Boosting the Intelligibility of Waveform Speech Enhancement Networks through Self-supervised Representations

Tao Sun*, Shuyu Gong*, Zhewei Wang*, Charles D. Smith†, Xianhui Wang‡, Li Xu‡, Jundong Liu*
*School of Electrical Engineering and Computer Science, Ohio University, Athens, OH 45701
†Department of Neurology, University of Kentucky, Lexington, KY 40536
‡Division of Communication Sciences, Ohio University, Athens, OH 45701

***Abstract*—The ultimate goal of speech enhancement is to improve speech quality and intelligibility. Integrating human speech elements into waveform denoising neural networks has proven to be a simple yet effective strategy for this purpose. Such integration, however, has mostly been carried out within supervised learning settings, without taking advantage of the power of the latest self-supervised learning models, which have demonstrated remarkable capability of extracting knowledge from large training sets.**

In this paper, we present *K-SENet*, a knowledge-assisted waveform framework for speech enhancement. Wave-U-Net is utilized as the baseline model and the foundation to build our framework. To achieve enhanced intelligibility, we propose a perceptual loss function that relies on self-supervised speech representations pretrained on large datasets, to provide guidance for the baseline network. Wav2vec and PASE are the choices of self-supervised models in this work. Our proposed perceptual loss is calculated upon the perceptual similarities captured by the speech representations. Minimizing this loss would ensure the denoised network outputs sound like clean human speeches. Experiments on the Noisy VCTK and modified TIMIT datasets demonstrate that our K-SENet can significantly improve the perceptual quality of network outputs.

***Index Terms*—Speech enhancement, self-supervised learning, speech representation, wav2vec, PASE**

I. Introduction

Speech enhancement (SE) aims to reduce additive disturbance components from noisy speech signals. Traditional SE solutions focus on extracting high-level features in the spectral domain to identify target audio patterns. In recent years, deep neural networks (DNN) have emerged as a popular paradigm to solve the SE problem, with early solutions mostly designed under the frequency domain.

The past three years has seen switched efforts to develop waveform-based DNN models [1]–[6]. Within this group of models, fully convolutional networks (FCNs) and their variants have become especially popular as they produced state-of-the-art performance on a variety of datasets. FCN models are built on a certain hierarchical configuration that relies on convolutional layers to extract discriminative features. Such setup equips the models with a remarkable capability of processing input data from multiple spatial or temporal scales. Many existing FCN models, however, focus on reducing the noise of generic types, lacking tailored considerations for speech data and tasks.

As the utmost goal of SE is to improve the quality and intelligibility of human speeches, efforts have been pushed forward to integrate speech elements into SE frameworks. The existing solutions can be generally grouped three categories. The first strategy uses intelligibility-related metrics (e.g., short-term objective intelligibility (STOI)) as the objective function to explicitly boost the quality of network outputs [5], [6]. The second group of solutions are commonly developed under certain GAN framework [7]–[10], where the discriminator is designed to differentiate denoised speeches from real clean signals, which makes it essentially act as an indirect quality enhancer to the generator's output. The third group of solutions [11]–[14] rely on the integration of multi-scale feature maps from certain pretrained network to provide guidance in boosting the intelligibility of the system output.

Within the third group, most pretrained networks [11]–[13] are trained for certain supervised learning tasks with rather limited data. Such *supervised representations*, specialized or biased to the considered problem, have relatively limited exportability to other tasks [15]. In recent years, self-supervised pretrained models have been increasingly utilized in many AI-related areas. For human speech tasks, self-supervised learning (SSL) models [15]–[17] have demonstrated remarkable capabilities of extracting massive amount of knowledge from large unlabeled datasets, where the high-level semantic information is commonly embedded into compact vectors, called *speech representations*. In [13] and [14], speech representations extracted by SSL models have been taken to guide the training of frequency-domain SE models.

