APSIPA Transactions on Signal and Information Processing, 2025, 14, e10

Original Paper

An Investigation of Noisy-to-noisy Voice Conversion Performance in Various Noisy Conditions

Chao Xie* and Tomoki Toda

Nagoya University, Japan

ABSTRACT

Voice conversion (VC) in a noisy-to-noisy (N2N) scenario aims to convert the speaker identity of noisy speech to a target speaker while preserving both the linguistic content and background noise. In our previous work, we proposed an N2N framework for this conversion. Notably, our VC approach relies solely on noisy speech data for training without requiring clean speech data from either the source or target speakers. Additionally, the framework enables the retention or removal of the noise component in the converted speech during conversion. However, significant performance degradation was observed in the N2N framework when certain noisy conditions were present in the training data. In this paper, we further investigate adverse noisy conditions affecting our framework's performance. We identify two key factors contributing to performance degradation: the lack of noise diversity leading to feature entanglement and noise bias during training. To address these issues, we introduce a mutual information approximation and a noise dropout strategy into the N2N framework. Objective and subjective evaluations validate the effectiveness of our approach

*Corresponding author: xie.chao@g.sp.m.is.nagoya-u.ac.jp.
Acknowledgements: This work was partly supported by JST CREST JPMJCR19A3 and JST AIP Acceleration Research JPMJCR25U5, Japan.

Received 13 February 2025; revised 26 March 2025; accepted 02 May 2025
ISSN 2048-7703; DOI 10.1561/116.20250008

in improving converted speech quality and mitigating VC performance degradation under adverse noisy conditions.

Keywords: Voice conversion (VC), noisy-to-noisy VC, noisy speech modeling, mutual information, noise dropout

1 Introduction

Voice conversion (VC) is a technique for converting non-/para-linguistic information of a source voice to a target one without changing its linguistic content. VC has been studied for several decades, evolving from early statistical methods to deep learning-based approaches. Recent achievements in VC research have led to various applications, such as noise-robust VC [54, 2, 38, 7], movie dubbing [14, 9], and singing voice conversion [31, 51, 21, 67]. However, these new applications necessitate additional requirements because the usage scenarios differ from experimental ones. For instance, both training and test speech data are relatively clean and high-quality in experimental environments, whereas the test data in real-world scenarios are often corrupted with various kinds of noise. Besides, as deep learning-based VC techniques are data-driven, Web-crawled speech data are also an important resource for training, although they often contain undesired background noise extremely degrading the performance of the VC model in terms of speech naturalness and similarity.

Although background noise is often treated as interference in many studies, it can also be valuable in certain tasks to be preserved. For example, in movie dubbing and singing voice conversion, background sound and accompaniment are typically removed beforehand to ensure the quality of the vocal conversion, but they should be retained during inference. Furthermore, recent VC techniques have been employed for data augmentation in downstream tasks, such as low-resource text-to-speech (TTS) [23, 57, 46], automatic speech recognition (ASR) [52, 48, 62], and speaker verification [50, 49, 43, 13, 17]. The original speech datasets used in these tasks often contain inherent background noise, which can enhance the models robustness and should therefore be retained as a valuable training resource after conversion.

In our previous work [64], we proposed a noisy-to-noisy (N2N) VC framework capable of converting the speaker identity of noisy speech while also allowing control over the noise component. Notably, the VC model within the N2N framework does not require clean speech data for training. However, in our subsequent work [63], we observed significant degradation in VC performance under specific noisy training conditions. We hypothesized that the degradation was caused by the entanglement between speaker identity and noise conditions. Although a data augmentation method was proposed to mitigate this issue, subjective evaluation results suggest that it is not sufficiently effective. Moreover, how noise influences the modeling of noisy speech remains unclear.

In this paper, we categorize noise conditions and conduct targeted experiments to better analyze the causes of VC performance degradation and the impact of noise on noisy speech modeling. Our findings reveal that the previous hypothesis about speaker-noise entanglement is inaccurate: Speech-noise entanglement is the primary cause of VC performance degradation. To mitigate this entanglement, we use a mutual information (MI) approximation [5] jointly trained with the VC model without requiring additional data. Furthermore, we identify noise bias in training as another major factor. To address this, we apply noise dropout in training to reduce the models excessive focus on noise reconstruction. Integrating the MI estimator and noise dropout into the N2N framework mitigates VC performance degradation, as confirmed by objective and subjective evaluations. The main contributions of this paper are summarized as follows:

- We investigate the causes of VC performance degradation identified in our previous work [63] and analyze the impact of noise on the modeling of noisy speech. To this end, we segment the previously used noisy training sets by noise categories and conduct separate experiments to analyze the individual effects of each noise type on the VC model.
- Our experimental results indicate that the entanglement between speech content and noise, along with noise bias during training, are the primary contributing factors to the degradation of VC performance.
- To improve the speech naturalness and similarity of the N2N VC framework, we introduce two methods: an MI approximation network to mitigate the entanglement issue and a noise dropout strategy to counter noise bias.
- We conduct both subjective and objective experiments to evaluate the effectiveness of our methods in improving VC performance. Additionally, an ablation study is performed to highlight the individual contributions of the MI estimator and noise dropout to the observed performance gains.

2 Related Work

VC techniques have been extensively studied for decades, even before the advent of deep learning. Early approaches primarily relied on the statistical modeling of speech signals. Many proposed methods, such as exemplar-based sparse representation [55], vector quantization (VQ) [1], and Gaussian mixture modeling [53] have established the foundation for modern approaches. With the emergence of deep learning, neural network-based methods have continuously advanced the naturalness and similarity of synthesized speech [69]. Numerous approaches have been the focus of recent VC studies, including generative adversarial networks (GANs) [25, 26, 27, 28, 11], variational autoencoders (VAEs) [15, 24, 47, 59, 66, 36], automatic speech recognition (ASR) combined with text-to-speech (TTS) [30, 56, 22, 37, 40], and diffusion probabilistic models [32, 42, 70, 6].

The rapid advancement of VC technology in recent years has also driven efforts to apply it in real-world scenarios. These practical applications introduce new challenges, such as addressing environmental interferences like noise. However, compared to conventional VC which has been extensively studied, research on noise-robust VC systems remains limited. Moreover, most existing studies treat noise as an interference to be removed, with only a few focusing on noise-robust VC approaches that preserve the background noise.

2.1 Noise-robust VC

Before the advent of deep learning, researchers had already explored noise-robust VC. Takashima *et al.* [54] proposed a sparse-representation-based VC method using non-negative matrix factorization to optimize source and target basis matrices with a shared activity matrix. By representing both source and noise components with separate dictionaries, the method isolates source speech features effectively while minimizing noise interference.

With the advancements in deep learning, its application to complex tasks has become increasingly prevalent, surpassing the performance of traditional methods in most tasks, particularly in speech enhancement (SE). Consequently, employing deep learning-based SE models for noise-robust VC is a straightforward and practical way. Valentini-Botinhao *et al.* [60] proposed a pioneer research on neural network-based TTS for noisy environments. A recursive neural network (RNN)-based SE model is used as a preprocessing stage to the TTS system to effectively remove noise before it is passed to the TTS model. Similarly, Chan *et al.* [2] proposed a noise-robust VC framework that incorporated a lightweight SE component ahead of the VC model to mitigate the effects of noise. Miao *et al.* [38] proposed a noise-robust VC method that improved voice clarity by leveraging high-quefrency boosting through sub-band cepstrum conversion and fusion. By separating the speech signal into sub-bands and applying cepstrum-based conversion, this method selectively boosts high-quefrency components, which are less affected by noise. Choi *et al.* [7] proposed a cascading VC framework that employs two sequential SE modules to address background noise and reverberation separately, enabling independent control over both factors in the converted speech.

