MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Investigating continuous autoregressive generative speech enhancement

Yang, Haici; Wichern, Gordon; Aihara, Ryo; Masuyama, Yoshiki; Khurana, Sameer; Germain, François G; Le Roux, Jonathan

TR2025-119 August 20, 2025

Abstract

Following the success of autoregressive (AR) language models in predicting discrete tokens, it has become common practice for autoregressive audio and speech models to use discrete to- kens generated by a neural audio codec. However, recent work has demonstrated that replacing discrete token probability modeling in an AR model with a continuous diffusion procedure can improve both model performance and efficiency for image generation. In this paper, we explore applying such a diffusion loss to replace discrete token modeling in an AR generative speech enhancement model. We explore several important design choices, including comparing standard AR models with masked AR models, and mel spectrograms with learned latents as the continuous feature representation. Our results demon- strate the potential of continuous AR speech enhancement, particularly in cases of severe noise.

Interspeech 2025

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Investigating continuous autoregressive generative speech enhancement

Haici Yang[1,2], *Gordon Wichern*[1], *Ryo Aihara*[1,3], *Yoshiki Masuyama*[1], *Sameer Khurana*[1], *François G. Germain*[1], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Indiana University, Bloomington, IN, USA
[3]Information Technology R&D Center, Mitsubishi Electric Corporation, Kanagawa, Japan

hy17@iu.edu, Aihara.Ryo@dx.MitsubishiElectric.co.jp,
{masuyama, germain, wichern, leroux}@merl.com

Abstract

Following the success of autoregressive (AR) language models in predicting discrete tokens, it has become common practice for autoregressive audio and speech models to use discrete tokens generated by a neural audio codec. However, recent work has demonstrated that replacing discrete token probability modeling in an AR model with a continuous diffusion procedure can improve both model performance and efficiency for image generation. In this paper, we explore applying such a diffusion loss to replace discrete token modeling in an AR generative speech enhancement model. We explore several important design choices, including comparing standard AR models with masked AR models, and mel spectrograms with learned latents as the continuous feature representation. Our results demonstrate the potential of continuous AR speech enhancement, particularly in cases of severe noise.

Index Terms: Speech Enhancement, Diffusion Loss, Generative Modeling

1. Introduction

Improving the quality of recorded speech from a single-channel microphone has numerous practical applications, from acting as a front-end for downstream processing tasks such as speech recognition, to improving perceptual quality for media applications, and even cleaning training data for other speech models. The range of situations where degraded speech can be repaired with high quality has increased significantly as the field of speech enhancement has evolved from classical approaches such as spectral subtraction [1] to powerful deep learning models such as time-frequency masking [2,3], spectral mapping [4], or time-domain masking [5]. However, challenges remain, especially in the case of extreme degradations where it is difficult for a discriminative model trained to always predict a single clean speech target, to handle the intrinsic uncertainty of resynthesizing clean speech from a very low-quality input.

This has led to a series of generative speech enhancement approaches that resynthesize the clean speech conditioned on either the noisy input signal or features of the noisy signal. In addition to common degradations such as additive background noise and reverberation, generative models have exhibited success restoring a much broader class of degradations, including tasks such as declipping, bandwidth extension, and removal of codec artifacts [6–8]. Most popular generative modeling techniques have been previously applied to the speech enhancement task, including generative adversarial networks (GANs) [9, 10], diffusion models [6, 8, 11–13], and flow-related techniques [14, 15]. Following the success of transformer-based autoregressive (AR) large language models (LLMs) in generating text, these types of models have also been employed for speech enhancement [16–19].

