

# On the Design of Diffusion-based Neural Speech Codecs

Pietro Foti and Andreas Brendel  
Fraunhofer IIS Erlangen (pietro.foti@iis.fraunhofer.de)

**Abstract**—Recently, neural speech codecs (NSCs) trained as generative models have shown superior performance compared to conventional codecs at low bitrates. Although most state-of-the-art NSCs are trained as Generative Adversarial Networks (GANs), Diffusion Models (DMs), a recent class of generative models, represent a promising alternative due to their superior performance in image generation relative to GANs. Consequently, DMs have been successfully applied for audio and speech coding among various other audio generation applications. However, the design of diffusion-based NSCs has not yet been explored in a systematic way. We address this by providing a comprehensive analysis of diffusion-based NSCs divided into three contributions. First, we propose a categorization based on the conditioning and output domains of the DM. This simple conceptual framework allows us to define a design space for diffusion-based NSCs and to assign a category to existing approaches in the literature. Second, we systematically investigate unexplored designs by creating and evaluating new diffusion-based NSCs within the conceptual framework. Finally, we compare the proposed models to existing GAN and DM baselines through objective metrics and subjective listening tests.

**Index Terms**—Neural Speech Coding, Diffusion Models

## I. INTRODUCTION

Neural Speech Codecs (NSCs) have been significantly advanced in recent years, offering improved audio quality and compression efficiency compared to traditional codecs, especially for low and very low bitrates. Most state-of-the-art (SOTA) NSCs [1]–[5] follow similar design patterns consisting of an end-to-end trained convolutional encoder-decoder architecture with quantization at the bottleneck. The work horse for SOTA neural speech coding is the Generative Adversarial Network (GAN) training paradigm which enjoyed great popularity, especially in the computer vision field where it has been applied to various image generation tasks [6]–[8]. Recently, Diffusion Models (DMs) surpassed GAN performance on image generation [9]. Moreover, DMs do not suffer from the well-known training issues of GANs such as mode collapse and vanishing gradients, making them an attractive alternative to GANs for generative tasks.

In the audio domain, DMs have been applied to several fields, including audio synthesis [10]–[12] and audio denoising [13]–[15]. Recently, the first diffusion-based audio and speech codecs started to emerge and showed promising results: LaDiffCodec (LDC) [16] upsamples and dequantizes low-bitrate EnCodec (EC) [2] tokens with a latent DM to produce a continuous latent that is decoded by an EC decoder pretrained

on continuous input data, i.e., without quantization. Similarly, Multi-Band Diffusion (MBD) [17] conditions on the EC latent but directly generates decoded waveforms by independently processing different frequency bands.

New speech and audio coding approaches have been proposed by combining iterative sampling methods, such as DMs or Conditional Flow Matching (CFM) models, with other advanced deep learning techniques, notably semantic embeddings: SemantiCodec [18] and MuCodec [19] target the ultra-low bitrate regime (0.3 to 1.4 kbps) by combining, semantic embeddings, DM or CFM, pretrained Variational Autoencoder (VAE) for mel-spectrogram reconstruction and vocoding. In this paper, we exclude the computationally complex approaches that employ semantic embeddings, which seem to be crucial when specifically targeting ultra-low bitrates.

Diffusion-based NSCs have not yet attracted the research interest that would be expected from the success of DMs in image and audio synthesis and are underrepresented in the literature compared to GANs. Furthermore, the design space of DM-based speech codecs has not been explored systematically: the existing diffusion-based codecs [16]–[19] are the result of arguably one of the most substantial design choices, namely the conditioning and output domains of the DM, whose impact on speech generation quality is unclear.

We systematically explore the design space of diffusion-based codecs by three contributions:

- 1) We propose a categorization based on the DM conditioning/output domain, where we consider waveform (*wav*), mel-spectrogram (*mel*) and latent space (*lat*) representations.
- 2) All possible combinations of conditioning/output domain pairs from the representations mentioned above are systematically explored, except for using *wav* as conditioning, which is infeasible for low bitrates due to the high dimensionality of time-domain signals.
- 3) We evaluate the proposed models and compare them to GAN-based and DM-based baselines.

## II. DIFFUSION MODEL-BASED SPEECH CODECS

In the following, we give a brief overview that covers the main principles of generative DMs [20] for neural speech coding. Notably, there exist two formulations of DMs: a continuous-time description based on Stochastic Differential Equations (SDEs) and a discrete-time framework based on Markov Chains which is typically referred to as Denoising

This work has been supported by the Free State of Bavaria in the DSgenAI project.

