Enhancing Neural Speech Embeddings for Generative Speech Models

Doyeon Kim * and Yanjue Song† and Nilesh Madhu† and Hong-Goo Kang*
* Dept. of Electrical and Electronic Eng., Yonsei University, Seoul, Korea
E-mail: ehyeon24@dsp.yonsei.ac.kr, hgkang@yonsei.ac.kr
† IDLab, Ghent University - imec, Ghent, Belgium
E-mail: Yanjue.Song, Nilesh.Madhu@ugent.be

***Abstract*—We explore a speech enhancement framework where neural speech embeddings, derived from pre-trained self-supervised learning (SSL) models applied to noisy signals, are used as inputs to a neural vocoder to generate the corresponding clean speech. The primary innovation lies in enhancing these latent neural embeddings to mitigate distortions caused by noise and reverberation, resulting in a superior quality of the synthesized signal. By dividing the process into Separate phases for embedding enhancement and speech generation, the approach allows for greater flexibility in network design. We also examine the advantage of integrating hidden states from the SSL model in a learnable manner to create a more robust embedding for the vocoder input. Additionally, we investigate various loss functions for training the neural vocoder. Experimental results confirm the effectiveness of our proposed approach, particularly in environments with simultaneous background noise and reverberation.**

***Index Terms*—embedding enhancement, speech embeddings, generative speech enhancement, automatic speech recognition, self-supervised learning**

I. Introduction

The limited availability of large labeled and paired datasets in deep learning for speech has led to a surge in the development of self-supervised learning (SSL) models [1]–[3], which are designed to extract meaningful representations from unlabeled data. Pre-trained SSL models have shown their potential as feature extractors across various tasks, allowing a small back-end network to generalize effectively with minimal training for different downstream applications, as reported in the SUPERB benchmark [4].

Several prior studies have explored the use of SSL models for speech enhancement. For instance, Zhao et al. [5] conducted a denoising task by initializing encoder and bottleneck parameters with a pre-trained WavLM model. In [6], [7], SSL features were fine-tuned specifically for speech enhancement enabling the estimation of masks from noisy SSL embeddings to recover clean speech. In general, SSL models typically transform speech input into low-dimensional embedding features. Neural vocoders, commonly used in text-to-speech (TTS) systems, generate synthesized speech waveforms from low-dimensional speech features such as log-mel spectrograms [8]–[10]. To leverage the benefits of pre-trained SSL models for generating compact representations and the neural vocoder's capabilities in speech generation, the Denoising Vocoder [11] was proposed. This approach directly synthesizes clean speech from the distorted embeddings extracted by the pre-trained HuBERT model [3] from noisy input signals. The vocoder utilizes the HiFi-GAN architecture [12] to produce a time-domain signal through a series of upsampling blocks from the embedding features.

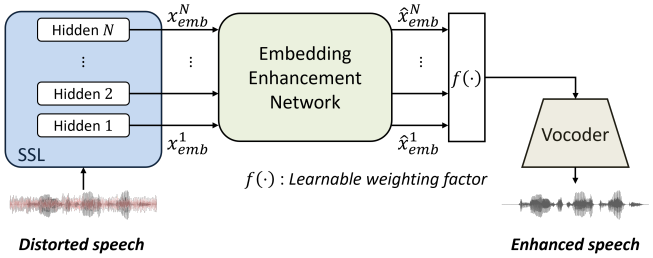


Fig. 1: Proposed framework overview

A recent study [13] explored the semantic information encoded by SSL models by tokenizing and de-tokenizing wav2vec2.0 embeddings. The HiFi-GAN vocoder was then used to synthesize the clean speech estimate from these processed embeddings. Miipher [14] also utilized SSL feature information for speech synthesis, but this model generates speech by conditioning enhanced features on a combination of the ground truth transcript and the SSL embeddings. In this paper, we investigate the model's performance when only the noisy speech signal is available.