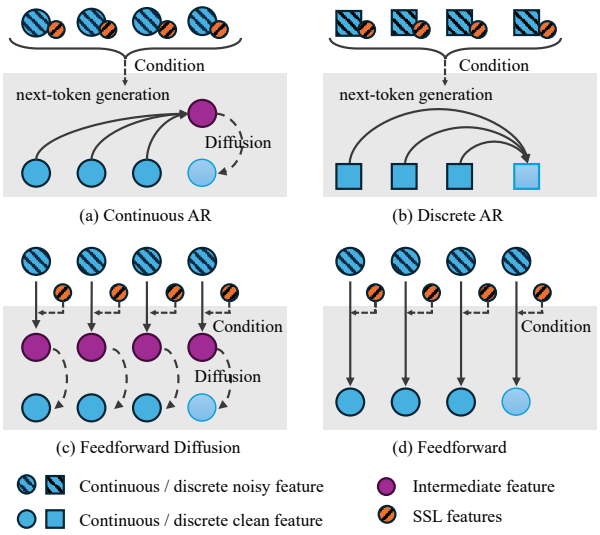


Figure 1: *Overview of various inference regimes for speech enhancement*

A typical transformer-based language model generates sequences of discrete tokens, which while quite natural for text, are a much less natural fit for continuous modalities such as audio signals. Yet, neural codecs [20–24] based on vector quantized variational autoencoders and residual vector quantization (RVQ) have become the standard for obtaining discrete representations for audio signals. While undoubtedly influential, these approaches also have significant drawbacks that may limit performance for speech applications, such as low codebook utilization, fine resolution codebooks that may learn to quantize only noise, and the non-differentiability of the vector quantization operation [25, 26].

Similar drawbacks have also been observed when using discrete representations for image generation, leading to the authors of [27] to focus on replicating two advantages of discrete representations when used for AR transformers: 1) a categorical distribution that is easy to sample from, and 2) a well-defined loss function (i.e., cross entropy). The authors then argue that these benefits can also be achieved with continuous features via a diffusion loss. As such, they propose replacing the discrete classification head of a transformer with a small diffusion model that generates the continuous feature, thus avoiding the drawbacks of discrete features. Continuous AR transformers have recently been successfully applied for speech [28–30] and music [31] synthesis along with multimodal language models [32, 33].

This paper explores the application of these recent advances

in continuous AR transformer models to the speech enhancement task. For the continuous features, we experiment with 1) predicting mel spectrograms directly and using a HiFi-GAN model to generate the speech signal, or 2) predicting continuous latent features learned by a variational autoencoder (VAE) and using a decoder to convert these features to an audio signal. In addition to the standard autoregressive next-step prediction, we also explore masked approaches that simultaneously predict multiple output tokens. Experiments demonstrate the effectiveness of continuous AR models when compared to discrete, feed-forward, and discriminative baselines.

2. Continuous autoregressive speech enhancement

We introduce here the implementation of our continuous AR speech enhancement framework, including the diffusion loss and masked autoregressive (MAR) model. As shown in Fig. 1(a), it essentially consists of an LLM-type model autoregressively running next-clean-token prediction on continuous features, with a diffusion model supervising the prediction.

2.1. Continuous tokenizer

Our continuous autoregressive speech enhancement model is built upon pre-trained tokenizers. We consider two feature domains: 1) mel spectrogram with a HiFi-GAN as vocoder; and 2) a latent feature space learned from a VAE.

Mel spectrogram: The mel spectrogram is a versatile representation that summarizes effectively the acoustic, semantic, and speaker information of an audio signal, with some relationship to human perception. More recently, however, it has been slightly losing ground to learned features in the context of generative tasks. Tasks like speech synthesis tend to favor more tailored features, such as HuBERT [34], for longer-span phoneme consistency. Recent advances in next-token prediction [26] have primarily benefited discrete audio tokens, outperforming traditional continuous representations like mel spectrograms. Nevertheless, the mel spectrogram is a very convenient feature, as it is simple to extract and encoder-free. In this work, we bring mel spectrogram features back to the next-token generation model. We compute mel-spectrogram features with 80 dimensions and a hop size of 256, leading to a 62.5 Hz frame rate. A HiFi-GAN [35] pre-trained on 16 kHz speech[1] is used to vocode the mel spectrogram. When adapted to diffusion modeling, we convert mel-spectrogram features to dB and normalize to zero mean with a range of [-1,1] for stable training.

VAE: The VAE produces bottleneck latent features with an explicit Gaussian prior and guarantees high reconstruction quality. It has been widely used as a tokenizer in latent generative models [36]. We train a speech-specialized VAE using 16 kHz speech waveforms, and design the latent features to have a frame rate of 62.5 Hz.

2.2. Diffusion loss for next-token-generation

LLM-inspired next-token-generative models have been successfully applied to intrinsically non-discrete data modalities, such as images and audio. So far, these models have been limited to discrete features because discrete features naturally lead to categorical distributions that are straightforward to define, optimize, and sample from for the purpose of probability modeling.

