IMPROVING PERCEPTUAL QUALITY BY PHONE-FORTIFIED PERCEPTUAL LOSS FOR SPEECH ENHANCEMENT

Tsun-An Hsieh[1] , Cheng Yu[1], Szu-Wei Fu[1] , Xugang Lu[2], and Yu Tsao[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]National Institute of Information and Communications Technology, Japan

ABSTRACT

Speech enhancement (SE) aims to improve speech quality and intelligibility, which are both related to a smooth transition in speech segments that may carry linguistic information, e.g. phones and syllables. In this study, we took phonetic characteristics into account in the SE training process. Hence, we designed a phone-fortified perceptual (PFP) loss, and the training of our SE model was guided by PFP loss. In PFP loss, phonetic characteristics are extracted by *wav2vec*, an unsupervised learning model based on the contrastive predictive coding (CPC) criterion. Different from previous deep-feature-based approaches, the proposed approach explicitly uses the phonetic information in the deep feature extraction process to guide the SE model training. To test the proposed approach, we first confirmed that the *wav2vec* representations carried clear phonetic information using a t-distributed stochastic neighbor embedding (t-SNE) analysis. Next, we observed that the proposed PFP loss was more strongly correlated with the perceptual evaluation metrics than point-wise and signal-level losses, thus achieving higher scores for standardized quality and intelligibility evaluation metrics in the Voice Bank–DEMAND dataset.

Index Terms— Speech enhancement, perceptual loss, contrastive predictive coding, representation learning, unsupervised learning

1. INTRODUCTION

In real-world speech-related applications, speech signals may be distorted by environmental noise, and thus constrain the performance achievable on a target task. To address this issue, speech enhancement (SE) has been studied for decades. Numerous signal processing-based methods [1, 2, 3, 4] have been proposed. These methods are based on the assumed statistical properties of speech and noise signals. When these assumed properties are unfulfilled, SE performance may drop considerably. With recent advances in neural network (NN) models, SE performance has increased notably. Well-known NN models, such as deep denoising autoencoder (DDAE) [5], deep neural networks (DNNs) [6], recurrent neural networks (RNNs) [7], long short-term memory (LSTM) [8], convolutional neural networks (CNNs) [9], fully convolutional networks (FCNs) [10, 11], convolutional recurrent neural networks (CRNNs) [12], and generative adversarial networks (GANs) [13, 14, 15, 16, 17, 18, 19] have made notable improvements over traditional signal processing-based SE methods.

For these NN-based SE approaches, designing a suitable objective function is quite important. Traditionally, point-wise distances have often been used as an objective function. Point-wise distances are calculated as L1 and/or L2 norms between paired noisy-clean speech signals, attempt to recover information on a signal level. Recent studies have indicated that objective functions based on point-wise distances may not fully reflect the perceptual difference between noisy and clean speech signals. As the purpose of SE is to recover speech quality and intelligibility, objective functions that use perceptual metrics have been investigated for NN-based SE. In those studies, perceptual metrics were modified in their differentiable alternatives for convenient gradient calculation in NN parameter optimization. Some notable works are the perceptual evaluation-based loss function [20], joint signal distortion ratio (SDR) perceptual evaluation for speech quality optimization [21], and modified short-time objective intelligibility (STOI) loss functions for network optimization [10, 22, 23]. Along this line, several studies focus on training NN models with target metrics for SE tasks [24], as well as which in GAN approaches like HiFi-GAN [18] and MetricGAN [19].

Other than direct optimizations for evaluation metrics, objective functions can be designed to minimize loss based on representations in latent spaces, where the latent spaces are from a pre-trained model that is given paired noisy-clean speech signals. For example, in style transfer studies of computer vision, [25] proposed training feed-forward networks based on perceptual loss. In [26], the authors proposed utilizing an acoustic scene recognition network's latent spaces for the loss function, termed deep feature loss (DFL), and obtained promising results. We believe that such objective function can be further improved by using latent representations extracted using an NN model that is more relevant to the SE task.