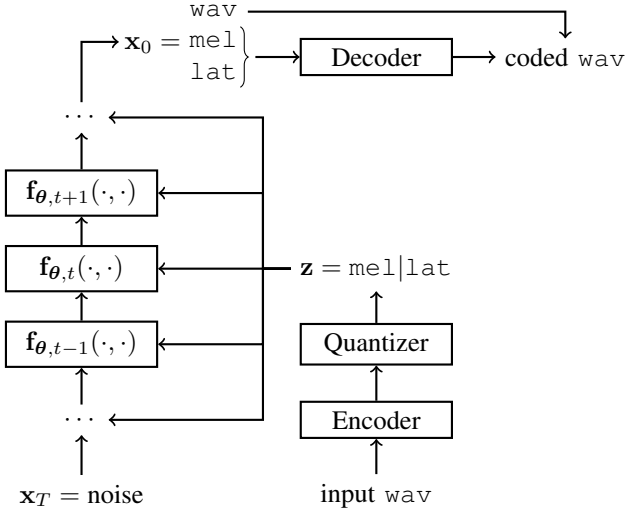


Fig. 1: Sampling scheme of the proposed DM-based NSCs. The encoder can be fixed (mel) or learned (lat).

Diffusion Probabilistic Models (DDPMs). We refer the reader to [21] and [22] for further details regarding DDPMs and SDE-based DMs, respectively.

DMs for neural speech coding model a stochastic process transforming speech samples (or derived representations such as mel-spectra or latent embeddings)  $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$  into standard Gaussian noise samples  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . This transformation, typically called the forward diffusion process, may be expressed by

$$\mathbf{x}_t = a_t \mathbf{x}_0 + b_t \epsilon, \text{ with } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1)$$

where  $t$  is a time index (not related to the speech signal but to the diffusion process), which can be discrete  $t \in \{1, \dots, T\}$  (DDPMs) or continuous  $t \in [0, T]$  (SDE-based DMs). Here,  $a_t, b_t \in \mathbb{R}_{\geq 0}$  are time-dependent coefficients chosen according to a user-defined noise schedule such that  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ .

DMs are trained to reverse the forward diffusion process by transforming samples from the standard Gaussian noise prior  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  into speech samples, i.e., samples following the data distribution  $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ . This so-called reverse diffusion process is modeled by a Deep Neural Network (DNN)-based function  $\mathbf{f}_{\theta,t}$  that is parameterized by  $\theta$  and is dependent on the time step  $t$ . The DNN  $\mathbf{f}_{\theta,t}$  is trained to estimate the speech sample  $\mathbf{x}_0$  from a noisy version of it  $\mathbf{x}_t$ . Equivalently [21], the DNN can be trained to either predict the noise added in the forward process, i.e.,  $\epsilon$  in Eq. (1), or to estimate the ‘score’ function, i.e.,  $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ . For generating new speech samples, we sample from the standard Gaussian noise prior  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and apply the following steps iteratively until  $t = 0$

$$\mathbf{x}_t \leftarrow \mathbf{f}_{\theta,t}(\mathbf{x}_t) + c_t \epsilon, \text{ with } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2)$$

$$t \leftarrow t - \Delta t. \quad (3)$$

Here,  $c_t \in \mathbb{R}_{\geq 0}$  denotes a coefficient that increases in  $t$  depending on the noise schedule and  $\Delta t$  denotes the discretiza-

tion step for the diffusion time axis ( $\Delta t = 1$  in the discrete-time case). Hence, the noisy speech sample  $\mathbf{x}_t$  is denoised step by step in the reverse diffusion process.

