Universal Discrete-Domain Speech Enhancement

Fei Liu, Yang Ai, *Member, IEEE*, Ye-Xin Lu, Rui-Chen Zheng, Hui-Peng Du, Zhen-Hua Ling, *Senior Member, IEEE*

***Abstract*—In real-world scenarios, speech signals are inevitably corrupted by various types of interference, making speech enhancement (SE) a critical task for robust speech processing. However, most existing SE methods only handle a limited range of distortions, such as additive noise, reverberation, or band limitation, while the study of SE under multiple simultaneous distortions remains limited. This gap affects the generalization and practical usability of SE methods in real-world environments. To address this gap, this paper proposes a novel Universal Discrete-domain SE model called UDSE. UDSE primarily works to enhance speech by predicting clean discrete tokens that are quantized by the residual vector quantizer (RVQ) of a pre-trained neural speech codec and are predicted in accordance with the RVQ rules. Specifically, UDSE first extracts global features from the degraded speech. Guided by these global features, the clean token prediction for each VQ follows the rules of RVQ, where the prediction of each VQ relies on the results of the preceding ones. Finally, the predicted clean tokens from all VQs are decoded to reconstruct the clean speech waveform. During training, the UDSE model employs a teacher-forcing strategy, and is optimized with cross-entropy loss. Experimental results confirm that the proposed UDSE model can effectively enhance speech degraded by various conventional and unconventional distortions, e.g., additive noise, reverberation, band limitation, clipping, phase distortion, and compression distortion, as well as their combinations. These results demonstrate the superior universality and practicality of UDSE compared to advanced regression-based SE methods.**

***Index Terms*—universal speech enhancement, neural speech codec, residual vector quantizer, discrete token**

I. Introduction

SPEECH signals inevitably suffer from various types of distortion in practical applications [1]. For instance, in outdoor environments, speech is frequently corrupted by different types of additive noise. In transmission scenarios, when speech is transmitted between narrowband communication devices, only the low-frequency components are typically preserved, resulting in bandwidth limitation and the introduction of additional compression noise [2]. These interferences significantly degrade the quality of the speech signal, posing challenges to the understanding of the target speech. Speech enhancement (SE) aims to process degraded speech signals using specific techniques to remove noise, reverberation, and other interfering components, with the goal of improving speech clarity and intelligibility [3]. The enhanced high-quality speech can better serve downstream tasks such as speech communication [4]–[6], hearing aid devices, and automatic speech recognition (ASR) [7]–[9]. High-quality, efficient, and generalizable SE techniques have become key research objectives for scholars worldwide, carrying significant academic and practical value.

In the past decade, with the development of deep learning, neural network-based SE models have significantly surpassed traditional algorithms in terms of enhanced speech quality and have gradually become the mainstream solutions for SE [10]. At present, most neural SE models operate in the continuous domain, formulating SE as a regression task in which the neural network directly predicts clean speech waveforms or continuous spectral features and employs regression-based loss functions. Regression-based continuous-domain neural SE models [11]–[27] can be roughly categorized into waveform-modeling-based and spectrum-modeling-based approaches according to their prediction targets and modeling objectives. Waveform-modeling-based models directly predict clean speech waveforms from degraded ones using a single neural network [11]–[18]. For example, SEGAN [12] is an early representative model that employs a fully convolutional network to directly predict clean speech waveforms and introduces a discriminator to enable adversarial training. DEMUCS [17] employs a multi-layer convolutional encoder-decoder architecture with U-Net-style skip connections, along with a sequence modeling module applied to the encoder output, to directly predict the clean speech waveform.

However, direct prediction of clean speech waveforms often leads to low generation efficiency and high computational complexity. In contrast, spectrum-model-based models [19]–[27] predict continuous clean spectral features and focus on enhancing speech at the frequency level, thereby avoiding direct manipulation of waveforms. Early spectrum-modeling-based SE models only enhance the amplitude spectrum while ignoring phase degradation, resulting in limited speech quality improvement [21]–[23]. To further improve phase enhancement, recent studies have proposed directly enhancing the short-time complex spectrum, thereby enabling the implicit enhancement of both amplitude and phase components. For instance, CMGAN [25] uses a Conformer-based [28] backbone to predict the amplitude spectrum and the real and imaginary parts of the short-time complex spectrum, which are then used to reconstruct the enhanced speech. However, these methods still fall short in achieving explicit and accurate phase enhancement. To address this, in our previous work, we proposed MP-SENet [26], [27], a model that performs parallel and direct enhancement of both the amplitude and phase spectra based on anti-wrapping phase prediction [29], [30], demonstrating impressive results.

Recently, the rapid development of neural speech coding

This work was funded by the National Nature Science Foundation of China under Grant 62301521, the Anhui Provincial Natural Science Foundation under Grant 2308085QF200.

F. Liu, Y. Ai, Y.-X. Lu, R.-C. Zheng, H.-P. Du and Z.-H. Ling are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China (e-mail: fliu215@mail.ustc.edu.cn, yangai@ustc.edu.cn, {yxlu0102, zhengruichen, redmist}@mail.ustc.edu.cn, zhling@ustc.edu.cn).

Corresponding author: Yang Ai.

technologies has fostered innovation in the field of SE. A few researchers have begun to explore the use of acoustic discrete tokens generated by neural speech codecs for SE, introducing novel discrete-domain approaches that aim to address classification problems rather than regression problems. In this framework, a neural network predicts the quantized acoustic discrete tokens (i.e., classification categories) of clean speech from degraded speech waveforms or continuous features. The enhanced speech is then reconstructed by decoding these predicted discrete tokens. The loss function is defined as the classification loss between the predicted and the ground truth clean discrete tokens. However, these methods are still in the early stages of development and face numerous challenges. For example, most existing work [31]–[35] relies on the capabilities of large language models (LLMs) for token predictions, resulting in excessive model complexity that hinders the practical deployment of SE systems. Genhancer [36] employs a relatively simple DF-Conformer [37] architecture but still relies on auxiliary components, such as ASR tools and self-supervised speech models, to extract text and semantic tokens for assisting acoustic token predictions, resulting in a complex and impractical solution.

In the past one to two years, researchers have started to place greater emphasis on universal SE, aiming to improve the robustness and adaptability of SE models across various distortion types and complex scenarios. For example, the URGENT Challenge [38], which was launched in 2024 and has been held for two consecutive editions, aims to advance universal SE techniques targeting multiple distortion types under challenging real-world conditions. However, current regression-based continuous-domain SE models primarily focus on addressing conventional distortions, such as additive noise, reverberation, and band limitation [39]–[41]. These models often struggle to handle unconventional distortion types or scenarios involving multiple simultaneous interferences. This is because many continuous-domain models formulate the task as a regression problem [24], aiming to learn the underlying noise patterns and to fit an explicit input–output mapping through neural networks. Such models tend to perform poorly when facing correlated noise (e.g., compression artifacts), where the noise shares strong characteristics with the underlying speech signal and cannot be explicitly modeled or fitted. In contrast, classification-based discrete-domain approaches theoretically hold promise for achieving universal SE. These methods formulate the problem as a classification task that maps the input into a latent discrete space. They leverage the strong generative capability of generative models to learn data distributions and then decode waveforms from the classified tokens. Furthermore, the cross-entropy loss used in the discrete domain penalizes incorrect predictions more severely, which helps the model better distinguish between correct and incorrect outputs. This theoretically makes them more flexible in handling diverse and complex distortions. Nonetheless, current research on discrete-domain methods remains largely focused on conventional additive noise, with limited exploration of their universality across a wider range of distortions.

Motivated by the aforementioned challenges and prior works, we propose a novel Universal Discrete-domain SE (UDSE) model guided by a pre-trained neural speech codec with a residual vector quantizer (RVQ). The UDSE model treats SE as a discrete-domain classification task, focusing on the prediction of clean acoustic discrete tokens quantized by the neural speech codec without requiring additional guidance from text or semantic tokens, and eliminating the need for LLMs. Specifically, starting from a randomly initialized sequence of discrete tokens, UDSE predicts clean acoustic discrete tokens for each VQ in sequence according to the rules of RVQ, where the prediction of each VQ depends on the outputs of the preceding ones. This process is guided by global features extracted from the degraded speech as a conditioning mechanism. Finally, the predicted tokens from all VQs are decoded to reconstruct the enhanced speech waveform. In our experiments, we simulated six types of distortions, i.e., noise, reverberation, band-limiting, clipping, phase distortion, and compression distortion, as well as three mixed distortion scenarios. Both objective and subjective experimental results confirm that our proposed UDSE exhibits robust performance across nine distortion scenarios, consistently restoring high-quality clean speech. This demonstrates its superior universality and practicality compared to advanced continuous-domain SE models.

The main contributions of this work are as follows:

- The proposed UDSE model introduces a novel token prediction paradigm that aligns with the RVQ scheme and may also provide inspiration for other discrete-domain speech generation tasks.
- The proposed UDSE model is capable of handling a wider range of distortion types, demonstrating superior universality and promoting the practicality of SE methods in complex real-world scenarios.

This paper is organized as follows. In Section II, we briefly review the advanced neural speech codecs and their applications in discrete-domain speech generation methods. In Section III, we provide the details of our proposed UDSE model. In Section IV, we present our experimental results. Finally, we give conclusions in Section V.

II. Related Work

The proposed UDSE integrates speech coding techniques and operates within the discrete domain. Therefore, this section primarily reviews two relevant areas: neural speech coding, as UDSE leverages neural speech codecs for discrete token representations, and discrete-domain speech generation methods, since UDSE also falls under the broader category of discrete-domain speech generation.

