Model as Loss: A Self-Consistent Training Paradigm

Saisamarth Rajesh Phaye[1], *Milos Cernak*[1], *Andrew Harper*[1]

[1]Audio Machine Learning, Logitech

(sphaye, mcernak, aharper)@logitech.com

Abstract

Conventional methods for speech enhancement rely on handcrafted loss functions (e.g., time or frequency domain losses) or deep feature losses (e.g., using WavLM or wav2vec), which often fail to capture subtle signal properties essential for optimal performance. To address this, we propose Model as Loss, a novel training paradigm that utilizes the encoder from the same model as a loss function to guide the training.

The Model as Loss paradigm leverages the encoder's task-specific feature space, optimizing the decoder to produce output consistent with perceptual and task-relevant characteristics of the clean signal. By using the encoder's learned features as a loss function, this framework enforces self-consistency between the clean reference speech and the enhanced model output. Our approach outperforms pre-trained deep feature losses on standard speech enhancement benchmarks, offering better perceptual quality and robust generalization to both in-domain and out-of-domain datasets.

Index Terms: speech enhancement, noise reduction, deep feature loss, loss functions

1. Introduction

Speech enhancement has long been a challenging problem, with applications in telecommunication, hearing aids, and robust automatic speech recognition (ASR) [1, 2, 3, 4, 5]. A critical component in training enhancement models is the choice of loss function, which directly influences the quality and generalization of enhanced output [6, 7]. Conventional loss functions, such as time or spectrogram-domain losses [8], often do not fully capture the complex relationships between noisy and clean speech. For instance, spectrogram loss treats all frequency bins equally, which can result in overemphasis of less perceptually relevant regions while underweighting crucial frequencies important for speech intelligibility [9]. Although there are perceptually sensitive losses such as mel-spectral losses [8] or PMSQE [9], they overly compress the signal, limiting the preservation of fine-grained details crucial for maintaining speech intelligibility, achieving suboptimal performance [10].

Pre-trained deep feature losses [6], such as those derived from WavLM [11] or Wav2Vec [12], have gained popularity due to their ability to incorporate perceptual and contextual information into training objectives. Recent work by Babaev *et al.* [3] shows that WavLM's intermediate convolutional features have a high correlation to speech enhancement, as compared to its transformer layers. WavLM has also been shown to be superior to Wav2Vec 2.0 [13] when used as a loss function. However, these methods are often optimized for tasks such as ASR or phoneme recognition, which may not align with speech enhancement objective. Such losses might prioritize linguistic content while ignoring residual noise components critical to the enhancement task. Moreover, these pre-trained neural networks, when used as loss functions, can suffer from limited sensitivity to noise, as the extracted features may focus on abstract representations rather than task-specific properties [14].

To address these limitations, we propose a novel training paradigm, *Model as Loss (MAL)*, which uses the encoder of the same model as a loss function to guide the training of the decoder. It involves training a model with conventional loss functions and then using the trained encoder's embeddings as a loss function for the next stage. This approach aligns the loss function with the downstream task, leveraging the encoder's ability to extract task-specific features while ensuring contextual and hierarchical understanding of the signal. Unlike traditional methods that rely on external pre-trained models or handcrafted losses, this paradigm exploits the encoder's specialization in processing noisy speech and its inherent ability to prioritize perceptually relevant signal components.

Our proposed method has several advantages. First, the encoder's feature space is inherently tailored to the enhancement task, capturing both global and local signal features crucial for noise suppression. Second, using the encoder as a loss function enforces a feedback loop that aligns training and inference dynamics, improving generalization to unseen noise types. Third, this approach ensures relevance across all frequency bins by leveraging the encoder's weighting of spectral components based on their contribution to the task, avoiding the pitfalls of uniformly weighted losses. Finally, the self-consistency of the model provides a robust foundation for optimizing the decoder, leading to superior performance and perceptual quality.

In this paper, we present a detailed analysis of the *MAL* paradigm, including its theoretical foundations and practical implementation. Through extensive experiments, we demonstrate that our approach outperforms conventional pre-trained deep feature losses and hand-crafted loss functions on standard speech enhancement benchmarks. The results highlight the potential of using the model itself as a loss function, offering a new perspective on loss design in machine learning.

2. Methodology

In the realm of speech enhancement, models typically comprise an encoder and one or more decoders [5]. For simplicity, we will refer to this setup as a single encoder-decoder system moving forward. The encoder's job is to extract relevant features from the noisy signal, which are then used by the decoder to synthesize the enhanced signal.

