the speaker's identity and the noise condition from the noisy speech. The factorization is further improved by domain adversarial training. Therefore, two latent factors of speaker characteristics and background noise are controllable. However, the quality of

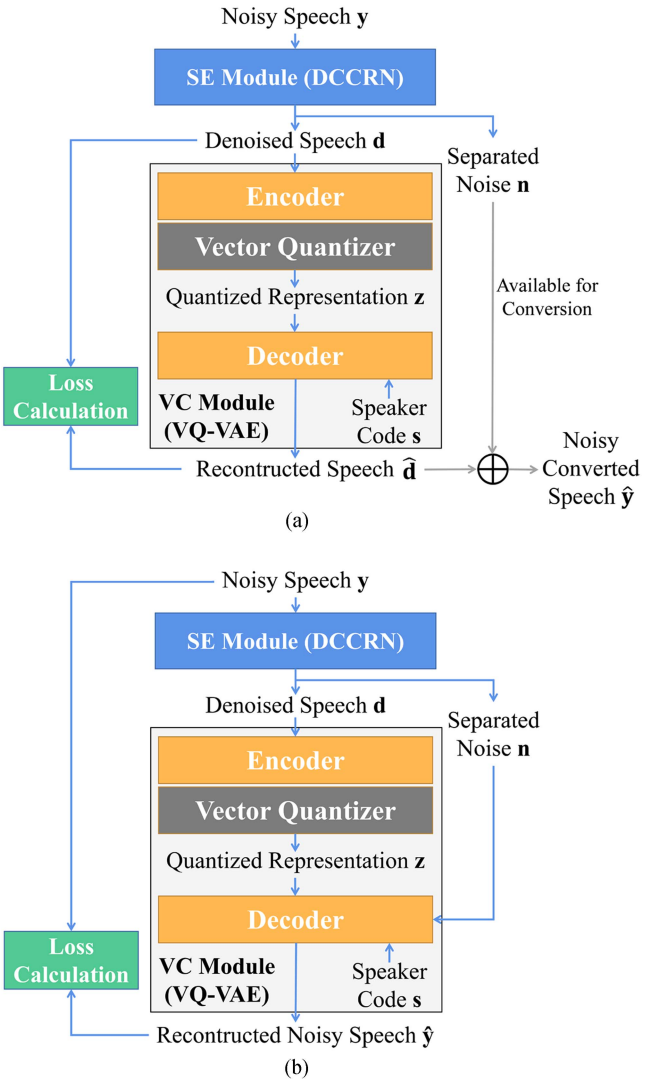


Fig. 1. Overall workflow of the proposed N2N VC framework. (a) Baseline framework. (b) Noise-conditioned VC framework.

the generated noise is subpar, and clean speech data remains required.

III. Proposed Noisy-to-Noisy VC Methods

A. Baseline Framework

We have proposed a naive N2N VC in [18], which serves as the baseline and whose workflow is illustrated in Fig. 1(a). Given that only noisy training data are available, and recent SE methods have achieved significant success [31] [32] showing promising applications in assisting downstream tasks in noisy scenarios, the baseline follows the cascading design consisting of off-the-shelf SE and VC modules. The SE module functions to separate the speech and noise signals in the time domain:

$$\mathbf{n} = \mathbf{y} - \mathbf{d}, \tag{1}$$

where $\mathbf{y} = \{y_1, \ldots, y_T\}$ denotes the noisy speech waveform, $\mathbf{d} = \{d_1, \ldots, d_T\}$ denotes the denoised speech waveform estimated by the SE module, and $\mathbf{n} = \{n_1, \ldots, n_T\}$ is the separated noise waveform. Note that only background noise is focused on in this article. We left the consideration of reverberation for future work.

As a case study, the Deep Complex Convolution Recurrent Network (DCCRN) [33] and a VQ-VAE-based non-parallel VC method [34] are chosen as the SE module and VC module, respectively.

DCCRN is a single-channel denoising model based on the convolution recurrent network. Complex CNN and RNN structures are implemented to simulate the complex-valued operation. More details about DCCRN can be found in [33]. In our framework, the SE module needs to separate the estimated speech and noise signals as defined in (1). However, the commonly used scale-invariant signal-to-noise ratio (SI-SNR) loss will result in a power mismatch. Therefore, scale-dependent signal-to-distortion (SD-SDR) loss [35] is adopted, which has comparable performance as the SI-SNR while sensitive to the down-scaling and up-scaling of the estimated speech.

The VC module is implemented using a self-supervised VQ-VAE-based VC method proposed in [34], which is capable of non-parallel conversion. Fig. 2(a) illustrates the structure of VQ-VAE in the baseline. The model comprises an encoder, a vector quantizer, and a decoder. The encoder is a stack of one-dimensional convolutional layers, batch normalization layers, and ReLU activation functions. The vector quantizer manages a learnable codebook and quantizes the output of the encoder using the nearest discrete vectors from the codebook. The decoder is a WaveRNN structured vocoder [36], which predicts the μ-law decoded denoised waveform $\mathbf{d}$ on the basis of the quantized representation $\mathbf{z}$ from the quantizer, speaker code $\mathbf{s}$, and the previous samples in an autoregressive (AR) manner, which can be described as a conditional joint probability distribution:

$$p(\mathbf{d} \mid \mathbf{s}, \mathbf{z}) = \prod_{t=1}^{T} p(d_t \mid d_1, \ldots, d_{t-1}, \mathbf{s}, \mathbf{z}). \tag{2}$$

The SE module is pre-trained and fixed during the training stage, providing the VC module with the denoised speech as input, and the reconstruction loss is also calculated using this denoised speech. During the inference stage, the VC module generates the converted speech based on this denoised speech and the identity od a target speaker. It is an optional step to either add back the separated noise or discard it.

B. Defects of the Baseline

The performance of the baseline framework has been evaluated in [18]. We found that the common metrics for evaluating the SE methods did not accurately reflect their contributions to the VC downstream. Moreover, even a state-of-the-art SE method could introduce unavoidable distortions when suppressing background noise. While inconspicuous to perceptual listening, this additional distortion could negatively affect the quality of the converted speech in terms of naturalness and similarity. Other factors, such as residual noise in the denoised speech, also affect the VC performance. Regrettably, the VC module is trained to reconstruct the distorted speech data, which further degrades the VC performance. Another limitation is that