A. Neural Speech Coding

With the rise of deep learning, end-to-end neural speech codecs now markedly outperform traditional methods. They learn adaptive, compact representations that retain essential information and remove redundancy even at low bitrates, making them the dominant direction in speech coding. A central component of neural speech codecs is the learnable quantizer [42]–[45]. High quantization quality, reflected by

the close alignment between the encoder features and their quantized counterparts, enables the decoder to generate higher-quality reconstructed speech. In early work, researchers proposed a vector quantization (VQ) method that utilizes a learnable codebook, where each feature vector is assigned to the codeword with the minimum Euclidean distance, allowing effective representation learning within a differentiable quantization process [46]. Although this alleviates gradient-continuity issues, it often converges unstably and yields higher distortion. To address these limitations, RVQ was introduced and successfully applied in building the neural speech codec SoundStream [42]. RVQ connects multiple VQ modules in a residual manner by iteratively computing and quantizing the residual errors from previous quantization steps. Compared to single-VQ methods, RVQ further reduces overall quantization loss through this residual quantization process.

Building on SoundStream [42], which is the predecessor of neural speech codecs with RVQ, researchers have introduced further improvements in model architecture and training strategies, leading to the development of impressive codecs such as EnCodec [43] and DAC [44]. Recently, we have also proposed improved strategies regarding the coding targets, motivated by the observation that directly coding speech waveforms requires multiple downsampling and upsampling steps, leading to high computational complexity. For example, we introduced APCodec [45], which encodes and decodes the amplitude and phase spectra of speech, and further proposed MDCTCodec [47], which regards the MDCT spectrum of speech as coding target. By shifting the coding target from the waveform level to the spectral level, these approaches significantly reduce computational complexity while maintaining high reconstructed speech quality.

B. Universal Speech Enhancement

In real-world scenarios, distortion is highly diverse, and SE models designed for a single distortion type have limited effectiveness in practice. Therefore, exploring universal SE methods is of great importance. Recently, several generative SE approaches based on diffusion models have demonstrated strong performance [48]–[52]. For example, UniverSE++ [52] combines score-based diffusion and adversarial training: the score-based diffusion model uses stochastic differential equations to gradually transform clean speech into noise, while in the reverse process, it learns to recover the clean speech. Adversarial training further helps the model generate more natural-sounding speech, achieving excellent performance across various distortion types.

In addition to the abovementioned diffusion-based SE methods, recent studies have started to focus on applying neural speech codecs to the SE field, constructing novel discrete-domain SE approaches. This type of method has the potential to achieve universal SE by equivalently converting enhancement tasks of any distortion type into a classification problem over discrete representations. The discrete-domain SE leverages the discrete tokens produced by neural speech codecs as a bridge between degraded and clean speech, enabling SE within a discrete representation space. For example, Genhancer [36] adopts a DF-Conformer as its core architecture and combines discrete tokens with a self-supervised speech model to perform SE in the discrete domain. GenSE [34] employs a hierarchical modeling framework. It first uses an LLM to refine the semantic tokens extracted from degraded speech, producing enhanced semantic representations. These refined semantic tokens, along with the degraded semantic tokens and acoustic tokens of the degraded speech, are then jointly processed by another LLM to generate enhanced acoustic tokens, which are finally decoded to produce the enhanced speech. However, existing approaches often depend on large-scale models and complicated processing pipelines, which undermines their practical applicability, and there has also been insufficient exploration of their effectiveness for universal SE.

III. Proposed Method

A. Overview

Suppose a clean speech $\boldsymbol{x} \in \mathbb{R}^T$ is affected by distortions $\theta \subset \Theta$, resulting in the degraded speech $\boldsymbol{y} \in \mathbb{R}^T$, i.e.,

$$\boldsymbol{y} = \theta(\boldsymbol{x}), \tag{1}$$

where T is the length of speech waveform. Θ is the set of all possible distortions, and θ is a subset of Θ. The SE aims to recover the clean speech $\hat{\boldsymbol{x}} \in \mathbb{R}^T$ from $\boldsymbol{y}$ by constructing enhancement method ξ to minimize the gap between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$, i.e.,

$$\hat{\boldsymbol{x}} = \xi(\boldsymbol{y}). \tag{2}$$

The ideal solution for universal SE is that ξ is unique for all $\theta \subset \Theta$. The proposed UDSE model is designed to address the challenge of universal SE.

The proposed UDSE achieves SE in the discrete domain by solving a classification problem using a pre-trained neural speech codec with RVQ. As shown in Figure 1, UDSE consists of multiple modules. The global feature extraction module extracts global features from $\boldsymbol{y}$ to guide the prediction of each VQ token in the token prediction module, which are finally decoded by the speech decoding module to generate $\hat{\boldsymbol{x}}$. As shown in Figure 2, the training phase of UDSE additionally incorporates a training data generation module, which provides input data and defines the classification loss.

B. Global Feature Extraction Module

The global feature extraction module is designed to extract global feature $\boldsymbol{G}^{(y)} \in \mathbb{R}^{C \times L}$ from the degraded speech $\boldsymbol{y}$, where C and L represent the global feature dimension and the number of frames, respectively. This module consists of an encoder ϕ_E and an RVQ with N VQs $\phi_{Q_1}, \ldots, \phi_{Q_N}$ (each with the same codebook size) of a neural speech codec ϕ_{Codec}, as well as a feature processor ϕ_{FP}.

Specifically, $\boldsymbol{y}$ is first passed through ϕ_E to obtain a downsampled encoded feature by a factor of T/L, i.e.,

$$\boldsymbol{E}^{(y)} = \phi_E(\boldsymbol{y}), \tag{3}$$

where $\boldsymbol{E}^{(y)} \in \mathbb{R}^{K \times L}$ and K denotes the encoded feature dimension. Next, $\boldsymbol{E}^{(y)}$ is quantized by RVQ, and the quantization results of each VQ, $\tilde{\boldsymbol{Q}}_1^{(y)}, \ldots, \tilde{\boldsymbol{Q}}_N^{(y)} \in \mathbb{R}^{K \times L}$, are

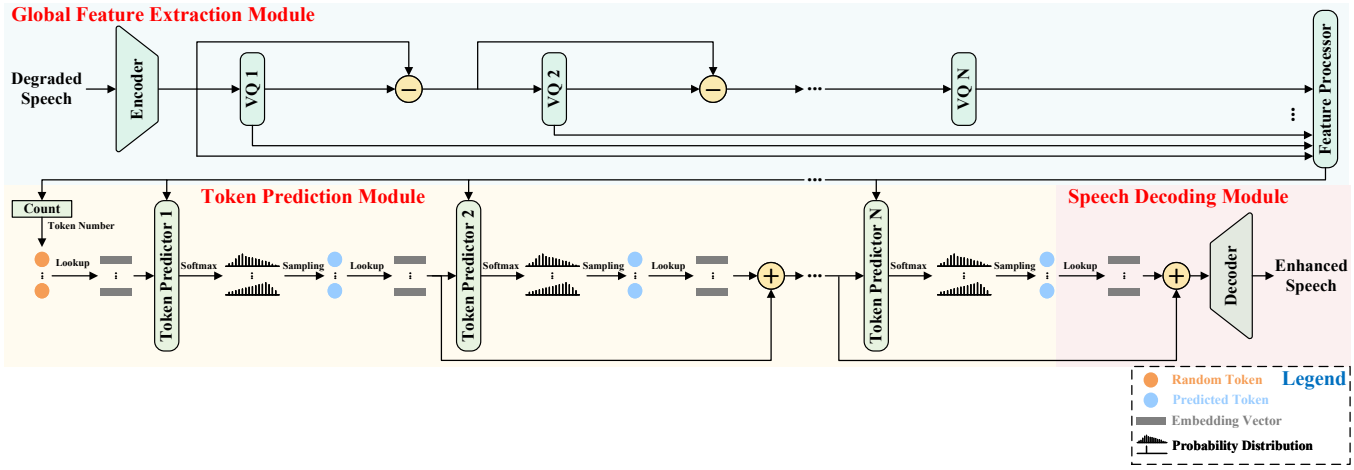

Fig. 1. Overview of the proposed UDSE's inference process.

sequentially outputted. During the quantization process, after the first VQ quantization, the difference between its input and output was used as the input to the second VQ. In other words, each subsequent VQ quantized the residual error of the previous one, forming a progressively refined quantization process. For ϕ_{Q_1}, its input is the encoded feature $\boldsymbol{E}^{(y)}$, let $\boldsymbol{Q}_1^{(y)} = \boldsymbol{E}^{(y)}$, i.e.,

$$\tilde{\boldsymbol{Q}}_1^{(y)} = \phi_{Q_1}(\boldsymbol{Q}_1^{(y)}), \tag{4}$$

whereas for ϕ_{Q_n}, its input is the quantization residual of $\phi_{Q_{n-1}}$, where $n = 2, \ldots, N$, i.e.,

$$\boldsymbol{Q}_n^{(y)} = \boldsymbol{Q}_{n-1}^{(y)} - \tilde{\boldsymbol{Q}}_{n-1}^{(y)}, \tag{5}$$

$$\tilde{\boldsymbol{Q}}_n^{(y)} = \phi_{Q_n}(\boldsymbol{Q}_n^{(y)}). \tag{6}$$

Finally, to fully preserve the original information, the quantization outputs from each VQ along with $\boldsymbol{E}^{(y)}$ are fed into the feature processor ϕ_{FP}. Within ϕ_{FP}, the input features are concatenated, dimension-reduced, and passed through B_G Conformer blocks [28], with residual connections applied to each block to link its input and output. This process generates the global feature $\boldsymbol{G}^{(y)}$, i.e,

$$\boldsymbol{G}^{(y)} = \phi_{FP}(\boldsymbol{E}^{(y)}, \tilde{\boldsymbol{Q}}_1^{(y)}, \ldots, \tilde{\boldsymbol{Q}}_N^{(y)}). \tag{7}$$

C. Token Prediction Module

The token prediction module starts from a randomly initialized sequence of discrete tokens, i.e.,

$$\boldsymbol{d}_0 = [d_{0,1}, \ldots, d_{0,L}]^\top, \tag{8}$$

and conditioned on the global feature $\boldsymbol{G}^{(y)}$, sequentially predicts the clean speech's token sequences generated through $\phi_{Q_1}, \ldots, \phi_{Q_N}$ by quantizing $\boldsymbol{x}$, i.e.,

$$\hat{\boldsymbol{d}}_n^{(x)} = [\hat{d}_{n,1}^{(x)}, \ldots, \hat{d}_{n,L}^{(x)}]^\top, n = 1, \ldots, N, \tag{9}$$

where $d_{0,*}/\hat{d}_{*,*}^{(x)} \in \{1, 2, \ldots, M\}$. The token number L is determined by the number of frames in $\boldsymbol{G}^{(y)}$ and M is the codebook size of a VQ. The token prediction module consists of N token predictors $\phi_{TP_1}, \ldots, \phi_{TP_N}$, each of which has a backbone composed of B_T Conformer blocks [28] with residual connections.

