Genhancer: High-Fidelity Speech Enhancement via Generative Modeling on Discrete Codec Tokens

Haici Yang[1,2], *Jiaqi Su*[2], *Minje Kim*[3], *Zeyu Jin*[2]

[1]Indiana University, USA
[2]Adobe Research, USA
[3]University of Illinois at Urbana-Champaign, USA

hy17@iu.edu, jsu@adobe.com, minje@illinois.edu, zejin@adobe.com

Abstract

We present a high-fidelity generative speech enhancement model, Genhancer, which generates clean speech as discrete codec tokens while conditioning on the input speech features. Discrete codec tokens provide an efficient latent domain in place of the conventional time or time-frequency domain of signals, so as to enable complex modeling of speech and allow generative modeling to enforce speaker consistency and content continuity. We provide insights into the best-fit generation scheme for enhancement among parallel prediction, autoregression, and masking to demonstrate the benefits of conditioning on both pre-trained and jointly learned speech features. Subjective and objective tests show that Genhancer significantly improves audio quality and speaker-identity retention over the SOTA baselines, including conventional and generative ones while preserving content accuracy. Audio samples and supplement materials are available at https://minjekim.com/research-projects/genhancer

Index Terms: speech enhancement, discrete codec token, generative models, language models

1. Introduction

Conventional speech enhancement separates clean signal from a noisy mixture in time or time-frequency domain. Although denoising has been extensively researched, dereverberation and enhancement for other physical and digital degradation remain challenging, particularly in real-world scenarios. We attribute the limitations to three main factors. First, conventional methods consider enhancement as a many-to-one mapping problem (i.e., there is only one fixed and deterministic clean signal for each degraded signal); however, the problem becomes multimodal when dealing with reverberation and equalization, thus causing mode collapse in the conventional setups. Second, as is often modeled as a deterministic transformation from the input degraded signal, the output can hardly maintain fidelity if significant information is missing from the input due to corruption. Lastly, conventional methods are often agnostic to contextual priors such as content and speaker consistency, whereas content and speaker information is shown to benefit enhancement quality in recent generative work [1].

Therefore, this paper approaches the general speech enhancement problem from a generative perspective – we call it generative speech enhancement. The goal is to generate plausible, high-quality, clean speech that maintains perceptual consistency of speaker and spoken content with the lower-quality input recordings. If the generated speech is accurate with the speaker identity and content (velocity, prosody, and phonemes), it effectively represents a plausible version of the many possible clean signals corresponding to the degraded signal, avoiding mode collapse. In this paper, we specifically apply generative models with discrete neural codec tokens [2] as generation units in place of raw waveform or spectrogram, as they demonstrates strength in modeling modalities while allowing efficient decoding back to high-fidelity audio signals.

The idea of adopting generative models for enhancement has received increasing attention in recent years. Early attempts to incorporate elements of generative modeling include the introduction of perceptual losses [3], self-supervised learning [4], and generative adversarial networks (GAN) [5, 6, 7]. These approaches shift the focus of the optimization target from sample accuracy to perceptual similarity, relieving the requirement of reconstructing the exact signal as the ground-truth clean signal.

Researchers also investigated the use of vocoders. Several work conditions vocoders on enhanced spectral features (e.g., mel-spectrogram) to re-synthesize clean waveform [8]. The results tend to carry fewer artifacts and deliver improved consistency of phase; yet the quality is bounded by the enhancement of the spectral features. Parallel work utilizes generative processes such as variational autoencoders [9, 10] and diffusion models [11, 12] to directly shift the data distribution from the input degraded speech domain to clean speech domain; yet the task is still challenging as the data typically does not conform to simple distributions, especially when filter-like degradation such as reverberation and equalization distortion is involved.

