IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 22 September 2025, accepted 6 October 2025, date of publication 9 October 2025, date of current version 15 October 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3619782

RESEARCH ARTICLE

Foundation Models for Speech Enhancement Leveraging Consistency Constraints and Contrast Stretching

MUHAMMAD SALMAN KHAN[1], (Member, IEEE), VALERIO MARIO SALERNO[1], MORENO LA QUATRA[1], (Member, IEEE), KUO-HSUAN HUNG[2], SZU-WEI FU[3], YU TSAO[2], (Senior Member, IEEE), AND SABATO MARCO SINISCALCHI[4,5], (Senior Member, IEEE)

[1]University of Enna "Kore," 94100 Enna, Italy
[2]Academia Sinica, Taipei 115, Taiwan
[3]Nvidia Corporation, Taipei 114, Taiwan
[4]University of Palermo, 90133 Palermo, Italy
[5]Norwegian University of Science and Technology (NTNU), 7034 Trondheim, Norway

Corresponding author: Moreno La Quatra (moreno.laquatra@unikore.it)

This work was supported in part by the "DARE–Digital Lifelong Prevention" Project under Grant PNC0000002 and Grant CUP: B53C22006450001, and in part by the Italian Complementary National Plan PNC-I.1 Research Initiatives for Innovative Technologies and Pathways in the Health and Welfare Sector under Grant D.D. 931 of 06/06/2022.

ABSTRACT Foundation models (FM) have proven effective in many speech applications except for speech enhancement (SE), where FM-based SE solutions still fall short with respect to specialized deep architectures. This work seeks to close this gap by systematically assessing and contrasting leading pre-trained FM architectures on a commonly used SE task, namely VoiceBank-Demand, and on the complex Deep Noise Suppression (DNS) challenge. Furthermore, three main ideas will be leveraged to boost FM-based SE models, namely: 1) Attention-based mask generation, 2) consistency-preserving loss, and 3) perceptual contrast stretching (PCS). Specifically, frame-level representations are effectively modeled using conformer layers, which leverage an attention mechanism. Inconsistency effects of signal reconstruction from the spectrogram are mitigated by incorporating consistency in the loss function. Finally, PCS is employed to improve the contrast of input and target features according to perceptual importance. All FM-based models generate an Ideal Ratio Mask (IRM) from which the estimated clean speech is obtained. Experimental results on the VoiceBank-DEMAND task demonstrate that our approach helps close the gap between FM-based and SOTA SE solutions. When tested on the DNS challenge, the proposed FM-based SE solution compares favorably with previously proposed approaches on perceptual quality metrics, using only 10% of the available training material, though with a trade-off in signal fidelity (SI-SDR) when PCS preprocessing is applied.

INDEX TERMS Speech enhancement, foundation models, consistency loss.

I. INTRODUCTION

Background noise is known to potentially severely degrade both quality and intelligibility of the audible speech [1] and negatively affects the accuracy of human and automatic speech recognition systems [2]. Therefore, it is not surprising that speech enhancement (SE) has attracted considerable research attention for decades [3] as an important front-end of speech processing systems. SE is a complex task focused on improving intelligibility and quality of the signal by eliminating undesired background noise [4], [5], [6], [7]. The handling of the SE task is indeed important, as it plays a key role in applications targeting (i) human-auditory-related tasks, such as communications, e.g. [8], hearing aids, e.g. [9],

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

and cochlear implants [10], [11], and (ii) automatic speech recognition (ASR), e.g., [12], [13], [14], [15], [16], [17], [18], [19].

Conventional SE techniques, such as Wiener filtering [20], spectral subtraction [21], and minimum mean squared error (MMSE) estimation [22], struggle to adapt to non-stationary noise in real-world environments with unpredictable acoustic conditions. Additionally, they often introduce an undesirable artifact known as musical noise [23]. In contrast, SE models based on neural architectures have demonstrated significantly superior performance compared to classical methods [24], [25], [26]. However, their effectiveness in addressing the musical noise issue and outperforming traditional signal processing techniques has only been widely recognized in recent years, driven by advancements in high-performance computing and breakthroughs in deep learning training methodologies [27], [28], [29]. Over the years, various deep neural network models have been proposed for SE, including (i) feed-forward deep neural networks (DNNs) [7], [24], [30], (ii) recurrent neural networks (RNNs) [31], [32], [33], and (iii) convolutional neural networks (CNNs) [34], [35]. Regardless of the specific architecture, deep SE models can generally be classified into two categories: discriminative and generative.

In the discriminative approach, models are trained to minimize the discrepancy between enhanced and clean speech using either a masking-based or a mapping-based strategy. In masking-based methods, an ideal binary mask (IBM) or a smoothed ideal ratio mask (IRM) [7] is estimated from a set of complementary features extracted from noisy speech. The noisy features are then modified using the predicted mask to obtain enhanced features [36]. Conversely, in the mapping-based approach, a deep model learns a high-dimensional vector-to-vector transformation and is trained in a regression framework [24], [37]. In both methods, the difference between the enhanced and reference speech is typically measured using a point-wise L_p-norm distance [38] or a perceptual metric [39]. In the second category, the clean speech distribution is incorporated into the objective function. Specifically, generative approaches focus on aligning with the speech signal's distribution rather than minimizing a point-wise loss, aiming to produce more natural-sounding speech. In recent years, several deep generative models have been explored for SE. For instance, SEGAN [40], CMGAN [41], and SCP-GAN [42] utilize generative adversarial networks (GANs) [43] for implicit density estimation. Alternatively, explicit density estimation has been investigated in works such as [44] and [45], where a variational autoencoder (VAE) [46] is employed. Diffusion-based generative models [47] have also been studied to address SE, e.g., [48], [49], [50]; the key idea is to gradually transform clean speech into noise, and training a deep model to invert this process for different noise scales. Flow matching generative models have been successfully tested in [51]. Current generative approaches, though aimed at estimating clean speech distributions for robustness to unseen noise, fall short compared to discriminative solutions. It should also be mentioned that deep SE models for processing both magnitude and phase information have also been proposed, e.g., [52], [53].

Despite the significant advancements and active research in deep SE models, developing effective SE solutions remains an ongoing challenge. This is highlighted by the continual release of novel speech datasets, such as VoiceBank-DEMAND [54], created to benchmark SE techniques, and the recurring organization of specialized competitions, like the Deep Noise Suppression (DNS) Challenge [26], now in its fifth edition. Furthermore, the increasing complexity of deep SE architectures poses additional difficulties, particularly in reproducing results when the associated source code is not publicly accessible. Foundation models (FM) comprise both self-supervised learning (SSL) approaches, such as Wav2vec2.0 [55], HuBERT [56], and WavLM [57], and weakly supervised learning (WSL) solutions like Whisper [58]. These deep neural architectures are pre-trained on diverse acoustic data and can extract speech representations without requiring detailed knowledge of their internal configuration and training process.

FMs have attracted a lot of interest in the speech community and have been successfully used in downstream speech tasks, as demonstrated on several benchmarks [59], [60], [61]. They have shown strong performance across diverse speech applications, including speech emotion recognition [62], [63], [64], [65], speaker and language identification [66], [67], and language understanding tasks [68], [69].

Nevertheless, those FMs attained poor results on the SE task, which were not as good as those attained in the other speech tasks. The underperformance of FMs in SE tasks stems from several fundamental challenges. First, most SSL models are pre-trained with objectives focused on high-level semantic representations (e.g., phonetic content, speaker identity) rather than the fine-grained spectral details crucial for noise suppression. Second, the temporal resolution mismatch between FM outputs (typically 20ms frames) and the precise spectral information needed for SE creates alignment issues. Third, SE requires modeling complex spectro-temporal relationships between clean speech and noise, which differs significantly from the contrastive or reconstruction objectives used in FM pre-training. Finally, the magnitude-phase decomposition inherent in SE is not explicitly addressed in current FM architectures, limiting their direct applicability to enhancement tasks.

In [70], the authors analyzed SSL-based solutions and argued that the key issue was the lack of fine-grained information in speech embeddings and tried to address the problem by introducing cross-domain features. In [71] several data-strategies have been evaluated to improve the system performance. In [72], clean SSL were used for latent space modeling within the Conditional Variational Autoencoder (CVAE) framework, ensuring the model fully utilizes existing knowledge without noise interference. Despite the above

efforts, FM-based SE solutions still lag behind state-of-the-art (SOTA) SE methods.

In this work, we provide a systematic evaluation of FM-based SE models, both SSL and WSL, with the aim at closing the gap with SOTA SE solutions. While it is known that SSL models are effective in learning representations from large amounts of unlabelled data [55], [57], [73], WSL-based foundational models have also recently shown to learn generalizable representations from large amounts of curated data [58], [74]. Specifically, we first consider the following backbones: (i) Wav2Vec2.0 [55], (ii) WavLM [57], (iii) HuBERT [56], (iv) UniSpeech [75], (v) Mockingjay [76], (vi) Tera [77], and (vii) Whisper [58] encoder. Then we assess SE based on those FMs on both the VoiceBank-DEMAND task [54], and the DNS Challenge [26]. Finally, we deploy a FM-based SE model based on the recently proposed BEST-RQ architecture [78]. Our BEST-RQ model is the first version adapted for magnitude Short-time Fourier Transform (STFT) processing. To further boost FM-based SE, we introduce specific architectural solutions along with an ad-hoc consistency-preserving loss when deploying our solutions. In particular, perceptual contrast stretching (PCS) is applied as a spectral pre-processing step on both noisy and target speech data before training our models to enhance perceptually important frequency regions. As a key architectural choice, we employ the convolution-augmented Transformer (Conformer) [79] to generate the ideal ratio mask needed in the enhancement stage. Conformers have proven effective across various speech applications [79], [80], including speech enhancement [81], [82]. Specifically, Conformer-based heads have demonstrated superior performance in modeling contextual information within speech signals compared to both bidirectional LSTM (biLSTM) and standard Transformer-based architectures [83]. To optimize the neural parameters, consistency-preserving training [42] is leveraged. More specifically, the effect of the inverse STFT (iSTFT) reconstruction, which causes inconsistencies between signals, is taken into account when computing the loss, e.g., [42], [84]. Finally, perceptual contrast stretching (PCS) [85] was employed as a pre-processing step on both the noisy and target waveforms before training our model to better exploit the peculiar sensitivity of the human auditory systems and improve the final speech perceptual quality. In testing, PCS was applied only on the noisy waveform. Experimental evidence on the VoiceBank-DEMAND dataset reveals that a careful design of FM-based SE models closes the gap with SOTA SE solutions. Most importantly, the FM-based SE solution trained only on 300-hour material outperforms SOTA SE models trained on the entire 3000-hour material provided with the DNS task. In more detail, the proposed WavLM-Large-based SE solution achieves SOTA performance on VoiceBank-DEMAND with a PESQ score of 3.54, while on the DNS Challenge, it attains a WB-PESQ of 3.55 and NB-PESQ of 3.89, surpassing specialized architectures like MFNET (WB-PESQ: 3.43) and HPNET (WB-PESQ: 3.40) that were trained on significantly more data.