Specifically, for the first token predictor ϕ_{TP_1}, its input is the result $\hat{\boldsymbol{Q}}_0^{(x)} \in \mathbb{R}^{K \times L}$ obtained by looking up a random codebook $\mathbb{W}_0 = \{\boldsymbol{w}_{0,m} \in \mathbb{R}^K | m = 1, \ldots, M\}$ using the random token $\boldsymbol{d}_0$, and the global feature $\boldsymbol{G}^{(y)}$. For subsequent token predictors ϕ_{TP_n} $(n = 2, \ldots, N)$, following the dependency relationships among VQs in RVQ, their input is the sum of the results $\hat{\boldsymbol{Q}}_1^{(x)}, \ldots, \hat{\boldsymbol{Q}}_{n-1}^{(x)}$, which are obtained by looking up the codebooks $\mathbb{W}_1, \ldots, \mathbb{W}_{n-1}$ of $\phi_{Q_1}, \ldots, \phi_{Q_{n-1}}$ using the previously predicted tokens $\hat{\boldsymbol{d}}_1^{(x)}, \ldots, \hat{\boldsymbol{d}}_{n-1}^{(x)}$, respectively, and the global feature $\boldsymbol{G}^{(y)}$. This also clearly reflected that our prediction was based on the RVQ principle. Therefore, the output $\hat{\boldsymbol{U}}_n^{(x)} = [\hat{\boldsymbol{u}}_{n,1}^{(x)}, \ldots, \hat{\boldsymbol{u}}_{n,L}^{(x)}] \in \mathbb{R}^{M \times L}$ of ϕ_{TP_n} $(n = 1, \ldots, N)$ can be expressed as

$$\hat{\boldsymbol{U}}_n^{(x)} = \begin{cases} \phi_{TP_n}(\hat{\boldsymbol{Q}}_{n-1}^{(x)}, \boldsymbol{G}^{(y)}), & n = 1, \\ \phi_{TP_n}(\sum_{n_0=1}^{n-1} \hat{\boldsymbol{Q}}_{n_0}^{(x)}, \boldsymbol{G}^{(y)}), & n = 2, \ldots, N. \end{cases} \tag{10}$$

Next, each frame of $\hat{\boldsymbol{U}}_n^{(x)}$ is passed through a softmax layer to compute the classification probability distribution. Taking the l-th $(l = 1, \ldots, L)$ frame as an example, we can get

$$\hat{\boldsymbol{p}}_{n,l}^{(x)} = softmax(\hat{\boldsymbol{u}}_{n,l}^{(x)}), \tag{11}$$

where $\hat{\boldsymbol{p}}_{n,l}^{(x)} \in \mathbb{R}^M$. The token is sampled from $\hat{\boldsymbol{p}}_{n,l}^{(x)}$ based on the maximum probability, selecting the most likely category from the M categories, i.e.,

$$\hat{d}_{n,l}^{(x)} = argmax(\hat{\boldsymbol{p}}_{n,l}^{(x)}). \tag{12}$$

The above process is executed from $n = 1$ to $n = N$ to achieve the prediction of discrete tokens $\hat{\boldsymbol{d}}_1^{(x)}, \ldots, \hat{\boldsymbol{d}}_N^{(x)}$.

D. Speech Decoding Module

The speech decoding module uses the decoder ϕ_D of the neural speech codec ϕ_{Codec} to convert token prediction

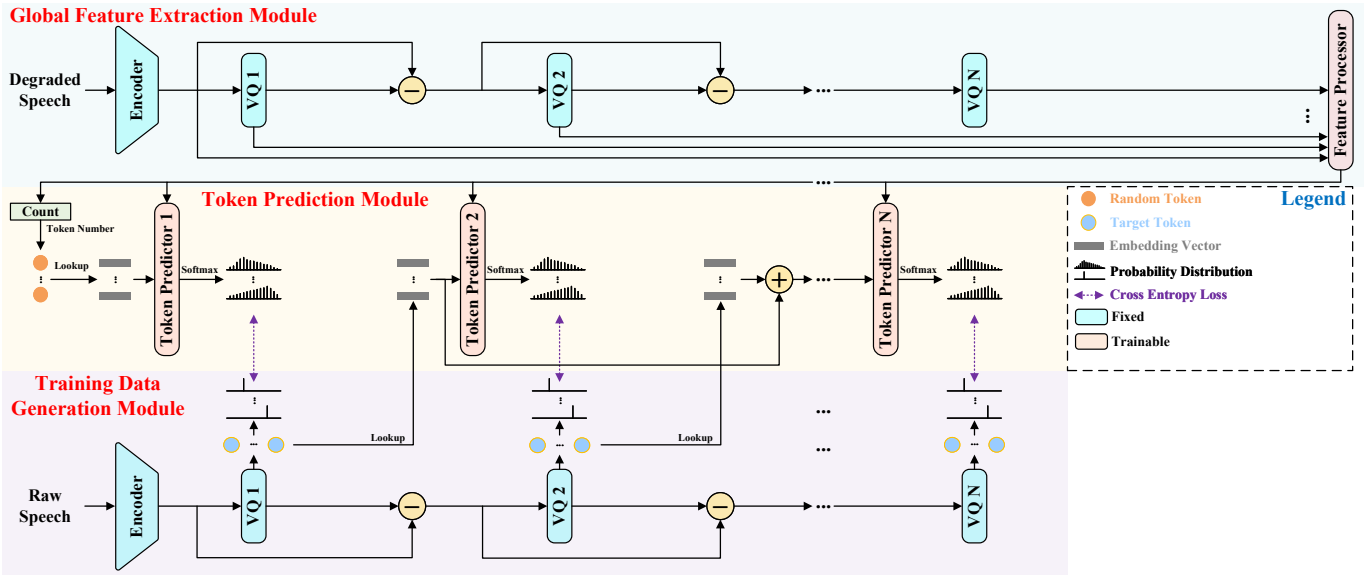

Fig. 2. Overview of the proposed UDSE's training process.

results $\hat{\boldsymbol{d}}_1^{(x)}, \ldots, \hat{\boldsymbol{d}}_N^{(x)}$ into the enhanced speech $\hat{\boldsymbol{x}}$. Specifically, $\hat{\boldsymbol{d}}_1^{(x)}, \ldots, \hat{\boldsymbol{d}}_N^{(x)}$ first retrieve $\hat{\boldsymbol{Q}}_1^{(x)}, \ldots, \hat{\boldsymbol{Q}}_N^{(x)}$ from the codebooks $\mathbb{W}_1, \ldots, \mathbb{W}_N$ of $\phi_{Q_1}, \ldots, \phi_{Q_N}$, respectively, and then add them together as the input to ϕ_D, which outputs $\hat{\boldsymbol{x}}$, i.e.,

$$\hat{\boldsymbol{x}} = \phi_D(\sum_{n=1}^{N} \hat{\boldsymbol{Q}}_n^{(x)}). \tag{13}$$

E. Training Strategy

During the training phase, we first train the neural speech codec ϕ_{Codec} which includes the encoder ϕ_E, quantizers $\phi_{Q_1}, \ldots, \phi_{Q_N}$, and decoder ϕ_D. As shown in Figure 2, we then froze all the parameters of the codec ϕ_{Codec} and proceeded to train the remaining components, i.e., ϕ_{FP} and $\phi_{TP_1}, \ldots, \phi_{TP_N}$. Compared to the inference process, UDSE introduces a training data generation module to provide training target and jointly train ϕ_{FP} and $\phi_{TP_1}, \ldots, \phi_{TP_N}$ at the training phase. To reduce the impact of error propagation on model training and enable the model to learn the correct patterns more quickly, we adopt a teacher forcing strategy during training. In other words, each token predictor takes the ground-truth clean token as input rather than the predicted one.