[1] https://huggingface.co/speechbrain/tts-hifigan-libritts-16kHz

The recently proposed diffusion loss [27] and diffusion forcing [37] de-coupled AR models from vector-quantized features. Combining the diffusion process with an AR model, the diffusion models provide tractable posterior estimation for the lower bound of the likelihood, making the AR model compatible with arbitrary continuous distributions.

Specifically, an AR model produces intermediate features z^i for the next token, $z^i = f_\theta(x^{<i})$, via a neural network parameterized by θ (a transformer in this work), where i indexes the order of the autoregression process. A forward diffusion process is introduced to add noise to the clean next token x_0^i, leading to $x_t^i = \sqrt{\bar{\alpha}_t} x_0^i + \sqrt{1-\bar{\alpha}_t}\epsilon$ at diffusion step t, where $\bar{\alpha}_t$ defines a noise schedule [38]. The intermediate feature z^i is used to condition the denoising process. We use the ϵ-parameterization loss for diffusion training, via a noise estimator ϵ_ϕ parameterized by ϕ [38]. When jointly training the diffusion process with the AR model, the diffusion loss serves as a parametric loss for the entire model:

$$\mathcal{L}(x^i, z^i) = \mathbb{E}_{\epsilon,t}(\|\epsilon - \epsilon_\phi(x_t^i | t, z^i = f_\theta(x^{<i}))\|^2). \quad (1)$$

We also adopt the training trick suggested in [27], where we randomly sample four different values of t at each training step to increase the update frequency of the diffusion model.

2.3. Generalized autoregression

A standard AR model processes input samples causally in temporal order. When using a transformer, this is implemented with causal attention. Recent masked generative methods, such as MaskGIT [39], use bi-directional attention in generative modeling by iteratively masking subsets of tokens for parallel generation. The MAR model [27] bridges these two approaches, using bi-directional attention for AR modeling. It eases the strict pre-defined order of an AR model to a randomized order during training. At inference time, an order is randomly generated at the beginning and fixed for the following sampling. In audio, both MaskGIT and MAR can be used to consider future information when making current predictions. But MAR still keeps the autoregressive manner along a certain order and avoids iterative generation. Like other masked-based models, MAR can generate more than one token for each sampling step. In this work, we experimented with both standard AR and MAR to understand the impact of bi-directional attention in speech enhancement tasks. We run 64 sampling steps for every 3-second sequence during inference time for MAR models.

2.4. Relation to stochastic regeneration method (StoRM)

A particularly effective application of diffusion models for speech enhancement is stochastic regeneration [40]. These models combine diffusion models with a discriminative predictor. As the predictor methods learn a mapping to the posterior mean, they are incapable of reproducing fine-grained details. A diffusion model applied on top of the predictor uses generative modeling to correct the bias, using a computationally heavy diffusion-based model. Both our architecture and StoRM use a diffusion model to generate the final enhanced output starting from a previous model's learned feature. However, our model is conceptually very different from StoRM. The diffusion model in this work is designed to be lightweight because it does not serve as a primary generator; instead, it supervises the preceding AR model's generation. Meanwhile, the diffusion model does not take any noisy component directly as input and thus is agnostic to the enhancement process.

3. Baselines

We separately built three different types of baseline models with the same model architecture, model size, and condition mechanism as the main model.

Discrete autoregressive model (Fig. 1(b)) is built following Genhancer [17] with BigCodec [41] as a replacement for its original tokenizer DAC [22]. Designed as a non-streamable codec, BigCodec leverages large-sized models to empower the decoder, such that it can achieve high-quality audio reconstruction with a single codebook. In contrast to its multi-codebook counterparts, BigCodec has only one token to predict for each frame. It avoids the tedious design for multi-codebook generation, and greatly improves training and synthesis efficiency. More importantly, we believe it provides a fairer comparison to the continuous features, because the required number of AR predictions remains similar.

Feedforward diffusion model (Fig. 1(c)) aims to understand the effect of AR modeling in the presence of diffusion modeling, by eliminating AR modeling from the main model. From an implementation perspective, it can be viewed as a feedforward model with diffusion loss. Compared with the common diffusion-based speech enhancement models, this implementation features heavy condition processing with a relatively lighter network within the diffusion model for score computing, akin to one of the model variants proposed in UNIVERSE [6]. Note that this baseline also differs from StoRM because this is an end-to-end trained model, and the feedforward model doesn't directly output enhanced features.