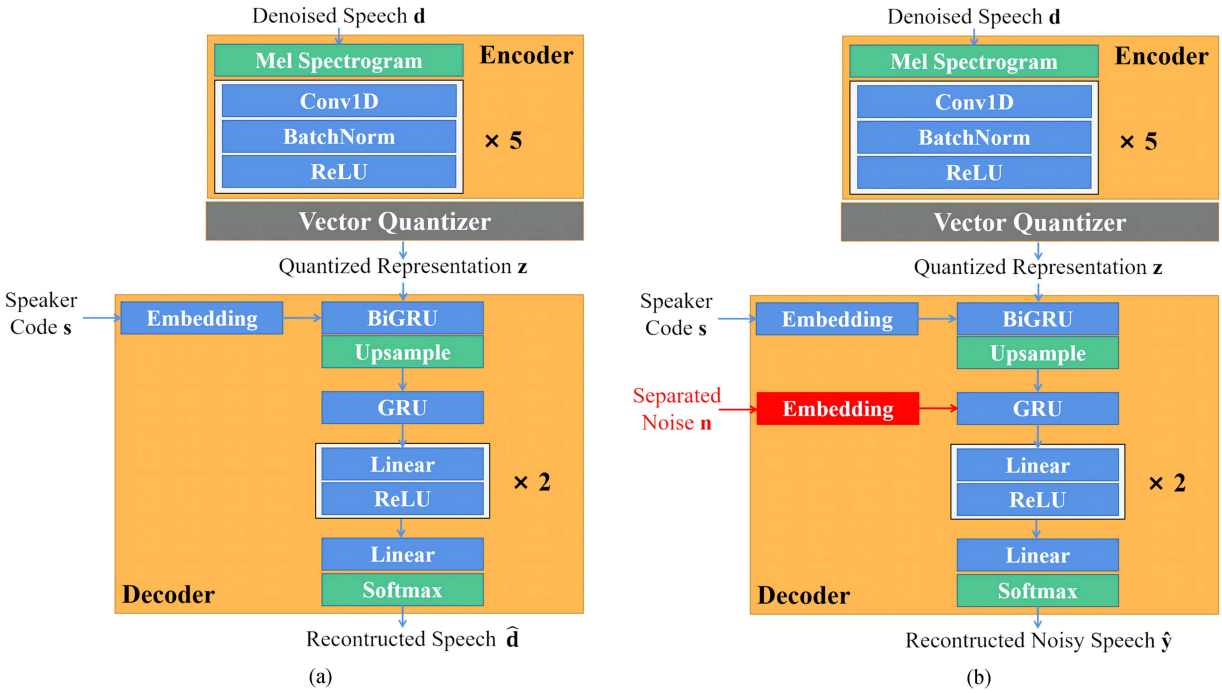


Fig. 2. Detailed structures of the VQ-VAE-based VC module. (a) Original VQ-VAE in Baseline. (b) Noise-conditioned VQ-VAE in N2N VC framework.

converting the noisy target speech is redundant: the converted speech is generated first, then the separated background noise is superimposed.

C. Noise-Conditioned VC Framework

Among the three signals from (1), only the noisy speech is not affected by distortion from the SE module. As a result, we consider leveraging this noisy speech in our training process. Specifically, the VC module still receives the denoised data as inputs but is designed to reconstruct the noisy speech data. This allows for the loss calculation to be conducted using the original noisy data, bypassing the distortion introduced by the SE module. Moreover, modeling the noisy speech enables the VC module to generate the noisy converted speech directly, thereby eliminating the unnecessary two-stage generation found in the baseline.

Modeling noisy speech data is indeed challenging, as affirmed by the experiments conducted in [18]. To simplify this, we utilize the separated noise as a condition for the VC module, as illustrated in Fig. 1(b). During the training stage, the VC module is given denoised speech as input and separated noise as a condition for the decoder to reconstruct the noisy speech. In the conversion stage, sending a noise signal to the condition results in noisy converted speech generation, whereas replacing the noise signal with zero sequences leads to a clean converted speech. Fig. 2(b) shows the modification to the VQ-VAE: An embedding layer transforming the μ-law decoded noise signal along the time axis to a sequence of high-dimensional vectors is added and connected to the Gated Recurrent Unit (GRU) layer. On the basis of the conditional joint probability distribution of the baseline described in (2), the noise-conditioned version is modified as

$$p\left(\mathbf{y} \mid \mathbf{n}, \mathbf{s}, \mathbf{z}\right) = \prod_{t=1}^{T} p\left(y_t \mid y_1, \ldots, y_{t-1}, n_1, \ldots, n_t, \mathbf{s}, \mathbf{z}\right)$$
$$\text{s.t.} \quad \mathbf{y} = \mathbf{d} + \mathbf{n}. \tag{3}$$

By introducing the separated noise signal as a condition, we enable the decoder to learn the distribution of the background noise more easily. This significantly simplifies the generation of noisy speech.

D. Pre-Training Strategies for VC Module

In the pre-training phase for the VC module of the baseline, a clean dataset is used. The VC module takes the clean speech as the input and reconstructs the same. The reconstruction loss is calculated with this clean input speech. As for the pre-training of the VC module in the noise-conditioned framework, since the VC module takes noise as a condition to the decoder, it receives clean speech as the input and the original noise signal as a condition to reconstruct the noisy speech. The reconstruction loss is calculated with the noisy speech, which equals the clean speech superimposed with the original noise signal. Note that pre-training strategies are used to improve the VC modules' performance but not prerequisites. As reported in [26], our framework does not rely on clean training data for the VC.

E. Data Augmentation

In the previous experiments [26], the noisy training dataset involved mixing utterances from a single speaker with various types of background noise at multiple SNRs, which can be considered SI noisy conditions. Results showed that the

noise-conditioned method improved the baseline significantly, reducing the naturalness gap between the baseline and the upper bound of the framework by up to 65% [26]. However, when training data are extended to SD noisy conditions where the noise category and SNR are correlated with the speaker, the VC performance noticeably degrades, even descending below the baseline, as shown in Section (V-A-3).

The degradation of the VC performance arises from the assumption that the speaker code $\mathbf{s}$ and the noise signal $\mathbf{n}$ are independent in (3). When training under SI noisy conditions, the decoder of the VQ-VAE has seen various noisy conditions within each speaker, enabling it to learn the independence of speaker information and noise condition in a self-supervised manner without any additional loss term. However, in the case of SD noisy environments, due to that speaker identity being inherently linked to the noise conditions, the decoder only encounters limited and fixed combinations of each speaker code $\mathbf{s}$ and the noise signal $\mathbf{n}$. This limitation impedes its learning to disentangle speaker information and the noise condition, leading to the leakage of speaker information into the noise condition. Consequently, during inference, changing the speaker code $\mathbf{s}$ to the target code only partially transforms speaker characteristics, as residual characteristic information remains in the separated noise. This leads to the quality degradation of the converted sample in terms of both naturalness and similarity.

The most straightforward approach to addressing the lack of diversity is through data augmentation. Thus, while adhering to the premise that only noisy data from source/target speakers are available for VC training, we implemented three augmentation strategies, namely, Data-Aug, Noise-Aug I, and Noise-Aug II, as illustrated in Fig. 3.

1) Data-Aug: Fig. 3(a) demonstrates the workflow of Data-Aug. In this process, the original noisy training data are duplicated and mixed with augmented noise at various SNRs to expand the diversity of the noisy conditions from SD to SI. The SE module then separates the augmented noisy training data into estimated speeches and the separated background noise, both of which are involved in the self-supervised training of the VC module. The operation of Data-Aug can be formulated as

$$\mathbf{y_{aug}} = \mathbf{d} + \mathbf{n} + \mathbf{n_{aug}} \tag{4a}$$

$$\mathbf{n_{augsep}} = \mathbf{y_{aug}} - \mathbf{d_{aug}}, \tag{4b}$$

where $\mathbf{y_{aug}}$, $\mathbf{n_{aug}}$, $\mathbf{n_{augsep}}$ and $\mathbf{d_{aug}}$ represent the augmented noisy speeches, noise for augmentation, augmented separated noise and denoised augmented noisy speech, respectively.

However, Data-Aug compromises the quality of the training set, as the augmented noisy data also undergoes processing through the SE module. This could result in further distortion in both the denoised speeches and the separated noise.