In this paper, we propose to take advantage of the power of self-supervised speech models and integrate their representations to boost the output intelligibility of time-domain SE networks. Wav2vec [16] and PASE [15], [17], both trained on LibriSpeech [18], are the choices of SSL models in this work. A new perceptual loss function is designed to ensure the phonetic similarities captured by speech representations. Minimizing this loss would ensure our FCN outputs sound like clean human speeches, as in the training data of the pretrained models. We name our framework *K-SENet*. To the best of our knowledge, this is the first work that explores and compares the integration of wav2vec and PASE into time-domain SE networks.

II. Method

Machine learning-based speech enhancement systems convert noisy input signals $\mathbf{x}_n$ into denoised outputs $\mathbf{x}_d$ to match the ground-truth signals $\mathbf{x}_{gt}$. In both traditional and DNN approaches, training such systems has often been formulated as an optimization problem, where the differences between $\mathbf{x}_d$ and $\mathbf{x}_{gt}$ are set as the objectives to be minimized. Among waveform-based solutions, FCNs have increasingly become the state-of-the-art solutions, where *mean squared error* (MSE) is commonly used as the network objective functions. In this work, we take Wave-U-Net, an FCN variant with excellent performance on speech enhancement, as our baseline models to build our knowledge-assisted SE networks.

A. Baseline model

FCN [19] and its variants, including U-Net [20], were originally developed to solve 2D image segmentation problems. Typical FCNs are built with an encoder-decoder architecture. In the encoding path, input signals are processed through a number of layers combining convolution and pooling operations. The generated high-level latent features in the encoder are then progressively upsampled along the decoding path to match the target ground-truth. The fundamental goals of FCNs are to find mappings, with certain desired property, between paired signal sources; therefore, they are well-suited for many signal processing tasks, including waveform-based speech enhancement [1], [3], [6], [21].

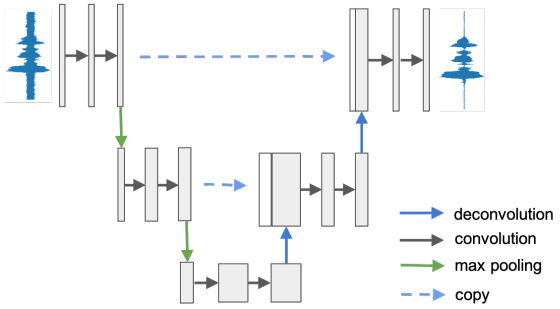

Fig. 1. An illustration of the architecture of FCN models.

Our FCN baseline model is the Wave-U-Net [2] developed by Stoller *et al*. Wave-U-Net is a one-dimensional variant of U-Net with a similar encoder-decoder architecture illustrated in Fig. 1. Two major modifications were made from the original U-Net. First, decimate operations (keep features in every other time points) are used for downsampling purposes. Second, linear interpolation, rather than deconvolution, is adopted in the decoder to reduce aliasing artifacts in the upsampling process. The inputs to Wave-U-Net are fixed-length frames split from utterances, where the corresponding utterance outputs are formed through a concatenation of the frame-level outputs in the original temporal order. In our implementation, there are 12 down-sampling layers and the filter number for the i^{th} convolutional layer are $F \cdot i$ with $F = 24$. The output frame for the models are approximately 1s ($\approx$ 16000 points).

B. Self-supervised representation learning

In self-supervised learning, models are trained to predict one part of the data from other parts [22]. SSL models for speech data and tasks commonly aim to output speech representations in the form of compact vectors that capture high-level semantic information of the raw speech data [15]–[17], [23], [24]. In this paper, we use the speech representations generated from the following SSL models.

Wav2vec model [16] is trained on LibriSpeech corpus through the contrastive predictive coding (CPC) loss [25] to pretrain speech representations for ASR tasks. Experiment results show that wav2vec can significantly improve the performance over the chosen baseline solutions. Wav2vec model consists of two networks, an encoder and a context network. The former is a seven-layer convolutional network, and its functionality is to extract latent features from the inputs. The context network combines multiple outputs from the encoder into a contextualized tensor, which then could be fed into the downstream tasks.