Other generative enhancement work extracts meaningful features from input speech and reconstructs clean signals upon them. Polyak et al. [13] investigates the use of phoneme-related features, fundamental frequency, loudness level, and a global speaker embedding. Miipher [1] restores high-quality clean speech from enhanced W2v-Bert features while conditioning on speaker embedding and phoneme features. Similarly, SELM [14] regenerates the discretized WavLM features for signal synthesis. Meanwhile, DeVo [15] proposes a de-noising vocoder to avoid extra processing on the noisy SSL features. Xue et al. [16] runs an auto-regressive model on discrete codec tokens, but focusing on low latency rather than high quality.

This paper proposes *Genhancer*, a high-fidelity generative speech enhancement approach using discrete tokens of a neural codec. A generative transformer model akin to language models generates discrete tokens corresponding to clean signals while conditioning on the speech features of input signals. The contribution of this paper is as follows: 1) We propose a discrete-token-based generative enhancement approach that achieves state-of-the-art performance on various types of degraded input in terms of both sound quality and consistency with the original speaker and spoken content. 2) We extensively experiment with various types of generation schemes and conditioning features, and offer insights into the optimal design among the experimented models for speech enhancement.

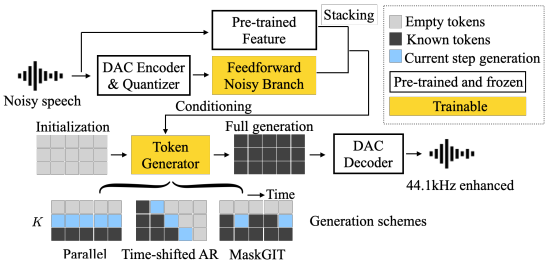

Figure 1: *Overview of Genhancer, featuring 1) generative modeling of discrete codec tokens, 2) trainable feed-forward branch to extract input features, and 3) pre-trained speech features. Generation schemes differ in generation orders and masking patterns. Black tokens are ground truth during training and generated tokens during synthesis. Blue and grey tokens are areas being masked and are agnostic to generation*

2. Generative Enhancement Model

2.1. Model Overview

Given a degraded speech input, speech enhancement aims to output a signal perceptually consistent with the original clean speech. A generative enhancement model in a discrete latent space predicts the discrete representation of clean audio conditioned on the features of the degraded input. The generated representation is then decoded into audio waveform using a pre-trained decoder. Figure 1 overviews the proposed model.

2.2. Tokenization

We use the discrete code space learned end-to-end by a high-fidelity neural audio codec, DAC [2]. It uses an autoencoder structure with the bottleneck features quantized by residual vector quantization (RVQ). During quantization, each continuous bottleneck feature vector is represented by an index that exclusively links to a D-dimensional codeword from the codebooks. We refer to the codeword-replaced D-dimensional features as discrete tokens. In particular, RVQ uses $K(K > 1)$ codebooks of M codewords each, where later codebooks take care of the coding residuals from the earlier ones. We use the discrete tokens from all the codebooks as the generation target.

2.3. Generative Modeling on Discrete Tokens

Our model runs as conditional discrete token generation, which is optimized by a cross-entropy loss under teacher-forcing setup during training. All of the K quantizers are equally considered for optimization. During inference, the generation progresses from earlier quantizers to later ones, following the hierarchy of acoustic significance in RVQ. We experiment with three generative modeling designs: parallel token prediction, auto-regression, and MaskGIT-based [17] modeling (Figure 1), differing in generation orders and masking patterns.

Parallel token predictions generates all the tokens within one quantizer in parallel. Meanwhile, causal generation across the quantizers enables the later quantizer layer to remedy any inconsistency of sampled modes in the earlier layers' generation.

Auto-regressives (AR) modeling generates tokens in temporal order. It masks the future frames and predicts the next token in a sequence based on the previously generated tokens. We use the time-shifted auto-regression pattern as in MusicGen [18]. It enables efficient prediction of K tokens in parallel by stacking the quantizers and introducing a time shift in each quantizer.