Although using an SE model for noise preprocessing is straightforward, it can introduce additional distortions to the denoised speech features. As a result, several noise-robust approaches have been proposed that avoid relying on SE models. Du *et al.* [12] adopted domain adversarial training (DAT) to achieve noise-robust VC. Their approach builds on the zero-shot VC framework AdaIN-VC [8], which is trained in a denoising manner: it takes both clean and noisy data as input but predicts only the clean reconstructed output during training. DAT is applied to the encoders to extract noise-invariant speaker and content representations. Chen *et al.* [3] proposed a noise-robust VC by conducting adversarial training to suppress noise components. Two noise decoupling discriminators are employed to extract noise-invariant content and speaker identity representations. Huang *et al.* [19] combined denoising and adversarial training to develop a generalized degradation-robust VC model. The training dataset is augmented using the adversarial examples generated by embedding attacks [20], along with degradations randomly selected from background noise, reverberation, and band rejection. During training, the VC model processes clean speech, speech with augmented distortions, and adversarial examples, while the loss is computed based on the corresponding clean speech. Xue *et al.* [33] proposed a noise-robust VC method based on noise-controllable Glow-WaveGAN [10]. The training data is augmented by superimposing noise onto clean speech to create paired clean and noisy samples. A robust feature extractor is then trained to obtain noise-independent acoustic representations of speech, along with a vocoder incorporating additional embeddings to control the clean or noisy attributes of the generated speech.

2.2 Noise-Robust VC with Background Noise Preservation

With the growing application of VC techniques to various tasks, such as speech data augmentation and singing VC, there is a growing need to preserve informative background sounds as a resource. As a pioneering effort, Hsu *et al.* [16] proposed a background noise-controllable VC method based on a TTS model with data augmentation and adversarial factorization. The training data is augmented by adding noise to clean speech while retaining the original transcripts and speaker labels. A VAE model is jointly trained with the TTS model to disentangle speaker identity and noise conditions from noisy speech, with domain adversarial training further enhancing this factorization. During inference, two latent factors representing speaker characteristics and background noise are extracted and fed into the TTS model, enabling control over the noise characteristics in the converted speech. However, the quality of the generated noise remains poor, often resembling white noise.

In our primary work [65], we proposed an N2N baseline framework that follows a cascading design involving SE and VC models. The SE model is pre-trained to extract the denoised speech as well as the background noise by subtracting the denoised speech from a noisy speech in the time domain. The VC model is then trained on the denoised speech. The separated noise can be superimposed on the converted speech during inference. However, similar to other approaches employing SE models, the use of an SE model introduces additional distortion to the denoised speech, which subsequently degrades the performance of the VC model trained to reconstruct this distorted data. Furthermore, in our N2N task, a key constraint is the unavailability of clean speech data for VC training, which restricts the methods that can be employed.

To address this limitation, in our subsequent work [64], we improved the N2N framework by incorporating separated noise as an input to the VC model during training, enabling the VC model to directly reconstruct noisy speech that possesses full information on speech content, speaker identity, and background noise. Experimental results show that our improved method reduces the performance gap between the original VC approach and its upper bound by up to 60%.

Chen *et al.* [4] proposed a noise-robust VC framework that supports noise preservation. The architecture is similar to the baseline of our proposed method [65], consisting of a SE model for separation and a VC model. The VC model is trained on denoised speech separated by the SE model, while the noise component is excluded from the training process and is only used during inference by being superimposed to the converted speech.

In another work [68], Yao *et al.* proposed a noise-robust VC method that cascades a SE model and a VC model. The SE model is used to separate speech and background noise from the noisy input. Unlike our proposed approach, which is limited to using only noisy speech data for VC training, Yao *et al.* do not have this constraint, allowing the SE and VC models to be jointly trained in a multi-task learning framework. Another key difference is that, in their method, the separated noise is superimposed externally onto the reconstructed clean speech to calculate the loss with respect to the noisy speech. In contrast, our method directly inputs the noise into the decoder of the VC model, enabling it to learn the reconstruction of the noisy waveform.

3 Analysis of N2N-VC Performance Degradation

In this section, we first introduce the previously proposed N2N method and analyze its VC performance degradation under specific noisy conditions. To identify the cause of this degradation, we conduct a series of experiments based on noise categories to assess their individual impact on VC performance. Finally, we conclude the causes of performance degradation based on objective evaluation results.

3.1 Proposed Noisy-to-Noisy VC Methods

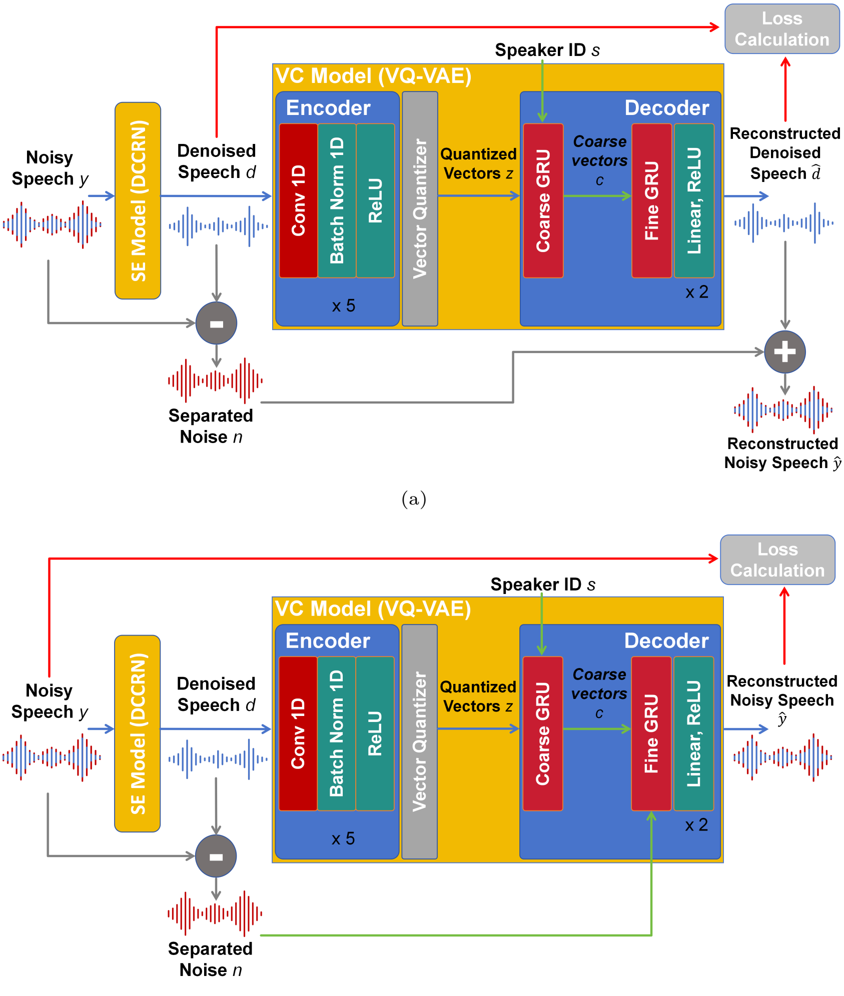
Figure 1 (a) illustrates the original N2N framework from [65] that serves as the baseline. The framework follows a cascaded design comprising off-the-shelf SE and VC models. The SE model decomposes the noisy input into speech and noise components, and the VC model is trained on denoised speech. Although noise is excluded from training, it can be superimposed onto the converted speech during conversion.

The SE model is implemented using the Deep Complex Convolution Recurrent Network (DCCRN) [18], which is a single-channel speech denoising model. It is pre-trained on the DNS Challenge 2020 [45] dataset using scale-dependent signal-to-distortion ratio (SD-SDR) loss [29], which offers performance comparable to the scale-invariant signal-to-noise ratio (SI-SNR) loss while retaining sensitivity to scaling variations in the estimated speech. This work focuses exclusively on background noise, with reverberation modeling reserved for future research.

