Exploiting Consistency-Preserving Loss and Perceptual Contrast Stretching to Boost SSL-based Speech Enhancement

Muhammad Salman Khan[†], Moreno La Quatra[†], Kuo-Hsuan Hung[‡], Szu-Wei Fu[*],
Yu Tsao[‡], *Sabato Marco Siniscalchi*[–]

[†]Kore University of Enna, [‡]CITI, Academia Sinica, Taiwan, [*]NVIDIA , [–] Università degli Studi di Palermo

Objectives

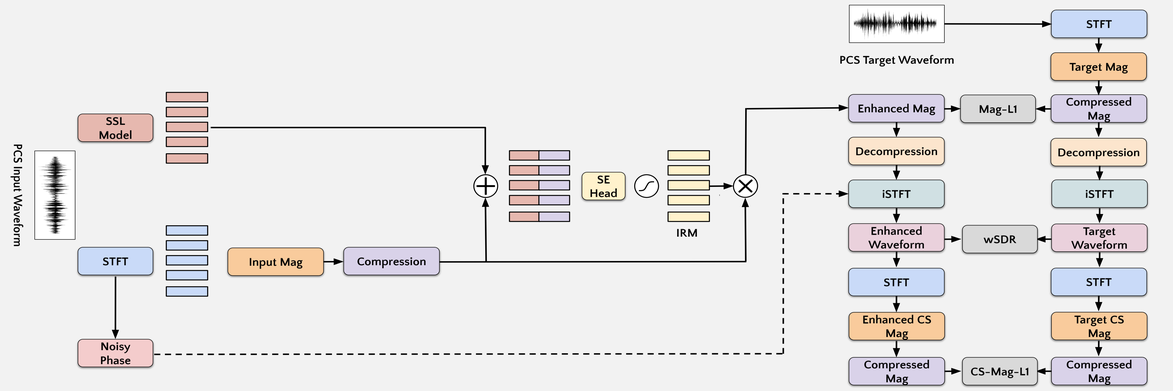
Self-supervised representation learning (SSL) has attained SOTA results on several downstream speech tasks, but SSL-based speech enhancement (SE) solutions still lag behind. To address this issue, we exploit three main ideas.

- **Mask Generation with Conformer** We adopt Conformer layers to predict the Ideal Ratio Mask (IRM), using both STFT magnitude and SSL embeddings.
- **Consistency-preserving loss**: Handles signal reconstruction inconsistencies by incorporating iSTFT in the loss calculation.
- **perceptual contrast stretching (PCS)** PCS is a spectral processing technique that aims to improve the perceptual quality of a speech signal.

Dataset

The VoiceBank-DEMAND dataset was used for this study, comprising noisy speech recordings created by mixing clean speech from the VoiceBank corpus with noise from the DEMAND dataset. It includes recordings from 30 speakers, with 28 used for training and 2 for testing. The audio was downsampled from 48 kHz to 16 kHz. Four SNR levels [0, 5, 10, 15 dB] were used for training, and [2.5, 7.5, 12.5, 17.5 dB] for testing. The training set contains 11,572 utterances, while the test set includes 824, with no overlap in speakers, noises, or SNRs.

Overall architecture



Overall architecture of the proposed SSL-based speech enhancement (SE) model. The magnitude of the STFT is indicated with Mag. CS stands for consistent, and PCS indicates perceptural contrast stretching.

Comparison with best SE solutions

Table: Comparison between proposed CS-WavLM SE model and best SE solutions tested on the VoiceBank+DEMAND dataset. SSL-based solutions are indicated by a *.

Model	PESQ	CSIG	CBAK	COVL	STOI
CMGAN [1]	3.41	4.63	3.94	4.12	**0.96**
TridentSE [2]	3.47	4.70	3.81	4.10	**0.96**
MP-SENet [3]	3.50	4.73	**3.95**	4.22	**0.96**
*BSSE [4]	3.20	4.53	3.78	4.04	**0.96**
*SSF-CVAE [5]	3.04	4.38	2.91	3.72	0.95
*CS-WavLM	3.29	4.64	3.80	4.05	**0.96**
*PCS-BSSE	3.46	**4.75**	3.49	4.20	0.95
*PCS-CS-WavLM	**3.54**	**4.75**	3.54	**4.25**	**0.96**

Ablation studies

Table: Effect on SSL-based SE performance of varying components and processes.

Model	PESQ	CSIG	CBAK	COVL	STOI
SE head					
BiLSTM	3.47	4.70	3.50	4.18	0.95
Transformer	3.52	4.73	3.53	4.23	0.95
PCS pre-processing					
w/o PCS input	3.17	4.16	3.27	3.71	0.94
w/o PCS target	3.31	4.65	**3.70**	4.06	0.96
PCS-CS-WavLM	**3.54**	**4.75**	3.54	**4.25**	**0.96**

Conclusion

We proposed PCS-CS-WavLM, a SSL based speech enhancement model that incorporates PCS and a consistency-preserving loss to reduce the performance gap with SOTA solutions. The model uses a pre-trained WavLM backbone and Conformer-based SE head, optimized with waveform-based and magnitude-based losses, including a consistency-preserving loss. Evaluated on the VoiceBank+DEMAND task, PCS-CS-WavLM showed competitive results, with ablation studies confirming the benefits of the Conformer architecture and PCS. Future work will focus on SSL pre-training objectives tailored to speech enhancement.

References

[1] Ruizhe Cao, Sherif Abdulatif, and Bin Yang.
Cmgan: Conformer-based metric gan for speech enhancement.

[2] Dacheng Yin, Zhiyuan Zhao, Chuanxin Tang, Zhiwei Xiong, and Chong Luo.
Tridentse: Guiding speech enhancement with 32 global tokens.

[3] Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling.
Mp-senet: A speech enhancement model with parallel denoising of magnitude and phase spectra.

[4] Kuo-Hsuan Hung, Szu-wei Fu, Huan-Hsin Tseng, Hsin-Tien Chiang, Yu Tsao, and Chii-Wann Lin.
Boosting self-supervised embeddings for speech enhancement.

[5] Yoonhyung Lee and Kyomin Jung.
Boosting speech enhancement with clean self-supervised features via conditional variational autoencoders.