Noisy-to-Noisy Voice Conversion Capable of Controlling Background Noise

Chao XIE

Contents

Abstract

Voice conversion (VC) is a technique for modifying the non-linguistic features of speech while preserving its linguistic content. This thesis focuses on noise-robust VC for speaker identity conversion, while enabling explicit control over background noise. Specifically, the VC model is trained solely on noisy speech data from source and target speakers to convert the speaker identity while retaining both speech content and background noise. The proposed VC framework, referred to as Noisy-to-Noisy (N2N) VC, derives its name from two aspects: the first "noisy" indicates that the model is trained exclusively on noisy speech from source and target speakers, and the second "noisy" signifies that the background noise can either be retained or removed during inference.

We first propose a baseline framework for the N2N-VC task, which adopts a cascade design commonly used in conventional noise-robust VC systems. Specifically, it consists of an off-the-shelf speech enhancement (SE) module for front-end preprocessing and a VC module. Experimental results indicate that despite employing a state-of-the-art SE approach, the baseline framework still suffers from significant distortion introduced by the SE model.

To mitigate this distortion, we train the VC model to directly predict noisy speech, which is the only available data unaffected by SE-induced artifacts in the N2N task. Specifically, the separated noise is fed to the VC model as a condition to facilitate noisy

speech reconstruction. Introducing noise conditioning to assist noisy speech modeling leads to significant improvements in naturalness and speaker similarity, compared with the baseline without noise conditioning. Moreover, the noise-conditioned VC model can generate high-quality noise components in the converted noisy speech.

However, when the range of noise conditions in the training set is further expanded, the noise-conditioned N2N framework even underperforms the baseline. At first, we attribute it to the limited coverage of noise types in the noisy training set. Therefore, we adopt noise augmentation with pre-training strategies. Nevertheless, the improvement remains limited. Experimental results show that the performance drop is alleviated to some extent, the noise-conditioned method still performs worse than the baseline.

To further identify the underlying causes of the observed performance degradation, we first investigate the effect of noise sampling strategies, which was assumed to be attributed to the performance degradation. However, the results suggest that the performance degradation is primarily determined by the characteristics of the noise sources rather than sampling strategies. We therefore conduct targeted ablation studies across different noise categories to quantify how noise properties affect model performance. Based on these findings, we attribute the performance degradation to two primary factors: noise bias, where the model focuses more on reconstructing the noise component over speech, and the feature entanglement due to insufficient noise diversity.

To address these issues, we introduce a noise dropout strategy and a mutual information (MI) approximation network into the N2N framework. Noise dropout strategy aims to mitigate the model's focus on the noise component during training, while the MI approximation serves as a regularizer to enhance disentanglement between the latent feature and the noise representation. Experimental results show that either the MI approximation or noise dropout can alleviate the performance degradation. When

combined, the N2N framework achieves the best performance and outperforms the baseline.

In summary, this thesis proposes an N2N VC framework that directly models the noisy speech and preserves the background noise in the converted speech. The impact of the noise on the VC performance is analyzed, and several approaches are proposed to improve the N2N-VC framework to achieve better performance in terms of naturalness and similarity.

1 Introduction

1.1 Research Background

Voice conversion (VC) is a technique for converting non-linguistic features of a source speech, i.e., speaker identity, accent, and emotion, to a target one without changing its linguistic content. Figure 1.1 presents a typical neural-network-based VC pipeline, which consists of multiple encoders, an acoustic conversion model, and a neural vocoder. Specifically, a content encoder extracts linguistic representations from the source utterance, whereas other encoders extract target-side conditioning signals, such as speaker embeddings, prosodic factors, and emotional attributes, from a target reference utterance. The acoustic model predicts the target acoustic features conditioned on the source content and the target attribute representations. Finally, the vocoder synthesizes the converted waveform in the time-domain from the predicted features.

With the development of VC for decades, VC techniques have been extended to various applications such as noise-robust VC [1–4], movie dubbing [5, 6], singing voice conversion [7–10], and speech restoration [11]. Those new applications also introduce new requirements for the VC methods. Take noise-robust VC as an example: Unlike many existing VC methods that require both training and test speech data to be relatively clean and high-quality, noise-robust VC deals with real-world data often corrupted by various types of noise. Moreover, as deep learning-based VC techniques are data-driven, web-crawled speech data are also an important training resource, al-

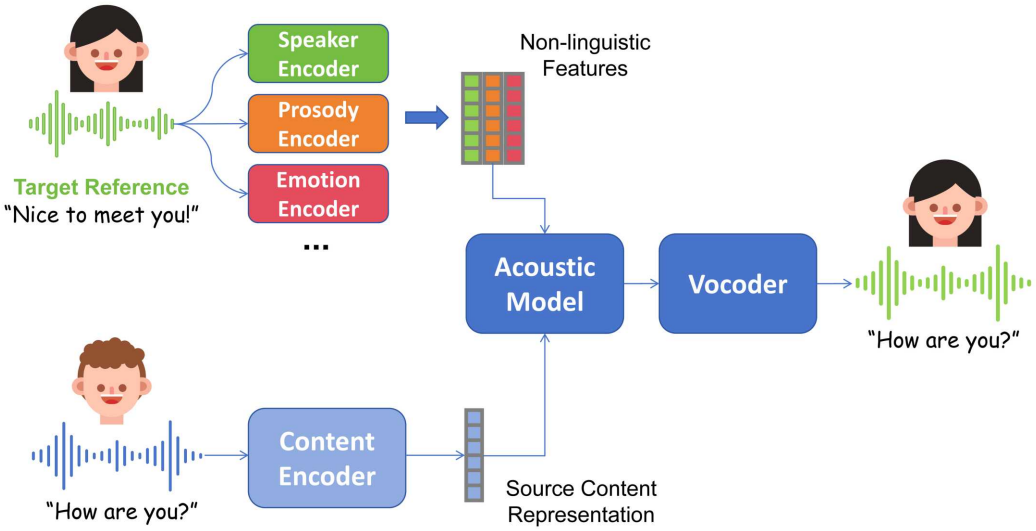


Figure 1.1: *Overview of a neural voice conversion system*

though they often contain undesired background noise that extremely degrades the performance of the VC model in terms of speech naturalness and similarity. Therefore, noise robustness has been recognized as a key attribute for real-world VC applications.

On the other hand, although background sound is often treated as interference in many noise-robust VC studies, it can also be valuable to be preserved in certain tasks. For example, recent VC techniques have been employed for data augmentation for low-resource text-to-speech (TTS) [12–14], automatic speech recognition (ASR) [15–17], and speaker verification [18–22]. The original speech datasets used in these tasks often contain inherent background noise, which can enhance the model's robustness and should be retained as a valuable training resource after conversion.

However, preserving background sounds is still not well supported by existing VC research. While conventional VC has been extensively studied, relatively few studies focus on noise-robust VC, as discussed in Section 2.2.1. Nevertheless, most of these studies treat noise as interference to be removed, and only a few can preserve back-

ground noise during conversion. In addition, these approaches typically assume access to clean training speech, which is often unavailable in the low-resource or in-the-wild settings targeted by downstream applications.

1.2 Research Scope

In this thesis, a novel noisy-to-noisy (N2N) VC framework is proposed. The first "noisy" indicates that the VC model is trained solely on noisy speech data from both source and target speakers. The second "noisy" indicates that the method models noisy speech directly, allowing for the retention or removal of background noise, or adding new noise records during inference. Unlike previous methods that treat the noise as the interference to be discarded or use clean speech data for training, the proposed VC method models the noisy speech conditioned on the separated noise. Moreover, it is noise-robust and capable of controlling the noise component in the converted waveform, and does not require clean corpora for VC training.

We first proposed a naive N2N framework following the cascading design for the noise-robust VC, which comprises an off-the-shelf speech enhancement (SE) module and a VC module. To further improve the VC performance by reducing the distortion introduced by the SE model, the VC module directly reconstructs the noisy speech as the training target conditioned on the separated noise to ease the modeling difficulty. However, subsequent experimental results show that the noise-conditioned N2N framework exhibits worse performance under certain noise conditions, compared to the baseline without noise conditioning. To understand and address this degradation, we investigate how noise conditioning affects the modeling of noisy speech and propose several methods to improve N2N-VC performance. The remainder of this section presents the exact problem formulation, the research questions investigated in this thesis, and

the in-scope and out-of-scope boundaries.

1.2.1 Problem Definition

Unlike conventional VC settings that use clean speech as the training target, N2N-VC targets noisy speech, where background noise serves as an explicit conditioning to the VC model. The objective is to convert speaker identity from source to target while preserving linguistic content and maintaining robustness across diverse noise conditions. During inference, the background sound can be either retained or removed, enabling explicit control over the noise component. Notably, both training and test data consist solely of noisy speech. To investigate the impact of introducing noise conditioning into VC, we conduct a systematic analysis to quantify how noise conditions, such as SNR level, noise category, noise diversity, and sampling strategy, affect VC quality, intelligibility, and robustness.

1.2.2 Research Questions

The research questions (RQs) in this thesis can be summarized as follows:

- **RQ1:** As a cascading approach composed of an off-the-shelf SE module followed by VC, to what extent does upstream SE quality affect downstream VC performance?
- **RQ2:** Does introducing noise conditioning into the VC module improve N2N-VC performance, compared to the baseline framework without noise conditioning?
- **RQ3:** What is the impact of limited noise diversity in training on N2N-VC performance?

- **RQ4:** How do specific noise factors, such as SNR level, noise category, and noise sampling strategies, affect the N2N-VC naturalness and speaker similarity?
- **RQ5:** Among the proposed strategies for addressing the degradation, which ones most effectively mitigate the observed degradation?

1.2.3 In-scope

This thesis is scoped to the N2N-VC framework implemented using DCCRN [23] as the SE module and a VQ-VAE-based non-parallel VC method [24] with noise conditioning as the VC module. The framework is trained and evaluated on multiple noisy datasets that are constructed using English read speech with additive noise from standard corpora across various SNR levels. Experiments are conducted first at 8 kHz and then extended to 16 kHz. The VC module is assessed on the many-to-many non-parallel VC configurations. Objective evaluations and subjective listening tests with statistical significance are conducted to evaluate the N2N-VC framework. Moreover, we conduct targeted experiments on noise sampling strategies and noise categories to investigate the effects of noise conditions on N2N-VC performance and the observed degradation. Based on these analyses, we further propose and evaluate methods based on data augmentation and mutual information estimation to mitigate the observed degradation.

1.2.4 Out-of-scope

For clarity of boundary, this work does not address the reverberant scenarios, which are left for future work. Cross-lingual, zero- and few-shot VC, emotional or style transfer, and singing voice are not included. We do not pursue performance gains by

swapping in more advanced SE or VC modules. Experimental results indicate that the observed degradation is invariant to the SE module, and the noise-conditioning implementation in the VC module already adopts the most straightforward form. Increasing architectural and conditioning complexity would likely introduce additional sources of degradation. Targeted experiments further suggest that the degradation arises from the limitation of the N2N task other than the VC backbone itself.

1.3 Main Contributions

The main contributions of this thesis can be summarized as follows:

- We propose a novel N2N-VC framework that directly models noisy speech by introducing separated noise as an explicit condition. The proposed framework improves robustness to real-world noise and can retain the background sounds during inference.
- We establish an evaluation of N2N-VC under diverse noise settings to systematically assess conversion quality in terms of speech naturalness and speaker similarity, together with the perceptual quality of the noise component in the converted samples.
- Based on extensive experimental results, we characterize the performance boundaries of N2N-VC and observe consistent degradation under specific noise conditions.
- To investigate the causes of observed degradation, we conduct a series of experiments based on key noise-related factors, including noise diversity, noise category, SNR levels, and noise sampling strategies, and clarify how the noise conditioning

affects model performance.

- Building on the above findings, we propose effective methods, including noise data augmentation and pre-training strategies, and further introduce regularization-based methods to mitigate VC performance degradation and improve robustness.

1.4 Thesis Overview

The remaining contents of the thesis are arranged as illustrated in Figure 1.2. Chapter 2 surveys related work in SE and VC, especially in noise-robust VCs and those capable of retaining the background noise. Chapter 3 presents the proposed baseline and the noise-conditioned N2N-VC framework, where the impact of the SE performance on the VC downstream is also assessed. Chapter 4 expands the noise conditions by varying noise sources and sampling strategies. Performance degradation is first observed under certain noise conditions. An initial trial to address the VC performance degradation is proposed, which combines a pre-training strategy and noise data augmentation for the VC model. Chapter 5 investigates the causes of degradation through targeted experiments. Based on these findings, Chapter 6 proposes the methods combining noise dropout and mutual information regularization to mitigate the degradation. Chapter 7 concludes the thesis and presents the future work.

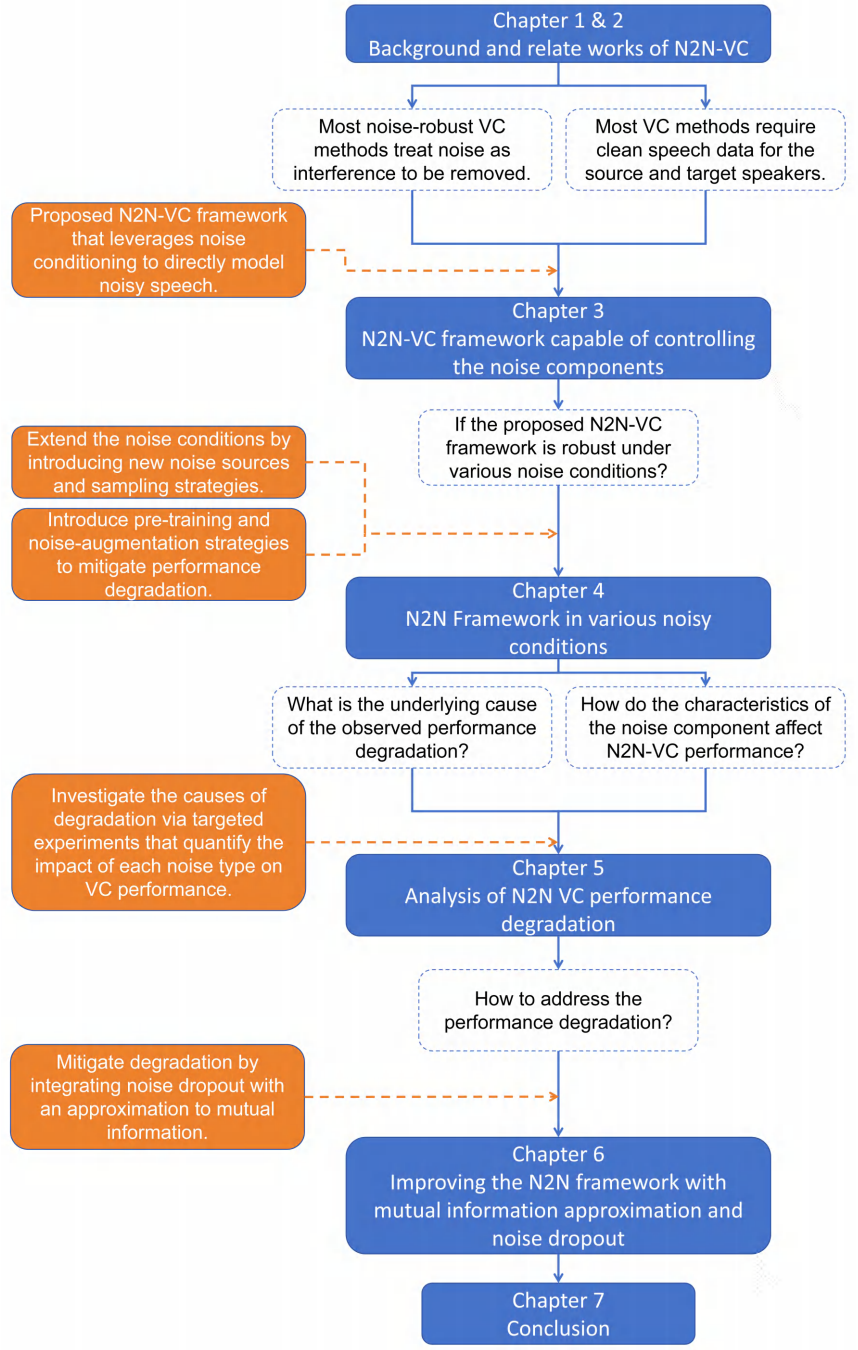


Figure 1.2: *Thesis overview*

2 Related Work

2.1 Speech Enhancement

Speech enhancement (SE) aims to improve the quality of contaminated speech signals in terms of naturalness, intelligibility, and perceptual attributes. With the rapid development of speech technologies such as automatic speech recognition (ASR), VC, and speech communication, there is an increasing demand for deploying these techniques in real-world environments. Consequently, SE has become a fundamental upstream component in building robust speech processing systems.

In real-world scenarios, speech signals are often contaminated by various types of interference, such as additive noise and reverberation. A typical noisy speech signal in such environments can be formulated as:

$$y(t) = x(t) * h(t) + n(t), \tag{2.1}$$

where $y(t)$, $x(t)$, and $n(t)$ are time-domain signals representing the noisy speech, clean speech, and additive noise, respectively. $h(t)$ represents the room impulse response (RIR), and $*$ denotes the convolution operation that models reverberation effects.

Given the observed noisy speech signal $y(t)$, the objective of the SE task is to estimate the underlying clean speech $x(t)$. The process can be formally defined as:

$$\begin{aligned} \hat{x}(t) &= \mathrm{SE}_\theta(y(t)), \\ \theta^* &= \arg\min_\theta \; d\big(\mathrm{SE}_\theta(y(t)),\, x(t)\big), \end{aligned} \tag{2.2}$$

where SE_θ represents the SE method with parameters θ, and $\hat{x}(t)$ denotes the estimated clean speech. The function $d(\cdot,\cdot)$ denotes the predefined distance metric used to compute the loss between the estimated and reference clean signals. The optimal parameters θ^* are obtained by minimizing the distance $d(\cdot,\cdot)$ to achieve a better estimated $\hat{x}(t)$.

Early SE approaches, such as spectral subtraction [25], Wiener filtering, minimum mean square error estimation [26–28], and subspace techniques [29], typically rely on several assumptions. For instance, background noise is often assumed to be short-term stationary to simplify the noise analysis and modeling. However, the mismatch between these assumptions and real-world conditions can lead to speech distortion and reduced robustness.

To improve the performance, many statistical model-based SE methods have been proposed, such as non-negative matrix factorization-based methods [30–32], Gaussian mixture models (GMMs)-based [33–38], hidden Markov models (HMMs)-based [39–44], and factor analysis [45, 46]. These methods improve robustness in non-stationary noisy environments, however, they often depend on carefully designed dictionaries or priors.

Recent advances in deep learning have substantially improved SE performance under diverse and unseen noisy conditions, as well as in low-SNR scenarios. Leveraging the power of data-driven deep neural networks (DNNs), many approaches have been proposed that surpass traditional methods in most aspects, particularly in speech quality and intelligibility.

Beyond the classification by methodological paradigm (traditional statistical methods or recent deep learning-based approaches), SE techniques can also be categorized based on other characteristics, such as the number of microphones used for speech signal acquisition (single-channel or multi-channel) and the processing domain of the

speech signal (time-domain or frequency-domain). However, as discussed in Chapter 1, the N2N-VC framework adopts the DNN-based single-channel SE method as the pre-processing component. Therefore, this chapter focuses on introducing DNN-based single-channel SE. In general, based on the training target, DNN-based SE methods can be categorized into mapping-based methods and masking-based methods.

2.1.1 Mapping-based SE Methods

The objective of mapping-based SE methods is to learn a regression function F_θ that directly maps the noisy speech to its clean counterpart:

$$\hat{\mathbf{x}} = \mathrm{F}_\theta(\mathbf{y}), \tag{2.3}$$

where $\mathbf{y}$ denotes the features of the noisy speech, such as the log power spectrum (LPS), Mel spectrogram, and magnitude spectrogram (Mag), or the raw noisy waveform, and $\hat{\mathbf{x}}$ denotes the estimated clean speech in the corresponding representation. The regression function F_θ is learned by optimizing the neural network to minimize a loss function, commonly including the mean absolute error (MAE)/mean squared error (MSE) on LPS or Mag, complex MSE, multi-resolution short-time Fourier transform (MR-STFT), and signal-to-distortion ratio (SDR)/scale-invariant SDR (SI-SDR).

As a pioneering work, Lu *et al.* [47] proposed a mapping-based SE method based on a deep denoising autoencoder to map the Mel frequency power of the noisy utterances to their clean counterparts, which establishes the core paradigm of noisy-to-clean regression using paired data. Similarly, Xu *et al.* [48] proposed a DNN-based SE method mapping the LPS features of the noisy speech to its clean counterpart. The network is trained to minimize the MSE loss between the estimated and reference LPS features. During inference, the magnitude spectrum derived from the estimated LPS and the

phase of the noisy speech are combined to reconstruct the time-domain waveform.

With the advancement of deep learning, network architectures beyond MLPs with linear layers, such as convolutional and recurrent neural networks (RNNs), have been explored. Tan and Wang [49] proposed a convolutional recurrent network (CRN)-based SE method that maps noisy STFT magnitudes to clean magnitudes in real time. The model adopts a U-Net structure consisting of a convolutional encoder–decoder with skip connections and a recurrent bottleneck. The encoder comprises five 2-D causal convolutional layers, each followed by batch normalization [50] and an exponential linear unit (ELU) activation. The decoder mirrors the encoder but replaces each 2-D convolutional layer with its transposed counterpart. A two-layer long short-term memory (LSTM) bottleneck is set between the encoder and decoder. The model is trained to minimize the MSE between the estimated and the target STFT magnitudes.

In conventional magnitude-only prediction, the phase of the noisy speech is reused during synthesis, which degrades perceptual quality, especially at low SNRs where phase errors become dominant. Therefore, complex spectral mapping and time-domain waveform mapping have been investigated, as they retain the complete information of the signal.

Défossez *et al.* [51] proposed an end-to-end SE method working in the time domain. The network adopts a U-Net–based causal DEMUCS architecture [52], which is similar to that in [49]. The encoder comprises multiple blocks that progressively downsample the noisy speech waveform. Each block contains a 1-D convolutional layer with a ReLU activation [53], and a “1×1” 1-D convolutional layer with a gated linear unit (GLU) [54] activation. The decoder has a similar structure but uses 1-D transposed convolutional layer. Skip connections are implemented between corresponding encoder and decoder blocks. The LSTM at the bottleneck conditions the current predictions

on long-range representations, reducing inter-block discontinuities and enhancing perceptual continuity. The training objective is to minimize a combination of the MAE and the MR-STFT losses between the estimated and reference waveforms.

Tan and Wang [55] proposed a gated CRN-based method capable of complex spectral mapping. The architecture extends their previous U-Net design [49] by introducing dual decoders. The encoder takes both the real and imaginary parts of the noisy spectrum as input, while the two decoders estimate the real and imaginary parts of the clean STFT, respectively. The network is trained by minimizing the MSE between the real and imaginary parts of the predicted and reference clean speech. Li *et al.* [56] proposed a two-stage SE method mapping the complex spectrum of the noisy speech to its clean counterpart. In the first stage, a network is trained to predict the clean magnitude spectrogram from the noisy magnitude, which is then combined with the phase of the noisy speech to construct a coarse complex spectrogram. In the second stage, a refinement network further processes this coarse spectrogram to predict the clean complex spectrogram.

With the development of deep learning, generative modeling has become a core technique by shifting the objective from regression to distribution alignment and perceptual quality. Many SE works also adopt the advantages of generative modeling for stronger naturalness and robustness under complex noise.

Pascual *et al.* [57] proposed a speech enhancement generative adversarial network (SEGAN). The model consists of a U-Net-structure generator that maps the noisy utterance to the estimated clean one, and a binary discriminator judging whether the paired inputs are real (target clean speech, noisy speech) or fake (estimated speech, noisy speech). A least-squares adversarial loss and an MAE for the reconstruction term are used as the loss term. In [58], the performance of SEGAN was further improved

by incorporating a self-attention mechanism into both the generator and discriminator. Yu *et al.* [59] adopted cycle-consistent GAN (CycleGAN) and an adaptive attention-in-attention module to develop a spectrogram-domain SE framework. Li *et al.* [60] proposed a perception-guided GAN for the SE task, where the discriminator predicts perceptual quality scores directly, rather than the real/fake decision.

Lu *et al.* [61] proposed DiffuSE, a pioneering work that introduced diffusion probabilistic model into SE task. In a subsequent extension, Lu *et al.* [62] introduced conditional DiffuSE, where the noisy speech is provided as a condition at every step of both the forward and reverse diffusion processes, yielding stronger metric gains and better cross-domain robustness. Welker *et al.* [63] proposed the score-based generative model for SE (SGMSE) operating on the complex STFT. Richter *et al.* [64] extended SGMSE by starting the diffusion from the noisy recording instead of Gaussian noise and training the forward model to transform clean speech into the noisy counterpart. Lemercier *et al.* [65] proposed StoRM, a two-stage diffusion-based SE model. The first stage adopts a mapping-based SE method, while the second stage refines the estimated features through a short reverse-diffusion process to mitigate artifacts and preserve speech details. Shi *et al.* [66] proposed a unified SE framework with a shared encoder, a diffusion-based generative decoder, and a predictive decoder. The predictive branch initializes the first reverse-diffusion step, and both branches are fused again at the output to exploit complementary strengths. Hu *et al.* [67] introduced a noise-classification module into diffusion-probabilistic SE, using noise embeddings as conditional inputs to the reverse diffusion so that denoising is tailored to the noise type.

2.1.2 Masking-based SE Methods

Let $Y(t, f)$, $X(t, f)$, and $N(t, f)$ denote the STFT representations of the noisy, clean, and noise signals, respectively. Their relations in the time–frequency (T–F) domain can be expressed as:

$$Y(t, f) = X(t, f) + N(t, f), \tag{2.4}$$

where t and f represent the frame-center time and the linear frequency (Hz), respectively. The core idea of the masking-based SE method is to learn a T–F mask M(t,f) that suppresses the noise component $N(t, f)$ in the noisy STFT $Y(t, f)$:

$$\hat{X}(t, f) = M(t, f) \odot Y(t, f), \tag{2.5}$$

where $\hat{X}(t, f)$ denotes the estimated clean STFT and $\odot$ represents element-wise multiplication.

As shown in Equation 2.5, the key challenge in masking-based SE methods lies in defining the mask $M(t, f)$. Based on different assumptions about the speech–noise mixture and corresponding modeling strategies, various masks have been proposed.

The ideal binary mask (IBM) is the most basic form in masking-based SE works [68–71]. For each T-F bin, if speech dominates noise that the local SNR is larger than a predefined threshold, the bin is retained; otherwise, it is suppressed. The IBM can be formulated as:

$$\mathrm{IBM}(t, f) = \begin{cases} 1, & \text{if } \mathrm{SNR}(t, f) > \theta, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

where θ denotes SNR threshold and the $\mathrm{SNR}(t, f)$ refers to the local SNR between $X(t, f)$ and $N(t, f)$, computed as:

$$\mathrm{SNR}(t, f) = 10 \log_{10} \frac{|X(t, f)|^2}{|N(t, f)|^2}. \tag{2.7}$$

However, IBM has been reported to introduce musical noise due to binary holes and discontinuities in the T–F domain [72, 73]. To address this issue, a continuous form, referred to as the ideal ratio mask (IRM), has been proposed and modeled with DNNs [73–78]. The IRM can provide smooth attenuation, which can substantially mitigate the musical noise to improve the overall perceptual naturalness. The IRM is formally defined as:

$$\mathrm{IRM}(t,f) = \left(\frac{X(t,f)^2}{X(t,f)^2 + N(t,f)^2}\right)^{\beta}, \tag{2.8}$$

where β is a tunable parameter to scale the mask. Another continuous form of mask, the ideal amplitude mask (IAM) [75], is also proposed to overcome the limitation of the IBM:

$$\mathrm{IAM}(t,f) = \frac{X(t,f)}{N(t,f)}. \tag{2.9}$$

Although the IRM and IAM improve upon the IBM in SE performance, they are real-valued mask defined only on magnitudes and thus retains the noisy phase during synthesis. To overcome this magnitude–phase inconsistency, the complex IRM (cIRM) generalizes masking to the complex domain, enabling joint enhancement of magnitude and phase [79, 80]. The cIRM can be defined as:

$$\mathrm{cIRM} = \frac{Y_r X_r + Y_i X_i}{Y_r^2 + Y_i^2} + i\,\frac{Y_r X_i - Y_i X_r}{Y_r^2 + Y_i^2}, \tag{2.10}$$

where X_r, X_i and Y_r, Y_i are the real and imaginary parts of the clean and noisy speech, respectively. The real part of the cIRM is also called the phase-sensitive mask (PSM).

As a pioneer, Williamson *et al.* [79] applied the cIRM for speech denoising and dereverberation, and later demonstrated its superiority over magnitude-domain masking [81]. Choi *et al.* [82] proposed a deep U-net with complex convolutions/normalization to estimate a polar-form complex ratio mask to reflect the distribution of the cIRM.

Hu *et al.* [23] proposed a deep complex convolution recurrent network (DCCRN) estimating cIRM-style masks, which achieves top performance in the real-time track of DNS Challenge 2020 [83]. Hao *et al.* [84] developed the FullSubNet, which cascades a full-band network using the noisy magnitude spectrogram as input and a parameter-sharing sub-band network that ingests local spectral neighborhoods to estimate cIRM directly. In [85], it is further improved by introducing phase-aware complex inputs, multi-scale temporal channel-attention components, and a TCN-based full-band network in place of LSTMs. Zhao *et al.* [86] proposed FRCRN, which augments a CRN with frequency-axis recurrence. The model operates in the complex domain to predict a cIRM and is optimized with a joint objective that combines time-domain SI-SNR with MSE losses on the real and imaginary mask components.

Unlike masks defined by formulas based on clean/noise STFT pairs and used as explicit training targets, some methods treat the mask as a multiplicative gate over a latent representation. In these approaches, the mask range is fixed by design via activation functions such as tanh, sigmoid, softmax, or ReLU with normalization, while the mask itself is learned implicitly in an end-to-end manner under reconstruction-based objectives.