2) Noise-Aug I: To mitigate the additional distortion introduced by Data-Aug, Noise-Aug I is proposed, where both the noisy training data and the separated noise are duplicated and augmented with additional noise clips, as depicted in Fig. 3(b). Unlike in Data-Aug, the SE module no longer processes the augmented noisy data, thus eliminating the additional distortion. These augmented noisy speeches are only used as the ground truth for loss calculation and the AR process when the noise condition receives the augmented separated noise. The denoised speeches are from the original noisy data to avoid further distortion as in the case of Data-Aug. The operation of Noise-Aug I can be formulated as

$$\mathbf{y_{aug}} = \mathbf{d} + \mathbf{n} + \mathbf{n_{aug}} \tag{5a}$$

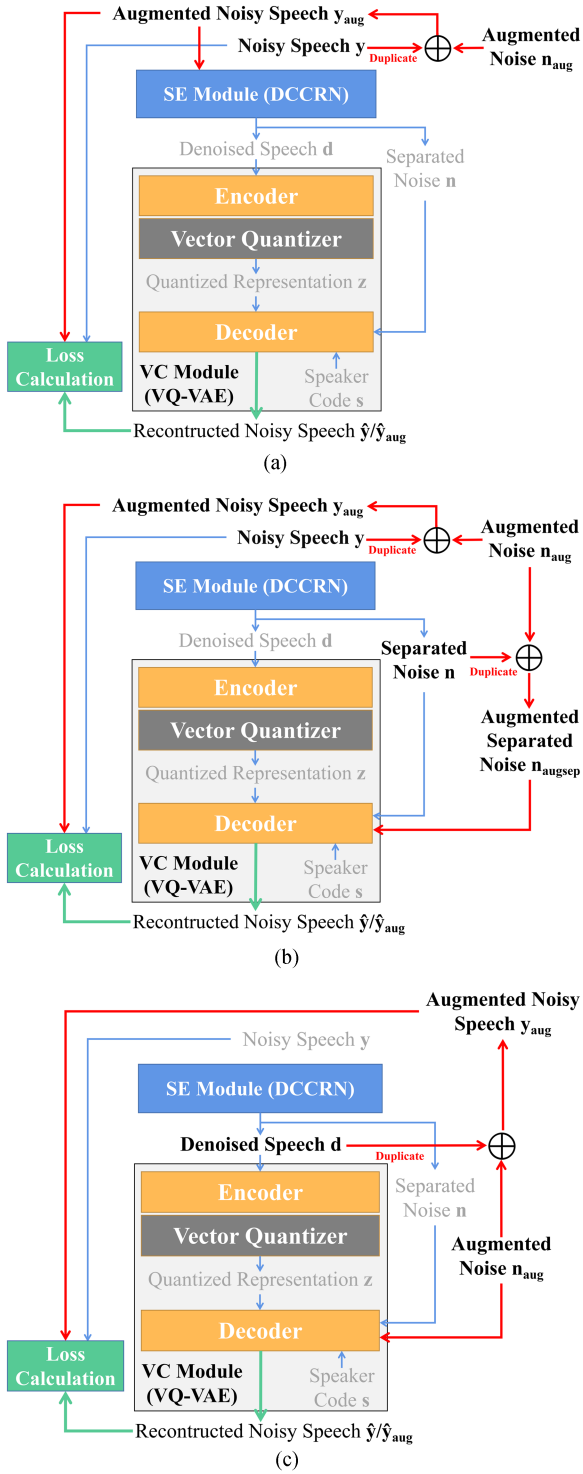


Fig. 3. Illustration of the three data augmentation strategies. (a) Data-Aug. (b) Noise-Aug I. (c) Noise-Aug II.

$$\mathbf{n_{augsep}} = \mathbf{y_{aug}} - \mathbf{d} = \mathbf{n} + \mathbf{n_{aug}}. \tag{5b}$$

$\mathbf{y_{aug}}$ has the augmented noise part $\mathbf{n_{augsep}}$ as $\mathbf{n} + \mathbf{n_{aug}}$, which appears to increase the variety of the noisy conditions. However, we are concerned that since the speaker-dependent noise $\mathbf{n}$ still persists in each utterance within one speaker, the disentanglement cannot be well learned.

3) Noise-Aug II: To address our concerns about Noise-Aug I, we implemented another strategy Noise-Aug II demonstrated in Fig. 3(c). In this approach, the denoised speeches, rather than the original noisy speeches, are duplicated and superimposed with the augmented noise clips to compose the augmented noisy speeches. During training, the noise condition in the decoder receives either speaker-dependent separated noise $\mathbf{n}$ or augmented noise $\mathbf{n_{aug}}$, depending on whether the current ground truth is the original noisy speech or augmented noisy speech. Consequently, the decoder encounters a greater variety of noisy conditions and a broader combination of noise and speaker identity. The noisy conditions are extended from SD to SI, and speaker-dependent separated noise $\mathbf{n}$ is absent from the augmented noisy data. The operation of Noise-Aug II can be formulated as

$$\mathbf{y_{aug}} = \mathbf{d} + \mathbf{n_{aug}} \tag{6a}$$

$$\mathbf{n_{augsep}} = \mathbf{n_{aug}}. \tag{6b}$$

However, the augmented noisy speech is based on the denoised speech $\mathbf{d}$, which is already distorted. This implies that Noise-Aug II compromises the advantage of using undistorted noisy data. Theoretically, as the volume of augmented training data increases, the performance of the noise-conditioned method will degrade to the baseline level. In Section V-A-3 we evaluate the ideal number of the augmented data samples to use.

IV. Experimental Conditions

A. Dataset for SE Module

The SE module was trained on the DNS Challenge 2020 dataset [31], which provides 500 hours of speeches from 2,150 speakers in multiple languages and 70,000 noise clips across 150 classes. A validation set was prepared by sampling 10,000 clean speeches and 8,000 noise clips. The noisy dataset was constructed by mixing the clean speech and noise clips at uniformly sampled SNRs ranging from 0 to 20 dB.

B. Dataset for VC Module

The VCC 2018 dataset [37] was selected as the clean corpus for training and testing. VCC 2018 dataset includes 12 speakers with a balanced gender distribution; eight of these speakers are designated as sources and the remaining four as targets.

ESC-50 [38] and DEMAND [39] were used for simulating SI and SD noisy conditions, respectively. ESC-50 is a dataset for environmental sound classification comprising 2,000 records across 50 classes. Its high intelligibility is valuable given that our work also involves the generation and quality evaluation of background noise. DEMAND is an environmental noise dataset containing six main categories and 18 subcategories. Each subcategory has a five-minute, 16-channel recording, from which we selected ch01 for each subcategory.

For SI noisy conditions, voice activity detection (VAD) was conducted by WebRTC VAD[1] to trim the noise clips of ESC-50. Nine categories were randomly chosen to mix with the VCC testing set, and the rest were for the training set. When building the noisy training set, the clean speech was overlaid with the randomly chosen noise clips at uniformly sampled SNRs from 0, 5, 10, 15, and 20 dB. The noisy VCC testing set was a parallel dataset where the same utterance across speakers was mixed with the same noise clip. Several noisy testing sets were constructed in parallel at each of the following SNRs: −5, 0, 5, 10, 15, 20, and 25 dB. All the VC models were evaluated on these noisy testing datasets with SI noisy conditions, regardless of whether the training dataset incorporated the SI noisy conditions.

To simulate SD noisy conditions, the corpus from each speaker in the VCC 2018 dataset was overlaid with a randomly chosen noise clip from a unique subcategory within the DEMAND dataset at an SNR of 5 dB. To investigate the impact of different SNRs on our proposed method, we constructed additional noisy conditions, namely, semi-speaker-dependent (SSD) noisy conditions. For the SSD noisy conditions, we used the same sampled noise clips as in the SD conditions to create the noisy training dataset. However, unlike the SD conditions where only an SNR of 5 dB was considered, the SSD conditions involved superimposing the corpus with noise at uniformly sampled SNRs from 0, 5, 10, 15, and 20 dB. This procedure was performed such that in SSD noisy conditions, speaker identity correlates with the noise category but not with the SNR, while in SD noisy conditions, speaker identity correlates with both the noise category and SNR. Note that both the SD and SSD noisy conditions were used exclusively for the training dataset.