NSCs aim to generate speech signals that are natural, i.e., signals that follow the data distribution  $\mathbf{x}_0 \sim p_{\text{data}}(\mathbf{x})$ , while also sounding as similar as possible to the signal to be transmitted. To that end, we implement a control mechanism to guide the sampling procedure, by providing the DM with conditioning information  $\mathbf{z}$  both at training and inference time, i.e.,  $\mathbf{f}_{\theta,t}(\mathbf{x}_t, \mathbf{z})$ . As commonly done in NSCs, we employ quantization methods to learn compact discrete representations that are suitable for transmission, which will also be the basis for the mentioned conditioning information  $\mathbf{z}$ . Several NSCs [1]–[5] use Residual Vector Quantization (RVQ) [23], i.e., a cascade of vector quantizers each encoding the residual of its predecessor. However, RVQ suffers from well-known drawbacks: it requires reinitialization and decision procedures to avoid underutilized codevectors (codebook collapse), careful hyperparameter tuning, and extra training losses. Scalar Quantization (SQ) addresses the shortcomings of RVQ [24] and has been successfully applied in the image domain [25]. For training, SQ can be approximated by noise addition (‘NoiseSQ’) which simplifies training while achieving results comparable to RVQ, as shown in [26]. Due to the noise addition, training NoiseSQ end-to-end with a neural codec yields a smoother latent distribution, which is desirable for latent space modeling with generative models.

### III. EXPERIMENTAL SETUP

#### A. Model Design

We investigate the choice of the DM conditioning and output domains for diffusion-based NSCs. Waveform (wav), mel-spectrogram (mel) and latent embeddings (lat) are popular representations of speech signals and are considered as output domain choices, whereas mel and lat are examined for DM conditioning. Since NSCs require discrete conditioning information, we always assume that the mel/latent representation used for conditioning is quantized, e.g., with SQ. Based on this conceptual framework, we identify six conditioning/output configurations: mel2wav, lat2wav, mel2mel, lat2mel, mel2lat, lat2lat. To the best of our knowledge, only two out of the six model designs have been already explored in the literature: MBD [17] and LDC [16]. MBD belongs to the lat2wav category since the DM conditions on the EC latent and outputs waveforms, whereas LDC is a lat2lat model. In general, we refer to mel2wav and lat2wav as ‘waveform diffusion’ approaches, since the output of the DM is a signal in the time domain. Similarly, mel2mel and lat2mel are grouped under ‘mel diffusion’ and mel2lat and lat2lat under ‘latent diffusion’. Fig. 1 provides a schematic overview of speech generation with the proposed design setup using the terminology introduced in Sec. II. Note that the quantizer can include learnable projections, e.g., as in SQ [26].

The choice of the DM output domain has several implications, e.g., waveform diffusion is more computationally complex compared to mel or latent diffusion. The latter

approaches, on the other hand, require an additional model, a mel vocoder or a latent decoder for mel-spectra and latent, respectively. Mel diffusion provides better interpretability compared to latent diffusion and does not require a neural encoder. On the other hand, we expect a latent representation specifically learned for coding to be more powerful and thus achieve better results compared to the very generic speech representation by mel-spectra.

In this paper, the focus is primarily on investigating which choice of conditioning and output domain yields the best performance for neural speech coding. To that end, we employ well-known DM backbones from audio synthesis. In order to make the comparison as fair as possible, we use the same DNN architecture for latent and mel diffusion, which is possible due to the similar dimensionality of these speech data representations, and use a different model only for waveform diffusion. DiffWave [11], a SOTA diffusion-based vocoder, is chosen as the main building block to realize `mel2wav` and `lat2wav`. Similarly, we use GradTTS [27], which, in addition to text-to-speech, has also been applied for speech denoising [14], for `mel2mel`, `mel2lat`, `lat2mel` and `lat2lat`. Following [16], [17], we leverage EC as GAN-based baseline and to learn an expressive latent representation to be used as conditioning for the DMs. We consider BigVGAN-base [28] and HiFiGAN V1 [29] as vocoder models. Table I gives an overview of these models as they were proposed in the literature.

### B. Training and Evaluation

A general training setup applies to all the models: the models were trained for 1 million steps on clean speech signals with a fixed segment length of 1 second (convergence of the models has been confirmed for each training). The training data comprise the LibriTTS [30] and VCTK [31] datasets at 16 kHz. For each model architecture, we followed the recommended training hyperparameter choices (optimizer parameters, batch size, noise schedule, etc.) indicated in the respective publications.

An internal English test set consisting of 28 speech signals, 14 female and 14 male utterances of 8 seconds duration, was used for assessing the models' performance based on objective and subjective evaluation. For objective evaluation, we used ViSQOL [32] and SCOREQ [33], a perceptual-based and a learning-based speech quality objective metric, respectively. Details about subjective Listening Test (LT) evaluation are given in Sec. III-E.