Embeddings extracted from noisy inputs are likely to be distorted, resulting in lower-quality synthesized speech. To address this, we propose dividing the process into two stages: a) enhancing the distorted embeddings and b) synthesising clean speech from these improved embeddings. We aim to optimize the respective networks for each stage individually. The *embedding enhancement* network transforms SSL features extracted from distorted signals into features that resemble those from clean speech. Since the network preserves the dimension of the pre-trained SSL features, these enhanced features can seamlessly replace the SSL embeddings from noisy inputs in downstream tasks, thereby straightforwardly improving the system's robustness to noise.

To achieve high-quality speech, we investigate methods to further optimize the HiFi-GAN vocoder within the context of speech enhancement (SE). We begin by comparing the baseline mel-spectrogram loss with alternative loss functions commonly

used in speech enhancement, such as signal-to-distortion ratio (SDR) [15], compressed linear spectrogram [16], and Huber loss [17]. Regarding the *input* to the vocoder, it can theoretically be the latent representation from each hidden state of the SSL model. Thus, we analyze the importance of these latent representations at each hidden state and investigate the impact of a learnable weighted combination of all states. This approach offers deeper insights into the information from each layer that is pertinent to the task of speech enhancement.

The remainder of the paper is organized as follows. Section II describes the selection of pre-trained SSL models based on environmental distortions. In Section III, We outline the proposed framework and the optimization of the vocoder for speech enhancement. Section IV demonstrates the effectiveness of the approach and related metrics, while Section V provides the conclusions drawn from our study.

II. Selection of the pre-trained SSL model

Given the growing application of SSL models for generative speech enhancement, our recent study [18] analyzed these neural embeddings to gain insights into the information they contain. Specifically, we quantified the preservation of speaker and phonetic information, as well as the robustness of embedding features to noise and reverberation, extracted from various SSL speech models, including TERA [19], wav2vec2.0 [1], wavLM [2], and HuBERT [3]. The analysis revealed that HuBERT and wavLM preserved more phonetic information compared to TERA and wav2vec2.0, while TERA demonstrated the highest preservation of speaker information among the models compared. For assessing robustness, we considered two types of input distortions: additive background noise and reverberation, achieved by convolving with room impulse responses. The benchmark employed instrumental metrics, including cosine similarity and normalized mean square error (MSE), to compare embeddings extracted from clean speech with those from the corresponding distorted versions. The main finding was that additive noise caused more significant distortion to the embedding features than reverberation, with the distortion worsening at lower signal-to-noise ratios (SNRs). Among the four pre-trained models, TERA generated the most robust embedding features across all conditions. In the context of speech enhancement, where the input speech is subject to similar distortions, this robustness offers a significant advantage, making TERA a strong candidate as the SSL feature extractor. This choice is further supported by our results for the speech enhancement task, as described in Section IV-D.

III. Methodology

Figure 1 provides a schematic overview of the proposed method. First, we segment the signal and extract pre-trained SSL feature embeddings x_{emb}^{N}, where N denotes the number of layers, from the distorted speech data x for each frame. For simplicity, N will be omitted in the subsequent sections. The embedding enhancement network is then employed to transform the distorted features into their clean counterparts. Subsequently, a learnable weighted sum function within the $f(\cdot)$ layer combines the enhanced representations from the various hidden layers of the SSL model into a single refined feature per frame. These refined features, denoted as $\hat{x}_{emb}$, are then fed into the vocoder network to produce the enhanced speech signal.

A. Embedding Enhancement Network

The embedding enhancement network comprises four convolutional gated recurrent unit (ConvGRU) blocks, with each block featuring two bidirectional GRU layers and one convolution layer. The bidirectional GRU layers capture long-term temporal dependencies, while the convolutional layer adjusts the input dimension to match the feature dimension required for the vocoder input. We train the network by minimizing the mean squared error (MSE) between the enhanced embeddings and the clean embeddings, as follows:

$$L_{emb} = \frac{1}{B}\sum_{i=0}^{B}(\hat{x}_{emb} - y_{emb})^2, \tag{1}$$

where $\hat{x}_{emb}$ and y_{emb} represent the feature embeddings from the enhanced and clean speech signals, and B denotes the batch size.