Another dataset VCTK [40] was used as the pre-training corpus for the VC model. All 110 speakers from VCTK participated in the training. For validation, we selected 20 utterances from each speaker. A noisy version of VCTK was built to pre-train the noise-conditioned VC model. Noise clips from the validation set used during SE module training were mixed with the VCTK corpus at uniformly sampled SNRs from 0, 5, 10, 15, and 20 dB. Thus, the noisy VCTK processed the SI noisy conditions.

As for data augmentation, to control variable factors, we reused the sampled noise clips for the noisy VCC 2018 with SI noisy conditions to conduct the proposed augmentation strategies.

C. Methods to Be Evaluated

Given that the methods were similar, we annotated them as listed in Table I. Generally, two types of VC modules were involved: the conventional VQ-VAE and the noise-conditioned one, which were abbreviated as VQ and NC-VQ, respectively. Three noisy VCC training datasets with speaker-independent, semi-speaker-dependent, and speaker-dependent noisy conditions were denoted as SI, SSD, and SD, respectively. Lastly,

[1] [Online]. Available: https://github.com/wiseman/py-webrtcvad

TABLE I
SYSTEMS TO BE EVALUATED

Name of Method	VC Model	Type of Input	VC Training Set	Description
Upper Bound	VQ	Clean speech	Clean VCC2018	The true upper bound of the N2N VC framework.
	NC-VQ	Clean speech; Original noise	Noisy VCC2018 (SI noisy conditions)	The comparable upper bound of the N2N VC framework.
Baseline-SI	VQ	Denoised speech	Noisy VCC2018 (SI noisy conditions)	-
Baseline-SD	VQ	Denoised speech	Noisy VCC2018 (SD noisy conditions)	-
N2N-SI	NC-VQ	Denoised speech; Separated noise	Noisy VCC2018 (SI noisy conditions)	-
N2N-SSD	NC-VQ	Denoised speech; Separated noise	Noisy VCC2018 (SSD noisy conditions)	-
N2N-SD	NC-VQ	Denoised speech; Separated noise	Noisy VCC2018 (SD noisy conditions)	-
N2N-Data-Aug	NC-VQ	Denoised speech; Separated noise	Noisy VCC2018 (SD noisy conditions) with data augmentation	The augmentation strategy is shown in Fig. 3 (a)
N2N-Noise-Aug I	NC-VQ	Denoised speech; Separated noise	Noisy VCC2018 (SD noisy conditions) with noise augmentation	The augmentation strategy is shown in Fig. 3 (b)
N2N-Noise-Aug II	NC-VQ	Denoised speech; Separated noise	Noisy VCC2018 (SD noisy conditions) with noise augmentation	The augmentation strategy is shown in Fig. 3 (c)

Data-Aug, Noise-Aug I, and Noise-Aug II represent the data augmentation strategies discussed in Section III-E.

The methods to be evaluated fall into three primary categories: **Upper Bound**, **Baseline**, and **N2N**. **Upper Bound** represents the theoretical best performance of the framework, where the VC module is trained using the original VCC dataset rather than the denoised one. Note that there exist two types of Upper Bound using VQ and NC-VQ, respectively. The **Upper Bound** using VQ receives and models the clean speech, whereas the **Upper Bound** using NC-VQ takes clean speech as input and the original noise clip as the condition to model the noisy speech with SI noisy condition. We will prove in Section V-A-1 that these two **Upper Bound**s yield comparable results for clean converted speech generation. **Baseline** utilizes VQ as the VC module and **N2N** utilizes NC-VQ as the VC module. Both are trained on the denoised VCC dataset. More specifically, **Baseline** receives and models the denoised speech, while **N2N** receives denoised speech as input and uses separated noise as the condition to model the noisy speech. The abbreviation following the names of **Baseline** and **N2N** indicates their corresponding training set.

D. Evaluation Metrics

1) Objective Evaluation: Since the metrics for the SE tasks cannot well reveal their impacts on the downstream, we focus on evaluating the performance of VC. Mel cepstral distortion (MCD) [41] was employed to measure speech quality. Additionally, the word error rate (WER) was used to measure the quality of linguistic content and calculated using a publicly available ASR model.[2] As for similarity, an open-source speaker verification method[3] was utilized to compute the score by comparing the converted sample with its target reference. Since these objective metrics are exclusively applicable to clean speech data, all the systems generated clean-converted speeches, and we left the quality assessment of the generated noise and noisy speech to the subsequent subjective evaluation.

2) Subjective Evaluation: Although all the speakers from VCC2018 participated in the training, four source speakers (VCC2SF3, VCC2SF34, VCC2SM3, and VCC2SM4) and two target speakers (VCC2TF2 and VCC2TM2) were selected for the testing set to reduce the total number of converted samples for subjective evaluation. All the subjective evaluations were carried out on Amazon MTurk with 15 participants. The converted samples can be found at the demo page.[4]

The preference evaluation between methods with and without pre-training was conducted to demonstrate the effectiveness of the pre-training strategies. Two testing sets with SNRs of 5 and 15 dB were prepared, based on which we parallelly sampled 32 converted samples from each set. Listeners were first asked to compare the naturalness of two converted samples - one from a method with pre-training and one without pre-training. Subsequently, they compared which sample sounded more similar to the provided reference from the target speaker. To maintain consistency with the objective evaluation, all the methods generated clean converted speeches. As our primary focus is on the noise-conditioned framework and SSD serves as transitional noisy conditions compared to SD, we chose **Upper Bound**, **N2N-SI**, and **N2N-SD** in the preference evaluation, resulting in 192 pairs for each listener.

To evaluate the data augmentation strategies, We conducted the Mean Opinion Score (MOS) and Similarity (SIM) tests [37] to measure the naturalness and similarity of the converted samples, respectively. Two testing sets with SNRs of 5 and 15 dB were used.

The MOS test was divided into two groups: clean MOS and noisy MOS. To better depict and compare the performance of each method, six systems were evaluated in the two MOS tests: **Upper Bound**, **Baseline-SI**, **Baseline-SD**, **N2N-SI**, **N2N-SD**, and **N2N-Noise-Aug II**. **Upper Bound** represents the theoretically best performance our framework can attain. **N2N-SI** paired with **Baseline-SI**, and **N2N-SD** with **Baseline-SD** constitute two evaluation pairs trained with different noisy datasets. SSD noisy conditions were designed to illustrate the impact of SNRs on the disentanglement of noise conditions and speaker identity. They serve as transitional conditions in comparison to SD. Consequently, we did not evaluate methods using training sets with SSD noisy conditions during subjective evaluation. According to the results of the objective evaluation in Section V-A-3, **N2N-Noise-Aug II** outperformed other augmentation strategies, and thus it was selected for subjective evaluation.

The clean MOS test was carried out in a conventional manner. The systems generated clean converted speeches, the naturalness of which was scored by the participants from 1 to 5. For each method, 64 samples were randomly selected from the converted samples: 32 with an SNR of 5 dB and 32 with an SNR of

[2][Online]. Available: https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self.

[3][Online]. Available: https://github.com/resemble-ai/Resemblyzer.

[4][Online]. Available: https://chaoxiefs.github.io/Samples-of-Noisy-to-Noisy-Voice-Conversion/

15 dB, except for **Upper Bound**, from which only 32 samples with an SNR of 5 dB were chosen, because **Upper Bound** has been proven in Section V-A-1 as noise-invariant. In general, a listener would be presented with 364 samples in total, of which 12 samples were from the target speech as the ground truth.