Specifically, in the training data generation module, clean speech $\boldsymbol{x}$ is processed through ϕ_E and $\phi_{Q_1}, \ldots, \phi_{Q_N}$ to generate ground-truth clean tokens, i.e.,

$$\boldsymbol{d}_n^{(x)} = [d_{n,1}^{(x)}, \ldots, d_{n,L}^{(x)}]^\top, n = 1, \ldots, N, \tag{14}$$

for both training targets and token predictors' input, where $d_{*,*}^{(x)} \in \{1, 2, \ldots, M\}$. Then, $\boldsymbol{d}_1^{(x)}, \ldots, \boldsymbol{d}_{N-1}^{(x)}$ lookup codebooks $\mathbb{W}_1, \ldots, \mathbb{W}_{N-1}$ to obtain quantized results $\tilde{\boldsymbol{Q}}_1^{(x)}, \ldots, \tilde{\boldsymbol{Q}}_{N-1}^{(x)}$ as inputs to $\phi_{TP_2}, \ldots, \phi_{TP_N}$ for teacher forcing training, and the output is calculated according to Equation (10), i.e.,

$$\tilde{\boldsymbol{U}}_n^{(x)} = \begin{cases} \phi_{TP_n}(\hat{\boldsymbol{Q}}_{n-1}^{(x)}, \boldsymbol{G}^{(y)}), & n = 1, \\ \phi_{TP_n}(\sum_{n_0=1}^{n-1} \tilde{\boldsymbol{Q}}_{n_0}^{(x)}, \boldsymbol{G}^{(y)}), & n = 2, \ldots, N. \end{cases} \tag{15}$$

Next, each frame of $\tilde{\boldsymbol{U}}_n^{(x)}$ is passed through a softmax layer to compute the classification probability distribution. Taking the l-th ($l = 1, \ldots, L$) frame as an example, according to Equation (11), we can get

$$\tilde{\boldsymbol{p}}_{n,l}^{(x)} = softmax(\tilde{\boldsymbol{u}}_{n,l}^{(x)}), \tag{16}$$

where $\tilde{\boldsymbol{u}}_{n,l}^{(x)}$ is the l-th column vector in $\tilde{\boldsymbol{U}}_n^{(x)}$. Finally, token $d_{n,l}^{(x)}$ is one-hot encoded to generate the target probability distribution $\boldsymbol{p}_{n,l}^{(x)}$, and the cross-entropy loss is defined with respect to $\tilde{\boldsymbol{p}}_{n,l}^{(x)}$ to minimize the distance between two distributions, used for model training, i.e.,

$$\begin{aligned} \mathcal{L} &= \frac{1}{NL} \sum_{n=1}^{N} \sum_{l=1}^{L} CrossEntropy(\tilde{\boldsymbol{p}}_{n,l}^{(x)}, \boldsymbol{p}_{n,l}^{(x)}) \\ &= -\frac{1}{NL} \sum_{n=1}^{N} \sum_{l=1}^{L} \log \tilde{\boldsymbol{p}}_{n,l}^{(x)}(d_{n,l}^{(x)}), \end{aligned} \tag{17}$$

where $\tilde{\boldsymbol{p}}_{n,l}^{(x)}(d_{n,l}^{(x)})$ denotes the $d_{n,l}^{(x)}$-th element in $\tilde{\boldsymbol{p}}_{n,l}^{(x)}$.

IV. Experiments

A. Dataset Construction and Task Definition

We constructed clean speech dataset from the VoiceBank corpus [53], with 23,075 utterances from 56 speakers for training set and 824 utterances from 2 unseen speakers for test set. The speech sampling rate is 44.1 kHz. We constructed task-specific degraded datasets based on a clean speech dataset, incorporating three conventional distortions (i.e., noise, reverberation, and band-limiting), three unconventional distortions

(i.e., clipping, phase distortion, and compression distortion), and three mixed distortions. Each task is defined as follows.

1) **Denoising (DN)**: This task aims to enhance the degraded speech caused by additive noise. We constructed the noisy dataset by adding noise from the DEMAND dataset [54] to the clean speech. For the training set, we used 5 types of noise, with the signal-to-noise ratio (SNR) ranging from 0 to 15 dB, in 5 dB intervals. For the test set, we used 5 types of unseen noise, with the SNR ranging from 2.5 to 17.5 dB in 5 dB intervals.

2) **Dereverberation (DR):** This task aims to enhance the degraded speech caused by reverberation. We adopted the room impulse response (RIR) dataset from the DNS Challenge [55], which included 248 real RIRs and approximately 60,000 simulated RIRs. We constructed the reverberant dataset by randomly selecting RIRs and convolving RIRs with clean speech. The RIRs of the test set were unseen in the training set.

3) **Bandwidth Extension (BWE):** This task aims to enhance the degraded speech caused by band-limiting. We constructed the band-limited dataset by downsampling the clean wideband speech to a 2 kHz sampling rate for both training and testing.

4) **Declipping (DC):** This task aims to enhance the degraded speech caused by clipping. We constructed the clipping dataset by restricting the amplitude of the original clean speech waveform to the range between 0.1 and 0.9 of their maximum values.

5) **Phase Distortion Restoration (PDR):** This task aims to enhance the degraded speech caused by inaccurate phase. We constructed the phase distortion dataset by replacing phase spectra of clean speech with the ones predicted by an NSPP[1] model [29].

6) **Compression Distortion Restoration (CDR):** This task aims to enhance the degraded speech caused by compression from a speech codec. We constructed the compression distortion dataset by compressing clean speech using an APCodec[2] [45] with a single VQ.

7) Mixed Distortion Restoration: We combined the six types of distortions mentioned above to create three tasks, represented as **DN+DR+BWE**, which stands for denoising + dereverberation + bandwidth extension (band-limited speech sampling rate is 8 kHz), **DN+DR+DC**, which stands for denoising + dereverberation + declipping, and **DN+PDR+CDR**, which stands for denoising + phase distortion restoration + compression distortion restoration. We did not downsample the speech to 2 kHz in the **DN+DR+BWE** task, as in the standalone **BWE** task, because further adding noise and reverberation to 2 kHz speech would render the speech nearly unintelligible and the SE solution unrealistically ill-posed.

[1]https://github.com/YangAi520/NSPP.

[2]https://github.com/YangAi520/APCodec.

B. Model Details

The UDSE[3] utilized a DAC[4] [44] as ϕ_{Codec}, which included 9 VQs (i.e., $N = 9$), with both the codebook size and the code vector dimension set to 1024 (i.e., $K = M = 1024$). The feature processor ϕ_{FP} used 8 Conformer blocks (i.e., $B_G = 8$), while each token predictor $\phi_{TP_1}, \ldots, \phi_{TP_N}$ used 4 Conformer blocks (i.e., $B_T = 4$). The number of channels and attention heads in each block were 512 and 8, respectively (i.e., $C = 512$). We trained the UDSE using an AdamW optimizer on a single NVIDIA A800 GPU, with $\beta_1 = 0.9, \beta_2 = 0.95$, and a weight decay of 0.01 for 100 epochs. The initial learning rate was set to 0.0005, with a cosine annealing strategy for decay and a warm-up training scheduler for the first 4k steps.

We compared UDSE with several advanced continuous-domain SE models, including waveform-regression-based DEMUCS[5] [17], spectrum-regression-based CMGAN[6] [25] and MP-SENet[7] [26], [27], and diffusion-based UniverSE++[8] [52]. In addition, we have also reproduced the Genhancer [36] as a discrete-domain baseline. For task **BWE**, we also compared UDSE with the advanced continuous-domain AP-BWE[9] [56], specifically designed for speech bandwidth extension. Except for Genhancer, we reproduced all the models using the official codes they provided, whereas for Genhancer, we had to reproduce it manually due to the lack of open-source code. All models were conducted on our 44.1 kHz dataset.

C. Evaluation Metrics

For objective evaluation, since UDSE predicts the distribution of discrete tokens corresponding to clean speech (i.e., it formulates SE as a classification task rather than directly generating the clean speech as in regression-based models), traditional intrusive indicators such as perceptual evaluation of speech quality (PESQ) are not well-suited for evaluating our model, as suggested in [34]. Therefore, we used three non-intrusive metrics commonly adopted in SE field, including NISQA [57], DNSMOS [58] and UTMOS [59], to assess overall speech quality. NISQA is a full-band metric that can evaluate the overall quality of 44.1 kHz speech, while DNSMOS and UTMOS are both designed for 16 kHz, focus more on the low-frequency quality. All three metrics use the same scoring scale as traditional mean opinion score (MOS), ranging from 1 to 5. Both NISQA and UTMOS directly reflect the perceived acceptability and naturalness of speech. DNSMOS, on the other hand, follows the ITU-T P.835 subjective test framework and provides multi-dimensional scores: SIG represents speech quality and primarily reflects overall signal distortion; BAK measures the perceptual intensity of background noise, with higher scores indicating better noise suppression; and OVRL provides an overall rating, indicating the general naturalness of the speech.

[3]Codes and speech samples are available at: https://fliu215.github.io/UDSE/.

[4]https://github.com/descriptinc/descript-audio-codec.

[5]https://github.com/facebookresearch/denoiser.

[6]https://github.com/ruizhecao96/CMGAN.

[7]https://github.com/yxlu-0102/MP-SENet.

[8]https://github.com/line/open-universe.

[9]https://github.com/yxlu-0102/AP-BWE.

TABLE I
OBJECTIVE EVALUATION RESULTS AMONG UDSE AND BASELINES FOR DIFFERENT SE TASKS.