PASE [15], [17] model consists of a single neural encoder that encodes each raw speech waveform into a sequence of latent embeddings, which are then fed into multiple designated self-supervised tasks in parallel. These tasks include reconstructions of waveform, mel-frequency cepstrum (MFC), prosody, Log power spectrum, and other binary discrimination signals. These tasks are designed to ensure prior knowledge to be distilled into the encoder, leading to insightful and robust embeddings.

C. Architecture of our proposed K-SENet

Extensively trained on large datasets, speech representations out of SSL models are expected to contain certain reliable insights or knowledge that well characterize the training samples. As a result, these representations can potentially be used as informative inputs to train certain downstream task, where data are limited, or the labels are difficulty or expensive to obtain. We take advantage of the power of the pretrained speech representations with a different approach. Our approach is based on the thought that knowledge embedded in the representations can potentially be transferred to our task of interest, speech enhancement, to provide a broad, informative and robust guidance. More specifically, we design a pairwise loss function to ensure valuable information from the representations to be integrated into the training of our models. This loss function penalizes the dissimilarities between the representations of the denoised network outputs and those of ground-truth speeches.

Fig. 2 shows the architecture of our proposed *K-SENet* framework. Noisy inputs are fed into Wave-U-Net, which seeks to produce denoised outputs $\mathbf{x}_d$ similar to the clean ground-truth $\mathbf{x}_{gt}$. In many FCN-based SE models, including our baseline Wave-U-Net, such similarity is enforced by minimizing a pointwise MSE loss:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathbf{x}_d^{(i)}|}||\mathbf{x}_d^{(i)} - \mathbf{x}_{gt}^{(i)}||_2^2,$$

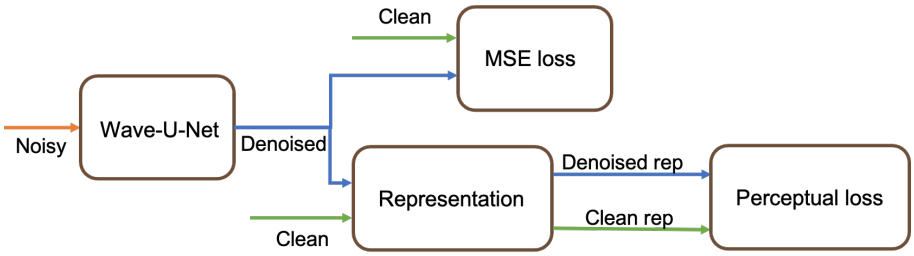

Fig. 2. Architecture of our proposed K-SENet.

where N is the number of training examples and $|\mathbf{x}|$ is the number of elements in $\mathbf{x}$.

1) Perceptual loss: It should be noted that minimization of MSE, as in Wave-U-Net, is mathematically convenient and can directly increase the signal-to-noise ratio (SNR) of the inputs. However, waveform SE networks designed with point-wise losses, including MSE, often do not have specific mechanism or components to target speech perception. In addition, minimizing MSE tends to seek point-wise averages, which leads to overly-smooth outputs [26]. As a result, phonetic information may be easily distorted along the network, resulted in outputs with poor intelligibility [27].

As mentioned in a previous section, we intend to rely on pretrained speech representations to provide a remedy, boosting the intelligibility of the denoised outputs. In this work, such remedy is designed as an auxiliary loss to enforce $\mathbf{x}_d$ and $\mathbf{x}_{gt}$ to be close in terms of phonetic similarity. We call it *perceptual loss*, which is calculated based on the L2-norm of the difference between representations of $\mathbf{x}_d$ and $\mathbf{x}_{gt}$:

$$L_{\text{perc}}(\mathbf{x}_d, \mathbf{x}_{gt}) = \sum_j \frac{1}{|\phi_j(\mathbf{x}_d)|} ||\phi_j(\mathbf{x}_d) - \phi_j(\mathbf{x}_{gt})||_2^2$$

where $\phi_j(\mathbf{x})$ represents the j^{th} vector in the representation of $\mathbf{x}$.