Our preliminary results reported in [83], limited to the WavLM backbone, are expanded and systematically organized, leading to the following core contributions:

- We provide a comprehensive evaluation framework for adapting foundation models to speech enhancement, systematically comparing SSL and WSL approaches across multiple architectures and identifying the key factors limiting their performance.
- We demonstrate that properly designed FM-based SE models can achieve state-of-the-art results while using only 10% of typical training data, with the proposed WavLM-Large solution achieving WB-PESQ of 3.55 on DNS Challenge, outperforming specialized architectures trained on 10× more data.
- We introduce a dual-stream architecture combining foundation model representations with spectral features, enhanced by consistency-preserving training and perceptual contrast stretching, along with the first BEST-RQ adaptation for magnitude STFT processing. Unlike previous approaches that apply these techniques independently, our framework systematically integrates them within a unified FM adaptation strategy.

Furthermore, our systematic evaluation demonstrates that foundation models are particularly effective in limited data scenarios. The proposed framework is evaluated using at most 300 hours of training data (i.e., on DNS dataset) to achieve state-of-the-art performance, making it practical for domains where large-scale labeled datasets are unavailable. We speculate that this data efficiency stems from the rich speech representations learned during pre-training, which transfer effectively to the enhancement tasks.

The rest of the work is organized as follows. Section II introduces the self-supervised and weakly supervised foundation models used in our study. The proposed enhancement architecture is detailed in Section III. Section IV describes the datasets and experimental setup. Results and analysis are presented and discussed in Section V. Finally, Section VI concludes our work.

II. SSL AND WSL BACKBONE

The foundation models employed in this work can be broadly categorized into two groups based on their training paradigm: SSL and WSL models. Both types learn speech representations that can be leveraged for SE, though they differ in their architectures and training objectives.

A. SELF-SUPERVISED MODELS

Several SSL models in our study share a common architectural foundation: a CNN-based feature extractor processing raw waveforms, followed by a stack of specialized encoder layers. The original architecture, introduced by Wav2vec 2.0 [55], employs transformer encoder layers and a contrastive learning approach, where frames are masked and replaced

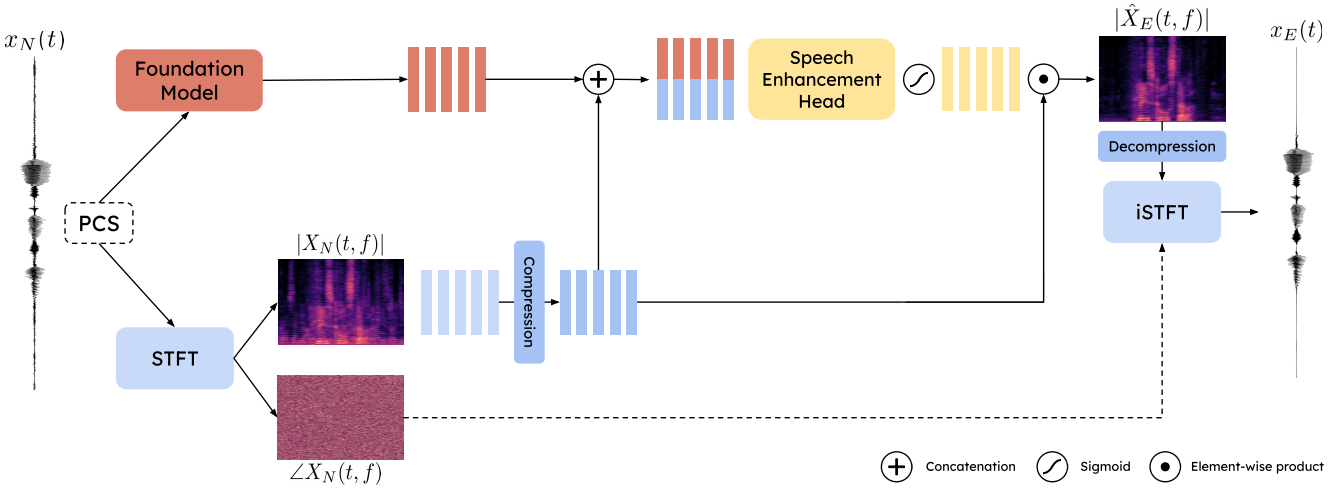

FIGURE 1. Overall architecture of the proposed foundation model-based SE framework. Input noisy speech $x_N(t)$ is processed through two parallel streams: the foundation model extracts frame-wise feature representations (shown as sequential bars), while STFT decomposition produces magnitude $|X_N(t,f)|$ and phase $\angle X_N(t,f)$ spectrograms. The magnitude spectrum undergoes logarithmic compression before feature concatenation. The frame-wise representations from both streams are concatenated and fed to the Speech Enhancement Head, which generates an ideal ratio mask applied to the log-compressed magnitude spectrum. The red blocks represent foundation model feature extraction, while the blue blocks show spectral processing with logarithmic compression. The enhanced magnitude $|X_E(t,f)|$ is decompressed and combined with the original phase to reconstruct the enhanced speech $x_E(t)$ via inverse STFT (iSTFT). PCS indicates perceptual contrast stretching preprocessing.

with learned feature vectors. The model is trained to distinguish true frame representations from distractors, establishing a framework for self-supervised speech representation learning.

This architecture has evolved in different directions. HuBERT [56] and WavLM [57] keep the CNN-transformer structure but shift to predicting cluster assignments derived from the MFCC-based clustering of k-means. This assimilates the training paradigm of these models to BERT's masked token prediction in NLP [86]. WavLM further enhances this approach by incorporating noisy and overlapped speech during training, making it particularly suitable for SE tasks. Conformer [79] takes a different direction by replacing transformer layers with conformer layers, which integrate convolution modules within the self-attention mechanism. This architectural choice enables the model to better capture local and global dependencies in the speech signal. UniSpeech [75] represents a hybrid approach, combining self-supervised learning with phonetic-aware supervision. While maintaining WavLM's core architecture, it incorporates a supervised CTC objective and a quantizer module to obtain discrete latent representations, bridging the gap between purely self-supervised and supervised approaches. All these waveform-based models originally generate frame-level representations with a temporal resolution corresponding to a 20 ms hop length. To align these representations with our STFT features for SE, we modified the stride of the final convolutional layer from 2 to 1, achieving the desired temporal resolution without any information loss.

A fundamentally different approach is taken by MockingJay [76] and TERA [77], which operate directly on spectrogram representations rather than raw waveforms. These models use log-scale mel spectrograms as input and are trained to reconstruct masked portions of the input using L1 loss. Although MockingJay employs a standard masking strategy, TERA introduces a more sophisticated approach with multiple augmentation techniques applied to the input spectrogram, including temporal and frequency masking, magnitude perturbation, and their combinations. However, both models rely primarily on reconstruction-based objectives without using additional learning characteristics of speech signals.

BEST-RQ [78] introduces a simplified approach to speech SSL, employing a random-projection quantizer to generate discrete labels for masked prediction tasks. In its original form, the architecture uses a stack of Conformer layers and processes mel-spectrograms through a convolutional front-end that reduces the temporal resolution by a factor of four. The key innovation lies in its training objective: while waveform-based models use contrastive or clustering approaches and spectrogram-based models rely on reconstruction objectives, BEST-RQ uses randomly initialized and fixed projections to map speech signals to discrete labels. For our SE task, we introduce the first BEST-RQ model adapted for magnitude STFT processing. We design the architecture with 12 Conformer layers (comparable to other base models) and remove the CNN front-end, allowing direct processing of STFT features with precise temporal alignment. Due to computational constraints, we only evaluate the base variant of BEST-RQ in this study. We pre-train the model using the aforementioned architecture on LibriSpeech 960h [87], making BEST-RQ a viable backbone for SE while preserving its core random-projection quantization principle.

B. WEAKLY SUPERVISED MODEL

In contrast to SSL approaches, Whisper [58] represents a WSL paradigm, trained end-to-end on a large-scale ASR task. Its encoder-decoder architecture processes 80-dimensional log-magnitude mel spectrograms, first through a CNN encoder that traditionally reduces temporal resolution to 20 ms, followed by Transformer-based encoder and decoder stages. For SE, we use only the encoder portion, modifying the CNN's final stride similarly to our SSL adaptations to maintain temporal alignment with STFT features.

