Phoneme-Preserving Mamba-Based Frontend for Noise-Robust Automatic Speech Recognition

Reshma K R, Anil Kumar Vuppala
LTRC, International Institute of Information Technology Hyderabad, India
reshmak.r@research.iiit.ac.in, anil.vuppala@iiit.ac.in

***Abstract*—Speech enhancement models optimised for perceptual quality metrics such as PESQ do not necessarily improve automatic speech recognition (ASR) performance, as aggressive noise suppression can distort the phonemic cues that acoustic models rely on. We propose a time-domain speech frontend based on the Mamba selective state space model, trained exclusively with ASR-oriented loss functions to preserve phonetically discriminative structure rather than maximise perceptual quality. The model follows a U-Net encoder-decoder architecture with $4\times$ temporal downsampling and six stacked bidirectional Mamba blocks operating on the raw waveform, trained with a multi-component loss combining multi-resolution spectral supervision, log-mel feature matching, speech correlation preservation, and ideal-ratio-mask spectral objectives. On VoiceBank-DEMAND, the proposed model achieves 10.26% WER with `wav2vec2-base-960h` and 7.68% with `whisper-base`, achieving statistically equivalent WER to SEMamba (10.30% / 7.73%) while producing a lower PESQ of 3.10 versus 3.69. Phoneme-class analysis confirms that the proposed model matches or outperforms SEMamba on five of seven articulatory classes, demonstrating that ASR-oriented training objectives achieve equivalent robustness to perceptual enhancement models without modifying the downstream ASR system.**

***Index Terms*—Noise-robust speech recognition, Mamba, selective state space models, time-domain processing, ASR frontend, phoneme preservation**

I. Introduction

Robust automatic speech recognition (ASR) in noisy environments remains a difficult problem. Recent self-supervised learning (SSL) models such as wav2vec 2.0 [1] and HuBERT [2] have shown strong results on clean speech benchmarks. However, their performance often drops when the input speech is corrupted by environmental noise, reverberation, or channel distortions. Large-scale weakly supervised models such as Whisper [26] generally show better robustness because they are trained on more diverse speech data. However, their performance still degrades at low signal-to-noise ratios.

A common approach to mitigate this problem is to use a speech enhancement (SE) frontend that reduces noise before the signal is passed to the ASR system. Neural SE models such as SEGAN [3], MetricGAN+ [4], CMGAN [5], MP-SENet [27], and SEMamba [6] achieve strong results on standard benchmarks measured by perceptual metrics such as PESQ [16] and STOI [17]. However, these metrics reflect human intelligibility and do not reliably predict ASR performance. Iwamoto et al. [7] demonstrated on VoiceBank-DEMAND that PESQ improvement does not consistently correlate with WER reduction, over-suppression artifacts can distort the stop bursts, fricative energy, and nasal formants that acoustic models rely on for phoneme discrimination, simultaneously raising PESQ while degrading WER.

Several recent works have tried to close this gap more directly through end-to-end ASR objectives [9], joint frontend–backend training [11], [12], post-hoc bridging modules [24], or MAP-based regularisation [25], though these methods require differentiable ASR access or joint optimisation, limiting generalisation across backends. Data augmentation strategies such as SpecAugment [13] and multi-condition training [14] improve robustness without a frontend but require retraining the ASR system. Dissen et al. [8] propose a lightweight U-Net trained with frozen Whisper cross-entropy loss, directly targeting WER without modifying the ASR model. The model of Dissen et al. is retrained on VoiceBank-DEMAND and obtains 11.59% WER with Whisper-base, demonstrating that ASR-gradient-based front-ends optimised for specific noise types do not generalise to complex environmental noise.

This work proposes an *independent* time-domain frontend trained exclusively with ASR-oriented signal-level losses, requiring no ASR model access during training and compatible with any downstream recogniser without modification. The model operates on raw waveforms throughout, motivated by the fact that SSL-based ASR models process raw waveforms through a convolutional feature extractor making temporal fine structure directly relevant to recognition performance. The Mamba selective state space model [10] is adopted as the sequence modelling backbone, which provides linear-time complexity over long audio sequences, and has shown strong capability in recent speech processing work [6]. The training loss is structured to preserve acoustic features relevant to both CTC-based SSL models and attention encoder–decoder models, without optimising for perceptual quality.