The usual approach for training encoder-decoder models involves minimizing a loss function $\mathcal{L}$ between the clean reference speech and the model output. For example, an L1 loss

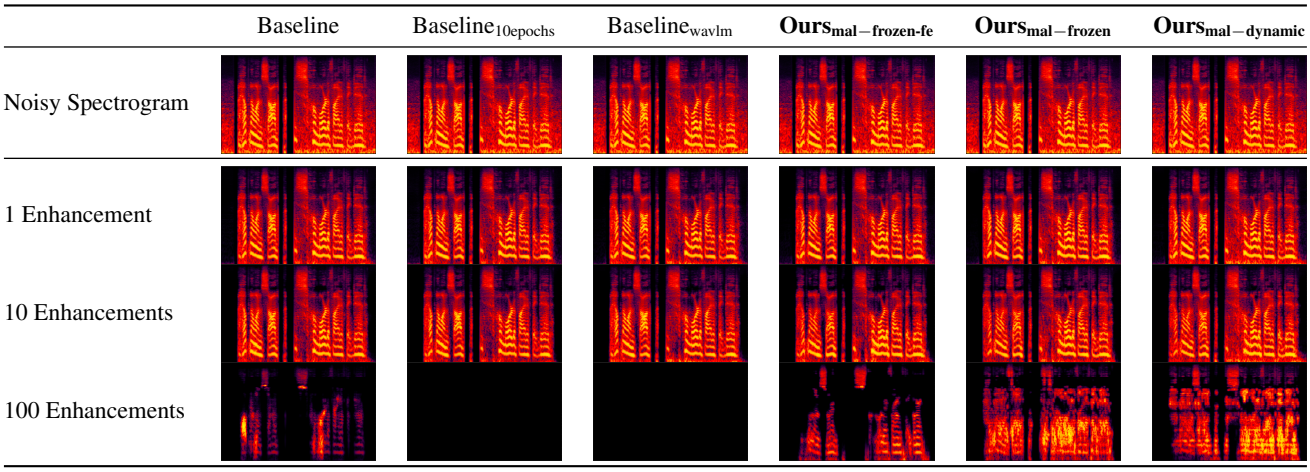

Figure 1: *Comparison of models using iterative enhancement. Each row represents the number of iterative enhancements applied, where the output of the previous enhancement step is used as input for the next step. The columns show different models being compared.*

between the clean and enhanced signal in spectral domain using Short-time Fourier Transform (STFT) can be represented as:

$$\mathcal{L}_{\text{spectral}} = \|\text{STFT}(\mathbf{y}_{\text{clean}}) - \text{STFT}(\mathbf{y}_{\text{enhanced}})\|_1 \quad (1)$$

where y_{clean} is the clean reference speech and $y_{enhanced}$ is the model output when the noisy speech is used as input to the model. The objective is to minimize the difference between the model's output and the clean speech, typically measured on a spectrogram or multiple spectrograms of different resolutions.

Once we train a model with some loss $\mathcal{L}$ to convergence, we know that for any given pair ($y_{enhanced}$, y_{clean}), $\mathcal{L}$ will be minimal. However, this does not imply that any other loss $\mathcal{L}_{new}$ will also be minimal. Our search is to find the ideal loss function $\mathcal{L}_{ideal}$, such that, once trained to convergence, for any given mathematical function $\mathcal{F}$, we get the minimum loss:

$$\mathcal{L}_{\text{F}} = \|\mathcal{F}(\mathbf{y}_{\text{clean}}) - \mathcal{F}(\mathbf{y}_{\text{enhanced}})\|_1 \quad (2)$$

Babaev *et al.* [3] propose a *Signal-to-Noise (SNR) Rule*, which suggests that as more noise is added to speech (lowering the SNR), feature representations should move farther apart in the embedding space. Extending this idea, intuitively, once the model is trained to satisfy equation (2), $y_{enhanced}$ and y_{clean} should be identical points in the embedding space. Assuming equation (2) holds, if we input y_{clean} to the trained model, it should return y_{clean}. Moreover, $y_{enhanced}$ when used as input should again return $y_{enhanced}$, which is y_{clean}. This would ensure the stability of the model in its embedding space. However, this is not the case for most models. Figure 1 illustrates the impact of iteratively enhancing the noisy input multiple times with various trained models, which are described later in Section 3. All models show a clear degradation in speech quality with each iteration, suggesting a loss of fine-grained features.