Figure 1 (a) also illustrates the VC model architecture. It adopts a self-supervised VQ-VAE approach proposed in [39], which enables non-parallel conversion and end-to-end generation. The model comprises three main components: a content encoder, a vector quantizer, and a decoder. The content encoder is composed of a series of one-dimensional convolutional layers, batch normalization layers, and ReLU activation functions, taking the Mel spectrogram of denoised speech $\mathbf{d}$ as input. The vector quantizer employs a learnable codebook to map the encoder's output to a discrete representation $\mathbf{z}$ by selecting the nearest vectors from the codebook. The decoder is a WaveRNN-based vocoder [34], which generates the μ-law decoded $\mathbf{d}$ conditioned on $\mathbf{z}$ from the quantizer, speaker code $\mathbf{s}$, and the past samples in an autoregressive (AR) manner. The behavior of the decoder can be characterized as a conditional joint probability distribution:

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

Figure 1 (b) illustrates the improved N2N framework proposed in [64]. A major enhancement is training the VC model to reconstruct the noisy speech so that the distortion introduced by the SE model can be alleviated. To facilitate noisy speech modeling, the separated noise is incorporated during training as a conditioning input for the VC model's decoder. As shown in Figure 1(b), the decoder comprises two recurrent structures. The first gated recurrent unit (GRU) extracts the global audio features based on $\mathbf{z}$ and $\mathbf{s}$, producing a coarse speaker-related representation $\mathbf{c}$. The second GRU refines $\mathbf{c}$ by capturing finer details to enhance synthesis precision. To ensure high-quality noise synthesis, noise vectors derived from the μ-law decoded separated



(a)

(b)

Figure 1: Overall workflow of the proposed N2N VC framework. (a) Baseline framework [39]. (b) N2N framework [64].

noise $\mathbf{n}$ via an embedding layer are concatenated with $\mathbf{c}$ and fed into the second GRU. On the basis of the baseline's conditional joint probability distribution in Equation (1), we modify the noise-conditioned version as follows:

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{n}, \mathbf{s}, \mathbf{z}) &= \prod_{t=1}^{T} p(y_t \mid y_1, \ldots, y_{t-1}, n_1, \ldots, n_t, \mathbf{s}, \mathbf{z}) \\ \text{s.t.} \quad \mathbf{y} &= \mathbf{d} + \mathbf{n}. \end{aligned} \tag{2}$$

The VC model is trained by minimizing the loss:

$$L_{VC} = -\log p\left(\mathbf{y} \mid \mathbf{z}, \mathbf{n}, \mathbf{s}\right) + \beta \|E(\mathbf{d}) - \mathrm{sg}(\mathbf{e})\|^2, \tag{3}$$

where the first term represents the reconstruction loss, and the second term corresponds to the commitment loss. $E()$ denotes the encoder of the VC model, $sg()$ represents the stop-gradient operator in the vector quantizer, and $\mathbf{e}$ is the nearest embedding of $E(x)$ indexed from the codebook. β is the weight for the commitment loss, which is set to 0.25 as in the original VQ-VAE [61].

3.2 Experimental Setup and VC Performance Degradation

In our previous works [64, 63], we used the VCC2018 dataset [35] as the clean corpus, and ESC-50 [41] and DEMAND [58] as noise sources. The VCC2018 comprises 972 utterances in the training set and 420 utterances in the test set from 12 speakers with a balanced gender distribution. Of these, eight speakers were designated as sources, while the remaining four served as targets. The ESC-50 dataset provides diverse noise types with 2,000 recordings spanning 50 categories. In contrast, the DEMAND dataset includes six noise categories, further divided into 18 subcategories, while effectively representing diverse real-world environments. Each subcategory contains a five-minute, 16-channel recording, where we used channel 01 for all subcategories in our experiments.

Based on the characteristics of ESC-50 and DEMAND, we employed two noise sampling strategies to construct the noisy datasets, referred to as speaker-independent (SI) and speaker-dependent (SD), respectively.

In the SI strategy, for each utterance in VCC2018, we uniformly sampled a noise clip from the noise dataset and a signal-to-noise ratio (SNR) level between 0 and 20 dB to synthesize noisy speech. In contrast, in the SD strategy, a noise category was first assigned to each speaker in VCC2018 to associate each speaker's identity with a specific noise type. Then, each utterance from a given speaker was mixed with a randomly sampled noise clip from the assigned noise category at a fixed SNR of 5 dB. Considering the differences in noise diversity between ESC-50 and DEMAND, we applied the SI strategy to ESC-50 (denoted as E-SI) and the SD strategy to DEMAND (denoted as D-SD). In the test set, noise clips from unseen E-SI categories were sampled using the SI strategy and mixed with VCC2018 test utterances.

Figure 2 presents the Mel cepstral distortion (MCD) results for models trained on E-SI and D-SD. N2N refers to the improved method with noise conditioning illustrated in Figure 1 (b). The upper bound refers to the baseline VC model trained on the clean VCC2018 dataset, representing the N2N's theoretical maximum performance. The red frame in Figure 2 highlights VC performance degradation where N2N fails to surpass the baseline with the D-SD training set. When trained on the E-SI training set, N2N significantly

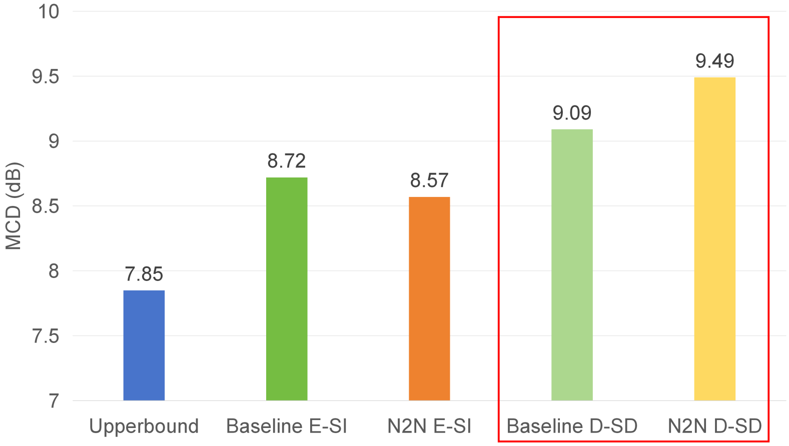


Figure 2: MCD results for methods trained on the E-SI and D-SD datasets. The red frame highlights the performance degradation observed when N2N is trained on the D-SD training set.

outperforms the baseline. However, when using the D-SD training set, noise conditioning in N2N hinders the improvements in VC performance. The baseline, trained solely on denoised speech, achieves an MCD of 9.09, whereas noise-conditioned N2N achieves a subpar MCD of 9.49.

In our previous work [63], we attributed the observed performance degradation to speaker-noise entanglement. We reached this conclusion based on the noise dataset characteristics and sampling strategy. The DEMAND dataset has limited noise diversity comprising only 18 noise subcategories, each with a single noise recording, while the D-SD dataset includes just 12 noise subcategories. Moreover, the SD sampling strategy assigns each noise category to a specific speaker and further exacerbates this lack of diversity. Together, these factors lead to speaker-noise entanglement. Therefore, we proposed a noise augmentation strategy [63] to enhance noise diversity. However, the improved N2N method still remains inferior to the baseline.

3.3 Investigating the Causes of VC Performance Degradation

We conduct a series of experiments to identify key factors affecting VC performance degradation. First, we evaluate the impact of the noise sampling strategy, which is employed to establish speaker-noise entanglement. In the previous experiment, SI and SD strategies were exclusively applied to ESC-50 and DEMAND. We expand the setup by applying the SI strategy on DEMAND and the SD strategy on ESC-50. Additionally, while earlier experiments used denoised speech and separated noise for training and testing, we now use the original noisy dataset instead. In this setup, the noise-conditioned VC model in N2N uses clean speech as input and conditions on raw noise to reconstruct noisy speech. By removing the influence of the SE model on the downstream VC task, the adjustment allows a more targeted analysis of additional contributing factors.

Figure 3 illustrates the MCD results for models trained on noisy datasets using SI and SD strategies. The upper bound remains the same as in the previous case, representing the original VC model trained solely on clean speech without noise conditioning, which serves as the theoretical performance upper bound of the N2N framework. ESC-50 and DEMAND denote the noise sources of the training data.

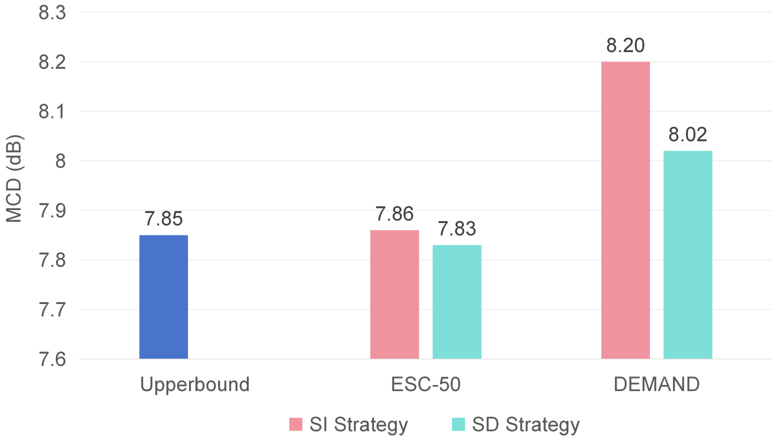


Figure 3: MCD results for models trained and tested on the original noisy datasets using SI and SD noise sampling strategies.