The *combined loss* of our K-SENet is formulated as a weighted summation of the MSE loss and perceptual loss:

$$L = \lambda_1 L_{\text{MSE}} + \lambda_2 L_{\text{perc}}$$

where λ_1 and λ_2 is weighting coefficients, which can be set manually or empirically in experiments.

III. Experiments and Results

In this section, we conduct experiments to evaluate the effectiveness of our proposed K-SENet framework. We first introduce the datasets, preprocessing steps, our training strategy, and evaluation metrics. Then, the results of our model and the corresponding baseline models are compared and analyzed. We also conduct experiments to compare our models with SE models guided by pretrained networks on some supervised learning tasks [11], [28], [29].

A. Data

Our experiments are conducted on two datasets: *Noisy VCTK* and *TIMIT with speech-shaped noises (SSNs)*.

Noisy VCTK dataset [30], [31] consists of around 400 sentences with 30 speakers (28 for training set, 2 for test set). Ten noise signals, including two artificially generated noise signals and 8 real noises, are used in training set of this database. The training set involves 4 SNRs (15, 10, 5, and 0 dB) for each noise signal, which means that there are 40 noise conditions. For each speaker of the set, around 10 different sentences are available with each noise condition. In the testing set, there are five real noise signals and 5 SNRs (17.5 dB, 12.5 dB, 7.5 dB and 2.5 dB). For each test speaker, there are around 20 sentences. All the recordings are sampled at 48kHz with 24 bits/sample. In the preprocessing, they are resampled into 16KHz. During our training, two speakers in the training set are held out for validation purpose.

TIMIT with SSNs dataset is modified from the TIMIT dataset [32], which consists of 630 speakers with 8 different dialects of American English and a total of 6300 utterances. The noisy data were generated by adding SSNs with 6 SNR levels (6 dB, 3 dB, 0dB, -3 dB, and 6 dB, respectively) onto the TIMIT utterances. For training purposes, we split the data into training set (20790 pairs), validation set (2310 pairs) and testing set (8400 pairs). All utterances are 16-bit long, 16kHz single-channel waveforms. Note that SSN, which was designed to ensure maximum masking effect, is "one common type of steady noise marker" used by audiologists in the speech-in-noise tests to evaluate human speech intelligibility in noise [33]. The average SNR level in this dataset is lower than the average SNR level in the previous noisy VCTK dataset, and its noises are randomly generated for each clear utterance. Compared with the noisy VCTK dataset, this dataset can be regarded as more challenging for the denoising task.

B. Training and evaluation

All the proposed models are implemented in PyTorch. The training is performed on an Nvidia GeForce Titan Xp GPU. Each configuration is trained by an Adam optimizer with a learning rate 0.0001 for 150 epochs. In the experiments with wav2vec representations, we set $\lambda_1 = 0.8$ and $\lambda_2 = 0.2$ in the combined loss function; In the PASE experiments, we set $\lambda_1 = 1.00$ and $\lambda_2 = 0.01$ in the combined loss function. We test and compare multiple SE networks in our experiments.

TABLE I
EXPERIMENTAL RESULTS ON NOISY VCTK DATASET.

Model	PESQ	CSIG	CBAK	COVL	STOI
Noisy	1.97	3.35	2.44	2.63	0.91
WaveCRN [34]	2.64	3.94	3.37	3.29	-
Attention_waveUNet [21]	2.62	3.91	3.35	3.27	-
D+M [35]	2.73	3.94	3.35	3.33	-
UNet [36]	2.90	4.22	3.32	3.58	-
Wave-U-Ne+wave2vec (ours)	2.93	4.22	**3.49**	3.58	**0.945**
Wave-U-Net+PASE (ours)	**2.95**	**4.23**	3.43	**3.60**	0.943

TABLE II
EXPERIMENTAL RESULTS ON TIMIT-WITH-SSNS DATASET. BEST PERFORMANCE IN EACH SNR LEVEL IS HIGHLIGHTED WITH BOLD FONT.