In the case of using the last hidden feature, the loss is computed using only this final hidden feature. For fixed summation and learnable weighted summation, the loss values are averaged across all layers.

Note that the embedding enhancement network is trained in a *task-agnostic* manner to maximize flexibility, enabling its application to various downstream tasks without the need for re-training. Thus, the network directly estimates the clean SSL features. To enable fine-tuning for specific tasks, a learnable weighting factor is introduced and trained according to the requirements of the downstream task.

B. Vocoder

We utilize the HiFi-GAN [12] neural vocoder to synthesize speech from the SSL feature embeddings. Unless stated otherwise, the vocoder architecture adheres to the original HiFi-GAN design. We adjust the vocoder parameters to match the window size and frame shift of the pre-trained SSL model, as described further in Section IV. Additionally, we align the dimension of SSL feature embeddings with the input dimension of the vocoder through a one-dimensional matching layer and adapt the upsampling ratio for each block in the neural vocoder.

We use the original adversarial loss as proposed in [12] and optimize the generator loss specifically for the speech enhancement task. In addition to the original **L1 Mel-spectrogram loss**, we also investigate other loss functions commonly used in speech enhancement, including the **SDR loss**:

$$L_{\text{SDR}} = 10\log_{10}\frac{\|x\|^2}{\|\hat{x}-x\|^2}, \tag{2}$$

where x is the target signal and $\hat{x}$ is the synthesized estimate; the **compressed spectrogram loss**:

$$L_{\text{spec}} = \||\hat{X}|^c - |X|^c\|_H, \tag{3}$$

where $\hat{X}$ is the synthesized speech magnitude spectrogram, X is the clean speech magnitude spectrogram, and c is a compression factor; and the Huber loss [17], $\|\cdot\|_H$, on the compressed spectrogram, which calculates L1 loss when the error values exceed a predefined threshold and switches to mean squared error (MSE) loss below this threshold, thereby combining the advantages of both loss functions. Lastly, we investigate the **Time-domain Huber loss** applied to the synthesized and clean speech in the time domain to determine the optimal loss function.

Each hidden layer of the SSL model provides latent speech representations, with each layer containing valuable information about the underlying speech that can be utilized by the vocoder. Consequently, the method of combining embeddings is crucial, as the speech generation process involves upsampling this input. We investigate three approaches: (1) utilizing a single, optimal, hidden feature; (2) employing a fixed, equally weighted summation of all embedding layers; and (3) incorporating a learnable weighted summation of all embedding layers. These approaches are tested with different SSL models (see Section IV), with a distinct set of weights independently learned for each model.

IV. Experiments

A. Experimental setup

The vocoder and the proposed embedding network are trained using the Deep Noise Suppression (DNS) Challenge 2021 training set [20]. This dataset includes clean speech samples from the Librivox dataset [21], which features recordings from 10,000 audiobooks in various languages. Additionally, noise samples are sourced from Audioset [22], with supplementary augmentation from noise datasets such as Freesound [23] and DEMAND [24]. For our experiments, we utilized the English language subset of the aforementioned datasets for model training. To simulate reverberation, we employed the openSLR26 and openSLR28 datasets [25]. We adhered to the SNR, reverberation time (RT60), and other hyperparameters specified in the original DNS challenge paper [26]. For the comparison of model performance and analysis, the test set included both background noise and reverberation. Background noise was sourced from the Noisex-92 [27] dataset, with SNR levels set at (-7, 0, 5, 10, 15) dB. We use the MIT Impulse Response Survey dataset [28], which contains recorded room impulse responses, to generate reverberation. For analysis, we utilized three distinct datasets: a noisy test set (*noise*), a reverberant test set (*reverb*), and a combined test set (*noise+reverb*).