### C. Preliminary Experiments

a) *Mel-spectrogram diffusion*: Vocoder models are essential for the DM configurations that generate mel-spectra, namely `mel2mel` and `lat2mel`. Thus, we run a set of preliminary experiments to choose the vocoder architecture and hop size. We retrained BigVGAN-base [28] and HiFiGAN V1 [29] in various configurations following the official implementation<sup>1</sup>

and found that they showed similar performance. Furthermore, larger hop sizes seem to yield more robust results when the vocoders take degraded mel-spectra as input (which represents a training test mismatch). Thus, we selected HiFiGAN V1 with a hop size of 256 as the vocoder for the following experiments.

b) *Latent diffusion*: In order to choose the quantizer model, we trained from scratch six EC models with either SQ or RVQ, at 1.5, 3 and 6 kbps. In this and all subsequent experiments, SQ was trained as an autoencoder with noise addition at the bottleneck, as for NoiseSQ in [26]. We followed the official EC model implementation<sup>2</sup> and training hyperparameters [2], except for modifying the hop size and latent dimension of the original model from 320 and 128 to 256 and 80 respectively, to match the hop size of and number of mel-bands of HiFiGAN, thereby allowing for easier comparison. The modified EC model has downsampling/upsampling ratios of [8, 4, 4, 2] and [2, 4, 4, 8] respectively. As SQ performed at least as good as RVQ for all bitrates, we chose SQ as the quantizer for the following experiments. This also allows for learning smooth latent representations for latent DMs as argued above.

Model	Framework	Input/Cond.	Out	Param. (M)	GMACs
EnCodec	GAN	wav	wav	14.42	1.66
HiFiGAN	GAN	mel	wav	13.93	19.35
BigVGAN	GAN	mel	wav	13.94	19.72
DiffWave	DM	mel	wav	2.66	41.78
GradTTS	DM	mel	mel	91.41	16.57

TABLE I: Overview of models from the literature. The complexity values (GMACs) refer to a single step forward-pass.

Design	Encoder	DM	Decoder/Vocoder
<code>mel2wav</code>	Mel	DiffWave	–
<code>lat2wav</code>	EnCodec	DiffWave	–
<code>mel2mel</code>	Mel	GradTTS	HiFiGAN
<code>lat2mel</code>	EnCodec	GradTTS	HiFiGAN
<code>mel2lat</code>	Mel	GradTTS	EnCodec
<code>lat2lat</code>	EnCodec	GradTTS	EnCodec

TABLE II: Overview of proposed designs.

### D. Experiments

a) *Exp. 1 - Evaluation of proposed designs*: We assess the impact of the DM conditioning/output design choice by evaluating the proposed diffusion-based NSC configurations. In this experiment, all DM models condition on a discrete representation quantized at 3 kbps. For `mel2wav`, `mel2mel` and `mel2lat`, the DM and SQ are trained end-to-end. We found it beneficial to use an additional reconstruction loss (sum of  $L_1$  and  $L_2$  losses) to train SQ. `lat2wav`, `lat2mel` and `lat2lat` are trained using the quantized latent embeddings of a pretrained EC model as conditioning. The jointly pretrained EC encoder and quantizer are frozen when training the DMs. For the latent diffusion configurations, the ceiling quality is determined by the decoder. Thus, we pretrain a high-bitrate EC with SQ at 8 kbps, which achieves very good speech