Luo *et al.* [87] proposed TasNet, which is an end-to-end time-domain method for speech separation and enhancement. The model comprises an encoder mapping the noisy waveform into a latent representation, a separation module predicting element-wise gating masks for the latent representation, and a decoder for waveform reconstruction. Subsequently, Luo *et al.* [88] introduced Conv-TasNet to further improve the performance, which replaces the LSTM components with a temporal convolutional network (TCN) along with residual and skip connections. Fu *et al.* [89] proposed MetricGAN, which integrates GANs into mask prediction. Based on this framework,

several variants have been developed: MetricGAN+ [90] that further improves SE performance, MetricGAN-U [91] that extends the approach to a self-supervised training manner, and CMGAN [92] that incorporates a Conformer backbone. Westhausen *et al.* [93] proposed the dual-signal transformation LSTM network (DTLN) for real-time SE task. The model includes two LSTM-based separation modules, each estimating latent masks in an end-to-end manner. Park *et al.* [94] proposed an end-to-end SE framework based on a multi-scale autoencoder that learns a latent mask over the encoder output.

2.1.3 Evaluation Metrics for SE

This section introduces the commonly used objective and subjective metrics for SE tasks. These metrics are designed to capture different aspects, including perceived quality and intelligibility. Relying on a single metric is often insufficient. Therefore, in most cases, objective and subjective metrics are typically combined to provide a more comprehensive and robust evaluation of the SE performance.

Objective Metrics for SE

For perceived quality, the perceptual evaluation of speech quality (PESQ; ITU-T P.862) is a widely used metric. PESQ compares the estimated speech with its clean reference and produces a mean opinion score-listening quality objective (MOS-LQO) score in the range of approximately 1 to 4.5. The perceptual objective listening quality assessment (POLQA; ITU-T P.863) is also an intrusive metric. POLQA extends PESQ to wideband and super-wideband conditions and exhibits greater robustness to time warping and spectral distortions. Beyond these intrusive measures, non-intrusive DNN-

based predictors, Quality-Net [95], MOSNet [96], NISQA [97, 98], and DNSMOS [99, 100], estimate MOS-like scores directly from the signal without a clean reference.

For intelligibility, the short-time objective intelligibility (STOI) is commonly adopted. STOI measures the extent to which the estimated utterance preserves the short-time temporal envelopes of the clean reference and outputs a score in $[0, 1]$. The extended STOI (ESTOI) extends STOI by evaluating longer spectro-temporal segments with cross-band dependencies, which is more robust in modulated noise conditions.

In addition, some metrics for downstream tasks are also used to assess SE performance. The most common are word error rate (WER) and characteristic error rate (CER), which are primarily designed for ASR. However, the transcription text is required. Several metrics of signal fidelity and distortion are also used, including the signal-to-distortion ratio (SDR), acale-invariant SDR (SI-SDR), and scale-invariant SNR (SI-SNR).

Subjective Metrics for SE

Absolute category rating MOS (ACR MOS, ITU-T P.800) is widely conducted to assess the overall speech quality. The test set typically covers multiple speakers, noise types, and SNRs. Listeners are provided with a single 6–12 second-long utterance at a time and rate its quality on a five-point scale (1 – Bad, 2 – Poor, 3 – Fair, 4 – Good, 5 – Excellent).

ITU-T Recommendation P.835 evaluates perceptual quality along three aspects: speech signal quality (SIG), background noise intrusiveness (BAK), and overall quality (OVRL). Similar to MOS, it follows a single-stimulus paradigm with five-point categorical scales. For SIG: 1 – Very distorted to 5 – Not distorted; for BAK: 1 – Very intrusive to 5 – Not noticeable; and for OVRL: the same scale as MOS.

MUSHRA (ITU-R BS.1534-3) is a standardized multi-stimulus listening test originally designed for codec evaluation but now widely applied to SE when distortions are moderate and multiple systems must be compared efficiently. In each rating task, listeners assess multiple versions of the same utterance, which includes a labeled reference, a hidden reference, one or more degraded references by low-pass filters, and denoised samples from systems under test. The assessing scores are continuous from 0–100, corresponding to quality levels from bad to excellent.

2.2 Voice Conversion

Voice conversion (VC) aims to convert the timbre and other speaker-specific characteristics of a source utterance to those of a target speaker, while preserving the original linguistic content. VC has been extensively studied for decades before the advent of deep learning. Early approaches were primarily based on statistical modeling of speech signals. Many proposed methods, such as Gaussian mixture model [101–104] and hidden Markov model [105–109], vector quantization [110–112], and exemplar-based sparse representation [113–115], have established the foundation for modern approaches. However, most statistical methods depend on parallel corpora, where source–target sentence pairs must be carefully aligned by forced alignment or dynamic time warping. Even minor alignment errors can lead to spectral mismatch and audible artifacts. Their representational capacity is also limited under multi-speaker, cross-lingual, and noisy/reverberant conditions. Moreover, most statistical approaches only model the spectral envelope, while F0 and phase are handled with simple linear transforms or reused from the source, which constrains the perceptual quality in terms of naturalness and similarity.

With the emergence of deep learning, neural network-based methods have continu-

ously improved the naturalness and similarity of synthesized speech [116–118]. Moreover, modern deep learning approaches enable any-to-any, zero-shot, and real-time conversion that are challenging for statistical models. A typical deep learning-based VC framework often consists of a content encoder that extracts speaker-invariant content representations, a decoder that conditions on speaker embeddings to predict acoustic features, and a neural vocoder that synthesizes the waveform. In some cases, a speaker encoder is also introduced to enhance the representation of speaker information and better capture speaker-specific style characteristics. According to the development timeline of deep learning, neural network-based VC can be broadly categorized into autoencoder-based, GAN-based, text-to-speech (TTS)-based, flow-based, and diffusion-based.

Autoencoder-based VC aims to disentangle the linguistic content and speaker identity. To prevent the information leakage between content and speaker, information bottlenecks and regularization are adopted, such as discrete bottleneck [119–124], architectural bottlenecks [125–129], and cycle consistency [130–134]. As a special subclass of autoencoders, variational autoencoders (VAEs) [135–139] replace the disentangling feature representations with distributions and introduce a prior-regularized latent space via Evidence Lower Bound (ELBO) optimization, which differentiates them from conventional reconstruction-driven autoencoders.

In the deep learning-based techniques, GANs provide a powerful framework for generative modeling, where a generator learns to synthesize samples that fool a discriminator trained to distinguish real from synthetic data. Since GAN is basically a training strategy rather than a single architecture, it has been integrated into various VC frameworks to improve perceptual naturalness and preserve target speaker characteristics, such as CycleGAN-VC [140–143], StarGAN-VC [144–146], and VAE-GAN [147, 148].

Moreover, GANs are also widely used in neural vocoders [149–155], which also play an important role in VC frameworks.

TTS refers to techniques that map symbolic or learned linguistic representations, such as text, phonemes, or self-supervised discrete units, to acoustic features or waveforms, optionally conditioned on controllable attributes like speaker identity, prosody, style, emotion, and language. Because these controls naturally coincide with the objectives of VC, TTS has been widely adopted as a content-to-speech pipeline. In TTS-based VC [156–159], a speaker-independent content representation is first extracted from the source utterance, typically using an ASR model or self-supervised speech representations. This representation is then re-synthesized in the target timbre through a conditioned TTS back end [159–162]. Compared with previous VC approaches, TTS-based VC achieves stronger disentanglement of linguistic content from speaker and channel factors, resulting in state-of-the-art naturalness and flexibility. Furthermore, it provides finer control over prosody and style, even in non-parallel and cross-lingual settings.

Flow-based methods [163] belong to the probabilistic side of generative modeling. Different from VAE-based approaches which optimize a variational lower bound with an approximate posterior, flow-based models compute tractable densities and are trained by exact maximum likelihood estimation (MLE), thereby avoiding the inference gap. In flow-based VC [164–169], a single invertible and differentiable transformation maps speech features to a simple latent prior, which conditions on speaker identity, prosody, and content representations. During inference, the model first encodes the source utterance into a latent code under the source condition and then decodes the same code under the target condition to convert the target timbre while preserving linguistic content. Because flow-based VC relies purely on MLE, training is more stable than VAE-

or GAN-based approaches. Moreover, due to the reversible and non-autoregressive mapping, inference can be executed in parallel, enabling low latency for real-time applications.

Diffusion models [170] are also a probabilistic branch of generative modeling, which are trained to reverse a stochastic noising process by predicting noise under a likelihood-style objective. Diffusion models have become popular since 2021 and have been shown to match or surpass GANs on image synthesis. [171–173]. In diffusion-based VC [174–178], a forward process gradually mixes the clean target features with Gaussian noise based on a variance schedule. Then, a DNN-based denoiser is optimized across time steps to recover the clean target from noisy counterparts conditioned on content, speaker, and optional prosody representations. During inference, the model reconstructs target speaker features by iterative denoising from Gaussian noise or distilled few-step solvers conditioned on the source content and target speaker representations to preserve source content while controlling prosody. Last, a neural vocoder synthesizes the waveform from the converted speech feature.

Despite recent DNN-based VC approaches demonstrating strong generative capability and being capable of any-to-any, zero-shot, and real-time tasks, most of them rely on high-quality training data and become fragile when deployed to noisy environments. This limitation significantly constrains practical deployment. Compared to these conventional VCs that have been extensively studied, research on noise-robust VC has received comparatively less attention.

2.2.1 Noise-robust Voice Conversion

In the realm of statistic-based methods, Takashima *et al.* [113] proposed an exemplar-based noise-robust VC method, in which noisy speech is represented as a weighted

combination of parallel source/target exemplars and noise exemplars extracted from non-speech regions. During inference, the clean target speech is reconstructed from the target speech exemplars and weighted source exemplars while ignoring the noise exemplars, thereby achieving conversion and denoising simultaneously. Both clean source and target speech data are needed for training this method. Subsequently, Takashima *et al.* [1] further improved the approach by introducing compact non-negative matrix factorization (NMF) speech bases that share a common activity matrix across source and target to replace large exemplar bases.

With the advent of deep learning, the performance of noise-robust VC has been further enhanced. For deep learning-based methods, there are basically two strategies to achieve noise-robust VC. The first is a cascading framework, where an SE module is applied as a preprocessing step to discard the noise components in a noisy mixture before the VC module. The second is a direct mapping strategy, where the VC model itself learns to convert noisy source speech into clean target speech, thereby performing conversion and denoising simultaneously.

Cascading Framework

Valentini-Botinhao *et al.* [179] investigated an RNN-based SE method for noise-robust TTS. An SE front end implemented by LSTMs is used to map noisy acoustic features to their clean counterparts before training a TTS acoustic model. The authors compared two training strategies for the SE component, each corresponding to a different training target. The first strategy trained the SE module to enhance the vocoder features directly, including Mel-Cepstrum (MCEP), F0, and aperiodicity. The second enhanced only MCEP from the magnitude spectrum. Then, an estimated clean waveform was reconstructed by combining the enhanced MCEP with the noisy phase,

from which the vocoder features were re-extracted for TTS training. Experimental results show that the second route achieves higher subjective scores comparable to those obtained with clean training data.

Following a similar idea, Chan *et al.* [2] proposed an SE-assisted StarGAN framework. A bidirectional LSTM (BLSTM)-based SE serves as a denoising front end, while the VC back end is implemented as StarGAN-VC as the back end. The SE module takes noisy log-Mel spectrograms as input and outputs the estimated clean spectrograms, which are then fed into the VC module. Since both SE and VC modules are jointly trained, paired noisy–clean utterances are required for training. Experimental results show that the framework with joint training achieves higher subjective quality and improved robustness on unseen noises, which outperforms the variants without the SE module or joint training.

Miao *et al.* [3] proposed a noise-robust VC based on BLSTM. The method combines two lightweight filters for pre- and post-processing. The pre-filters apply time-domain low-pass processing on the waveform to suppress high-frequency noise introduced during recording. The MCEPs of the filtered waveform are split into low- and high-quefrency bands, and two BLSTM converters are trained to map each sub-band from source to target. The outputs from both sub-bands are integrated via a weighted fusion to smooth the band boundary. A statistical post-filter normalizes the mean and variance of the converted MCEPs to match the target speaker ' s training statistics. During inference, the post-filtered MCEPs are fed to a vocoder to synthesize the converted waveform.

Choi *et al.* [4] proposed a three-module cascading VC framework consisting of one VC and two SE modules to handle background noise and reverberation separately. The first SE module, implemented as DCCRN [23], performs denoising, while the second,

implemented as TasNet [87], performs dereverberation. This design allows each SE module to control noise and reverberation independently in the converted samples, which is further investigated in the subsequent work [180]. The study further found that the processing order of denoising and dereverberation had only a minor impact on VC performance. The VC module is based on a VQ-VAE [24] operating on log-Mel spectrograms. Given that the two SE modules are pre-trained and fixed, the framework does not require clean training data for VC.

Noisy-to-clean Mapping Methods

Although using an SE method for noise preprocessing is straightforward, it can introduce additional distortions to the denoised speech features. As a result, a direct noisy-to-clean mapping strategy has been proposed that avoids relying on SE models.

Mottini*et al.* [181] proposed a noise-robust non-parallel VC framework based on Auto-VC [125]. The denoising training strategy is conducted to guarantee the robustness against noisy and reverberant conditions. The training set is augmented by introducing systematic data exposure to additive noise and realistic room impulse responses across SNRs and T60s. During training, the model is always optimized to reconstruct the clean counterpart, regardless of whether the input is noisy or reverberant.

Du *et al.* [182] incorporated domain adversarial training (DAT) into a disentanglement-based VC model, which is built on a zero-shot AdaIN-VC framework [183]. Each of the two domain discriminators, judging whether the input is noisy or clean, is attached to the content encoder and the speaker encoder, respectively, via separate gradient reversal layers. In this way, encoders are explicitly forced to learn noise-invariant representations of speech content and speaker identity, based on which the decoder learns to synthesize target speech. The training set is augmented with noise recordings across

SNRs to construct paired noisy-clean corpora. The entire model is trained in a denoising manner without the SE module. During inference, the learned encoders output noise-robust latent representations regardless of whether the source or target reference is noisy. Compared with its non-DAT baseline, the proposed model exhibits cleaner and better generalization to unseen noise conditions.

Xue *et al.* [184] proposed a two-part noise-robust VC framework based on noise-controllable Glow-WaveGAN [185] and a flow-based conversion model [186]. In Glow-WaveGAN, the encoder maps a raw waveform to a latent code intended to capture only speech content and speaker identity while excluding noise information. To achieve this, a feature-wise linear modulation (FiLM) is first adopted for the decoder switch to specify whether the output is supposed to be a clean or noisy waveform. Then, during training, the model sees paired clean and noisy utterances and reconstructs both from the same latent code but with different decoder switches. In this way, noise is factorized into a controllable condition rather than being entangled with linguistic and speaker information. In the conversion part, a flow-based VC method is trained to map phoneme posteriorgrams (PPG) to that noise-independent latent space. During inference, the decoder ' s switch is set to " clean " so that the system generates clean converted utterances.

Chen *et al.* [187] proposed a noise-robust VC framework that disentangles content, speaker, and pitch representations while enhancing noise invariance through adversarial training. The framework consists of three encoders, extracting the content, speaker information, and pitch representations separately. Based on these representations, the AutoVC decoder [125] reconstructs the Mel spectrograms of the input speech, which is then converted to the waveform by a neural vocoder. Mutual information minimization with a vCLUB estimator [188] is applied across content, speaker, and

pitch representations to enhance feature disentanglement. Noise robustness is achieved by attaching a noise-decoupling discriminator to the content and speaker encoders via gradient reversal layers. During training, the model receives paired noisy-clean utterances as the inputs, while the reconstruction is always referenced to the clean Mel spectrogram.

He *et al.* [189] proposed a noise-robust VC that targets noisy reference utterances. A diffusion-based VC method [190] is adopted as the VC backbone, which consists of a source encoder receiving clean source speech as the input and a diffusion acoustic model predicting target Mel spectrograms. The source encoder includes a semantic extractor implemented with a fixed HuBERT model and a pitch extractor by WORLD [191]. A dual-branch reference encoder is introduced to extract a noise-invariant speaker representation. During training, one branch encodes clean reference speech and the other encodes its noisy counterpart. Noise-agnostic contrastive speaker loss is applied to maximize similarity between noisy–clean pairs from the same speaker while minimizing similarity across different speakers. In this way, the reference encoder learns to extract robust speaker representation from the target regardless of the acoustic environment.

Igarashi [192] incorporated recording quality and acoustic environment characteristics of noisy source speech into conventional denoising training to improve the noise-robust VC. Specifically, an utterance-wise or frame-wise recording quality vector an environment vector are extracted via pre-trained NISQA [193] and PaSST [194] models, respectively. These vectors are concatenated with the source and target SSL features and fed into the baseline S2VC model [195]. By explicitly conditioning the VC model on degradation factors during training, the approach improves generalization to unseen noise compared to standard denoising training. Experimental results show that frame-wise PaSST conditioning is the most effective, which yields better objective scores and

higher subjective naturalness and similarity than the unconditional baseline.

2.2.2 Noise-robust Voice Conversion with Noise Preservation

As discussed in Section 2.2.1, recently proposed noise-robust VC methods can be categorized into cascading frameworks and noisy-to-clean mapping approaches, depending on whether they leverage SE methods. However, these approaches typically treat noise components as interference to be discarded. With the growing application of VC techniques to tasks, such as speech data augmentation and singing VC, there is a growing demand to preserve informative background sounds as a resource. Consequently, some recent studies have focused on noise-robust VC methods that are capable of retaining background noise.

As a pioneering effort, Hsu *et al.* [196] proposed a method that combines noise augmentation with adversarial invariance to address the entanglement between speaker identity and recording conditions in multi-speaker TTS. The model is formulated as conditional generation with two latent factors representing speaker characteristics and background noise. Two independent encoders, a speaker encoder and a residual encoder, are introduced to extract these latent factors. To enforce disentanglement effectively, the training data are augmented with various noise categories across SNR levels so the model observes the same speaker under multiple environments. In addition, a noise classifier is attached to the speaker encoder via a gradient-reversal layer, which encourages the speaker latent representation to be noise-invariant and forces noise information into the residual encoder ' s output. Trained with standard reconstruction objectives plus the adversarial constraint, the system learns two independent factors controlling speaker and noise, enabling consistent clean synthesis for all speakers and editable noise conditions. However, the quality of the generated noise remains poor:

fine-grained details are lost, and the model tends to produce an averaged noise across the training set, which is perceived as white noise.

In another work [197], Yao *et al.* proposed an end-to-end noise-robust VC framework that cascades SE and VC models. The SE model is implemented as DCCRN [198] optimized with a power-law compressed phase-aware and asymmetric losses, which predicts clean speech and background noise separately. The clean speech is then passed to an ASR-based extractor to obtain content representations, which are fed into the VC module consisting of a convolutional CLSTM encoder and a HiFi-GAN-style decoder [152] to generate the target clean waveform. A discriminator receiving the predicted clean waveform is introduced to further improve perceptual quality. To encourage background preservation, the separated noise is externally superimposed onto the reconstructed clean waveform for unified reconstruction loss computation, which ensures simultaneous voice conversion and noise retention. During training, the SE and VC modules are separately trained first and then jointly fine-tuned to reduce their mismatch.

Chen *et al.* [187] proposed a controllable background sounds VC framework following the cascading approach. The SE module is implemented as an enhanced DCCRN with a dual-branch structure and an inter-branch bridge, enabling simultaneous prediction of speech and background noise. The VC module is based on VQMIVC [122], which comprises three encoders for content, speaker timbre, and F0 with mutual-information minimization, and a decoder that reconstructs the target Mel spectrogram. A cycle-consistency regularizer is further employed to enhance VC performance. During training, the SE and VC models are coupled with a unified reconstruction objective to mitigate the cascade mismatch. It should be noted that the estimated noise component from the SE model is excluded from the VC training process. During inference,

the SE model separates the noisy mixture into speech and noise components, the VC operates on the estimated speech only, and the background noise can be remixed with the converted speech.

2.2.3 Evaluation Metrics for VC

Objective Metrics for VC

Mel-cepstral distortion (MCD) is a widely used objective metric in VC that quantifies frame-aligned spectral-envelope error between a reference utterance and a converted one. It is defined as:

$$\mathrm{MCD} = \frac{10}{\ln 10}\sqrt{2\sum_{d=1}^{D}(c_d - \hat{c}_d)^2} \tag{2.11}$$

where c_d and $\hat{c}_d$ denote the d-dim Mel-cepstral coefficient (MCEP) of the reference and converted speech, respectively, and D is the number of coefficients. Lower MCD indicates closer spectral envelopes and correlates with reduced spectral distortion. However, it does not reliably predict perceived naturalness and speaker similarity. Therefore, it is often calculated as a reference and should be complemented with subjective listening tests.

ASR-based evaluation is widely adopted to assess the content preservation. The ASR system first recognizes the converted speech, then its transcript is compared with the reference text to compute WER or CER. Lower error rates indicate better preservation of lexical content. In addition, in recent years, DNN-based estimators, such as Quality-Net [95], MOSNet [96], and DNSMOS [99,100], are often adopted to estimate MOS-like scores for evaluating the naturalness of the converted speech.

Automatic speaker verification (ASV) is often adopted to quantify similarity. In practice, an off-the-shelf ASV model is used to obtain an averaged target embedding

as the reference. Then, each converted utterance is embedded and scored with the reference via cosine similarity or PLDA. Because ASV focuses on the features of the speaker identity rather than perceptual quality, conversions with artifacts may still be overestimated with high similarity scores. Therefore, subjective listening tests remain necessary to assess similarity.

Subjective Metrics for VC

ACR MOS (ITU-T P.800) is widely used to assess the naturalness of the converted speech. Listeners are provided with a single 6–12 second-long utterance at a time on a five-point scale (1 – Bad, 2 – Poor, 3 – Fair, 4 – Good, 5 – Excellent). Tests should be conducted in a quiet environment, and listeners must wear headphones.

A MOS-like speaker similarity test (SIM) [117] is often used to evaluate the similarity of converted utterances. Listeners are provided with an original utterance from the target speaker as a reference and a converted utterance at a time. The evaluation score is often a four-point scale(1 – Definitely the same, 2 – Maybe the same, 3 – Maybe different, 4 – Definitely different). The SIM score is typically reported as the percentage of responses rated 1 or 2.

Comparative MOS tests, also referred to as AB or XAB tests, are also adopted to compare VC systems in terms of naturalness and speaker similarity. In the AB test, listeners are presented with two converted utterances (A and B) in a single trial and indicate which sounds more natural. In the XAB test, a reference utterance X from the target speaker and two converted utterances from two VC systems are presented at a time. Listeners are then asked which of the utterance sounds more natural compared to the reference, or which sounds more similar to the reference speaker. Results are typically reported as the percentage of preferences for each system.

2.3 Summary of This Chapter

This chapter briefly introduced related work on single-channel SE and VC.

In single-channel SE, approaches can be broadly categorized into mapping-based and masking-based methods depending on the training target. Mapping-based methods directly map noisy waveforms or complex spectra to their clean counterparts, while masking-based methods learn to estimate a time–frequency mask, such as IRM, cIRM, and PSM, which is applied to the noisy representation.

Early studies on mapping and masking typically adopted magnitude-only objectives and reused the noisy phase for inverse STFT reconstruction. This implicitly limited the potential quality due to phase errors. To overcome this limitation, phase-aware training objectives were introduced. Mapping-based methods adopted direct waveform reconstruction or complex-valued networks for the complex domain, while the masking-based method predicted the complex-domain masking to estimate real and imaginary components to correct both magnitude and phase.

More recently, generative objectives have also been incorporated into the SE task, which yields higher perceptual quality ceilings and improved robustness to non-stationary noise. In general, mapping-based methods are more straightforward and tend to be more robust to wide SNR variability, often achieving more aggressive noise suppression under low-SNR conditions. By contrast, masking-based methods yield more stable training and greater interpretability, and typically produce more natural perceptual quality [199].

With the rapid development of deep learning, VC techniques have experienced substantial progress, particularly in perceptual quality and speaker similarity. The improved generative capability has also encouraged the application of VC to new scenarios, which in turn brings new challenges. Among these, noise-robust VC is of particular

importance, as it directly affects the deployment of VC techniques in real-world environments, their use in data augmentation, and the need to leverage abundant but low-quality web data, given the data-driven nature of deep learning.

However, compared with the extensive research on conventional VC, only a limited number of studies have focused on noise-robust VC. Based on whether an SE model is adopted, noise-robust VC methods can be categorized into cascading frameworks and noisy-to-clean mapping approaches. A cascading framework consists of an SE module followed by a VC module, while a noisy-to-clean mapping approach is often achieved by denoising training with data augmentation. Huang *et al.* [200] compared these two implementations of noise-robust VC. Their experimental results showed that cascading frameworks typically yield higher perceptual naturalness but may suffer from additional distortions and mismatches introduced by the SE module, whereas noisy-to-clean mapping approaches require no extra modules, exhibit stronger generalization to unseen degradations, and often preserve speaker identity more effectively.

3 Baseline and the Improved N2N Frameowrk

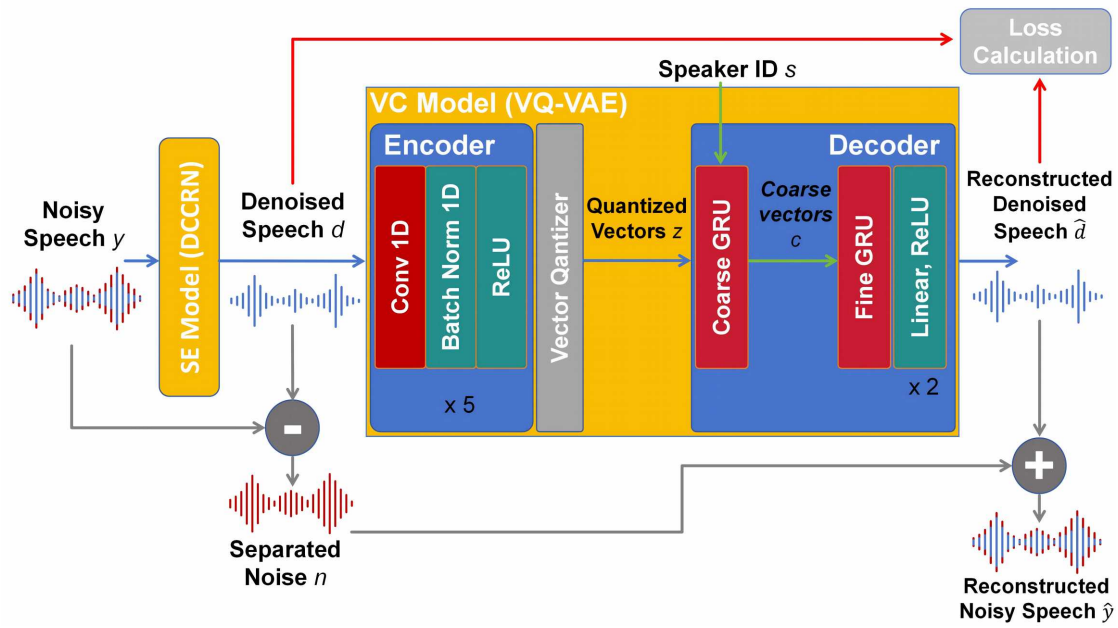
3.1 Introduction

As mentioned in Section 2.2.1, most existing noise-robust VC methods treat noise as interference to be removed. Only a few methods are capable of retaining background noise. However, those methods require clean speech for training, making them unsuitable for the N2N scenario.

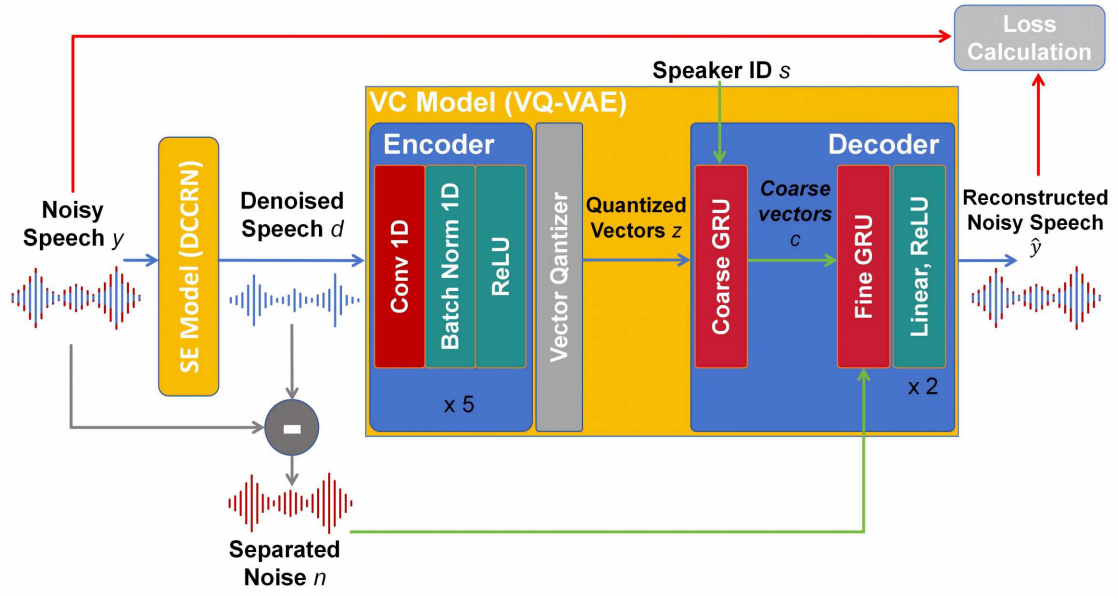
This chapter presents the proposed N2N framework, which does not require clean training data and enables background noise retention during inference. First, a primitive framework [201] following the conventional cascading structure is introduced, which serves as the baseline for the N2N task. To improve the VC performance, a new strategy is proposed based on the baseline, where the VC method is modified to model the noisy speech directly. Both subjective and objective evaluations were conducted to demonstrate the effectiveness of the proposed method.