SNR	FCN	Loss	PESQ	STOI
-6dB	Noisy	/	1.052	0.503
	Wave-U-Net	MSE	1.351	0.738
		MSE+wav2vec	**1.387**	**0.767**
		MSE+PASE	1.363	0.760
-3dB	Noisy	/	1.061	0.578
	Wave-U-Net	MSE	1.543	0.821
		MSE+wav2vec	**1.602**	**0.840**
		MSE+PASE	1.568	0.829
0dB	Noisy	/	1.090	0.662
	Wave-U-Net	MSE	1.775	0.876
		MSE+wav2vec	**1.860**	**0.888**
		MSE+PASE	1.811	0.875
3dB	Noisy	/	1.142	0.744
	Wave-U-Net	MSE	2.032	0.912
		MSE+wav2vec	**2.121**	**0.919**
		MSE+PASE	2.078	0.908
6dB	Noisy	/	1.229	0.816
	Wave-U-Net	MSE	2.260	0.934
		MSE+wav2vec	**2.364**	**0.938**
		MSE+PASE	2.326	0.930

For each model, the best-performing network setup in training, measured by validation PESQ, is loaded in testing.

The implementation of wav2vec is downloaded from the project GitHub site and the pretrained model is *Wav2Vec large*. The outputs of its encoder are taken as the speech representations in this work. For PASE, we choose PASE+ model [17] and download it from Google Drive. Both models are trained on the full 960-hour LibriSpeech training set [18]. PESQ and STOI [6] are two widely used metrics for perceptual evaluation. We utilize them in our experiments to evaluate the perceptual performances.

In our VCTK experiments, composite scores introduced in [37], which include MOS predictor of speech distortion (CSIG), MOS predictor of intrusiveness of background noise (CBAK), and MOS predictor of overall processed speech quality (COVL), have also been used to evaluate the competing models.

C. Results and analysis

Noisy VCTK Table I shows the results of our models (bottom two lines) and those of some state-of-the-art solutions on the Noisy VCTK dataset. As evident, our models outperform the state-of- the-art models in all metrics. Comparing our two models, the model guided by PASE (*Wave-U-Net+PASE*) has a better performance measured with PESQ and the composite scores, and the model guided by wave2vec (*Wave-U-Net+wave2vec*) produces better results measured with STOI.

TIMIT with SSNs Table II reports the results of the baseline Wave-U-Net model and those of the two proposed knowledge-assisted models on the TIMIT-with-SSNs dataset. Wave-U-Net uses MSE as its objective function. The corresponding K-SENets are trained with the combined loss based on wav2vec and PASE representations, respectively. It is evident that our *MSE+wav2vec* model outperforms the baselines with significant margins, measured by STOI and PESQ. It is worth noting that the performance improvements have been demonstrated at every SNR level. Our *MSE+PASE*

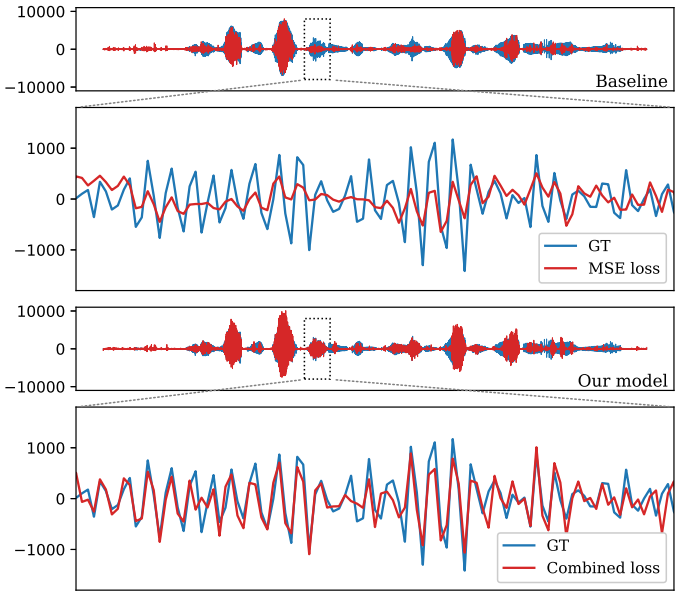

Fig. 3. Denoised outputs of a sample utterance from Wave-U-Net (MSE-only) network (top two rows) and our MSE+wav2vec model (bottom two rows).