<sup>1</sup><https://github.com/NVIDIA/BigVGAN>

<sup>2</sup><https://github.com/facebookresearch/encodec>

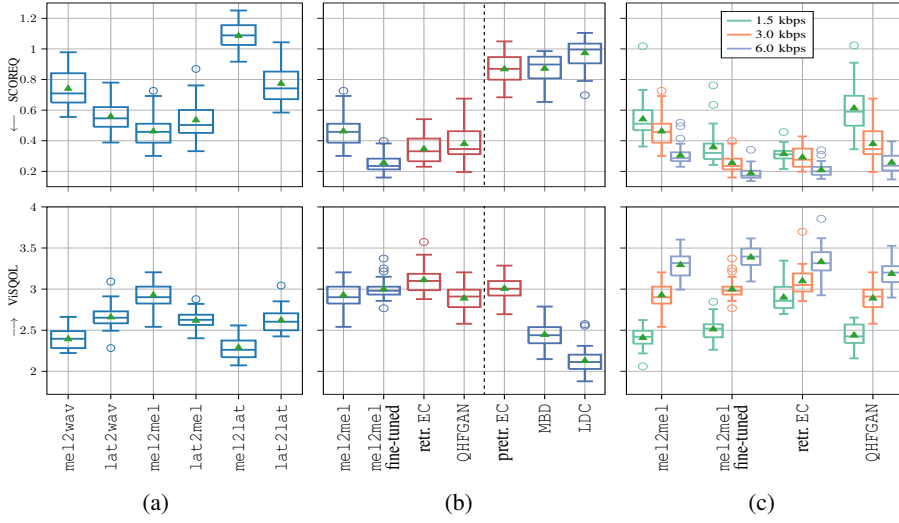


Fig. 2: Objective evaluation of Exp. 1 (a), 2 (b) and 3 (c). (b) shows retrained models on the left of the dashed line, pretrained ones on the right, with GAN and DM models colored in red and blue respectively.

quality, and subsequently train `mel2lat` and `lat2lat` to generate the latent embeddings of the high-bitrate EC. The main difference of our approach compared to LDC [16] is that the EC model, whose latent space is to be learned by the DM, is trained with SQ to enforce a smooth latent representation thereby facilitating the DM generative task.

Table II provides an overview of the proposed designs.

*b) Exp. 2 - Best proposed design vs baselines:* Here, we compare the best performing configuration from Exp. 1 to SOTA DM and GAN-based baselines at 3 kbps. The following baselines are included:

- Pretrained EC (GAN-based) and MBD (DM-based) (checkpoints and inference code available<sup>3</sup>), both trained for coding general audio at 24 kHz with variable bitrate (1.5, 3, 6 kbps for MBD, 1.5, 3, 6, 12, 24 kbps for EC). Since our models are trained only on speech for a single fixed bitrate, we expect better results than these baselines.
- Pretrained LDC, 3 kbps model trained on Librispeech [34] (clean-100) at 16 kHz (checkpoint and inference code available<sup>4</sup>).
- Retrained EC with SQ at 3 kbps (see Sec. III-C).
- Retrained HiFiGAN V1 with SQ at 3 kbps (quantizer and vocoder trained end-to-end). Similarly to `mel2wav`, `mel2mel` and `mel2lat`, SQ is trained with a reconstruction loss. We refer to this baseline as QHFGAN.

It is worth to emphasize that mel and latent diffusion are evaluated without fine-tuning the decoder/vocoder model, i.e., with a ‘non-matched’ condition. Intuitively, fine-tuning is expected to improve performance as the decoder/vocoder can learn to adapt to the input generated by the DM. Since `mel2mel` will prove to be the best performer of Exp. 1, we

<sup>3</sup><https://huggingface.co/facebook/multiband-diffusion>

<sup>4</sup><https://github.com/haiciyang/LaDiffCodec>

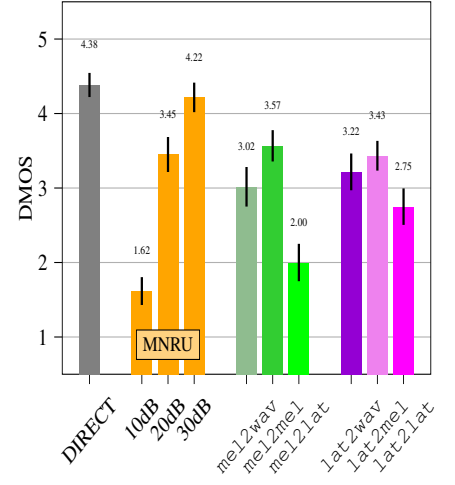


Fig. 3: P.808 DCR test results (including 15 listeners) comparing the proposed DM-based NSCs at 3 kbps.

include in the comparison the ‘matched’ condition by fine-tuning the pretrained HiFiGAN vocoder on VCTK [31].

*c) Exp. 3 - Best performers at different bitrates:* `mel2mel`, the best performer of Exp. 1, is compared to the best performing baselines of Exp. 2, namely the retrained EC and QHFGAN. All models are evaluated at 1.5, 3, and 6 kbps.

#### E. Listening Tests

To support and validate the objective metrics evaluation, we run two subjective LTs with Degradation Category Ratings (DCR) following the ITU-T P.808 principles [35] on the test set described in Sec. III-B. The first LT includes the results of Exp. 1, while the second LT comprises all results of Exp. 2 and 3. To represent the distribution of the results, we show the mean (at the top of each bar) and the 95% confidence intervals (as black vertical lines) in Figs. 3 and 4.