The proposed system is evaluated against two baselines. The first baseline, *Clean Vocoder*, is trained to synthesize clean speech signals from embeddings extracted from *clean* data. This naïve neural vocoder serves as an *upper* bound for the proposed system's performance, representing the best available speech quality when the embedding enhancement network perfectly maps distorted embeddings to clean speech embeddings. The second baseline is the Denoising Vocoder [11], which synthesizes clean speech from distorted embeddings derived from noisy input signals.

B. System and training details

We employ the AdamW optimizer [29] with an initial learning rate of 1e-4 and momentum parameters (β_1, β_2) set to (0.8, 0.99). Exponential learning rate scheduling is applied with a decay rate of 0.999. The parameters of the HiFi-GAN vocoder were adjusted to match the input specifications of the SSL model. For example, to align with the pre-trained TERA model and accommodate 16 kHz signals, the DFT size, window size, and hop size in HiFi-GAN are set to 400, 400, and 160 samples, respectively. After the feature matching layer, the channel dimension of the feature embeddings is reduced from 768 to 512, consistent with the original HiFi-GAN vocoder. This reduction allows us to align the SSL model with the baseline models for comparison. For loss and input analysis, we first reduce the input embeddings to 80 dimensions to match the dimension of the original vocoder input, which uses 80-dimensional mel spectrograms. Subsequently, the channel is upsampled using ratios of (8, 5, 2, 2) and kernel sizes of (16, 15, 4, 4). The spectrogram compression factor c in Equation (3) is set to 0.6, as recommended in [16], and the threshold value for the Huber loss is set to 1.0.

C. Metrics

Performance is evaluated using non-intrusive neural network-based metrics, specifically DNSMOS [20] and NISQAv2 [30], as intrusive metrics are less suitable for assessing generative models [31]. DNSMOS is a robust instrumental metric utilized to evaluate model performance in the DNS Challenge. It involves training a combination of convolutional and fully connected layers to estimate perceptual mean opinion score (MOS) results. DNSMOS provides instrumental MOS scores for overall quality (**OVRL**), signal quality (**SIG**), and background noise (**BAK**) of the input signal, with higher scores indicating better performance. Another non-intrusive speech quality assessment metric, NISQAv2, provides a more detailed evaluation of speech quality. It evaluates the given signal based on **MOS**, noisiness (**NOIS.**), discontinuity (**DIS.**), coloration (**COL.**), and loudness (**LOUD.**), with higher scores indicating better speech quality.

In addition, we use word error rate (**WER**) and character error rate (**CER**) [32] to measure the speech intelligibility of the enhanced signals. For this purpose, we employ two different pre-trained ASR backends: wav2vec2.0[1] and Hubert[2]. Both WER and CER follow a "lower is better" principle, meaning that lower values indicate better recognition of the speech signal.

D. Results

Choice of SSL model. To determine the optimal SSL features for vocoder input, we implemented the vocoder using three different pre-trained SSL models: HuBERT, WavLM and TERA.

[1]https://huggingface.co/facebook/wav2vec2-base-960h
[2]https://huggingface.co/facebook/hubert-large-ls960-ft

TABLE I: Vocoder performance with different SSL models, trained with the *generator loss only*. Use learnable weight summation of each model as input.

SSL Model	DNSMOS			NISQAv2				
	OVRL	**SIG**	**BAK**	**MOS**	**NOIS.**	**DIS.**	**COL.**	**LOUD.**
Target	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476
HuBERT	3.435	3.725	4.131	3.256	3.637	3.284	3.331	3.777
WavLM	3.477	3.755	4.158	3.354	3.760	3.355	3.373	3.817
TERA	**3.607**	**3.922**	**4.160**	**4.021**	**3.893**	**4.184**	**3.823**	**4.224**

TABLE II: Analysis with various types of generator loss functions using *noise+reverb*.

