Noise Robustness Project Report, Spring 2025

Mike Qu

May 22, 2025

1 Introduction

While speech enhancement models have seen rapid improvements, they can still struggle in real-world environments where noise is unpredictable or overlaps heavily with the speech signal. These failures often manifest as incomplete noise suppression, destruction of clean speech, or both, leading to inconsistent performance on downstream tasks like sentiment or emotion detection. This project focuses on studying when and why these models break down. We evaluate their robustness across different types of noise and speaker conditions, using both perceptual metrics and task-specific performance to pinpoint failure cases. Once identified, we explore how these weaknesses can be addressed—whether through changes in training objectives, model inputs, or architectural tweaks. The goal is to better understand how noise enhancement models behave under stress and how to make them more reliable in practice.

2 Related Works

2.1 Speech Enhancement Models

The objective of Speech Enhancement (SE) Models is to remove "unwanted" background noise from noisy speech signals. Below, we briefly describe a few commonly-used state-of-the-art SE models developed over the past several years. An up-to-date benchmark of Speech Enhancement models on the VoiceBank + DEMAND dataset (a dataset commonly used in SE model training) can be found in `https://paperswithcode.com/sota/speech-enhancement-on-demand`.

CMGAN [Cao et al., 2022] uses a conformer-based metric generative adversarial network for speech enhancement in the time-frequency domain. CMGAN's architecture features a two-stage conformer used in the generator to capture spatial and temporal dependencies of the magnitude and phase information of the audio spectrogram. A metric discriminator is used to improve the quality of the generator's output with respect to a corresponding evaluation score like PESQ.

SEMamba [Chao et al., 2024] uses a Mamba-based regression model for Speech Enhancement. The model architecture consists of a convolutional layer to learn adjacent relationships, bidirectional Mamba blocks to capture temporal dependencies, and a fully connected decoder layer in the STFT domain. Another interesting result presented by the authors is that using Transformer layers over Mamba modules offers no significant performance gains in terms of the quality of the enhanced speech.

SGMSE [Richter et al., 2023] approaches SE using diffusion-based generative models. However, unlike standard conditional generation methods that typically begin the reverse process from pure Gaussian noise, SGMSE initializes the reverse diffusion from a mixture of clean speech and Gaussian noise. Meanwhile, its forward process gradually corrupts the original noisy speech over multiple steps. This allows for highly-efficient inference performance to generate high-quality clean speech from their noisy counterparts.

2.2 Speech Evaluation Metrics

To evaluate whether a particular SE model has successfully enhanced a noisy speech signal, a number speech evaluation metrics can be employed. Widely-used intrusive metrics like PESQ [Rix et al., 2001] and STOI [Taal et al., 2010] are effective, but require a clean audio waveform as reference. In contrast, non-intrusive metrics (often deep-learning-based) do not have such a requirement and are much more versatile.

NISQA [Mittag et al., 2021] first segments the speech using a mel spectrogram, then uses a CNN-based encoder to compute features suitable for speech quality prediction. A self-attention-based framework then models time-dependency before being aggregated over time in a pooling layer for metric prediction. NISQA provides five distinct submetrics: MOS (Mean Opinion Score), Noisiness, Coloration, Discontinuity, and Loudness.

Meta Audiobox Aesthetics [Tjandra et al., 2025] proposes finer-grained speech evaluation through more specific metrics such as PQ (Production Quality), PC (Production Complexity), CE (Content enjoyment), and CU (Content Usefulness). PQ and PC quantify the technical aspects of quality, focusing on fidelity and the number of audio components, respectively. CE and CU are both subjective axes, focusing on the perceived subjective experience and usefulness for content creation, respectively.

SCOREQ [Ragano et al., 2025] is an extension on top of NISQA, focusing solely on MOS (Mean Opinion Score). A triplet loss function for contrastive regression is used to improve generalization performance on unobserved data and datasets.