In conclusion, SI or SD noise sampling strategies do not lead to VC performance degradation when ESC-50 is used as the noise source. In particular, the N2N framework trained on E-SI and E-SD achieves MCD scores of 7.86 and 7.83, respectively, approaching the upper bound of 7.85. In contrast, significant performance degradation is observed with DEMAND as the noise source, regardless of the noise sampling strategy. The N2N framework trained on D-SD and D-SI achieves MCD scores of 8.02 and 8.20, respectively, both exceeding the upper bound. These results suggest that the VC performance degradation is attributed to the characteristics of the noise dataset rather than speaker-noise entanglement introduced by the sampling strategy. Moreover, the performance trends, where the N2N framework performs well when trained on the E-SI dataset but experiences performance degradation when trained on the D-SD dataset, are consistent with those shown in Figure 2, even when clean speech and raw noise are employed. This consistency indicates that the SE model is not responsible for the observed performance degradation.

In Section 3.2, we discussed that the DEMAND dataset has limited diversity in noise patterns. In contrast, the ESC-50 provides significantly greater noise diversity. Although speaker-noise entanglement caused by the SD strategy has been ruled out as a factor in performance degradation, DEMANDs limited noise diversity may hinder the VC model generalization, potentially contributing to other forms of feature entanglement. In our previous work [63], we implemented a noise augmentation approach by sampling noise clips from ESC-50 using the SI strategy to increase D-SD's noise diversity. While the augmentation alleviated the performance degradation to some extent, the improvements were limited, and the N2N framework still underperformed compared to the baseline. This suggests that, aside from the noise diversity issue, other aspects of the noise dataset also affect the observed performance degradation.

To further analyze the impact of noise on VC performance, we focus on the E-SI dataset, which provides a wide variety of noise types. First, we examine the distribution of noise categories in the E-SI dataset, as illustrated in Figure 4. The horizontal axis denotes the noise types sampled from the ESC-50 dataset to construct the E-SI training set, and the vertical axis shows the number of sampled noise clips per category. Our analysis focuses on the impact of the top 20 sampled noise categories highlighted in the red frame. Then, we construct multiple noisy training sets for each noise category, applying two distinct strategies to train the N2N framework:

- **Multi-clip sampling**: Multiple noise clips from the category are uniformly sampled and mixed with utterances from the VCC2018 training set at an SNR of 5 dB.
- **Single-clip sampling**: A single noise clip is uniformly sampled from the category. To provide sufficient data for training, the sampled clip is temporally duplicated in the time domain to generate an extended recording, from which a random segment is extracted to mix with utterances at an SNR of 5 dB.

The test set remains consistent with previous experiments using E-SI settings. Since using clean speech and raw noise clips results in relatively small MCD differences across models, and the SE model does not influence the observed VC performance degradation, we use the denoised speech as input and separated noise as the conditioning signal during training and testing.

Figure 5 illustrates the MCD results for the baseline and the N2N framework trained on a series of noisy training sets, each corresponding to a specific noise category. Within each noise category, both multi-clip and single-clip strategies were employed. Overall, within the same noise category, the N2N

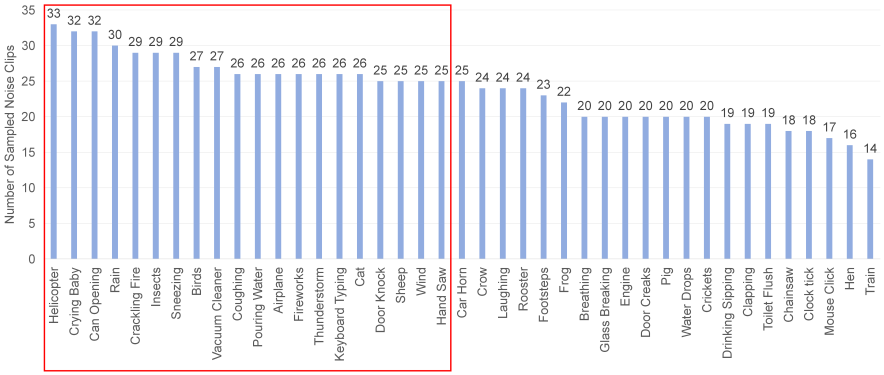
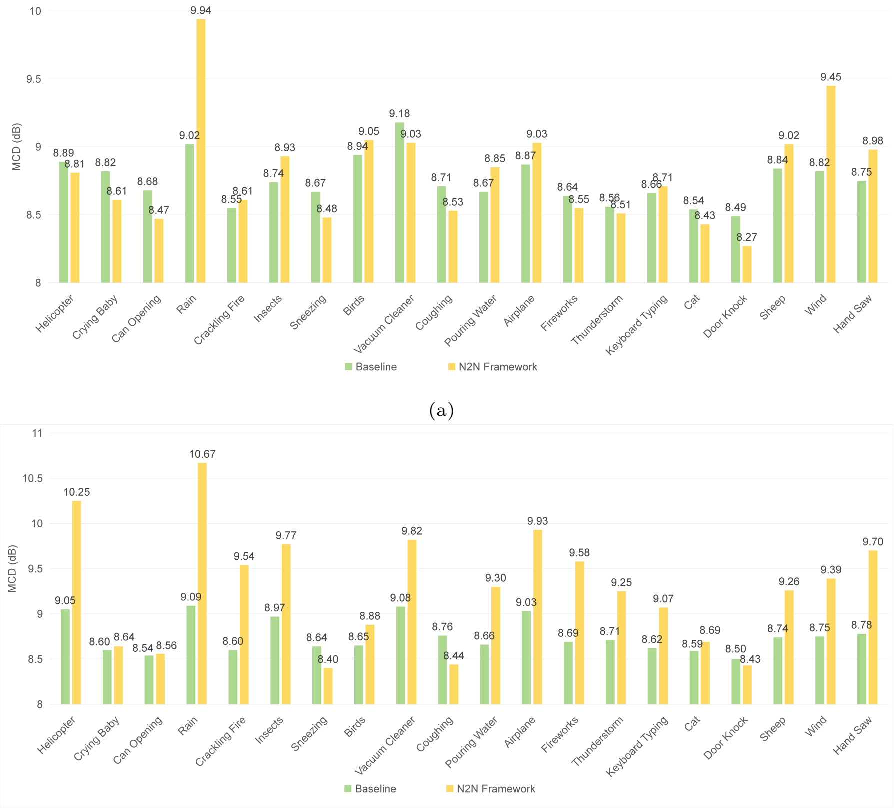


Figure 4: Noise distribution in the E-SI dataset sorted by the number of sampled noise clips. The red frame highlights the 20 most sampled noise categories.

framework trained on datasets using the multi-clip strategy consistently outperforms those using the single-clip one. Furthermore, performance degradation of the VC model is more prevalent in the single-clip group. These findings highlight the importance of noise diversity in the training set to improve VC performance. Even when the noise data comes exclusively from one noise category, a lack of noise diversity can cause feature entanglements that degrade VC performance.

As the single-clip strategy is overly restrictive and results in VC performance degradation of the N2N framework across nearly all noise categories, as shown in Figure 5 (b), we shift our focus to the experiments of the multi-clip groups in Figure 5 (a). Although previous experiments demonstrated that the N2N framework trained on the E-SI dataset does not exhibit performance degradation compared to that trained on the D-SD dataset, certain noise categories from the ESC-50 dataset, such as *rain*, *wind*, *hand saw*, and others still lead to varying degrees of performance degradation.

Initial observations of the results from the multi-clip groups indicate that stationary noise types are more likely to contribute to VC performance degradation. For example, the N2N framework experiences performance degradation when trained on *rain* and *insects*. The noises in *rain* exhibit broadband noise with wide frequency coverage and evenly distributed energy, and the noises in *insects* exhibit high-frequency dominance and consistently stable temporal distribution. In contrast, the noises in *can opening* and *crying baby* demonstrate dynamic and complex temporal properties, trained on which the N2N framework outperforms the baseline. However, although the noises in *vacuum cleaner* are wide-band and temporally stable, reflecting characteristics of stationary noise, the N2N framework trained on *vacuum cleaner* achieves



(a)

(b)

Figure 5: MCD results for the baseline and N2N trained on noisy datasets with individual noise categories using different noise sampling strategies. The horizontal axis represents the noise category involved in the training set. (a) Multi-clip noise sampling strategy. (b) Single-clip noise sampling strategy.

an MCD of 9.03, surpassing the baseline score of 9.13. This suggests that factors beyond stationarity possibly contribute to performance degradation. However, quantifying and further analyzing these characteristics remains a significant challenge.