In the noisy MOS test, listeners were asked to evaluate the noisy converted speech and rate the naturalness of both the speech and the noise part on a scale from 1 to 5. Since evaluating the naturalness of background noise is challenging, the original noise clip was provided as the reference. Similar to the clean MOS test, 64 samples were randomly selected from each system: half with an SNR of 5 dB and the remaining half with an SNR of 15 dB. For each listener, 408 samples were prepared, which included 24 samples of noisy target speech with SNRs of 5 and 15 dB to serve as the noisy ground truth.

In the SIM test, a listener was presented with a converted speech and a target speech, then asked to judge whether these two samples were from the same speaker. The SIM score was represented as the acceptance rate: The answers in the SIM test were rated from "Definitely the same", "Maybe the same", "Maybe different", and "Definitely different", whereas the samples deemed "Definitely the same" and "Maybe the same" were considered as "accepted." From the results of the MOS tests, we evaluated five systems in the SIM test: **Upper Bound**, **Baseline-SD**, **N2N-SI**, **N2N-SD**, and **N2N-Noise-Aug II**. The SIM test had the same sampling pattern as the clean MOS test; thus, 288 samples were prepared for each listener.

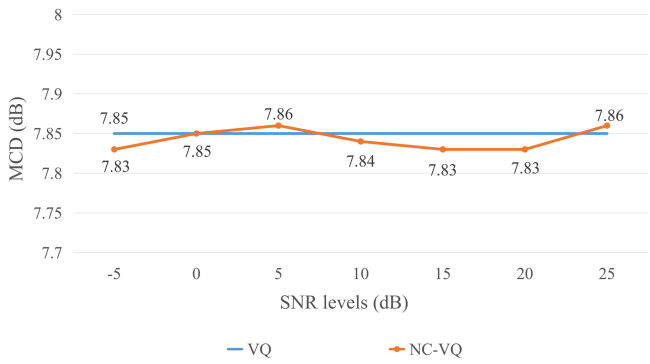


Fig. 4. Results of MCD as a function of SNRs for NC-VQ and VQ.

TABLE II
RESULTS OF OBJECTIVE EVALUATIONS AMONG METHODS W/ AND W/O PRE-TRAINING STRATEGIES UNDER SNRS OF 5 AND 15 DB

Methods	Status	MCD (dB) ↓		WER (%) ↓		Similarity ↑	
		5 dB	15 dB	5 dB	15 dB	5 dB	15 dB
Noisy Testing Set	Denoised	-	-	6.41	3.81	-	-
Upper Bound	**w/ pre-training**	7.86	**7.83**	**9.55**	**10.56**	**0.824**	0.823
	w/o pre-training	7.84	7.84	14.57	14.93	0.821	0.823
Baseline-SI	**w/ pre-training**	**8.76**	**8.39**	**32.92**	**16.02**	**0.772**	**0.798**
	w/o pre-training	8.89	8.64	56.62	40.59	0.757	0.766
N2N-SI	**w/ pre-training**	**8.58**	**8.33**	**29.22**	**17.01**	**0.786**	**0.798**
	w/o pre-training	8.62	8.38	39.27	27.62	0.777	0.786
N2N-SSD	**w/ pre-training**	9.17	**8.92**	**28.61**	**15.60**	0.750	**0.762**
	w/o pre-training	9.16	8.94	44.49	32.04	0.752	0.756
N2N-SD	**w/ pre-training**	**9.27**	**9.08**	**31.63**	**19.72**	**0.737**	**0.744**
	w/o pre-training	9.46	9.22	45.43	31.52	0.719	0.725

V. EVALUATION RESULTS

A. Results of Objective Evaluation

1) Noise-Conditioned VQ-VAE vs VQ-VAE: Theoretically, if the noisy training set processes SI noisy conditions, NC-VQ should function in the same way as the original VQ when generating clean converted speech because the noise condition, speaker information, and content representations are independent of each other, as indicated by (3). We carried out experiments to substantiate this hypothesis. VQ was initially pre-trained on VCTK, and then fine-tuned on VCC2018. For a fair comparison, NC-VQ was also pre-trained on noisy VCTK described in Section IV-B and later fine-tuned on noisy VCC2018 with SI noisy conditions.

Fig. 4 presents a comparison of MCD scores between NC-VQ and VQ on the clean VCC2018 testing set. As VQ is independent of noise, its MCD remains stable at 7.85 across all SNRs. Since NC-VQ involves using noise as the condition during training, it is evaluated using the clean corpora for building their noisy versions at SNRs from −5 to 25 dB. NC-VQ achieves similar MCD scores of around 7.85, with an average score of 7.84. This confirms that introducing noise as conditions into VQ does not compromise the quality of the clean converted samples when the training set is under SI noisy conditions. Consequently, we employ NC-VQ as the VC module of **Upper Bound**.

2) Pre-Training Strategies: Table II presents the results of evaluating the methods with/without pre-training strategy, as measured by MCD, WER, and similarity under SNRs of 5 and 15 dB. The WER of the denoised testing set is also provided as the reference. In general, pre-training improves the performance of all methods. The effectiveness is significant in terms of WER, whereas results for MCD and similarity also achieve minor improvements, except for **N2N-SD** with the most strict training conditions, which also achieves noticeable improvements in MCD. It can be concluded that pre-training enhances the robustness of the methods against distortion, particularly those methods that utilize denoised data as input and are trained on limited training sets (SSD and SD noisy conditions). Therefore, all the methods employed the pre-trained VC model in the subsequent experiments.

3) Data Augmentation Methods: As evidenced in Table II, it is obvious that the performance of the N2N framework degrades when the training set possesses SSD noisy conditions. Under SD noisy conditions where both the type of noise and the SNR are related to the speaker, the method yields the poorest results in terms of MCD, WER, and similarity, the reason for which has been explained in Section III-E.

Fig. 5 shows the results of the objective evaluation of three augmentation strategies on the noisy VCC2018 testing set with SI noisy conditions at an SNR of 5 dB. As discussed in Section II-I-E, **N2N-Noise-Aug II** carries out the augmentation based on the distorted data; therefore, we represent **N2N-Noise-Aug II** at various degrees of augmentation by using the number of augmented data for one speaker (80 utterances per speaker) and their respective percentages relative to the original data, as depicted in Fig. 5.