3 Experiments and Analysis (Noise Robustness)

3.1 Analysis of Enhancement Models

We evaluate the performance of three state-of-the-art enhancement models: CMGAN, SEMamba, and SGMSE using a variety of intrusive and non-intrusive metrics including PESQ, STOI, Audiobox CE (Content Enjoyment), and SCOREQ MOS. The evaluation is performed on a synthetic dataset, which adds noisy miscellaneous sounds derived from the Audioset dataset [Gemmeke et al., 2017], at 4 dB and 8 dB SNR, onto clean MSP-Podcast audio [Lotfian and Busso, 2019] samples.
The results of our experiments are shown below in fig. 2.

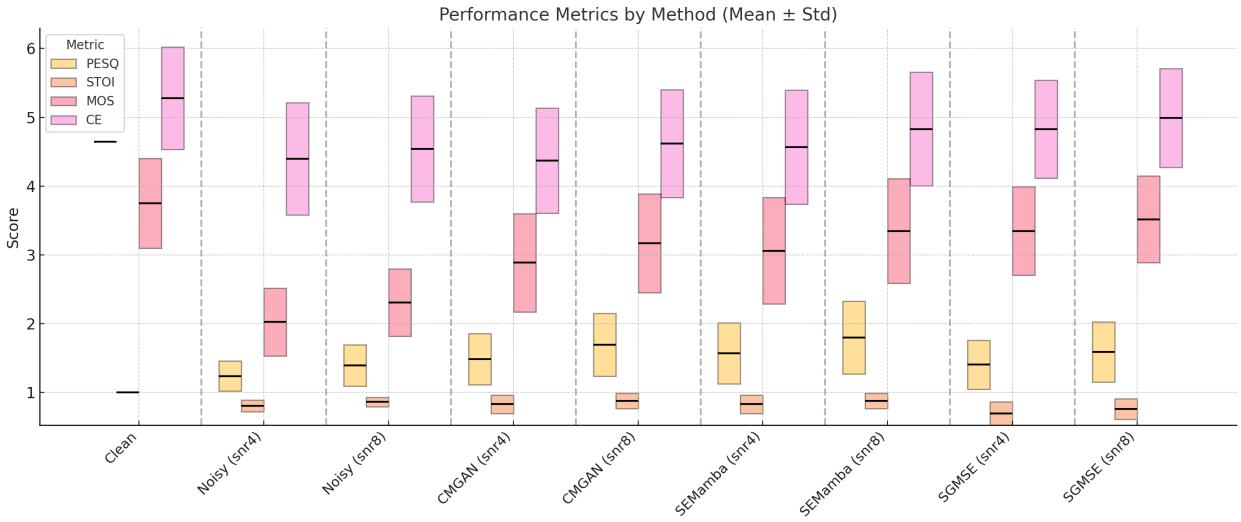

Figure 1: MSP PODCAST SE model and evaluation metric baselines

As expected, clean speech consistently scored highest across all metrics, while the noisy baselines, particularly at 4 dB, performed the worst. Increasing the SNR to 8 dB led to noticeable improvements across the board. Both CMGAN and SEMamba showed solid gains in PESQ and STOI compared to the noisy inputs, with SEMamba slightly ahead at the higher SNR. In contrast, SGMSE lagged behind on STOI, especially at 4 dB, suggesting that its diffusion-based approach may not preserve intelligibility as well in more challenging conditions.
That said, SGMSE did particularly well on subjective metrics like MOS and CE, outperforming the other models in both cases at 8 dB. This indicates that SGMSE produces outputs that appear to be more natural-sounding and consistent, even if intelligibility is slightly weaker. The variability in its performance was also lower, indicating a higher level of robustness. SEMamba also performed well across all metrics, delivering good performance for intrusive metrics and slightly worse in non-intrusive metrics. Overall, the results reflect a trade-off between intelligibility and subjective quality. The right choice of enhancement model depends on whether the downstream goal is audio clarity or subjective quality.