Feedforward model (Fig. 1(d)) is transformer-powered but does not involve any generative modeling. It takes noisy tokens as input and outputs clean versions in one feedforward step. In line with the conventional speech enhancement models, it handles the speech enhancement task discriminatively, endeavoring to learn noisy-to-clean feature mapping. We experimented with feedforward models on both mel spectrogram and VAE features.

We also considered the following established baselines.

ConvTasNet [5] was used as a well-developed baseline of the time-domain end-to-end approach. We follow the implementation in `Asteroid` [42], but modified the kernel size and stride for the 1D convolution encoder to 32 and 16 samples, respectively, to handle 16 kHz audio.

DCCRN [43] predicts a complex mask in the time-frequency domain and has shown promising performance on several benchmarks with a moderate model size. We again followed the implementation in `Asteroid`.

StoRm [40] runs a diffusion model to regenerate enhanced speech from the denoised output produced by a discriminative predictor. We fine-tuned the provided Voicebank-DEMAND checkpoints[2] on our datasets.

4. Experiments

4.1. Model design

Conditioning: In line with Miipher [7] and Genhancer [17], we include w2v-BERT [44] as a condition to the transformer models. The w2v-BERT feature is interpolated to the same length as other features before being attached along the channel dimension. For the base model setups, we run cross-attention conditioning for the AR models and prefix conditioning for MAR. We compare additional options in the ablation study.

Transformer: We use a decoder-only transformer with 16 attention blocks. Each block contains a 12-head self-attention layer. When conditioning with cross-attention, we reduce the number of blocks to 12. The latent units have 768 dimensions. Both type of transformers land at roughly 114M parameters.

Diffusion: We follow MAR's [27] design for the diffusion model, employing a simple MLP consisting of 3 residual blocks and a width of 1024 channels. The condition z is added to the time embedding of the diffusion step t, and conditions the MLP in the layer normalization via AdaLN [45]. This model has 22M parameters. We run 100 steps during inference.

VAE. We adapt the architecture of the audio convolution-based VAE module of Stable Audio [36][3], and re-train a speech VAE model. The downsampling rate of the 4 convolution blocks of the encoder is [2, 4, 4, 8]. The bottleneck dimension is 64.

4.2. Training and Evaluation

Dataset: The training set of the speech enhancement models combines WHAM! 16 kHz [46], Voicebank-DEMAND [47], and mixtures of LibriTTS-R [48] speech and WHAM! noise, with a signal-to-noise ratio (SNR) randomly sampled in [-10,10] dB. Each training sequence is 3 seconds long. The evaluation set includes 200 samples randomly selected from the WHAM! test set and 200 from the LibriTTS-R+WHAM! test set. The speech VAE is trained on LibriSpeech.

Training: All speech enhancement models are trained with batch size 32 on four 48GB A40 GPUs, for 120K steps. The VAE converged at around 100K steps, with a batch size of 8.

Ablations: We also experiment with several variants for the AR and MAR models, as listed in Table 2. For the standard AR models, we test 1) replacing diffusion loss with MSE loss; 2) conditioning by prefixing the noisy features to the input; 3) switching from transformer to conformer by adding one convolution block to the end of the attention layers; The kernel size of the convolution layers is 31; 4) removing SSL features in the condition. For the MAR model, we test 1) including a Masked Autoencoder (MAE)-style [49] encoder transformer, with the transformer dimension adjusted to 512 to maintain a similar model size; 2) enforcing a temporal order.

Evaluation All the metrics are calculated using the VERSA [50] toolkit. Word error rate (WER) uses whisper-base[4] for ASR and is computed on the LibriTTS-R+WHAM! test set, using the original text transcripts.

5. Results

Generative vs. feedforward models: On average, generative modeling does not necessarily lead to better speech enhancement. From Table 1, we find that MAR and feedforward models overall outperform the AR models on both continuous tokenizers. When enforcing the temporal order in the MAR model (Table 2), the performance also significantly drops for all metrics. Similarly, comparing AR base and AR w/ MSE loss, we see that replacing generative modeling with discriminative MSE loss does not impact performance. This indicates that temporal-level AR generative modeling is not effective in this model. We believe that, compared to feedforward models, AR models spend too much computational power on self-attention modeling, which could help maintain speaker and content consistency but does not directly contribute to the enhancement task.