The main contributions are:

- A time-domain U-Net frontend with a bidirectional Mamba bottleneck, trained with ASR-oriented signal losses and compatible with any downstream ASR system without modification.
- A structured multi-component loss combining multi-resolution STFT supervision, log-mel feature matching, waveform correlation preservation, and ideal-ratio-mask spectral supervision, each motivated by the acoustic features exploited by SSL-based and attention-based ASR models.

- Empirical evidence on VoiceBank-DEMAND that the proposed model (PESQ 3.10) achieves statistically equivalent WER to SEMamba (PESQ 3.55) under wav2vec2-base-960h (10.26% vs. 10.30%), and outperforms the ASR-gradient-trained frontend of Dissen et al. [8] on Whisper-base (7.68% vs. 11.59%), showing that signal-domain ASR-oriented losses generalise more effectively to complex noise.
- A phoneme-class error analysis across major articulatory classes under two ASR backends: wav2vec2-base-960h (CTC-based SSL) and Whisper-base (attention encoder–decoder).

II. Proposed Method

A. Time-Domain Processing

A key design choice is to operate entirely in the *time domain*. SSL-based ASR models such as wav2vec 2.0 process raw waveforms through a convolutional feature extractor before the transformer encoder. The features most relevant to phoneme discrimination – temporal envelope modulations, fine structure, onset transients are naturally represented in the waveform. Operating in the time domain avoids the lossy phase reconstruction step required by frequency-domain models and preserves temporal structure that STFT-domain processing may distort. The proposed model takes a raw waveform as input and produces a raw waveform as output; all spectral analysis occurs only within the loss functions during training.

B. Architecture

The overall pipeline is illustrated in Figure 1. Given a noisy waveform $\mathbf{x} \in \mathbb{R}^T$, the model generates an enhanced waveform $\hat{\mathbf{x}} \in \mathbb{R}^T$ of identical length.

1) Encoder: The encoder consists of three 1-D convolutional layers with GELU activation [22]. The first layer expands to $C/2 = 64$ channels; the second and third layers use stride 2, reducing the temporal dimension by a factor of four ($T \to T/4$) while expanding to $C = 128$ channels. The final representation is then transposed and layer-normalized before it enters the Mamba bottleneck.

2) Bidirectional Mamba Bottleneck: The bottleneck uses $N = 6$ stacked bidirectional Mamba blocks. Each processes the sequence in both directions to capture complete temporal context essential for resolving phonemes masked by noise.

Each Mamba block [10] maintains a hidden state updated through a discretized state space model with input-dependent selection, letting the model choose what information to keep or discard. d_{state} is set to 16 with expand factor 4.

Each bidirectional block runs a forward (causal) and backward (sequence-flipped) Mamba sub-block in parallel, concatenates their outputs, projects back to dimension C through a linear layer, then adds a residual connection with layer normalization. These six blocks are stacked sequentially.

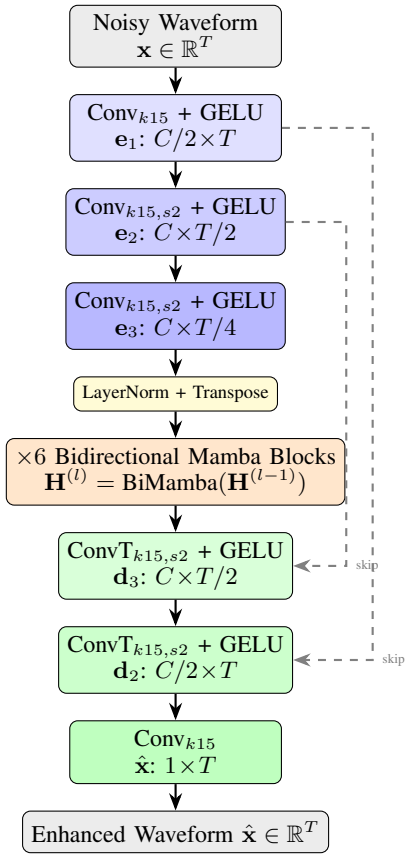

Fig. 1. Architecture of the proposed time-domain frontend. Skip connections from encoder to decoder are via concatenation.

3) Decoder: Mirroring the encoder, the decoder uses two transposed convolutional layers (stride 2, GELU) to recover the original temporal resolution. Skip connections from $\mathbf{e}_2$ and $\mathbf{e}_1$ are concatenated at corresponding decoder stages following the U-Net architecture [21], preserving fine temporal details lost in downsampling. A final convolution produces the single-channel output $\hat{\mathbf{x}} \in \mathbb{R}^T$.