The foundation models explored in this work represent different approaches to speech representation learning, ranging from waveform-based SSL to spectrogram-based weakly supervised approaches. These models are typically available in different sizes, with base versions trained on smaller datasets (∼1,000 hours) and larger variants trained on extensive speech collections (53K-90K hours). The exception is Whisper, which is trained on a proprietary dataset of 680,000 hours of automatically collected and filtered ASR data, making it more narrowly focused on speech recognition compared to the other models. By adapting their temporal resolutions and combining their learned representations with STFT features, our framework enables a systematic investigation of how different pre-training strategies and architectures contribute to SE performance, allowing us to identify the most effective approaches for this task.

III. FOUNDATION SE ARCHITECTURE

SE requires complex transformations to map noisy speech signals to their clean counterparts. We propose a framework that leverages pre-trained FMs while incorporating both perceptual and consistency aspects in the processing pipeline.

A. OVERALL ARCHITECTURE

The proposed architecture consists of two parallel processing streams that extract complementary features from the input signal (Figure 1). The first stream leverages a pre-trained FM, either SSL or WSL, to extract rich contextual speech representations. These representations capture high-level speech characteristics learned during pre-training on large-scale datasets. The second stream processes the signal in the spectral domain through STFT, providing explicit frequency decomposition. The FM stream processes the raw waveform, spectrogram or mel-spectrogram (depending on the chosen backbone) to generate frame-level representations $\mathbf{F}_{\text{fm}} \in \mathbb{R}^{T \times D_{\text{fm}}}$, where T is the number of frames and D_{fm} is the feature dimension. Concurrently, the spectral stream computes the STFT to obtain magnitude $|\mathbf{X}(f,t)|$ and phase $\angle\mathbf{X}(f,t)$ spectrograms. To effectively combine these representations, we apply logarithmic compression to the magnitude spectrum:

$$\mathbf{X}_{\log}(f,t) = \log(|\mathbf{X}(f,t)| + 1) \tag{1}$$

where $\mathbf{X}_{\log}(f,t)$ represents the compressed magnitude spectrum. This compression step helps effectively capture both low and high frequency energy components.

To optimally combine the complementary information from both feature streams, we systematically evaluated multiple fusion strategies. The foundation model features $\mathbf{F}_{\text{fm}}$ and compressed magnitude features $\mathbf{X}_{\log}(f,t)$ can be integrated through several approaches:

1) CONCATENATION

The log magnitude features are directly concatenated with the FM features along the feature dimension:

$$\mathbf{F}_{\text{concat}} = [\mathbf{F}_{\text{fm}} \| \mathbf{X}_{\log}(f,t)] \tag{2}$$

where $\|$ denotes concatenation and $\mathbf{F}_{\text{concat}} \in \mathbb{R}^{T \times (D_{\text{fm}}+F)}$.

2) GATING

A learnable gating mechanism controls the mixture between the two feature types, where spectral features are first projected to match the FM dimension:

$$\mathbf{F}_{\text{gate}} = \sigma(\mathbf{F}_{\text{fm}}) \odot \mathbf{F}_{\text{fm}} + (1 - \sigma(\mathbf{F}_{\text{fm}})) \odot \text{Linear}(\mathbf{X}_{\log}(f,t)) \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid activation function and $\odot$ denotes element-wise multiplication.

3) ADDITIVE FUSION

The spectral features are linearly projected to the FM dimension and added element-wise:

$$\mathbf{F}_{\text{add}} = \mathbf{F}_{\text{fm}} + \text{Linear}(\mathbf{X}_{\log}(f,t)) \tag{4}$$

4) MULTIPLICATIVE FUSION

Similar to additive fusion but using element-wise multiplication:

$$\mathbf{F}_{\text{mul}} = \mathbf{F}_{\text{fm}} \odot \text{Linear}(\mathbf{X}_{\log}(f,t)) \tag{5}$$

5) CROSS-ATTENTION

A cross-attention mechanism where FM features serve as queries to attend to the spectral information, while projected spectral features act as both keys and values:

$$\mathbf{F}_{\text{ca}} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{6}$$

where $\mathbf{Q} = W_Q\mathbf{F}_{\text{fm}}$, $\mathbf{K} = W_K\text{Linear}(\mathbf{X}_{\log}(f,t))$, $\mathbf{V} = W_V\text{Linear}(\mathbf{X}_{\log}(f,t))$, and W_Q, W_K, W_V are learnable projection matrices.

Based on empirical evaluation (Section V-C), concatenation proved most effective, achieving the best balance between performance and computational efficiency while ensuring both feature streams contribute equally to the enhancement process. Therefore, we adopt concatenation as our primary fusion strategy:

$$\mathbf{F}_{\text{combined}} = \mathbf{F}_{\text{concat}} = [\mathbf{F}_{\text{fm}} \| \mathbf{X}_{\log}(f,t)] \tag{7}$$

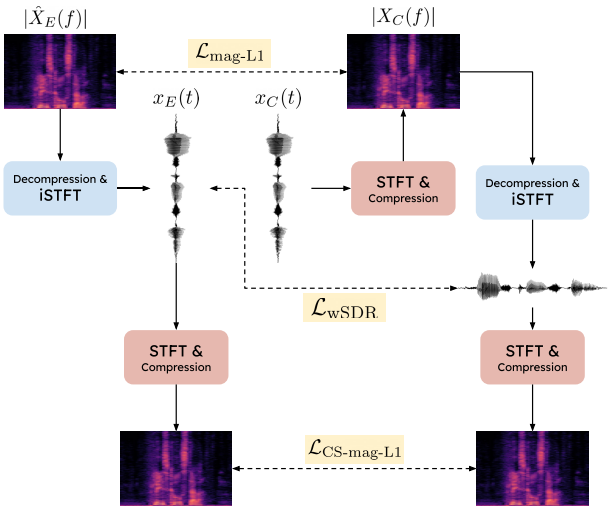

FIGURE 2. Diagram illustrating the computation of the three losses ($\mathcal{L}_{\text{mag-L1}}$, $\mathcal{L}_{\text{wSDR}}$, and $\mathcal{L}_{\text{CS-mag-L1}}$) with consistency-preserving reconstruction steps. Enhanced speech $x_E(t)$ and clean reference $x_C(t)$ undergo identical STFT transformation and compression operations to compute magnitude-domain loss $\mathcal{L}_{\text{mag-L1}}$. Time-domain loss $\mathcal{L}_{\text{wSDR}}$ directly compares the waveform signals. Consistency-preserving loss $\mathcal{L}_{\text{CS-mag-L1}}$ ensures both signals experience the same reconstruction artifacts by applying full STFT-iSTFT cycles before comparison. The process highlights the use of STFT, iSTFT, compression, and decompression operations to ensure alignment between input and reference signals.

B. HEAD OF SE MODEL

The enhancement process centers around estimating an IRM [7] from the combined features. We implement the enhancement head using two Conformer layers [79], selected based on empirical findings from our previous work [83] and the established effectiveness of shallow architectures for mask estimation tasks [70], [88]. The Conformer architecture combines self-attention with convolution operations, allowing it to capture both global dependencies and local feature patterns. Each Conformer layer processes the input sequence while maintaining temporal resolution. The output features from the final Conformer layer are projected to the frequency dimension F through a linear layer, then transformed into the log-magnitude IRM with [0, 1] bounds through a sigmoid activation:

$$\text{Mask}(f,t) = \sigma(\text{Linear}(\text{Conformer}(\mathbf{F}_{\text{combined}}))) \quad (8)$$

where $\sigma(\cdot)$ is the sigmoid function ensuring mask values between 0 and 1. Note that this mask operates in the log-compressed magnitude domain rather than the conventional linear magnitude domain used in traditional Ideal Ratio Masks (IRM). Since our features are in the log-compressed domain (as described in Section III-A), the estimated mask is applied to the log-compressed noisy magnitude spectrum through element-wise multiplication $\odot$:

$$|\mathbf{X}_{\text{log-enhanced}}(f,t)| = \text{Mask}(f,t) \odot |\mathbf{X}_{\text{log}}(f,t)| \quad (9)$$

The enhanced magnitude spectrum is then decompressed using the inverse of the logarithmic compression:

$$|\mathbf{X}_{\text{enhanced}}(f,t)| = \exp(|\mathbf{X}_{\text{log-enhanced}}(f,t)|) - 1 \quad (10)$$

Finally, the enhanced speech signal is reconstructed by combining the decompressed enhanced magnitude with the original noisy phase through iSTFT:

$$x_{\text{enhanced}}(t) = \text{iSTFT}(|\mathbf{X}_{\text{enhanced}}(f,t)|e^{j\angle\mathbf{X}_{\text{noisy}}(f,t)}) \quad (11)$$

Here, we retain the original noisy phase for reconstruction, following established practices in masking-based SE approaches [70], [88], [89]. Our approach focuses on magnitude enhancement, which has been shown to provide substantial perceptual improvements while maintaining computational tractability.

C. MULTI-OBJECTIVE LOSS COMPUTATION

Our training objective combines three complementary losses that operate at different signal representation levels (Figure 2):

1) **Magnitude Spectrum L1 Loss** ($\mathcal{L}_{\text{mag-L1}}$): Computes the L1 distance between enhanced and clean compressed magnitude spectrograms:

$$\mathcal{L}_{\text{mag-L1}} = \frac{1}{FT}\sum_{f,t} ||\mathbf{X}_{\text{log-enhanced}}(f,t)| - |\mathbf{X}_{\text{log-clean}}(f,t)|| \quad (12)$$

where $\mathbf{X}_{\text{log-clean}}(f,t)$ is the log-compressed magnitude spectrum of the clean signal, while F and T are the total number of frequency bins and time frames.