[2] https://github.com/sp-uhh/storm

[3] https://huggingface.co/stabilityai/stable-audio-open-1.0

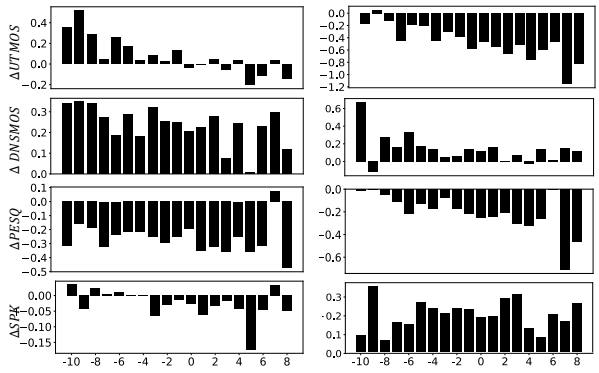
[4] https://huggingface.co/openai/whisper-base

Table 1: *Performance comparison of continuous autoregressive models and baselines.*

	Tokenizer	WER % ↓	UTMOS↑	DNSMOS↑	PESQ↑	STOI↑	SPK_SIM↑
Noisy input	—	31 ± 40	1.36 ± 0.18	2.46 ± 0.20	1.11 ± 0.20	0.73 ± 0.09	0.78 ± 0.12
Clean	—	3 ± 8	4.15 ± 0.19	3.67 ± 0.24	4.64 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
ConvTasNet [5]	—	35 ± 38	3.46 ± 0.51	3.28 ± 0.23	$\mathbf{1.91} \pm 0.37$	$\mathbf{0.93} \pm 0.06$	$\underline{0.71} \pm 0.13$
DCCRN [43]	—	43 ± 43	3.03 ± 0.56	3.35 ± 0.26	1.71 ± 0.33	$\underline{0.89} \pm 0.07$	$\mathbf{0.75} \pm 0.12$
StoRm [40]	—	52 ± 50	3.05 ± 0.61	3.60 ± 0.20	1.34 ± 0.24	0.85 ± 0.10	0.70 ± 0.13
Feedforward + MSE	mel	$\mathbf{28} \pm 29$	3.48 ± 0.40	3.58 ± 0.23	$\underline{1.88} \pm 0.33$	0.10 ± 0.07	0.67 ± 0.13
	VAE	$\underline{29} \pm 33$	3.14 ± 0.52	3.49 ± 0.27	1.57 ± 0.28	0.88 ± 0.07	0.60 ± 0.13
Feedforward + diffusion	VAE	33 ± 34	2.99 ± 0.55	3.43 ± 0.24	1.72 ± 0.30	0.88 ± 0.06	0.67 ± 0.11
AR	mel	45 ± 36	3.16 ± 0.36	3.52 ± 0.26	1.58 ± 0.28	0.11 ± 0.07	0.66 ± 0.14
	VAE	70 ± 47	2.75 ± 0.61	3.46 ± 0.25	1.56 ± 0.29	0.83 ± 0.10	0.61 ± 0.14
	BigCodec	58 ± 34	$\mathbf{4.01} \pm 0.30$	3.62 ± 0.28	1.62 ± 0.33	0.87 ± 0.09	0.41 ± 0.10
MAR	mel	34 ± 34	3.46 ± 0.30	$\mathbf{3.75} \pm 0.19$	1.54 ± 0.25	0.10 ± 0.07	0.61 ± 0.12
	VAE	34 ± 34	2.70 ± 0.58	2.79 ± 0.18	1.37 ± 0.22	0.78 ± 0.12	0.45 ± 0.14
	BigCodec	46 ± 38	$\underline{3.91} \pm 0.39$	$\underline{3.63} \pm 0.26$	1.74 ± 0.35	0.88 ± 0.07	0.39 ± 0.10

