Towards Unsupervised Speech Enhancement Using Neural Audio Codecs

Ing. Dominik Klement, Supervised By: Doc. Ing. Lukáš Burget, Ph.D.

BRNO UNIVERSITY OF TECHNOLOGY

Motivation and Introduction

Speech signal present in the real world often gets distorted or corrupted by noise, compression, and transmission. This often complicates the day-to-day communication between two or more parties (especially for people with impaired hearing) due to decreased intelligibility, and hinders the performance of automatic speech processing systems (e.g., voice assistants).

Speech Enhancement (SE) aims to suppress/remove the present noise, while retaining as much speech information as possible. Nowadays, it is done using a machine learning (ML) model that takes noisy audio as an input and produces the cleaned version on its output, as shown in Figure 1.

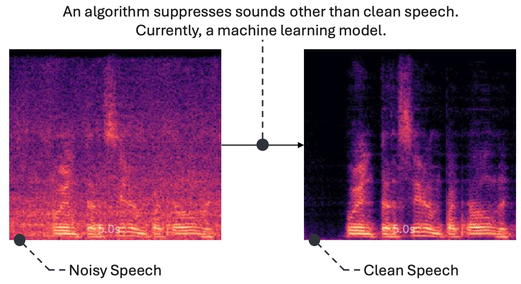


Figure 1. Example of noisy and clean (enhanced) spectrograms.

Problems with Current Solutions

To train SE ML models, we usually need examples of noisy speech and the corresponding clean speech. Therefore, majority of the current SE methods are trained using synthetic data, created by summing a randomly sampled clean speech and noise (shown in Figure 2), as it is highly challenging (often impossible) to collect parallel noisy and clean speech data in real world.

However, this creates discrepancy between training and testing, as the model does not see real noisy data during training, potentially resulting in less robustness towards real-world noisy signals.

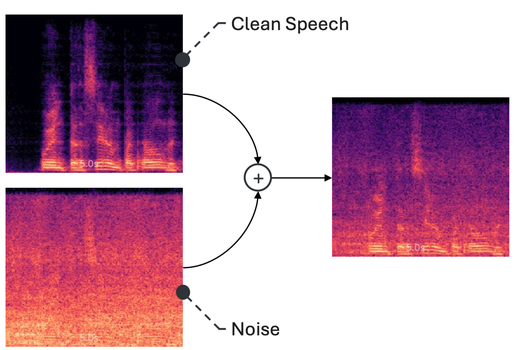


Figure 2. Schema of the synthetic training data creation process.

Proposed Method

Compared to prior approaches [1] that pre-trained speech quality estimation models, our method [2] is based solely on data — real-world noisy speech, clean speech, and noise.

The method works as follows:

1. Input noisy audio is encoded using a convolutional encoder into a sequence of embeddings (vectors).
2. The embedding sequence is divided into two parallel branches: noise and clean speech.
3. Each branch transforms the embedding sequence and decodes it back to audio.
4. Branch-wise audio sequences are combined to produce reconstructed noisy input.

The clean speech branch is adversarially trained using clean speech discriminator to produce audio sounding like clean speech (analogously for noise). Furthermore, the noisy input reconstruction forces the two branches to add up to the input, ensuring consistency — i.e., clean speech is an enhanced version of the input.

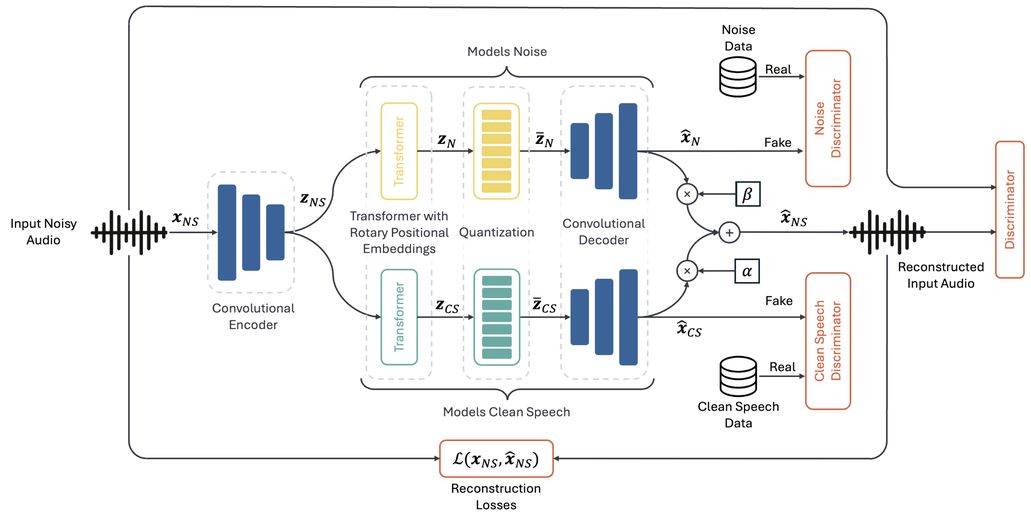
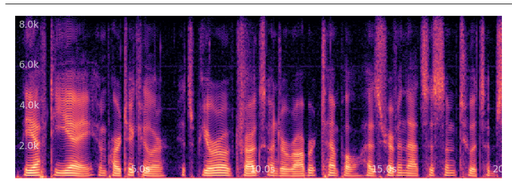
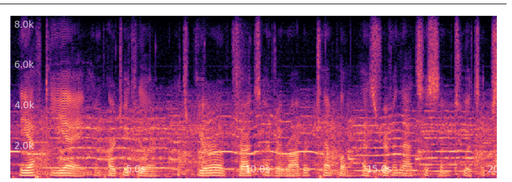