2) **Weighted Signal-to-Distortion Ratio Loss** ($\mathcal{L}_{\text{wSDR}}$): Operates in the time domain [90], combining speech and noise prediction terms:

$$\mathcal{L}_{\text{wSDR}} = -\alpha\frac{\langle x_{\text{clean}}, x_{\text{enhanced}}\rangle}{\|x_{\text{clean}}\|\|x_{\text{enhanced}}\|} - (1-\alpha)\frac{\langle n_{\text{clean}}, n_{\text{enhanced}}\rangle}{\|n_{\text{clean}}\|\|n_{\text{enhanced}}\|} \quad (13)$$

where $n_{\text{clean}} = x_{\text{noisy}} - x_{\text{clean}}$ and $n_{\text{enhanced}} = x_{\text{noisy}} - x_{\text{enhanced}}$ are noise components, and $\alpha = \|x_{\text{clean}}\|^2/(\|x_{\text{clean}}\|^2 + \|n_{\text{clean}}\|^2)$ is the energy ratio between clean speech and noise following the standard formulation in [90].

3) **Consistency-Preserving Magnitude L1 Loss** ($\mathcal{L}_{\text{CS-mag-L1}}$): Following [42], this loss addresses potential inconsistencies from STFT/iSTFT operations by ensuring both enhanced and clean signals undergo identical transform sequences before comparison (see Figure 2).

The total loss is the sum of these components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mag-L1}} + \mathcal{L}_{\text{wSDR}} + \mathcal{L}_{\text{CS-mag-L1}} \quad (14)$$

D. PERCEPTUAL CONTRAST STRETCHING

To improve perceptual quality, we incorporate PCS [85] as a pre-processing step. For each input waveform $x(n)$, PCS applies:

$$\tilde{Y}(\omega, t) = \log(|\text{STFT}(x)| + 1) \tag{15}$$

$$\hat{Y}(\omega, t) = \tilde{Y}(\omega, t) \cdot W(\omega) \tag{16}$$

where $W(\omega)$ represents frequency-dependent perceptual weights derived from the Band Importance Function [91]; $\tilde{Y}(\omega, t)$ represents the compressed magnitude spectrum and $\hat{Y}(\omega, t)$ is the perceptual weighted magnitude. During training, we apply PCS to both input and target signals, while at inference, only the input signal undergoes PCS processing. Exponential decompression and iSTFT are used to obtain the pre-processed waveform combining the $\hat{Y}(\omega, t)$ with the original phase information $\angle\text{STFT}(\hat{x})$. This perceptual enhancement particularly benefits regions of the spectrum that are most important for human perception, leading to improved subjective quality of the enhanced speech. The effectiveness of this approach is empirically demonstrated in our experiments (Section V).

IV. DATASETS AND EXPERIMENTAL SETUP

In the following sections, we describe the experimental setup, including datasets, evaluation metrics, training details, and model configurations.

A. DATASETS

We conduct evaluations on two datasets that represent different noise conditions and scenarios.

The first dataset is VoiceBank-DEMAND [54], which combines clean speech from VoiceBank [92] with noise samples from the DEMAND collection [93]. It contains recordings from 30 speakers, of which 28 are used for training and 2 for testing, ensuring no overlap in speakers between these sets. The training portion has 11,572 utterances mixed at four signal-to-noise ratio (SNR) levels: 0, 5, 10, and 15 dB. The test set contains 824 utterances mixed at SNRs of 2.5, 7.5, 12.5, and 17.5 dB. These different SNR values allow us to test how well our models generalize under changing noise intensities. All audio in VoiceBank-DEMAND is originally sampled at 48 kHz, and we downsample it to 16 kHz in our experiments, following standard practice [40].

The second dataset is a subset of the DNS 2020 Challenge dataset [26], a large-scale collection designed for real-world SE tasks. It comprises 500 hours of clean speech recorded from 2,150 speakers and 65,000 noise clips spanning 150 noise categories from Audioset and Freesound. For training, we generate a 300-hour subset of clean-noisy speech pairs following the official DNS dataset generation pipeline,[1] with noise added at SNR levels ranging from -5 dB to 20 dB. This subset maintains the original noise diversity and speaker distribution of the full dataset, as we follow the official DNS generation pipeline with proportional sampling across all noise categories and SNR ranges. For evaluation, we use the provided non-blind synthetic test set without reverberation, which consists of 150 noisy-clean utterance pairs with SNR values between 0 dB and 25 dB.

[1] https://github.com/microsoft/DNS-Challenge/tree/interspeech2020/master

B. MODEL CONFIGURATIONS

As discussed in Section II, we evaluate a broad set of foundation models as the backbone of our SE framework. These include waveform-based SSL models: Wav2Vec2.0 (W2V2) [55], Wav2Vec2.0 with Conformer layers (W2V2C) [79], HuBERT [56], WavLM [57], and UniSpeech [75]. We also evaluate spectrogram-based SSL models: Mockingjay [76], TERA [77], and BEST-RQ [78]. Additionally, we include the weakly supervised Whisper model [58] to provide a comprehensive comparison across different pre-training strategies. For the SSL models W2V2, HuBERT, WavLM, and UniSpeech, we use both the Base and Large variants. Whisper is evaluated using only the encoder portion of the Medium (M) and Large (L) configurations, as the decoder is tailored for ASR tasks and not relevant for SE. We adopt publicly available pre-trained weights for all models, with the exception of BEST-RQ, which we adapt to SE as discussed in Section II. All models are trained end-to-end using the framework presented in Section III.

C. TRAINING SETUP

Our implementation ensures exact temporal alignment between the foundation model outputs and the STFT representations. All experiments consistently use the same STFT configuration (400 FFT points, 160-sample hop) across different foundation model backbones to ensure fair comparison. This unified spectral processing approach allows us to isolate the impact of different pre-trained representations while maintaining consistent frequency-domain analysis. For foundation models with CNN front-ends (specifically Wav2Vec2.0, HuBERT, WavLM, UniSpeech, and Whisper), we adjust the stride of the final convolutional layer from 2 to 1 to achieve a 10ms frame resolution. This modification preserves all temporal information while achieving precise alignment with STFT frames, avoiding the need for frame duplication strategies used in prior work [70], [71].

After computing the STFT, we obtain 201 frequency bins for the magnitude features. We then concatenate these 201-dimensional STFT features with the output features $\mathbf{F}_{\text{fm}}$ from the selected foundation model. Base versions of the SSL models typically have $\mathbf{F}_{\text{fm}} = 768$, and Large versions have $\mathbf{F}_{\text{fm}} = 1024$. During training, we limit the input length to a maximum of 10 seconds for VoiceBank-DEMAND and 5 seconds for DNS, applying zero-padding if needed. In inference mode, we remove these length restrictions to handle variable-length recordings. We train all models end-to-end for 50 epochs on a single NVIDIA A100 GPU with a batch size of 4 for both VoiceBank-DEMAND and DNS datasets. We use the Adam optimizer [94] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a fixed learning rate of 10^{-4} throughout

training. All foundation model backbone parameters are fine-tuned end-to-end using equal weights for the three loss components ($\mathcal{L}_{\text{mag-L1}}$, $\mathcal{L}_{\text{wSDR}}$, $\mathcal{L}_{\text{CS-mag-L1}}$). No data augmentation beyond the training-time PCS preprocessing is employed. For PCS implementation, we leverage the perceptual scores $W(\omega)$ derived by the Band Importance Function (BIF) proposed in [85]. From preliminary tests, WavLM Large showed particularly strong performance, so we use this model backbone in our ablation studies. To ensure reproducibility and facilitate future research, we have made the source code publicly available,[2] including model implementations, training scripts, and evaluation protocols.

D. EVALUATION METRICS

To enable direct comparison with previous work, we adopt widely-used evaluation metrics specific to each dataset. For VoiceBank-DEMAND, we employ five complementary metrics: Wide-band Perceptual Evaluation of Speech Quality (WB-PESQ) [95], which evaluates perceptual speech quality on a scale from -0.5 to 4.5, following ITU-T P.862.2 recommendations; three Mean Opinion Score (MOS) predictions [96] - CSIG for speech signal distortion, CBAK for background noise intrusiveness, and COVL for overall quality, each ranging from 1 to 5; and Short-Time Objective Intelligibility (STOI) [97], which assesses speech intelligibility on a scale from 0 to 1 (reported as percentage values in our experiments). This set of metrics aligns with those used in recent SE studies [40], [53], [70], [83], facilitating fair comparison with SOTA approaches.

For the DNS dataset evaluation, we follow the challenge guidelines and prior work by using four key metrics: Narrow-band PESQ (NB-PESQ) and Wide-band PESQ (WB-PESQ) for perceptual quality assessment at different frequency ranges, STOI for intelligibility measurement, and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) for objective evaluation of enhancement quality.

E. ABLATION STUDIES

Our ablation studies systematically evaluate three key aspects of the proposed framework. First, we analyze the multi-component loss function by individually removing each loss term (magnitude L1, consistency-preserving loss, and wSDR). Second, we examine architectural design choices through specific tests:

1) **Feature Integration**: We compare models using only foundation model features, only STFT features, and their combination to quantify each feature type's contribution.
2) **Temporal Alignment**: We evaluate our front-end adaptation approach against the frame duplication strategy from previous work [70], [71].
3) **Sequential Processing**: We assess the importance of temporal modeling in the SE head by replacing Conformer layers with simple linear layers.
4) **Feature Fusion Strategy**: We systematically compare the five fusion approaches described in Section III-A to determine the optimal method for combining foundation model and spectral representations.

Finally, we also investigate the impact of training data volume on the DNS Challenge task by comparing models trained on 100, 200, and 300 hours of data. These ablation studies are designed to understand the relative importance of each component and identify the most effective configurations for the speech enhancement task.

V. RESULTS AND DISCUSSION

We present our experimental results across three dimensions: performance on the VoiceBank-DEMAND dataset (Section V-A), evaluation on the DNS Challenge (Section V-B), and ablation studies (Section V-C).