3.2 Baseline Framework

Figure 3.1 (a) illustrates the overall workflow of the baseline framework, which follows the conventional noise-robust VC method by cascading an SE model with a VC model. Since clean speech is unavailable for VC training, the SE module is pre-trained



(a)



(b)

Figure 3.1: *Overall workflow of the proposed N2N VC framework. (a) Baseline framework [24]. (b) N2N framework [202].*

on a large-scale dataset and kept fixed during VC training to ensure stable denoising performance. During training, the pre-trained SE model processes the noisy input to separate speech from background noise, and the VC model is trained on the resulting denoised speech. During inference, the noisy speech is similarly processed by the SE model, and the denoised output is fed into the VC model. The separated background noise can optionally be added back to the converted speech, depending on the application requirements.

To investigate the impact of the SE model on the VC task, two state-of-the-art single-channel SE methods are implemented: DCCRN [23] and Conv-TasNet [88]. Since these models are used for speech-noise separation, it is essential that the power of the estimated speech matches that of the clean target. Hence, the original SI-SNR training loss is replaced by scale-dependent signal-to-distortion ratio (SD-SDR) loss [203] to remain sensitive to scaling variations in the estimated speech. The SD-SDR loss is formulated as:

$$\text{SD-SDR} = 10\log_{10}\left(\frac{\|\alpha s\|^2}{\|s-\hat{s}\|^2}\right), \tag{3.1}$$

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

$$\alpha = \frac{\hat{s}^\top s}{\|s\|^2}. \tag{3.2}$$

The VC model is implemented using a self-supervised VQ-VAE-based method [24] that supports non-parallel conversion. Its structure is illustrated in Figure 3.1 (a). The model consists of three components: an encoder, a vector quantizer, and a decoder. The encoder comprises multiple convolutional blocks, each containing a one-dimensional convolutional layer, a batch normalization layer, and a ReLU activation function. The vector quantizer employs a learnable codebook to discretize the encoder's output by selecting the nearest vectors. The decoder is a WaveRNN structured vocoder [204], which

predicts the μ-law decoded denoised waveform $\mathbf{d}$ based on the quantized representation $\mathbf{z}$ from the quantizer, speaker code $\mathbf{s}$, and the previous samples in an autoregressive (AR) manner. The decoding process can be expressed as a conditional joint probability distribution:

$$p\left(\mathbf{d} \mid \mathbf{s}, \mathbf{z}\right) = \prod_{t=1}^{T} p\left(d_t \mid d_1, \ldots, d_{t-1}, \mathbf{s}, \mathbf{z}\right). \tag{3.3}$$

During training, the VC model is optimized by minimizing the loss:

$$L_{VC} = -\log p\left(\mathbf{d} \mid \mathbf{s}, \mathbf{z}\right) + \beta \|E(\mathbf{d}) - \mathrm{sg}(\mathbf{e})\|^2, \tag{3.4}$$

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 [205].

3.3 Improved N2N Framework with Noise-conditioning

Based on the evaluation results of the baseline framework in Section 3.5.2, it is evident that SE evaluation metrics do not accurately reflect the contributions of SE methods to downstream VC performance. Even a state-of-the-art SE method can introduce unavoidable distortions, which degrade the naturalness and similarity of the converted speech. Moreover, the residual noise in the denoised speech is also a major factor impairing the VC performance. A key limitation of the baseline lies in its use of distorted, denoised speech as the training target for the VC model, which exacerbates performance degradation. Another drawback is the redundancy in converting noisy target speech: the converted speech is generated first, then the separated background noise is superimposed.

Among all the signals available for N2N tasks: denoised speech, noisy speech, and separated noise, only the noisy speech remains unaffected by distortions from the SE model. Therefore, incorporating noisy speech as the training target is expected to reduce distortion and improve VC performance. In this approach, the VC model is trained to reconstruct the corresponding noisy speech. This enables loss computation against the original noisy signal, thereby mitigating the adverse effects introduced by the SE model. Additionally, modeling the noisy speech allows the VC model to directly generate the noisy converted speech, thereby avoiding the two-stage generation process used in the baseline.

Modeling noisy speech is a challenging task. To ease this process, the separated noise is introduced as an additional conditioning input to the VC model, as shown in Figure 3.1 (b). An embedding layer is added to transform the μ-law decoded noise signal along the time axis to a sequence of high-dimensional vectors, which is then fed into the Gated Recurrent Unit (GRU) layer in the decoder. During training, the VC model receives denoised speech as input and separated noise as a condition to reconstruct the corresponding noisy speech. During inference, providing a noise signal to the decoder results in noisy converted speech generation, whereas replacing it with a zero vector produces clean converted speech. Based on the conditional joint probability distribution defined for the baseline in Equation (3.3), the noise-conditioned version is modified as follows:

$$\begin{aligned} &p\left(\mathbf{y} \mid \mathbf{n}, \mathbf{s}, \mathbf{z}\right) = \prod_{t=1}^{T} p\left(y_t \mid y_1, \ldots, y_{t-1}, n_1, \ldots, n_t, \mathbf{s}, \mathbf{z}\right) \\ &\text{s.t.} \quad \mathbf{y} = \mathbf{d} + \mathbf{n}. \end{aligned} \tag{3.5}$$

The training loss is also modified based on the original loss in Equation (3.4):

$$L_{VC} = -\log p\left(\mathbf{y} \mid \mathbf{n}, \mathbf{s}, \mathbf{z}\right) + \beta \|E(\mathbf{d}) - \mathrm{sg}(\mathbf{e})\|^2. \tag{3.6}$$

Introducing the separated noise signal as a condition simplifies the generation of noisy speech, enabling the decoder to more effectively learn its distribution. Additionally, the VC model can generate the noisy converted speech directly in a single step.

3.4 Experimental Setup

Due to the original research motivation is to augment the noisy speech data for speaker recognition of telephone speech, N2N VC was initially developed with an 8 kHz sampling rate. The following experiments in this thesis adopt 16 kHz.

3.4.1 Dataset for SE Model

Deep Noise Suppression (DNS) Challenge 2020 dataset [83] is used to train the SE models. It comprises 500 hours of carefully selected clean speech from 2,150 multilingual speakers and 70,000 noise clips spanning 150 categories. A total of 6,000 speech recordings and 500 noise clips are randomly sampled to form the validation set, and the remaining data are used for training. Noisy speech is generated by mixing clean utterances with noise clips at uniformly sampled SNRs between 5 and 20 dB.

3.4.2 Dataset for VC Model

For the VC model, VCC2018 dataset [117] and PNL 100 Nonspeech Sounds [206] are used as the clean corpus and the noise source, respectively.

The VCC2018 contains 972 utterances (81 utterances per speaker) in the training set and 420 (35 utterances per speaker) in the test set from 12 speakers with a balanced gender distribution. Of these, eight speakers (VCC2SF1, VCC2SF2, VCC2SF3,

VCC2SF4, VCC2SM1, VCC2SM2, VCC2SM3, VCC2SM4) are designated as source speakers, and the remaining four (VCC2TF1, VCC2TF2, VCC2TM1, VCC2TM2) as target speakers. All source and target speakers are included in the objective evaluation. However, for subjective evaluation, the number of tested utterances is reduced to ensure the perceptual quality of listening tests. Therefore, the non-parallel (SPOKE) task of VCC 2018 [117] is adopted, with four source speakers (VCC2SF3, VCC2SF4, VCC2SM3, and VCC2SM4) and two target speakers (VCC2TF2, VCC2TM2) forming eight conversion pairs in total.

The PNL 100 Nonspeech sounds contains 100 environmental recordings across 20 categories. Noise clips N1-N85 (85 clips in total, from 9 categories) are used to construct the noisy training set by mixing with the clean corpus from VCC2018 at uniformly sampled SNR levels of 6, 8, 10, 12, 14, 16, 18, and 20 dB. The remaining clips N86-N100 (15 clips in total, from the remaining 11 categories) are used to construct multiple parallel test sets, with each set corresponding to a single fixed SNR level chosen from -5, 0, 5, 7, 10, 11, 15, 19, 20, 25, or 30 dB.

3.4.3 Methods Under Evaluation

The evaluation is organized into two categories. The first assesses the performance of the baseline denoted as *Baseline-D* and *Baseline-CT*, which use DCCRN and Conv-TasNet as the SE module, respectively. The second evaluates the improvements introduced by the noise-conditioned N2N framework, denoted as *N2N*. *Upper bound* refers to the VC model trained and evaluated on the original clean corpus, representing the theoretical maximum performance of the N2N framework. For comparison, a VC model trained directly on noisy speech without SE processing is also included, denoted as *Lower Bound*. In general, the evaluated methods are: *Upper Bound*, *Lower Bound*,

Baseline-D, *Baseline-CT*, and *N2N*.

3.4.4 Evaluation Metrics

To show and compare the performance of the SE methods in *Baseline*, several objective metrics are computed: SI-SDR, signal-to-artifact ratio (SAR), PESQ, and STOI. For the VC model, MCD is used as the objective evaluation metric.

In the subjective evaluation, the MOS is applied to assess the naturalness of the converted speech. Participants rate each sample on a 5-point scale (1 = Bad, 5 = Excellent). In the first evaluation category (baseline evaluation), six utterances are randomly selected for each conversion pair: three with an SNR of 7 dB and three with an SNR of 15 dB, resulting in 48 utterances per method. In the second evaluation category (N2N evaluation), the number of utterances per conversion pair is increased to ten, with five from each SNR level, yielding 80 utterances per method.

To evaluate speaker similarity, a SIM test [117] is conducted, where each participant listens to a converted sample and a reference sample from the target speaker, and rates the similarity on a 4-point scale: 1: Definitely the same; 2: Probably the same; 3: Probably different; 4: Definitely different. The similarity score is calculated as the percentage of responses rated 1 or 2. Four converted utterances are randomly selected for each conversion pair: two with an SNR of 7 dB and two with an SNR of 15 dB. This results in a total of 32 utterances per method.

Additionally, in the second evaluation category, two XAB tests are conducted to directly compare speaker similarity: one between the *Baseline* and *N2N*, and the other between the *Upper Bound* and *N2N*. The converted utterances used in the XAB test are identical to those used in the MOS test.

Since the goal of this study is the N2N task, all converted samples in the subjective

evaluation retain background noise. For *Upper Bound*, the clean converted speech is superimposed with the original noise clip to construct the noisy converted speech. Participants are instructed to disregard the noise and focus solely on speech quality. However, as MCD is not suitable for assessing noisy speech, all methods produce clean converted speech for objective evaluation.

3.4.5 Training Details

DCCRN is implemented by Asteroid [207], with the "DCCRN-CL" configuration applied. The window length, hop size, and FFT length are set to 50 ms, 12.5 ms, and 512, respectively. The model is trained with a batch size of 64 using the Adam optimizer. The initial learning rate is set to 1e-4, and decays by a factor of 0.5 if the validation loss does not decrease within four epochs. An early stopping mechanism is employed to select the optimal model based on validation loss. For Conv-TasNet, a pre-trained model provided by Asteroid is used, which is trained on the single-speaker speech enhancement task from the Libri3Mix dataset [208].

The VC model follows the implementation in [24]. The window length, hop size, and FFT size are set to 20 ms, 5 ms, and 1024, respectively. The model is trained with a batch size of 64 using the Adam optimizer, with an initial learning rate of 2e-4. The learning rate is halved after 300k steps, and training is conducted for a total of 600k steps.

Table 3.1: *Objective evaluation results of SE methods on the noisy VCC2018 training set.*

Methods	SI-SDR (dB)	SAR (dB)	PESQ	STOI
DCCRN	**21.49**	**22.11**	**3.47**	**0.98**
Conv-TasNet	19.16	19.80	3.13	0.96

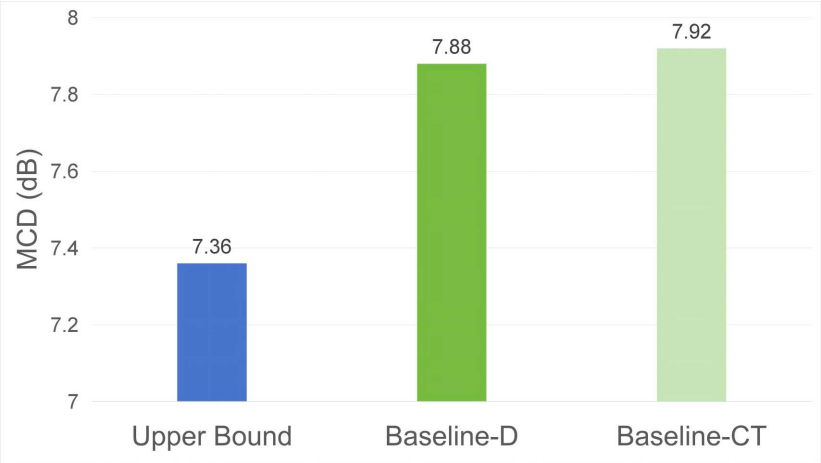


Figure 3.2: *Average MCD results for Upper Bound, Baseline-D, and Baseline-CT. Upper Bound is evaluated on the original clean corpus as a reference.*

3.5 Experimental Results

3.5.1 Evaluation Results for Baseline

As the first step, the performance of two SE methods is evaluated on the noisy VCC2018 training set. The noisy VCC2018 training set is used instead of the test set because the former includes a greater variety and larger number of noise clips from the PNL100 dataset.

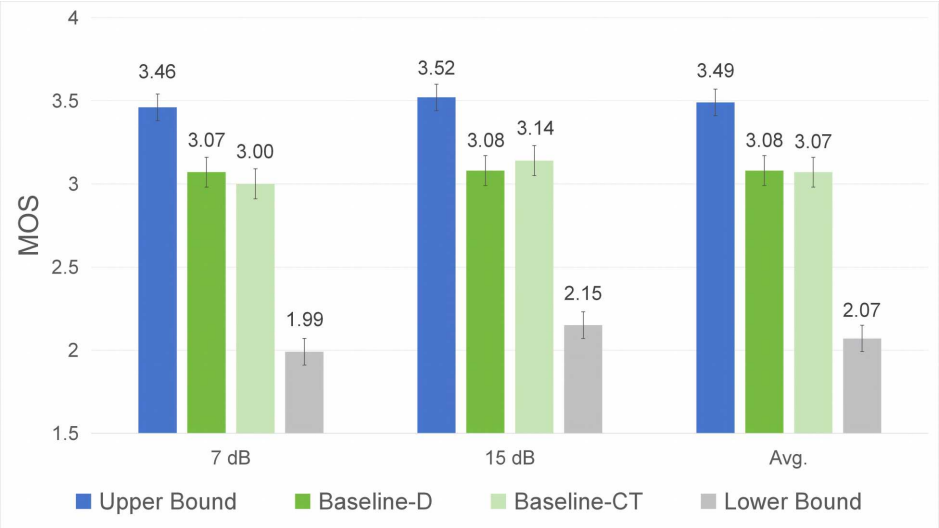


Figure 3.3: *MOS results with 95% confidence intervals for Upper Bound, Lower Bound, Baseline-D, and Baseline-CT.*

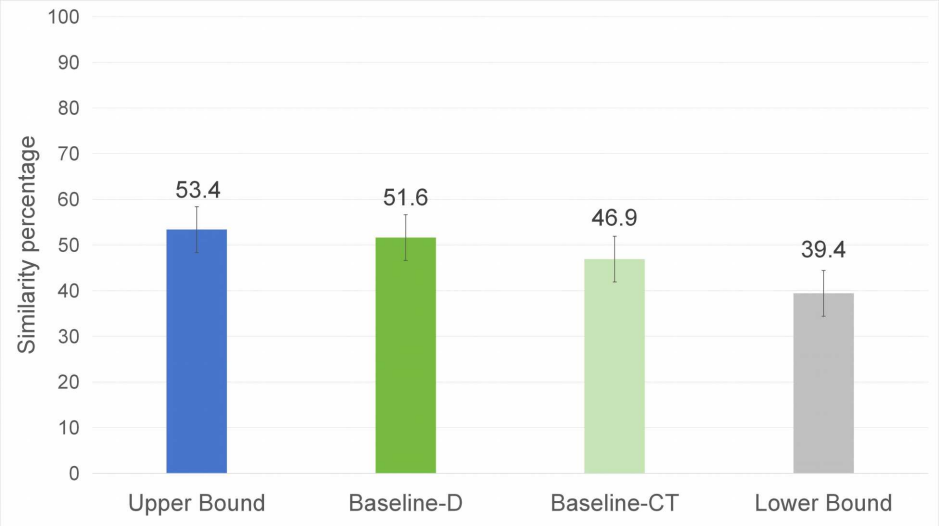


Figure 3.4: *SIM test results with 95% confidence for Upper Bound, Lower Bound, Baseline-D, and Baseline-CT.*

As demonstrated in Table 3.1, DCCRN outperforms Conv-TasNet across all SE metrics, with a particularly notable advantage in PESQ, where DCCRN achieves a

score of 3.47 compared to Conv-TasNet of 3.13. However, these improvements in SE performance do not contribute to better VC outcomes, as shown in Figure 3.2. *Baseline-D* (using DCCRN as SE model) and *Baseline-CT* (using Conv-TasNet as SE model) achieve similar MCD scores of 7.88 and 7.92, respectively, indicating no significant advantage from the better SE performance. *Upper bound* achieves the best performance with an MCD of 7.36, highlighting a clear gap between it and the two *Baseline* methods.

Figure 3.3 presents the MOS results for *Upper Bound*, *Lower Bound*, *Baseline-D*, and *Baseline-CT*. *Baseline-D* and *Baseline-CT* achieve MOS of 3.07 and 3.08, respectively, significantly outperforming the *Lower Bound* of 2.07. However, a noticeable gap remains between the two *Baseline* methods and the *Upper Bound*, which achieves the highest average MOS of 3.49. Although *Baseline-CT* slightly outperforms *Baseline-D* at 15 dB SNR (3.14 vs. 3.00), their overall performance is still similar.

Figure 3.4 shows the SIM test results for the same four methods. The trends are consistent with the MOS results: both *Baseline* methods significantly outperform the *Lower Bound* (39.4%) but fall short of the *Upper Bound* (53.4%). Notably, *Baseline-D* achieves a higher similarity score (51.6%) than *Baseline-CT* (46.9%), indicating better speaker similarity.

In general, *Baseline-D* and *Baseline-CT* significantly outperform the *Lower Bound*, but a clear gap remains compared to the *Upper Bound*. Although DCCRN outperforms Conv-TasNet across all SE metrics, its advantages do not fully carry over to the downstream VC task: *Baseline-D* and *Baseline-CT* exhibit comparable performance in terms of MCD and MOS, with *Baseline-D* outperforming only in the SIM test. These results suggest that even the state-of-the-art SE method can introduce distortions that hinder VC performance. Based on the overall results, DCCRN is adopted

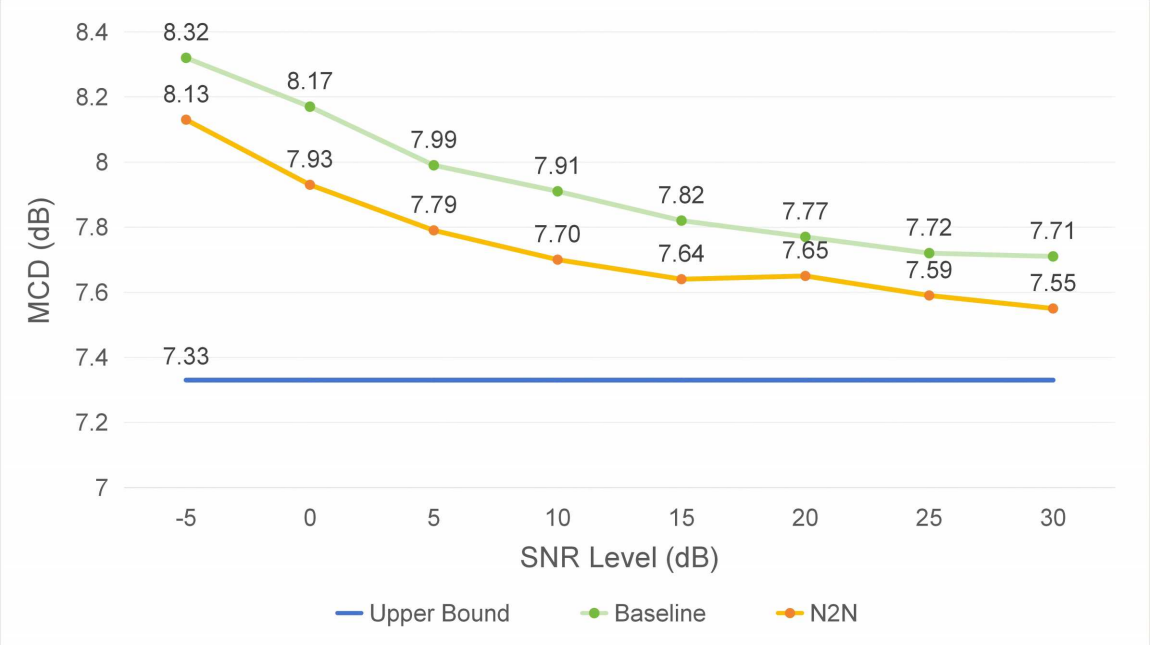


Figure 3.5: *MCD results for Upper Bound, Baseline, and N2N across multiple test sets, each constructed at a fixed SNR level.*

as the SE module for the proposed N2N framework in subsequent experiments. For brevity, *Baseline-D* is hereafter referred to as *Baseline*.

3.5.2 Evaluation Results: Baseline vs. Noise-conditioned N2N

Figure 3.5 presents the MCD results of *Baseline* and *N2N* across multiple test sets with SNR levels ranging from -5 dB to 30 dB, where the MCD is plotted as a function of SNR level. The *Upper Bound*, trained and evaluated on clean speech, achieves the best MCD score of 7.33. As the SNR increases, both *Baseline* and *N2N* exhibit improved performance. Notably, *N2N* consistently outperforms the *Baseline* across all SNR conditions, effectively narrowing the performance gap with the *Upper Bound*.

Figure 3.6 presents the MOS results. The *Ground Truth*, referring to the original

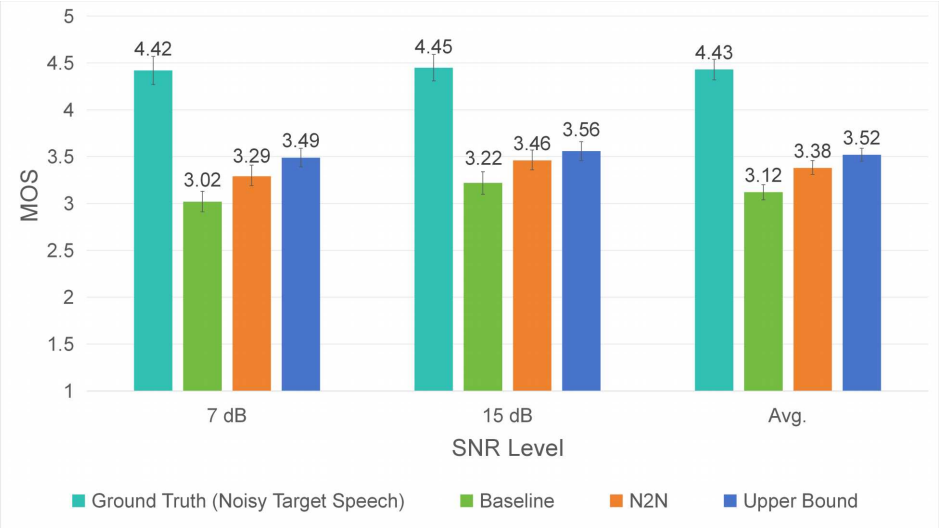
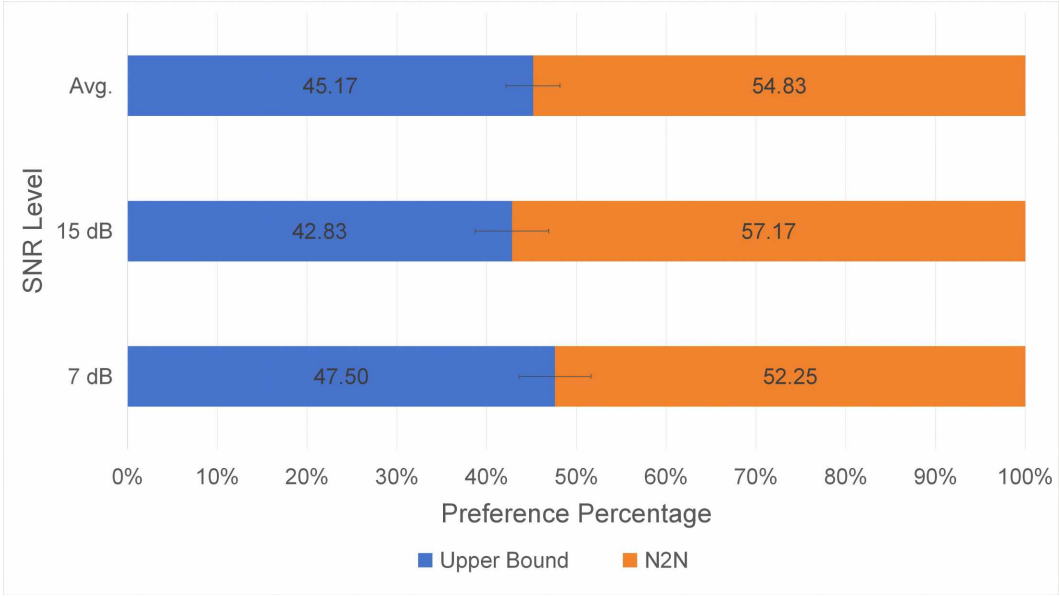


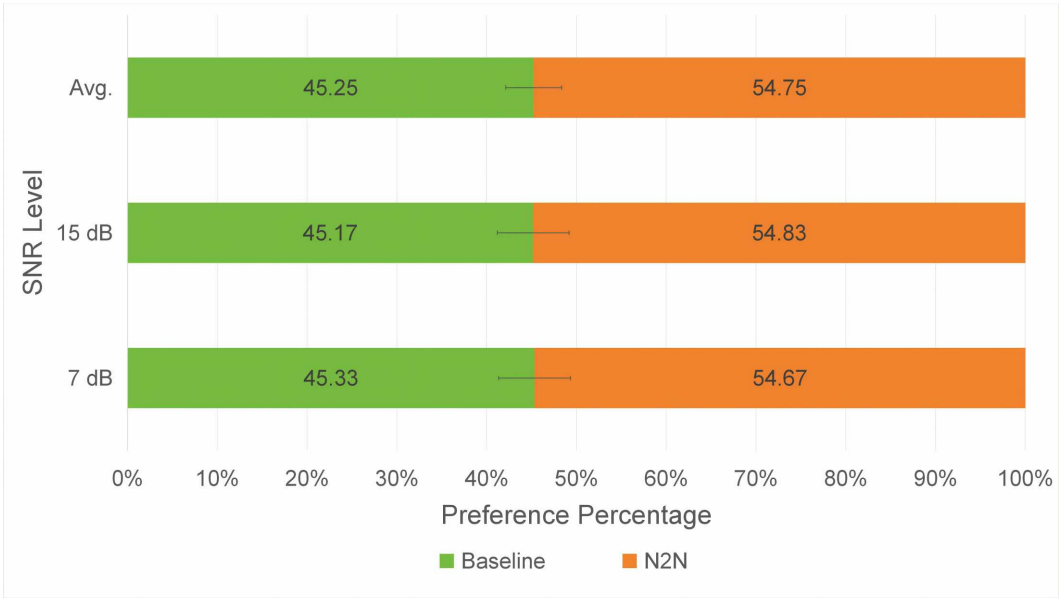
Figure 3.6: *MOS results with 95% confidence intervals for Upper Bound, Baseline, and N2N. Ground Truth refers to the original noisy speech from the target speaker.*

noisy speech from the target speaker, achieves the best average MOS of 4.43. This indicates that the participants are capable of reliably evaluating noisy speech samples. Consistent with objective evaluations, both *N2N* and *Baseline* exhibit improved performance as the SNR increases. Notably, *N2N* consistently outperforms *Baseline*, achieving MOS of 3.29 and 3.46 at 7 dB and 15 dB SNR, respectively, compared to 3.02 and 3.22 for *Baseline*. On average, *N2N* attains MOS of 3,38, significantly surpassing *Baseline* of 3.12 and reducing the gap to the *Upper Bound* of 3.52. These results prove the effectiveness of noise conditioning in enhancing the performance of the proposed N2N framework.

Figure 3.7 illustrates the XAB results for similarity. When comparing *N2N* and *Baseline*, as shown in Figure 3.7 (b), *N2N* achieves an average similarity percentage of 54.75%, over *Baseline* of 45.52%. In Figure 3.7 (a) comparing *N2N* and *Upper Bound*, *N2N* attains a slightly higher preference of 54.83%, compared to 45.17% for *Upper Bound*. These marginal differences suggest that the three methods yield similar



(a)



(b)

Figure 3.7: *Similarity preference scores with 95% confidence intervals. (a) Upper Bound versus N2N. (b) Baseline versus N2N.*

performance in terms of speaker similarity preference.

3.6 Summary of This Chapter

This chapter presented the baseline and the improved noise-conditioned method for the N2N task. The baseline framework adopts the conventional cascading approach in noise-robust VC, where an SE model is followed by a VC model. Experimental results indicated that the SE module introduces substantial distortion to the downstream VC model, severely limiting performance, even when state-of-the-art SE methods were employed. To address this issue, noise conditioning was introduced into the baseline framework to enable direct modeling of noisy speech, thereby mitigating the distortion caused by the SE model. Experimental results demonstrated that noise conditioning significantly enhanced VC performance and narrowed the gap in naturalness between the baseline and the upper bound.

4 N2N Framework in Various Noise Conditions

4.1 Introduction

In Chapter 3, a baseline and an improved noise-conditioned N2N framework are introduced. Initial experimental results have demonstrated the effectiveness of the proposed noise-conditioned method in enhancing VC performance. However, several issues remain unresolved, and certain characteristics of noise conditioning within the VC model require further investigation.