SE Task						Model	Objective Metrics				
DN	**DR**	**BWE**	**DC**	**PDR**	**CDR**		**NISQA↑**	**DNSMOS↑** SIG	BAK	OVRL	**UTMOS↑**
						Clean Speech	4.62	3.51	4.04	3.22	4.10
✓	✗	✗	✗	✗	✗	**DEMUCS**	3.45	3.37	3.96	3.07	3.63
						CMGAN	4.51	**3.51**	4.03	**3.21**	3.99
						MP-SENet	4.56	**3.51**	**4.04**	**3.21**	**4.01**
						UniverSE++	4.44	3.46	4.03	3.17	3.88
						Genhancer	3.97	3.27	3.94	2.96	3.43
						UDSE	**4.60**	3.49	3.99	3.18	3.88
✗	✓	✗	✗	✗	✗	**DEMUCS**	1.43	2.69	3.91	2.44	1.30
						CMGAN	1.13	2.62	3.99	2.28	1.55
						MP-SENet	4.34	**3.50**	**4.06**	**3.21**	3.65
						UniverSE++	3.90	3.02	3.97	2.74	2.59
						Genhancer	2.05	2.01	3.46	1.83	1.68
						UDSE	**4.60**	**3.50**	4.01	3.19	**3.82**
✗	✗	✓	✗	✗	✗	**DEMUCS**	1.91	3.30	3.97	2.99	1.80
						CMGAN	2.57	3.35	3.98	3.03	2.94
						MP-SENet	1.59	3.26	3.95	2.94	2.38
						AP-BWE	4.04	3.45	4.01	3.15	3.17
						UniverSE++	3.79	3.39	3.99	3.08	3.50
						Genhancer	3.44	3.25	3.93	2.94	3.01
						UDSE	**4.48**	**3.49**	**4.03**	**3.19**	**3.87**
✗	✗	✗	✓	✗	✗	**DEMUCS**	3.11	3.49	4.03	3.19	3.79
						CMGAN	**4.57**	3.49	4.03	3.20	4.04
						MP-SENet	4.53	**3.51**	4.04	**3.22**	**4.08**
						UniverSE++	3.98	3.48	**4.05**	3.19	3.94
						Genhancer	3.81	3.26	3.96	2.96	3.14
						UDSE	4.49	3.50	4.03	3.21	4.03
✗	✗	✗	✗	✓	✗	**DEMUCS**	2.03	3.04	3.98	2.77	2.55
						CMGAN	4.43	3.32	3.98	3.02	3.74
						MP-SENet	**4.64**	3.48	**4.04**	3.19	**4.04**
						UniverSE++	4.59	3.44	4.03	3.15	3.95
						Genhancer	4.15	3.27	3.95	2.97	3.44
						UDSE	**4.64**	**3.52**	4.03	**3.22**	4.02
✗	✗	✗	✗	✗	✓	**DEMUCS**	1.99	3.14	3.99	2.85	2.28
						CMGAN	3.26	3.28	3.99	2.97	3.25
						MP-SENet	2.47	3.36	4.02	3.06	3.14
						UniverSE++	4.35	3.35	**4.03**	3.06	3.60
						Genhancer	4.31	3.22	3.96	2.93	3.19
						UDSE	**4.67**	**3.49**	**4.03**	**3.20**	**3.90**
✓	✓	✓	✗	✗	✗	**DEMUCS**	0.77	1.83	3.54	1.51	1.33
						CMGAN	0.85	2.30	3.71	1.94	1.38
						MP-SENet	2.33	2.79	3.80	2.46	2.20
						UniverSE++	3.89	2.95	3.97	2.68	2.51
						Genhancer	2.05	2.13	3.62	1.93	1.65
						UDSE	**4.37**	**3.46**	**4.01**	**3.16**	**3.60**
✓	✓	✗	✓	✗	✗	**DEMUCS**	0.78	1.86	3.51	1.51	1.34
						CMGAN	1.04	2.05	3.39	1.73	1.34
						MP-SENet	1.54	3.11	3.94	2.79	1.93
						UniverSE++	3.74	2.92	3.97	2.65	2.36
						Genhancer	2.00	2.11	3.62	1.92	1.62
						UDSE	**4.25**	**3.43**	**4.01**	**3.13**	**3.47**
✓	✗	✗	✗	✓	✓	**DEMUCS**	1.92	2.92	4.05	2.68	1.65
						CMGAN	2.93	3.29	3.85	2.93	2.96
						MP-SENet	2.26	3.38	4.05	3.08	2.62
						UniverSE++	4.41	3.33	4.06	3.05	3.52
						Genhancer	3.87	3.13	4.04	2.89	2.87
						UDSE	**4.57**	**3.56**	**4.09**	**3.29**	**3.96**

For the subjective evaluation, we conducted ABX preference tests on the Amazon Mechanical Turk[10] platform to compare the differences between UDSE and the baseline model with the best objective performance. In each ABX test, 20 speech samples enhanced by both compared models were randomly selected from the test set. These samples were evaluated by at least 30 native English-speaking listeners. The listeners were asked to judge which of the two speech samples in each pair had better speech quality or whether there was no preference. In addition to calculating the average preference score, we also used the p-value from a t-test to assess the statistical significance of the differences between the two models.

D. Preliminary Experimental Results

The objective and subjective evaluation results are shown in Tables I and II, respectively. To provide an intuitive visualization of the enhancement performance, we also drew the spectrograms of degraded speech, clean speech, and speech

[10]https://www.mturk.com/.

TABLE II
AVERAGE PREFERENCE SCORES (%) OF ABX SUBJECTIVE TESTS ON SPEECH QUALITY BETWEEN UDSE AND A BASELINE (I.E., MODEL B) ON DIFFERENT SE TASKS, WHERE N/P STANDS FOR "NO PREFERENCE" AND p DENOTES THE p-VALUE OF A t-TEST BETWEEN TWO MODELS. FOR TASKS **DN**, **DR**, **DC** AND **PDR**, "MODEL B" IS MP-SENET, AND FOR TASKS **CDR**, **DN+DR+BWE**, **DN+DR+DC** AND **DN+PDR+CDR**, "MODEL B" IS UNIVERSE++, WHILE FOR TASK **BWE**, "MODEL B" IS AP-BWE.

SE Task									
DN	**DR**	**BWE**	**DC**	**PDR**	**CDR**	**UDSE**	Model B	N/P	p
✓	✗	✗	✗	✗	✗	41.96	41.25	16.79	0.85
✗	✓	✗	✗	✗	✗	**56.92**	36.35	6.73	$<$ **0.01**
✗	✗	✓	✗	✗	✗	**54.60**	38.60	6.80	$<$ **0.01**
✗	✗	✗	✓	✗	✗	26.35	32.50	41.15	0.07
✗	✗	✗	✗	✓	✗	**46.83**	37.01	16.16	$<$ **0.01**
✗	✗	✗	✗	✗	✓	**79.93**	12.65	7.42	$<$ **0.01**
✓	✓	✓	✗	✗	✗	**83.52**	6.11	10.37	$<$ **0.01**
✓	✓	✗	✓	✗	✗	**86.55**	8.28	5.17	$<$ **0.01**
✓	✗	✗	✗	✓	✓	**66.67**	24.50	8.83	$<$ **0.01**

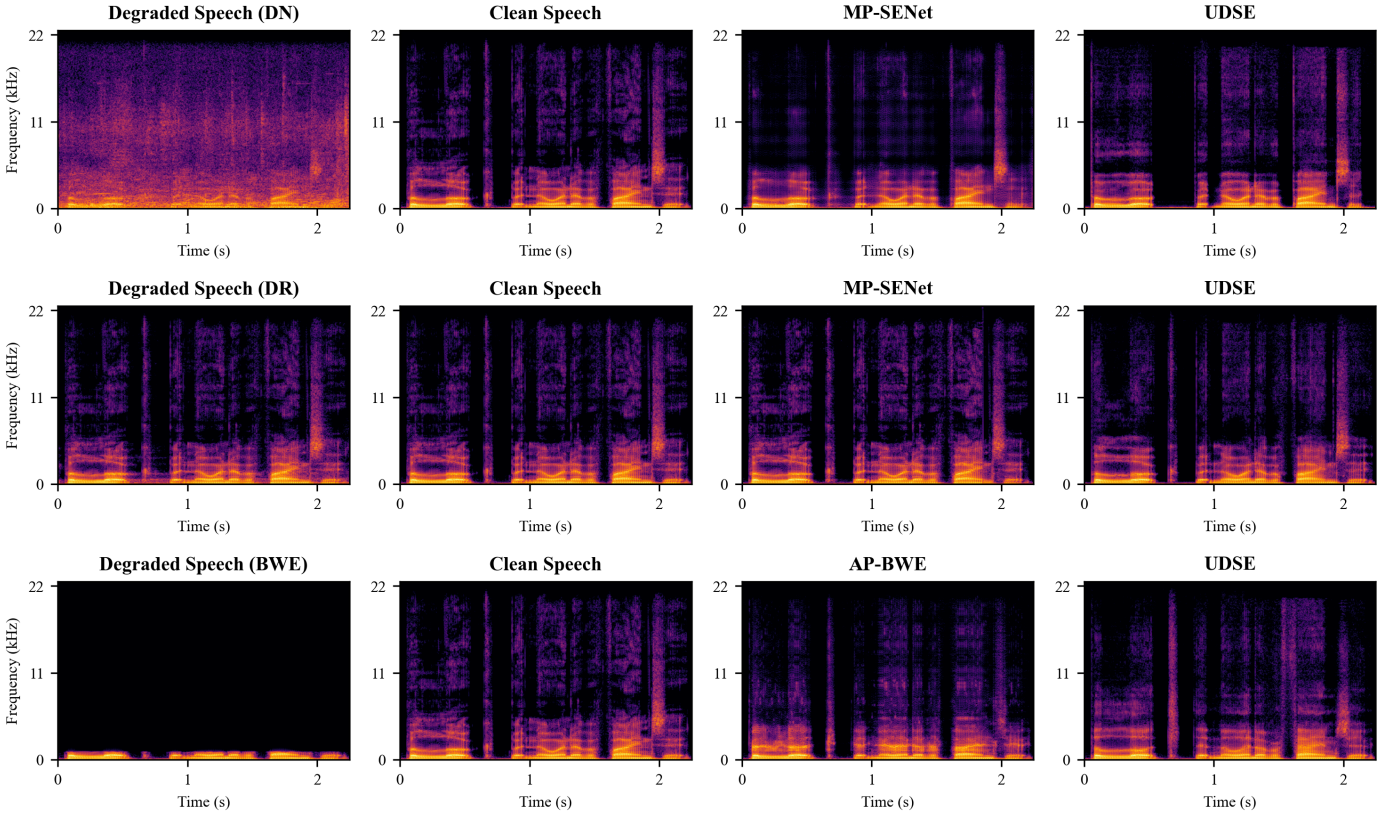

Fig. 3. Spectrogram comparison among degraded speech, clean speech, and speeches enhanced by the baseline with the best objective scores and UDSE for conventional **DN**, **DR** and **BWE** tasks, respectively.

enhanced by both the baseline model with the best objective scores and UDSE across three task types (i.e., conventional, unconventional and mixed), as shown in Figures 3, 4, and 5, respectively.