Loss function	DNSMOS			NISQAv2
	OVRL	**SIG**	**BAK**	**MOS**
Distorted	1.819	2.545	1.812	1.659
Mel-spectrogram	2.893	3.241	3.859	2.035
SDR	1.841	2.047	4.014	1.117
Time-domain Huber	2.878	3.218	3.842	1.589
Compressed spec.	**3.170**	**3.483**	**4.044**	**2.145**

We utilized a clean input waveform and employed hidden features obtained through learnable weighted summation as input to the vocoder network. The results are summarized in Table I. It is important to note that the vocoders consistently utilize the learnable weighted summation of all available hidden states, including all four layers of TERA or all twelve layers of HuBERT. Speech generated from pre-trained TERA embeddings demonstrates superior scores in terms of speech intelligibility, quality, and denoising, even though HuBERT extracts three times more hidden features than TERA. This supports the findings from the task-agnostic comparison of SSL models in [18], which again highlighted TERA as a highly suitable option. Based on these results, TERA is used in the subsequent studies.

Generator loss function. Table II presents the instrumental metric scores for different loss functions used in the vocoder generator loss. The evaluation is conducted on both reverberant and noisy data. It can be observed that the compressed spectrogram loss yields the best results across all metrics. Thus, this loss function was selected as the generator loss function, replacing the mel-spectrogram loss. Due to space constraints, only the MOS score from NISQAv2 is presented. However, we confirmed that the other NISQAv2 metrics showed similar trends to those observed with DNSMOS.

Consequently, both the Clean Vocoder and Denoising Vocoder baselines are trained using this generator loss with the original HiFi-GAN adversarial loss. It is important to reiterate their main difference: the Clean Vocoder synthesizes speech using embeddings extracted from clean speech, while the Denoising Vocoder aims to synthesize the underlying clean speech from embeddings extracted directly from the noisy input signal.

Embedding enhancement network. Table III benchmarks the performance of the proposed two-stage framework, which includes the embedding enhancement network. This network comprises four ConvGRU blocks, a configuration that has been shown to provide the most significant improvement in speech signal quality and noise reduction compared to other setups. To ensure that the enhanced embeddings are accurately mapped to clean speech, we use the pre-trained Clean Vocoder model to generate the underlying clean speech.

Overall, the **Proposed.** framework consistently outperforms the Denoising Vocoder, demonstrating that the task-agnostic embedding enhancement is indeed beneficial. The ASR performance results in Table III focus only on background noise, as prior research [18] has shown that background noise has a more pronounced impact on embeddings than reverberation. We verified this finding: using wav2vec2.0 as backend, the noise test set yielded WER and CER scores of 47.60% and 35.95%, respectively, whereas the reverberant test set yielded scores of 17.10% and 7.53%. This trend is also observed with the HuBERT-based backend. The proposed system outperforms the Denoising Vocoder and generally provides better than directly using the noisy input.

Layer-wise analysis. To assess the impact of different layers, we measure the distortion of the representation at *each* layer using normalized MSE against the representation from clean speech, following the methodology outlined in [18]. As shown in Fig. 2, noise affects the embeddings more significantly than reverberation, as indicated by larger deviations in the distances. The representation from the last layer shows the lowest normalized MSE.

Next, we investigate the layer-wise weighting factors relative to the input distortion, as detailed in Table IV. The first row shows the learnable weighting factors for distorted speech input, while the second row represents the weighting factors for clean speech input. The robustness analysis of the last hidden feature, as depicted in Figure 2, aligns with the weighting factor results, which show that the last hidden feature is given the highest weight. Therefore, for evaluating model performance using a *single* embedding, we select the embedding from the last layer.