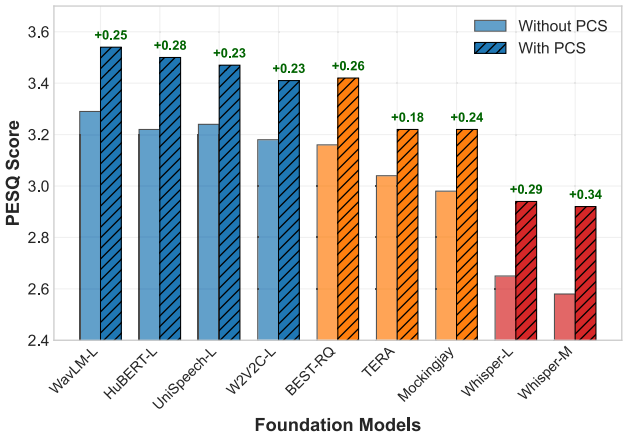

FIGURE 3. Impact of PCS preprocessing across different architectures on VoiceBank-DEMAND dataset.

A. VOICEBANK-DEMAND RESULTS

Table 1 shows a comprehensive comparison of FM-based approaches against previous SOTA methods on the VoiceBank-DEMAND dataset. Several key findings emerge from these results. First, WavLM-Large model with PCS preprocessing achieves the highest PESQ score of 3.54 and COVL score of 4.23, surpassing both specialized deep learning approaches and previous FM-based methods. Second, the application of PCS preprocessing consistently improves perceptual quality metrics (PESQ and COVL) across all foundation models. For instance, WavLM-Large shows an improvement in PESQ from 3.29 to 3.54 when PCS is applied. This pattern applies to all other models as well, with improvements ranging from +0.18 to +0.34 PESQ points, as illustrated in Figure 3. Third, the Whisper model, despite its extensive supervised pre-training on speech recognition tasks, performs notably worse than SSL-based models, with

[2]https://github.com/K-STMLab/FM-SE

TABLE 1. Performance comparison on VoiceBank-DEMAND dataset. PCS-PP denotes PCS pre-processing while B and L suffixes denote base and large model variants.

Model Name	PCS-PP	PESQ-WB	CSIG	CBAK	COVL	STOI
SEGAN [40]	×	2.16	3.48	2.94	2.80	92.0
DEMUCS [98]	×	3.07	4.31	3.40	3.63	95.0
SE-Conformer [82]	×	3.13	4.45	3.55	3.82	95.0
Metric-GAN [39]	×	2.86	3.99	3.18	3.42	-
Metric-GAN+ [99]	×	3.15	4.14	3.16	3.64	-
DPT-FSNet [100]	×	3.33	4.58	3.72	4.00	**96.0**
CMGAN [41]	×	3.41	3.63	3.94	4.12	**96.0**
TridentSE [101]	×	3.47	4.70	3.81	4.10	**96.0**
MP-SENet [53]	×	3.50	4.73	**3.95**	4.22	**96.0**
Previous FM-based Methods						
BSSE [70]	×	3.20	4.53	3.78	4.04	**96.0**
BSSE [70]	✓	3.46	**4.75**	3.49	4.20	95.0
SSF-CVAE [72]	×	3.04	4.38	2.91	3.72	95.0
BSS-CFFMA [102]	×	3.21	4.55	3.70	3.91	94.0
Huang et al. [89]	×	2.80	-	-	-	94.0
Close et al. [103]	×	2.79	4.10	2.68	3.44	94.0
PFPL [104]	×	3.15	4.18	3.60	3.67	95.0
Proposed FM-based Architecture						
W2V2-B	×	3.10	4.56	3.70	3.90	95.0
W2V2-L	×	2.95	4.44	3.60	3.75	95.0
W2V2C-L	×	3.18	4.58	3.74	3.95	95.0
W2V2-B	✓	3.40	4.65	3.46	4.11	95.0
W2V2-L	✓	3.18	4.51	3.35	3.91	95.0
W2V2C-L	✓	3.41	4.66	3.48	4.12	95.0
HuBERT-B	×	3.13	4.57	3.68	3.92	95.0
HuBERT-L	×	3.22	4.63	3.76	4.00	**96.0**
HuBERT-B	✓	3.40	4.62	3.47	4.13	95.0
HuBERT-L	✓	3.50	4.70	3.51	4.20	**96.0**
Whisper-M	×	2.58	4.16	3.30	3.42	94.0
Whisper-L	×	2.65	4.22	3.4	3.5	94.0
Whisper-M	✓	2.92	4.35	3.79	3.69	90.0
Whisper-L	✓	2.94	4.34	3.7	3.98	94.0
Tera	×	3.04	4.46	3.67	3.82	95.0
Tera	✓	3.22	4.52	3.37	3.95	95.0
Mockingjay	×	2.98	4.45	3.63	3.79	94.0
Mockingjay	✓	3.22	4.54	3.37	3.95	94.0
Unispeech-B	×	2.91	4.45	3.60	3.74	95.0
Unispeech-L	×	3.24	4.67	3.76	4.04	95.0
Unispeech-B	✓	3.36	4.68	3.44	4.10	95.0
Unispeech-L	✓	3.47	3.73	3.50	4.20	**96.0**
BEST-RQ	×	3.16	4.55	3.73	3.93	95.0
BEST-RQ	✓	3.42	4.64	3.46	4.12	95.0
WavLM-B	×	2.97	4.40	3.61	3.76	95.0
WavLM-L	×	3.29	4.66	3.79	4.08	**96.0**
WavLM-B	✓	3.43	4.66	3.48	4.13	95.0
WavLM-L	✓	**3.54**	4.73	3.53	**4.23**	**96.0**

PESQ scores below 3.0 even with PCS preprocessing. This suggests that the nature of pre-training, rather than just its scale, plays a crucial role in transfer learning for SE. Interestingly, our pre-trained BEST-RQ model achieves the highest performance among base-sized models without PCS preprocessing (PESQ 3.16), outperforming both spectrogram-based approaches like TERA (3.04) and waveform-based models like WavLM-B (2.97). This suggests that combining direct STFT processing with random-projection quantization may provide an effective solution for learning SE-related features. Examining model scale effects across foundation model architectures reveals that larger variants generally outperform their base counterparts. WavLM demonstrates the most substantial benefit from increased scale, with the Large model achieving PESQ 3.54 compared to Base model's 3.43 when using PCS. HuBERT and UniSpeech show similar improvements with their Large variants reaching PESQ 3.50 and 3.47, respectively, compared to 3.40 and 3.36 for their Base versions. Wav2Vec2.0 presents an exception where the Base model (PESQ 3.40 with PCS) outperforms the Large variant (PESQ 3.18), suggesting that increased model capacity does not universally guarantee better SE performance and may depend on the specific pre-training strategy.

TABLE 2. Performance on the DNS3 dataset. Training data size (data) is reported in hours. PCS-PP denotes PCS pre-processing while B and L suffixes denote Base and large model variants.

Model Name	PCS-PP	Data	PESQ-NB	PESQ-WB	STOI	SI-SDR
MFNET [105]	×	3000h	3.72	3.43	97.78	**20.31**
HPNET [106]	×	100h	3.79	3.40	**98.10**	-
TaEr [107]	×	3000h	3.60	3.26	97.56	19.40
INHSS [108]	×	300h	3.72	3.35	97.91	17.02
FRCNN [109]	×	3000h	3.60	3.23	97.69	19.78
HGCN+ [110]	×	3500h	3.65	3.19	97.20	-
GaGNet(Pro.) [111]	×	300h	3.56	3.17	97.13	18.91
Proposed FM-based Architecture						
W2V2-B	×	300h	3.69	3.25	97.60	18.60
W2V2-L	×	300h	3.70	3.28	97.64	19.00
W2V2C-L	×	300h	3.83	3.42	98.00	19.37
W2V2-L	✓	300h	3.77	3.40	97.47	10.58
W2V2C-L	✓	300h	3.85	3.49	97.80	10.63
HuBERT-B	×	300h	3.75	3.33	97.80	19.20
HuBERT-L	×	300h	3.84	3.44	98.00	19.20
HuBERT-L	✓	300h	3.87	3.52	97.77	10.62
Whisper-M	×	300h	2.95	2.36	94.70	14.90
Whisper-L	×	300h	3.03	2.46	95.30	15.90
Tera	×	300h	3.73	3.32	97.60	19.16
Mockingjay	×	300h	3.71	3.29	97.60	19.16
Unispeech-B	×	300h	3.76	3.34	97.70	19.14
Unispeech-L	×	300h	3.86	3.46	98.06	19.41
Unispeech-L	✓	300h	3.88	3.53	97.70	10.61
BEST-RQ	×	300h	3.80	3.41	98.00	19.40
BEST-RQ	✓	300h	3.84	3.49	97.60	10.60
WavLM-B	×	300h	3.67	3.22	97.50	18.91
WavLM-L	×	300h	3.88	3.48	**98.10**	19.48
WavLM-L	✓	300h	**3.89**	**3.55**	97.82	10.63

B. DNS CHALLENGE RESULTS

All DNS Challenge results reported in this work are evaluated on the DNS 2020 synthetic, non-reverberant test set (non-blind track) as specified in the dataset description (Section IV-A). Our FM-based models are trained on 300 hours of data, which represents 10% of the full 3000-hour DNS training corpus. Results on the DNS Challenge dataset (Table 2) reveal two significant findings. First, our FM-

based solutions achieve competitive results while using only 300 hours of training data, compared to previous approaches trained on 3,000 hours. WavLM-Large confirms to be the best-performing model and, when combined with the PCS pre-processing step, achieves a WB-PESQ of 3.55 and NB-PESQ of 3.89, approaching or outperforming specialized architectures. Second, PCS preprocessing demonstrates an interesting trade-off: while it consistently improves perceptual quality metrics (PESQ-NB and PESQ-WB), it leads to a significant decrease in SI-SDR scores. For instance, WavLM-Large sees its SI-SDR drop from 19.48 to 10.63 when PCS is applied, despite improvements in PESQ scores. This apparent contradiction reflects fundamentally different optimization objectives: SI-SDR measures signal reconstruction fidelity, while PESQ evaluates perceptual quality. PCS intentionally modifies the signal's spectral characteristics to enhance perceptually important frequency regions by design, which inherently leads to deviations from the original signal structure measured by SI-SDR. Given that SE primarily aims to improve human listening experience, this trade-off between perceptual quality and signal fidelity is often acceptable in practical applications where human perception is the ultimate evaluation criterion.