Firstly, the previously utilized clean corpus and noise clips were sampled at 8 kHz, which may limit the reliability of perceptual evaluations. Moreover, there was a lack of suitable metrics and evaluations specifically designed to measure the quality of noise components in the converted speech. Additionally, a significant performance gap remains between the noise-conditioned method and the upper bound, highlighting the need for further improvement.

Moreover, the limited diversity of noise data used in prior experiments makes it unclear whether noise-conditioned VC is suitable for scenarios involving complex or diverse acoustic environments. However, when expanding to more realistic noisy environments, the performance of the noise-conditioned VC method significantly degrades. To address this issue, three data augmentation methods are proposed. Among these, one method has demonstrated efficacy in alleviating performance degradation, as val-

idated by experimental results presented in Section 4.5. The main contents of this chapter can be summarized as follows:

- Two new noise datasets with greater diversity and two noise sampling strategies are introduced to construct the training sets with more diversity of noise conditions. Additionally, the sampling rate of both the corpus and noise source is increased from 8 kHz to 16 kHz.
- The impact of different noisy training conditions on the performance of the N2N method is investigated, and results show that the performance significantly degrades under certain noise scenarios.
- Pre-training of the VC model and noise data augmentation are proposed and applied to improve the performance.
- Additional objective metrics are adopted to provide a more comprehensive evaluation of the N2N framework. Moreover, subjective evaluation is modified to explicitly assess the naturalness of noise components in the converted speech.

4.2 More Diverse Noise Conditions

In the previous experiments, PNL100 was used as the noise source, which consists of 100 noise clips across 20 categories. In this chapter, two new noise datasets: ESC-50 [209] and DEMAND [210] are adopted as noise sources. The ESC-50 is designed for environmental sound classification. It provides a wide range of noise types, comprising 2,000 recordings across 50 categories, with 40 clips per category. The high perceptual clarity of these environmental sounds makes the dataset particularly valuable for this work, which involves both noise generation and the evaluation of noise quality. 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, with channel 01 used for all subcategories in the experiments.

In addition to the noise source, the noise sampling strategies to construct the noisy dataset are also considered. In previous experiments, noisy speech datasets were created by superimposing uniformly sampled noise clips onto clean speech with uniformly sampled SNR levels. As a result, each speaker in the dataset was associated with noise clips from various categories and multiple SNR levels. This strategy is denoted as the speaker-independent (SI) sampling strategy, as the noise conditions in the dataset are unrelated to speaker identity.

However, in realistic scenarios, A speaker is often associated with a single acoustic environment. To better reflect this, a speaker-dependent (SD) sampling strategy is introduced, in which each speaker is assigned a unique noise category and a fixed SNR level, thereby correlating noise conditions with speaker identity. To further investigate the entanglement between speaker identity and noise conditions, a semi-speaker-dependent (SSD) sampling strategy is also employed, where each speaker is assigned a unique noise category but with multiple SNR levels.

Considering the characteristics of ESC-50 and DEMAND, the SI strategy is applied to ESC-50, while the SD and SSD strategies are employed with DEMAND to construct the noisy training sets. For the test set, the SI strategy and ESC-50 are employed to ensure that the evaluations cover a wide range of noise conditions.

In the previous experiments, each utterance was superimposed with various types of background noise at multiple SNRs, which can be considered the SI sampling strategy. Experimental results demonstrated that the noise-conditioned method significantly im-

proved the baseline. When trained on the noisy dataset constructed with the SI strategy, the VC model is exposed to diverse noise conditions across speakers, allowing it to learn the independence between speaker information and noise conditions in a self-supervised manner. Under such conditions, as shown in Equation (3.5), the speaker code **s** and the noise signal **n** are independent.

However, as shown in Section 4.5, when the noise-conditioned framework is trained using DEMAND as the noise source and SD/SSD sampling strategies, its performance degrades markedly, even descending below the baseline. Given the characteristics of the noise source and the SD/SSD sampling strategy, speaker identity is entangled with the noise conditions. As a result, the VC model encounters only limited and fixed combinations of speaker codes **s** and noise signals **n**, leading to speaker-noise entanglement. To address this issue, VC model pre-training and data augmentation techniques are adopted.

4.3 Proposed Method

4.3.1 Pre-training Strategies for VC Model

Pre-training is a frequently used technique in deep learning-based model training. By leveraging large and diverse datasets, pre-training allows models to learn useful and generalizable feature representations, thereby alleviating data scarcity constraints in low-resource settings, such as cases using SD/SSD strategies with limited noise patterns. Furthermore, the pretrained model converges faster during fine-tuning, which can reduce the overall training time and computational cost.

For the baseline VC model, pre-training is conducted using a large-scale clean speech corpus, where the model takes clean speech as input and is optimized using a recon-

struction loss calculated on the same clean speech. As for the noise-conditioned N2N framework, pre-training is performed on a noisy training set constructed from a large noise dataset using the SI sampling strategy. It should be noted that while pre-training enhances the generalization ability of the VC model, it is not a prerequisite for the framework.

4.3.2 Data Augmentation

As the performance degradation is attributed to the entanglement between speaker identity and noise conditions, one of the most straightforward approaches to mitigate this issue is to apply data augmentation to enhance the diversity of the training data. Given that only noisy data from source and target speakers are available for VC training, three augmentation strategies are proposed and denoted as *Data-Aug*, *Noise-Aug I*, and *Noise-Aug II*.

Data-Aug

Figure 4.1 illustrates the workflow of *Data-Aug*. In this process, a portion of the original noisy utterances is randomly sampled and duplicated to be superimposed with the augmented noise at various SNRs to increase the diversity of noise conditions. The SE module then separates the augmented and original noisy utterances into estimated speech signals and the separated background noise, which are used for VC model training.

In general, *Data-Aug* is the most straightforward method for augmenting the training set with noise. However, it may compromise the quality of the training set, as the augmented noisy utterances must also be processed through the SE model. This can

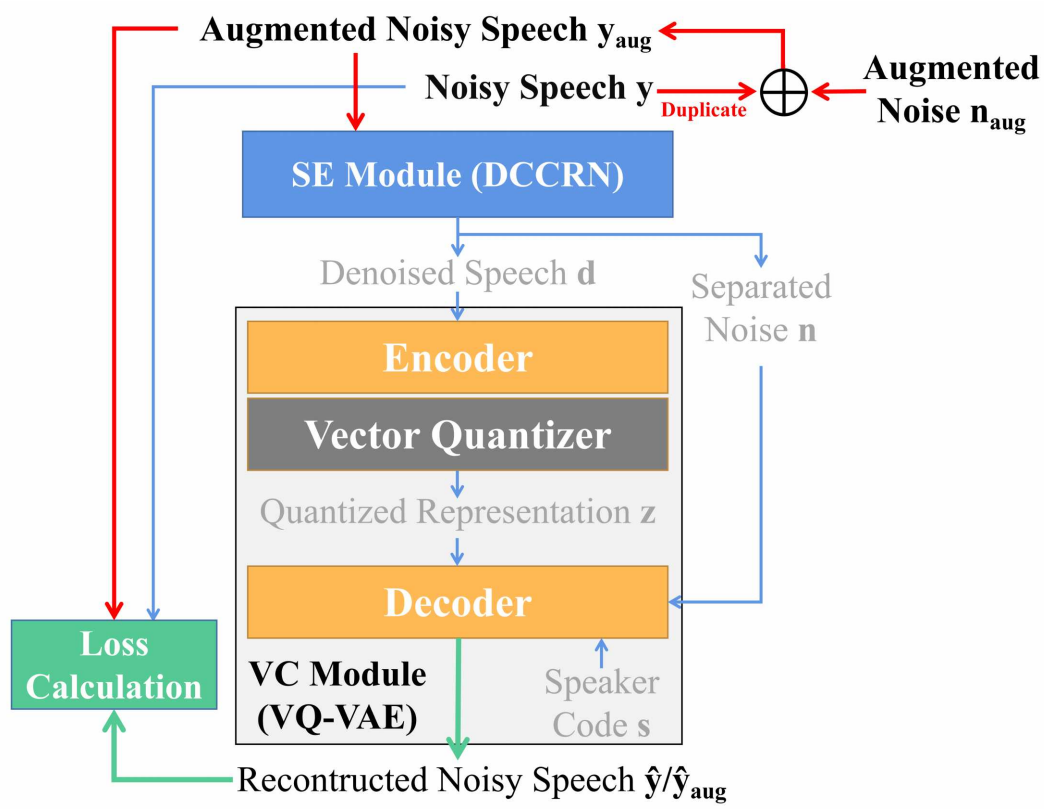


Figure 4.1: *The workflow of Data-Aug on noise-conditioned N2N framework.*

result in further distortion of both the denoised speech and the separated noise.

Noise-Aug I

To overcome the limitations of *Data-Aug*, an improved noise augmentation method *Noise-Aug I* is proposed, as illustrated in Figure 4.2. Unlike *Data-Aug*, where the augmented noise is only superimposed with the noisy speech, *Noise-Aug I* duplicates both the noisy utterances and their corresponding separated noise components, and then superimposes augmented noise clips onto them. In this way, the augmented noisy utterances are not processed by the SE model and are only used as targets for loss calculation during training. As a result, additional distortion introduced by the

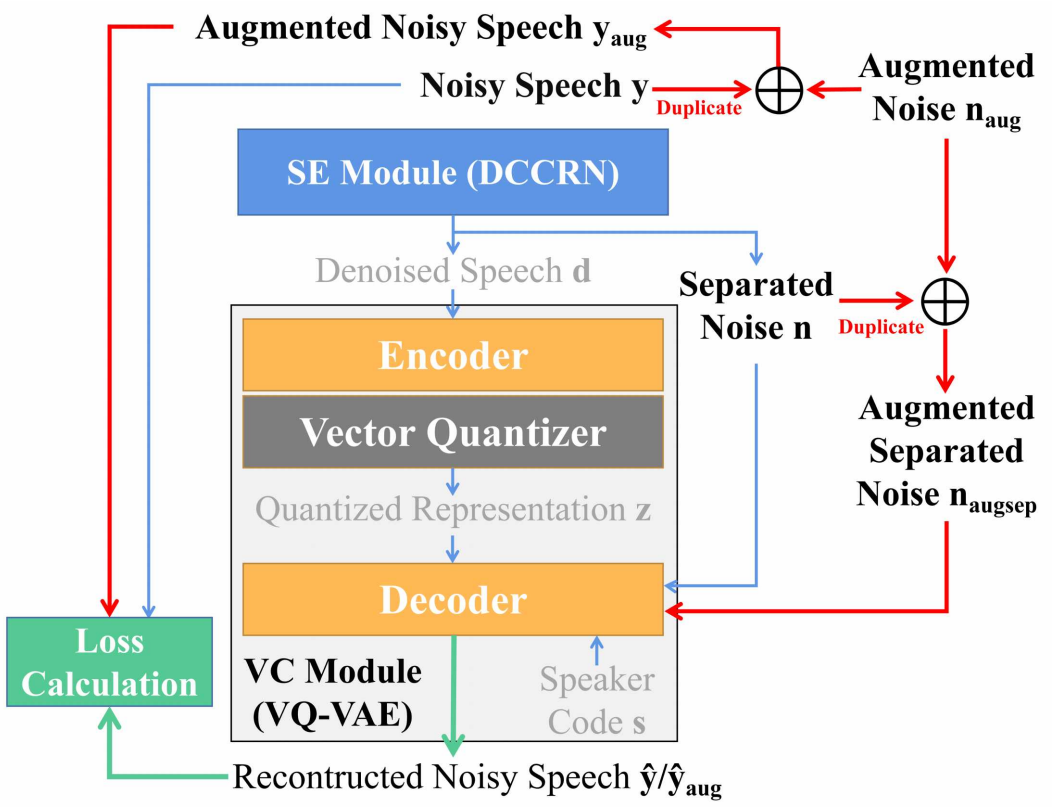


Figure 4.2: *The workflow of Noise-Aug I on noise-conditioned N2N framework.*

augmented noise can be avoided.

Noise-Aug II

For the former two augmentation methods, although additional noise is introduced to increase diversity, the inherent speaker-dependent noise components remain in the augmented noisy speech, which may hinder disentanglement learning. To address this, another augmentation strategy *Noise-Aug II* is proposed, as illustrated in Figure 4.3.

In this approach, a portion of the denoised utterances, rather than the original noisy utterances, is duplicated and superimposed with the augmented noise clips to compose the augmented noisy speeches. During training, the VC model receives either speaker-

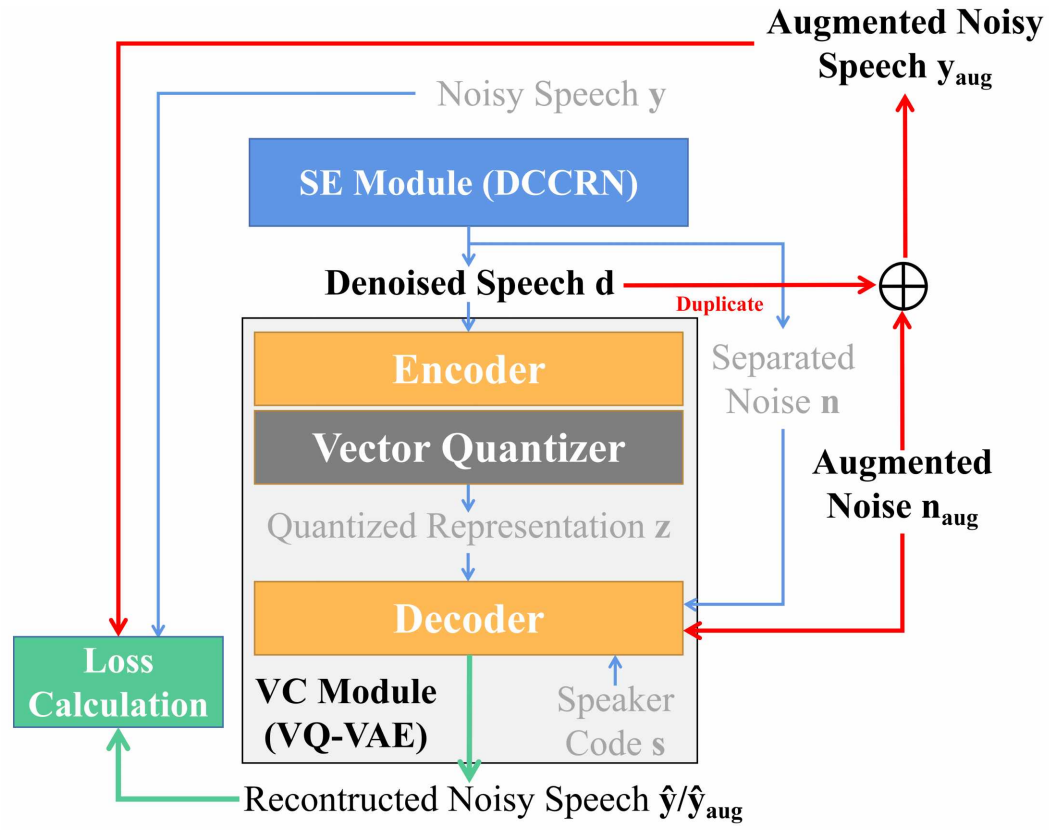


Figure 4.3: *The workflow of Noise-Aug II on noise-conditioned N2N framework.*

dependent separated noise or augmented noise, depending on whether the current training target is the original noisy speech or the augmented noisy speech.

However, since the augmented noisy speech is based on the denoised speech, which is already distorted, *Noise-Aug II* inevitably compromises the benefit of using undistorted noisy data as the training target. In theory, as the volume of augmented training data increases, the performance of the noise-conditioned method will degrade toward the baseline level. The optimal ratio of augmented data samples is empirically investigated in Section 4.5.1.

4.4 Experimental Setup

4.4.1 Dataset for SE Model

The same DNS Challenge 2020 dataset [83] used in the previous experiments is adopted here with 16 kHz sampling rate. From this dataset, 10,000 clean utterances and 8,000 noise clips are sampled to construct the validation set, while the remaining data are used for the training set. The noisy speech is generated by superimposing a noise clip onto a clean utterance at uniformly sampled SNR ranging from 0 to 20 dB.

4.4.2 Dataset for VC Model

In the pre-training stage of the VC model, VCTK [211] is adopted as the clean corpus. All 110 speakers from VCTK are included in the training process, with 20 utterances randomly selected per speaker to create the validation set. For pre-training the noise-conditioned VC model, a noisy version of VCTK is constructed. The noise clips from the SE model validation set are used as the noise source and superimposed onto the VCTK corpus at uniformly sampled SNRs from 0, 5, 10, 15, and 20 dB, following the SI sampling strategy.

For the N2N task, as described in Section 4.2, VCC2018 with the same setups in the previous experiments is used as the clean corpus, while ESC-50 and DEMAND are used as the new noise sources. In total, six noisy training sets are constructed using different combinations of the noise sources and noise sampling strategies.

ESC-50 Using SI Sampling Strategy

Before constructing the noisy dataset, ESC-50 noise clips are trimmed using voice activity detection (VAD) via WebRTC VAD1[1]. Among the 50 noise categories, nine are assigned to the VCC test set, and the remained categories are adopted for the training set. The noisy training set is created by uniformly sampling the noise clips and superimposing them onto the clean utterances at uniformly sampled SNRs from 0, 5, 10, 15, and 20 dB.

The noisy VCC test set is constructed as a parallel dataset, where the same utterances across speakers are superimposed with the same noise clip. Multiple noisy test sets are created in parallel, with each test set at a single SNR level of -5, 0, 5, 10, 15, 20, and 25 dB. All methods in this chapter are evaluated on these noisy test sets, regardless of whether the noisy training dataset adopts the SI sampling strategy.

DEMAND Using SD/SSD Sampling Strategy

When using DEMAND as the noise source, SD and SSD sampling strategies are applied to construct the noisy training set. Specifically, a noise subcategory from DEMAND is randomly assigned to each speaker in the VCC training set, so that each speaker is associated with a single noise class. For the SD strategy, each utterance is superimposed with a noise clip sampled from the assigned subcategory at a constant SNR of 5 dB. For the SSD strategy, the process is identical except for the SNR levels: instead of a constant 5 dB, SNR levels are uniformly sampled from 0, 5, 10, 15, and 20 dB. As a result, the SSD strategy ensures that speaker identity is associated with the noise category but not with the SNR level, while under SD conditions, speaker identity is associated with both the noise category and SNR level. Note that both the SD and

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

Table 4.1: *Summary of the evaluated methods.*

Method	VC Model	Input Type	Noise Source	Noise Sampling Stratrgy
Upper Bound	VQ-VAE	Clean speech	-	-
	Noise-Conditioned VQ-VAE	Clean speech; Original noise	ESC-50	SI
Baseline-SI	VQ-VAE	Denoised speech	ESC-50	SI
Baseline-SD	VQ-VAE	Denoised speech	DEMAND	SD
N2N-SI	Noise-Conditioned VQ-VAE	Denoised speech; Separated noise	ESC-50	SI
N2N-SSD	Noise-Conditioned VQ-VAE	Denoised speech; Separated noise	DEMAND	SD
N2N-SD	Noise-Conditioned VQ-VAE	Denoised speech; Separated noise	DEMAND	SD
N2N-Data-Aug	Noise-Conditioned VQ-VAE	Denoised speech; Separated noise	DEMAND + ESC-50 (Augmented)	SD
N2N-Noise-Aug I	Noise-Conditioned VQ-VAE	Denoised speech; Separated noise	DEMAND + ESC-50 (Augmented)	SD
N2N-Noise-Aug II	Noise-Conditioned VQ-VAE	Denoised speech; Separated noise	DEMAND + ESC-50 (Augmented)	SD

SSD noise conditions are used exclusively for constructing the training set.

The proposed noise augmentation methods are applied to the noisy training set constructed using DEMAND as the noise source and the SD sampling strategy. To control variable factors during augmentation, the same noise clips from ESC-50 under the SI sampling strategy are randomly sampled across all proposed noise augmentation methods.

4.4.3 Methods under Evaluation

Table 4.1 summarizes the methods evaluated in this chapter. The name of each method generally follows the format "TypeOfModel-TypeOfTrainingSet", except for *Upper Bound*.

The "TypeOfModel" is defined according to the naming scheme used in prior experiments and includes three types: *Upper Bound*, *Baseline*, and *N2N*. *Baseline* denotes the baseline N2N framework where the original VC model is applied. *N2N* refers to the improved framework, in which noise conditioning is introduced into the VC model to directly model noisy speech. As for *Upper Bound*, to investigate whether noise con-

ditioning affects the VC performance, the original VC model and the noise-conditioned version are compared. Both the original VC model and the noise-conditioned version are trained and evaluated on clean speech, with the latter additionally conditioned on the original noise clips.

The term “TypeOfTrainingSet” refers to the specific configuration of the noisy training set. Since the sampling strategy is associated with the noise source, the suffixes *SI*, *SD*, and *SSD* are used to distinguish different training sets. *SI* refers to the noisy training set constructed using the ESC-50 as the noise source with the SI sampling strategy, while *SD* and *SSD* refer to training sets using DEMAND as the noise source with corresponding SD and SSD sampling strategies. Three proposed noise augmentation methods based on *SD* are represented as *Data-Aug*, *Noise-Aug I*, and *Noise-Aug II*.

In total, nine methods are evaluated. *Upper Bound* represents the theoretical best performance of the N2N framework, which is trained and tested on the original dataset. Two *Upper Bound* methods are evaluated: one using the original VC model and the other using the noise-conditioned VC model. This comparison aims to assess the impact of noise conditioning on VC performance while avoiding potential interference introduced by the SE model. Two *Baseline* methods trained on different training sets are denoted as *Baseline-SI* and *Baseline-SD*. To evaluate the performance of *N2N* under varying noise conditions, six methods using different noisy training sets are tested and denoted as *N2N-SI*, *N2N-SSD*, *N2N-SD*, *N2N-Data-Aug*, *N2N-Noise-Aug I*, and *N2N-Noise-Aug II*, respectively.

4.4.4 Evaluation Metrics

Objective Evaluation Metrics

In the previous experiments in Section 3.5.1, the impact of the SE models on the VC downstream has been evaluated. Experimental results reveal that the objective metrics for the SE tasks can not well reveal their impact on the VC downstream. As a result, this chapter focuses more on evaluating VC performance. Except for MCD measuring the overall quality of the converted speech, WER was used to measure the quality of linguistic content, calculated using a publicly available ASR model[2]. An open-source speaker verification method[3] is also adopted to assess the similarity between the converted sample and its target reference. Since these objective metrics are exclusively applicable to clean speech data, all the methods generate clean converted speech. The quality assessment of the generated noise component and the noisy speech is left to the subsequent subjective evaluation.

Subjective Evaluation Metrics

The subjective evaluation setup follows that of previous experiments, employing the non-parallel (SPOKE) task of VCC2018. The non-parallel task includes four source speakers (VCC2SF3, VCC2SF4, VCC2SM3, and VCC2SM4) and two target speakers (VCC2TF2, VCC2TM2), resulting in eight conversion pairs in total. All subjective evaluations are conducted via Amazon Mechanical Turk (MTurk) with 15 participants.

As the first step, the effectiveness of the pre-training strategy is assessed via a preference test. Two noisy testing sets with SNRs of 5 and 15 dB are prepared, where 32 converted utterances per test set are parallelly sampled. Participants are asked to

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

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

compare the naturalness and similarity of two samples, with one generated by a VC model with pre-training and the other without. An original target utterance is provided as a reference for assessing similarity. For consistency with objective evaluations, all utterances in this test are clean.

The proposed noise augmentation strategies are assessed through an extended MOS test and a SIM test (see Section 3.4.4), both conducted on two test sets at 5 dB and 15 dB SNR. The extended MOS test includes two categories: clean MOS and noisy MOS. Six methods are evaluated: *Upper Bound*, *Baseline-SI*, *Baseline-SD*, *N2N-SI*, *N2N-SD*, and *N2N-Noise-Aug II*. Although *N2N-SSD* is also examined via objective evaluation, it is excluded from subjective evaluation, as it mainly serves to analyze the impact of SNR variation on the disentanglement between speaker identity and noise conditions.

The clean MOS test follows the standard procedure used in conventional MOS evaluations. All methods generate clean converted speech, the naturalness of which is scored by the participants on a 5-point scale (1 = Bad, 5 = Excellent). Each method provides 64 utterances: 32 at SNR 5 dB and 32 at SNR 15 dB.

In the noisy MOS test, participants rate the naturalness of both the speech and noise components in a noisy converted utterance, also on a 5-point scale. To facilitate the evaluation of the background noise, the original noise clip is provided as a reference. Similar to the clean MOS test, 64 noisy converted utterances are randomly sampled for each method, equally split between the two SNR levels.

The SIM test follows the previous procedure described in Section 3.4.4, where participants rate the similarity between a converted utterance and a target utterance on a 4-point scale. Scores of 1 or 2 are counted toward the similarity percentage. To minimize noise-related perceptual interference, the SIM test uses clean converted speech, following the same sampling procedure as the clean MOS test (64 utterances

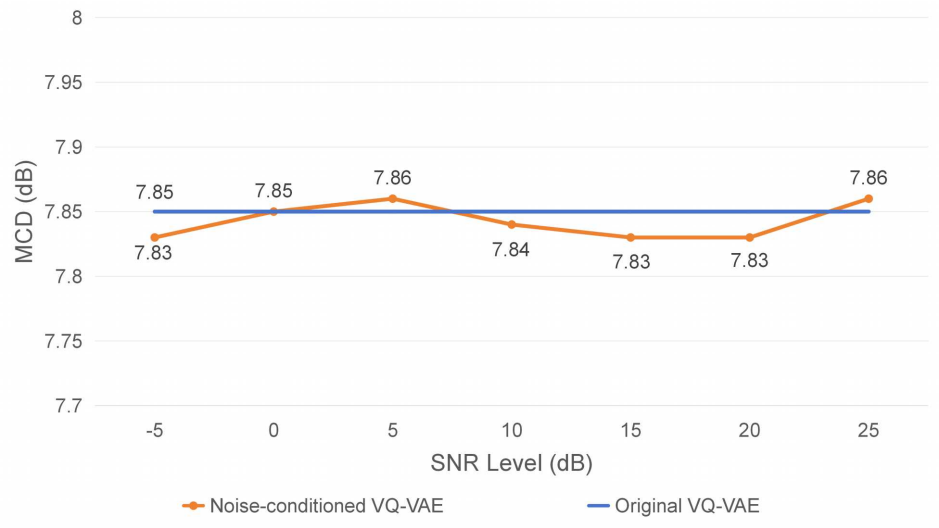


Figure 4.4: *MCD results for noise-conditioned and original VQ-VAE across test sets, each evaluated at a fixed SNR.*

per method). Based on MOS test results, five methods are selected for the SIM test: *Upper Bound*, *Baseline-SD*, *N2N-SI*, *N2N-SD*, and *N2N-Noise-Aug II*.

4.5 Experimental Results

4.5.1 Objective Evaluation Results

Comparison of Noise-Conditioned VQ-VAE and Standard VQ-VAE

Theoretically, if speaker identity and noise conditions are independent from each other, the noise-conditioned VQ-VAE operates in the same way as the original model when generating clean converted speech, as indicated by Equation (3.5).

Figure 4.4 presents the MCD scores across various SNR levels for the noise-conditioned and original VQ-VAE models. Both models are pre-trained following the process mentioned in Section 4.3.1. Since the original VQ-VAE is not conditioned on noise, it is

Table 4.2: *Objective evaluation results of the methods with/without pre-training on two test sets with fixed SNRs of 5 and 15 dB.*

Methods	Status	MCD (dB) ↓		WER (%) ↓		Similarity ↑	
		5 dB	15 dB	5 dB	15 dB	5 dB	15 dB
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

trained and evaluated solely on clean data, so that its MCD remains stable at 7.85. The noise-conditioned VQ-VAE achieves comparable MCDs around 7.85 across all SNRs, with an average score of 7.84. These results confirm that incorporating noise as a conditioning factor into the VC model does not degrade the quality of clean converted speech when the speaker identity and the noise conditions are disentangled. Consequently, the noise-conditioned VC model is adopted in the *Upper Bound* for subsequent experiments.

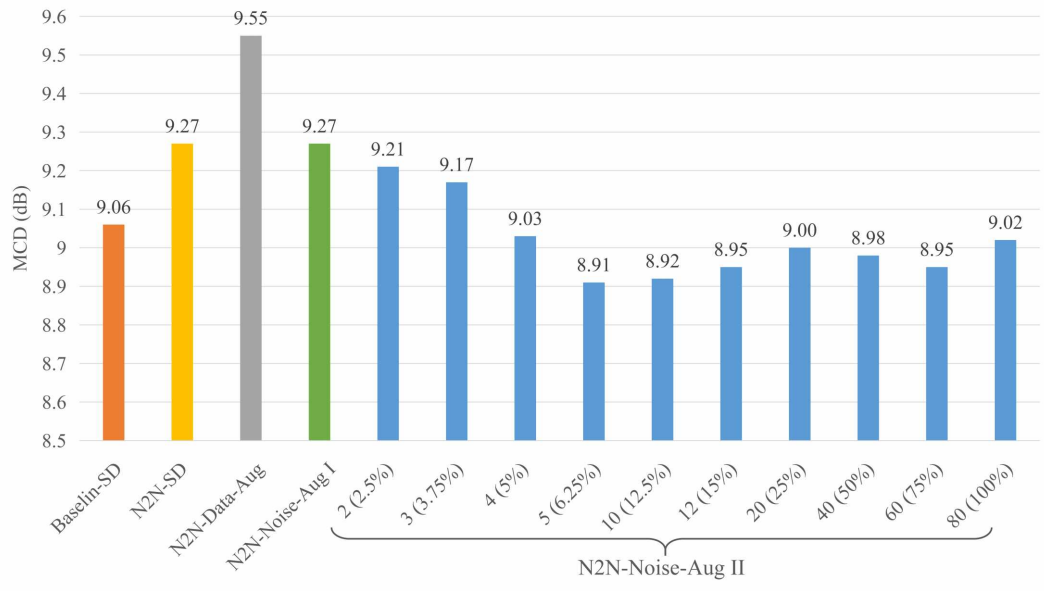


Figure 4.5: *MCD results for Baseline-SD, N2N-SD, and N2N-SD with data augmentation strategies on the noisy VCC2018 test set.*

Pre-training Strategies

Table 4.2 presents the objective evaluation results of the methods with and without pre-training, assessed by MCD, WER, and similarity score on separate noisy test sets at SNR 5 dB and 15 dB. Overall, pre-training consistently improves the performance of all methods. The improvement is most pronounced in WER, while MCD and similarity scores also show moderate gains. Among all methods, *N2N-SD* exhibits noticeable improvements across all metrics, demonstrating the effectiveness of pre-training in enhancing model robustness. Based on these findings, all methods adopt the pre-trained VC model in subsequent experiments.