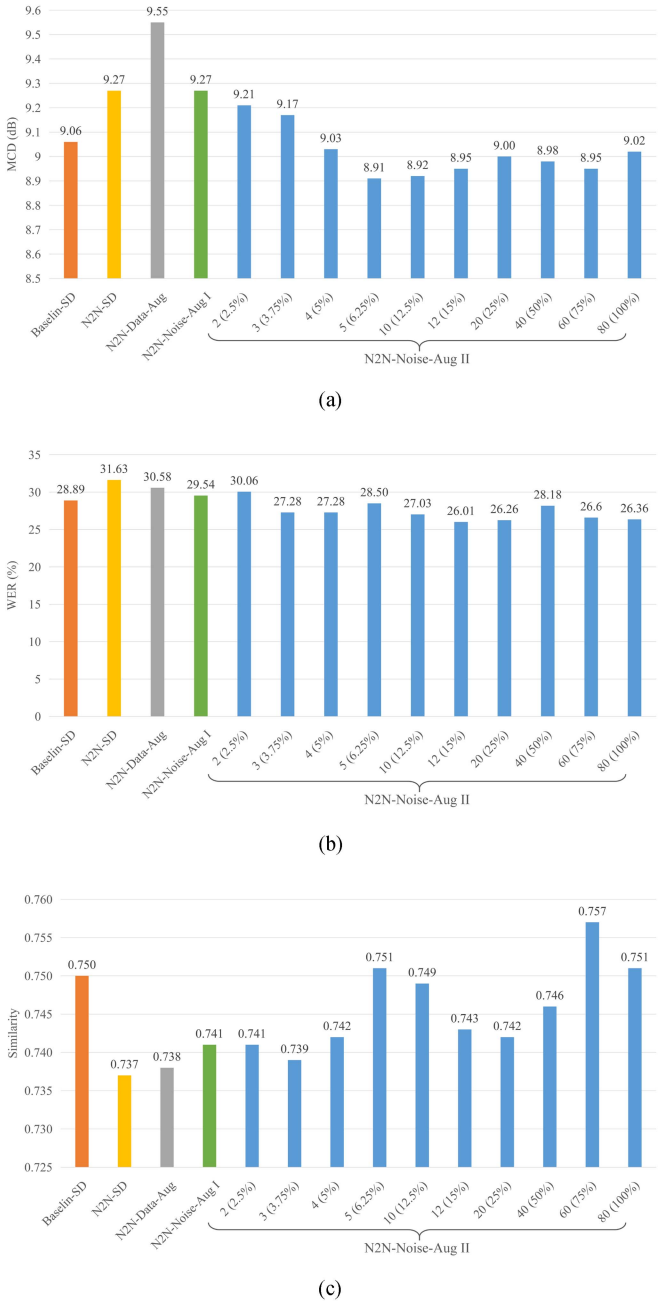


Fig. 5. Result of the objective evaluation of methods with data augmentation strategy. (a) MCD. (b) WER. (c) Similarity.

Baseline-SD reaches an MCD of 9.06, a WER of 28.89%, and a similarity score of 0.750, outperforming both **N2N-Data-Aug** and **N2N-Noise-Aug I**. This supports the concerns raised in Section III-E-2. **N2N-Data-Aug** reaches an MCD of 9.55, a WER of 30.58%, and a similarity score of 0.738, which are worse than those in **N2N-Noise-Aug I** because **N2N-Data-Aug** introduces additional distortions as explained in Section III-E-1. **N2N-Noise-Aug I** achieves an MCD of 9.27, a WER of 29.54%, and a similarity score of 0.741. These are comparable to the scores of **N2N-SD**, which are an MCD of 9.27, a WER of 31.63%, and a similarity score of 0.737. This suggests that Noise-Aug I is not effective in addressing the disentanglement issues.

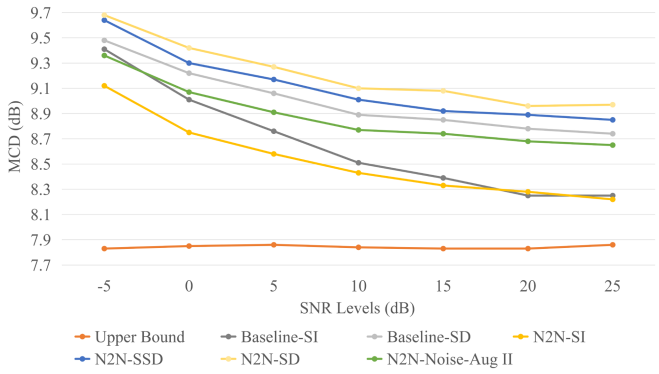
For the **N2N-Noise-Aug II** group, most methods outperform **Baseline-SD** in both MCD and WER, while attaining similar scores in similarity. In terms of MCD, **5 (6.25%)** ranks first with an achievement of 8.91; **10 (12.5%)** is in second place with a score of 8.92; **12 (15%)** and **60 (75%)** share third place with a score of 8.95. For similarity, **60 (75%)** takes the lead with a score of 0.757. **5 (6.25%)** and **80 (100%)** are tied for second place with a score of 0.751, while **10 (12.5%)** takes the third place with a score of 0.749.In the case of WER, **5 (6.25%)** achieves 28.50% only, whereas **10 (12.5%)** and **60 (75%)** gain 27.05% and 26.6%, respectively. As discussed in Section III-E, since **Noise-Aug II** benefits most from a smaller volume of augmented data, **5 (6.25%)** is chosen as **Noise-Aug II**.

We summarize the objective evaluation results as a function of SNRs in Fig. 6. **Upper Bound** takes the lead in all metrics across all SNRs. A clear gap exists between **Baseline-SI** and **Upper Bound**, which widens at lower SNRs. **N2N-SI** significantly improves upon **Baseline-SI** when the SNRs are lower than 20 dB. However, under SD noisy conditions, the N2N framework cannot achieve the same level of performance as it does under SI noisy conditions: **N2N-SD** shows degradation, while **Baseline-SD** delivers better results in all metrics. Moreover, **N2N-SSD** consistently outperforms **N2N-SD** in all metrics, showing that the variations in both SNRs and noise categories affect the disentanglement of the speaker information and noise condition in our N2N VC framework. **N2N-Noise-Aug II** improves the performance of **N2N-SD** and achieves better results in all metrics, and outperforms **Baseline-SD** in terms of MCD and similarity across all SNRs. However, **Baseline-SD** still holds the advantage in WER at 15 dB by 1.07%. Despite these improvements, the gap reduction is not as significant compared to that under SI noisy conditions, indicating potential room for further enhancement.

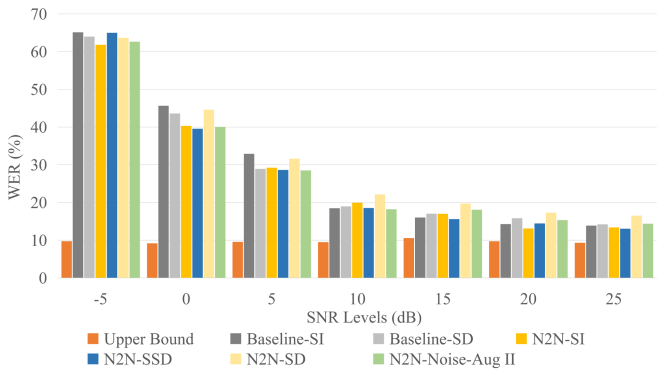
B. Experimental Results for Subjective Evaluation

Fig. 7 shows the results of the preference evaluation comparing the methods with and without pre-training. Across all three N2N frameworks at SNR levels of 5 and 15 dB, the methods using pre-training take minor advantages over their non-pre-training counterparts in terms of the naturalness and similarity of the converted samples from 4.38% to 15.42%. Different from the results of objective evaluation where pre-training only demonstrated effectiveness with methods using denoised speech with distortion as input during training, **Upper Bound**, which uses original speech and noise clips, also achieves better results with pre-training. These findings echo the conclusion observed in the objective evaluation: Pre-training can enhance the performance of the N2N framework in terms of the naturalness and similarity of the clean converted samples.