Table 2: *Ablation study around the autoregressive models.*

	WER % ↓	UTMOS ↑	DNSMOS ↑	PESQ ↑
AR	45 ± 36	3.16 ± 0.36	3.52 ± 0.26	1.58 ± 0.28
AR w/ MSE Loss	46 ± 37	3.26 ± 0.36	3.33 ± 0.25	1.63 ± 0.31
AR w/ Prefix	58 ± 37	3.04 ± 0.41	3.55 ± 0.25	1.47 ± 0.23
AR w/ Conv	54 ± 37	3.02 ± 0.41	3.57 ± 0.29	1.54 ± 0.27
AR w/o SSL	66 ± 34	2.86 ± 0.42	2.79 ± 0.17	1.38 ± 0.25
MAR	34 ± 34	3.46 ± 0.30	3.75 ± 0.19	1.54 ± 0.25
MAR w/ MAE	30 ± 30	3.54 ± 0.31	3.81 ± 0.20	1.61 ± 0.26
MAR w/ Temporal	55 ± 44	2.98 ± 0.46	3.51 ± 0.28	1.46 ± 0.22



(a) *MAR(mel) v.s. FF(mel)* (b) *MAR(mel) v.s. MAR(BC)*

Figure 2: *Score difference on UTMOS, DNSMOS, PESQ, and speaker similarity between models at different noise levels. The horizontal axis indicates the noisy speech input SI-SDR [dB]. FF stands for feedforward, and BC stands for BigCodec.*

Conversely, MAR performs similarly to the feedforward models on average and has a clear advantage on the more severely noisy cases, i.e., samples with low input SI-SDR, as shown in Fig. 2a. We can observe different preferences from each metric. In general, PESQ favors non-generative models, while DNSMOS tends to rate generative models higher. Nevertheless, all the metrics indicate the same score difference pattern along input SI-SDR. Clearly, when the input SI-SDR is lower than zero, MAR(mel) outperforms FF(mel) on most metrics. The performance discrepancy increases as the input SI-SDR further decreases. This aligns with the motivation and observation of many other generative speech enhancement models. Generative speech enhancement models are beneficial under severe conditions where traditional discriminative models fail to learn the mapping. It is also interesting to note that the MAE encoder is helpful. While we did not further explore this direction, it underlines the potential of a specialized transformer encoder in this task, especially for mel-spectrogram features.

Notably, as all generative models (including the baseline StoRm) suffer from higher WER, MAR + continuous features achieve relatively low WERs, beating ConvTasNet and DCCRN, and are close to the feedforward models. We attribute this to the bi-directional attention of MAR models.

Tokenizers Continuous features (mel spectrogram and VAE) lead to better speaker similarity, and better WER in most cases. The result on speaker similarity aligns with a comparison study between mel spectrogram and Encodec features [51], which reports that discrete features are slightly inferior in representing speaker information. BigCodec excels at output quality, as reflected by UTMOS. We believe this is because all the tokens from the BigCodec are trained from clean speech, and because only one codebook is used, not much non-speech sound can be coded in the codebook. However, on the flip side, the model can be subject to more content errors and higher WER. Mel spectrogram tends to preserve speaker features better than VAE features. Because HiFi-GAN on mel does not preserve phase information, all the models involving mel spectrogram have low STOI scores.

A similar pattern is observed from Fig. 2b, where the mel spectrogram performs better than BigCodec on the low input SNR cases. We believe this is because the continuous space is capable of more complex modeling.

6. Conclusion

In this work, we explored continuous generative speech enhancement using diffusion loss and AR models. By designing related baselines and exploring several important design choices, we show that our generative speech enhancement is particularly effective and valuable in severely noisy cases compared to the discriminative and discrete counterparts. In the future, we plan to further scale up continuous AR and MAR models for generative speech enhancement and explore additional applications such as extreme restoration tasks or more challenging use cases involving multiple speakers and languages.

Acknowledgments—We thank Jiaqi Su from Adobe Research for thoughtful advice regarding feedforward models.

7. References

[1] P. Loizou, "Speech enhancement: Theory and practice," 2007.

[2] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[4] Z.-Q. Wang *et al.*, "TF-GridNet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[5] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.

[6] 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.

[7] Y. Koizumi *et al.*, "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," in *Proc. WASPAA*, 2023.

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

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

[10] 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.

[11] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. ICASSP*, 2022.

[12] H. Yen, F. G. Germain, G. Wichern, and J. Le Roux, "Cold diffusion for speech enhancement," in *Proc. ICASSP*, 2023.

[13] J.-M. Lemercier *et al.*, "Diffusion models for audio restoration: A review," *IEEE Signal Process. Mag.*, vol. 41, no. 6, pp. 72–84, 2024.

[14] M. Strauss and B. Edler, "A flow-based neural network for time domain speech enhancement," in *Proc. ICASSP*, 2021.