Additionally, we also observe that the models loss function L_{VC} in Equation 3 evaluates the reconstruction of noisy speech as a whole. The lack of an additional loss term for speech reconstruction suggests that the model assigns equal importance to speech and noise components during training. However, when the noise component dominates the noisy speech or exhibits complex and hard-to-learn patterns, the VC model may allocate more capacity to modeling noise, potentially at the expense of the speech component. Therefore, we realize that the difficulty in noise modeling could also impact VC performance.

Although SNR is a direct metric for quantifying noise interference in a signal, it does not effectively reflect the level of noise dominance in our tasks. As shown in Figure 5 (a), despite all noisy utterances having a consistent SNR of 5 dB, datasets of some noise categories still cause performance degradation to a certain extent. Therefore, we explored and computed three additional metrics: MCD, PESQ, and STOI, to assess the degree of noise dominance. Specifically, those metrics were calculated between the clean utterances and their noisy counterparts from the noisy training sets in Figure 5 (a), E-SI, and D-SD. A higher MCD value indicates greater noise dominance, whereas higher PESQ and STOI scores suppose lower noise dominance.

Detailed results for these metrics can be found in Figure A.1 in the appendix. Here, MCD_N denotes the MCD between the clean corpus and its noisy counterpart in the training sets, and MCD_{VC} denotes the MCD for the converted samples produced by N2N framework. To demonstrate the relationship between noise dominance and VC performance, we computed the Pearson correlation coefficients between MCD_N, PESQ, STOI, and MCD_{VC}.

As presented in Figure 6, PESQ and STOI show relatively strong negative correlations with MCD_N, which is in our expectation that these metrics can reflect the noise dominance of the noisy dataset to some extent. In the cases of MCD_{VC}, MCD_N demonstrates the strongest correlation, PESQ shows a moderate negative relation, and STOI exhibits only a weak negative relation. This suggests that MCD_N is more effective in explaining the relationship between VC performance and the original speech distortion in training data, *i.e.*, the level of noise dominance. Overall, the correlation coefficients indicate that when the level of noise dominance is high, the N2N framework performs worse in speech conversion, because the noise-conditioned model tends to focus more on modeling the noise component. As a result, we refer to this phenomenon as noise bias.

4 Proposed Method

As discussed in Section 3.3, the primary factors contributing to the degradation in VC performance are the limited diversity of noise leading to feature entanglements and noise bias during training. To address these issues, we use a mutual information approximation for feature disentanglement and a noise dropout to mitigate noise bias. Finally, these two approaches are integrated into the N2N framework to enhance VC performance.

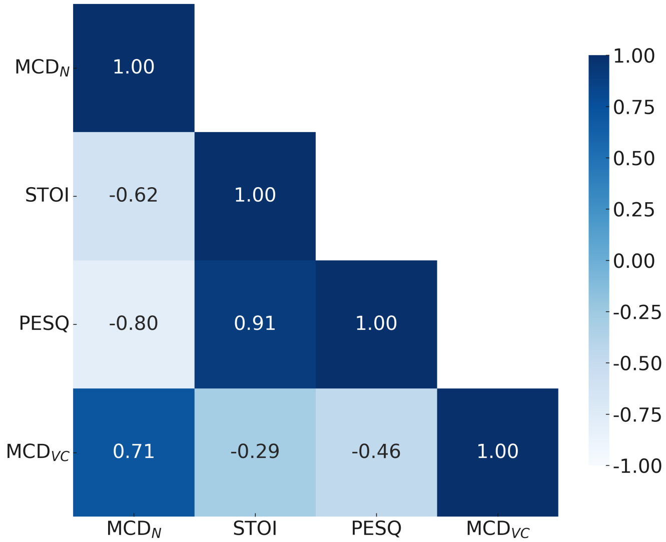


Figure 6: Pearson correlation coefficient between MCDs, PESQ, and STOI.

4.1 Mutual Information Approximation

As one of the primary factors contributing to VC performance degradation, the limited diversity of noise can lead to potential feature entanglement. However, simply increasing noise diversity through noise augmentation is problematic, as some noise types could further degrade VC performance, which is shown in Figure 5 (a). Although Section 3.3 discusses metrics for evaluating noise dominance, the threshold at which VC performance degrades in the N2N framework remains unclear and hard to quantify. Consequently, we explore an alternative approach to mitigate feature entanglement to avoid introducing additional noise data.

Mutual information (MI) is a fundamental metric that is used to quantify the dependency or shared information between two random variables. Formally, the MI between variables $\mathbf{X}$ and $\mathbf{Y}$ is defined as:

$$I(\mathbf{X};\mathbf{Y}) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right], \tag{4}$$

where $p(x,y)$ is the joint probability distribution of $\mathbf{X}$ and $\mathbf{Y}$, $p(x)$ and $p(y)$ are the marginal probability distributions of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

However, directly computing MI is often intractable, because it involves estimating the joint and marginal probability densities $p(x,y)$ and $p(x)p(y)$ in high-dimensional spaces, which is a particularly hard task. To address this, Cheng *et al.* [5] proposed a variational contrastive log ratio upper bound (vCLUB) to estimate an upper bound on MI using contrastive learning and a reformulation of the log-ratio of probabilities. The vCLUB reformulates the Equation 4 as:

$$\begin{aligned} I_{vCLUB}(\mathbf{X};\mathbf{Y}) = \mathbb{E}_{p(x,y)}\left[\log q_\theta(y \mid x)\right] \\ -\mathbb{E}_{p(x)p(y)}\left[\log q_\theta(y \mid x)\right], \end{aligned} \tag{5}$$

where the variational distribution $q_\theta(y \mid x)$ is the estimation to $p(y \mid x)$ by an approximation network with parameters θ.

In our task, we adopt vCLUB to estimate the upper bound of MI between the coarse content representation $\mathbf{c}$ and the noise vectors $\mathbf{n_v}$ to reduce their dependency, as illustrated in Figure 7. The representation $\mathbf{c}$ is the output of the first GRU, which encapsulates speaker identity and content vectors. Meanwhile, $\mathbf{n_v}$ is the continuous representation of the discrete noise input $\mathbf{n}$ obtained through an affine transformation. Based on Equation 5, the unbiased estimation for vCLUB between $\mathbf{c}$ and $\mathbf{n_v}$ is given by:

$$\begin{aligned} I_{vCLUB}(\mathbf{C};\mathbf{N_v}) = \mathbb{E}_{p(\mathbf{c},\mathbf{n_v})}\left[\log q_\theta(\mathbf{c} \mid \mathbf{n_v})\right] \\ -\mathbb{E}_{p(\mathbf{c})p(\mathbf{n_v})}\left[\log q_\theta(\mathbf{c} \mid \mathbf{n_v})\right]. \end{aligned} \tag{6}$$

The variational approximation $q_\theta(\mathbf{c} \mid \mathbf{n_v})$ is implemented with a simple neural network consisting of a stack of CNNs and linear layers. During the training process, approximation network is trained first to maximize the log-likelihood:

$$L_{MI} = \mathbb{E}_{p(\mathbf{c},\mathbf{n_v})}\left[\log q_\theta(\mathbf{c} \mid \mathbf{n_v})\right]. \tag{7}$$

After the optimization of the approximation network, its parameters are fixed, and the VC model is subsequently trained to minimize the total loss:

$$L_{Total} = L_{VC} + \lambda I_{vCLUB}(\mathbf{C};\mathbf{N_v}), \tag{8}$$

where I_{vCLUB} is the estimated upper bound of MI by the approximation network, and λ represents the weight used to control the disentanglement level and is set to 1e−3 in our experiments.

4.2 Noise Dropout Strategy

As discussed in Section 3.3, another key factor contributing to the degradation of VC performance is noise bias, which refers to the tendency of the noise-conditioned VC model to focus excessively on reconstructing noise components during training under specific noise categories. Consequently, the VC model does not sufficiently capture speech features, which degrades the overall quality of the reconstructed speech.