Examining the relative performance across model scales, BEST-RQ confirms its efficiency among base models, achieving competitive results (PESQ-WB 3.41, SI-SDR 19.40) that approach those of larger architectures. Similar to the VoiceBank-DEMAND results, we observe that SSL-based models generally outperform the WSL models, with both medium and large Whisper models achieving significantly lower scores across all metrics. The poor performance of this family of models can be attributed to the fundamental mismatch between ASR pre-training objectives and SE requirements. ASR tasks focus on extracting high-level semantic and phonetic content while discarding acoustic details that may be considered ''noise'' for recognition purposes. In contrast, SE requires preserving fine-grained spectral information and distinguishing between speech and noise at the acoustic level. Furthermore, Whisper's training focuses on clean speech transcription, potentially leading to representations less robust to spectral variations in noisy speech. This explains its consistent underperformance compared to SSL models trained on diverse acoustic conditions.

This establishes SSL models as more suitable for SE tasks than WSL approaches, confirming that pre-training objectives focused on diverse acoustic patterns rather than semantic content extraction are better aligned with enhancement task requirements.

C. ABLATION STUDY RESULTS

Tables 3, 4 and 5 report specialized analyses across key aspects: loss function components, architectural design choices, fusion methods and training data requirements. Our ablation studies use WavLM-Large as the foundation model backbone, as it demonstrated the strongest performance in our previous experiments.

TABLE 3. Ablation studies on the VoiceBank-DEMAND dataset.

Model Name	PCS-PP	PESQ-WB	CSIG	CBAK	COVL	STOI
WavLM-Large	×	3.29	4.66	**3.79**	4.08	**96.0**
WavLM-Large	✓	**3.54**	**4.73**	3.53	**4.23**	**96.0**
Loss Function Analysis						
w/o Mag. L1 Loss	×	3.25	4.62	**3.79**	4.02	**96.0**
w/o Mag. L1 Loss	✓	3.48	4.68	3.51	4.18	95.0
w/o Consistent Loss	×	3.23	4.63	3.77	4.01	**96.0**
w/o Consistent Loss	✓	3.45	4.65	3.49	4.15	95.0
w/o wSDR Loss	×	3.26	4.59	3.76	3.98	95.0
w/o wSDR Loss	✓	3.50	4.72	3.50	4.21	95.0
Architectural Components Analysis						
Linear SE Head	×	3.23	4.64	3.75	4.02	**96.0**
Linear SE Head	✓	3.47	4.71	3.49	4.19	95.0
FM Feat. Only	×	3.21	4.59	3.76	3.98	**96.0**
FM Feat. Only	✓	3.44	4.68	3.48	4.15	95.0
STFT Feat. Only	×	2.92	4.42	3.56	3.72	95.0
STFT Feat. Only	✓	3.20	4.52	3.36	3.93	95.0
Frame Duplication	×	3.19	4.59	3.73	3.97	95.0
Frame Duplication	✓	3.45	4.68	3.50	4.16	94.0

Loss Function Analysis: Ablation results in Table 3 reveal three key findings regarding loss components. First, removing the magnitude L1 loss causes a moderate decrease in PESQ (3.54 to 3.48) and CSIG (4.73 to 4.68) with PCS preprocessing. Second, removing the consistency loss leads to the most substantial degradation across metrics (PESQ drops to 3.45, CSIG to 4.65). This suggests that addressing signal reconstruction inconsistencies is crucial for high-quality SE. Third, the wSDR loss removal shows the smallest impact (PESQ 3.50 with PCS), suggesting this term provides a more limited contribution compared to other losses in our architecture, though it still contributes to the overall enhancement quality.

Architectural Component Analysis. The architectural components section of Table 3 demonstrates the impact of three key architectural decisions. First, replacing Conformer layers with linear layers in the enhancement head reduces PESQ from 3.54 to 3.47 with PCS preprocessing, demonstrating the importance of temporal modeling for SE. The relatively modest performance drop suggests that the FM's rich contextual information partially compensates for the simpler enhancement head, though the results still confirm that self-attention and convolution operations benefit the enhancement task. Feature integration analysis shows a clear hierarchy: FM features alone achieve a PESQ of 3.44 with PCS, substantially outperforming STFT features alone (PESQ 3.20). This 0.24 PESQ difference underlines the effectiveness of foundation models' learned representations. The further 0.10 PESQ improvement to 3.54 with combined features demonstrates clear complementarity: while FM features excel at modeling temporal dependencies and speech characteristics, STFT features provide explicit frequency-domain information crucial for precise spectral manipulation. This complementarity suggests that FM representations do not fully encode the fine-grained spectral details neces-

sary for optimal enhancement, validating our dual-stream architectural design. Finally, the adopted temporal alignment strategy outperforms the frame duplication approach used in previous work, improving PESQ from 3.45 to 3.54 with PCS. This 0.09 PESQ improvement shows that preserving distinct representations for each frame leads to better enhancement quality than duplicating the same representation across multiple frames, likely because it allows more precise modeling of the speech signal and its dynamics.

TABLE 4. Comparison of different fusion strategies for combining foundation model and spectral features on VoiceBank-DEMAND using WavLM-Large backbone. PCS-PP indicates whether perceptual contrast stretching preprocessing was applied.

Fusion Method	PCS-PP	PESQ-WB	CSIG	CBAK	COVL	STOI
Concatenation	×	3.29	4.66	**3.79**	4.08	**96.0**
Concatenation	✓	**3.54**	**4.73**	3.53	**4.23**	**96.0**
Gating	×	3.26	4.63	3.76	4.03	**96.0**
Gating	✓	3.46	4.68	3.50	4.17	95.0
Addition	×	3.28	4.63	3.78	4.05	**96.0**
Addition	✓	**3.54**	4.72	3.53	4.21	95.0
Multiplication	×	3.22	4.64	3.75	4.00	**96.0**
Multiplication	✓	3.46	4.71	3.53	4.17	95.0
Cross-Attention	×	3.27	4.64	3.78	4.05	**96.0**
Cross-Attention	✓	3.47	4.68	3.50	4.16	95.0

TABLE 5. Impact of training data size and PCS on DNS challenge performance using wavLM-large as backbone.

PCS-PP	Data	PESQ-NB	PESQ-WB	STOI	SI-SDR
×	100h	3.78	3.32	97.8	18.07
✓	100h	3.82	3.44	97.5	10.43
×	200h	3.82	3.41	97.9	19.18
✓	200h	3.86	3.51	97.6	10.56
×	300h	3.88	3.48	**98.1**	**19.48**
✓	300h	**3.89**	**3.55**	97.8	10.63

Feature Fusion Strategy Analysis: Table 4 presents a systematic comparison of different fusion strategies for combining foundation model and spectral features. Concatenation consistently achieves the highest performance across most metrics, with PESQ scores of 3.29 without PCS and 3.54 with PCS preprocessing. The gating mechanism, designed to learn adaptive combinations of the two feature types, performs slightly worse (PESQ 3.26/3.46), suggesting that the learned gates may not provide significant advantages over simple concatenation in this context. Additive fusion achieves competitive results (PESQ 3.28/3.54), indicating that the linear combination of features can be effective, though it doesn't surpass concatenation. Multiplicative fusion shows more substantial degradation (PESQ 3.22/3.46), likely because element-wise multiplication can suppress important information when feature magnitudes vary significantly between the two streams. Cross-attention fusion, despite its theoretical appeal for learning complex feature interactions, achieves moderate performance (PESQ 3.27/3.47), possibly due to the additional parameters requiring more training data to optimize effectively. These results validate our architectural choice of concatenation, which provides the best performance while maintaining computational simplicity and ensuring both feature streams contribute equally to the enhancement process.

Training Data Analysis: Table 5 examines how training data volume affects performance on the DNS Challenge. Increasing training data from 100h to 300h improves results on all metrics, with the largest gains occurring between 100h and 200h (PESQ-WB improves from 3.44 to 3.51). Importantly, the PCS preprocessing trade-off remains consistent across all training sizes: while it improves perceptual quality metrics (PESQ-NB and PESQ-WB), it reduces SI-SDR by approximately 8-9 dB regardless of the amount of training data, confirming this is an inherent characteristic of perceptual enhancement rather than a data-dependent effect.

PCS Application Timing Analysis: We investigate the impact of applying PCS at different stages of the training and inference pipeline. Specifically, we evaluate three configurations: (1) applying PCS during both training and inference (our proposed approach), (2) applying PCS during training only, and (3) using PCS during inference only. Results show that applying PCS during both training and inference (PESQ 3.54) outperforms training-only (PESQ 3.31) and inference-only (PESQ 3.49) approaches. The superior performance of the combined approach suggests that consistent perceptual enhancement during both phases enables the model to learn representations that effectively align with perceptual importance.