1) Experimental Results for Conventional Distortions: For the classic speech denoising task **DN**, our proposed UDSE achieved objective results comparable to the advanced continuous-domain baseline models, as shown in Table I. It can be seen that UDSE outperformed all baselines in NISQA scores and was similar to MP-SENet and CMGAN in the three indicators of DNSMOS. However, the UTMOS score of UDSE was slightly lower than that of MP-SENet and CMGAN. This may be attributed to the fact that UTMOS can only evaluate the performance of speech in the low-frequency band (0∼8 kHz). In conjunction with the fullband NISQA results, it can be inferred that the proposed UDSE possesses a stronger ability to reconstruct the high-frequency components of speech. This is also verified in the first row of spectrograms in Figure 3, where it is clearly observed that MP-SENet struggles to recover high-frequency information (8 kHz∼22.05 kHz) under low-SNR noisy conditions, while UDSE successfully preserves these components. Considering subjective perception, the ABX preference test results presented in Table II indicate no significant perceptual difference between the noisy speech enhanced by

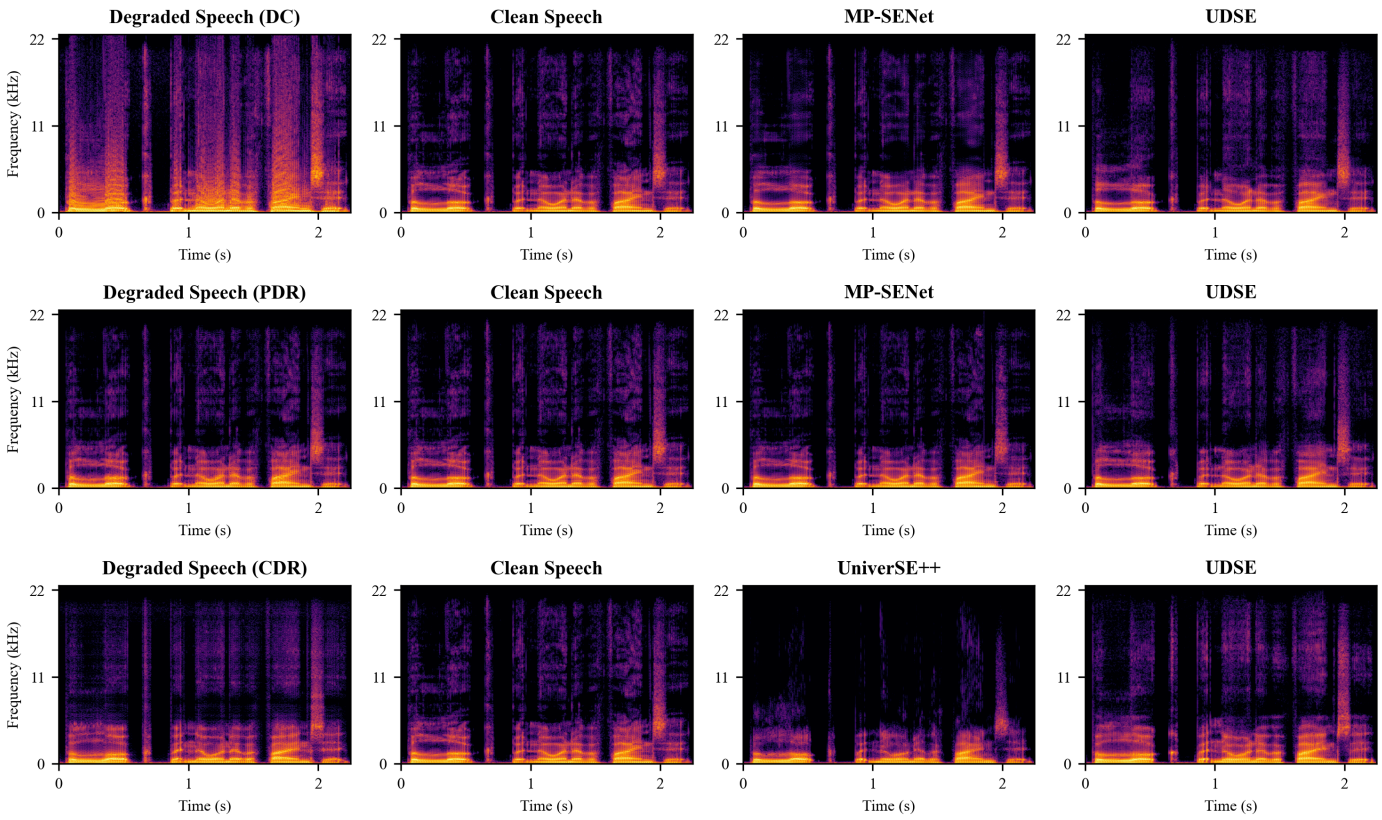

Fig. 4. Spectrogram comparison among degraded speech, clean speech, and speeches enhanced by the baseline with the best objective scores and UDSE for unconventional **DC**, **PDR** and **CDR** tasks, respectively.

UDSE and that enhanced by MP-SENet ($p = 0.85$). These findings suggest that the proposed discrete-domain UDSE model can perform on par with continuous-domain models for **DN** task.

For the classic dereverberation task **DR**, our proposed UDSE significantly outperformed the advanced continuous-domain baseline models in most objective metrics, as shown in Table I. Specifically, UDSE significantly outperformed all baselines in both NISQA and UTMOS scores, and achieved a comparable DNSMOS score to MP-SENet, effectively suppressing speech distortion caused by reverberation. It was evident that although CMGAN performed well in denoising tasks, it lagged behind in all objective metrics when dealing with dereverberation. In contrast, the proposed UDSE consistently achieved robust performance across both distortion types. The second row of Figure 3 shows the spectrograms of reverberant speech enhanced by MP-SENet and UDSE under low reverberation time conditions. Interestingly, the spectrogram of the speech enhanced by MP-SENet is more similar to that of the clean reference speech. This may be attributed to the fact that regression-based continuous-domain SE models directly model the speech waveform or features. In particular, MP-SENet explicitly predicts the amplitude spectra, which makes the enhanced speech more "similar" to the reference speech, especially when the distortion is not severe (e.g., under high SNR or low reverberation time conditions). In contrast, classification-based discrete-domain models indirectly predict discrete tokens of speech and then decode them into speech waveforms, which may result in enhanced speech that is "not similar" to the reference speech. This is also why intrusive indicators that rely on reference speech are not suitable for evaluating discrete-domain models, as stated in Section IV-C. At this point, subjective evaluation becomes the gold standard for assessing speech quality. As shown in the results in Table II, this lack of "similarity" does not negatively impact perceptual quality — reverberant speech enhanced by UDSE is even significantly better than that enhanced by MP-SENet ($p < 0.01$).

For the bandwidth extension task **BWE**, as suggested in Table I, DEMUCS, CMGAN, MP-SENet and Genhancer, which were specifically designed for denoising and reverberation, all failed. The more general-purpose UniverSE++ also performed poorly. Although the objective results of AP-BWE, which was specifically designed for task **BWE**, outperformed the five models mentioned above, there is still a significant gap compared to our proposed UDSE. Subjectively, UDSE also significantly outperformed AP-BWE ($p < 0.01$) as shown in Table II. According to listener feedback, the speech extended by AP-BWE contained noticeable pronunciation errors, whereas UDSE did not. The last row of Figure 3 intuitively shows the results of bandwidth extension. It is evident that UDSE achieves a richer and more accurate reconstruction of high-frequency speech details compared to AP-BWE. For example, within the time range of 1 to 1.2

seconds and the frequency band of 1 kHz to 6 kHz, the harmonic details of the speech extended by UDSE are more similar to those of the clean speech. This indicates that UDSE has potential in speech bandwidth extension for extremely narrow bands (e.g., from 2 kHz to 44.1 kHz, representing an expansion of over 22 times in bandwidth).

2) Experimental Results for Unconventional Distortions: For the declipping task **DC**, the results in Table I showed that for this type of distortion, both our proposed UDSE and the baseline models achieved satisfactory performance, with only minor differences in the scores across various objective metrics. As demonstrated in the first row of Figure 4, the observation is consistent with the **DR** task: UDSE produces declipped speech that is slightly less similar to the reference than MP-SENet. Nevertheless, subjective evaluation results presented in Table II reveal no significant perceptual difference between the two approaches ($p = 0.07$). This indicates that clipping is a relatively simple type of distortion, which can be effectively addressed by most SE approaches.

For the phase distortion restoration task **PDR**, UDSE achieved objective results comparable to those of MP-SENet, while outperforming DEMUCS, CMGAN, UniverSE++ and Genhancer, as shown in Table I. MP-SENet, by introducing explicit phase prediction, was more suitable for phase distortion recovery compared to DEMUCS, CMGAN, UniverSE++ and Genhancer. As shown in the second row of Figure 4, speech affected by phase distortion exhibited prominent horizontal stripes — perceptually experienced as an irritating buzzing sound. In contrast, the spectrogram of the UDSE-enhanced speech showed no such patterns, indicating that UDSE effectively eliminated interference caused by phase distortion. Since phase restoration effects are difficult to visualize clearly in spectrograms, subjective evaluation is necessary to provide additional evidence. As shown in the subjective evaluation results in Table II, UDSE-enhanced speech was generally preferred by listeners ($p < 0.01$).

For the compression distortion restoration task **CDR**, our proposed UDSE significantly outperformed all baseline models in both objective and subjective aspects ($p < 0.01$), as shown in Tables I and II. Interestingly, we found that these baseline models even had a negative effect on the enhancement of compressed speech. As clearly shown in the last row of Figure 4, UDSE effectively restored spectral details, with the enhanced harmonics closely resembling those of the clean speech. Although the speech enhanced by UniverSE++ showed decent restoration of the low-frequency harmonics, it lost many spectral details in the high-frequency regions, which significantly degraded the listening experience. These results suggest that baseline SE models are less effective in handling codec-induced compression distortions, likely due to the strong correlation between compressed and clean speech signals. In comparison, the proposed UDSE tackles this challenge from a discrete-domain perspective.