MaskGIT-based modeling circumvents the strict assumption of sequential temporal dependency and allows generation in arbitrary orders via masking. During training, it randomly masks a portion of tokens and predicts them using context from both sides. At inference time, synthesis begins by simultaneously generating all the masked tokens. A portion of the generated tokens is confirmed with high confidence according to a cosine schedule, while the rest are masked again. As this process repeats, the model iteratively refines the generation by progressively expanding the confirmed context. We follow a similar training paradigm of MaskGIT as used in SoundStorm [19], except that we introduce 20% probability of no prompt in training to adapt to speech enhancement scenarios. We apply MaskGIT within each quantizer, and move to the next quantizer once all the tokens in the current quantizer are confirmed.

The formulation of MaskGIT is particularly well-suited for speech-processing tasks. The strong conditioning from the input speech significantly reduces the number of possible modes and alleviates the need for the mode selection process to follow temporal dependencies. Instead, generation can happen parallel at multiple tokens without sacrificing consistency in the generated audio. Parallel generation further reduces hallucination of content, which often appears in auto-regressive models. Specifically for speech enhancement problem, a MaskGIT model can first confirm the generated results for input audio frames that contain relatively less degradation as well as clearer content and voice, and iteratively expand the range of enhanced frames until the whole audio is completed.

2.4. Conditioning Features

The generation of clean tokens conditions on the input audio. A simple design is to use the raw input as a condition. However, when the signal-to-noise (SNR) ratio is low, the generative model may fail to identify the speech content in the input mixture, resulting in hallucinations. Thus, we opt to provide the generative model with cleaner and more stable conditioning.

Feed-forward noisy branch. We incorporate a feed-forward branch to provide extra capacity for processing noisy input conditions. To align with the temporal resolutions of the discrete tokens, the branch takes in the stacked codewords from the neural codec computed from the input audio, i.e., of size $K \times D$. The branch serves a similar role as the feature cleaner in Miipher [1] or the conditioning networks [7, 8], which enhance the representations of the input audio. The branch is jointly learnt with the main generation branch, granting it flexibility in tailoring its output representations to the generation task.

Pre-trained speech features. While the jointly learnt noisy branch provides abundant information about the input audio, some details of the spoken content could still be obliterated due to lack of prior knowledge about speech or language. Therefore, we utilize pre-trained speech features of the input audio to guide the generative model with content understanding. We consider two types of pre-trained speech features: (1) the ASR features, which extract content-related information from the audio while being robust to noise in preparation for ASR tasks, e.g., in Whisper [20]; (2) SSL speech representations such as W2v-BERT 2.0 [21] and WavLM [22] that leverage large-scale pre-training and generalize well across speech tasks.

2.5. Model Structure

DF-Conformer blocks [23] form the backbones of both the main generation branch and the noisy branch, with demonstrated strength in speech enhancement tasks. The main branch takes in a masked sequence of token embeddings summed across

quantizers, using a learnt embedding table for each quantizer, and predicts the probability distribution of token indices for the masked locations. The main branch consists of N blocks, each containing a linear module, a layer norm, a self-attention layer, and a dilated convolution module. We grant the noisy branch with fewer blocks and a smaller channel size than the main branch, as feature cleaning and extraction typically require a smaller capacity than language modeling. The network layers use non-causal convolutions, except for the generation branch in the auto-regressive design. The pre-trained speech features are linearly interpolated to the same length as the discrete token sequence. Then, the output of the feed-forward noisy branch and the pre-trained speech features are stacked to form the final condition for the main branch. Our preliminary study experimented with different ways of merging conditions with the main generation branch. The results show that the auto-regressive model benefits from conditioning via cross-attention layer (added after self-attention layer with similar structure) in each DF-Conformer block in order to access the full context of the input audio, while the MaskGIT model benefits from additive conditioning at the token sequence for stronger temporal alignment. We stick to those designs in our experiments.