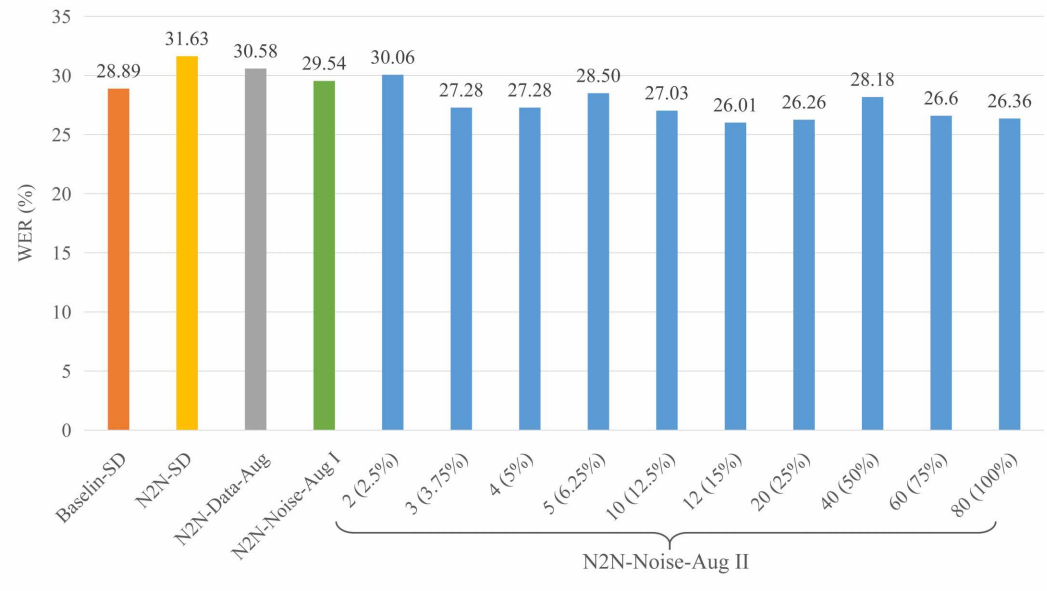


Figure 4.6: *WER results for Baseline-SD, N2N-SD, and N2N-SD with data augmentation strategies on the noisy VCC2018 test set.*

Data Augmentation Strategies

Figure 4.5, 4.6, and 4.7 present the objective evaluation results of the three augmentation strategies on the noisy VCC2018 test set, constructed using the SI sampling strategy at an SNR of 5 dB. As discussed in Section 4.3.2, *N2N-Noise-Aug II* generates augmented data based on denoised speech, thereby compromising the advantage of using undistorted noisy data as training targets. To investigate the impact of the augmentation volume, multiple variants of *N2N-Noise-Aug II* are evaluated. The number of augmented utterances per speaker (out of 80) and their corresponding percentages are shown on the x-axis in Figure 4.5, 4.6, and 4.7.

Among all methods, *Baseline-SD* achieves 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*. These results support the concerns raised in Section 4.3.2 that the speaker-dependent

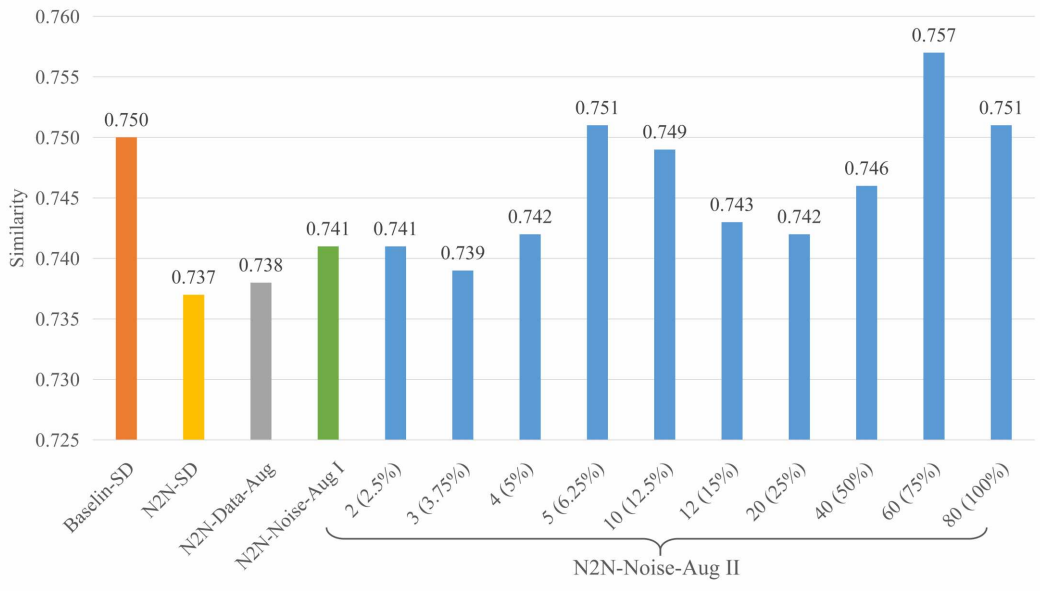


Figure 4.7: *SIM scores for Baseline-SD, N2N-SD, and N2N-SD with data augmentation strategies on the noisy VCC2018 test set.*

noise component persists in the augmented data and hinders disentanglement learning. *N2N-Data-Aug* achieves an MCD of 9.55, a WER of 30.58%, and a similarity score of 0.738, which are worse than *N2N-Noise-Aug I*. This is due to the fact that *N2N-Data-Aug* introduces additional distortions to degrade the VC performance further, as explained in Section 4.3.2. *N2N-Noise-Aug I* reaches an MCD of 9.27, a WER of 29.54%, and a similarity score of 0.741, offering limited improvement compared to *N2N-SD*, which scores an MCD of 9.27, a WER of 31.63%, and a similarity score of 0.737.

In contrast, most variants of *N2N-Noise-Aug II* outperform *Baseline-SD* in both MCD and WER, while maintaining comparable similarity scores. Specifically, the variant *5 (6.25%)* achieves the best MCD of 8.91, followed by *10 (12.5%)* and *60 (75%)*, with scores of 8.92 and 8.95, respectively. For similarity, *60 (75%)* ranks highest at 0.757. *5 (6.25%)* and *80 (100%)* achieve second place with a score of 0.751, while *10*

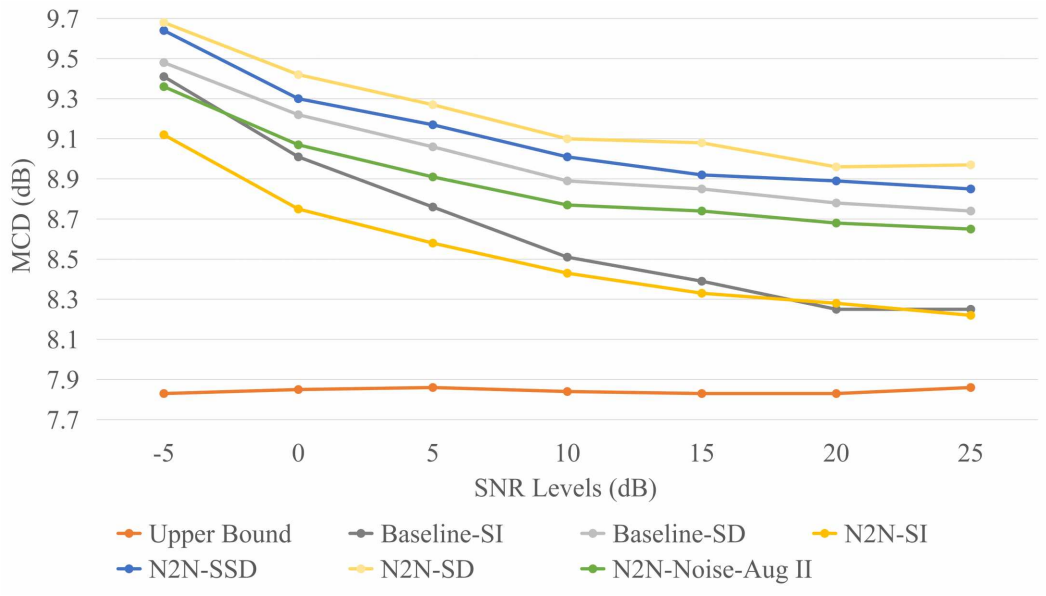


Figure 4.8: *MCD results for the methods under varying SNR conditions on the noisy VCC2018 test set.*

(12.5%) ranks third with 0.749. Regarding WER, *10 (12.5%)* and *60 (75%)* perform best, achieving 27.05% and 26.60%, respectively, while *5 (6.25%)* obtains 28.50%. Given the balance between performance gains and minimal augmentation, the variant *5 (6.25%)* is selected as the final configuration of *N2N-Noise-Aug II*.

The objective evaluation results of different methods under varying SNR conditions are summarized in Figure 4.8, 4.9, and 4.10. In general, *Upper Bound* achieves the best performance across all metrics and SNR levels. A notable gap exists between *Baseline-SI* and *Upper Bound*, which further widens at lower SNRs. *N2N-SI* significantly improves upon *Baseline-SI*, which is consistent with previous findings demonstrated in Section 3.5.2, highlighting the effectiveness of introducing noise conditioning into N2N task. However, *Baseline-SI*outperforms *N2N-SD* across all metrics, indicating performance degradation when speaker identity and noise conditions are entangled. Additionally, *N2N-SSD* consistently surpasses *N2N-SD* in all metrics. These results

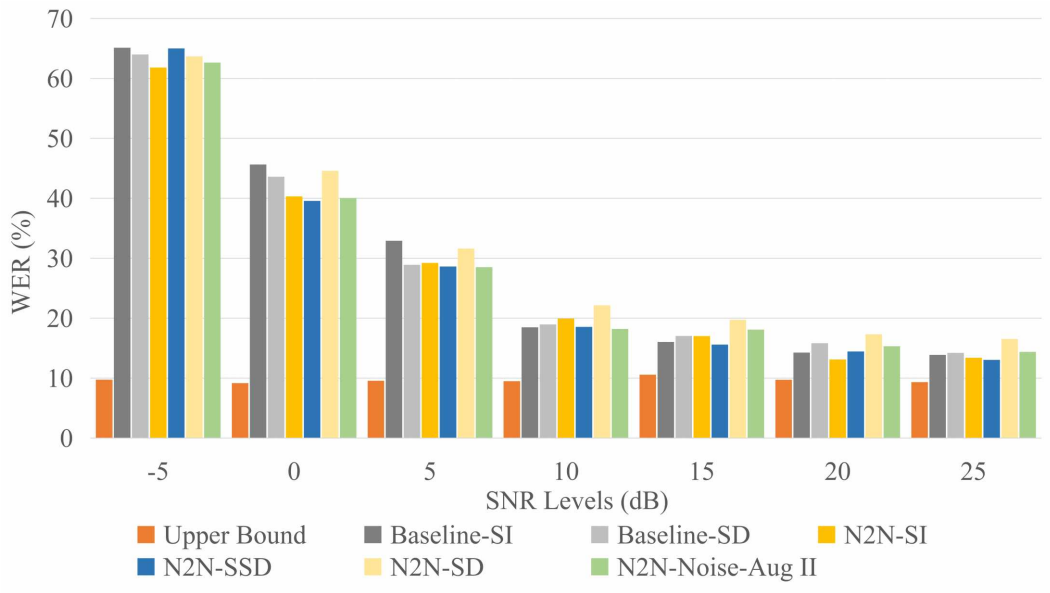


Figure 4.9: *WER results for the methods under varying SNR conditions on the noisy VCC2018 test set.*

highlight that not only the diversity of noise categories, but also variation in SNRs, contributes to disentanglement within the N2N framework.

For the proposed noise augmentation methods, *N2N-Noise-Aug II* improves upon *N2N-SD* and yields better results than *Baseline-SD* in terms of MCD and similarity across all SNRs. However, *Baseline-SD* still achieves a lower WER by 1.07% at 15 dB. Despite these improvements, the performance gap remains larger compared to that between *N2N-SI* and *Baseline-SI*, indicating room for further enhancement.

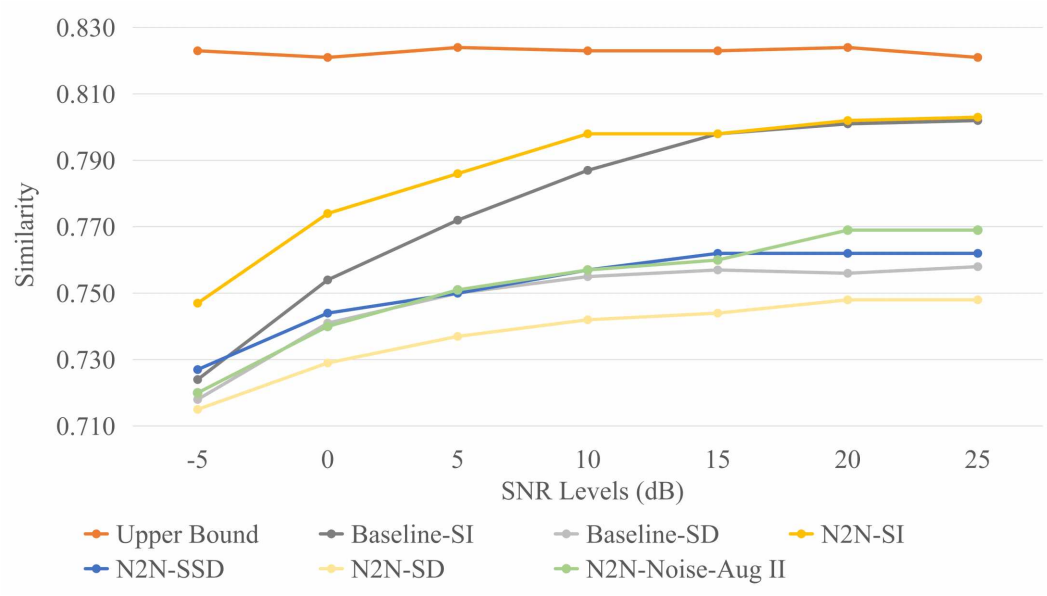
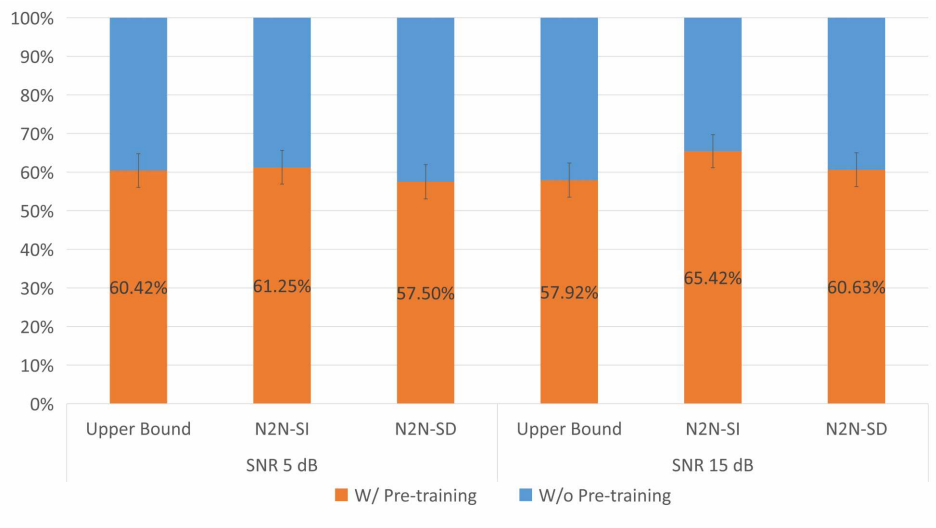


Figure 4.10: *SIM scores for the methods under varying SNR conditions on the noisy VCC2018 test set.*

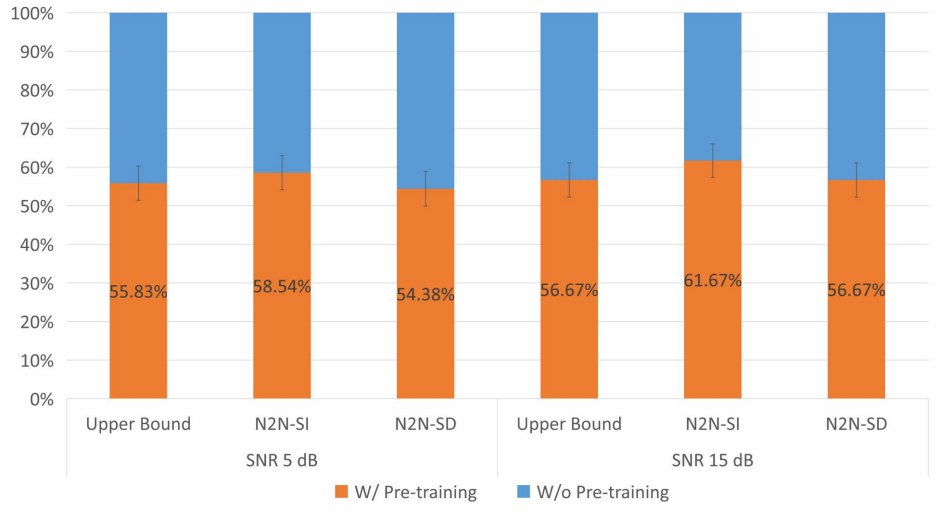
4.5.2 Subjective Evaluation Results

Pre-training Strategies

Figure 4.11 shows the preference evaluation results comparing *Upper Bound*, *N2N-SI*, and *N2N-SD*, each with and without pre-training. In general, pre-training consistently leads to improvements in both naturalness and similarity, with observed gains ranging from 4.38% to 15.42%. Notably, even *Upper Bound*, which is trained using clean speech and original noise clips without SE-induced distortion, benefits from pre-training, highlighting its broader effectiveness beyond scenarios involving denoised inputs. These findings reinforce the earlier conclusion that pre-training enhances the N2N framework by improving the naturalness and similarity of clean converted speech.



(a)



(b)

Figure 4.11: *Preference evaluation results of the N2N frameworks with/without pre-training for clean converted samples. Error bars show 95% confidence intervals. (a) Naturalness. (b) Similarity.*

Data Augmentation Strategies

Figure 4.12 illustrates the MOS results for methods generating clean converted speech. Since both *Upper Bound* and *Ground Truth* (referred to as the original clean

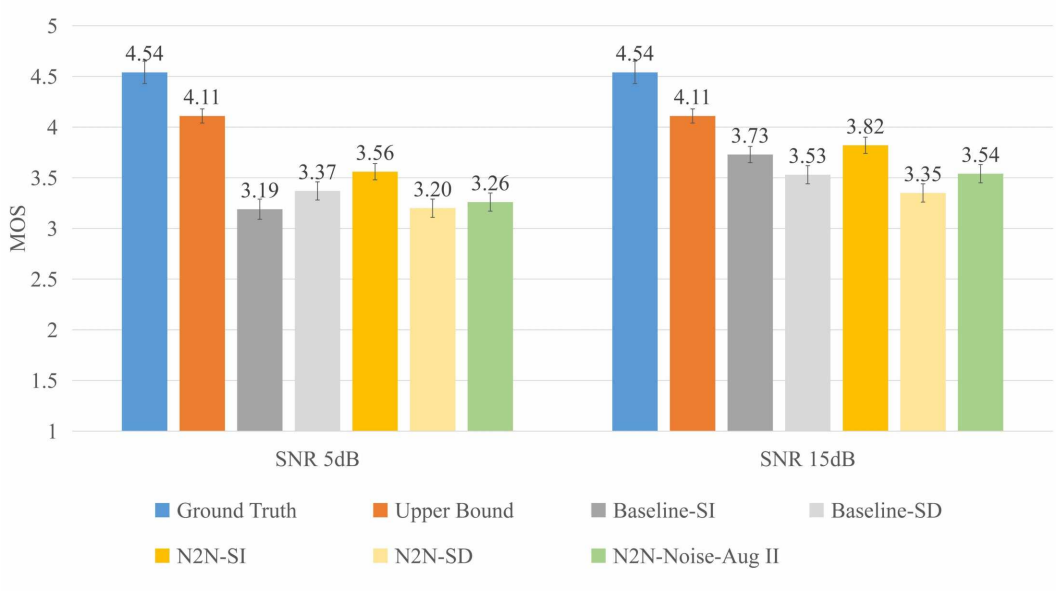
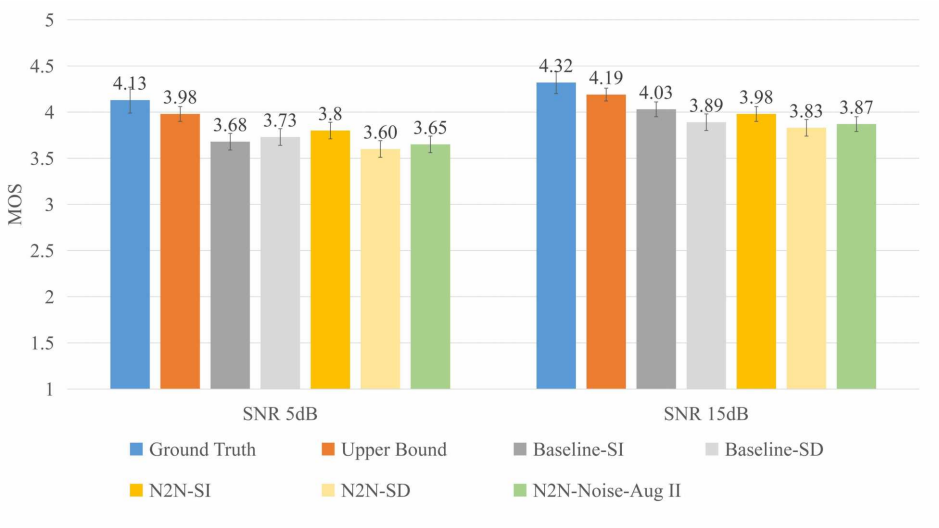


Figure 4.12: *MOS evaluation results for methods producing clean converted speech, with 95% confidence intervals.*

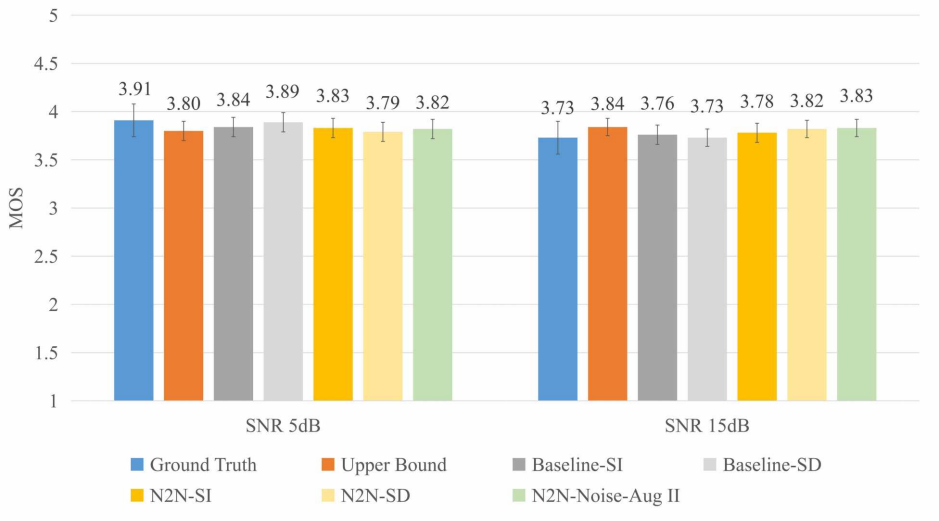
speech from the target speaker) are independent of noise, their MOS scores remain constant under SNRs of 5 and 15 dB. In contrast, the noise-dependent methods show higher MOS at 15 dB, as expected due to reduced noise interference.

Excluding *Ground Truth*, *Upper Bound* achieves the highest MOS of 3.98. *N2N-SI* obtains MOS of 3.56 and 3.82 at SNRs of 5 and 15 dB, respectively, outperforming *Baseline-SI* achieving 3.19 and 3.73. These results align with the objective evaluation, demonstrating the benefit of noise conditioning in the *N2N* framework.

However, consistent with the objective evaluation results, VC performance of the N2N framework deteriorates when using DEMAND with the SD sampling strategy: *N2N-SD* yields MOS of only 3.20 and 3.35 at SNRs of 5 and 15 dB, which are outperformed by *Baseline-SD* with 3.37 and 3.53. The proposed *N2N-Noise-Aug II* helps mitigate this degradation, raising the scores to 3.26 and 3.54, with improvements of 0.06 and 0.19 over *N2N-SD*. However, its performance at an SNR of 5 dB remains



(a)



(b)

Figure 4.13: *MOS evaluation results for methods producing noisy converted speech, with 95% confidence intervals: (a) MOS of the speech component. (b) MOS of the noise component.*

slightly below *Baseline-SD*, with a gap of 0.11.

Figure 4.13 (a) illustrates the MOS results for the speech component of the noisy

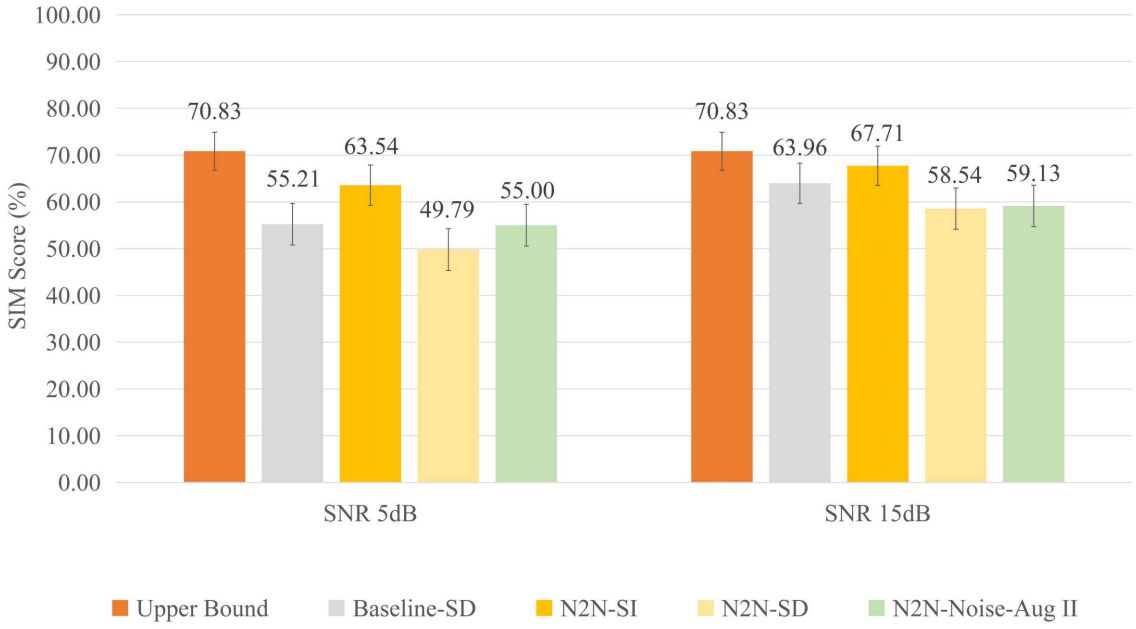


Figure 4.14: *SIM scores for methods producing clean converted speech, with 95% confidence intervals.*

converted samples. In general, similar trends to those for MOS for clean converted speech in Figure 4.12 can be observed. *Upper Bound* achieves the highest MOS of 3.98 and 4.19 at SNRs of 5 and 15 dB. *N2N-SI* ranks second, with MOS of 3.8 and 3.98, outperforming *Baseline-SI* by 0.12 at 5 dB, while *Baseline-SI* achieves a slightly higher MOS of 4.03 at 15 dB. *N2N-SD* obtains the lowest MOS of 3.6 and 3.83 at SNRs of 5 and 15 dB, falling behind *Baseline-SD* with 3.73 and 3.98. Although the proposed *N2N-Noise-Aug II* does not achieve the same level of improvement as observed in the objective evaluation in Figure 4.5, it slightly improves over *N2N-SD* by 0.05 and 0.04, respectively, but still lags behind *Baseline-SD* by 0.08 and 0.02.

Figure 4.13 (b) shows the MOS for the noise component of the noisy converted samples. Overall, all methods yield similar scores of approximately 3.8 at both SNR

levels, slightly lower than the *Ground Truth* score of 3.91 at 5 dB but higher than its 3.73 at 15 dB. Note that *Upper Bound*, *N2N-SI*, *N2N-SD*, and *N2N-Noise-Aug II* generate the noise components via the neural networks, whereas the remaining methods use superimposed noise clips. These results demonstrate the effectiveness and reliability of the N2N framework in generating realistic noise components.

Figure 4.14 presents the SIM scores for methods producing clean converted speech. Overall, the trends are similar to the MOS for clean converted samples illustrated in Figure 4.12. *Upper Bound* achieves the highest SIM score of 70.83%. *N2N-SI* obtains SIM scores of 63.54% and 67.71% at SNRs of 5 and 15 dB, respectively, outperforming *N2N-SD* with scores of 49.79% and 58.54%. *N2N-SD* is outperformed by *Baseline-SD*, which achieves 55.21% and 63.96%. The proposed *N2N-Noise-Aug II* improves upon *N2N-SD* by 5.21% at 5 dB, while the improvement is less significant at 15 dB by only 0.59%. Nevertheless, *Baseline-SD* yields better results than *N2N-Noise-Aug II* at both SNRs.