3.2 Analysis of Speech Evaluation Methods for Failure prediction

We also investigate which speech evaluation methods are most capable of identifying instances where speech enhancement models fail. For each evaluation metric, we look for outlier instances where that evaluation metric excels and other metrics underperform and vice versa. By listening to the actual enhanced audio, we can identify and reconcile which evaluation metric provides a better reflection for the quality of the audible enhanced speech signal.

SCOREQ MOS achieves consistently good performance throughout our investigation. We notice that average MOS on clean data is approximately 3.75, while average MOS on noisy synthetic data at SNR 8 is approximately 2.31. The enhancement models CMGAN, SEMamba, and SGMSE improve MOS scores to 3.16, 3.34, and 3.51 respectively.
Interestingly, we observe that a high MOS score almost always corresponds to a high STOI score but not necessarily a high PESQ score, whereas a low MOS score almost always correspond to a low PESQ score but not necessarily a low STOI score.
This phenomenon arises likely because a high MOS score typically indicates that the speech is both intelligible and perceptually pleasant, which tends to require high intelligibility (reflected in STOI), even if signal-level distortions exist (which PESQ would penalize). On the other hand, low MOS scores usually imply that the signal is severely degraded and distorted, strongly impacting PESQ. However, distorted speech audio can still be intelligible despite sounding unnatural and unpleasant, leading to a reasonably high STOI score.

Audiobox PC provides a useful measure of the structural complexity of an audio clip. In our dataset, clean speech obtained an average PC score of 1.81, while synthetic noisy speech at 4 dB SNR reaches a much higher average of 4.37. Speech enhancement models like CMGAN, SEMamba, and SGMSE successfully reduce this complexity, with average PC scores of 2.04, 1.82, and 1.70, respectively. This trend is intuitive: effective enhancement should remove irrelevant or noisy components, resulting in a simpler, cleaner signal. However, PC alone is not sufficient for evaluating model performance, as it does not distinguish between the removal of noise and the unintended loss of important speech content. It should therefore be interpreted alongside other metrics, particularly those that directly assess intelligibility or perceptual quality.

Audiobox CE , like SCOREQ MOS, provides a subjective opinion metric but on how much the audience would "enjoy" the content of a speech signal. In general, CE exhibits a strong, positive correlation with MOS. Clean speech data obtained an average CE of 5.28, while noisy speech data at SNR 8 obtained an average CE of 4.54. CMGAN, SEMamba, and SGMSE improve CE metric performance to 4.62, 4.83, 4.99 respectively. These results suggest that CE is a useful complementary metric for evaluating speech enhancement models, particularly when listener engagement or perceived appeal is relevant. Unlike purely intelligibility-based metrics, CE captures a broader sense of quality that is important in downstream applications such as content creation, emotion detection, and so on.

3.3 Robust Enhancement Models

To address speech enhancement failures, one versatile approach is to blend the enhanced output with the original noisy input. This approach leverages the fact that, in cases where the enhancement model fails to suppress noise but retains the speech signal, useful linguistic information may still be preserved for downstream tasks. In contrast, when the model over-suppresses and removes portions of the actual speech signal, the resulting degradation is more detrimental, as it leads to irreversible loss of semantic content. NRSER [Chen et al., 2023] proposes to include an NN-based SNR block to determine the optimal mixing between the original noisy audio and the enhanced audio. The SNR block outputs a mixing coefficient, which is used during reconstruction by taking the weighted sum of amplitudes in the time domain.
However, one major downside of using a global mixing coefficient is that, because over-enhancement may only happen over a certain small interval of the audio clip, noise will be inadvertently introduced to sections where the enhancement model actually performed well.
Therefore, instead of having a global mixing constant, we propose ConvMixer, which uses an encoder-decoder 1D Convolution model to output a **time-dependent** mixing coefficient. The loss function includes a reconstruction component: the L1 distance between the final mixed audio waveform to the original clean sample, a smoothness component: which prevents mixing coefficient from fluctuating too quickly. We also incorporate the weighted sum of negative SCOREQ MOS score, negative Audiobox CE score, and positive Audiobox PC score. In simpler words, we seek to penalize low content enjoyment, low opinion score, and high audio complexity. We also incorporate experimental results using a 50/50 mix of the enhanced output and the noisy input for reference.