3. Experiments

3.1. Training setup

Training dataset. We trained the models with public datasets at 44.1k sample rate. The clean speech data contains DAPS [24] and LibriTTS-R [25] up-sampled to 44.1k sample rate by bandwidth extension [26]. The noise samples include the TAU Urban Audio-Visual Scenes 2021 dataset [27], DNS Challenge [28], Isolated Urban Sound Database [29], Wham 48k noise [30], and SFS-Static-Dataset [31]. The impulse response (IR) data includes MIT IR Survey [32], EchoThief [33] and OpenSLR28 [34]. The degraded utterances are created using a previously proposed data simulation and augmentation procedure [**?**]. We used a signal-to-noise ratio (SNR) of -10 $\sim$ 30 dB. Additionally, we applied random bandwidth limitation to the degraded signal, reducing frequencies down to 1kHz at maximum to simulate input at various sample rates.

Hyperparameters. DAC achieves high fidelity for 44.1k audio at 8kbps with a frame rate of 86Hz and 9 quantizers. We stack the 8-dimension tokens from all quantizers ($D = 8, K = 9$) for the input audio, resulting in a 72-dimension input condition to the noisy branch. For the main branch, we built a DF-Conformer containing 12 blocks, with recurring dilation rates of [1, 2, 4, 8], a channel size of 512, and 8 attention heads. We use 8 blocks and a channel size of 256 for the noisy branch. Rotary positional embedding [35] is applied to all attention layers. The codec and the pre-trained network were frozen during training. We train our models with 300k steps using a batch size of 80 $\times$ 8 seconds distributed on 8 A100 GPUs. We use AdamW optimizer with betas of (0.9, 0.95) and weight decay of 0.01, and an initial learning rate of 1e-4 that gradually decays to 1e-5 with cosine scheduling. During generation, we use a temperature of 0.1 for sampling and [16, 16, 16, 8, 8, 8, 1, 1, 1] iterations for MaskGIT across the quantizer layers.

3.2. Evaluation setup

Evaluation dataset. Our evaluations involve three sets of audio samples, covering a wide range of acoustic scenarios.

- **DAPS Real**: The DAPS Dataset [24] provides pairs of recordings of studio-quality speech re-recorded under twelve different room environments, thus capturing real-world acoustic conditions such as noise, reverberation and recording device differences. One male voice (*m10*) and one female voice (*f10*), as well as 2 minutes of the script (*script5*) are held out from training for evaluation purposes.
- **Real-world Speech Content:** We use the same set of real-world consumer-grade recordings as in HiFi-GAN-2 [7], which contains audio samples collected from TED Talks (`www.ted.com`) and VoxCeleb1 [36]. This set reflects the typical audio quality captured in speech content creation.
- **Demo**: We also collected around 100 publicly shared demo samples from the web pages of six enhancement methods, namely SELM [14], StoRm [12], UNIVERSE [8], Miipher [1], DeVo [15], and low-latency SE [16]. This evaluation set contains more challenging acoustic scenarios as well as a variety of types of degradation that are not necessarily involved in our training simulation. This allows listeners to compare our method to a range of approaches not covered by the discussion of this paper.

Comparison models. We compare with following methods:

- **HiFi-GAN-2 [7]**: A GAN-based enhancement model inspired by vocoders. Its enhancement module outputs at 16k sample rate and has $30M$ trainable parameters. We train the model on our data, and apply bandwidth extension following the original paper when high sample-rate input is available.
- **StoRm [12]**: A score-based diffusion model of $55M$ trainable parameters for signal-level synthesis at 16k sample rate. We fine-tuned the WSJ0+Chime3 checkpoint[1] on our data.
- **Miipher [1]**: Re-synthesis of enhanced SSL speech features conditioned on speaker embedding and optionally on text transcription. Note that SSL features in Miipher play a different role from in our model, where we use SSL features to condition generation. We modified an unofficial open-source implementation[2] to use DF-Conformer as the enhancement backbone following the original paper and match its model size ($122M$ trainable parameters) to our model. There are still two differences from the original paper due to limitations in implementation: (1) WavLM-Large[3] for SSL speech features in place of W2v-BERT; (2) HiFi-GAN [37] for vocoding in place of WaveFit [38], resulting in 22.05k output in our version. We trained Miipher on our data using the transcription-free setup for a fair comparison.