#### IV. EXPERIMENTAL RESULTS

In Exp. 1, the proposed diffusion-based NSC designs are compared at a bitrate of 3 kbps. Based on the objective metrics, the best performing configuration is `mel2mel` as shown in Fig. 2a. In general, SCOREQ evaluates mel diffusion as the best paradigm, followed by waveform diffusion and latent diffusion, a conclusion which is supported by the results of the first LT shown in Fig. 3, whereas ViSQOL scores show a less definite trend. Fig. 2b presents the results of Exp. 2, showing that `mel2mel` significantly outperforms the pretrained baselines, while being slightly worse than the retrained QHFGAN and EC models. However, we observe that, as expected, fine-tuning improves `mel2mel` performance, and that the fine-tuned `mel2mel` achieves better or on-par results to the retrained GAN models. Fig. 2c depicts the outcome of Exp. 3, where the best performers of Exp. 2, `mel2mel`, the retrained EC and QHFGAN, are evaluated at 1.5, 3 and 6 kbps. Overall, we find that `mel2mel` without fine-tuning

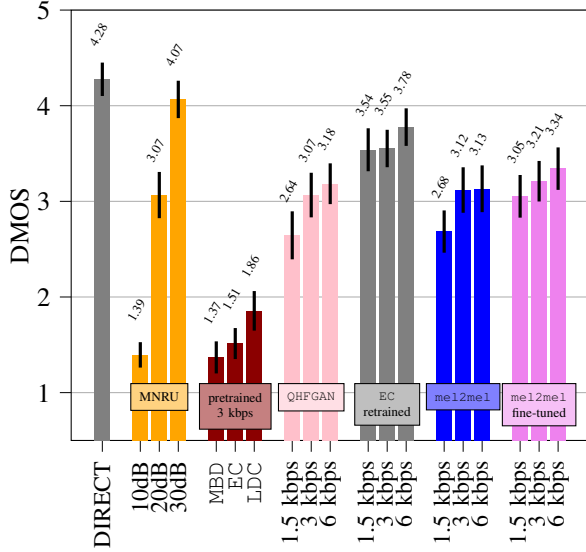


Fig. 4: P.808 DCR test results (including 19 listeners) comparing mel2mel to GAN and DM baselines.

(‘non-matched’ condition) achieves a comparable performance to QHFGAN (‘matched’ condition), but performs worse than EC. However, fine-tuning mel2mel significantly reduces the performance gap to EC, even yielding better scores for the 6 kbps models. Fig. 4 shows the outcome of a second LT that comprises the models from Exp. 2 and 3. Consistent with the objective metrics evaluation, we observe that the retrained models outperform the pretrained ones by a large margin. Moreover, mel2mel, both with and without fine-tuning is shown to achieve better or comparable ratings compared to QHFGAN. The retrained EC appears to be the best performing model, which is in line with the objective evaluation.

## V. CONCLUSION

In this paper, we explored the design space of diffusion-based NSCs by investigating which conditioning/output configuration produces the best speech quality. The proposed designs were compared to SOTA GAN and DM baselines through objective and subjective evaluation. According to our findings, the best proposed design was mel2mel, where a DM generates enhanced mel-spectra from quantized mel-spectra. mel2mel performed better than other DM-based baselines proposed in the literature, but fails to improve on the results of EC, a SOTA GAN-based codec.