In this paper, we explicitly take phonetic characteristics into consideration for SE. To emphasize these characteristics, a phone-fortified perceptual (PFP) loss is designed for SE model optimization. Experimental results have shown that *wav2vec*-encoded phonetic characteristics are representative of phonetic information. We can conclude that our proposed framework, which incorporates PFP loss, presents great improvement over other perceptual optimization approaches for SE tasks.

2. RELATED WORKS

In this section, we first present DFL, which was mentioned in the previous section, with a more detailed discussion in Section 2.1. We then review the perceptual metrics approximated with trained networks. Such a network can work as a discriminator in a GAN or a stand-alone metric. Last but not least, we review methods that maximize the mutual information within a context in Section 2.3.

2.1. Deep Feature Loss

The idea to incorporate acoustic scene recognition in SE was proposed in DFL [26]. According to [18], the latent features from a pre-trained recognition network (used for machine perception) are
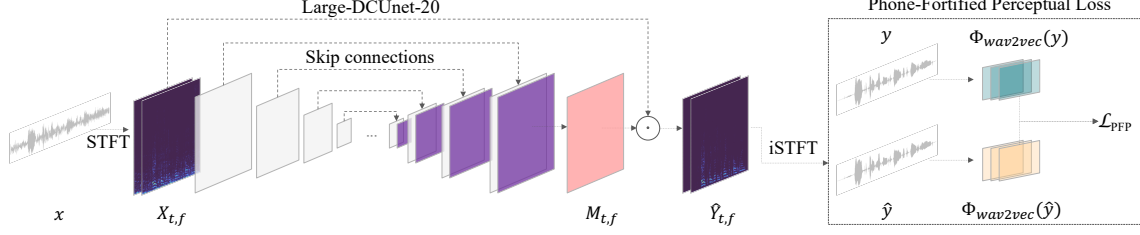
Fig. 1. **Block diagram of our proposed network**. The enhancement model estimates a cRM by the noisy spectrum, and consequently produces an enhanced spectrum.

used to approximate human perception (SE, in this case). However, acoustic scene recognition seems to be lacking in phonetic characteristic information, which we believe is the key to optimizing SE with respect to human perception.

2.2. MetricGAN and HiFi-GAN

MetricGAN [19] applies a discriminator (also called Quality-Net [27]) to approximate the behavior of the evaluation functions of interest. The predicted score can also be treated as a special case of perceptual loss, with an embedding dimension equal to 1. Due to the limited dimensions, Quality-Net is easily fooled by the speech generated by the updated generator. Therefore, MetricGAN needs to iteratively train between the generator and the discriminator which slows down its training efficiency. HiFi-GAN [18] incorporates the idea of GAN training and deep feature loss. However, its deep feature loss is based on the discriminator, which may not be highly related to human perception.

2.3. Contrastive Predictive Coding (CPC) and *wav2vec*

To make the best use of phonetic characteristics, we surveyed representation learning methods that could automatically discover representative features directly from raw data. Some notable approaches have been proposed recently in this research field. For instance, CPC [28] is an unsupervised method that proposes to extract information-rich features from high-dimensional data. It is the probabilistic contrastive loss in CPC approaches that help capture latent features which maximally promote the precision of predictions for future samples. Then, we focused on speech-related applications utilizing representation learning approaches. The unsupervised automatic speech recognition (ASR) *wav2vec* [29], utilizing the CPC technique, shows great performance in recognition accuracy. In practice, speech signals are first encoded with an encoder network that extracts features that are rich in phonetic characteristics. An ASR decoder is then trained based on these features as inputs. Based on our understanding in speech acoustics, phonetic characteristic information can be greatly distorted due to noise contamination.

To our instinct, it is essential that the objective functions for SE tasks are more focused on the integrity of speech enhanced with respect to the feature domain of phonetic characteristics (or phonetic characteristic level) rather than simply considering those at the signal level. That is, SE that aims to minimize signal-level losses does not guarantee integrity at the phonetic characteristic level, and thus suffers from observable mismatches with respect to perceptual metrics, despite satisfactory signal-level losses. As a result, our proposed PFP loss adopts the latent space of the *wav2vec* encoder network, which extracts phonetic characteristic-rich features to create a loss domain. We then design a PFP loss that performs distance calculation at phonetic characteristic level. We also incorporate mean ab-

solute error (MAE) with PFP loss to ensure optimization at both the phonetic characteristic level and at the signal level for SE.