In general, the experimental results demonstrate that the N2N framework significantly outperforms the baseline when speaker identity and noise conditions are properly disentangled. Incorporating the pre-training strategy can further improve the performance of the N2N framework. However, when using DEMAND and SD sampling strategy, which leads to entanglement between speaker identity and noise conditions, the VC performance of the N2N framework degrades substantially in terms of naturalness and similarity. Although the proposed noise augmentation improves the performance of N2N under such entangled conditions, it does not yield clear advantages over the baseline, indicating room for further enhancement. Regarding the noise components in the converted noisy speech, the N2N framework achieves comparable scores to the ground truth (original noisy speech from the target speaker), demonstrating its ability

to generate high-quality noise components.

4.6 Summary of This Chapter

In this chapter, the N2N VC framework was evaluated under a broader range of noise conditions. The results demonstrated that the framework was effective in generating high-quality noise components. However, when the speaker identity and noise conditions were entangled, the VC performance of the N2N framework degrades significantly. To address this issue, a pre-training strategy was introduced and shown to improve the model ' s performance. Furthermore, three noise augmentation strategies were proposed to mitigate the performance drop. Among them, *N2N-Noise-Aug II* enabled the N2N framework to outperform the baseline in objective evaluations. However, subjective evaluation results further indicated that there was still room for the entanglement between speaker identity and noise conditions.

5 Analysis of N2N VC Performance Degradation

5.1 Introduction

In the previous chapter, the VC performance of the N2N framework under various noise conditions was examined. Two new datasets, ESC-50 and DEMAND, were introduced as the noise sources. Compared to PNL100 used in Chapter 3, ESC-50 offers a broader diversity of noise categories and a larger number of noise clips. In contrast, the DEMAND dataset contains only 18 categories but effectively represents common real-world noise environments. Based on the characteristics of ESC-50 and DEMAND, two primary noise sampling strategies, denoted as speaker-independent (SI) and speaker-dependent (SD), were also employed to construct noisy training sets with specific noisy conditions.

Experimental results from the previous chapter revealed that using DEMAND with the SD sampling strategy leads to notable VC performance degradation due to entanglement between speaker identity and noise conditions. Although pre-training and noise augmentation strategies were proposed to mitigate this issue, the improved N2N framework still failed to outperform the baseline, suggesting the presence of additional factors contributing to the degradation.

In this chapter, the phenomenon of the VC performance degradation under specific noise conditions is further investigated. To identify the cause of this degradation, a

series of experiments is conducted on a per-category basis to evaluate the individual impact of each noise type on VC performance. The analysis is grounded in objective evaluation results, based on which the causes of degradation are summarized.

5.2 Experimental Setup

In this chapter, the experimental setup largely follows that of the previous experiments. The VCC2018 corpus is used as the clean speech source. ESC-50 and DEMAND are also adopted as the noise sources. Here, *E* and *D* are used as shorthand for ESC-50 and DEMAND, respectively. The previously defined SI and SD noise sampling strategies are also employed to construct the noisy training set. The same noisy test set used in the previous chapter is also retained for consistency.

To identify the factors contributing to performance degradation, experiments are conducted to compare the performance of the *Baseline* and *N2N* methods under various noise conditions. The *Baseline* refers to the original N2N framework, which cascades an SE model and a VC model, whereas the *N2N* method introduces noise conditioning into the *Baseline* to directly model noisy speech. *Upper Bound* with the same configuration as the previous one is included as a reference to indicate the theoretically best performance of *N2N*. This chapter particularly focuses on the quality of the speech component and how it is affected by different noise conditions in the noisy training set. Accordingly, all methods are evaluated using clean converted samples.

To reduce the evaluation cost in terms of time and resources, only objective metrics are considered. The MCD metric is primarily used to assess the overall quality of the converted samples.

5.3 Investigating the Causes of VC Performance Degradation

5.3.1 VC Performance Degradation

In the previous experiments, SI and SD noise sampling strategies were applied to ESC-50 and DEMAND, respectively, based on the characteristics of each dataset. In the SI strategy, for each utterance in VCC2018, a noise clip and an SNR level between 0 and 20 dB are uniformly sampled to generate a noisy speech, thereby keeping the speaker identity and the noise conditions uncorrelated. In contrast, the SD strategy first assigns a noise category to each speaker in VCC2018 to associate each speaker's identity with a specific noise type. Then, each utterance from a given speaker is mixed with a randomly sampled noise clip from the assigned noise category at a fixed SNR of 5 dB. The abbreviations E-SI and D-SD are used to denote the resulting noisy training sets based on ESC-50 with SI sampling and DEMAND with SD sampling, respectively.

Figure 5.1 presents the MCD results for models trained on E-SI and D-SD. The *Upper Bound* is trained and evaluated on the original clean speech data, while *Baseline* and *N2N* are based on the denoised data. When trained on the E-SI training set, *N2N-SI* outperforms the *Baseline-SI*. In contrast, when using the D-SD training set, noise conditioning in *N2N* fails to provide performance gains. Specifically, the *Baseline-SD* achieves an MCD of 9.09, whereas noise-conditioned *N2N-SD* achieves a subpar MCD of 9.49. In Figure 5.1, the red frame highlights this degradation, where *N2N-SD* fails to outperform the *Baseline-SD*.

In the previous chapter, the observed performance degradation was primarily attributed to speaker-noise entanglement, which was concluded based on the characteristics of the noise dataset and the noise sampling strategy. The DEMAND dataset

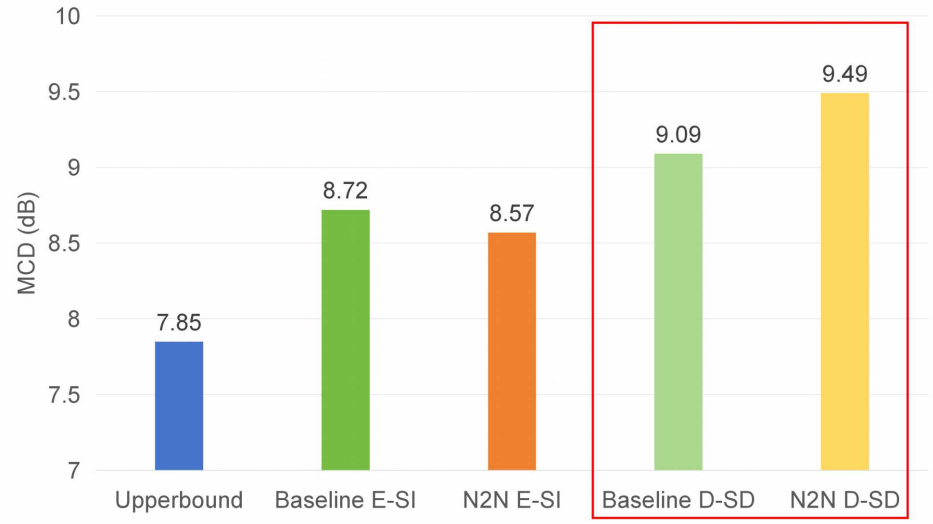


Figure 5.1: *MCD results for methods trained on the E-SI and D-SD datasets. The red frame highlights the performance degradation observed when the N2N framework is trained on the D-SD training set.*

exhibits limited noise diversity, containing only 18 noise subcategories, each with a single noise recording. Furthermore, the SD sampling strategy assigns each noise category to a specific speaker, which further amplifies this lack of variability. These combined factors contribute to the entanglement between speaker identity and noise conditions. To address this, pre-training and noise augmentation strategies were proposed. However, the improved *N2N* still remains inferior to the *Baseline*. These results indicate that, except for speaker-noise entanglement, additional factors beyond speaker–noise entanglement also contribute to the performance degradation.

5.3.2 Effects of Noise Sampling Strategies on VC Performance

Sampling strategy plays a crucial role in establishing speaker-noise entanglement. In the previous experiment, the SI and SD strategies were exclusively applied to ESC-50

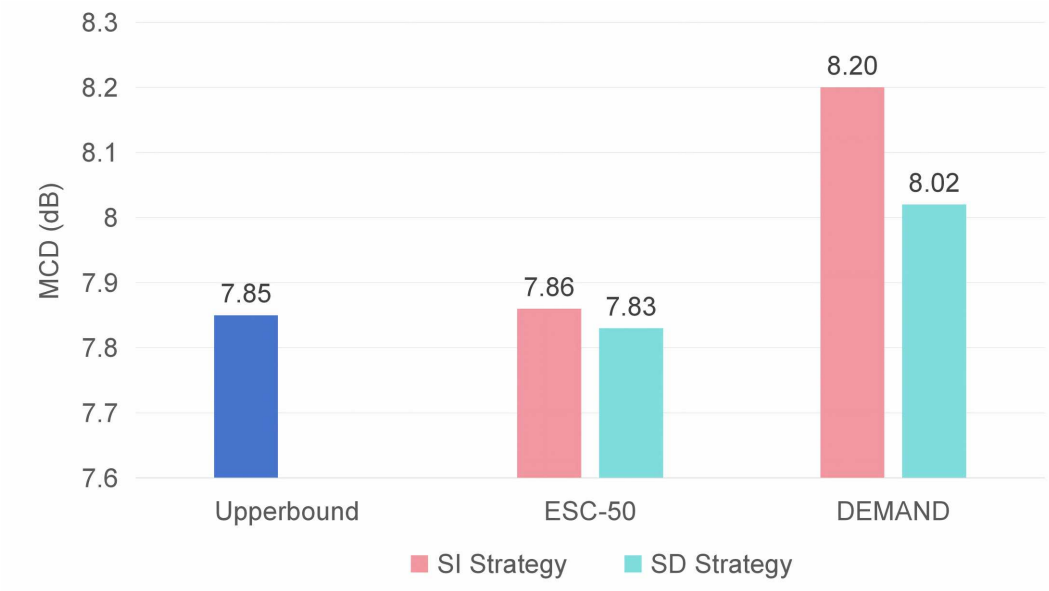


Figure 5.2: *MCD results for N2N trained on the original noisy datasets using SI and SD noise sampling strategies.*

and DEMAND, respectively. To further investigate the impact of sampling strategy independent of the dataset, this section extends the setup by applying the SI strategy to DEMAND and the SD strategy to ESC-50. To avoid the additional distortion introduced by the SE model resulting from the change in noise sampling strategy, the original noisy datasets are used for training and testing, instead of using denoised speech and separated noise as model inputs. Specifically, the noise-conditioned *N2N* takes clean speech as input and raw noise as conditions.

Figure 5.2 illustrates the MCD results for *N2N* trained on noisy datasets using SI and SD strategies. The *Upper Bound* refers to the original VC model trained solely on clean speech without noise conditioning. The labels "ESC-50" and "DEMAND" on the x-axis indicate the noise sources of the training data.

Theoretically, in the absence of degradation, *N2N* should achieve comparable MCD to the *Upper Bound*. As shown in Figure 5.3, no significant degradation is observed

when ESC-50 is used as the noise source, regardless of the sampling strategy: *N2N* models trained on E-SI and E-SD achieve MCDs of 7.86 and 7.83, closely matching the *Upper Bound* of 7.85. In contrast, when DEMAND is used as the noise source, substantial degradation is observed for both strategies. The *N2N* trained on D-SD and D-SI obtain MCDs of 8.02 and 8.20, respectively, both clearly inferior to 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 established by the sampling strategy.

Furthermore, the performance trends, where the *N2N* 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 5.1, even when clean speech and raw noise are employed. This consistency indicates that the SE model is not the primary cause of the observed performance degradation.

5.3.3 Effects of SNR Levels on VC Performance

In Chapter 4, the SSD noise sampling strategy was employed to examine the impact of SNR level, as SNR is a critical factor in defining noise conditions. Compared to the SD sampling strategy, which assigns each speaker a fixed SNR and a unique noise category, the SSD strategy introduces multiple SNR levels per speaker while keeping the noise category fixed. Experimental results show that *N2N-SSD* consistently outperforms *N2N-SD* across all metrics, suggesting that SNR variation may contribute to the observed performance degradation.

This section investigates the impact of SNR levels on the VC performance of the *N2N* by constructing two groups of parallel noisy training sets using different noise sources. To eliminate the distortion introduced by the SE model due to varying SNR

levels, the VC model is trained with clean speech and original noise clips, and evaluated by generating clean converted speech from clean inputs.

The first group adopts ESC-50 as the noise source, which is consistent with the previously used E-SI training set, where the *N2N* exhibits clear advantages over the *Baseline*. For each utterance in VCC2018, a noise clip is uniformly sampled and mixed with the clean speech at a fixed SNR level. Using this approach, ten parallel noisy training sets are constructed. In these parallel sets, each utterance is paired with the same noise clip across all datasets. The only difference lies in the SNR level used for mixing, with each set a fixed SNR from 25 dB, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB, -10 dB, -15 dB, and -20 dB. This setup is similar to the E-SI construction, except that each dataset contains a unique and fixed SNR level, enabling a controlled investigation of SNR-specific effects on VC performance.

The second group adopts DEMAND as the noise source. Its construction follows the procedure of the D-SD dataset. For each speaker in VCC2018, a noise category is randomly assigned. Each utterance is then mixed with a randomly selected noise clip from the assigned category at a fixed SNR. Unlike the previous D-SD set using only 5 dB, ten datasets are constructed here, each corresponding to one of the aforementioned SNR levels.

Figure 5.3 illustrates MCD results for *N2N* trained on original noisy datasets using ESC-50 and DEMAND as the noise sources. The *Upper Bound*, with an MCD of 7.85, is indicated by the blue dotted line for reference. As discussed in Section 5.3.2, *N2N* is expected to achieve comparable MCDs to the *Upper Bound* if no performance degradation occurs.

When trained with ESC-50, *N2N* maintains an MCD close to 7.85 for SNR levels ranging from 25 dB to -10 dB, showing comparable performance to the *Upper Bound*.

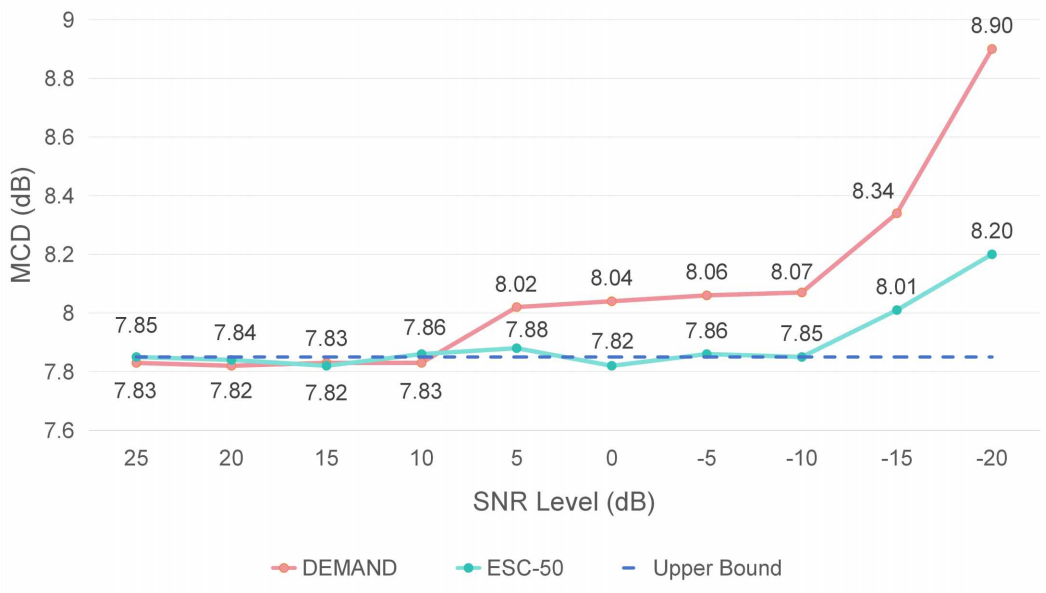
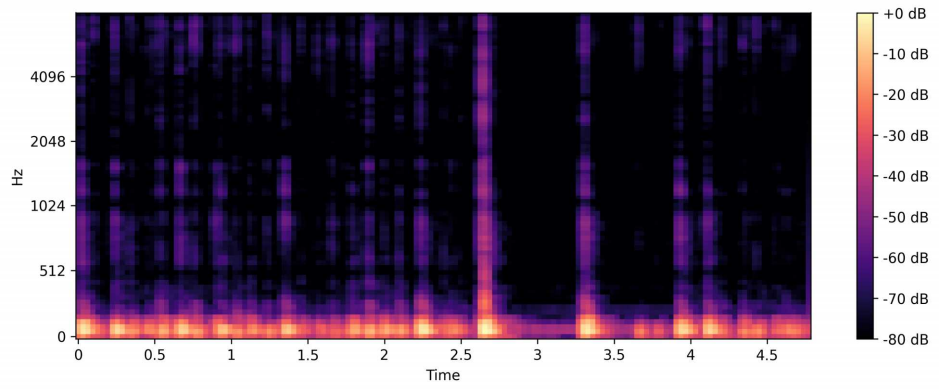


Figure 5.3: *MCD results for N2N trained on the original noisy datasets using ESC-50 and DEMAND as the noise sources, respectively.*

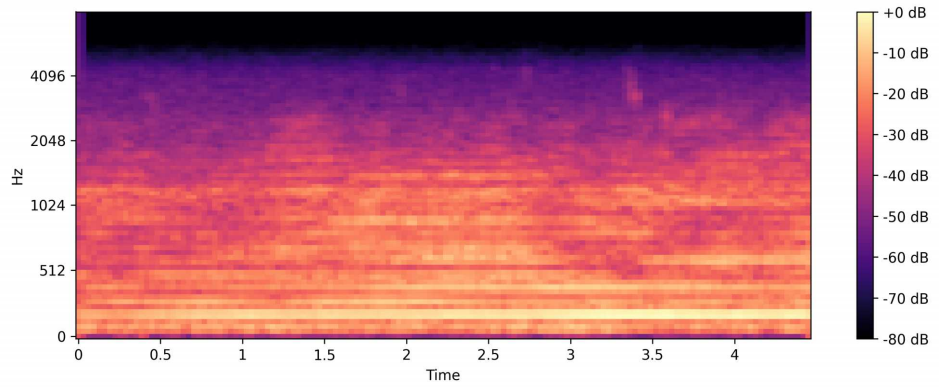
However, degradation becomes evident at lower SNRs. At −20 dB, *N2N* yields an MCD of 8.20, showing significant performance loss. This finding challenges the earlier conclusion in Section 4.5.1, which suggested that noise conditioning does not degrade the quality of clean converted speech when speaker identity and noise conditions are disentangled.

For the *N2N* trained using DEMAND as the noise source, a similar trend is observed. The MCDs remain comparable to the *Upper Bound* between 25 dB and 10 dB. Starting from an SNR of 5 dB, *N2N* yields higher MCD values than the *Upper Bound*, revealing signs of performance degradation.

Overall, the SNR level of the noisy training set contributes to the VC performance. Performance degradation occurs under both ESC-50 and DEMAND, though with different SNR thresholds. When using DEMAND as the noise source, degradation begins at an SNR of 10 dB, implying that factors beyond speaker-noise entanglement, such as



(a)



(b)

Figure 5.4: *Mel-spectrogram examples of the noise from keyboard and airplane categories. (a) Keyboard. (b) Airplane.*

noise characteristics, may contribute more significantly. In contrast, when using ESC-50 as the noise source, *N2N* shows degradation only below SNR of -15 dB, further supporting the hypothesis that the intrinsic properties of the noise dataset are critical contributors to performance degradation.

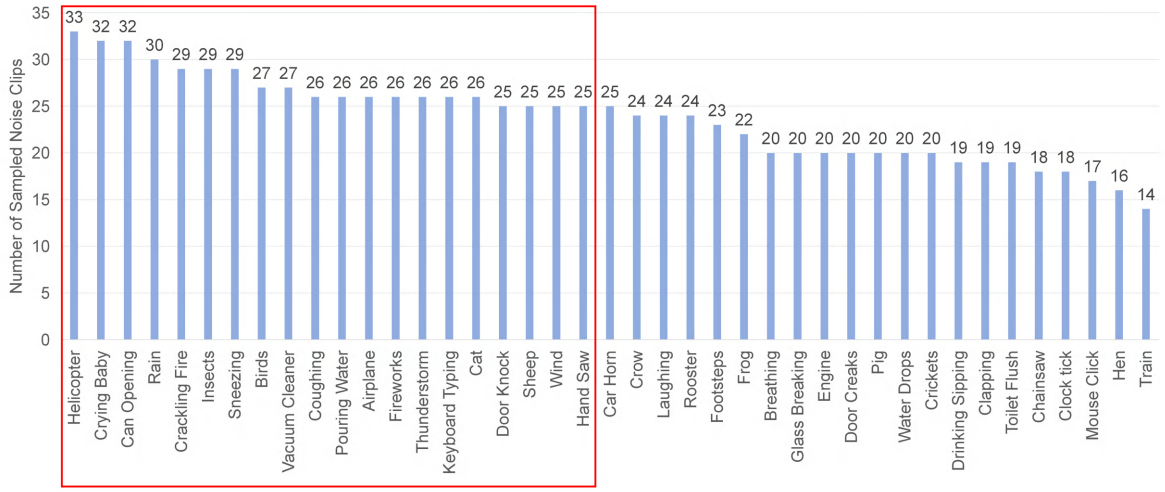


Figure 5.5: *Noise distribution in the E-SI dataset sorted by the number of sampled noise clips. The red frame highlights the 20 most sampled noise categories.*

5.3.4 Effects of Noise Categories on VC Performance

Besides the SNR level, another crucial aspect of the noise condition is the noise category. The type of noise not only offers a coarse-grained classification of background interference but also reflects intrinsic properties such as stationarity and temporal structure. For instance, impulsive noises like keyboard typing are typically non-stationary and sparse in the time domain, whereas steady noises like airplane engines exhibit more consistent spectral and temporal patterns, as demonstrated in Figure 5.4.

In Section 5.3.3, it is observed that VC performance degrades when the SNR level falls below a certain threshold, with the threshold differing between the two noise sources: ESC-50 and DEMAND. Since two *N2N* methods using different noise sources take the same clean speech as input and differ only in the noise conditioning component, this suggests that the modeling of the noise component plays a critical role in determining performance.

Given that *N2N* is trained to reconstruct the noisy speech as a whole, and the loss

function does not explicitly differentiate between speech and noise components, it can be hypothesized that the difficulty in modeling the noise component becomes the primary factor affecting VC performance. Specifically, at lower SNR levels, the noise becomes more dominant and potentially more complex, which increases the modeling difficulty that leads to performance degradation. The fact that different noise datasets (ESC-50 and DEMAND) exhibit different thresholds further supports the assumption that the characteristics of the noise, such as variability, stationarity, and spectral structure, impact the difficulty of modeling and thus influence the overall VC performance.

However, there is a lack of metrics to describe the characteristics of the noise and the associated modeling difficulty. Moreover, as both the N2N task and noise-conditioned VC models are relatively novel, the influence of noise characteristics on VC performance remains underexplored. To further investigate the impact of noise category on VC performance of the N2N framework, the E-SI dataset is selected as a representative case, as it includes a wide variety of noise types.

The distribution of noise categories in the E-SI dataset is first investigated, as illustrated in Figure 5.5. 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. Figure 5.6 illustrates the t-SNE visualization of BEATs-based [212] noise embeddings of these noise categories. Different from the count-based histogram in Figure 5.5, Figure 5.6 characterizes acoustic similarity between noise samples rather than the number of noise clips per category.

The red frame in Figure 5.5 highlights the top 20 sampled noise categories. Based on these, multiple noisy training sets are constructed, each using only a single noise category as the noise source. The *N2N* and *Baseline* are then trained on these category-specific datasets to evaluate whether the *N2N* can outperform the *Baseline* under

Figure 5.6: *Two-dimensional t-SNE visualization of noise feature representations across the noise categories in the E-SI dataset. The representations are extracted using BEATs [212] from 4-second segments and summarized with mean pooling and standard-deviation pooling over time.*

varying noise types. Additionally, to assess whether noise variety influences VC performance, two distinct strategies are employed for constructing the noisy training sets:

- **Multi-clip sampling**: Each utterance from VCC2018 is mixed with a noise clip uniformly sampled from the given noise category at an SNR of 5 dB. Consequently, multiple noise clips from the category are used to construct the noisy training set.

- **Single-clip sampling**: A single noise clip is uniformly sampled from the cat-

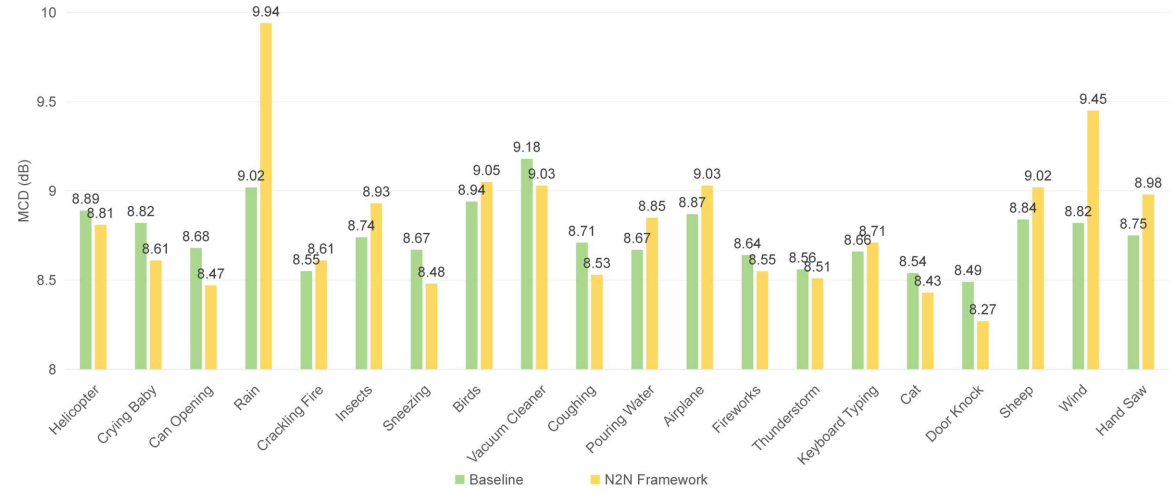
egory and used as the sole noise source for constructing the entire training set. To ensure sufficient training data, the sampled clip is temporally duplicated in the time domain to form a longer recording, from which random segments are extracted and mixed with utterances at an SNR of 5 dB.

The test set remains consistent with that used in previous experiments, as mentioned in Section 4.4.2, where ESC-50 serves as the noise source under the SI sampling strategy. As the use of clean speech and raw noise clips results in relatively small MCD differences across models, denoised data are employed for both training and evaluation to enhance sensitivity to performance variations.

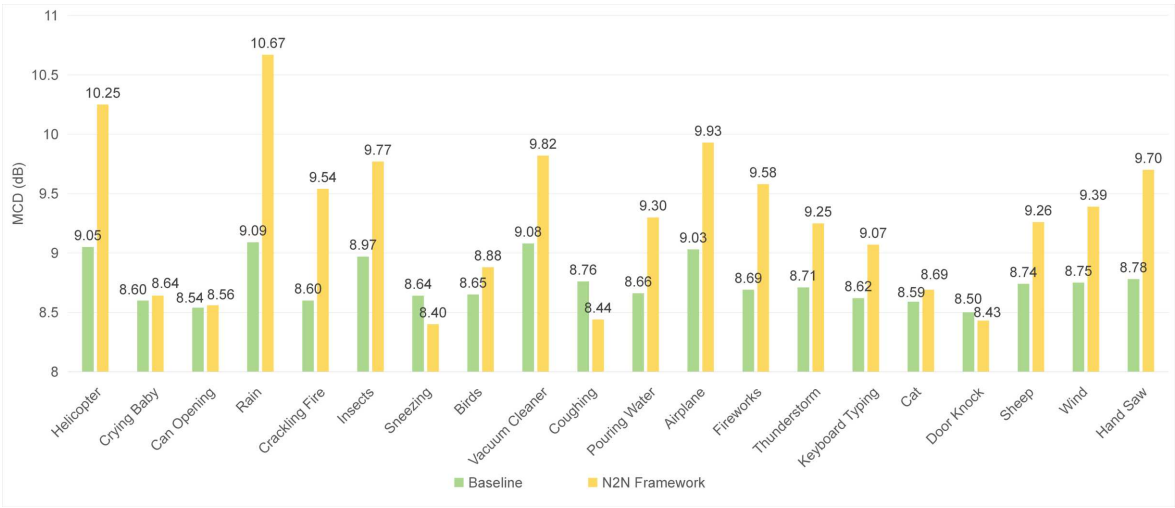
Figure 5.7 illustrates the MCD results for the *Baseline* and the *N2N* trained on a series of noisy training sets, each corresponding to a specific noise category with an SNR of 5 dB. Within each noise category, both multi-clip and single-clip strategies are applied.

Overall, for the same noise category, the *N2N*trained with the multi-clip strategy consistently outperforms its single-clip counterpart. Moreover, the VC performance degradation of the VC model is more prominent in the single-clip group. These findings underscore the significance of noise diversity in the training set to enhance VC performance. Even when all the noise components come from a single noise category, insufficient variability in the noise can lead to feature entanglement that degrades VC performance.

Compared to *N2N*, the *Baseline* yields similar MCD scores across both sampling strategies for most noise categories. Since the *Baseline* is trained on denoised speech, its MCD can be seen as partially reflecting the difficulty of modeling the speech component. A higher MCD indicates lower quality of the converted samples, which may result from greater residual noise or distortions introduced by the SE model.