## REFERENCES

- [1] N. Zeghidour et al., “Soundstream: An end-to-end neural audio codec,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2021.
- [2] A. Défossez et al., “High Fidelity Neural Audio Compression,” *Trans. on Machine Learning Research*, 2023.
- [3] R. Kumar et al., “High-fidelity audio compression with improved rvqgan,” *Advances in Neural Information Processing Systems*, 2023.
- [4] Y.-C. Wu et al., “Audiodec: An open-source streaming high-fidelity neural audio codec,” in *Inter. Conf. on Acoustics, Speech and Signal Processing*. IEEE, 2023.
- [5] N. Pia et al., “NESC: Robust Neural End-2-End Speech Coding with GANs,” *arXiv 2207.03282*, 2022.
- [6] A. Brock, J. Donahue, and K. Simonyan, “Large scale GAN training for high fidelity natural image synthesis,” in *Inter. Conf. on Learning Representations*, 2019.
- [7] T. Karras et al., “Progressive growing of gans for improved quality, stability, and variation,” in *Inter. Conf. on Learning Representations*, 2018.
- [8] —, “A style-based generator architecture for generative adversarial networks,” in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019.
- [9] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” in *Advances in Neural Information Processing Systems*, 2021.
- [10] N. Chen et al., “Wavegrad: Estimating gradients for waveform generation,” in *Inter. Conf. on Learning Representations*, 2021.
- [11] Z. Kong et al., “Diffwave: A versatile diffusion model for audio synthesis,” in *Inter. Conf. on Learning Representations*, 2021.
- [12] R. Huang et al., “Fastdiff: A fast conditional diffusion model for high-quality speech synthesis,” *arXiv 2204.09934*, 2022.
- [13] J. Serrà et al., “Universal Speech Enhancement with Score-based Diffusion,” *arXiv 2206.03065*, 2022.
- [14] Y. Tian, W. Liu, and T. Lee, “Diffusion-Based Mel-Spectrogram Enhancement for Personalized Speech Synthesis with Found Data,” in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023.
- [15] S. Welker, J. Richter, and T. Gerkmann, “Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain,” in *Interspeech*. ISCA, 2022.
- [16] H. Yang, I. Jang, and M. Kim, “Generative De-Quantization for Neural Speech Codec Via Latent Diffusion,” in *Inter. Conf. on Acoustics, Speech and Signal Processing*, 2024.
- [17] R. S. Roman et al., “From Discrete Tokens to High-Fidelity Audio Using Multi-Band Diffusion,” in *Advances in Neural Information Processing Systems*, 2023.
- [18] H. Liu et al., “SemantiCodec: An Ultra Low Bitrate Semantic Audio Codec for General Sound,” *IEEE J. of Selected Topics in Signal Processing*, 2024.
- [19] Y. Xu et al., “MuCodec: Ultra Low-Bitrate Music Codec,” *arXiv 2409.13216*, 2024.
- [20] J. Sohl-Dickstein et al., “Deep Unsupervised Learning using Nonequilibrium Thermodynamics,” in *Proc. of Inter. Conf. on Machine Learning*, 2015.
- [21] C. Luo, “Understanding diffusion models: A unified perspective,” *arXiv:2208.11970*, 2022.
- [22] Y. Song et al., “Score-based generative modeling through stochastic differential equations,” in *Inter. Conf. on Learning Representations*, 2021.
- [23] Y. Chen, T. Guan, and C. Wang, “Approximate Nearest Neighbor Search by Residual Vector Quantization,” *Sensors*, 2010.
- [24] F. Mentzer et al., “Finite Scalar Quantization: VQ-VAE Made Simple,” *arXiv 2309.15505*, 2023.
- [25] J. Ballé et al., “Nonlinear Transform Coding,” *IEEE J. of Selected Topics in Signal Processing*, 2021.
- [26] A. Brendel et al., “Neural speech coding for real-time communications using constant bitrate scalar quantization,” *IEEE J. of Selected Topics in Signal Processing*, 2024.
- [27] V. Popov et al., “Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech,” in *Proc. of Inter. Conf. on Machine Learning*, 2021.
- [28] S. Lee et al., “Bigvgan: A universal neural vocoder with large-scale training,” *arXiv 2206.04658*, 2022.
- [29] J. Kong, J. Kim, and J. Bae, “HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis,” in *Advances in Neural Information Processing Systems*, 2020.
- [30] H. Zen et al., “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech,” *arXiv 1904.02882*, 2019.
- [31] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English multi-speaker corpus for voice cloning toolkit,” 2019.
- [32] A. Hines et al., “ViSQOL: The Virtual Speech Quality Objective Listener,” in *Inter. Workshop on Acoustic Signal Enhancement*, 2012.
- [33] A. Ragano, J. Skoglund, and A. Hines, “Scoreq: Speech quality assessment with contrastive regression,” *arXiv 2410.06675*, 2024.
- [34] V. Panayotov et al., “Librispeech: An ASR corpus based on public domain audio books,” in *Inter. Conf. on Acoustics, Speech and Signal Processing*, 2015.
- [35] ITU, “Subjective evaluation of speech quality with a crowdsourcing approach,” ITU-T Recommendation P.808, 2018.