TABLE III
COMPARISONS OF PERCEPTUAL LOSS AND DEEP FEATURE LOSS.

Dataset	Model	PESQ	CSIG	CBAK	COVL	STOI
VCTK	Noisy	1.97	3.35	2.44	2.63	0.91
	Deep Feature Loss [11]	-	3.86	3.33	3.22	-
	Deep Feature Loss (our run)	2.59	3.85	3.32	3.22	0.931
	Speaker Embedding	2.60	3.95	3.23	3.27	0.936
	Acoustic Event Classification	2.82	3.68	3.38	3.25	0.942
	Wave-U-Net + wave2vec	2.93	4.22	**3.49**	3.58	**0.945**
	Wave-U-Net + PASE	**2.95**	**4.23**	3.43	**3.60**	0.943
TIMIT	Noisy	1.11	1.99	1.43	1.47	0.66
	Deep Feature Loss (our run)	1.54	2.69	2.17	2.05	0.84
	Wave-U-Net + wave2vec	**1.87**	2.91	2.14	2.31	**0.87**
	Wave-U-Net + PASE	1.83	**3.16**	**2.30**	**2.46**	0.86

model performs better than the baseline model in PESQ at every SNR level and STOI in low SNR levels (-3 and -6 dB). Overall, these results provide a strong evidence that the proposed perceptual loss can indeed substantially and consistently enhance the perceptual properties of the denoised utterances.

Fig. 3 shows the comparison of the models on a particular audio clip. Blue lines show the ground-truth waveform and red lines are the denoised network outputs. The top two rows show the ground-truth and denoised output obtained from the baseline Wave-U-Net (with MSE loss). The bottom two rows show the corresponding waveform output generated from our *MSE+wav2vec* network. The second row and the fourth row are the amplified views of the highlighted segments in the first and third rows, respectively. The clip produced by the baseline model, as shown in the top two rows, has fewer spikes and smaller amplitudes, which might be due to the tendency of MSE loss in producing overly-smooth results [26]. In contrast, the result generated by our K-SENet matches the ground-truth rather well, thanks to the speech information brought by the speech representations.

D. Perceptual loss vs. deep feature loss

In [11], a supervised network is trained jointly on scene classification and audio tagging targets. A loss named *deep feature loss* based on a multitude of features at different scales of the network is proposed to provide a speech guidance for the SE tasks. Comparing with [11], our framework has an advantage that the adopted speech representations were both trained on large speech datasets with the self-supervised learning paradigm. This combination greatly enhanced the knowledge to be distilled into the representations. Table III illustrates the comparison of our combined loss based on the perceptual loss and the deep feature loss on the Noisy VCTK dataset and the TIMIT-with-SSNs dataset. Note that the model marked as *our run* are trained by us with the original code provided by [11]. In [13], deep feature loss

functions based on the speaker embedding model [28] and the acoustic event classification model [29] are proposed to train frequency-domain SE networks. We also design models with these two deep feature loss functions to replace the perceptual loss in our setup and trained them on the noisy VCTK dataset. As shown in table III, our models guided by the speech representations, whether it is wav2vec or PASE, achieve much better performances than models based on deep feature loss functions in almost each metrics listed in the table.

IV. Conclusion

In this paper, we propose a knowledge-assisted framework to enhance the perceptual properties of the denoised outputs of waveform SE networks. Our approach relies on speech representations trained on large speech datasets to provide valuable insights and guidance regarding what clean speeches sound like. Experiments on both Noisy VCTK and TIMIT with SSNs datasets show that our models achieve significant perceptual performance gains to both the baseline and state-of-the-art models. The take-home message is that pretrained speech representation models, if properly integrated, do provide great help for SE. To explore integration of pretrained models with more speech enhancement networks, as well as their applications to other speech tasks, are our ongoing efforts.