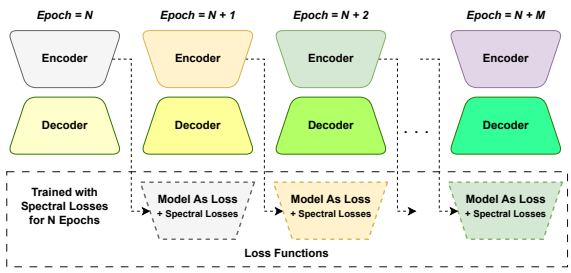

Figure 2: *An illustration of the Model as Loss paradigm, showcasing the $\mathcal{L}_{mal-dynamic}$ variation.*

Since the model being trained is itself a mathematical function, we propose using its encoder as a loss function. After training a model with traditional loss functions, the encoder inherently learns to represent the input data's structure in its feature space. This feature space encodes rich information beyond what traditional loss functions capture. Building on this, we introduce a novel training strategy: first, train the model with conventional losses, and then, using the trained encoder, add a loss term based on the encoder's latent feature space. This loss function compares the encoder's bottleneck embeddings of the enhanced output and the clean signal and can be expressed as:

$$\mathcal{L}_{\text{mal}} = \|Encoder(\mathbf{y}_{\text{clean}}) - Encoder(\mathbf{y}_{\text{enhanced}})\|_1 \quad (3)$$

where *Encoder* represents the function that returns the bottleneck features, denoted as the *MAL-encoder*. This ensures that the decoder's output aligns with the clean signal not only at the spectral level but also in the encoder's learned feature space. With this formulation, the decoder is guided to preserve the rich feature representation learned by the encoder. As a result, the decoder learns to produce $y_{enhanced}$, which, when fed back into the model, reproduce close to $y_{enhanced}$. This is evident in Figure 1, where Ours_{mal} models, trained with the *MAL* paradigm, preserve more speech harmonics after 100 iterations.

As shown in Figure 2, we train a model with conventional losses for N epochs. Then, for the next M epochs, we add the $\mathcal{L}_{mal}$ loss to further refine the model. Depending on how we use MAL-encoder, there are three possible $\mathcal{L}_{mal}$ variations:

1. $\mathcal{L}_{mal-frozen-fe}$: Freeze the trained Encoder (FE) of N^{th} epoch and use it as the *MAL-encoder* to train only the decoder for subsequent epochs.
2. $\mathcal{L}_{mal-frozen}$: Use the trained encoder of the N^{th} epoch as *MAL-encoder* for all subsequent epochs and train the full encoder-decoder model.
3. $\mathcal{L}_{mal-dynamic}$: Use the trained encoder of the n^{th} epoch as *MAL-encoder* for $(n+1)^{th}$ epoch for $n \geq N$ and train the full model. Hence, *MAL-encoder* is updated with every epoch.

Table 1: *NISQA and ScoreQ metrics are presented for both in-domain (In) and out-of-domain (Out) datasets. All proposed MAL-based models demonstrate superior performance, with* $\text{Ours}_{mal-dynamic}$ *achieving the highest results across all NISQA metrics.*

Model / loss variants	NISQA (↑)										ScoreQ MOS (↑)			
	Overall		Noisiness		Discontinuity		Coloration		Loudness		Natural		Synthetic	
	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out	In	Out
Baseline	3.50	2.93	4.06	3.82	3.87	3.45	3.45	2.97	3.90	3.50	2.91	2.50	2.10	2.01
$\text{Baseline}_{10epochs}$	3.54	2.99	4.10	3.85	3.92	3.53	3.45	2.99	3.89	3.53	2.93	2.51	2.09	2.00
Baseline_{wavlm}	3.56	3.02	4.06	3.85	3.94	3.55	3.47	3.03	3.89	3.57	2.93	2.52	2.07	1.99
$\text{Baseline}_{wavlm-fe}$	3.52	2.97	4.07	3.86	3.88	3.49	3.45	2.97	3.91	3.54	2.94	2.52	2.11	2.02
$\text{Ours}_{mal-frozen-fe}$	3.65	3.12	**4.14**	3.93	4.01	3.60	3.57	3.10	3.98	3.62	**3.00**	**2.60**	2.11	2.02
$\text{Ours}_{mal-frozen}$	3.66	3.13	4.09	3.90	4.06	3.64	3.58	3.15	3.99	**3.65**	2.97	2.58	2.08	1.99
$\text{Ours}_{mal-dynamic}$	**3.72**	**3.17**	**4.14**	**3.96**	**4.10**	**3.67**	**3.62**	**3.16**	**4.01**	3.62	2.96	2.58	2.10	**2.03**
$\text{Ours}_{wavlm-mal}$	3.65	3.13	4.13	3.95	4.02	3.61	3.57	3.11	3.97	3.61	3.00	**2.60**	**2.12**	**2.03**
DeepFilterNet2 [5]	3.46	2.82	3.97	3.66	3.87	3.47	3.43	2.88	3.86	3.48	2.90	2.49	2.08	1.98
$\text{Ours}^{ablation}_{mal-frozen-fe}$	3.62	3.01	4.12	3.87	3.97	3.51	3.55	3.04	3.97	3.59	2.95	2.56	2.10	2.00