3. PROPOSED FRAMEWORK

Two aspects are important in designing an effective NN-based SE system. One is the model architecture, by which complex mapping functions between noisy and clean speech can be efficiently approximated. The other is the objective function, by which necessary speech information should be retained for SE during model optimization.

3.1. Model Architecture

Inspired by the deep complex U-Net (DCUnet) in [30], we designed a modified framework that estimates complex ratio masks (cRM) for a noisy complex spectrum with a different normalization mechanism. More specifically, as shown in Fig. 1, a noisy speech signal is first converted to a complex spectrum through short-time Fourier transform (STFT), and the enhancement model generates a cRM. Subsequently, the noisy spectrum is point-wise multiplied by the cRM to derive the final enhanced spectrum, and is transformed to a waveform by inverse STFT (iSTFT). Here, according to [30], a scheme that produces cRM with a complex neural network (cRM$\mathbb{C}$n) is used in this work. We take the Large-DCUnet-20 [30] as a reference architecture for our enhancement model. Because a number of previous works [31, 32] have indicated that instance normalization outperforms batch normalization on generation tasks by preserving the independence of samples in a mini-batch, we substitute the batch normalization layers in the Large-DCUnet-20 with instance normalization layers. To describe the enhancement process precisely, given the noisy input speech signal x the noisy spectrum $X_{t,f}$ is derived by STFT, such that $X_{t,f} = \text{STFT}(x)$. Then, the enhancement model generates a cRM $M_{t,f} = f_\theta(X_{t,f})$ to produce the enhanced spectrum that $\hat{Y}_{t,f} = M_{t,f} \cdot X_{t,f}$, and transforms it to the enhanced waveform $\hat{y}$ by iSTFT.

3.2. Phone-Fortified Perceptual Loss

As the outputs of an SE system are always speech signals, we propose the PFP loss considering attributes that match the human perception of speech. Because one-hot encoded labels assume that the similarity between two arbitrary labels is zero, and this may underestimate the correlations between phones and consequently restrict the feature representations, we thus employ an unsupervised trained model to compute the PFP loss. Because speech signals carry phones, while noise does not carry phones, we desire a loss model that generates features that represent phonetic characteristics. Additionally, because CNN is known for its shift-invariance, which is
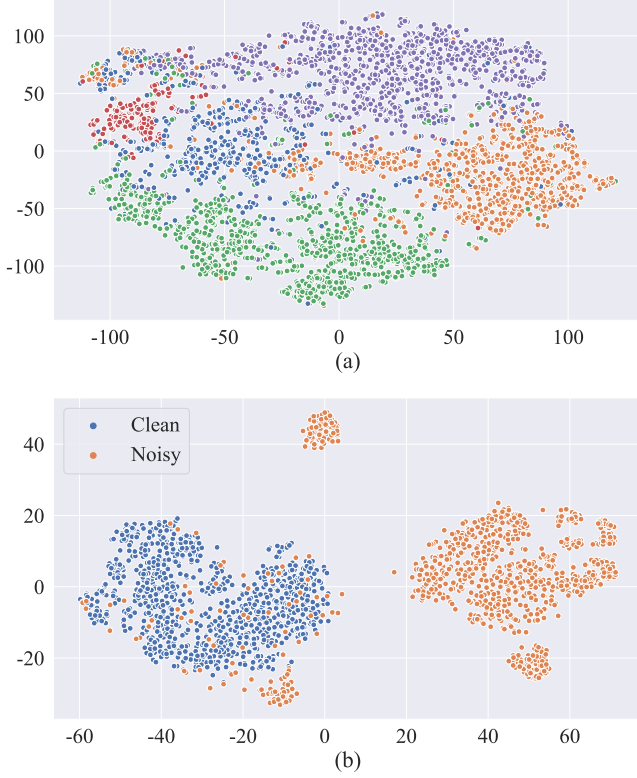
Fig. 2. t-SNE analysis on *wav2vec* encoded feature map. (a) Feature map of five phone classes. (b) Feature map of clean and noisy utterances.