Ablation study. We consider eleven variants of our proposed models to investigate the following three design aspects: 1) generative modeling strategies, i.e., parallel token prediction (*FFDiscrete*), auto-regression (*AR*), *MaskGIT*; 2) pre-trained speech features: the output of the audio encoder in Whisper-medium[4] (*WP*), the last hidden layer of W2v-BERT2.0[5] (*WB*), a learnt weighted sum of all the hidden layers from WavLM-Large (*WL*); 3) the role of conditioning features: *NoNoisy* – no use of the input condition at all, *NoDual* – using the raw representation of the input without the noisy branch. We carry out most of the ablations on MaskGIT and WP as they are efficient for training. Some variants are not included for DAPS evaluation due to limitation of resources.

Evaluation metrics. We consider three axes of quality in evaluations: content accuracy, sound quality, and speaker similarity.

[1] https://github.com/sp-uhh/storm
[2] https://github.com/Wataru-Nakata/miipher
[3] https://huggingface.co/microsoft/wavlm-large
[4] https://huggingface.co/openai/whisper-medium
[5] https://huggingface.co/facebook/w2v-bert-2.0

	Variants	WER %	S-MOS	Q-MOS			Avg Q-MOS
		DAPS	*DEMO* (16k)	*DAPS* (44.1k)	*AQECC* (16k)	*DEMO* (16k)	
Input	—	4.1	4.14 / 0.113	1.74 / 0.042	2.67 / 0.059	1.88 / 0.051	2.081
Clean	—	–	4.45 / 0.07	4.3 / 0.038	—	—	—
StoRm	—	8.6	2.95 / 0.109	2.51 / 0.055	3.43 / 0.056	2.56 / 0.063	2.814
HiFi-GAN-2	—	5.9	3.59 / 0.101	3.63 / 0.049	3.61 / 0.05	3.32 / 0.059	3.517
Miipher	—	6.1	3.25 / 0.097	3.3 / 0.053	3.40 / 0.053	3.13 / 0.057	3.272
AR	WP	6.2	3.72 / 0.095	3.97 / 0.052	4.0 / 0.048	3.59 / 0.058	3.843
	WL	5.4	3.72 / 0.099	—	4.01 / 0.046	3.74 / 0.053	—
MaskGIT	WP	5.9	3.97 / 0.087	4.02 / 0.046	3.86 / 0.044	**3.89 / 0.053**	**3.924**
	WL	5.8	3.8 / 0.098	3.97 / 0.05	4.0 / 0.045	3.76 / 0.056	3.908
	WB	6.3	3.94 / 0.087	**4.03 / 0.05**	3.96 / 0.044	3.76 / 0.053	3.912
	WP + NoDual	6.7	**3.99 / 0.087**	4.0 / 0.047	3.98 / 0.044	3.76 / 0.054	3.909
	WL + NoNoisy	7.0	3.6 / 0.104	3.89 / 0.053	3.92 / 0.046	3.66 / 0.057	3.819
FFDiscrete	WP	5.9	3.78 / 0.096	3.91 / 0.049	**4.07 / 0.043**	3.73 / 0.054	3.899

Table 1: *Mean and standard deviation of word error rate (WER), quality MOS score (Q-MOS), and similarity MOS score (S-MOS) reported on 3 evaluation sets, as well as Avg Q-MOS averaging over the three sets. The sample rates are noted by the side of the dataset names. All the models output at their maximum possible sample rates and down-sample to the dataset's sample rate when needed.*