Table 2: *SIGMOS Metrics are shown across in-domain (In) and out-of-domain (Out) datasets, while Intrusive Metrics are evaluated on the 2024 Urgent Challenge non-blind test set. All proposed MAL-based models outperform the other models.*

Model / loss variants	SIGMOS (↑)										Intrusive Metrics			
	Signal		Overall		Noise		Discontinuity		Coloration		PESQ (↑)	ESTOI (↑)	LSD (↓)	MCD (↓)
	In	Out	In	Out	In	Out	In	Out	In	Out	Out	Out	Out	Out
Baseline	3.48	3.15	3.05	2.74	4.07	4.08	3.83	3.73	3.56	3.22	1.96	0.75	5.12	5.06
$\text{Baseline}_{10epochs}$	3.52	3.19	3.09	2.77	**4.12**	4.10	3.89	3.82	3.56	3.20	2.01	0.76	5.14	4.99
Baseline_{wavlm}	3.48	3.22	3.05	2.80	4.03	4.08	3.87	3.87	3.53	3.22	**2.03**	0.76	4.99	4.91
$\text{Baseline}_{wavlm-fe}$	3.49	3.18	3.06	2.77	4.04	4.10	3.86	3.78	3.55	3.21	1.99	0.76	4.99	5.02
$\text{Ours}_{mal-frozen-fe}$	**3.57**	**3.31**	**3.14**	2.88	4.10	4.12	3.92	3.93	**3.60**	3.30	**2.03**	**0.77**	5.02	4.88
$\text{Ours}_{mal-frozen}$	3.53	**3.31**	3.10	**2.89**	4.04	4.11	3.91	**3.95**	3.58	**3.32**	**2.03**	**0.77**	**4.81**	**4.87**
$\text{Ours}_{mal-dynamic}$	3.56	3.30	3.13	2.88	4.08	**4.13**	3.91	3.91	3.59	**3.32**	2.00	0.76	4.88	**4.87**
$\text{Ours}_{wavlm-mal}$	3.55	3.30	3.11	2.87	4.08	4.11	**3.93**	3.92	3.58	3.29	**2.03**	**0.77**	4.98	4.92

Within this training paradigm, the encoder (applied to noisy input) functions solely as a feature extractor, while the decoder becomes the primary model for synthesizing the enhanced output. The *MAL-encoder*, acting as a loss function, ensures that the decoder produces an output closely aligned with the features of clean audio. The combination of supervised learning (matching the clean signal) and self-supervised learning (consistency within the encoder's feature space) ensures that the decoder's outputs are both accurate and perceptually meaningful.

The proposed method introduces a self-consistency feedback loop, where the decoder's output is evaluated not only against the clean signal but also through the encoder's feature space. Such a dual-objective structure reinforces meaningful learning at multiple levels of abstraction, leading to superior performance. This integration of self-supervised and supervised learning balances explicit and implicit learning objectives, making the model more robust in real-world scenarios.

3. Experimental setup

We base all our experiments on DeepFilterNet2 proposed by Schröter *et al.* [5]. It has an encoder that extracts relevant features and passes them into a first-stage decoder. The output of this decoder is passed into the deep filtering decoder, which predicts the deep filtering coefficients for each time frame. We train DeepFilterNet2 as the base model from the official GitHub repository [5], using a 960-point FFT, 480 hop size, Vorbis window and 2-frame lookahead. Training is done on English subset of the 48KHz DNS4 dataset [15], using the same loss function $\mathcal{L}_{\text{df}}$ as in the original article, combining multiple spectral losses. This model serves as the **Baseline** for our experiments.