similar to the behavior of the perceptual evaluation of speech quality (PESQ) [33], the *wav2vec* encoder appears to be suitable for our work. For PFP loss model, the CNN-based encoder $\Phi_{wav2vec}$ of *wav2vec large* transforms raw waveform inputs into a batch of sequence of 512 dimensional vectors. The parameters of the loss model are fixed during training. In contrast to previous works on perceptual loss that utilize activations in multiple layers, we merely extract the final outputs, and the PFP loss is defined as:

$$\mathcal{L}_{\mathrm{PFP}}(x, y) := \mathbb{E}_{x,y\sim\mathcal{D}}[\|\Phi_{wav2vec}(y) - \Phi_{wav2vec}(f_\theta(x))\|_1]. \quad (1)$$

For the given paired noisy speech x and clean speech y sampled from dataset $\mathcal{D}$, the PFP loss minimizes the phonetic distance between the clean speech and the enhanced one.

4. EXPERIMENTS

In this section, we begin with the selected dataset and the evaluation metrics that are used as a standard benchmark. Next, we provide visualizations that demonstrate that the features generated by the PFP loss model are correlated with PESQ and STOI. Finally, it is shown that the proposed modification achieves outstanding performance in terms of perceptual qualities.

4.1. Voice Bank–DEMAND Dataset

To compare our proposed SE system with other recent approaches, the Voice Bank–DEMAND dataset [34, 35] was used for evaluation. In this dataset, utterances recorded by 28 speakers out of a total 30 speakers are used for training and the utterances from the remaining
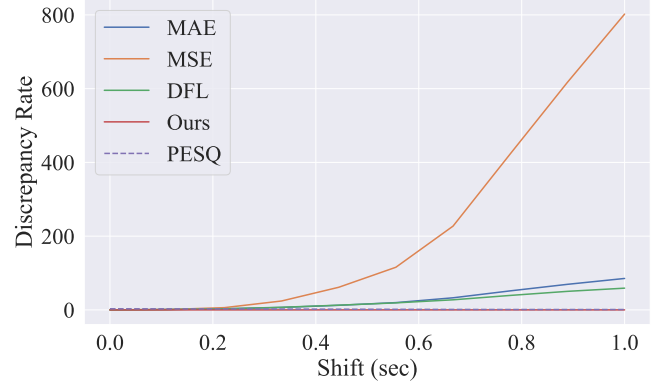

Fig. 3. Trending of different losses with its value reactions under different levels of time shift.

2 speakers are used for testing. In the training set, noisy mixtures were synthesized using 10 types of noise at 4 different SNR levels, ranging from 0 dB to 15 dB, and 5 types of unseen noises, ranging from 2.5 dB to 17.5 dB were added to the testing set.

4.2. Evaluation Metrics

Following prior works evaluated on the Voice Bank–DEMAND dataset, we used five metrics, which were CSIG, CBAK, COVL, introduced in [36], PESQ, and STOI to measure the performance of the proposed method. CSIG, CBAK, and COVL represent the signal distortion, background intrusiveness, and the overall quality with the same scale of mean opinion score, respectively. PESQ and STOI quantify the perceptual quality and the intelligibility of a speech signal, respectively.

4.3. Visualizing *wav2vec* Features

As mentioned in Section 3.2, features extracted by *wav2vec* are informative for phonetic characteristics. To verify that the generated features are competent at preserving phonetic information, we present an illustration that projects high-dimensional features into a two-dimensional space by t-SNE, which is a non-linear dimensionality reduction technique widely used for verifying the effectiveness of embeddings. As shown in Fig. 2 (a), five types of phones are separated into five groups. It is believed that features generated by *wav2vec* can represent phonetic characteristics. In Fig. 2 (b), most of the noisy and clean speech can be distinguished by the PFP loss model, meaning that the enhancement model is updated based on the phonetic distance between the estimation and its target.