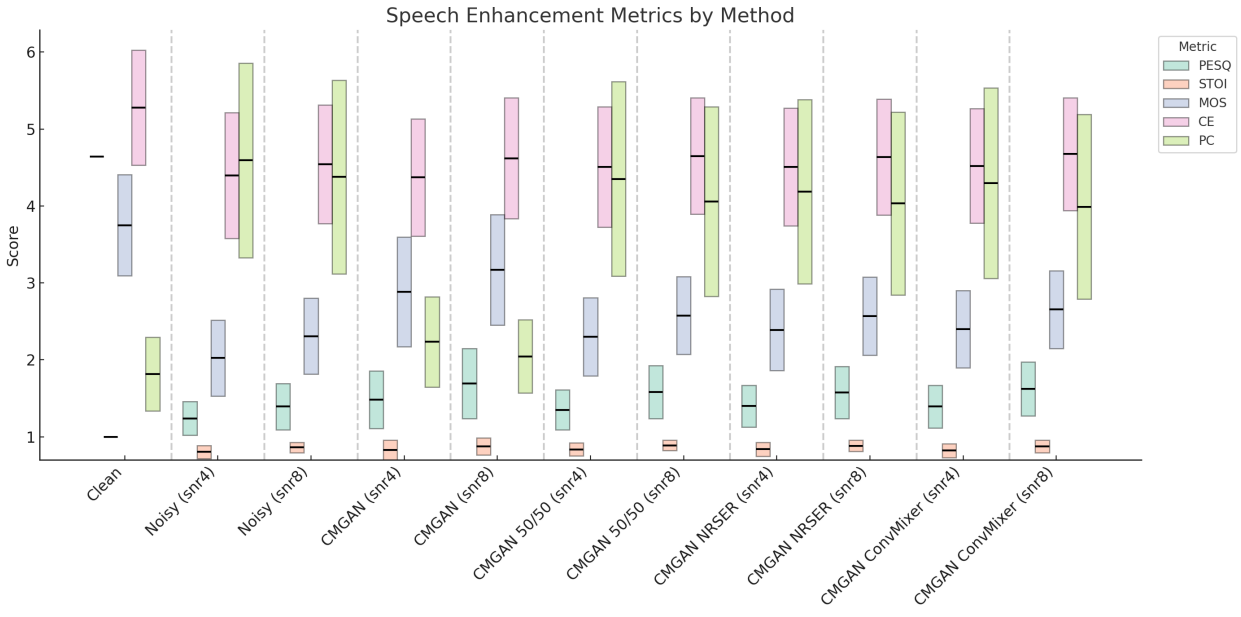

Figure 2: Speech Enhancement Robustness Benchmarks

Results show that ConvMixer obtains similar, if not better performance compared to our NRSER baseline for both SNR4 and SNR8 noisy data. However, the difference is not significant. One potential future improvement is to adopt an audio-specific encoder-decode architecture to capture more complex spatial and temporal dependencies across the two audio streams to mix them more effectively.

4 Experiments and Analysis (Downstream Emotion Detection)

An important dimension of evaluating speech enhancement models is their impact on downstream tasks, such as emotion detection. We evaluate performance across four setups: (1) training on clean speech and testing on noisy inputs; (2) training on noisy and enhanced speech and testing on enhanced inputs; (3) using

NRSER to adaptively mix clean and noisy signals; and (4) using a conditional version of NRSER, where the mixing decision is guided by the transcription of the enhanced speech signal.