3) Experimental Results for Mixed Distortions: In the three mixed distortion restoration tasks, our proposed UDSE had a significant advantage over all baseline models in both objective and subjective aspects ($p < 0.01$), as shown in Tables I and II. For task **DN+DR+BWE**, UDSE outperformed UniverSE++ by more than 1-point in the UTMOS score. In addition, for task **DN+DR+DC**, UDSE achieved a 0.4-point higher OVRL score in DNSMOS and a 1.1-point higher UTMOS score compared to UniverSE++. Consistently, for task **DN+PDR+CDR**, UDSE again exceeded UniverSE++ by over 0.4-point in the UTMOS metric. Subjective evaluations showed that, in all three mixed-distortion tasks, the number of participants preferring UDSE-enhanced speech was at least double that of those favoring UniverSE++ — highlighting UDSE's superior performance in managing complex distortions.

The spectrograms shown in Figure 5 further provide a clear visual illustration of the effectiveness of the proposed UDSE in enhancing speech with mixed distortions. We can observe that for the **DN+DR+BWE** task (the first row in Figure 5), UDSE recovered the high-frequency components much more accurately, bringing the result closer to the ground-truth clean speech, while UniverSE++ struggled with recovering the high-frequency part. For the **DN+DR+DC** task (the second row in Figure 5), the speech enhanced by UniverSE++ lost a lot of detail in the high-frequency range and had a noticeable fundamental frequency error around 1 second, resulting in a poor listening experience. For the **DN+PDR+CDR** task (the last row in Figure 5), the spectrogram of the speech recovered by UDSE retained more details, while UniverSE++ produced significant high-frequency loss and poor spectral detail recovery, significantly impacting the listening experience. All the above experimental results demonstrate that the proposed UDSE can effectively mitigate various types of distortion in degraded speech, making it more suitable for real-world applications.

In summary, our proposed UDSE can not only handle various single distortions but also effectively address mixed distortions. This confirms the universality of UDSE and the advantages of using discrete-domain classification solutions for SE, compared to regression-based continuous-domain models.

E. Analysis and Discussion

To further investigate the contribution of each component in the proposed UDSE model, we conducted a series of analytical experiments. For simplicity, the experiments were conducted under a commonly encountered mixed distortion condition, i.e., **DN+DR+BWE** task. To maintain a concise evaluation process, only subjective preference listening tests were carried out to assess the perceptual differences between UDSE and its variants. The results of the subjective evaluation are presented in Table III.

1) Backbone Network Selection Analysis: In the UDSE framework, we utilized the Conformer block as the core backbone network to support both feature processing and token predictions. To investigate the impact of different backbone networks on UDSE's performance, we replaced the Conformer blocks with Transformer blocks (denoted as UDSE w/ Transformer), which also incorporated multi-head self-attention. The results in Table III clearly indicate that UDSE with a Transformer backbone performed significantly worse than the original UDSE ($p < 0.01$). This may be attributed to the fact that the Transformer only models global dependencies,

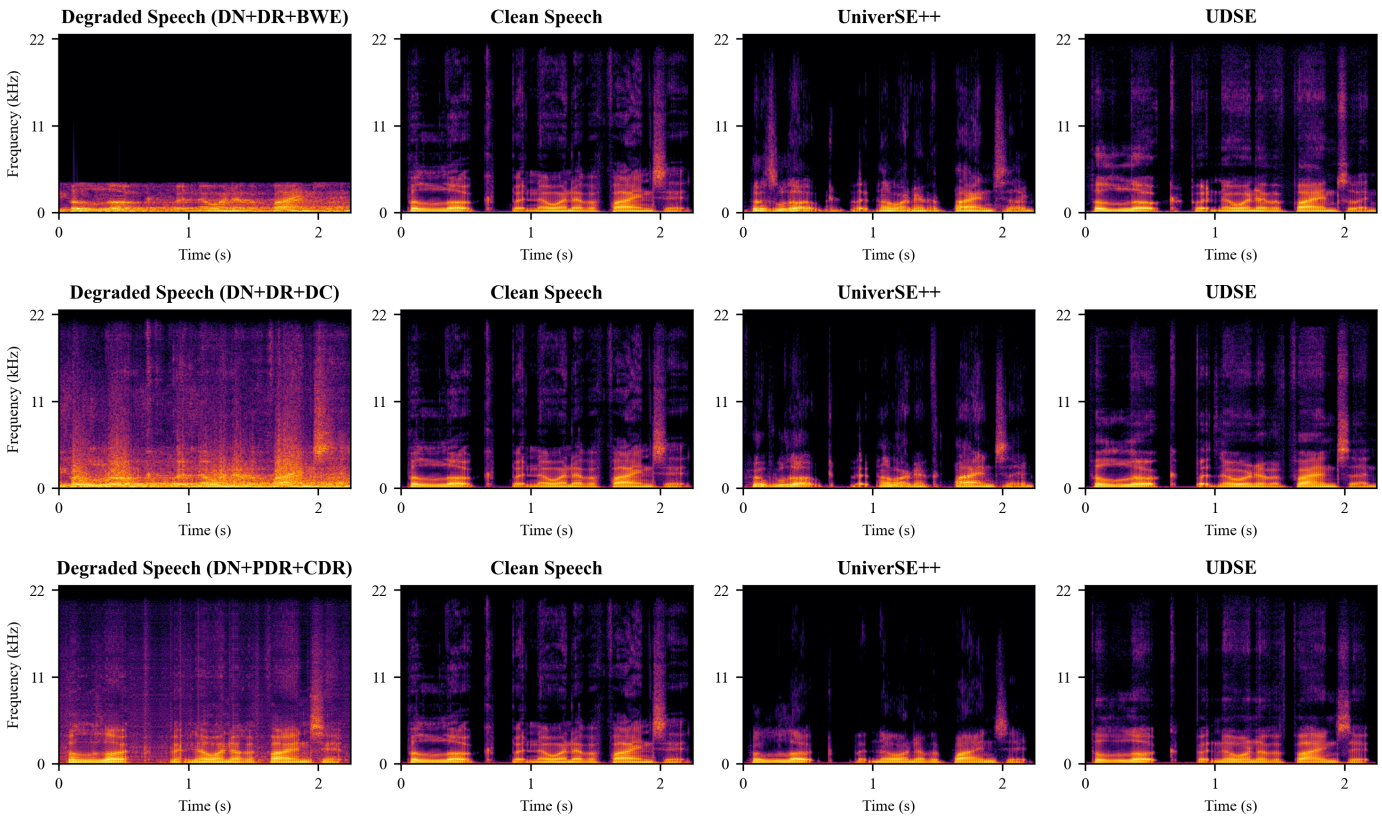

Fig. 5. Spectrogram comparison among degraded speech, clean speech, and speeches enhanced by the baseline with the best objective scores and UDSE for mixed **DN+DR+BWE**, **DN+DR+DC** and **DN+PDR+CDR** tasks, respectively.

TABLE III
AVERAGE PREFERENCE SCORES (%) OF ABX SUBJECTIVE TESTS ON SPEECH QUALITY BETWEEN UDSE AND ITS VARIANTS ON **DN+DR+BWE** TASK, WHERE N/P STANDS FOR "NO PREFERENCE" AND p DENOTES THE p-VALUE OF A t-TEST BETWEEN TWO MODELS.

UDSE	UDSE w/ Transformer	UDSE w/ Parallel Mode	UDSE w/o Global Condition	UDSE w/ MDCTCodec	N/P	p
73.00	25.33	-	-	-	1.67	$<$ **0.01**
74.83	-	24.31	-	-	0.86	$<$ **0.01**
58.00	-	-	40.00	-	2.00	$<$ **0.01**
47.87	-	-	-	49.50	2.63	0.64

whereas the Conformer enhances representational capacity by integrating convolutional layers for local feature extraction with self-attention mechanisms for capturing global context. Therefore, the Conformer, which captures both local and global features, is likely more suitable for our token prediction task than the Transformer. This advantage in representational capacity is the primary reason we selected it as the backbone network in UDSE.

2) Token Prediction Mode Analysis: In the UDSE framework, we proposed a novel token prediction mode based on the RVQ quantization rule, where clean tokens quantized by each VQ are predicted sequentially, with each prediction conditioned on all preceding prediction results. To evaluate the effectiveness of this token prediction mode, we conducted an ablation study in which the prediction of the current VQ's clean token was performed without relying on the outputs of previously predicted tokens. In this case, the clean tokens quantized by VQs were predicted in parallel by token predictors $\phi_{TP_1}, \ldots, \phi_{TP_N}$, with no interconnections among them (denoted as UDSE w/ Parallel Mode). The results in Table III show that UDSE significantly outperformed UDSE w/ Parallel Mode ($p < 0.01$). This performance gap indicates that the RVQ-based sequential prediction mode used in UDSE is beneficial, as it effectively leverages the correlations among different VQ stages. The parallel prediction mode, by ignoring these dependencies, leads to a notable degradation in enhancement performance. These experimental findings validate the effectiveness of the proposed sequential token prediction mode based on the RVQ rule.

3) Global Condition Analysis: In the UDSE framework, to reduce the difficulty of token prediction, global features extracted by a dedicated module are used as global conditioning for the prediction of each VQ. Given the characteristics of RVQ, the information provided to the first token predictor is naturally propagated to subsequent predictors. This raises an important question: can the global features be effectively treated as local, such that they only need to be provided to the first token predictor? To verify this, we constructed a

variant of UDSE (denoted as UDSE w/o Global Condition) in which the global features are provided only to the first token predictor ϕ_{TP_1} (i.e., removing $\boldsymbol{G}^{(y)}$ from the second line of Equation (10)). The results in Table III show that removing the global conditioning leads to a significant decline ($p < 0.01$) in the subjective quality of the enhanced speech produced by UDSE. This indicates that, although the first token predictor is conditioned on features extracted from the degraded speech, these features become increasingly diluted as they propagate through the network, ultimately reducing the effectiveness of subsequent token predictors. Providing global conditioning features explicitly to each token predictor is essential for preserving their prediction capabilities, and plays a critical role in ensuring the overall performance of UDSE.