C. ASR-Oriented Loss Functions

Let $\hat{\mathbf{x}}, \mathbf{x}_c \in \mathbb{R}^T$ denote the enhanced and clean waveforms respectively, and $\mathbf{x}_n \in \mathbb{R}^T$ the noisy input. $\hat{X}(f,t)$, $X_c(f,t)$, $X_n(f,t)$ denote their STFT representations, $|\cdot|$ the magnitude, $\bar{\cdot}$ the temporal mean, and σ the standard deviation. The total loss is a weighted sum:

$$\mathcal{L} = \sum_k w_k \, \mathcal{L}_k. \tag{1}$$

1) Multi-Resolution STFT Loss: Different phoneme classes concentrate energy at different temporal and spectral scales, stop bursts require fine temporal resolution while vowel formants require fine spectral resolution. Following [20], spectral losses are computed at three STFT resolutions $r \in$

$\{(512, 128), (1024, 256), (2048, 512)\}$ (FFT size, hop length):

$$\mathcal{L}_{\text{MRSTFT}} = \frac{1}{R}\sum_{r=1}^{R}\left[\| \, |\hat{X}_r| - |X_{c,r}| \, \|_1 + \frac{\| \, |X_{c,r}| - |\hat{X}_r| \, \|_F}{\| \, |X_{c,r}| \, \|_F + \epsilon}\right], \tag{2}$$

where $\| \cdot \|_F$ is the Frobenius norm. The spectral convergence term normalises by target energy, ensuring equal supervision across voiced and unvoiced phonemes.

2) Log-Mel Spectrogram Loss: The convolutional feature extractor of wav2vec 2.0 learns representations closely related to log-mel filterbank features. Minimising the distance in log-mel space therefore directly targets the input representation of the downstream SSL ASR model:

$$\mathcal{L}_{\text{mel}} = \| \log(\mathcal{M}(\hat{\mathbf{x}}) + \epsilon) - \log(\mathcal{M}(\mathbf{x}_c) + \epsilon) \|_1, \tag{3}$$

where $\mathcal{M}(\cdot)$ denotes 80-band mel filterbank extraction (FFT 512, hop 128). This loss provides targeted supervision for nasal formant structure and vowel quality in the lower mel bands.

3) Spectral Loss: A complementary log-mel loss computed at a coarser resolution reinforces frequency-level accuracy across the full spectral range:

$$\mathcal{L}_{\text{spec}} = \| \log(\mathcal{M}_s(\hat{\mathbf{x}}) + \epsilon) - \log(\mathcal{M}_s(\mathbf{x}_c) + \epsilon) \|_1, \tag{4}$$

where $\mathcal{M}_s(\cdot)$ denotes 80-band mel filterbank extraction at a coarser resolution (FFT 1024, hop 256), providing supervision at a longer temporal scale than $\mathcal{M}(\cdot)$. Together, $\mathcal{L}_{\text{mel}}$ and $\mathcal{L}_{\text{spec}}$ cover complementary temporal resolutions of the mel feature space used by both SSL and attention-based ASR encoders.

4) Speech Preservation Loss: Aggressive noise suppression can over-attenuate speech energy, distorting the temporal fine structure that SSL models rely on. Normalised cross-correlation is enforced between the enhanced and clean waveforms:

$$\mathcal{L}_{\text{sp}} = 1 - \frac{1}{T}\sum_{t=1}^{T}\frac{(\hat{x}_t - \bar{\hat{x}})\,(x_{c,t} - \bar{x}_c)}{\sigma_{\hat{x}}\,\sigma_{x_c} + \epsilon}, \tag{5}$$

where $\bar{\hat{x}}, \bar{x}_c$ are temporal means and $\sigma_{\hat{x}}, \sigma_{x_c}$ are standard deviations. The loss ranges from 0 (perfect correlation) to 2 (anti-correlation), penalising temporal misalignments and phase reversals that would disrupt the convolutional feature extractor of wav2vec 2.0 even when spectral energy appears preserved.

5) Spectral Masking Loss: Drawing on the ideal ratio mask (IRM) framework [19], the target mask and predicted mask are defined as:

$$M^*_{f,t} = \frac{|X_c(f,t)|}{|X_n(f,t)| + \epsilon}, \quad \hat{M}_{f,t} = \frac{|\hat{X}(f,t)|}{|X_n(f,t)| + \epsilon}, \tag{6}$$

both clamped to $[0, 1]$, and the loss is minimised:

$$\mathcal{L}_{\text{mask}} = \|M^* - \hat{M}\|_F^2. \tag{7}$$

This encourages the model to preserve speech-dominant time-frequency bins and suppress noise-dominant ones, directly aligned with how SSL models form representations from spectral patterns.