(a)



(b)

Figure 5.7: *MCD results for 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.*

However, the difficulty of modeling the speech component alone can not fully account for the performance degradation observed in the *N2N* method. For instance, in

the *rain* category, the *Baseline* achieves MCDs of 9.02 and 9.09 under multi-clip and single-clip settings, respectively, which are comparable to those in the *vacuum cleaner* category of 9.18 and 9.09, suggesting similar difficulty in modeling the speech component. However, *N2N* obtains significantly higher MCDs of 9.94 and 10.67 in the *rain* category, indicating clear degradation, whereas in the multi-clip setting of the *vacuum cleaner* category, it achieves a lower MCD of 9.03, which outperforms the *Baseline*.

As shown in Figure 5.7 (b), the single-clip strategy is overly restrictive, resulting in the *N2N* underperforming the *Baseline* across nearly all noise categories. Therefore, the results of the multi-clip groups in Figure 5.7 (a) are considered to be further investigated.

Although the experimental results in Fig 5.3.2 demonstrate that the *N2N* trained on the E-SI dataset does not exhibit performance degradation compared to that trained on the D-SD dataset, performance degradation is observed when individual ESC-50 noise categories are used as the sole noise source. Among the 20 most frequently sampled categories in ESC-50, obvious performance degradation occurs in eight cases. Specifically, when the noise source is from *rain*, *insects*, *birds*, *pouring water*, *airplane*, *sheep*, *wind*, or *hand saw*, the *N2N* underperforms the *Baseline*, suggesting that certain noise types are more prone to causing VC performance degradation.

Initial observations of these results indicate that stationary noise types are more likely to contribute to VC performance degradation. For instance, the *N2N* exhibits degraded performance when trained on *rain* and *insects*. The *rain* recordings contain broadband noise with wide frequency coverage and evenly distributed energy, and the *insects* recordings 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* outperforms the

Baseline. However, although the noises in *vacuum cleaner* are wide-band and temporally stable, reflecting characteristics of stationary noise, the *N2N* trained on *vacuum cleaner* achieves an MCD of 9.03, outperforming the *Baseline* of 9.13. This suggests that stationarity alone may not fully explain the performance degradation, and other noise characteristics may also play a critical role. However, quantifying and systematically analyzing these factors remains a challenging task.

Additionally, as mentioned above, the loss function of the VC model L_{VC} in Equation 3.6 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. Considering the SNR-level analysis in Section 5.3.3 and the observed performance variations across different noise categories, it is suggested that the difficulty of modeling the noise component may also contribute significantly to VC performance degradation.

Although SNR is a direct metric for quantifying noise interference in a signal and experimental results in Section 5.3.3 show that performance degradation occurs when the SNR falls below a certain threshold, it alone does not effectively capture the extent of noise dominance in the N2N tasks. As shown in Figure 5.7 (a), even when all noisy utterances are generated at a fixed SNR of 5 dB, certain noise categories still lead to notable performance degradation.

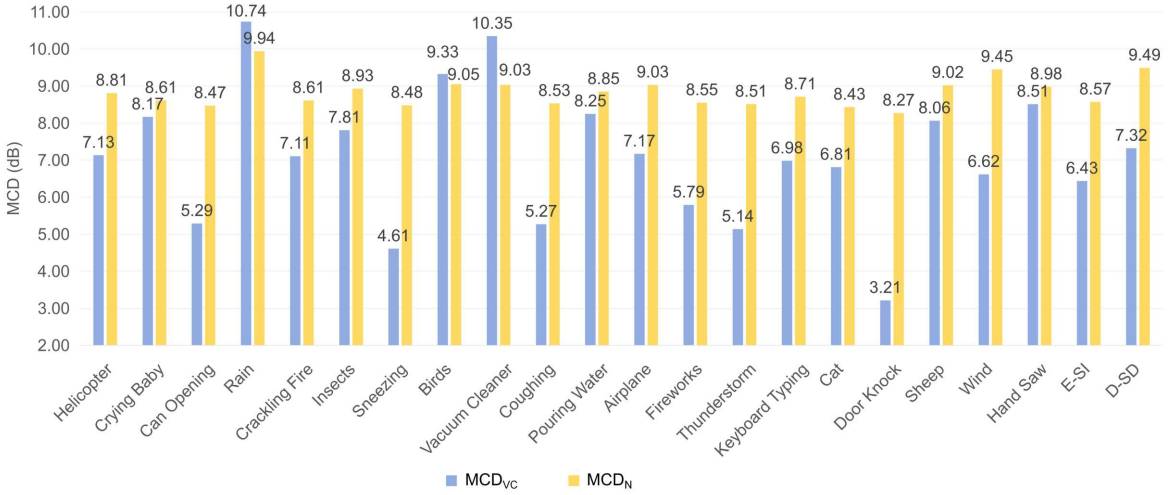
To better assess the degree of noise dominance in the training set, three additional objective metrics, MCD, PESQ, and STOI, are explored and computed. Specifically, those metrics are calculated between the clean utterances and their noisy counterparts

from the noisy training sets in Figure 5.7 (a), with E-SI and D-SD included as reference conditions. A higher MCD value reflects stronger spectral distortion and thus greater noise dominance, while higher PESQ and STOI scores indicate lower perceived distortion and better intelligibility, suggesting weaker noise dominance.

Figure 5.8 demonstrates the MCD, PESQ, and STOI results for the noisy training sets. Here, MCD_N denotes the MCD between the clean utterances and their noisy counterparts in the training sets, while MCD_{VC} denotes the MCD of the converted samples produced by the *N2N* trained on these noisy training sets. Since Figure 5.8 alone does not reveal the relationships among the metrics, Pearson correlation coefficients are calculated between MCD_N, PESQ, STOI, and MCD_{VC} to demonstrate the relationship between noise dominance and VC performance.

Figure 5.9 presents the Pearson Correlation between MCD_N, PESQ, STOI, and MCD_{VC}. In general, MCD_N exhibits strong negative correlations with PESQ and STOI, indicating that poorer scores reflect a higher level of noise dominance in the noisy speech. Notably, a strong positive linear relationship is observed between MCD_N and MCD_{VC}, evidenced by a Pearson correlation coefficient of 0.71 and a p-value of 3e-4, indicating high statistical significance. This is because both variables are mel-cepstral distances measured in decibels, capturing the same type of spectral envelope distortion. A higher MCD_N reflects a greater mismatch between the noisy utterance and its clean counterpart, suggesting increased noise dominance and greater modeling effort allocated to the noise component.

PESQ shows a moderate negative relation with MCD_{VC}, with a correlation coefficient of –0.46 and a corresponding p-value of 0.029. This can be attributed to the fact that PESQ penalizes not only envelope distortion but also time-domain clipping and bandwidth limitations. Since MCD_{VC} primarily reflects spectral-envelope mismatch,

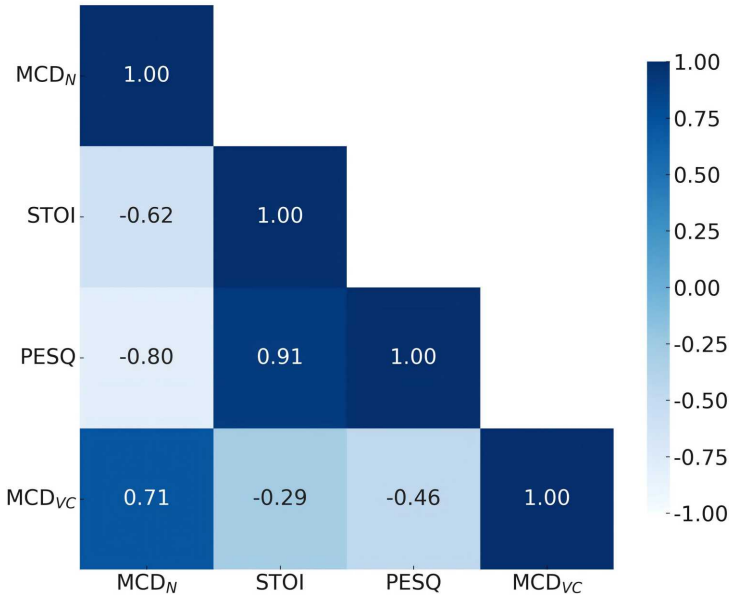


(a)

(b)

Figure 5.8: *MCD, PESQ, and STOI results for the noisy training sets in Figure 5.7 (a), E-SI, and D-SD. MCD_{VC} denotes the MCD for the converted samples produced by the N2N. (a) MCD results. (b) PESQ and STOI results.*

only the envelope-related errors captured by PESQ are relevant to the VC output, while other penalties, such as band limitations and clipping, do not manifest in the converted speech and thus introduce variance that weakens the linear relationship.



Predictor → MCD$_{VC}$	Pearson *r*	95% CI	*p*-value
MCD$_N$	+0.71	[+0.41, +0.87]	0.0003
PESQ	-0.46	[-0.74, -0.06]	0.0299
STOI	-0.29	[-0.62, +0.08]	0.1830

Figure 5.9: *Pearson Correlation between MCD_N, PESQ,j STOI, and MCD_{VC}.*

For STOI, a weak negative correlation with MCD_{VC} is observed, with a correlation coefficient of -0.295 and a p-value of 0.18, which does not reach statistical significance at the 0.05 level. This result aligns with the fact that STOI focuses on modulation fidelity in the 0–4 kHz band and is relatively insensitive to spectral colorations. As VC errors are mainly attributed to spectral-envelope deformation rather than envelope clipping, the correlation with STOI remains limited.

Overall, despite MCD, PESQ, and STOI can reflect aspects of noise dominance to some extent, the experimental results indicate that MCD_N provides a more effective explanation of the relationship between VC performance and the level of noise domi-

nance. A higher degree of noise dominance implies greater difficulty in modeling noisy speech for the noise-conditioned VC model. However, the VC model adopts the loss function defined in Equation 3.6, which evaluates the reconstruction of the noisy speech as a whole, without any explicit term for the speech component. As a result, the model tends to allocate more capacity to modeling the noise component instead of the speech part, leading to degradation in the quality of the converted speech. This phenomenon is referred to as noise bias, indicating the tendency of the VC model to prioritize noise modeling over speech reconstruction.

5.4 Summary of This Chapter

In the previous chapter, it was observed that the noise-conditioned *N2N* method suffered from VC performance degradation when trained on the D-SD (using DEMAND as the noise source with SD noise sampling strategy) noisy training set. Based on the characteristics of DEMAND and the SD strategy, this degradation was initially attributed to the entanglement between speaker identity and noise conditions during training.

In this chapter, the underlying causes of performance degradation in the *N2N* framework were further investigated. As a first step, the impact of the SI and SD noise sampling strategies on VC performance was assessed. Experimental results indicated that the sampling strategies had a limited effect on performance degradation compared to the influence of the noise source characteristics of the training sets.

Based on the E-SI dataset that included a diverse range of noise categories, the impact of the characteristics of the noise was investigated. To this end, *N2N* models were trained on noisy datasets constructed using individual noise categories from E-SI, each implemented with two distinct noise sampling strategies. In addition to the noise

category, the SNR level was also considered a crucial aspect of the noise condition and was accordingly evaluated.

Experimental results indicated that the performance degradation occurred when the SNR level dropped below a certain threshold. However, this threshold varied depending on the noise source, highlighting the significance of noise characteristics. Moreover, the results demonstrated that increasing noise diversity in the noisy training set enhanced VC performance, even when all noise samples originated from a single category. A lack of variability in noise could lead to poor generalization and feature entanglement, which negatively impacted the VC performance.

When the SNR level was fixed at SNR 5 dB, the performance degradation still occurred for certain noise categories used as the sole noise source. Given that the VC model computed the loss over the noisy speech as a whole without explicitly separating speech and noise components, the model might allocate more capacity to modeling noise when the noise component dominated the noisy speech or exhibited complex, hard-to-learn patterns. This phenomenon was referred to as noise bias.

However, there is a lack of metrics for directly quantifying the level of noise dominance. To address this, three metrics: MCD, PESQ, and STOI, were calculated between noisy training samples and their clean counterparts. Correlations between these metrics and the MCD of the converted speech produced by *N2N* were also analyzed. The experimental results suggested that MCD is the most effective metric for capturing the impact of noise dominance on VC performance.

In conclusion, two primary factors were identified as contributors to performance degradation in noise-conditioned *N2N* framework: limited noise diversity and noise bias. Greater noise diversity enhanced VC performance by covering a broader noise-speech distribution and mitigating feature entanglement, while noise bias arose when

the model overemphasized the noise component during training.

6 Improving the N2N Framework with Mutual Information Approximation and Noise Dropout

6.1 Introduction

In the previous chapter, the causes of performance degradation in the *N2N* framework were analyzed by evaluating multiple *N2N* methods trained on a series of noisy datasets, each constructed using a single noise category from the E-SI dataset. Two primary factors were identified: insufficient noise diversity and noise bias during training.

To enhance noise diversity, the most straightforward approach is data augmentation. However, as assessed in Chapter 4, although noise augmentation can improve the VC performance of *N2N*, it still fails to surpass the *Baseline*. One reason is that clean speech data are unavailable in the N2N task. When denoised speech is used as the corpus for augmentation, the resulting training targets are distorted, thereby undermining the advantage of the *N2N* framework, which is designed to use undistorted noisy speech as its target.

Another reason concerns noise categories that possess complex or hard-to-model

characteristics, which are more likely to cause performance degradation. Although MCD was shown to be an effective metric for indicating the level of noise dominance in noisy speech, this conclusion relies on the availability of clean reference speech and on comparative experiments between *N2N* and *Baseline*. In other words, there is currently no metric that can reliably predict whether a particular noise category will lead to degradation at a given SNR level. Consequently, augmenting with noise categories whose impact on VC performance cannot be reliably assessed is extremely risky, due to the lack of clean references and comparative benchmarks.

This chapter proposes two methods to address these issues: 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. Both subjective and objective experiments are conducted to evaluate the effectiveness of the proposed methods in improving VC performance. Additionally, an ablation study is performed to highlight the individual contributions of the mutual information approximation and noise dropout to the observed performance gains.

6.2 Proposed Methods

6.2.1 Mutual Information Approximation

The limited noise diversity has been identified in Section 5.3.4 as a primary factor contributing to VC performance degradation. However, directly increasing noise diversity via noise augmentation is risky, as certain noise types can further degrade VC performance. Moreover, as discussed in the same section, there is no established metric for quantifying the level of noise dominance, and the threshold at which VC performance degrades in the N2N framework remains uncertain and difficult to determine.

To address these challenges without introducing additional noise data, an alternative approach based on mutual information (MI) approximation is explored, aiming at mitigating the adverse effects of insufficient noise diversity.

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:

$$I(\mathbf{X};\mathbf{Y}) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right], \tag{6.1}$$

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 unavailable in most deep-learning-based cases, for it involves estimating the joint and marginal probability densities $p(x,y)$ and $p(x)p(y)$ in high-dimensional spaces. To address this, Cheng *et al.* [188] proposed a variational contrastive log ratio upper bound (vCLUB) to estimate an upper bound on MI defined in Equation 6.1 using contrastive learning and a reformulation of the log-ratio of probabilities. The vCLUB is defined as:

$$\begin{aligned} I_{vCLUB}(\mathbf{X};\mathbf{Y}) := {} & \mathbb{E}_{p(x,y)}\left[\log q_\theta(y \mid x)\right] \\ & -\mathbb{E}_{p(x)p(y)}\left[\log q_\theta(y \mid x)\right], \end{aligned} \tag{6.2}$$

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

The gap between $I_{\text{vCLUB}}(\mathbf{X};\mathbf{Y})$ and $I(\mathbf{X};\mathbf{Y})$ is calculated as:

$$\begin{aligned}
\tilde{\Delta} &:= I_{\text{vCLUB}}(\mathbf{X};\mathbf{Y}) - I(\mathbf{X};\mathbf{Y}) \\
&= \mathbb{E}_{p(x,y)}[\log q_\theta(y \mid x)] - \mathbb{E}_{p(x)p(y)}[\log q_\theta(y \mid x)] - \mathbb{E}_{p(x,y)}[\log p(y \mid x) - \log p(y)] \\
&= \mathbb{E}_{p(x)p(y)}\left[\log \frac{p(y)}{q_\theta(y \mid x)}\right] - \mathbb{E}_{p(x,y)}\left[\log \frac{p(y \mid x)}{q_\theta(y \mid x)}\right] \\
&= \mathbb{E}_{p(x)p(y)}\left[\log \frac{p(x)p(y)}{q_\theta(y \mid x)\, p(x)}\right] - \mathbb{E}_{p(x,y)}\left[\log \frac{p(y \mid x)\, p(x)}{q_\theta(y \mid x)\, p(x)}\right] \\
&= \mathrm{KL}\big(p(x)p(y) \,\|\, q_\theta(x,y)\big) - \mathrm{KL}\big(p(x,y) \,\|\, q_\theta(x,y)\big). \qquad (6.3)
\end{aligned}$$

Therefore, $I_{\text{vCLUB}}(\mathbf{X};\mathbf{Y})$ is an upper bound of $I(\mathbf{X};\mathbf{Y})$ if and only if $\mathrm{KL}\big(p(\mathbf{x})p(\mathbf{y}) \,\|\, q_\theta(\mathbf{x},\mathbf{y})\big) \geq \mathrm{KL}\big(p(\mathbf{x},\mathbf{y}) \,\|\, q_\theta(\mathbf{x},\mathbf{y})\big)$. Moreover, when $\mathbf{X}$ and $\mathbf{Y}$ are independent so that $p(\mathbf{x})p(\mathbf{y}) = p(\mathbf{x},\mathbf{y})$, we have $\mathrm{KL}\big(p(\mathbf{x})p(\mathbf{y}) \,\|\, q_\theta(\mathbf{x},\mathbf{y})\big) = \mathrm{KL}\big(p(\mathbf{x},\mathbf{y}) \,\|\, q_\theta(\mathbf{x},\mathbf{y})\big)$ and hence $\tilde{\Delta} = 0$.

In the N2N task, vCLUB is adopted 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 6.1. The representation $\mathbf{c}$ is the output of the first GRU, which encodes 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 6.2, vCLUB between $\mathbf{c}$ and $\mathbf{n_v}$ is given by:

$$\begin{aligned}
I_{\text{vCLUB}}(\mathbf{n_v};\mathbf{c}) &= \mathbb{E}_{p(c,n_v)}\left[\log q_\theta(c \mid n_v)\right] \\
&\quad - \mathbb{E}_{p(c)p(n_v)}\left[\log q_\theta(c \mid n_v)\right]. \qquad (6.4)
\end{aligned}$$

With sample pairs $\{(c_i,\, n_{v,i})\}_{i=1}^{N}$, $I_{\text{vCLUB}}(\mathbf{n_v};\, \mathbf{c})$ has an unbiased estimation as:

$$\begin{aligned}
\hat{I}_{\text{vCLUB}} &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\Big[\log q_\theta(c_i \mid n_{v,i}) - \log q_\theta(c_j \mid n_{v,i})\Big] \qquad (6.5) \\
&= \frac{1}{N}\sum_{i=1}^{N}\Big[\log q_\theta(c_i \mid n_{v,i}) - \frac{1}{N}\sum_{j=1}^{N}\log q_\theta(c_j \mid n_{v,i})\Big], \qquad (6.6)
\end{aligned}$$

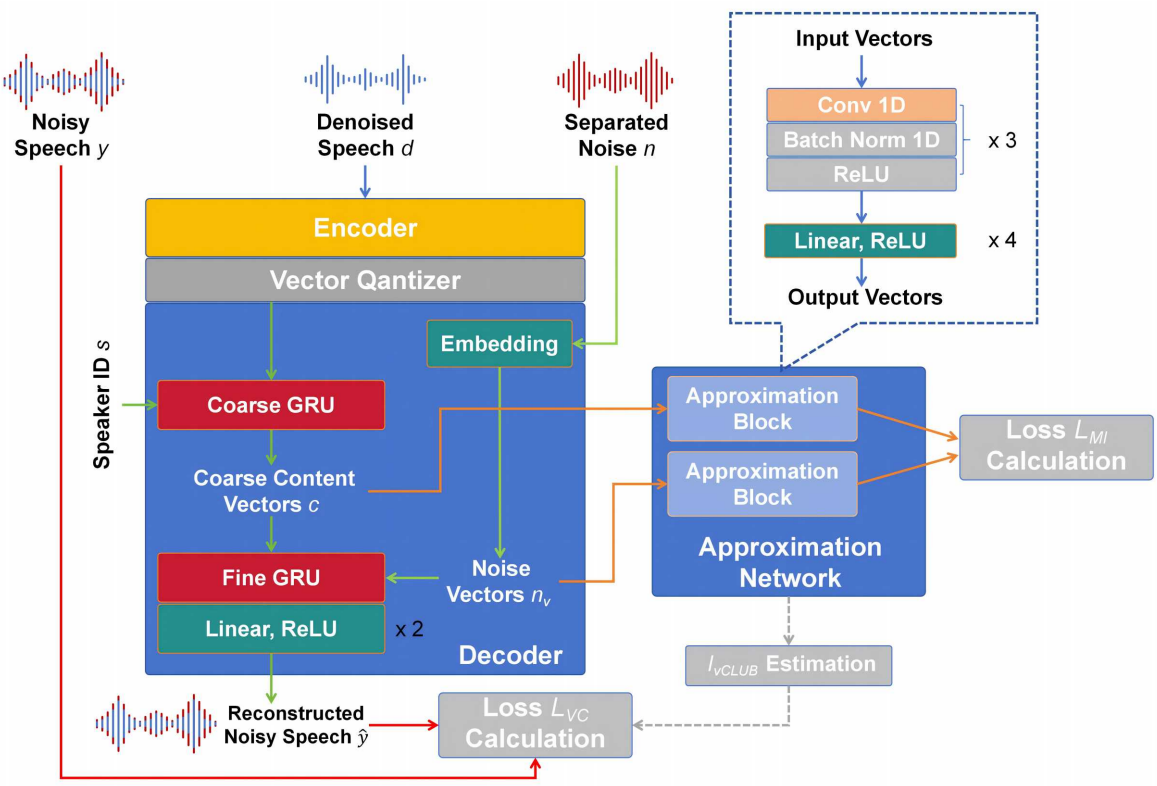


Figure 6.1: *Improved N2N framework with MI approximation network*

where 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, the approximation network is trained first to maximize the log-likelihood:

$$L_{MI} = \mathbb{E}_{p(c,n_v)} \left[\log q_\theta(c \mid n_v) \right]. \tag{6.7}$$

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

$$L_{Total} = L_{VC} + \lambda \hat{I}_{vCLUB}(\mathbf{n_v}; \mathbf{c}), \tag{6.8}$$

where $\hat{I}_{vCLUB}$ 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

the experiments.

Although MI approximation cannot directly resolve the problem of limited noise diversity, it can serve as an effective regularizer to alleviate the feature entanglement issue caused by the noisy training set's restricted noise variation. By penalizing redundant dependencies between $\mathbf{c}$ and $\mathbf{n_v}$, MI approximation encourages the disentanglement of speaker identity and speech content from interfering factors such as conditioned noise signal and residual artifacts in denoised speech, even when the training data lack diversity. Furthermore, MI-based regularization can be integrated into the unsupervised training process of the VC model and requires no additional training data, particularly clean speech data from source/target speakers and original noise clips from the noisy training set, which are typically unavailable in the N2N setting.

6.2.2 Noise Dropout Strategy

As discussed in Section 5.3.4, 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 components.

As formulated in Equation 6.2, the VC model's loss function primarily focuses on reconstructing the noisy speech as a whole. A straightforward solution is to introduce an additional loss term specifically for the reconstruction of the speech component. However, clean speech data from source and target speakers are unavailable for the VC task in the N2N setting, while such loss functions typically require clean utterances as the target. Similarly, improvements achieved by incorporating a discriminator for the

generated speech component may also be limited, as the discriminator has to rely on denoised speech as the real sample, forcing the VC model to approximate a distorted speech component.

Inspired by the dropout mechanism in deep learning, where a subset of neuron activations is randomly set to zero during each forward pass to promote redundant and robust representation learning, a noise dropout strategy is proposed to mitigate noise bias. During training, the entire noise signal is replaced with a zero sequence with a certain probability. This encourages the VC model to reconstruct the denoised speech serving as the loss target, thereby shifting the model ' s focus back to the speech component.

Similar to the noise augmentation discussed in Chapter 4, it is necessary to reduce the reliance on denoised speech as the training target. Over-reliance on denoised speech compromises the benefits of the noise-conditioned VC model, which uses noisy speech as the training target to alleviate the distortions introduced by the SE model. Therefore, the dropout rate is kept low to balance the trade-off effectively.

6.3 Experimental Setup

6.3.1 Experimental Datasets and Training Details

The same training and testing configurations as described in Chapter 4 are adopted. All audio data are sampled at 16 kHz.

The SE model is inherited from the previous experiments in Chapter 4, which is implemented using the DCCRN model and trained on the DNS Challenge 2020 dataset with SNRs ranging from 0 and 20 dB. The Adam optimizer is used to train the SE model with an initial learning rate of $2e{-}4$. An adaptive learning rate schedule is

applied based on validation performance, with a reduction factor of 0.5 and a patience of 3 epochs.

For the VC model, the D-SD dataset is used for training, which is constructed using the VCC2018 training set as the speech corpus and DEMAND as the noise source with the SD sampling strategy. The test set remains consistent with previous experiments, where noise clips from the categories excluded from the E-SI training set are sampled using the SI strategy and mixed with the VCC 2018 test set.

The VC model is also trained using the Adam optimizer, initialized with a learning rate of 2e−4. The training process is conducted for 500k steps with a step-based learning rate schedule, where the learning rate is halved at steps 100k and 200k to improve convergence. The MI approximation network follows the same training configuration as the VC model but uses a different learning rate schedule, with the rate halved at steps 50k and 150k. Based on the noise augmentation results reported in Chapter 4, the noise dropout rate is empirically set as 10%.

6.3.2 Methods to Be Evaluated

The experiments involve two main frameworks referred to as the *Baseline* and *N2N*. The 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 the condition to directly model the noisy speech. Two suffixes, *MI* and *ND*, are used for the mutual information approximation noise dropout strategy, respectively.

To differentiate method variations, we adopt the naming convention: *TypeOfModel ProposedMethod*. The objective evaluation is conducted as an ablation study including the *Baseline*, *N2N*, *N2N MI*, *N2N ND*, and *N2N NDMI*, where *N2N NDMI* denotes the

Table 6.1: *Objective evaluation results for the Baseline and the improved N2N 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**

N2N framework incorporating both noise dropout and MI approximation. Following the objective evaluation results, the *Baseline* and *N2N NDMI* are further assessed in the subjective evaluation.

6.3.3 Evaluation Metrics

Both objective and subjective evaluations are conducted to validate the effectiveness of the proposed methods. As Chapter 4 has demonstrated that the noise component can be consistently generated with high quality, and this chapter focuses on the speech component degradation, all noise-conditioned *N2N* methods generate the speech samples without background noise for evaluation.

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

For subjective evaluation, a preference test for naturalness and an XAB test for

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

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

similarity are conducted on Amazon MTurk with 12 participants. Based on the objective evaluation results presented in Table 6.1, two systems are chosen to be compared: *Baseline* and *N2N NDMI*. The evaluation investigates whether the proposed methods can enhance the *N2N* framework to outperform the *Baseline* under conditions where performance degradation is observed. Four source speakers (VCC2SF3, VCC2SF4, VCC2SM3, and VCC2SM4) and two target speakers (VCC2TF2 and VCC2TM2) are selected for evaluation. For each source-target pair, four converted utterances are uniformly sampled, resulting in 32 utterances per model.

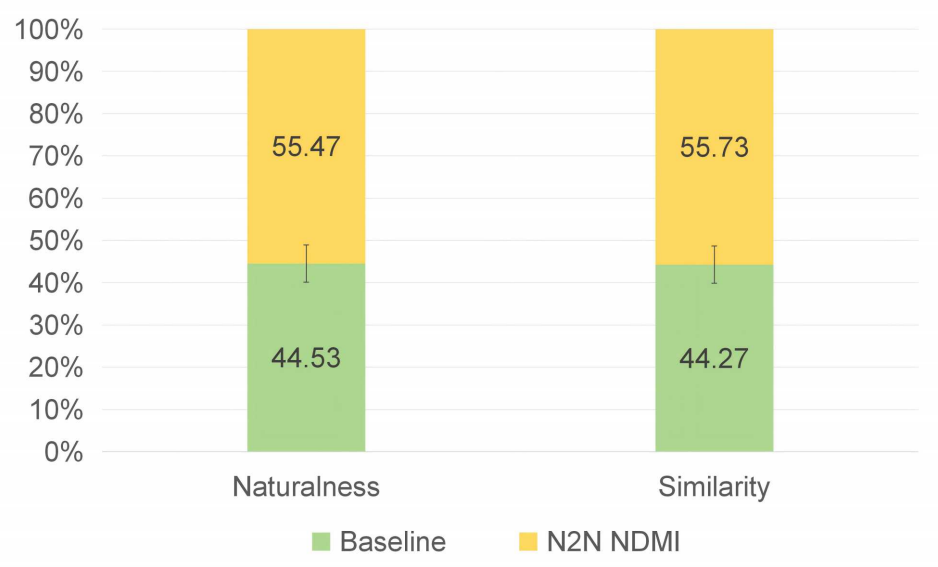
In the naturalness preference test, listeners are presented with paired samples from both *Baseline* and *N2N NDMI*, and asked to choose the more natural and higher-quality sample. In the XAB similarity test, listeners are provided with a reference sample from the target speaker and asked to select which of the two converted samples is more similar to the reference in terms of speaker identity.

6.4 Experimental Results

6.4.1 Results for Objective Evaluation

Table 6.1 shows the objective evaluation results from an ablation study comparing the *Baseline* and the improved *N2N* method with noise dropout and MI approximation. *N2N* achieves an MCD of 9.49, which is significantly higher than the *Baseline* of 9.09. Moreover, it shows lower performance in SIM, scoring 0.743 compared to the *Baseline* of 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



Metrics	P-values
Naturalness	0.032
Similarity	0.025

Figure 6.2: *Preference evaluation results in terms of naturalness and similarity with 95% confidence intervals for Baseline and N2N combined with noise dropout and MI approximation (P-values for naturalness and similarity are provided in the accompanying table).*

original N2N, achieving an MCD of 9.06, a SIM 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.