Table V presents the performance results for different combinations of TERA feature embeddings. The performance with the last hidden feature is comparable to that with fixed summation, indicating that fixed weighted summation might be less effective than using a single robust feature. The learnable sum approach achieved the highest performance, suggesting

TABLE III: Comparison with proposed embedding enhancement network. Evaluate with the combination of *noise+reverb* for † mark and tested with each metric of DNSMOS and NISQAv2, and use *noise* for $*$ mark and tested with pre-trained model of wav2vec2.0 and Hubert

Model	DNSMOS†			NISQAv2†					wav2vec2.0-based ASR*		Hubert-based ASR*	
	OVRL	**SIG**	**BAK**	**MOS**	**NOIS.**	**DIS.**	**COL.**	**LOUD.**	**WER (%)**	**CER (%)**	**WER (%)**	**CER (%)**
Clean	3.668	3.951	4.209	4.550	4.251	4.596	4.301	4.476	9.27	3.25	5.78	1.73
Distorted	1.819	2.545	1.812	1.659	1.698	3.073	2.328	2.505	47.60	35.95	31.73	21.69
Clean Vocoder	3.569	3.886	4.102	3.751	3.546	3.996	3.574	4.038	14.85	6.12	7.69	2.80
Denoising Vocoder	3.038	3.397	3.863	2.698	2.966	**3.579**	2.699	3.384	51.08	30.94	41.63	24.85
Proposed.	**3.285**	**3.606**	**4.027**	**2.779**	**3.204**	3.009	**2.737**	**3.597**	**45.73**	**27.50**	**37.34**	**22.68**

TABLE IV: Weighting of each latent feature from the two baselines, Denoising- and Clean Vocoder, for the corresponding inputs.

Input	1^{st} **layer**	2^{nd} **layer**	3^{rd} **layer**	4^{th} **layer**
Distorted	-0.018	0.113	-0.078	**1.360**
Clean	-0.025	0.171	0.244	**1.738**

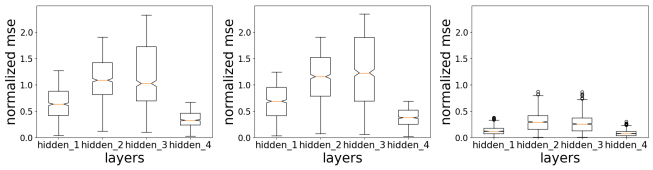
TABLE V: Analysis with various types of input embedding combinations, trained with all loss terms.

Input combination	DNSMOS			NISQAv2
	OVRL	**SIG**	**BAK**	**MOS**
Last hidden	2.718	3.059	3.673	2.008
Fixed sum	2.789	3.130	3.763	1.990
Learnable sum	**3.170**	**3.483**	**4.044**	**2.145**

that even though a single robust hidden feature is valuable, the model still benefits from incorporating information from multiple embeddings to synthesize speech more effectively.

V. Conclusions

We proposed an embedding enhancement network that maps distorted speech embeddings, obtained from pre-trained SSL models applied to noisy and reverberant inputs, to the embeddings of the underlying clean speech. These enhanced embeddings are then used by a vocoder to generate clean speech. After selecting TERA as the SSL model, experiments were conducted to identify the optimal generator loss function for the vocoder, resulting in the adoption of the compressed spectrogram loss. Additionally, a layer-wise analysis was performed to evaluate the robustness of embeddings from different hidden layers under various input distortions. This analysis revealed two key points: first, that additive noise has a greater impact on embeddings than reverberation, and second, that the final layer embedding is the most robust. These findings were supported by the analysis of weighting factor according to distortion type, which indicates contribution of each hidden layer to speech enhancement. The best performance was achieved using a learnable weighted combination of all layers, demonstrating that incorporating information from multiple embeddings remains beneficial. The proposed system, which includes the embedding enhancement, consistently outperforms the Denoising Vocoder baseline in both ASR tasks and speech quality metrics, even under challenging conditions.



(a) Noise+reverb (b) Noise (c) Reverb

Fig. 2: Standardized MSE between the clean and distorted embeddings from pre-trained TERA models. Three types of distortions (*noise+reverb*, *noise*, *reverb*) are simulated.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[2] S. Chen, C. Wang, Z. Chen, J. Li, and et al., "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. of Sel. Topics in Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[3] W. Hsu, B. Bolte, Y. Tsai, K. Lakhotia, and et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 29, pp. 3451–3460, 2021.

[4] S. Yang, P. Chi, Y. Chuang, C. Lai, and et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. INTERSPEECH*, 2021, pp. 1194–1198.