6) Auxiliary Losses: An L1 waveform loss $\mathcal{L}_{\text{L1}} = \frac{1}{T}\sum_t |\hat{x}_t - x_{c,t}|$ anchors sample-level reconstruction. A frequency-weighted spectral loss $\mathcal{L}_{\text{perc}}$ applies double weight to the 300–3400 Hz band containing the majority of speech energy. These are standard components shared with SE models [4], [5] and are included for signal stability rather than as primary ASR-oriented supervision.

The final weights were determined through iterative ablation on validation WER: components were added sequentially and weights tuned to maximise WER improvement, with w_{sp} assigned the highest weight as temporal waveform preservation was found most critical for ASR performance. The final weights are: $w_{\text{MRSTFT}} = 1.0$, $w_{\text{sp}} = 2.0$, $w_{\text{spec}} = 0.62$, $w_{\text{mel}} = 0.4$, $w_{\text{mask}} = 0.3$, $w_{\text{L1}} = 0.05$, $w_{\text{perc}} = 0.07$.

III. Experimental Setup

A. Dataset

We use the VoiceBank-DEMAND dataset [15]: 11,572 training utterances (28 speakers, SNR $\in \{0, 5, 10, 15\}$ dB) and 824 test utterances (SNR $\in \{2.5, 7.5, 12.5, 17.5\}$ dB) at 16 kHz. 10% of the training data is reserved for validation using a fixed random seed.

B. Training

The model is trained for up to 200 epochs using AdamW [23] (lr $= 2.5 \times 10^{-4}$, weight decay 5×10^{-5}) with a linear warmup cosine decay schedule (3 warmup epochs) and gradient accumulation over 4 steps (effective batch 4). Audio is chunked to 3-second segments during training. An exponential moving average (EMA) of model weights (decay 0.999) is maintained for evaluation. All training uses float32 precision on a single NVIDIA RTX 2080 Ti (11 GB).

C. Evaluation Protocol

Speech quality: PESQ [16], STOI [17], SI-SDR [18], and SNR, computed against clean references on full-length test utterances.

ASR performance: WER is computed using two frozen pretrained backends. (1) `facebook/wav2vec2-base-960h` [1]: a CTC-based SSL model pretrained on LibriSpeech, fine-tuned with CTC, operating directly on raw waveforms with greedy decoding. (2) `openai/whisper-base` [26]: a weakly-supervised encoder-decoder model trained on 680,000 hours, encoding log-mel spectrograms with attention-based decoding forced to English. Evaluating both backends tests whether the frontend is genuinely backend-agnostic; importantly, neither backend is fine-tuned or modified. Additionally, the Dissen et al. frontend (retrained on VoiceBank-DEMAND) and several SE baselines are evaluated using the same protocol.

Phoneme analysis: Word-level alignments are mapped to phoneme classes via the CMU Pronouncing Dictionary. Error rates (substitutions + deletions) are reported per class: stops, fricatives, affricates, nasals, liquids, glides, and vowels.

TABLE I
SPEECH QUALITY ON VOICEBANK-DEMAND TEST SET. † FROM RESPECTIVE PAPERS; OTHERS EVALUATED ON RELEASED MODELS.

Model	PESQ	STOI	SI-SDR	SNR
Noisy	1.97	0.92	8.45	9.45
MetricGAN+† [4]	3.15	0.93	7.96	8.75
MP-SENet† [27]	3.60	0.96	19.40	17.47
FlowMSE† [29]	3.12	0.88	18.95	19.03
StoRM† [28]	2.93	0.88	16.86	16.97
SEMamba† [6]	3.69	0.96	19.72	19.78
Proposed	**3.10**	**0.95**	**18.86**	**18.93**

TABLE II
WER (%) ON VOICEBANK-DEMAND TEST SET. ALL MODELS EVALUATED WITH THE SAME FROZEN ASR BACKENDS. † INDICATES MODEL RETRAINED ON VOICEBANK-DEMAND.