6.4.2 Results for Subjective Evaluation

Figure 6.2 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*, yet there remains room for further enhancement of the VC performance.

6.5 Summary of This Chapter

Based on the experimental analysis results in Chapter V, we identified two primary factors contributing to the VC performance degradation: the limited noise diversity leading to feature entanglement, and noise bias, where the noise-conditioned model tended to focus excessively on modeling the noise component rather than the speech part. To address the above issues, we proposed an MI approximation method to 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 mitigated the performance degradation of the N2N framework. When both MI approximation and noise dropout were combined, the N2N framework achieved the

best performance and outperformed the baseline. However, the subjective evaluations indicated that the improvements achieved through MI approximation and noise dropout were still limited, leaving room for further improvements.

7 Conclusions

7.1 Summary of This Thesis

In this thesis, we proposed a noise-robust N2N-VC framework that enabled explicit control of background noise in the converted output, without requiring clean speech from either source or target speakers. We reviewed recent advances in SE and VC, and identified open challenges in noise-robust VC while preserving the background noise.

In Chapter 3, we first revisited a baseline cascading an off-the-shelf SE front end and a VC module, and found that SE-induced distortions could severely impair downstream conversion, even offsetting the gains from a stronger SE model. To alleviate such distortions, we modified the VC module to predict noisy speech directly in the N2N task by introducing explicit noise conditioning using separated noise. Noise conditioning consistently improved conversion quality and narrowed the naturalness gap between the baseline and the upper bound.

In Chapter 4, we expanded the noisy conditions in the training set by using DEMAND as an additional noise source. However, we observed that the noise-conditioned VC degraded and even underperformed the baseline without noise conditioning. At first, we attributed this to the lack of noise diversity based on the characteristics of the noise source and the employed noise sampling strategy. To improve the VC performance, we introduced noise augmentation and pre-training strategies. However, experimental results indicated that noise augmentation and pre-training strategies could

alleviate the VC performance degradation, but the gained improvement was limited.

In Chapter 5, to explore the cause of the degradation, we conducted targeted experiments to assess how noise conditioning affected N2N-VC. Our analyses focused on the noise characteristics separately in terms of noise sampling strategy, noise diversity, SNR level, and noise category. The experimental results indicated that the degradation was not primarily explained by the noise sampling strategy. Instead, it was largely driven by two main factors: limited noise diversity, which promoted entanglement between speaker-related representations and noise conditions, and noise bias, where the model tended to prioritize reconstructing dominant noise components over the speech signal. In particular, performance dropped markedly at low SNRs and for stationary-like noise categories.

In Chapter 6, we proposed two regularization techniques to mitigate the performance degradation: an MI approximation-based regularizer to promote disentanglement and a noise dropout strategy to reduce noise bias. Experimental results showed that both the MI regularizer and noise dropout could improve N2N-VC performance, with noise dropout proving more effective. When both the MI regularizer and noise dropout were combined, the N2N-VC framework could achieve the best performance and outperformed the baseline. However, subjective evaluations showed that the gains remained modest, leaving scope for further improvement.

In conclusion, introducing noise conditioning with noisy speech modeling in the VC module for the N2N task yields better performance than the baseline without noise conditioning. However, when noise diversity in the training set is extremely limited, or when noise dominates more in the noisy mixture, the VC performance degrades and can even fall below the baseline. Two factors largely contribute to the noise dominance: SNR level and noise category. In most cases, when the SNR level is lower than 5 dB and

the noise is from a stationary-like category, the performance is likely to be degraded. Current metrics and criteria do not adequately quantify noise dominance, underscoring the need for better measures to assess it and to guide robustness improvements in N2N-VC.

7.2 Future Work

Compared with conventional noise-robust VC, N2N tasks are more constrained because paired noisy–clean data are unavailable, and only noisy speech is provided. Consequently, many standard strategies cannot be applied, such as denoising training and joint optimization of SE and VC. Rather than simply swapping in more advanced SE and VC models, a feasible alternative is to build a reliable front end for denoising and content representation: integrate an SE module with a self-supervised representation (SSL) encoder, pretrain them jointly on large-scale data for robustness, and then freeze this front end during N2N-VC training.

On the conversion back end, assessing and controlling the impact of noise characteristics on VC performance remains challenging. There is currently no reliable diagnostic to predict whether a particular noise recording will induce noise bias. Moreover, research on quantifying noise characteristics and using them as downstream conditioning signals is limited, as most prior work treats noise as interference to be removed. Developing robust evaluation protocols and noise characterization/quantization schemes for N2N-VC is therefore an important direction for future work.

Acknowledgments

For me, the years pursuing the doctorate were a long and difficult journey. But it is also unforgettable in my life. There were moments when I wanted to give up, but I chose to carry on and finally bring it to completion. I would like to express my deepest gratitude to my advisor, Prof. Toda. He is a kind and principled mentor with strong academic ethics. In research, Prof. Toda guided me with selflessness and clarity. Talking with him often led me to new ways of thinking, not just a single answer. In my studies, he also gave me patience and second chances: He never abandoned me, even after several graduation delays. Having such a mentor is one of the great honors of my life, and I wish to express my respect and gratitude once again.

I am also grateful to Ms. Noro, whose warm and efficient assistance covered everything from procuring supplies and equipment to preparing numerous documents. As an international student, I felt especially supported by her steady kindness.

I am especially grateful to my family. My wife's encouragement was constant. Every time I found myself uncertain about the way forward, she was the one who kept me going. I am also truly thankful to my parents for their unwavering support throughout this long journey.

I am also grateful to my labmates: YiChiao Wu, WenChin Huang, Choi, and Ding Ma, for their thoughtful discussions and help with my research.

Lastly, my sincere thanks to Prof. Toda once more. The years in Toda lab have

become part of who I am as a researcher. I will carry those lessons for life.

References

[1] R. Takashima, R. Aihara, T. Takiguchi, and Y. Ariki, "Noise-robust voice conversion based on spectral mapping on sparse space," in *SSW*, 2013, pp. 71–75.

[2] Y.-J. Chan, C.-J. Peng, S.-S. Wang, H.-M. Wang, Y. Tsao, and T.-S. Chi, "Speech enhancement-assisted voice conversion in noisy environments," in *Proceedings of 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2022)*, 2022, pp. 1533–1538.

[3] X. Miao, M. Sun, X. Zhang, and Y. Wang, "Noise-robust voice conversion using high-quefrency boosting via sub-band cepstrum conversion and fusion," *Applied Sciences*, vol. 10, no. 1, p. 151, 2019.

[4] Yeonjong Choi and Chao Xie and Tomoki Toda, "An Evaluation of Three-Stage Voice Conversion Framework for Noisy and Reverberant Conditions," in *Interspeech 2022*, 2022, pp. 4910–4914.

[5] 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.

[6] 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, pp. 14 687–14 697.

[7] 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, pp. 11 020–11 028.

[8] K. Shen, Z. Ju, X. Tan, E. 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," in *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., vol. 2024, 2024, pp. 698–722.

[9] 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, pp. 2525–2535.

[10] 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, pp. 12 161–12 165.

[11] K. Byun, J. Filos, E. Visser, and S. Moon, "Vc-enhance: Speech restoration with integrated noise suppression and voice conversion," *arXiv preprint arXiv:2409.06126*, 2024.

[12] 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, pp. 6593–6597.

[13] Ryo Terashima and Ryuichi Yamamoto and Eunwoo Song and Yuma Shirahata and Hyun-Wook Yoon and Jae-Min Kim and Kentaro Tachibana, "Cross-Speaker Emotion Transfer for Low-Resource Text-to-Speech Using Non-Parallel Voice Conversion with Pitch-Shift Data Augmentation," in *Interspeech 2022*, 2022, pp. 3018–3022.

[14] M. S. Ribeiro, J. Roth, G. Comini, G. Huybrechts, A. Gabryś, 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, pp. 6797–6801.

[15] 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, pp. 511–515.

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

[17] 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*, vol. 10, pp. 89 170–89 180, 2022.

[18] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "In-domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7554–7558.

[19] S. Shahnawazuddin, W. Ahmad, N. Adiga, and A. Kumar, "Children's speaker verification in low and zero resource conditions," *Digital Signal Processing*, vol. 116, p. 103115, 2021.

[20] 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, pp. 227–237.

[21] 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*, vol. 6, p. 100346, 2023.

[22] 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, pp. 1–5.

[23] 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 *Interspeech*, 2020, pp. 2472–2476.

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

[25] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[26] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[27] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[28] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108–110, May 2000.

[29] Y. Ephraim and H. L. V. Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[30] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008, pp. 4029–4032.

[31] E. M. Grais and H. Erdogan, "Single channel speech music separation using nonnegative matrix factorization and spectral masks," in *2011 17th International Conference on Digital Signal Processing (DSP)*. IEEE, 2011, pp. 1–6.

[32] K. Kwon, J. W. Shin, and N. S. Kim, "Nmf-based speech enhancement using bases update," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 450–454, 2014.

[33] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[34] G.-H. Ding, X. Wang, Y. Cao, F. Ding, and Y. Tang, "Speech enhancement based on speech spectral complex gaussian mixture model," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 1. IEEE, 2005, pp. I–165.

[35] A. Kundu, S. Chatterjee, A. S. Murthy, and T. Sreenivas, "Gmm based bayesian approach to speech enhancement in signal/transform domain," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4893–4896.

[36] J. Hao, H. Attias, S. Nagarajan, T.-W. Lee, and T. J. Sejnowski, "Speech enhancement, gain, and noise spectrum adaptation using approximate bayesian estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 24–37, Jan. 2009.

[37] Y. Yeminy, S. Gannot, and Y. Keller, "Speech enhancement using a multidimensional mixture-maximum model," in *Proceedings of the International*

Workshop on Acoustic Echo and Noise Control (IWAENC), 2010. [Online]. Available: https://www.iwaenc.org/proceedings/2010/HTML/Uploads/934.pdf

[38] J. Hao, T.-W. Lee, and T. J. Sejnowski, "Speech enhancement using gaussian scale mixture models," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 6, pp. 1127–1136, 2009.

[39] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, Apr. 1992.

[40] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "Hmm-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[41] D. Y. Zhao and W. B. Kleijn, "Hmm-based gain modeling for enhancement of speech in noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 882–892, Mar. 2007.

[42] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Proceedings of the 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*. Springer, 2010, pp. 140–148.

[43] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using hmm state-dependent super-gaussian priors," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 253–256, Mar. 2013.

[44] H. Veisi and H. Sameti, "Speech enhancement using hidden markov models in mel-frequency domain," *Speech Communication*, vol. 55, no. 2, pp. 205–220, Feb. 2013.

[45] Y. Wang and M. J. Gales, "Speaker and noise factorization for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2149–2158, 2012.

[46] L. K. Saul and M. G. Rahim, "Maximum likelihood and minimum classification error factor analysis for automatic speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 115–125, 2002.

[47] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[48] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 23, no. 1, pp. 7–19, 2014.

[49] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[50] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 37. PMLR, 2015, pp. 448–456.

[51] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech 2020*, 2020, pp. 3291–3295.

[52] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," *arXiv preprint arXiv:1911.13254*, 2019.

[53] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011, pp. 315–323.

[54] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[55] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2019.

[56] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1829–1843, 2021.

[57] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," in *Interspeech 2017*, 2017, pp. 3642–3646.

[58] H. Phan, H. Le Nguyen, O. Y. Chén, P. Koch, N. Q. Duong, I. McLoughlin, and A. Mertins, "Self-attention generative adversarial network for speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7103–7107.

[59] G. Yu, Y. Wang, C. Zheng, H. Wang, and Q. Zhang, "Cyclegan-based non-parallel speech enhancement with an adaptive attention-in-attention mechanism," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 523–529.

[60] Y. Li, M. Sun, and X. Zhang, "Perception-guided generative adversarial network for end-to-end speech enhancement," *Applied Soft Computing*, vol. 128, p. 109446, 2022.

[61] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 659–666.

[62] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7402–7406.

[63] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," in *Interspeech*, 2022, pp. 2928–2932.

[64] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.

[65] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2724–2737, 2023.

[66] H. Shi, K. Shimada, M. Hirano, T. Shibuya, Y. Koyama, Z. Zhong, S. Takahashi, T. Kawahara, and Y. Mitsufuji, "Diffusion-based speech enhancement with joint generative and predictive decoders," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 12 951–12 955.

[67] Y. Hu, C. Chen, R. Li, Q. Zhu, and E. S. Chng, "Noise-aware Speech Enhancement using Diffusion Probabilistic Model," in *Interspeech*, 2024, pp. 2225–2229.

[68] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[69] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[70] A. Narayanan and D. Wang, "The role of binary mask patterns in automatic speech recognition in background noise," *The Journal of the Acoustical Society of America*, vol. 133, no. 5, pp. 3083–3093, 2013.

[71] P. Wang and D. Wang, "Enhanced spectral features for distortion-independent acoustic modeling," in *Interspeech*, 2019, pp. 476–480.

[72] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 1, pp. 47–56, 2010.

[73] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind source separation: advances in theory, algorithms and applications*. Springer, 2014, pp. 349–368.

[74] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7092–7096.

[75] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[76] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 53–62, 2018.

[77] G. W. Lee and H. K. Kim, "Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection," *Applied Sciences*, vol. 10, no. 9, p. 3230, 2020.

[78] P. Selvaraj and E. Chandra, "Ideal ratio mask estimation using supervised dnn approach for target speech signal enhancement," *Journal of Intelligent & Fuzzy Systems*, vol. 42, no. 3, pp. 1869–1883, 2022.

[79] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[80] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5220–5224.

[81] D. S. Williamson and D. Wang, "Speech dereverberation and denoising using complex ratio masks," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5590–5594.

[82] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *International Conference on Learning Representations*, 2018.

[83] 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 *Interspeech*, 2020, pp. 2492–2496.

[84] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6633–6637.

[85] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement,"

in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7857–7861.

[86] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 9281–9285.

[87] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

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

[89] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *International Conference on Machine Learning*. PmLR, 2019, pp. 2031–2041.

[90] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," in *Interspeech*, 2021, pp. 201–205.

[91] S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "Metricgan-u: Unsupervised speech enhancement/dereverberation based only on noisy/reverberated speech," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7412–7416.

[92] Ruizhe Cao and Sherif Abdulatif and Bin Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Interspeech*, 2022, pp. 936–940.

[93] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation lstm network for real-time noise suppression," in *Interspeech*, 2020, pp. 2477–2481.

[94] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "Manner: Multi-view attention network for noise erasure," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7842–7846.

[95] S. wei Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," in *Interspeech*, 2018, pp. 1873–1877.

[96] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning-based objective assessment for voice conversion," in *Interspeech*, 2019, pp. 1541–1545.

[97] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE ICASSP*, 2019, pp. 7125–7129.

[98] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowd-sourced datasets," in *Interspeech*, 2021, pp. 2127–2131.

[99] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-*

2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 6493–6497.

[100] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 886–890.

[101] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 1. IEEE, 1998, pp. 285–288.

[102] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on speech and audio processing*, vol. 6, no. 2, pp. 131–142, 1998.

[103] H.-T. Hwang, Y. Tsao, H.-M. Wang, Y.-R. Wang, and S.-H. Chen, "Incorporating global variance in the training phase of gmm-based voice conversion," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2013, pp. 1–6.

[104] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 290–294.

[105] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Voice characteristics conversion for hmm-based speech synthesis system," in *1997 IEEE international*

conference on acoustics, speech, and signal processing, vol. 3. IEEE, 1997, pp. 1611–1614.

[106] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 805–808.

[107] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[108] H. Zen, Y. Nankaku, and K. Tokuda, "Probabilistic feature mapping based on trajectory hmms." in *INTERSPEECH*, 2008, pp. 1068–1071.

[109] T. Nose, Y. Ota, and T. Kobayashi, "Hmm-based voice conversion using quantized f0 context," *IEICE transactions on information and systems*, vol. 93, no. 9, pp. 2483–2490, 2010.

[110] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[111] K. Shikano, S. Nakamura, and M. Abe, "Speaker adaptation and voice conversion by codebook mapping," in *1991 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1991, pp. 594–597.

[112] L. M. Arslan and D. Talkin, "Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum." in *Eurospeech*, 1997, pp. 1347–1350.

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

[114] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[115] Z. Wu, C. E. Siong, and H. Li, "Joint nonnegative matrix factorization for exemplar-based voice conversion." in *INTERSPEECH*, 2014, pp. 2509–2513.

[116] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in *Proc. Interspeech 2016*, 2016, pp. 1632–1636.

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

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

[119] S. Ding and R. Gutierrez-Osuna, "Group latent embedding for vector quantized variational autoencoder in non-parallel voice conversion," in *Interspeech 2019*, 2019, pp. 724–728.

[120] D.-Y. Wu and H. yi Lee, "One-shot voice conversion by vector quantization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7734–7738.

[121] D.-Y. Wu, Y.-H. Chen, and H. yi Lee, "Vqvc+: One-shot voice conversion by vector quantization and u-net architecture," in *Interspeech*, 2020, pp. 4691–4695.

[122] D. Wang, L. Deng, Y. T. Yeung, X. Chen, X. Liu, and H. Meng, "Vqmivc: Vector quantization and mutual information-based unsupervised speech representation disentanglement for one-shot voice conversion," in *Interspeech*, 2021, pp. 1344–1348.

[123] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Avqvc: One-shot voice conversion by vector quantization with applying contrastive learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4613–4617.

[124] H. Tang, X. Zhang, J. Wang, N. Cheng, and J. Xiao, "Vq-cl: Learning disentangled speech representations with contrastive learning and vector quantization," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

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

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

[127] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H. yi Lee, and L. shan Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," in *ICASSP 2021 – IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5939–5943.

[128] J. hao Lin, Y. Y. Lin, C.-M. Chien, and H. yi Lee, "S2vc: A framework for any-to-any voice conversion with self-supervised pretrained representations," in *Interspeech*, 2021, pp. 836–840.

[129] C. H. Chan, K. Qian, Y. Zhang, and M. Hasegawa-Johnson, "Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6332–6336.

[130] Y. Zhou, X. Tian, Z. Wu, and H. Li, "Cross-lingual voice conversion with a cycle consistency loss on linguistic representation," in *Interspeech*, 2021, pp. 1374–1378.

[131] H. Du, X. Tian, L. Xie, and H. Li, "Optimizing voice conversion network with cycle consistency loss of speaker identity," in *2021 IEEE Spoken language technology workshop (SLT)*. IEEE, 2021, pp. 507–513.

[132] T. Dang, D. Tran, P. Chin, and K. Koishida, "Training robust zero-shot voice conversion models with self-supervised features," in *ICASSP 2022 – IEEE Int. Conf. Acoustics, Speech and Signal Processing*, 2022, pp. 6557–6561.

[133] W. Liang, L. Li, W. Du, and D. Wang, "Enhanced exemplar autoencoder with cycle consistency loss in any-to-one voice conversion," *arXiv preprint arXiv:2204.03847*, 2022. [Online]. Available: https://arxiv.org/abs/2204.03847

[134] J. Lian, P. Lin, Y. Dai, and G. Li, "Arbitrary voice conversion via adversarial learning and cycle consistency loss," in *International Conference on Intelligent Computing*. Springer, 2022, pp. 569–578.

[135] 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, pp. 1–6.

[136] 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*, vol. 27, no. 9, pp. 1432–1443, 2019.

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

[138] H. Lu, D. Wang, X. Wu, Z. Wu, X. Liu, and H. Meng, "Disentangled speech representation learning for one-shot cross-lingual voice conversion using ß-vae," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 814–821.

[139] K. Tanaka, H. Kameoka, and T. Kaneko, "Prvae-vc: Non-parallel many-to-many voice conversion with perturbation-resistant variational autoencoder," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[140] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. EUSIPCO*, 2018, pp. 2100–2104.

[141] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc2: Improved cyclegan-based non-parallel voice conversion," in *Proc. ICASSP*, 2019, pp. 6820–6824.

[142] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cyclegan-vc3: Examining and improving cyclegan-vcs for mel-spectrogram conversion," in *Interspeech*, 2020, pp. 2017–2021.

[143] S. Lee, B. Ko, K. Lee, I.-C. Yoo, and D. Yook, "Many-to-many voice conversion using conditional cycle-consistent adversarial networks," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6279–6283.

[144] 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)*, 2018, pp. 266–273.

[145] Y. A. Li, A. Zare, and N. Mesgarani, "Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," in *Interspeech*, 2021, pp. 1349–1353.

[146] R. Wang, Y. Ding, L. Li, and C. Fan, "One-shot voice conversion using star-gan," in *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7729–7733.

[147] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Interspeech*, 2017, pp. 3364–3368.

[148] D. Yook, I.-C. Yoo, and K. Lee, "Many-to-many voice conversion using cycle-consistent variational autoencoder with multiple decoders," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2020, pp. 215–222.

[149] K. Kumar, R. Kumar, T. De Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. De Brebisson, Y. Bengio, and A. C. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," *Advances in neural information processing systems*, vol. 32, 2019.

[150] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[151] Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda, "Quasi-periodic parallel wavegan vocoder: A non-autoregressive pitch-dependent dilated convolution model for parametric speech generation," in *Interspeech*, 2020, pp. 3535–3539.

[152] J. K. Kong, J. Kim, J. Son, Y. Byun, S. Yoon, and J. Kwon, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *NeurIPS 2020*, 2020, pp. 17 022–17 033.

[153] G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 492–498.

[154] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," in *International*

Conference on Learning Representations (ICLR), 2023. [Online]. Available: https://openreview.net/forum?id=iTtGCMDEzS_

[155] T. Bak, J. Lee, H. Bae, J. Yang, J.-S. Bae, and Y.-S. Joo, "Avocodo: Generative adversarial network for artifact-free vocoder," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 12 562–12 570.

[156] H. Lu, Z. Wu, R. Li, S. Kang, J. Jia, and H. Meng, "A compact framework for voice conversion using wavenet conditioned on phonetic posteriorgrams," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6810–6814.

[157] J.-X. Zhang, L.-J. Liu, Y.-N. Chen, Y.-J. Hu, Y. Jiang, Z.-H. Ling, and L.-R. Dai, "Voice conversion by cascading automatic speech recognition and text-to-speech synthesis with prosody transfer," in *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 121–125.

[158] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International conference on machine learning*. PMLR, 2022, pp. 2709–2720.

[159] H. Liu, T. Wang, R. Fu, J. Yi, Z. Wen, and J. Tao, "Unifyspeech: A unified framework for zero-shot text-to-speech and voice conversion," *arXiv preprint arXiv:2301.03801*, 2023. [Online]. Available: https://arxiv.org/abs/2301.03801

[160] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6706–6713.

[161] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=piLPYqxtWuA

[162] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in *International conference on machine learning*. PMLR, 2021, pp. 8599–8608.

[163] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *International conference on machine learning*. PMLR, 2015, pp. 1530–1538.

[164] J. Serrà, S. Pascual, and C. Segura Perales, "Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[165] P. Biliński, T. Merritt, A. Ezzerg, K. Pokora, S. Cygert, K. Yanagisawa, R. Barra-Chicote, and D. Korzekwa, "Creating new voices using normalizing flows," in *Interspeech*, 2022, pp. 936–940.

[166] M. Proszewska, G. Beringer, D. Sáez-Trigueros, T. Merritt, A. Ezzerg, and R. Barra-Chicote, "Glowvc: Mel-spectrogram space disentangling model for language-independent text-free voice conversion," in *Interspeech*, 2022, pp. 2973–2977.

[167] T. Merritt, A. Ezzerg, P. Biliński, M. Proszewska, K. Pokora, R. Barra-Chicote, and D. Korzekwa, "Text-free non-parallel many-to-many voice conversion using normalising flow," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6782–6786.

[168] L. Xu, R. Zhong, Y. Liu, H. Yang, and S. Zhang, "Flow-vae vc: End-to-end flow framework with contrastive loss for zero-shot voice conversion," in *Interspeech*, 2023, pp. 2293–2297.

[169] P. Ren, W. Guan, K. Wang, P. Chen, Q. Hong, and L. Li, "ReFlow-VC: Zero-shot Voice Conversion Based on Rectified Flow and Speaker Feature Optimization," in *Interspeech*, 2025, pp. 1388–1392.

[170] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[171] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162. PMLR, 2022, pp. 16 784–16 804.

[172] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[173] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[174] 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, pp. 741–748.

[175] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=8c50f-DoWAu

[176] 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, pp. 154–167.

[177] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, "Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation," in *Interspeech*, 2023, pp. 2283–2287.

[178] T. Kaneko, H. Kameoka, K. Tanaka, and Y. Kondo, "FastVoiceGrad: One-step Diffusion-Based Voice Conversion with Adversarial Conditional Diffusion Distillation," in *Interspeech*, 2024, pp. 192–196.

[179] 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.

[180] Y. Choi, C. Xie, and T. Toda, "Noise and reverberation-controllable voice conversion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 2430–2443, 2025.

[181] A. Mottini, J. Lorenzo-Trueba, S. V. K. Karlapati, and T. Drugman, "Voicy: Zero-Shot Non-Parallel Voice Conversion in Noisy Reverberant Environments," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 113–117.

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

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

[184] 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, pp. 2548–2552.

[185] 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, pp. 2182–2186.

[186] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *Advances in Neural Information Processing Systems*, vol. 33, pp. 8067–8077, 2020.

[187] 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*, vol. 131, p. 107807, 2024.

[188] 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, pp. 1779–1788.

[189] H. He, Y. Song, Y. Wang, H. Li, X. Zhang, L. Wang, G. Huang, E. S. Chng, and Z. Wu, "Noro: A noise-robust one-shot voice conversion system with hidden speaker representation capabilities," *arXiv preprint arXiv:2411.19770*, 2024.

[190] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. S. Kudinov, and J. Wei, "Diffusion-based voice conversion with fast maximum likelihood sampling scheme," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=8c50f-DoWAu

[191] H. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[192] T. Igarashi, Y. Saito, K. Seki, S. Takamichi, R. Yamamoto, K. Tachibana, and H. Saruwatari, "Noise-Robust Voice Conversion by Conditional Denoising Training Using Latent Variables of Recording Quality and Environment," pp. 2750–2754, 2024.

[193] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. Interspeech*, 2021. [Online]. Available: https://www.isca-archive.org/interspeech_2021/mittag21_interspeech.pdf

[194] Khaled Koutini and Jan Schlüter and Hamid Eghbal-zadeh and Gerhard Widmer, "Efficient Training of Audio Transformers with Patchout," in *Interspeech*, 2022, pp. 2753–2757.

[195] J. hao Lin, Y. Y. Lin, C.-M. Chien, and H. yi Lee, "S2vc: A framework for any-to-any voice conversion with self-supervised pretrained representations," in *Interspeech*, 2021, pp. 836–840.

[196] 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, pp. 5901–5905.

[197] 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, pp. 1–5.

[198] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dccrn: Deep complex convolution recurrent network for speech enhancement," in *Proceedings of Interspeech*, 2020, pp. 2472–2476.

[199] S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "Mapping and masking targets comparison using different deep learning based speech enhancement architectures," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[200] 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, pp. 6777–6781.

[201] 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, pp. 814–820.

[202] 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, pp. 6787–6791.

[203] 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, pp. 626–630.

[204] 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 *Interspeech*, 2019, pp. 181–185.

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

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

[207] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. Stöter, M. Hu, J. M. Martín-Doñas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, "Asteroid: The pytorch-based audio source separation toolkit for researchers," in *Interspeech*, 2020, pp. 2637–2641.

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

[209] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[210] 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, p. 035081.

[211] J. Yamagishi, C. Veaux, K. MacDonald *et al.*, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)," 2019.

[212] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

List of Publications

Journal Papers

1. C. Xie, T. Toda, "An Investigation of Noisy-to-noisy Voice Conversion Performance in Various Noisy Conditions," APSIPA Transactions on Signal and Information Processing, vol. 14: No. 1, e10, pp. 1–30, 2025.

2. C. Xie, T. Toda, "Noisy-to-Noisy Voice Conversion Under Variations of Noisy Condition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 3871–3882, 2023.

3. Y. Choi, C. Xie, T. Toda, "Noise and Reverberation-Controllable Voice Conversion," IEEE Transactions on Audio, Speech and Language Processing, vol. 33, pp. 2430–2443, 2025.

4. D. Ma, Y. Choi, T. Fujimura, F. Li, C. Xie, K. Kobayashi, T. Toda, *et al.*, "Sequence-to-sequence Voice Conversion-based Techniques for Electrolaryngeal Speech Enhancement in Noisy and Reverberant Conditions," APSIPA Transactions on Signal and Information Processing, vol. 14: No. 1, e8, pp. 1–40, 2025.

International Conferences

1. C. Xie, Y.-C. Wu, P.-L. Tobing, W.-C. Huang, T. Toda, "Noisy-to-noisy voice conversion framework with denoising model," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 814–820, 2021.

2. C. Xie, Y.-C. Wu, P.-L. Tobing, W.-C. Huang, T. Toda, "Direct noisy speech modeling for noisy-to-noisy voice conversion," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6787–6791, 2022.

3. C. Xie, T. Toda, "Noisy-to-noisy voice conversion with pretraining strategy," in *Proc. 24th International Congress on Acoustics (ICA)*, 2022.

4. Y. Choi, C. Xie, T. Toda, "An evaluation of three-stage voice conversion framework for noisy and reverberant conditions," in *Proc. Interspeech*, pp. 4910–4914, 2022.

5. Y. Choi, C. Xie, T. Toda, "Reverberation-Controllable Voice Conversion Using Reverberation Time Estimator," in *Proc. Interspeech*, pp. 2103–2107, 2023.

6. D. Ma, Y. Choi, F.-J. Li, C. Xie, K. Kobayashi, T. Toda, "Robust Sequence-to-sequence Voice Conversion for Electrolaryngeal Speech Enhancement in Noisy and Reverberant Conditions," in *Proc. IEEE Engineering in Medicine and Biology Society (EMBS)*, pp. 1–4, 2024.