The experimental setup consists of five evaluation systems. The first three configurations include finetuning the baseline model with one of the three $\mathcal{L}_{\text{mal}}$ loss variations, mentioned in Section 2. In each model, the $\mathcal{L}_{\text{mal}}$ loss is added in equal proportion to the original DeepFilterNet2 loss function:

i) **$\text{Ours}_{mal-frozen-fe}$** trained with $\mathcal{L}_{\text{df}} + \mathcal{L}_{\text{mal-frozen-fe}}$
ii) **$\text{Ours}_{mal-frozen}$** trained with $\mathcal{L}_{\text{df}} + \mathcal{L}_{\text{mal-frozen}}$
iii) **$\text{Ours}_{mal-dynamic}$** trained with $\mathcal{L}_{\text{df}} + \mathcal{L}_{\text{mal-dynamic}}$

In the next two setups, we introduce a loss function, $\mathcal{L}_{\text{wavlm}}$, which uses the final Conv-layer output from the pre-trained WavLM-Base-Plus (WavLM) [11], resembling equation (3). We again add this loss function in equal proportion to the original loss function, keeping the encoder either frozen or trainable:

iv) **Baseline_{wavlm}** trained with $\mathcal{L}_{\text{df}} + \mathcal{L}_{\text{wavlm}}$
v) **$\text{Baseline}_{wavlm-fe}$** trained with $\mathcal{L}_{\text{df}} + \mathcal{L}_{\text{wavlm-fe}}$

As an ablation experiment to evaluate the effect of $\mathcal{L}_{\text{wavlm-fe}}$ helps, we combine it with $\mathcal{L}_{\text{mal-frozen-fe}}$ and $\mathcal{L}_{\text{df}}$, adding all three in equal proportions, and denote this model as **$\text{Ours}_{wavlm-mal}$**.

In all experiments, the **Baseline** model is finetuned for ten epochs and the best epoch is chosen. For consistency, we also finetune the baseline model for ten epochs without introducing any new losses. This is **$\text{Baseline}_{10epochs}$**.

3.1. Evaluation data

The evaluation was carried out on a total of 7404 samples drawn from multiple test sets. We divide the test sets into two domains:

- **In-domain test set (3504 samples)**: Since the models are trained on DNS V4 training data, we aggregate the test samples from DNS Challenge V2 [2], V3 [15], and V5 [4] covering diverse acoustic scenarios such as mouse clicks, headset noise, speakerphone noise, and emotional speech.
- **Out-of-domain test set (3900 samples)**: We aggregated fully unseen test sets from 2024 and 2025 Urgent Challenges, combining the two nonblind and two blind sets [16].

This test setup enabled robust performance comparisons across diverse acoustic conditions.

3.2. Evaluation metrics

Models are evaluated using SIGMOS, NISQA v2.0, and ScoreQ (no-reference natural and synthetic MOS) metrics [17, 18, 19]. SIGMOS scores were computed with all enhanced samples normalized to a peak level of -10 dBFS to account for level dependency. For NISQA, enhanced samples were normalized to have an active speech level of -26 dBFS. Intrusive metrics such as PESQ, ESTOI, log-spectral distance (LSD), and Mel-cepstral distance (MCD) were also calculated. We perform ANOVA analysis and report only the metrics with statistical significance. For intrusive metrics, we use only the 2024 Urgent Challenge nonblind test set with its open-source evaluation pipeline [16].

4. Results

Tables 1 and 2 present all metrics, clearly demonstrating that the proposed models with $\mathcal{L}_{\text{mal}}$ losses outperform the others. $\text{Ours}_{\text{mal}-\text{dynamic}}$ achieves the best performance across all NISQA metrics, while $\text{Ours}_{\text{mal}-\text{frozen}}$ leads in all intrusive metrics. The $\text{Ours}_{\text{mal}-\text{frozen-fe}}$ model outperforms others in SIGMOS Signal and Overall metrics, while performing comparably on the remaining SIGMOS metrics.