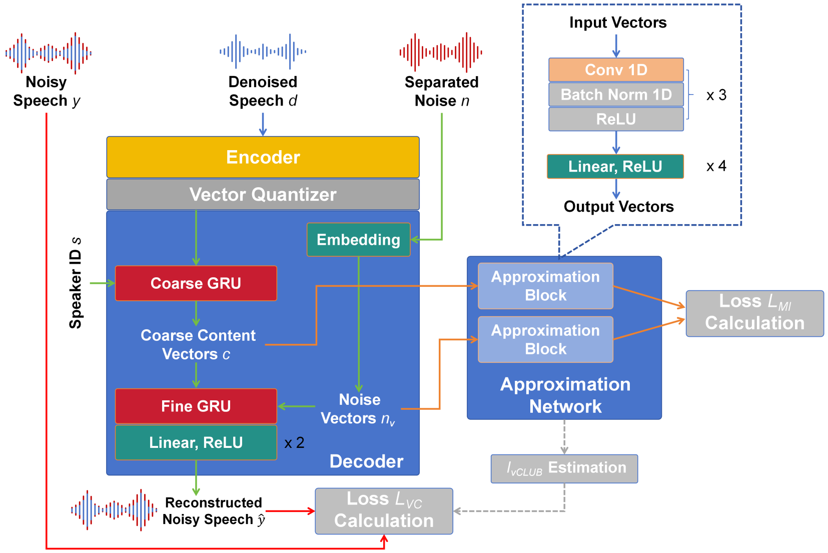


Figure 7: Improved N2N framework with MI approximation network

One possible cause of this issue is that the VC model's loss function L_{VC} in Equation 5 primarily focuses on reconstructing the noisy speech as a whole. A straightforward solution is to introduce an additional loss term specifically for the speech component reconstruction. However, clean speech data are unavailable for the VC task in our setting, yet such loss functions typically assume clean utterances as the target. In our previous work [63], we faced similar constraints and suggested reducing reliance on denoised speech as ground truth. Over-reliance on denoised speech compromises the benefits of the noise-conditioned model, which leverages noisy speech as ground truth to alleviate the distortions introduced by the SE model.

Inspired by the dropout mechanism in deep learning, we propose a noise dropout strategy to mitigate noise bias. During training, the entire noise signal is randomly replaced with a zero sequence with a certain probability. This encourages the model to reconstruct the denoised speech serving as the loss target, thus shifting the models focus back to the speech component. As previously discussed, minimizing the use of denoised speech is preferable. Therefore, the dropout rate is kept low to balance the trade-off effectively.

5 Experimental Setup

5.1 Experimental Datasets and Training Details

We adopted the same training and testing configurations as described in the previous work [63]. All audio data were sampled at a rate of 16 kHz.

The DNS Challenge 2020 dataset [44] was used to train the SE model implemented as DCCRN, as mentioned in Section 3.1. This dataset contains 500 hours of multilingual speech from over 2,000 speakers and 70,000 noise recordings spanning 150 categories. We uniformly sampled 10,000 clean utterances and 8,000 noise clips for the validation set, while the remaining data were used to construct the training set. Clean utterances and noise clips were mixed at SNRs uniformly sampled between 0 and 20 dB. The SE model was trained with the ADAM optimizer and an initial learning rate of 2e−4. The learning rate was adaptively adjusted based on validation performance, using a reduction factor of 0.5 and a patience parameter of 3 epochs. The SE model converged after 55 epochs.

As our focus is mitigating VC performance degradation, we trained the models on the D-SD datasets described in Section 3.2. The test set was consistent with previous experiments, where noise clips from categories excluded from the E-SI training set were sampled using the SI strategy and mixed with the VCC 2018 test set. The VC models were trained using the ADAM optimizer, initialized with a learning rate of 2e−4. The training phase lasted for 500k steps employing a step-based learning rate schedule. The learning rate was halved at step 100k and step 200k to improve convergence. The MI approximation network followed the same configuration as the VC models but featured a different learning rate schedule, with the rate halved at 50k and 150k steps. The noise dropout rate is empirically determined as 10%.

5.2 Methods to be Evaluated

The experiments involved two frameworks referred to as the baseline and N2N. The main difference between them is that the baseline employs the conventional VC model trained on denoised speech data, while N2N uses the noise-conditioned version trained with denoised speech as input and separated noise as a condition to model noisy speech. To differentiate method variations, we append the suffixes "MI" for mutual information approximation and "ND" for noise dropout strategy during training. We adopt the naming convention: *TypeOfModel ProposedMethod*. The objective evaluation was conducted as an ablation study including the baseline, *N2N*, *N2N MI*, *N2N ND*, and *N2N NDMI*, where *N2N NDMI* denotes the N2N framework incorporating both noise dropout and MI approximation. Following the objective evaluation results, the baseline and *N2N NDMI* were further assessed in the subjective evaluation.

5.3 Evaluation Metrics

We conducted both objective and subjective evaluations to validate the effectiveness of the proposed methods. Since our previous work has demonstrated the consistently high quality of the noise component generation and this study addresses speech component degradation, all noise-conditioned models generate the speech samples without background noise for evaluation.

For objective evaluation, we use three metrics: MCD, similarity score (SIM), and word error rate (WER). SIM is calculated using an open-source speaker verification method[1] between the converted sample and its target reference. WER is measured using a publicly available ASR model.[2]

For subjective evaluation, we conducted a preference test for naturalness and an XAB test for similarity. Based on the objective evaluation results presented in Section 3.2, we investigate whether the combination of our proposed methods could enhance the N2N framework to outperform the baseline under conditions of performance degradation. Thus, we evaluated two methods: Baseline and *N2N NDMI*. Four source speakers (VCC2SF3, VCC2SF4, VCC2SM3, and VCC2SM4) and two target speakers (VCC2TF2 and VCC2TM2) were selected for evaluation. For each conversion pair, we sampled four converted utterances, resulting in 32 utterances per model. The evaluation was conducted on Amazon MTurk with 12 participants. In the naturalness preference test, listeners were presented with paired samples from both models and asked to choose the more natural and higher-quality sample. Similarly, the XAB test for similarity has a similar procedure with an original target speaker sample as a reference. Based on this reference, listeners determined which sample sounded closer to the target speech.

6 Experimental Results

6.1 Results for Objective Evaluation

Table 1 shows the objective evaluation results from an ablation study comparing the baseline and the improved N2N framework with noise dropout and MI approximation. N2N achieves an MCD of 9.49, which is significantly higher than the baseline's 9.09. Moreover, it shows lower performance in SIM, scoring 0.743 compared to the baseline's 0.753, while the WER remains nearly identical at 30.81 and 30.80.

When MI approximation is applied to N2N, MCD improves from 9.49 to 9.27, with WER reduced from 30.81 to 29.09, while the SIM score remains nearly unchanged at 0.742. In contrast, *N2N ND* significantly improves performance compared to the original N2N, achieving an MCD of 9.06, a SIM

[1] https://github.com/resemble-ai/Resemblyzer.

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

Table 1: Objective evaluation results for the baseline and the improved N2N framework incorporating noise dropout and MI approximation, analyzed through an ablation study.

Methods	MCD (dB)	SIM	WER (%)
Baseline	9.09	0.753	30.80
N2N	9.49	0.743	30.81
N2N MI	9.27	0.742	29.09
N2N ND	9.06	0.750	29.52
N2N NDMI	**8.88**	**0.761**	**28.07**

of 0.750, and a WER of 29.52. However, *N2N ND* or *N2N MI* does not outperform the baseline across all metrics except for WER.

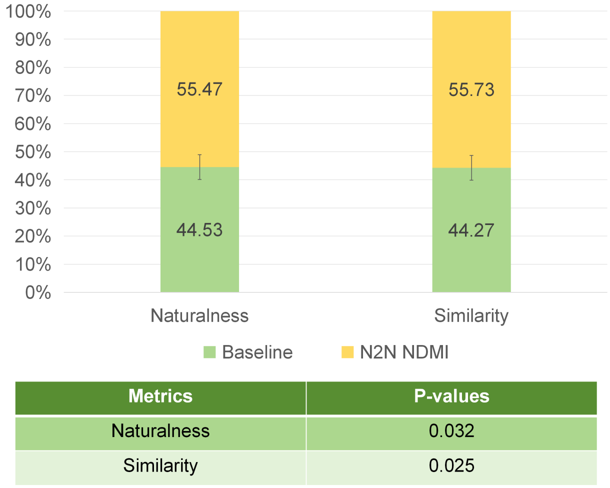
Finally, with the combination of noise dropout and MI approximation, *N2N NDMI* achieves an MCD of 8.88, a SIM of 0.761, and a WER of 28.07, outperforming the baseline across all metrics. These results demonstrate that the combined use of noise dropout and MI approximation effectively improves the N2N framework to mitigate performance degradation.