[15] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, "FlowAVSE: Efficient audio-visual speech enhancement with conditional flow matching," *arXiv preprint arXiv:2406.09286*, 2024.

[16] H. Erdogan *et al.*, "TokenSplit: Using discrete speech representations for direct, refined, and transcript-conditioned speech separation and recognition," in *Proc. Interspeech*, 2023.

[17] 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.

[18] Z. Wang *et al.*, "SELM: Speech enhancement using discrete tokens and language models," in *Proc. ICASSP*, 2024, pp. 11 561–11 565.

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

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

[21] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *TMLR*, 2023.

[22] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," *Proc. NeurIPS*, vol. 36, 2024.

[23] J. Shi *et al.*, "ESPnet-Codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech," in *Proc. SLT*, 2024.

[24] H. Yang, I. Jang, and M. Kim, "Generative de-quantization for neural speech codec via latent diffusion," in *Proc. ICASSP*, 2024, pp. 1251–1255.

[25] F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite scalar quantization: VQ-VAE made simple," *arXiv preprint arXiv:2309.15505*, 2023.

[26] P. Mousavi *et al.*, "DASB–discrete audio and speech benchmark," *arXiv preprint arXiv:2406.14294*, 2024.

[27] T. Li, Y. Tian, H. Li, M. Deng, and K. He, "Autoregressive image generation without vector quantization," in *Proc. NeurIPS*, 2024.

[28] L. Meng *et al.*, "Autoregressive speech synthesis without vector quantization," *arXiv preprint arXiv:2407.08551*, 2024.

[29] A. Turetzky *et al.*, "Continuous speech synthesis using per-token latent diffusion," *arXiv preprint arXiv:2410.16048*, 2024.

[30] X. Zhu, W. Tian, and L. Xie, "Autoregressive speech synthesis with next-distribution prediction," *arXiv preprint arXiv:2412.16846*, 2024.

[31] M. Pasini, J. Nistal, S. Lattner, and G. Fazekas, "Continuous autoregressive models with noise augmentation avoid error accumulation," *arXiv preprint arXiv:2411.18447*, 2024.

[32] Z. Yuan, Y. Liu, S. Liu, and S. Zhao, "Continuous speech tokens makes LLMs robust multi-modality learners," *arXiv preprint arXiv:2412.04917*, 2024.

[33] Y. Sun *et al.*, "Multimodal latent language modeling with next-token diffusion," *arXiv preprint arXiv:2412.08635*, 2024.

[34] W.-N. Hsu *et al.*, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451 – 3460, 2021.

[35] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[36] Z. Evans *et al.*, "Stable audio open," *arXiv preprint arXiv:2407.14358*, 2024.

[37] B. Chen *et al.*, "Diffusion forcing: Next-token prediction meets full-sequence diffusion," *arXiv preprint arXiv:2407.01392*, 2024.

[38] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[39] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "MaskGIT: Masked generative image transformer," in *Proc. CVPR*, 2022, pp. 11 315–11 325.

[40] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2023.

[41] D. Xin, X. Tan, S. Takamichi, and H. Saruwatari, "BigCodec: Pushing the limits of low-bitrate neural speech codec," *arXiv preprint arXiv:2409.05377*, 2024.

[42] M. Pariente *et al.*, "Asteroid: The PyTorch-based audio source separation toolkit for researchers," in *Proc. Interspeech*, 2020.

[43] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020.

[44] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, and J. Qin, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. ASRU*, 2021.

[45] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. ICCV*, 2023.

[46] G. Wichern *et al.*, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech*, 2019.

[47] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR), doi: 10.7488/ds/2117, 2017.

[48] Y. Koizumi *et al.*, "LibriTTS-R: A restored multi-speaker text-to-speech corpus," *arXiv preprint arXiv:2305.18802*, 2023.

[49] K. He *et al.*, "Masked autoencoders are scalable vision learners," in *Proc. CVPR*. IEEE, 2022, pp. 16 000–16 009.

[50] J. Shi *et al.*, "VERSA: A versatile evaluation toolkit for speech, audio, and music," *arXiv preprint arXiv:2412.17667*, 2024.

[51] K. C. Puvvada, N. R. Koluguri, K. Dhawan, J. Balam, and B. Ginsburg, "Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition," in *Proc. ICASSP*, 2024.