- **Word Error Rate (WER)**: We use Whipser-Large-v3[6] to transcribe the reference clean signals and the enhanced signals and compute WER on the DAPS Real set. Note that Whisper is designed to be noise-robust, so this score helps to identify if there is significant content hallucination while disregarding sound quality.
- **Quality mean-opinion-score (Q-MOS):** We conducted listening tests on sound quality for the three evaluation sets respectively using Prolific [39]. In each question, a listener rates the sound quality of an audio sample in terms of cleanness and naturalness on a scale from 1 (Bad) to 5 (Excellent) with a fixed clean reference provided. For each evaluation set, we recruited 150 unique workers at $16/hr, with each method condition receiving around 400 ratings. Specifically, we include 16k and 22k down-sampled versions of our model results (*MaskGIT + WP*) for the DAPS Real set to compare with baselines at the same sample rates.
- **Similarity mean-opinion-score (S-MOS)** We carved out a subset of 53 samples from the Demo set that provides clean reference, and conducted a MOS test on speaker similarity using Prolific. In each question, a subject listens to both a reference audio sample of the input quality and an enhanced audio sample by one of the method conditions, and rates how well the speaker's voice is preserved on a scale from 1 (Bad) to 5 (Excellent). We recruited 107 unique workers, with each method condition receiving around 150 ratings.

3.3. Results

Full-bandwidth generation quality. The proposed models outperform the baselines across all the evaluation sets, and achieve both improved audio quality and speaker voice preservation. Specifically on the DAPS Real set, our 16k down-sampled version of results is rated 3.50 for Q-MOS, which is considerably better than the baselines at their sample rates. Meanwhile, we observe that the 44.1k model achieves an improvement of 0.52 in Q-MOS from its 16k correspondence, showing effective full-bandwidth generation of our models. The most capable model among all, considering the three quality axes, incorporates our complete proposed designs: *MaskGIT + WP* – MaskGIT conditioning on a feedforward noisy branch processed input together with Whisper features.

Conditioning features. *NoNoisy*-model, conditioned purely on WavLM features, is considerably worse than other models in both Q-MOS and S-MOS. This shows input condition is necessary in faithfully recovering the spoken content of input speech. Yet it can still generate valid speech, showing the strong guidance pre-trained SSL features provide. This claim is also supported by a side observation we made where a model variant of (*MaskGIT + WP*) without using pre-trained SSL features scores much worse in WER (i.e., 15.3%). Meanwhile, the absence of the noisy branch harms the performance on the Demo set, showing that the noisy branch benefits content stability and sound quality particularly in scenarios where the input audio contains high entropy of noise and reverberation.

Generative modeling strategy. AR models achieve a higher quality upper bound as they deliver more consistent and pleasant-sounding results, but they can occasionally fail with hallucinated syllables and mumbled sounds in the harder cases of the Demo set. Meanwhile, MaskGIT achieves more stable performance across different model variants and datasets, i.e., higher quality-MOS on more difficult data (DAPS Real and Demo) and high S-MOS and WER universally, largely thanks to its bi-directional attention to the full context. Surprisingly, the simple paradigm of parallel token prediction *FFDiscrete* still achieved comparable evaluation scores as other models, especially on easier cases of AQECC. Based on our observation, we hypothesize that using discrete tokens already helps to prevent mode collapse compared to real-valued representations, especially when the input contains low noise and reverberation.

Pre-trained speech features. Subjective test results (Q-MOS and S-MOS) don't clearly indicate a preference for the different pre-trained features. However, we noticed that W2v-BERT and WavLM tend to retain more nuance in speech content, which aligns with the WER score. Since their pre-training was not designed for severe acoustic conditions, they show less robustness to environments in comparison to the Whisper features.