4.4. Shift Sensitivity

To evaluate the shift sensitivity of different losses, we illustrated the trending of losses over a time shift. Because losses have different scales, directly comparing loss change is impractical. Instead, we used the ratio of a loss to its second-smallest value, which is termed the discrepancy rate in Fig. 3. As shown in Fig. 3, PESQ is consistent at al level of time shift. Comparing all the losses listed, DFL and our proposed PFP loss are more insensitive than the other three signal-level losses, and more like auditory perception.

4.5. Relations Between Perceptual Metrics and Different Losses

To analyze the relation between perceptual metrics and other losses, we compared several different losses to the corresponding metric

Fig. 4. This figure illustrates the correlation of PESQ and STOI to different losses. To quantify how much a loss is correlated to a metric, we note the Pearson correlation coefficient in the parentheses. The higher absolute value of PCC indicates that the loss is more correlated to the metric.

Table 1. Our proposed method versus some well performing methods with respect to different metrics. DFL[†] shows the results from the official source code and released parameters.

Model	PESQ	CSIG	CBAK	COVL	STOI
Noisy	1.97	3.35	2.44	2.63	0.92
Wiener [3]	2.22	3.23	2.68	2.67	–
SEGAN [13]	2.16	3.48	2.94	2.80	0.93
DFL [26]	–	3.86	3.33	3.22	–
DFL[†]	2.58	3.80	2.72	3.19	0.93
MetricGAN [19]	2.86	3.99	3.18	3.42	**0.94**
HiFi-GAN [18]	2.94	4.07	3.07	3.49	–
SDR-PESQ [21]	3.01	4.09	3.54	3.55	–
T-GSA [37]	3.06	4.18	**3.59**	3.62	–
Ours	**3.09**	**4.22**	3.05	**3.67**	0.94

[†] https://github.com/francoisgermain/SpeechDenoisingWithDeepFeatureLosses.git.

scores on the testing set. Note that due to the limitations of the report layout, we report the relationship between PESQ, STOI, and different loss values. A report of the full comparisons and analyses can be accessed on our GitHub page[1]. In Fig. 4, we illustrate the relation of PESQ and STOI to five losses including, MAE, mean squared error (MSE), weighted signal-to-distortion-ratio (wSDR), Deep Feature Loss, and the PFP loss. In Fig. 4, MAE and MSE have a similar correlation to the two metrics, and the four groups of points represent the four SNR levels in the testing set. For the first three losses, there is no obvious relation to the two metrics. However, the more explicit tendencies are that DFL and PFP loss are correlated with PESQ and STOI. Here, the Pearson correlation coefficient (PCC) is utilized to quantify the correlation between metrics and losses. The PCCs of losses are shown inside the parentheses in Fig. 4. The PCC of PFP loss is much higher than the PCCs of all the other metrics. Comparing Table 1 and Table 2, although DFL is more highly correlated with PESQ than the other three signal-level metrics, it has similar results to wSDR, MAE, and MSE. However, because the PFP loss model measures how different the features are in terms of phonetic information, which is more salient to the human auditory system, it is reasonable that the PFP loss is highly correlated with PESQ and STOI.

Table 2. Our proposed proposed model using our proposed losses versus using different losses with respect to evaluation metrics.

Loss	PESQ	CSIG	CBAK	COVL	STOI
wSDR [30]	2.58	3.00	3.18	2.76	0.93
MSE	2.60	3.31	3.19	2.94	0.93
MAE	2.62	3.47	3.20	3.02	0.93
Ours	3.09	**4.22**	3.05	**3.67**	0.94
Ours + MAE	**3.11**	4.15	**3.52**	3.64	**0.95**

4.6. Qualitative Analysis

In Table 1, we compare prior approaches using non-signal-level objectives and studies on perception-specialized losses. We achieved the highest PESQ score of all the methods. Note that perceptual losses are inferior in dealing with background noise, as the signal level distance is not considered when the networks is updated, so CBAK remains low compared with other approaches. To circumvent this shortage, we incorporate MAE into the PFP loss, which forces the enhancement model to match the targets in the latent space and the signal level. In Table 2, we further compare several losses with the same enhancement model. Signal-level losses derived lower PESQ but higher CBAK compared with the PFP loss. Compared with the results that are optimized only by PFP loss, adding MAE substantially eliminates the noise intrusion, thereby achieving a higher CBAK score.