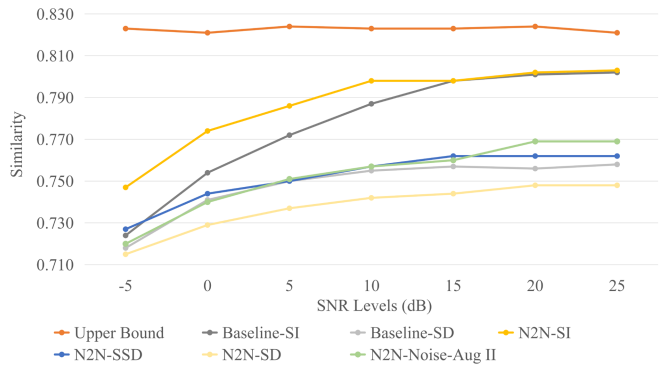
Fig. 8 presents the results of MOSs for clean converted samples. As **Upper bound** and **Ground Truth** are noise-independent, their MOSs for SNRs of 5 and 15 dB are identical. In contrast, the other methods, being noise-related, achieve higher MOSs at an SNR of 15 dB. Excluding **Ground Truth**,



(a)

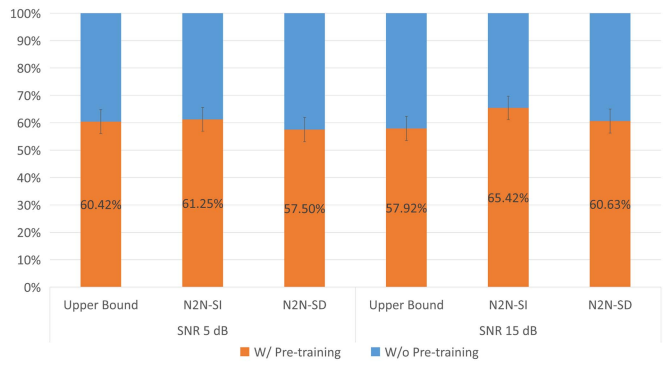


(b)

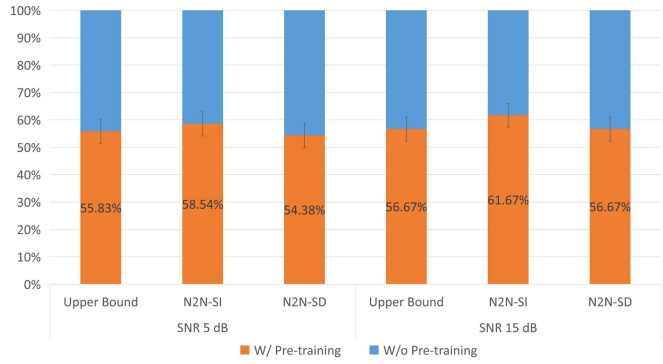


(c)

Fig. 6. Objective evaluation results as functions of SNRs. (a) MCD. (b) WER. (c) Similarity.



(a)



(b)

Fig. 7. Result of the preference evaluation for the clean converted samples by the N2N frameworks with/without pre-training. Error bars show 95% confidence intervals. (a) Preference evaluation on naturalness. (b) Preference evaluation on similarity.

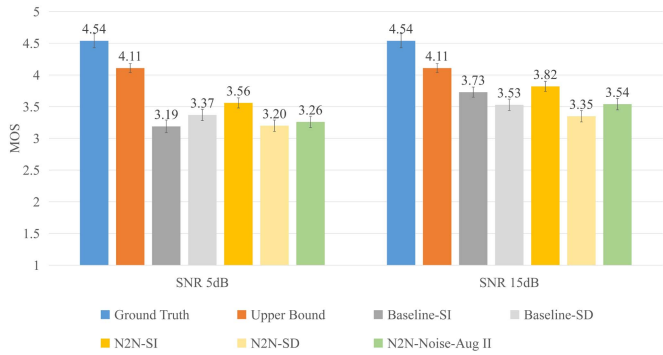


Fig. 8. Result of the subjective evaluation of naturalness for clean converted speech. Error bars show 95% confidence intervals.

Upper Bound secures the highest score of 3.98. **N2N-SI** achieves MOSs of 3.56 and 3.82 at SNRs of 5 and 15 dB, outperforming **Baseline-SI**, which obtains MOSs of 3.19 and 3.73. Consistent with the results of the objective evaluation, the performance of the N2N framework deteriorates under SD noisy conditions. **N2N-SD** attains MOSs of only 3.20 and 3.35 at SNRs of 5 and 15 dB, falling short of **Baseline-SD**, which scores MOSs of 3.37 and 3.53. With the noise augmentation strategy, the degradation is alleviated with **N2N-Noise-Aug II** achieving MOSs of 3.26 and 3.54 at SNRs of 5 and 15 dB, improving the performance of **N2N-SD** by 0.06 and 0.19, respectively. However, it still trails behind **Baseline-SD** at an SNR of 5 dB by 0.11.

Fig. 9(a) shows the MOSs for the speech part of the noisy converted samples. Similar trends as those for MOSs for clean converted speech are observed. **Upper Bound** achieves the highest MOSs of 3.98 and 4.19 at SNRs of 5 and 15 dB. **N2N-SI** ranks the second place with MOSs of 3.8 and 3.98, outperforming **Baseline-SI** by 0.12 at an SNR of 5 dB, while **Baseline-SI** obtains a marginally higher MOS of 4.03 at an SNR of 15 dB. With the lowest MOSs of 3.6 and 3.83 at SNRs of 5 and 15 dB respectively, **N2N-SD** falls short of **Baseline-SD**, which

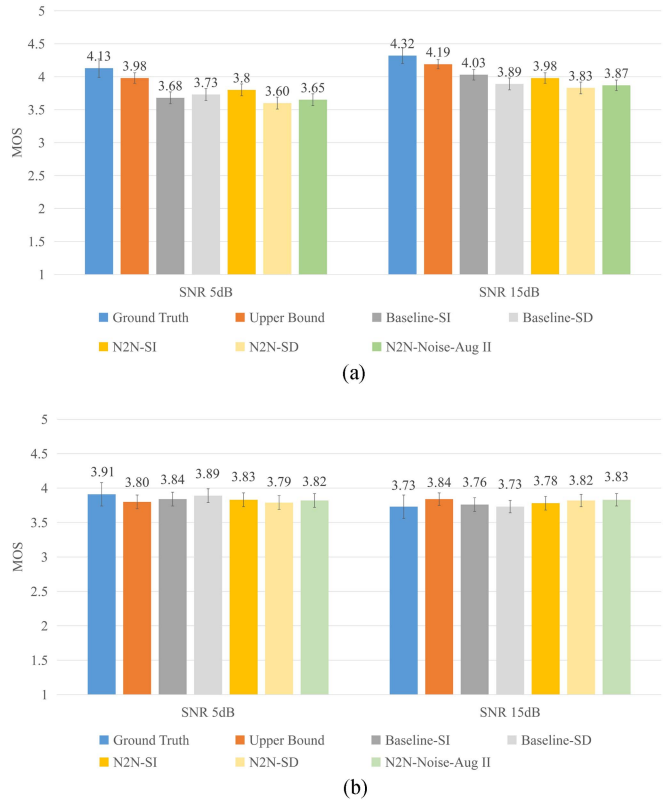


Fig. 9. Result of the subjective evaluation of naturalness for noisy converted speech. Error bars show 95% confidence intervals. (a) MOS for the speech part of the converted samples. (b) MOS for the noise part of the converted samples.

scores 3.73 and 3.98. However, noise augmentation does not achieve the same improvements as observed in the results of the objective evaluation. Compared to **N2N-SD**, **N2N-Noise-Aug II** reaches slightly higher MOSs of 3.65 and 3.87 at SNRs of 5 and 15 dB, which only improves the performance of **N2N-SD** by 0.05 and 0.04, respectively, and remains marginally lower than **Baseline-SD** by 0.08 and 0.02.

Fig. 9(b) shows the MOSs for the noise part of the noisy converted samples. All the methods achieve similar MOSs, approximately 3.8 at SNRs of 5 and 15 dB, which is lower than that of **Ground Truth** receiving 3.91 at the SNR of 5 dB but exceeds its score of 3.73 at an SNR of 15 dB. Note that **Upper Bound**, **N2N-SI**, **N2N-SD**, and **N2N-Noise-Aug II** generate the noise part via the neural networks, whereas the remaining methods superimpose the separated noise.