Model	wav2vec2-base	Whisper-base
Clean speech	8.26	6.02
Noisy speech	14.60	8.73
MetricGAN+ [4]	14.01	10.05
FlowMSE [29]	11.72	9.32
StoRM [28]	13.12	9.58
MP-SENet [27]	**10.03**	8.00
SEMamba [6]	10.30	7.73
Dissen et al.† [8]	—	11.59
Proposed	**10.26**	**7.68**

IV. RESULTS

A. Speech Quality Metrics

Table I shows that the model achieves PESQ 3.10 and STOI 0.95. As intended, PESQ is lower than SEMamba (3.69) because the model does not optimise for perceptual quality. The STOI score of 0.95 is close to SEMamba (0.96) and indicates strong speech intelligibility preservation, reflecting the degree of speech structure retained after enhancement [17].

B. ASR Performance

Table II shows ASR results across all models. The proposed model achieves 10.26% WER with wav2vec2-base, matching SEMamba (10.30%) within the evaluation variance for this benchmark (±0.1–0.15%), and 7.68% with Whisper-base versus SEMamba's 7.73%. The consistent results across both backends confirm that the multi-component loss design preserves acoustic features relevant to both recognition paradigms.

Notably, the Dissen et al. frontend [8], which is explicitly trained with ASR cross-entropy loss from a frozen Whisper model, achieves 11.59% with Whisper; underperforming the proposed model (7.68%) despite having direct access to ASR gradients. This demonstrates that their approach, designed for white noise and packet loss, does not generalise well to the complex environmental noise in VoiceBank-DEMAND, and that signal-domain ASR-oriented losses can be more effective than direct ASR loss minimisation in this setting.

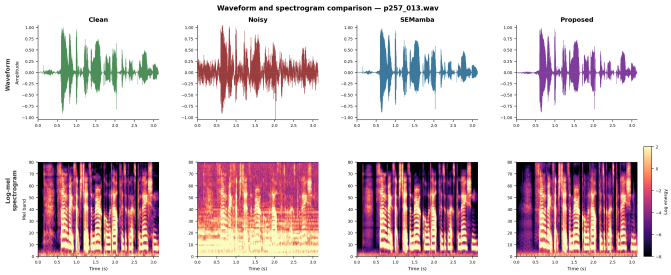

Fig. 2. Waveform and log-mel spectrogram for utterance *"some have accepted it as a miracle without physical explanation."* Proposed recovers the reference exactly (WER: 0%), while SEMamba introduces three word errors (WER: 50%). The proposed spectrogram preserves formant structure and phoneme boundaries closer to the clean reference, while SEMamba over-suppresses mid-frequency energy in the 0.5–1.5 s region.

TABLE III
PHONEME-CLASS ERROR RATES (%) WITH WHISPER-BASE. S = SUBSTITUTIONS, D = DELETIONS PER REFERENCE WORDS IN CLASS.

Class	SEMamba		Proposed	
	Err%	S/D	Err%	S/D
Stops	6.1	164/20	**5.7**	158/16
Fricatives	6.3	168/27	**6.2**	164/32
Affricates	4.6	9/0	6.6	13/0
Nasals	6.4	104/15	**6.3**	101/16
Liquids	6.9	94/18	**6.8**	95/16
Glides	9.7	88/13	9.9	91/12
Vowels	6.7	343/66	**6.6**	344/64

Among the compared models, MP-SENet achieves 10.03% WER with wav2vec2-base at a higher PESQ of 3.60, suggesting that phase-accurate frequency-domain reconstruction may also implicitly benefit ASR performance [27].

C. PESQ–WER Relationship

SEMamba achieves PESQ 3.69 with 10.30% WER while the proposed model achieves PESQ 3.10 with 10.26% WER. The 0.59-point PESQ gap corresponds to only 0.04% WER difference, well within benchmark variance (±0.1–0.15%), confirming that optimising for PESQ beyond a threshold does not translate to ASR gains [7].

D. Qualitative Analysis

Figure 2 illustrates the signal-level difference. SEMamba exhibits heavy waveform suppression (amplitude rarely exceeding ±0.25) with significant energy loss in low mel bands (10–30), attenuating vowel formants and nasal resonances. The proposed model maintains amplitude fidelity (±0.75) close to the clean reference, preserving harmonic structure in voiced regions and high-frequency content (mel bands 60–80), consistent with the speech preservation loss preventing over-attenuation across phoneme boundaries.