6.2 Results for Subjective Evaluation

Figure 8 shows the subjective evaluation results of the preference tests on naturalness and similarity. *N2N NDMI* achieves preference scores of 55.47% for naturalness and 55.73% for similarity, outperforming the baseline scores of 44.53% and 44.27%, respectively. The P-values for naturalness (0.032) and similarity (0.025) are below the significance threshold of 0.05, indicating statistical differences. However, the respective preference advantages are only 10.94% and 11.46%, and the lower bound of the confidence intervals is close to 50%. The observed improvements, while meaningful, are relatively limited. These results highlight that MI approximation and noise dropout contribute to improving the N2N framework, yet there remains room for further enhancement of the VC performance.

7 Conclusion

In this paper, we investigate the causes of performance degradation in the proposed N2N framework. A series of experiments were conducted to assess the impact of noise from different categories on the N2N framework. Based on the evaluation results, we identify two primary factors contributing to the VC performance degradation: the limited diversity of noise that leads to the feature entanglement, and noise bias where the noise-conditioned model tends to focus excessively on modeling the noise component rather than the speech part. To address the above issues, we propose an MI approximation method to



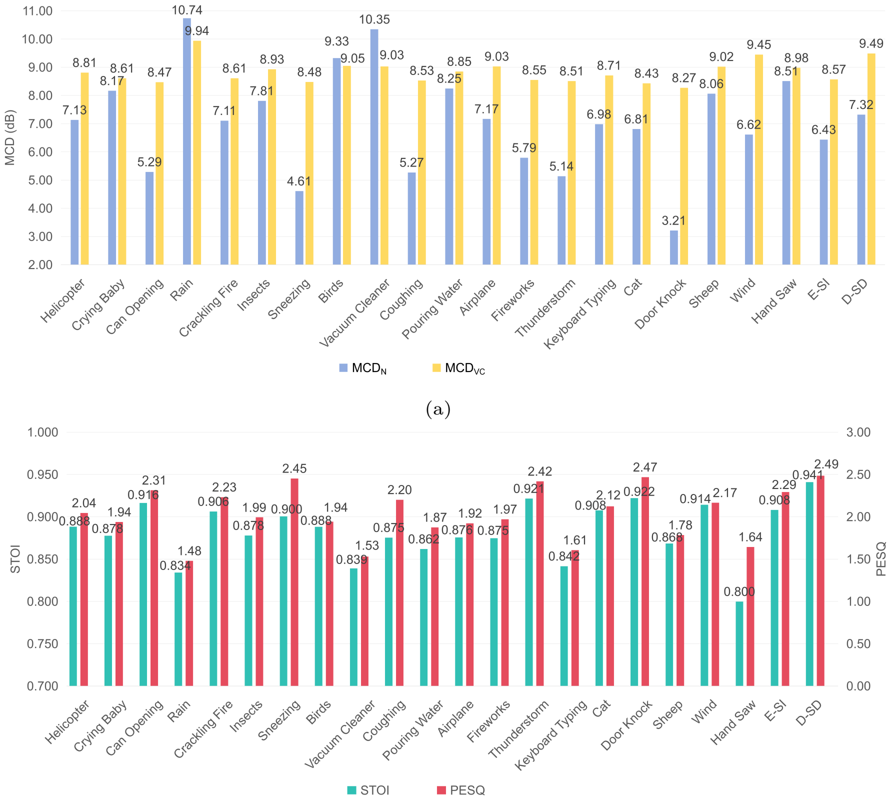
Metrics	P-values
Naturalness	0.032
Similarity	0.025

Figure 8: Preference evaluation results in terms of naturalness and similarity with 95% confidence intervals for the baseline and the N2N framework combined with noise dropout and MI approximation (P-values for naturalness and similarity are provided in the accompanying table).

enhance the feature disentanglement, and the noise dropout strategy during training to mitigate the model's focus on reconstructing the noise component. The objective evaluations were conducted in an ablation way, demonstrating the effectiveness of the proposed methods. Specifically, employing either MI approximation or noise dropout individually mitigates the performance degradation of the N2N framework. When both MI approximation and noise dropout are combined, the N2N framework achieves the best performance and outperforms the baseline. However, the subjective evaluations indicate that the improvements achieved through MI approximation and noise dropout are still limited, leaving room for further improvements. In future work, we plan to explore methods for quantifying noise characteristics and continue refining the N2N framework to address these challenges.

A Supplementary Evaluation Results

Figure A.1 shows the MCD, PESQ, and STOI results calculated between the noisy training sets in Figure 5 (a) and their clean counterparts to assess the degree of noise dominance.



(a)

(b)

Figure A.1: MCD, PESQ, and STOI results for the noisy training sets in Figure 5 (a), E-SI, and D-SD. MCD_{VC} denotes the MCD for the converted samples produced by N2N framework. (a) MCD results. (b) PESQ and STOI results.

References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice Conversion through Vector Quantization", *Journal of the Acoustical Society of Japan (E)*, 11(2), 1990, 71–6.

[2] Y. Chan, C. Peng, S. Wang, H. Wang, Y. Tsao, and T. Chi, "Speech Enhancement-assisted Stargan Voice Conversion in Noisy Environments", *CoRR*, abs/2110.09923, 2021, arXiv: 2110.09923, https://arxiv.org/abs/2110.09923.

[3] L. Chen, X. Zhang, Y. Li, and M. Sun, "Noise-robust voice conversion using adversarial training with multi-feature decoupling", *Engineering Applications of Artificial Intelligence*, 131, 2024, 107807.

[4] L. Chen, X. Zhang, Y. Li, M. Sun, and W. Chen, "A noise-robust voice conversion method with controllable background sounds", *Complex & Intelligent Systems*, 10(3), 2024, 3981–94.

[5] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information", in *International conference on machine learning*, PMLR, 2020, 1779–88.

[6] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion", in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, No. 16, 2024, 17862–70.

[7] Y. Choi, C. Xie, and T. Toda, "An Evaluation of Three-Stage Voice Conversion Framework for Noisy and Reverberant Conditions", in *Interspeech*, 2022, 4910–4, DOI: 10.21437/Interspeech.2022-10158.

[8] J.-c. Chou and H.-Y. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization", in *Proc. Interspeech 2019*, 2019, 664–8, DOI: 10.21437/Interspeech.2019-2663.

[9] G. Cong, L. Li, Y. Qi, Z.-J. Zha, Q. Wu, W. Wang, B. Jiang, M.-H. Yang, and Q. Huang, "Learning to dub movies via hierarchical prosody models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, 14687–97.

[10] J. Cong, S. Yang, L. Xie, and D. Su, "Glow-WaveGAN: Learning Speech Representations from GAN-Based Variational Auto-Encoder for High Fidelity Flow-Based Speech Synthesis", in *Interspeech 2021*, August 2021, 2182–6, DOI: 10.21437/Interspeech.2021-414.

[11] S. Dhar, N. D. Jana, and S. Das, "An adaptive-learning-based generative adversarial network for one-to-one voice conversion", *IEEE Transactions on artificial intelligence*, 4(1), 2022, 92–106.

[12] H. Du, L. Xie, and H. Li, "Noise-robust Voice Conversion with Domain Adversarial Training", *Neural Networks*, 148, 2022, 74–84.

[13] M. Dua, S. Joshi, and S. Dua, "Data augmentation based novel approach to automatic speaker verification system", *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 6, 2023, 100346.

[14] W. Gan, B. Wen, Y. Yan, H. Chen, Z. Wang, H. Du, L. Xie, K. Guo, and H. Li, "IQDUBBING: Prosody Modeling Based on Discrete Self-supervised Speech Representation for Expressive Voice Conversion", *arXiv preprint arXiv:2201.00269*, 2022.

[15] 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 *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, IEEE, 2016, 1–6.

[16] 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 *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 5901–5.

[17] H.-R. Hu, Y. Song, J.-T. Zhang, L.-R. Dai, I. McLoughlin, Z. Zhuo, Y. Zhou, Y.-H. Li, and H. Xue, "Stargan-vc Based Cross-Domain Data Augmentation for Speaker Verification", in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[18] 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, 2472–6, DOI: 10.21437/Interspeech.2020-2537.

[19] C.-y. Huang, K.-W. Chang, and H.-y. Lee, "Toward Degradation-Robust Voice Conversion", in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6777–81.