Given that $\text{Ours}_{\text{mal}-\text{frozen-fe}}$ is at par with $\text{Ours}_{\text{wavlm-mal}}$, using $\mathcal{L}_{\text{wavlm}-\text{fe}}$ with $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$ offers no advantage over $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$ alone. Notably, WavLM, with 95.1M parameters, is trained on 94,000 hours of speech data [11], while DeepFilterNet2, with only 2.31M parameters, is trained on 1,100 hours of speech data [5]. However, all models trained with $\mathcal{L}_{\text{mal}}$ outperform both models trained with $\mathcal{L}_{\text{wavlm}}$ loss variants.

Typically, finetuning leads to overfitting on in-domain data, resulting in degraded performance on out-of-domain samples [20]. This is particularly a risk when the *MAL-encoder* is optimized specifically for in-domain data. However, the results demonstrate that Ours_{mal} models not only avoid overfitting but also significantly outperform all other models.

4.1. Ablation Experiments

A key factor is the quality of the *MAL-encoder* for $\mathcal{L}_{\text{mal}-\text{frozen}}$ or $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$, as the effectiveness of the loss function depends on how well the encoder extracts features for enhancement. To answer this, we used the publicly available pre-trained DeepFilterNet2 [5], which performs slightly worse than our trained Baseline. We finetuned the pre-trained model with $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$ as previously described, resulting in $\text{Ours}^{\text{ablation}}_{\text{mal}-\text{frozen-fe}}$. Although $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$ significantly improved performance, it still performed worse than when applied to the superior Baseline model (see Table 1). The better the encoder, the more effective $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$ becomes.

We also trained with $\mathcal{L}_{\text{mal}-\text{dynamic}}$ per batch instead of per epoch, resulting in slightly worse metrics (e.g., LSD 5.03 vs. 4.88), likely due to loss instability from frequent updates.

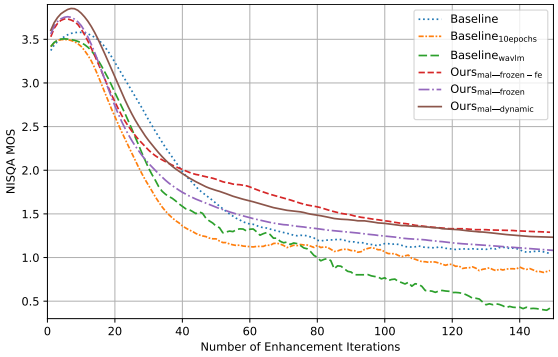

Figure 3: *NISQA MOS vs number of enhancement iterations*

Self-consistency Experiment: We take the first 200 samples of the 2025 Urgent Challenge nonblind test set and iteratively enhance them 150 times with every model. Figure 3 shows how the average NISQA MOS initially improves as noise is removed but later declines as speech quality degrades. Models trained with $\mathcal{L}_{\text{mal}-\text{frozen-fe}}$ or $\mathcal{L}_{\text{mal}-\text{dynamic}}$ better preserve speech, aligning with the self-consistency criterion, which $\mathcal{L}_{\text{mal}-\text{frozen}}$ lacks. In the 1st iteration, all *MAL* models achieve a high MOS, then reach a higher peak before converging to a higher MOS.

This aligns with Figure 1, where *MAL* models preserve more speech harmonics than non-*MAL* models. Notably, the Baseline model preserves slightly more speech than $\text{Baseline}_{\text{10epoch}}$ or $\text{Baseline}_{\text{wavlm}}$, despite having lower metrics (see Tables 1 and 2). Although the exact reason is unclear, we suspect that the latter two models are more aggressive than the Baseline in removing noise.

5. Conclusion

In this paper, we propose *Model as Loss (MAL)*, a novel training paradigm that leverages the encoder of an encoder-decoder model as a loss function to guide optimization. By aligning the loss with the model's task-specific feature space, *MAL* overcomes the limitations of traditional handcrafted and pre-trained deep feature losses. This approach offers key advantages such as task-specific feature extraction, self-consistency, and enhanced contextual understanding of input signals. Additionally, *MAL* reduces dependency on deep-feature losses derived from pre-trained models while achieving comparable or superior performance. This is particularly essential in domains such as medical imaging [21] or specialized signal analysis [22], where pre-trained models are scarce. Experiments demonstrate that *MAL* improves both perceptual quality and task-specific performance in speech enhancement.