E. Phoneme-Class Error Analysis

Table III breaks down errors by phoneme class using Whisper-base. On five of seven classes – stops (5.7% vs. 6.1%), nasals (6.3% vs. 6.4%), liquids (6.8% vs. 6.9%), and

vowels (6.6% vs. 6.7%) – the error rates are lower than those of SEMamba, with fricatives and glides approximately tied. The only class where SEMamba holds a clear advantage is affricates (4.6% vs. 6.6%), which account for only 197 of 17,029 reference words (1.2%). Affricates combine a stop burst and fricative release, making them acoustically complex transients. Their degradation suggests that the speech preservation loss, which enforces global waveform correlation, may insufficiently constrain brief transient events. Targeted high-frequency mel supervision in future work may address this. Fewer stop deletions (16 vs. 20) and fewer vowel deletions (64 vs. 66) are observed, confirming that the speech preservation and spectral masking losses prevent over-suppression of phoneme boundaries.

V. Conclusion

This paper presented a time-domain Mamba-based frontend for noise-robust ASR, trained with ASR-oriented signal losses rather than perceptual quality objectives. The model achieves 10.26% WER with wav2vec2-base and 7.68% with Whisper-base, achieving statistically equivalent WER to SEMamba (10.30% / 7.73%) while producing a lower PESQ of 3.10 versus 3.69, demonstrating that phoneme preservation through targeted signal supervision is sufficient for ASR robustness without aggressive perceptual enhancement. Phoneme-class analysis confirms that the proposed model matches or outperforms SEMamba on five of seven articulatory classes, providing direct evidence that the WER equivalence reflects genuine phoneme preservation. The model also outperforms the ASR-gradient-trained frontend of Dissen et al. on Whisper-base (7.68% vs. 11.59%); noting that their method was designed for white noise and packet loss, this result demonstrates that signal-domain ASR-oriented losses generalise more effectively across noise conditions than task-specific ASR gradient optimisation.

Future work will focus on generalisation to diverse and unseen noise conditions beyond the DEMAND noise types, and reverberation, model optimisation through quantisation and knowledge distillation for real-time inference, and evaluation under far-field and channel distortion conditions to broaden the applicability of the frontend across communication and recognition systems.

References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 12449–12460, 2020.

[2] W.-N. Hsu, B. Bolte, Y.-H. 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, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

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

[4] S.-W. Fu et al., "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, pp. 201–205, 2021.

[5] Y. Cao et al., "CMGAN: Conformer-based metric-GAN for monaural speech enhancement," in *Proc. Interspeech*, pp. 2790–2794, 2022.

[6] S. Li, Y. Xue, and H. Liu, "SEMamba: State-space model for speech enhancement," in *Proc. Interspeech*, 2024.

[7] K. Iwamoto et al., "How bad are artifacts? Analyzing the impact of speech enhancement errors on ASR," in *Proc. Interspeech*, pp. 4526–4530, 2022.

[8] Y. Dissen, S. Yonash, I. Cohen, and J. Keshet, "A front-end adaptation network for improving speech recognition performance in packet loss and noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 33, 2025.

[9] A. S. Subramanian et al., "Speech enhancement using end-to-end speech recognition objectives," in *Proc. IEEE WASPAA*, pp. 234–238, 2019.

[10] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[11] T. Ochiai, S. Watanabe, T. Hori, and J. R. Hershey, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1274–1288, 2017.

[12] Z. Zhang et al., "Closing the gap between time-domain multi-channel speech enhancement on high and low resource datasets," in *Proc. Interspeech*, 2021.

[13] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, pp. 2613–2617, 2019.

[14] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proc. IEEE ICASSP*, pp. 5220–5224, 2017.

[15] C. Valentini-Botinhao et al., "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. SSW*, pp. 146–152, 2016.

[16] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)," in *Proc. IEEE ICASSP*, vol. 2, pp. 749–752, 2001.

[17] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[18] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?" in *Proc. IEEE ICASSP*, pp. 626–630, 2019.

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

[20] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE ICASSP*, pp. 6199–6203, 2020.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, pp. 234–241, 2015.

[22] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv preprint arXiv:1606.08415*, 2016.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.

[24] H. Chen et al., "Reducing the gap between pretrained speech enhancement and recognition models using a real speech-trained bridging module," in *Proc. IEEE ICASSP*, 2025.

[25] S. Li et al., "Speech enhancement with MAP-based training for robust ASR," in *Proc. IEEE ICASSP*, 2025.

[26] A. Radford et al., "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, pp. 28492–28518, 2023.

[27] 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*, 2023.

[28] 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, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.

[29] S. Kang, H. Yun, and J. Nam, "FlowMSE: Speech enhancement using flow matching," in *Proc. IEEE ICASSP*, 2025.