SIM scores, depicted in Fig. 10, follow similar trends to the MOS scores for clean converted samples. **Upper Bound** attains the highest SIM score of 70.83. **N2N-SI** achieves SIM scores of 63.54 and 67.71 at SNRs of 5 and 15 dB respectively, outperforming **N2N-SD** which only achieves SIM scores of 49.79 and 58.54. These scores are lower than those of **Baseline-SD**, which achieves 55.21 and 63.96. **N2N-Noise-Aug II** outperforms **N2N-SD** by 5.21 at the SNR of 5 dB, while the improvement is less significant at the SNR of 15 dB by only 0.59. Nevertheless, **Baseline-SD** shows better results than **N2N-Noise-Aug II**.

In general, the results reveal the N2N framework's substantial superiority over the baseline in SI noisy conditions. Utilizing the pre-training strategy can further improve the performance of the

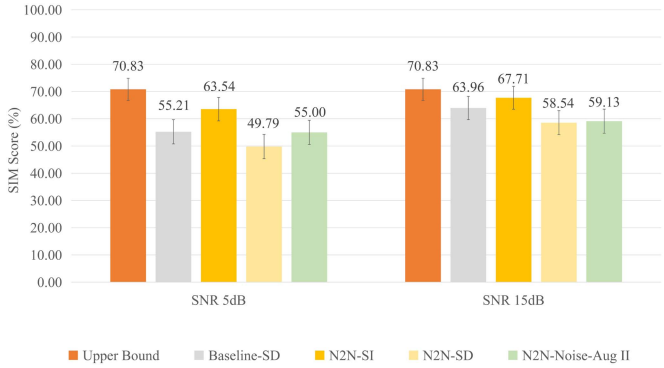


Fig. 10. SIM scores for clean converted speech. Error bars show 95% confidence intervals.

N2N framework. However, under SD noisy conditions, we observe a degradation in performance concerning both naturalness and similarity. While noise augmentation does enhance performance, it does not surpass the baseline under identical SD noisy conditions, indicating the potential for further improvements. As for the noise generation in the noisy converted samples, our methods achieve comparable scores to the ground truth, proving the high quality of the generated noise.

VI. Conclusion

In this article, we evaluated our N2N VC framework under extended noisy conditions from SI to stricter SD noisy conditions where the information of noise and speaker are correlated. The pre-training strategy was introduced to improve the framework's performance, and its benefit was confirmed by our evaluation results. Further experiments demonstrated the effectiveness of our framework under SI noisy conditions; however, performance suffered under SD conditions due to the entanglement of speaker information and noise conditions. To address these issues, we proposed three noise augmentation strategies. Objective evaluation results show that the noise augmentation strategy Noise-Aug II improves the N2N framework and outperforms the baseline, although the improvement is not as significant as that with the N2N framework under SI noisy conditions. However, subjective evaluation results show that there is still room to disentangle the correlation between speaker identity and noisy conditions. In future work, we aim to enhance the disentanglement performance while maintaining the premise that only noisy speech data are available. Additionally, since the commonly used metrics for VC are unsuitable for evaluating noisy speech, we will also explore the development of new metrics designed explicitly for assessing noisy speech quality.

References

[1] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2012, pp. 313–317.

[2] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *J. Acoust. Soc. Jpn.*, vol. 11, no. 2, pp. 71–76, 1990.

[3] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. speech audio Process.*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[4] Z. Yi et al., "Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion," in *Proc. Joint Workshop Blizzard Challenge Voice Convers. Challenge*, 2020, pp. 80–98.

[5] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 266–273.

[6] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. IEEE 26th Eur. Signal Process. Conf.*, 2018, pp. 2100–2104.

[7] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-VC2: Improved cyclegan-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6820–6824.

[8] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC3: Examining and improving CycleGAN-VCs for mel-spectrogram conversion," in *Proc. Interspeech*, 2020, pp. 2017–2021.

[9] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. IEEEAsia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2016, pp. 1–6.

[10] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 9, pp. 1432–1443, Sep. 2019.

[11] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5274–5278.

[12] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," in *Proc. Interspeech* 2019, pp. 674–678.

[13] W.-C. Huang, T. Hayashi, X. Li, S. Watanabe, and T. Toda, "On prosody modeling for ASR+ TTS based voice conversion," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2021, pp. 642–649.

[14] A. T. Liu, P. chun Hsu, and H. yi Lee, "Unsupervised end-to-end learning of discrete linguistic units for voice conversion," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 1108–1112.

[15] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "ATTS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6805–6809.

[16] W. Gan et al., "IQDUBBING: Prosody modeling based on discrete self-supervised speech representation for expressive voice conversion," 2022, *arXiv:2201.00269*.

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

[18] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy voice conversion framework with denoising model," in *Proc. IEEE Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 814-820.

[19] T.-h. Huang, J.-H. Lin, and H.-Y. Lee, "How far are we from robust voice conversion: A survey," in *Proc IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 514–521.

[20] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," in *Proc. 8th ISCA Workshop Speech Synth.*, 2013, pp. 71–75.

[21] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Workshop Speech Synth.*, 2016, pp. 146–152.

[22] Y. Chan, C. Peng, S. Wang, H. Wang, Y. Tsao, and T. Chi, "Speech enhancement-assisted stargan voice conversion in noisy environments," *CoRR*, 2021. [Online]. Available: https://arxiv.org/abs/2110.09923

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

[24] H. Du, L. Xie, and H. Li, "Noise-robust voice conversion with domain adversarial training," *Neural Netw.*, vol. 148, pp. 74–84, 2022.

[25] C.-y. Huang, K.-W. Chang, and H.-y. Lee, "Toward degradation-robust voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6777–6781.

[26] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Direct noisy speech modeling for noisy-to-noisy voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 6787–6791.

[27] Y. Choi, C. Xie, and T. Toda, "An evaluation of three-stage voice conversion framework for noisy and reverberant conditions," in *Proc. Interspeech*, 2022, pp. 4910–4914, doi: 10.21437/Interspeech.2022-10158.

[28] J. chieh Chou and H.-Y. Lee, "One-shot voice conversion by separating speaker and content representations with instance normalization," in *Proc. Interspeech* 2019, pp. 664–668.

[29] C.-y. Huang, Y. Y. Lin, H. -y. Lee, and L.-s. Lee, "Defending your voice: Adversarial attack on voice conversion," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2021, pp. 552–559.

[30] W.-N. Hsu et al., "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 5901–5905.

[31] C. K. Reddy et al., "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech* 2020, pp. 2492–2496.

[32] C. K. Reddy et al., "INTERSPEECH 2021 deep noise suppression challenge," in *Proc. Interspeech*, 2021, pp. 2796–2800.

[33] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020, pp. 2472–2476.

[34] 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, pp. 4836–4840.

[35] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 626–630.

[36] J. Lorenzo-Trueba et al., "Towards achieving robust universal neural vocoding," in *Proc. Interspeech*, 2019, pp. 181–185.

[37] J. Lorenzo-Trueba et al., "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Odyssey Speaker Lang. Recognit. Workshop*, 2018, pp. 195–202.

[38] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1015–1018.

[39] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-Channel Acoustic Noise Database (demand): A Database of Multichannel Environmental Noise Recordings," in *Proc. Meetings Acoustics*, 2013, Art. no. 035081.

[40] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multispeaker corpus for CSTR voice cloning toolkit (version 0.92)," [sound], Univ. Edinburgh, Centre Speech Technol. Res., 2019. [Online]. Available: https://doi.org/10.7488/ds/2645

[41] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process.*, 1993, pp. 125–128.