[20] C.-y. Huang, Y. Y. Lin, H.-y. Lee, and L.-s. Lee, "Defending Your Voice: Adversarial Attack on Voice Conversion", in *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2021, 552–9.

[21] R. Huang, C. Cui, F. Chen, Y. Ren, J. Liu, Z. Zhao, B. Huai, and Z. Wang, "Singgan: Generative adversarial network for high-fidelity singing voice generation", in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, 2525–35.

[22] W.-C. Huang, T. Hayashi, X. Li, S. Watanabe, and T. Toda, "On Prosody Modeling for ASR+ TTS based Voice Conversion", in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, 642–9.

[23] G. Huybrechts, T. Merritt, G. Comini, B. Perz, R. Shah, and J. Lorenzo-Trueba, "Low-resource expressive text-to-speech using data augmentation", in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, 6593–7.

[24] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel Voice Conversion with Auxiliary Classifier Variational Autoencoder", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(9), 2019, 1432–43.

[25] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, “StarGAN-VC: Non-parallel Many-to-many Voice Conversion using Star Generative Adversarial Networks”, in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, 266–73.

[26] T. Kaneko and H. Kameoka, “CycleGAN-VC: Non-parallel Voice Conversion using Cycle-consistent Adversarial Networks”, in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, 2018, 2100–4.

[27] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC2: Improved Cyclegan-based Non-parallel Voice Conversion”, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6820–4.

[28] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, “CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion”, in *Proc. Interspeech 2020*, 2020, 2017–21, DOI: 10.21437/Interspeech.2020-2280.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–Half-Baked or Well Done?”, in *Proc. 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 626–30.

[30] A. T. Liu, P.-c. Hsu, and H.-y. Lee, “Unsupervised End-to-End Learning of Discrete Linguistic Units for Voice Conversion”, in *INTERSPEECH*, 2019.

[31] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, “Diffsinger: Singing voice synthesis via shallow diffusion mechanism”, in *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36, No. 10, 2022, 11020–8.

[32] S. Liu, Y. Cao, D. Su, and H. Meng, “Diffsvc: A diffusion probabilistic model for singing voice conversion”, in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2021, 741–8.

[33] X. Liumeng, Y. Shan, H. Na, S. Dan, and X. Lei, “Learning Noise-independent Speech Representation for High-quality Voice Conversion for Noisy Target Speakers”, in *Interspeech 2022*, 2022, 2548–52, DOI: 10.21437/Interspeech.2022-570.

[34] J. Lorenzo-Trueba, T. Drugman, J. Latorre, T. Merritt, B. Putrycz, R. Barra-Chicote, A. Moinet, and V. Aggarwal, “Towards Achieving Robust Universal Neural Vocoding”, in *Proc. Interspeech 2019*, 2019, 181–5, DOI: 10.21437/Interspeech.2019-1424.

[35] 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, 195–202, DOI: 10.21437/Odyssey.2018-28.

[36] H. Lu, X. Wu, H. Guo, S. Liu, Z. Wu, and H. Meng, “Unifying one-shot voice conversion and cloning with disentangled speech representations”, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, 11141–5.

[37] C. Miao, Q. Zhu, M. Chen, J. Ma, S. Wang, and J. Xiao, “EfficientTTS 2: Variational end-to-end text-to-speech synthesis and voice conversion”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[38] 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*, 10(1), 2020, 151.

[39] 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, 4836–40, DOI: 10.21437/Interspeech.2020-1693.

[40] T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, “Convnext-tts and convnext-vc: Convnext-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion”, in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, 12456–60.

[41] K. J. Piczak, “ESC: Dataset for Environmental Sound Classification”, in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, 1015–8.

[42] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, “Diffusion-based voice conversion with fast maximum likelihood sampling scheme”, *arXiv preprint arXiv:2109.13821*, 2021.

[43] X. Qin, Y. Yang, Y. Shi, L. Yang, X. Wang, J. Wang, and M. Li, “VC-AUG: Voice Conversion Based Data Augmentation for Text-Dependent Speaker Verification”, in *National Conference on Man-Machine Speech Communication*, Springer, 2022, 227–37.

[44] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, “The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results”, in *Proc. Interspeech 2020*, 2020, 2492–6, DOI: 10.21437/Interspeech.2020-3038.

[45] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, *et al.*, “The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results”, *arXiv preprint arXiv:2005.13981*, 2020.

[46] M. S. Ribeiro, J. Roth, G. Comini, G. Huybrechts, A. Gabry, and J. Lorenzo-Trueba, “Cross-speaker style transfer for text-to-speech using data augmentation”, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6797–801.

[47] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, “Non-parallel Voice Conversion using Variational Autoencoders Conditioned by Phonetic Posteriorgrams and D-vectors”, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, 5274–8.

[48] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, “Voice Conversion Based Data Augmentation to Improve Childrens Speech Recognition in Limited Data Scenario”, *Proc. Interspeech 2020*, 2020, 4382–6.

[49] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, “Children’s speaker verification in low and zero resource conditions”, *Digital Signal Processing*, 116, 2021, 103115.

[50] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, “In-domain and out-of-domain data augmentation to improve childrens speaker verification system in limited data scenario”, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, 7554–8.

[51] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, “Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers”, *arXiv preprint arXiv:2304.09116*, 2023.

[52] D. K. Singh, P. P. Amin, H. B. Sailor, and H. A. Patil, “Data augmentation using cyclegan for end-to-end children asr”, in *2021 29th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, 511–5.

[53] Y. Stylianou, O. Cappé, and E. Moulines, “Continuous Probabilistic Transform for Voice Conversion”, *IEEE Transactions on speech and audio processing*, 6(2), 1998, 131–42.

[54] 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 on Speech Synthesis*, 2013.

[55] R. Takashima, T. Takiguchi, and Y. Ariki, “Exemplar-based Voice Conversion in Noisy Environment”, in *2012 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2012, 313–7.

[56] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, “AttS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms”, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, 6805–9.

[57] R. Terashima, R. Yamamoto, E. Song, Y. Shirahata, H.-W. Yoon, J.-M. Kim, and K. Tachibana, "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation", in *Proc. Interspeech 2022*, 2022, DOI: doi:10.21437/Interspeech.2022-11278.

[58] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-channel Acoustic Noise Database (demand): A Database of Multichannel Environmental Noise Recordings", in *Proceedings of Meetings on Acoustics ICA2013*, Vol. 19, No. 1, Acoustical Society of America, 2013, 035081.

[59] 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*, 2019, 674–8, DOI: 10.21437/Interspeech.2019-2307.

[60] 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 on Speech Synthesis*, 2016.

[61] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning", *Advances in neural information processing systems*, 30, 2017.

[62] Y. A. Wubet and K.-Y. Lian, "Voice conversion based augmentation and a hybrid CNN-LSTM model for improving speaker-independent keyword recognition on limited datasets", *IEEE Access*, 10, 2022, 89170–80.

[63] C. Xie and T. Toda, "Noisy-to-noisy voice conversion under variations of noisy condition", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2023, 3871–82.

[64] 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 *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, 6787–91.

[65] C. Xie, Y.-C. Wu, P. L. Tobing, W.-C. Huang, and T. Toda, "Noisy-to-noisy voice conversion framework with denoising model", in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, 2021, 814–20.

[66] J. Yang, Y. Zhou, and H. Huang, "Mel-s3r: Combining mel-spectrogram and self-supervised speech representation with vq-vae for any-to-any voice conversion", *Speech Communication*, 151, 2023, 52–63.

[67] Z. Yang, M. Chen, Y. Li, W. Hu, S. Wang, J. Xiao, and Z. Li, "ESVC: Combining Adaptive Style Fusion and Multi-Level Feature Disentanglement for Expressive Singing Voice Conversion", in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, 12161–5.

[68] J. Yao, Y. Lei, Q. Wang, P. Guo, Z. Ning, L. Xie, H. Li, J. Liu, and D. Xie, “Preserving Background Sound in Noise-Robust Voice Conversion Via Multi-Task Learning”, in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, 1–5, DOI: 10.1109/ICASSP49357.2023.10095960.

[69] 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, 80–98, DOI: 10.21437/VCC_BC.2020-14.

[70] X. Zhang, J. Wang, N. Cheng, and J. Xiao, “Voice conversion with denoising diffusion probabilistic gan models”, in *International Conference on Advanced Data Mining and Applications*, Springer, 2023, 154–67.