Figure 3. Schema of the proposed model architecture.

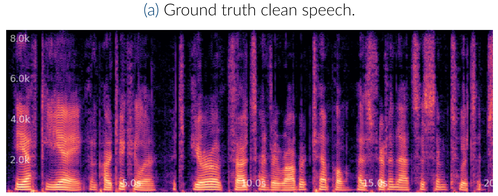
Example of Separated Speech and Noise



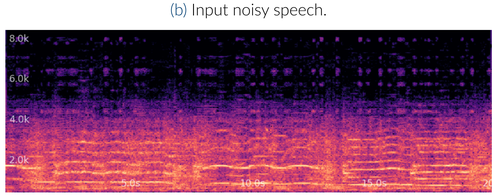
(a) Ground truth clean speech.



(b) Input noisy speech.



(c) Enhanced Speech.



(d) Estimated Noise.

Experiments

We performed the experiments using a well-established benchmark dataset VCTK+Demand [3]. Table 1 compares our model with noisy input speech and prior unsupervised SE models using non-intrusive metrics, reflecting the cleanliness and intelligibility. We can see that our model outperforms the input noisy speech and perform comparable to the prior unsupervised models, while only relying on data and not on external pre-trained models.

Table 1. Comparison of our unsupervised models with prior work. Our models are shown in the last two rows.

Model	DNSMOS↑	UTMOS↑	PESQ↑	STOI↑
Input Speech	2.54	3.10	1.97	0.92
Ground Truth	3.15	4.09	4.64	1.00
Wiener	2.54	3.05	1.93	0.92
NyTT	-	-	2.30	-
MetricGAN-U (half)	2.89	-	2.45	-
MetricGAN-U (full)	3.15	-	2.13	-
Ours	3.04	3.61	2.29	0.93

Next, Table 2 shows that our model helps downstream automatic speech recognition model Whisper, as the word error rate (WER) got decreased after using SE. Lastly, Table 3 proves that our model can be used in real-time even on CPU, making it a suitable option for on-the-fly speech denoising for online meeting software.

Table 2. Comparison of WER inferred by Whisper Large-V3 ASR model using the enhanced VCTK-Demand dataset.

Model	WER↓
Input Speech	0.07
Ground Truth	0.02
Ours	0.06

Table 3. Real time factor comparison to other SE models.

Model	RTF (CPU)↓	RTF (GPU)↓
HiFi-GAN2	-	0.500
FINALLY	-	0.030
Ours (1 CPU)	1.340	**0.023**
Ours (8 CPU)	0.378	**0.023**

Conclusion

We introduced a novel unsupervised speech enhancement architecture capable of learning to suppress noise from real-world noisy speech by performing noisy speech reconstruction and imposing data-defined priors on the model outputs. Furthermore, we showed that our method is capable of separating noise and clean speech and that it is comparable to the state-of-the-art unsupervised speech enhancement methods, while being simpler and data-driven only. This allows for easier scaling in terms of data and model size, resulting in a more robust SE model.

References

[1] Szu-Wei Fu, Cheng Yu, Kuo-Hsuan Hung, Mirco Ravanelli, and Yu Tsao. Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech. In *ICASSP 2022*.

[2] Dominik Klement, Maciejewski Matthew, Khudanpur Sanjeev, and Burget Lukáš. Unsupervised speech enhancement using data-defined priors. In *Submitted to ICASSP 2026*.

[3] Cassia Valentini-Botinhao. Noisy speech database for training speech enhancement algorithms and TTS models, 2017.