References

[1] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 IEEE APSIPA ASC*. IEEE, 2017, pp. 006–012.

[2] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[3] S. Gong, Z. Wang, T. Sun, Y. Zhang, C. D. Smith, L. Xu, and J. Liu, "Dilated fcn: Listening longer to hear better," in *2019 IEEE WASPAA*. IEEE, 2019, pp. 254–258.

[4] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *2019 IEEE ICASSP*. IEEE, 2019, pp. 6875–6879.

[5] ——, "A new framework for cnn-based speech enhancement in the time domain," *IEEE/ACM TALSP*, vol. 27, no. 7, pp. 1179–1188, 2019.

[6] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, no. 9, pp. 1570–1584, 2018.

[7] S. Pascual, A. Bonafonte, and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[8] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5024–5028.

[9] M. H. Soni, N. Shah, and H. A. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5039–5043.

[10] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *2018 IEEE ICASSP*. IEEE, 2018, pp. 5414–5418.

[11] F. G. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[12] M. Kegler, P. Beckmann, and M. Cernak, "Deep Speech Inpainting of Time-Frequency Masks," in *Interspeech*, 2020, pp. 3276–3280.

[13] S. Kataria, J. Villalba, and N. Dehak, "Perceptual loss based speech denoising with an ensemble of audio pattern recognition and self-supervised models," in *2021 IEEE ICASSP*. IEEE, 2021, pp. 7118–7122.

[14] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, "Improving perceptual quality by phone-fortified perceptual loss for speech enhancement," *arXiv preprint arXiv:2010.15174*, 2020.

[15] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks," in *2019 Interspeech*, 2019, pp. 161–165.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised Pre-Training for Speech Recognition," in *2019 Interspeech*, 2019, pp. 3465–3469.

[17] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *2020 IEEE ICASSP*. IEEE, 2020, pp. 6989–6993.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE ICASSP*. IEEE, 2015, pp. 5206–5210.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE CVPR*. IEEE, 2015, pp. 3431–3440.

[20] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *2015 MICCAI*. Springer, 2015, pp. 234–241.

[21] R. Giri, U. Isik, and A. Krishnaswamy, "Attention wave-u-net for speech enhancement," in *2019 IEEE WASPAA*. IEEE, 2019, pp. 249–253.

[22] X. Liu, F. Zhang, Z. Hou, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *arXiv preprint arXiv:2006.08218*, 2020.

[23] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *arXiv preprint arXiv:2007.06028*, 2020.

[24] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," in *2020 ICLR*, 2020.

[25] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[26] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE CVPR*. IEEE, 2017, pp. 4681–4690.

[27] A. Pandey and D. Wang, "A new framework for supervised speech enhancement in the time domain," in *2018 Interspeech*, 2018, pp. 1136–1140.

[28] J.-w. Jung, S.-b. Kim, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms," *arXiv preprint arXiv:2004.00526*, 2020.

[29] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.

[30] C. Valentini-Botinhao *et al.*, "Noisy speech database for training speech enhancement algorithms and tts models," *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*, 2017.

[31] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *2016 SSW*, 2016, pp. 146–152.

[32] J. S. Garofolo, "Timit acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[33] C. G. Le Prell and O. H. Clavier, "Effects of noise on speech recognition: Challenges for communication by service members," *Hearing research*, vol. 349, pp. 76–89, 2017.

[34] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *arXiv preprint arXiv:2004.04098*, 2020.

[35] J. Yao and A. Al-Dahle, "Coarse-to-fine optimization for speech enhancement," *arXiv preprint arXiv:1908.08044*, 2019.

[36] A. E. Bulut and K. Koishida, "Low-latency single channel speech enhancement using u-net convolutional neural networks," in *2020 IEEE ICASSP*. IEEE, 2020, pp. 6214–6218.

[37] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.