PCS Impact on Downstream Tasks: To further investigate this trade-off, we evaluated the impact of PCS on downstream ASR performance on VoiceBank-DEMAND test set.[3] We examined three conditions: applying PCS to original noisy speech, to clean reference speech, and to the complete enhancement pipeline where PCS was used during both training and inference. Notably, our WavLM-Large enhancement without PCS improves ASR performance, reducing word error rate from 5.33% (noisy input) to 4.90% (enhanced output). However, when PCS is applied, word error rates consistently increase across all conditions: noisy speech (5.33% to 5.35%), clean speech (4.44% to 4.68%), and enhanced speech (4.90% to 5.43%). The largest degradation occurs with enhanced speech (+0.53% absolute increase), indicating that while PCS improves perceptual SE metrics, the spectral modifications it introduces can disrupt the acoustic patterns that ASR systems rely on for accurate transcription. This demonstrates that perceptual enhancement and machine recognition accuracy represent different optimization objectives, emphasizing the need to consider specific application requirements when designing SE systems.

[3]ASR is run using `whisper-large-v3` pre-trained model.

VI. CONCLUSION

This work has established a systematic framework for adapting foundation models to speech enhancement, successfully closing the performance gap with specialized architectures through principled architectural design and training strategies. Rather than simply applying existing techniques, we provide a comprehensive methodology for understanding and addressing the fundamental challenges that limit foundation model effectiveness in SE tasks. Our results demonstrate that carefully adapted foundation models can match or exceed specialized SE architectures. The WavLM-Large backbone with perceptual contrast stretching achieved a PESQ score of 3.54 on VoiceBank-DEMAND and 3.55 on the DNS Challenge using only 10% of the standard training data. The strong performance of SSL-based models compared to supervised models suggests that self-supervised pre-training objectives better capture the acoustic features relevant to enhancement tasks. Our ablation studies revealed the importance of precise temporal alignment, feature combination between foundation model and spectral representations, and consistency-preserving training objectives.

These findings open two promising directions for future research. First, developing pre-training objectives specifically designed for speech quality and enhancement, rather than borrowing from ASR tasks, could substantially improve performance. Second, exploring architectures that can learn from multiple acoustic representations (waveform, spectrogram) simultaneously during pre-training could lead to better feature integration than our current approach of combining them afterwards. Overall, our results demonstrate that foundation models are effective building blocks for SE systems, offering competitive performance while reducing the need for specialized architectures.

REFERENCES

[1] I. Tashev, *Sound Capture and Processing*. New Jersey, NJ, USA: Wiley, 2009.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[3] X. Huang, A. Acero, and H-W. Hon, *Spoken Language Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[4] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*. Cham, Switzerland: Springer, 2005.

[5] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

[6] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

[7] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[8] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement for mobile communication based on dual-channel complex spectral mapping," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6134–6138.

[9] R. Sutherland, G. Close, T. Hain, S. Goetze, and J. Barker, "Using speech foundational models in loss functions for hearing aid speech enhancement," in *Proc. 32nd Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2024, pp. 421–425.

[10] R.-Y. Tseng, T.-W. Wang, S.-W. Fu, C.-Y. Lee, and Y. Tsao, "A study of joint effect on denoising techniques and visual cues to improve speech intelligibility in cochlear implant simulation," *IEEE Trans. Cognit. Develop. Syst.*, vol. 13, no. 4, pp. 984–994, Dec. 2021.

[11] N. Mamun and J. H. L. Hansen, "Speech enhancement for cochlear implant recipients using deep complex convolution transformer with frequency transformation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 32, pp. 2616–2629, 2024.

[12] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 5, pp. 1–28, Sep. 2018.

[13] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5024–5028.

[14] A. Pandey, C. Liu, Y. Wang, and Y. Saraf, "Dual application of speech enhancement for automatic speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 223–228.

[15] X. Chang, T. Maekaku, Y. Fujita, and S. Watanabe, "End-to-end integration of speech recognition, speech enhancement, and self-supervised learning representation," 2022, *arXiv:2204.00540*.

[16] Y.-J. Lu, X. Chang, C. Li, W. Zhang, S. Cornell, Z. Ni, Y. Masuyama, B. Yan, R. Scheibler, Z.-Q. Wang, Y. Tsao, Y. Qian, and S. Watanabe, "ESPNet-SE++: Speech enhancement for robust speech recognition, translation, and understanding," 2022, *arXiv:2207.09514*.

[17] Y. Koizumi, S. Karita, A. Narayanan, S. Panchapagesan, and M. Bacchiani, "SNRi target training for joint speech enhancement and recognition," in *Proc. Interspeech*, 2022, pp. 1173–1177, doi: 10.21437/Interspeech.2022-302.

[18] K. Kinoshita, T. Ochiai, M. Delcroix, and T. Nakatani, "Improving noise robust automatic speech recognition with single-channel time-domain enhancement network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7009–7013.

[19] I. Fedorov, M. Stamenovic, C. Jensen, L.-C. Yang, A. Mandell, Y. Gan, M. Mattina, and P. N. Whatmough, "TinyLSTMs: Efficient neural speech enhancement for hearing aids," 2020, *arXiv:2005.11138*.

[20] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, no. 3, pp. 197–210, Jun. 1978.

[21] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[23] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 1979, pp. 208–211.

[24] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[25] C. K. A. Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 deep noise suppression challenge," in *Proc. Interspeech*, 2021, pp. 2796–2800, doi: 10.21437/Interspeech.2021-1609.

[26] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, and R. Aichner, "ICASSP 2023 deep noise suppression challenge," in *Proc. ICASSP*, vol. 5, 2024, pp. 725–737.

[27] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[28] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[29] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 2015, pp. 436–444, Jul. 2006.

[30] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3734–3738.

[31] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.

[32] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[33] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proc. Int. Conf. Latent Variable Anal. Signal Separat.*, 2015, pp. 91–99.

[34] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Proc. Interspeech*, Sep. 2016, pp. 3768–3772.

[35] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 006–012.

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

[37] S. M. Siniscalchi, "Vector-to-vector regression via distributional loss for speech enhancement," *IEEE Signal Process. Lett.*, vol. 28, pp. 254–258, 2021.

[38] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[39] S. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019, pp. 2031–2041.

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

[41] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based metric GAN for speech enhancement," in *Proc. Interspeech*, Sep. 2022, pp. 936–940.

[42] V. Zadorozhnyy, Q. Ye, and K. Koishida, "SCP-GAN: Self-correcting discriminator optimization for training consistency preserving metric GAN on speech enhancement tasks," in *Proc. Interspeech*, Aug. 2023, pp. 2463–2467.

[43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, 2014, pp. 1–9.

[44] Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Statistical speech enhancement based on probabilistic integration of variational autoencoder and non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 716–720.

[45] Y. Bando, K. Sekiguchi, and K. Yoshii, "Adaptive neural speech enhancement with a denoising variational autoencoder," in *Proc. Interspeech*, Oct. 2020, pp. 2437–2441.

[46] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[47] J. Ho, A. N. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020, pp. 6840–6851.

[48] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7402–7406.

[49] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2351–2364, 2023.

[50] Z. Guo, J. Du, S. M. Siniscalchi, J. Pan, and Q. Liu, "Controllable conformer for speech enhancement and recognition," *IEEE Signal Process. Lett.*, vol. 32, pp. 156–160, 2025.

[51] P.-J. Ku, A. H. Liu, R. Korostik, S.-F. Huang, S.-W. Fu, and A. Jukic, "Generative speech foundation model pretraining for high-quality speech extraction and restoration," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[52] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A phase-and-harmonics-aware speech enhancement network," in *Proc. AAAI*, vol. 34, 2020, pp. 9458–9465.

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

[54] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Workshop Speech Synth. Workshop (SSW 9)*, Sep. 2016, pp. 146–152.

[55] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449–12460.

[56] W.-N. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 3451–3460, 2021.

[57] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022.

[58] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 28492–28518.

[59] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K.-T. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-Y. Lee, "SUPERB: Speech processing universal PERformance benchmark," 2021, *arXiv:2105.01051*.

[60] M. L. Quatra, A. Koudounas, L. Vaiani, E. Baralis, L. Cagliero, P. Garza, and S. M. Siniscalchi, "Benchmarking representations for speech, music, and acoustic events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops (ICASSPW)*, Apr. 2024, pp. 505–509.

[61] J. Turian et al., "HEAR: Holistic evaluation of audio representations," in *Proc. NeurIPS Competitions Demonstrations Track*, 2022, pp. 125–145.

[62] K. R. Bagadi and C. M. R. Sivappagari, "A robust feature selection method based on meta-heuristic optimization for speech emotion recognition," *Evol. Intell.*, vol. 17, no. 2, pp. 993–1004, Apr. 2024.

[63] K. R. Bagadi and C. M. R. Sivappagari, "An evolutionary optimization method for selecting features for speech emotion recognition," *TELKOMNIKA (Telecommun. Comput. Electron. Control)*, vol. 21, no. 1, pp. 159–167, Feb. 2023.

[64] A. Koudounas, M. La Quatra, S. M. Siniscalchi, and E. Baralis, "voc2vec: A foundation model for non-verbal vocalization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2025, pp. 1–5.

[65] K. R. Bagadi, "A comprehensive analysis of multimodal speech emotion recognition," *J. Phys., Conf. Ser.*, vol. 1917, no. 1, Jun. 2021, Art. no. 012009.

[66] V. Miara, T. Lepage, and R. Dehak, "Towards supervised performance on speaker verification with self-supervised learning by leveraging large-scale ASR models," in *Proc. Interspeech*, Sep. 2024, pp. 2660–2664.

[67] M. L. Quatra, A. Koudounas, E. Baralis, and S. M. Siniscalchi, "Speech analysis of language varieties in Italy," in *Proc. Joint Int. Conf. Comput. Linguistics*, May 2024, pp. 15147–15159.

[68] B. Lee, I. Calapodescu, M. Gaido, M. Negri, and L. Besacier, "Speech-MASSIVE: A multilingual speech dataset for SLU and beyond," in *Proc. Interspeech*, Sep. 2024, pp. 817–821.