4. Conclusion

In this work, we propose Genhancer, a high-fidelity generative enhancement model based on discrete codec tokens. The use of discrete tokens enables modeling complex distribution of speech and facilitates generative modeling that enforces context priors of speaker voice consistency and content continuity in the generated clean audio. We investigated multiple generation strategies and conditioning designs in the speech enhancement setting. Through objective and subjective evaluations, we show that Genhancer constantly outputs audio with superior sound quality and speaker-identity retention than the STOA benchmarks, while preserving content accuracy.

[6]https://github.com/openai/whisper

5. References

[1] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," *arXiv preprint arXiv:2303.01664*, 2023.

[2] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[3] J. Su, A. Finkelstein, and Z. Jin, "Perceptually-motivated environment-specific speech enhancement," in *ICASSP*, 2019, pp. 7015–7019.

[4] A. Sivaraman and M. Kim, "Efficient personalized speech enhancement through self-supervised learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1342–1356, 2022.

[5] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[6] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020.

[7] ——, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *WASPAA 2021*, Oct. 2021.

[8] 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.

[9] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *ICASSP*. IEEE, 2021, pp. 676–680.

[10] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.

[11] H. Yen, F. G. Germain, G. Wichern, and J. L. Roux, "Cold diffusion for speech enhancement," in *ICASSP*, 2023, pp. 1–5.

[12] 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.

[13] A. Polyak, L. Wolf, Y. Adi, O. Kabeli, and Y. Taigman, "High fidelity speech regeneration with application to speech enhancement," in *ICASSP*, 2021, pp. 7143–7147.

[14] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, N. Jiang, G. Zhao, and L. Xie, "SELM: Speech enhancement using discrete tokens and language models," *arXiv preprint arXiv:2312.09747*, 2023.

[15] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *ICASSP*, 2023, pp. 1–5.

[16] H. Xue, X. Peng, and Y. Lu, "Low-latency speech enhancement via speech token generation," *arXiv preprint arXiv:2310.08981*, 2023.

[17] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "MaskGIT: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 315–11 325.

[18] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[20] N. Cao, Y.-R. Lin, X. Sun, D. Lazer, S. Liu, and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.

[21] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, M. Duppenthaler, P.-A. Duquenne, B. Ellis, H. Elsahar, J. Haaheim *et al.*, "Seamless: Multilingual expressive and streaming speech translation," *arXiv preprint arXiv:2312.05187*, 2023.

[22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[23] 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 *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021, pp. 161–165.

[24] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments..." *IEEE Signal Proc. Letters*, vol. 22, no. 8, 2015.

[25] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," *arXiv preprint arXiv:2305.18802*, 2023.

[26] J. Su, Y. Wang, A. Finkelstein, and Z. Jin, "Bandwidth extension is all you need," in *ICASSP*. IEEE, 2021, pp. 696–700.

[27] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *ICASSP*, 2021, pp. 626–630.

[28] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, H. Gamper, M. Golestaneh, and R. Aichner, "ICASSP 2023 deep noise suppression challenge," in *ICASSP*, 2023.

[29] J.-R. Gloaguen, A. Can, M. Lagrange, and J.-F. Petiot, "Creation of a corpus of realistic urban sound scenes with controlled acoustic properties," in *Proceedings of Meetings on Acoustics*, vol. 30, no. 1. AIP Publishing, 2017.

[30] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, Sep. 2019.

[31] Z. Chen, X. Hu, and A. Owens, "Structure from silence: Learning scene structure from ambient sound," in *5th Annual Conference on Robot Learning*, 2021. [Online]. Available: https://openreview.net/forum?id=ht3aHpc1hUt

[32] J. Traer and J. H. McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[33] "Echothief [dataset]," http://www.echothief.com/echothief/, accessed: 2024-03-12.

[34] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *ICASSP*, 2017.

[35] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[36] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[37] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[38] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, "Wavefit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration," 2022.

[39] "Prolific," https://www.prolific.co/.