[5] X. Zhao, Q. Zhu, and J. Zhang, "Speech enhancement using self-supervised pre-trained model and vector quantization," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 330–334.

[6] K. Hung, S. Fu, H. Tseng, H. Chiang, and et al., "Boosting Self-Supervised Embeddings for Speech Enhancement," in *Proc. INTERSPEECH*, 2022, pp. 186–190.

[7] Z. Huang, S. Watanabe, S. Yang, P. García, and et al., "Investigating self-supervised learning for speech enhancement and separation," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 6837–6841.

[8] S. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and et al., "BigVGAN: A universal neural vocoder with large-

scale training," in *Proc. Intl. Conf. on Telecommun. and Signal Process.*, 2023.

[9] Y. Li, C. Han, V. Raghavan, G. Mischler, and et al., "StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 36, 2024.

[10] W. Jang, D. Lim, J. Yoon, B. Kim, and et al., "UnivNet: A Neural Vocoder with Multi-Resolution Spectrogram Discriminators for High-Fidelity Waveform Generation," in *Proc. INTERSPEECH*, 2021, pp. 2207–2211.

[11] B. Irvin, M. Stamenovic, M. Kegler, L. Yang, and et al., "Self-supervised learning for speech enhancement through synthesis," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2023, pp. 1–5.

[12] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Proc. Adv. in Neural Inf. Process. Syst.*, vol. 33, pp. 17 022–17 033, 2020.

[13] Z. Wang, X. Zhu, Z. Zhang, Y. Lv, and et al., "SELM: Speech enhancement using discrete tokens and language models," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2024, pp. 11 561–11 565.

[14] Y. Koizumi, H. Zen, S. Karita, Y. Ding, and et al., "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2023, pp. 1–5.

[15] R. Scheibler, "SDR — medium rare with fast computations," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 701–705.

[16] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *Proc. Intl. Conf. on Telecommun. and Signal Process.*, IEEE, 2021, pp. 72–76.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[18] Y. Song, D. Kim, N. Madhu, and H. Kang, "On the disentanglement and robustness of self-supervised speech representations," in *Intl. Conf. on Electron., Inf. and Commun.*, 2024, pp. 662–665.

[19] A. Liu, S. Li, and H. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE Trans. on Audio, Speech, and Language Process.*, vol. 29, pp. 2351–2366, 2021.

[20] C. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2021, pp. 6493–6497.

[21] J. Kearns, "Librivox: Free public domain audiobooks," *Reference Rev.*, vol. 28, no. 1, pp. 7–8, 2014.

[22] J. Gemmeke, D. Ellis, D. Freedman, A. Jansen, and et al., "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 776–780.

[23] E. Fonseca, J. P. Puig, X. Favory, F. F. Corbera, and et al., "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. Intl. Soc. for Music Inf. Retrieval*, ISMIR, 2017.

[24] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. of Meetings on Acoust.*, vol. 19, 2013.

[25] T. Ko, V. Peddinti, D. Povey, M. Seltzer, and et al., "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE Intl. Conf. on Acoust., Speech and Signal Process. (ICASSP)*, IEEE, 2017, pp. 5220–5224.

[26] C. Reddy, H. Dubey, K. Koishida, A. Nair, and et al., "INTERSPEECH 2021 Deep Noise Suppression Challenge," in *Proc. INTERSPEECH*, 2021, pp. 2796–2800.

[27] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech commun.*, vol. 12, no. 3, pp. 247–251, 1993.

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

[29] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Intl. Conf. on Learning Representations*, 2019.

[30] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. INTERSPEECH*, 2021, pp. 2127–2131.

[31] J. Pirklbauer, M. Sach, K. Fluyt, W. Tirry, and et al., "Evaluation metrics for generative speech enhancement methods: Issues and perspectives," in *SC; 15th ITG Conference*, VDE, 2023, pp. 265–269.

[32] A. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition," in *Proc. Intl. Conf. on Spoken Language Process.*, 2004.