[69] A. Koudounas, M. La Quatra, L. Vaiani, L. Colomba, G. Attanasio, E. Pastor, L. Cagliero, and E. Baralis, "ITALIC: An Italian intent classification dataset," in *Proc. Interspeech*, Aug. 2023, pp. 2153–2157.

[70] K.-H. Hung, S.-W. Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting self-supervised embeddings for speech enhancement," in *Proc. Interspeech*, Sep. 2022, pp. 186–190.

[71] H. Song, S. Chen, Z. Chen, Y. Wu, T. Yoshioka, M. Tang, J. W. Shin, and S. Liu, "Exploring WavLM on speech enhancement," in *Proc. IEEE Spoken Language Technol. Workshop (SLT)*, Jan. 2023, pp. 451–457.

[72] Y. Lee and K. Jung, "Boosting speech enhancement with clean self-supervised features via conditional variational autoencoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 12396–12400.

[73] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," 2021, *arXiv:2111.09296*.

[74] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *Proc. Interspeech*, Aug. 2023, pp. 2798–2802.

[75] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, "Unispeech-sat: Universal speech representation learning with speaker aware pre-training," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6152–6156.

[76] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6419–6423.

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

[78] C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 3915–3924.

[79] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020, *arXiv:2005.08100*.

[80] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," *Dim*, vol. 1, no. 4, pp. 1–5, 2020.

[81] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7118–7122.

[82] E. Kim and H. Seo, "SE-conformer: Time-domain speech enhancement using conformer," in *Proc. Interspeech*, Aug. 2021, pp. 2736–2740.

[83] M. S. Khan, M. L. Quatra, K.-H. Hung, S.-W. Fu, S. M. Siniscalchi, and Y. Tsao, "Exploiting consistency-preserving loss and perceptual contrast stretching to boost SSL-based speech enhancement," in *Proc. IEEE 26th Int. Workshop Multimedia Signal Process. (MMSP)*, Oct. 2024, pp. 1–6.

[84] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 900–904.

[85] R. Chao, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Perceptual contrast stretching on target feature for speech enhancement," in *Proc. Interspeech*, Sep. 2022, pp. 5448–5452.

[86] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2019, pp. 4171–4186.

[87] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 5206–5210.

[88] C.-H. Lee, C. Yang, R. S. Srinivasa, Y. Malur Saidutta, J. Cho, Y. Shen, and H. Jin, "Leveraging self-supervised speech representations for domain adaptation in speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 10831–10835.

[89] Z. Huang, S. Watanabe, S.-w. Yang, P. García, and S. Khudanpur, "Investigating self-supervised learning for speech enhancement and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6837–6841.

[90] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex U-Net," 2019, *arXiv:1903.03107*.

[91] *American National Standard: Methods for Calculation of the Speech Intelligibility Index*, American National Standards Institute, Washington, DC, USA, 1997.

[92] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCOSDA Held Jointly Conf. Asian Spoken Lang. Res. Eval.*, Nov. 2013, pp. 1–4.

[93] 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. Meetings Acoust.*, Jun. 2013, vol. 19, no. 1, Paper 035081. [Online]. Available: https://doi.org/10.1121/1.4799597

[94] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 2015, *arXiv:1412.6980*.

[95] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Nov. 2001, pp. 749–752.

[96] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[97] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4214–4217.

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

[99] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, Aug. 2021, pp. 201–205.

[100] F. Dang, H. Chen, and P. Zhang, "DPT-FSNet: Dual-path transformer based full-band and sub-band fusion network for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 6857–6861.

[101] D. Yin, Z. Zhao, C. Tang, Z. Xiong, and C. Luo, "TridentSE: Guiding speech enhancement with 32 global tokens," in *Proc. INTERSPEECH*, Aug. 2023, pp. 3839–3843.

[102] A. Mattursun, L. Wang, and Y. Yu, "BSS-CFFMA: Cross-domain feature fusion and multi-attention speech enhancement network based on self-supervised embedding," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2024, pp. 3589–3594.

[103] G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "Perceive and predict: Self-supervised speech representation based loss functions for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[104] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss using Wasserstein distance for speech enhancement," 2020, *arXiv:2010.15174*.

[105] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, and S. Ding, "A mask free neural network for monaural speech enhancement," in *Proc. Interspeech*, Aug. 2023, pp. 2468–2472.

[106] T. Wang, W. Zhu, Y. Gao, S. Zhang, and J. Feng, "Harmonic attention for monaural speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 2424–2436, 2023.

[107] A. Li, G. Yu, C. Zheng, W. Liu, and X. Li, "A general unfolding speech enhancement method motivated by Taylor's theorem," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 3629–3646, 2023.

[108] W. Jiang and K. Yu, "Speech enhancement with integration of neural homomorphic synthesis and spectral masking," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 1758–1770, 2023.

[109] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9281–9285.

[110] T. Wang, W. Zhu, Y. Gao, Y. Chen, J. Feng, and S. Zhang, "Harmonic gated compensation network plus for ICASSP 2022 DNS challenge," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 9286–9290.

[111] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Appl. Acoust.*, vol. 187, Feb. 2022, Art. no. 108499.

MUHAMMAD SALMAN KHAN (Member, IEEE) received the master's degree in electrical engineering from the Institute of Space and Technology, Islamabad, Pakistan, in 2020. He is currently pursuing the Ph.D. degree with the University of Enna "Kore." He has also worked in various positions in the telecom industry for three years. His research interests include speech processing and enhancement.

VALERIO MARIO SALERNO received the bachelor's degree in telematics engineering from the University of Catania, Catania, Italy, and the master's degree (cum laude) in telematics engineering and the Ph.D. degree from the University of Enna "Kore," Enna, Italy. He is currently an Associate Professor and the Laboratory Head of the Speech Technology and Machine Learning Laboratory (STMLab), University of Enna "Kore."

MORENO LA QUATRA (Member, IEEE) received the dual M.Sc. degree in computer engineering from Politecnico di Torino and ENSIMAG, in 2018, and the Ph.D. degree in computer engineering from Politecnico di Torino, Italy, in 2022. He is currently an Assistant Professor with the Speech Technology and Machine Learning Laboratory (STMLab), University of Enna "Kore," Italy. His research interests include deep learning, speech and audio processing, and natural language processing. He serves as a reviewer for several IEEE and ACM journals and actively contributes to the speech research community through workshop organization and conference activities. His recent work focuses on representation learning for audio signals, speech-based health assessment, and post-ASR generative error correction.

KUO-HSUAN HUNG received the B.S. degree from National Chiao Tung University, in 2015, and the M.S. degree from National Central University, Taiwan, in 2017. He is currently pursuing the Ph.D. degree with the Department of Biomedical Engineering, National Taiwan University. He is also a Research Assistant with the Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan. His research interests include biomedical and speech signal processing.

SZU-WEI FU received the M.S. degree from the Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, in 2014, and the Ph.D. degree from the Department of Computer Science and Information Engineering, National Taiwan University, in 2020. From 2021 to 2022, he was a Senior Applied Scientist with Microsoft, Vancouver, Canada. He is currently a Senior Research Scientist with Nvidia Corporation. His research interests include speech processing, speech generation and enhancement, and speech quality estimation. He was the first author of an article that received the 2021 IEEE Signal Processing Society (SPS) Young Author Best Paper Award.

YU TSAO (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taipei, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 2008. From 2009 to 2011, he was a Researcher with the National Institute of Information and Communications Technology, Tokyo, Japan, where he worked on research and product development for automatic speech recognition in multilingual speech-to-speech translation. He is currently a Research Fellow (Professor) and the Deputy Director of the Research Center for Information Technology Innovation, Academia Sinica, Taipei. Additionally, he serves as a Jointly Appointed Professor with the Department of Electrical Engineering, Chung Yuan Christian University, Taoyuan, Taiwan. His research interests include assistive oral communication technologies, audio coding, and bio-signal processing. In recognition of his contributions, he received the Outstanding Research Award from the National Science and Technology Council (NSTC)—the most prestigious research honor in Taiwan, in 2023. His other accolades include the Academia Sinica Career Development Award, in 2017, the multiple National Innovation Awards, from 2018 to 2021 and 2023, the Future Tech Breakthrough Award, in 2019, the Outstanding Elite Award from the Chung Hwa Rotary Educational Foundation, from 2019 to 2020, and the NSTC FutureTech Award, in 2022. Additionally, he is the corresponding author of an article that won the 2021 IEEE Signal Processing Society (SPS) Young Author Best Paper Award. He is an Associate Editor of IEEE TRANSACTIONS ON CONSUMER ELECTRONICS and IEEE SIGNAL PROCESSING LETTERS.

SABATO MARCO SINISCALCHI (Senior Member, IEEE) received the Laurea and Ph.D. degrees in computer engineering from the University of Palermo, Italy, in 2001 and 2006, respectively. He is currently a Professor with the University of Palermo, an Adjunct Professor with the Norwegian University of Science and Technology, and affiliated with the Georgia Institute of Technology. In 2006, he was a Postdoctoral Fellow with Georgia Institute of Technology, USA. From 2007 to 2009, he joined the Norwegian University of Science and Technology (NTNU) as a Research Scientist. From 2010 to 2015, he was an Assistant Professor, and an Associate Professor with the University of Enna "Kore." From 2017 to 2018, he joined as a Senior Speech Researcher with the Siri Speech Group, Apple Inc., Cupertino, CA, USA. In 2018, he joined the University of Enna "Kore" as a Full Professor. He is an elected member of the IEEE Speech and Language Technical Committee, from 2019 to 2021 and from 2024 to 2026. He acted as an Associate Editor of IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, from 2015 to 2019. He acts as a Senior Area Editor of IEEE SIGNAL PROCESSING LETTERS.