5. CONCLUSION

In this paper, based on prior works on perceptual objective functions, we proposed PFP loss to improve deep feature loss by incorporating the phonetic characteristics of deep feature extraction. As demonstrated by the experimental results, our SE mechanism outperforms numerous approaches using GANs or specialized objective function for auditory sense. We also presented multiple experiments showing that the improved loss has a higher correlation with PESQ than signal-level losses, and thus the enhancement model receives informative feedback. Additionally, to tackle the shortage of perceptual loss, we regulated PFP by MAE, so the enhancement model not only learned to distinguish latent representations, but also learned in the signal-level space. Consequently, the disruption by background noise was notably eliminated.

6. REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE TASSP*, vol. 27, no. 2, pp. 113–120, 1979.

[2] J.S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[3] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Proc. ICASSP*, 1987.

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

[5] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder.," in *Proc. Interspeech*, 2013, vol. 2013, pp. 436–440.

[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2014.

[7] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. ICASSP*, 2014.

[8] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Proc. LVA/ICA*, 2015.

[9] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *Proc. ICASSP*, 2018.

[10] 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.

[11] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. ICASSP*, 2019.

[12] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Proc. Interspeech*, 2018.

[13] Santiago S. Pascual, A. Bonafonte, and J. Serrà, ," in *SEGAN: Speech Enhancement Generative Adversarial Network*, 2017.

[14] M. Soni, N. Shah, and H. Patil, "Time-frequency masking-based speech enhancement using generative adversarial network," in *Proc. ICASSP*, 2018.

[15] A. Pandey and D. Wang, "On adversarial training and loss functions for speech enhancement," in *Proc. ICASSP*, 2018.

[16] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *Proc. ICASSP*, 2019.

[17] S. Qin and T. Jiang, "Improved wasserstein conditional generative adversarial network speech enhancement," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 181, 2018.

[18] J. Su, Z. Jin, and A. Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *Proc. Interspeech*, 2020.

[19] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin, "Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *Proc. ICML*, 2019.

[20] J. Martín-Doñas, A. Gomez, J. Gonzalez, and A. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[21] J. Kim, M. El-Kharmy, and J. Lee, "End-to-end multi-task denoising for joint sdr and pesq optimization," *arXiv preprint arXiv:1901.09146*, 2019.

[22] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure," in *Proc. ICASSP*, 2018.

[23] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *Proc. ICASSP*, 2018.

[24] S.-W. Fu, C.-F. Liao, and Y. Tsao, "Learning with learned loss function: Speech enhancement with quality-net to improve perceptual evaluation of speech quality," *IEEE Signal Processing Letters*, vol. 27, pp. 26–30, 2019.

[25] J. Johnson, A. Alahi, and F.-F. Li, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, 2016.

[26] F. Germain, Q. Chen, and V. Koltun, "Speech denoising with deep feature losses," *Proc. Interspeech*, 2019.

[27] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *Proc. Interspeech*, 2018.

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

[29] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *Proc. Interspeech*, 2019.

[30] H.-S. Choi, J.-H. Kim, J. Huh, A. Kim, J.-W. Ha, and K. Lee, "Phase-aware speech enhancement with deep complex u-net," in *Proc. ICLR*, 2018.

[31] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[32] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization.," in *Proc. ICCV*, 2017.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.

[34] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. O-COCOSDA/CASLRE*, 2013.

[35] J. Thiemann, N. Ito, and E. Vincent, "Demand: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proc. Meetings Acoust*, 2013.

[36] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE TSAP*, vol. 16, pp. 229–238, 2008.

[37] J. Kim, M. El-Khamy, and J. Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, 2020.