Although our evaluation focused on speech enhancement, *MAL*'s domain-agnostic design makes it applicable to tasks such as acoustic echo cancellation [23, 24], image denoising or super-resolution [25], and medical image analysis [26]. Ongoing research explores its broader potential. We hope that this work inspires further innovation in loss functions and training methodologies, advancing machine learning applications.

6. Acknowledgements

The authors thank Paul Kendrik, Tijana Stojkovic, and Andy Pearce for their valuable feedback and insights. We also thank Sai Dhawal Phaye for discussions during the early stages of *MAL*, and Kanav Sabharwal for his feedback on the writing.

7. References

[1] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[2] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun *et al.*, "The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," *arXiv preprint arXiv:2005.13981*, 2020.

[3] N. Babaev, K. Tamogashev, A. Saginbaev, I. Shchekotov, H. Bae, H. Sung, W. Lee, H.-Y. Cho, and P. Andreev, "FINALLY: fast and universal speech enhancement with studio-like quality," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=18RdkSv9h9

[4] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, H. Gamper, M. Golestaneh, and R. Aichner, "Icassp 2023 deep noise suppression challenge," 2023. [Online]. Available: https://arxiv.org/abs/2303.11510

[5] H. Schröter, A. N. Escalante-B., T. Rosenkranz, and A. Maier, "DeepFilterNet2: Towards real-time speech enhancement on embedded devices for full-band audio," in *17th International Workshop on Acoustic Signal Enhancement (IWAENC 2022)*, 2022. [Online]. Available: https://github.com/Rikorose/DeepFilterNet

[6] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *Proc. Interspeech 2019, 2723-2727*, 2018.

[7] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," in *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2021, pp. 72–76.

[8] C. J. Steinmetz and J. D. Reiss, "auraloss: Audio focused loss functions in PyTorch," in *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020.

[9] J. Martín-Doñas, A. Gomez, J. Gonzalez Lopez, and A. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. PP, pp. 1–1, 09 2018.

[10] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 825–838, 2020.

[11] 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, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *CoRR*, vol. abs/2110.13900, 2021. [Online]. Available: https://arxiv.org/abs/2110.13900

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[13] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[14] S. Maiti, Y. Peng, T. Saeki, and S. Watanabe, "Speechlmscore: Evaluating speech generation using speech language model," 2022. [Online]. Available: https://arxiv.org/abs/2212.04559

[15] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matusevych, S. Braun, S. E. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," 2022. [Online]. Available: https://arxiv.org/abs/2202.13288

[16] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, A. Kumar, J. Pirklbauer, M. Sach, S. Watanabe *et al.*, "Urgent challenge: Universality, robustness, and generalizability for speech enhancement," *arXiv preprint arXiv:2406.04660*, 2024.

[17] N. C. Ristea, A. Saabas, R. Cutler, B. Naderi, S. Braun, and S. Branets, "Icassp 2024 speech signal improvement challenge," 2024. [Online]. Available: https://arxiv.org/abs/2401.14444

[18] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," Aug. 2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2021-299

[19] A. Ragano, J. Skoglund, and A. Hines, "Scoreq: Speech quality assessment with contrastive regression," *arXiv preprint arXiv:2410.06675*, 2024.

[20] Y. Li, Y. Sun, K. Horoshenkov, and S. M. Naqvi, "Domain adaptation and autoencoder-based unsupervised speech enhancement," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 1, p. 43–52, Feb. 2022. [Online]. Available: http://dx.doi.org/10.1109/TAI.2021.3119927

[21] L. Gondara, "Medical image denoising using convolutional denoising autoencoders," in *2016 IEEE 16th international conference on data mining workshops (ICDMW)*. IEEE, 2016, pp. 241–246.

[22] K. Sabharwal, S. Ramesh, J. Wang, D. M. Divakaran, and M. C. Chan, "Enhancing lora reception with generative models: Channel-aware denoising of loraphy signals," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 2024, pp. 507–520.

[23] H. Zhang and D. Wang, "Neural cascade architecture for multi-channel acoustic echo suppression," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2326–2336, 2022.

[24] S. Braun and M. L. Valero, "Task splitting for dnn-based acoustic echo and noise removal," in *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE, 2022, pp. 1–5.

[25] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, vol. 131, pp. 251–275, 2020.

[26] Y. Li, B. Sixou, and F. Peyrin, "A review of the deep learning methods for medical images super resolution problems," *Irbm*, vol. 42, no. 2, pp. 120–133, 2021.