Metric	**S_{clean}**		**S_{en}'**		**NRSER**		**Conditional**	
	snr8	snr4	snr8	snr4	snr8	snr4	snr8	snr4
ACC	0.46055	0.45164	0.46088	0.45413	0.46446	0.46331	0.47978	0.47054
F1	0.13635	0.13180	0.15420	0.14615	0.15623	0.15566	0.13863	0.13439
A (CCC)	0.35488	0.32024	0.47509	0.46393	0.48028	0.47334	0.39219	0.36111
V (CCC)	0.31390	0.28381	0.32646	0.31089	0.34170	0.32958	0.40878	0.38883
D (CCC)	0.31722	0.29688	0.37666	0.36944	0.38956	0.38677	0.32665	0.30963

Table 1: Emotion detection performance across different enhancement strategies and SNR conditions.

These results reveal a persistent challenge in applying speech enhancement to downstream tasks: models trained to optimize perceptual quality often unintentionally remove signal components that are critical for semantic understanding. This is particularly problematic for tasks like emotion detection, where subtle linguistic cues may be lost during aggressive denoising. Mixing-based strategies such as NRSER provide a practical compromise by recovering emotional cues that are otherwise discarded by enhancement models. Building on this idea, the conditional variant of NRSER further refines the mixing decision by conditioning on a representation of the recognized speech. By incorporating information from the transcribed content, the model can better infer whether the enhanced or noisy segment is more likely to retain meaningful content. These results point to a promising future direction of research: models should consider not just acoustic fidelity but also downstream utility, and a good model should selectively preserve the parts of the signal most relevant to the task.

5 Conclusion and Future Work

This project investigated the limitations of current speech enhancement models in noisy conditions, focusing on both perceptual quality and performance on downstream tasks such as emotion detection. While models like CMGAN and SEMamba improve intelligibility and perceptual metrics, they can inadvertently remove signal components that are important for higher-level understanding. Our experiments show that mixing-based approaches, such as NRSER, provide a more balanced solution by preserving emotionally salient cues present in the noisy input. Extending this, we proposed a conditional mixing model that incorporates recognized speech content to guide mixing decisions. This allows the system to better assess when enhanced or noisy segments are more semantically informative. Although the gains over NRSER are modest, the conditional model demonstrates the promise of task-aware enhancement strategies that adaptively preserve information relevant to the end goal. Future work will explore more expressive encoders for capturing semantic and acoustic dependencies, and develop training objectives that more directly align enhancement with downstream task performance.

References

R. Cao, S. Abdulatif, and B. Yang. Cmgan: Conformer-based metric gan for speech enhancement. In *Proceedings of Interspeech 2022*. ISCA, 2022. URL http://dx.doi.org/10.21437/Interspeech.2022-517. DOI: 10.21437/Interspeech.2022-517.

R. Chao, W.-H. Cheng, M. L. Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao. An investigation of incorporating mamba for speech enhancement, 2024. URL https://arxiv.org/abs/2405.06573.

Y.-W. Chen, J. Hirschberg, and Y. Tsao. Noise robust speech emotion recognition with signal-to-noise ratio adapting speech enhancement, 2023. URL https://arxiv.org/abs/2309.01164.

J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

R. Lotfian and C. Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, October-December 2019. doi: 10.1109/TAFFC.2017.2736999.

G. Mittag, B. Naderi, A. Chehadi, and S. Möller. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Proceedings of Interspeech 2021*. ISCA, 2021. URL http://dx.doi.org/10.21437/Interspeech.2021-299. DOI: 10.21437/Interspeech.2021-299.

A. Ragano, J. Skoglund, and A. Hines. Scoreq: Speech quality assessment with contrastive regression, 2025. URL https://arxiv.org/abs/2410.06675.

J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models, 2023. URL https://arxiv.org/abs/2208.05830.

A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 2, pages 749–752 vol.2, 2001. doi: 10.1109/ICASSP.2001.941023.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4214–4217, 2010. doi: 10.1109/ICASSP.2010.5495701.

A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, C. Wood, A. Lee, and W.-N. Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound, 2025. URL https://arxiv.org/abs/2502.05139.