4) Codec Generalization Analysis: In the UDSE framework, the neural speech codec is a crucial component, serving multiple roles including extracting global feature condition, providing clean token targets for training, and decoding the final speech output. In our experiments, we adopted DAC [44] as the neural speech codec. The coding results of clean speech from DAC serve as the upper bound for the enhancement quality achievable by UDSE. Theoretically, UDSE can also work with other RVQ-based neural speech codecs that offer comparable coding quality. To verify the generalization ability of UDSE across different neural speech codecs, we replaced the DAC with a more lightweight RVQ-based MDCTCodec [47] in UDSE (denoted by UDSE w/ MDCTCodec). The results in Table III showed no significant difference between UDSE and UDSE w/ MDCTCodec ($p = 0.64$), indicating that replacing DAC with MDCTCodec did not obviously affect the performance of UDSE. This confirms that UDSE exhibits strong codec generalization to different RVQ-based codecs, laying a solid foundation for future efforts to reduce the complexity of UDSE by exploring more lightweight codec alternatives.

V. Conclusion

This paper proposed UDSE, a novel classification-based discrete-domain general universal SE model. Unlike conventional regression-based continuous-domain SE models that predict continuous features/waveforms, UDSE formulates the SE task as a classification problem over clean acoustic tokens quantized by a neural speech codec. The UDSE advances discrete-domain SE by introducing a self-contained framework that does not require any textual cues, semantic labels, or support from LLMs. The UDSE begins with a randomly initialized token sequence, extracts global feature conditions from degraded speech, and sequentially predicts clean tokens quantized by each VQ of a neural speech codec following the RVQ rule, where each prediction depends on the previous ones. The final clean speech is reconstructed by decoding all predicted tokens using the decoder of the neural speech codec. During training, UDSE is optimized using teacher forcing strategy with a cross-entropy classification loss. Both objective and subjective experimental results demonstrate that the proposed UDSE can effectively enhance speech affected by various single and combined distortions, confirming its strong universality and practical applicability. In future work, we will further optimize the model architecture and design to reduce computational complexity and latency, while continuing to explore enhancement performance under a wider range of distortion scenarios. In addition, we will explore applying the UDSE token prediction framework to other speech tasks, such as zero-shot TTS with LLMs, to extend its applicability.

References

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone array signal processing*. Springer Science & Business Media, 2008, vol. 1.

[2] R. G. ITU-T, "712. transmission performance characteristics of pulse code modulation channels," *International Telecommunication Union, Geneva, Switzerland*, 2001.

[3] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.

[4] L. R. Rabiner, "The impact of voice processing on modern telecommunications," *Speech Communication*, vol. 17, pp. 217–226, 1995.

[5] A. E. Mahdi and D. Picovici, "Advances in voice quality measurement in modern telecommunications," *Digital Signal Processing*, vol. 19, pp. 79–103, 2009.

[6] A. N. Ince, "Overview of voice communications and speech processing," *Digital Speech Processing: Speech Coding, Synthesis and Recognition*, pp. 1–42, 1992.

[7] P. Wang, K. Tan *et al.*, "Bridging the gap between monaural speech enhancement and recognition with distortion-independent acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 39–48, 2019.

[8] Z. Chen, S. Watanabe, H. Erdogan, and J. Hershey, "Integration of speech enhancement and recognition using long-short term memory recurrent neural network," in *Proc. Interspeech*, 2015, pp. 1–7.

[9] M. Delcroix, K. Kinoshita, T. Nakatani, S. Araki, A. Ogawa, T. Hori, S. Watanabe, M. Fujimoto, T. Yoshioka, T. Oba *et al.*, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Computer Speech & Language*, vol. 27, pp. 851–873, 2013.

[10] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

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

[12] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[13] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. ICASSP*, 2018, pp. 5069–5073.

[14] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain," in *Proc. ICASSP*, 2021, pp. 7098–7102.

[15] 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.

[16] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "Voicefixer: A unified framework for high-fidelity speech restoration," in *Proc. Interspeech*, 2022, pp. 4232–4236.

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

[18] R. Scheibler, Y. Fujita, Y. Shirahata, and T. Komatsu, "Universal score-based speech enhancement with high content preservation," in *Proc. Interspeech*, 2024, pp. 1165–1169.

[19] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. WASPAA*, 2021, pp. 166–170.

[20] 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, 2003.

[21] 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.

[22] J. Kim, M. El-Khamy, and J. Lee, "T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6649–6653.

[23] G. Yu, A. Li, C. Zheng, Y. Guo, Y. Wang, and H. Wang, "Dual-branch attention-in-attention transformer for single-channel speech enhancement," in *Proc. ICASSP*, 2022, pp. 7847–7851.

[24] S.-C. Chu, C.-H. Wu, and Y.-W. Lin, "Speech enhancement based on masking approach considering speech quality and acoustic confidence for noisy speech recognition," in *Proc. APSIPA*, 2021, pp. 536–540.

[25] S. Abdulatif, R. Cao, and B. Yang, "CMGAN: Conformer-based metric-GAN for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2477–2493, 2024.

[26] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A speech enhancement model with parallel denoising of magnitude and phase spectra," in *Proc. Interspeech*, 2023, pp. 3834–3838.

[27] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement," *Neural Networks*, p. 107562, 2025.

[28] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. Interspeech*, 2020, pp. 5036–5040.

[29] Y. Ai and Z.-H. Ling, "Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses," in *Proc. ICASSP*, 2023, pp. 1–5.

[30] Y. Ai and Z.-H. Ling, "Low-latency neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses for speech generation tasks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2283–2296, 2024.

[31] X. Wang, M. Thakker, Z. Chen, N. Kanda, S. E. Eskimez, S. Chen, M. Tang, S. Liu, J. Li, and T. Yoshioka, "SpeechX: Neural codec language model as a versatile speech transformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3355–3364, 2024.

[32] H. Xue, X. Peng, and Y. Lu, "Low-latency speech enhancement via speech token generation," in *Proc. ICASSP*, 2024, pp. 661–665.

[33] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, N. Jiang, G. Zhao, and L. Xie, "SELM: Speech enhancement using discrete tokens and language models," in *Proc. ICASSP*, 2024, pp. 11 561–11 565.

[34] J. Yao, H. Liu, C. CHEN, Y. Hu, E. Chng, and L. Xie, "Gense: Generative speech enhancement via language models using hierarchical modeling," in *Proc. ICLR*, vol. 2025, 2025, pp. 69 229–69 249.

[35] B. Kang, X. Zhu, Z. Zhang, Z. Ye, M. Liu, Z. Wang, Y. Zhu, G. Ma, J. Chen, L. Xiao, C. Weng, W. Xue, and L. Xie, "LLaSE-g1: Incentivizing generalization capability for LLaMA-based speech enhancement," in *Proc. ACL*, 2025, pp. 13 292–13 305.

[36] H. Yang, J. Su, M. Kim, and Z. Jin, "Genhancer: High-fidelity speech enhancement via generative modeling on discrete codec tokens," in *Proc. Interspeech*, 2024, pp. 1170–1174.

[37] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "Df-conformer: Integrated architecture of conv-tasnet and conformer using linear complexity self-attention for speech enhancement," in *Proc. WASPAA*, 2021, pp. 161–165.

[38] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt *et al.*, "URGENT Challenge: Universality, robustness, and generalizability for speech enhancement," in *Proc. Interspeech*, 2024, pp. 4868–4872.

[39] D. Kim, S. W. Chung, H. Han, Y. Ji, and H. G. Kang, "HD-DEMUCS: General speech restoration with heterogeneous decoders," in *Proc. Interspeech*, vol. 2023, 2023, pp. 3829–3833.

[40] D.-H. Yang, D. Kim, J.-H. Chang, J. Choi, and H.-g. Moon, "DM: Dual-path magnitude network for general speech restoration," *arXiv preprint arXiv:2409.08702*, 2024.

[41] A. A. Nair and K. Koishida, "Cascaded time+ time-frequency unet for speech enhancement: Jointly addressing clipping, codec distortions, and gaps," in *Proc. ICASSP*, 2021, pp. 7153–7157.

[42] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[43] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.

[44] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proc. NeurIPS*, vol. 36, 2023, pp. 27 980–27 993.

[45] Y. Ai, X.-H. Jiang, Y.-X. Lu, H.-P. Du, and Z.-H. Ling, "APCodec: A neural audio codec with parallel amplitude and phase spectrum encoding and Decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3256–3269, 2024.

[46] C. Gârbacea, A. van den Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with vq-vae and a wavenet decoder," in *Proc. ICASSP*, 2019, pp. 735–739.

[47] X.-H. Jiang, Y. Ai, R.-C. Zheng, H.-P. Du, Y.-X. Lu, and Z.-H. Ling, "MDCTCodec: A lightweight MDCT-Based neural audio codec towards high sampling rate and low bitrate scenarios," in *Proc. SLT*, 2024, pp. 540–547.

[48] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. ICASSP*, 2022, pp. 7402–7406.

[49] 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.

[50] 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.

[51] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[52] R. Scheibler, Y. Fujita, Y. Shirahata, and T. Komatsu, "Universal score-based speech enhancement with high content preservation," in *Proc. Interspeech 2024*, 2024, pp. 1165–1169.

[53] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. SSW*, 2016, pp. 146–152.

[54] 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*, vol. 19, 2013.

[55] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh *et al.*, "ICASSP 2023 deep noise suppression challenge," *IEEE Open Journal of Signal Processing*, 2024.

[56] Y.-X. Lu, Y. Ai, H.-P. Du, and Z.-H. Ling, "Towards high-quality and efficient speech bandwidth extension with parallel amplitude and phase prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 33, pp. 236–250, 2025.

[57] 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, pp. 2127–2131.

[58] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2022, pp. 886–890.

[59] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voice mos challenge 2022," in *Proc. Interspeech*, 2022, pp. 4521–4525.