



Improve Speech Enhancement using Perception-High-Related Time-Frequency Loss

Ding Zhao, Zhan Zhang, Bin Yu, Yuehai Wang

School of Information and Electronic Engineering, Zhejiang University, China
zhao_ding@zju.edu.cn, zhan_zhang@zju.edu.cn, yu_bin@zju.edu.cn,
wyuehai@zju.edu.cn

Abstract

Commonly used speech enhancement (SE) training losses like mean absolute error (MAE) loss and short-time Fourier transformation (STFT) loss suffer from the problem of mismatch with the speech quality, which leads to suboptimal training results. To tackle this problem, we propose a new loss named perception-high-related time-frequency (PHRTF) loss. The proposed loss modifies STFT loss by adding a trainable module named perceptual spectrum mask predictor (PSMP). This module can predict the perceptual spectrum mask (PSM) from the magnitude spectrum of enhanced and clean speech. Further, PHRTF loss multiplies the amplitude error spectrum (AES) with PSM to emphasize perception-relevant loss components to correlate highly with the speech quality. We conduct experiments on the VoiceBank-DEMAND dataset, and the results show that PHRTF loss has a significantly higher correlation with the speech quality than other losses. Meanwhile, PHRTF loss outperforms other losses and improves PESQ by 0.32 over MAE loss and 0.19 over STFT loss on the training of Wave-U-Net. We also apply PHRTF loss to a more advanced SE model, and the training result outperforms other competitive baselines.

Index Terms: speech enhancement, perceptual loss, masking, speech quality optimization

1. Introduction

Speech Enhancement (SE) aims to recover clean speech from noisy speech and improve the quality and intelligibility of speech. SE has a wide range of applications in daily life, such as automatic speech recognition (ASR) [1] and hearing aids [2]. Traditional SE methods are mainly statistical-based models. These methods can deal with stationary noise, but they perform poorly in the face of non-stationary noise due to their unreasonable assumptions about noise signals.

With the rapid development of deep learning (DL), more and more DL-based speech enhancement methods have been proposed in recent years. With the powerful feature extraction and modeling capabilities of neural networks (NN), DL-based speech enhancement methods have significantly improved over traditional methods. One current research focus is to propose better SE models with more efficient structures and more powerful NN components, such as convolutional neural networks (CNN) and recurrent neural networks (RNN) based models [3]–[6], transformer-based models [7]–[9] and complex neural networks based models [10]–[12].

Besides the model structure, loss functions for SE training have also received increasing attention in recent years. The commonly used losses like mean absolute error (MAE) loss and short-time Fourier transformation (STFT) loss simply

calculate the numerical deviation between the enhanced and clean speech in the time or time-frequency domain, which has an apparent mismatch with the perceptual quality of speech. As a result, a lower loss value may not guarantee a better speech quality, leading to suboptimal training results.

To solve this mismatch problem, many researchers have proposed new useful losses. These losses can be mainly divided into deep feature-based losses [13]–[16] and metric-mimic losses [17]–[21]. For deep feature-based losses, they first train a network under certain speech-related tasks. For example, [13] adopts the audio classification task, and [16] adopts the ASR task. Then they adopt the trained network as feature extractors to extract high-dimensional features from enhanced speech and clean speech and calculate the distance between two features as the final loss. These high-dimensional features often contain abstract semantic information, making the deep feature-based losses more related to the speech quality. Existing work shows that deep feature-based losses can bring better training results. However, not every kind of deep feature works out for SE [15], and the feature extraction networks are usually large and deep, which increases the hardware resource consumption during training. As for metric-mimic losses, they try to mimic perceptual evaluation metrics with differentiable functions and train the SE model with these functions directly. Researchers in [17] modify the calculation process of perceptual evaluation of speech quality (PESQ) [22] to obtain an approximate but derivable calculation function and use it to guide the training of the SE model directly. Researchers in [18]–[21] use the NN to learn and imitate the computation of evaluation metrics and then take the NN as the loss function to guide the training of the SE model. However, these metric-mimic losses are not entirely consistent with actual metrics, limited by training data and unreasonable approximations.

Unlike the existing methods, we hope to get better loss by improving the correlation between the loss and the speech quality. STFT loss naturally has a relatively higher correlation with the speech quality than MAE loss as it calculates the loss in the time-frequency domain, which is more relevant to human hearing perception. However, it applies equal weights to the errors on all frequency components, ignoring that the human ear has different sensitivities to different frequency components. Taking this into account, we propose the perception-high-related time-frequency (PHRTF) loss based on STFT loss. We add a perceptual spectrum mask predictor (PSMP) in PHRTF loss, which can predict the perceptual spectrum mask (PSM) using the enhanced and clean speech spectrum. The perception-relevant loss components are emphasized with PSM, making PHRTF loss high relevant to

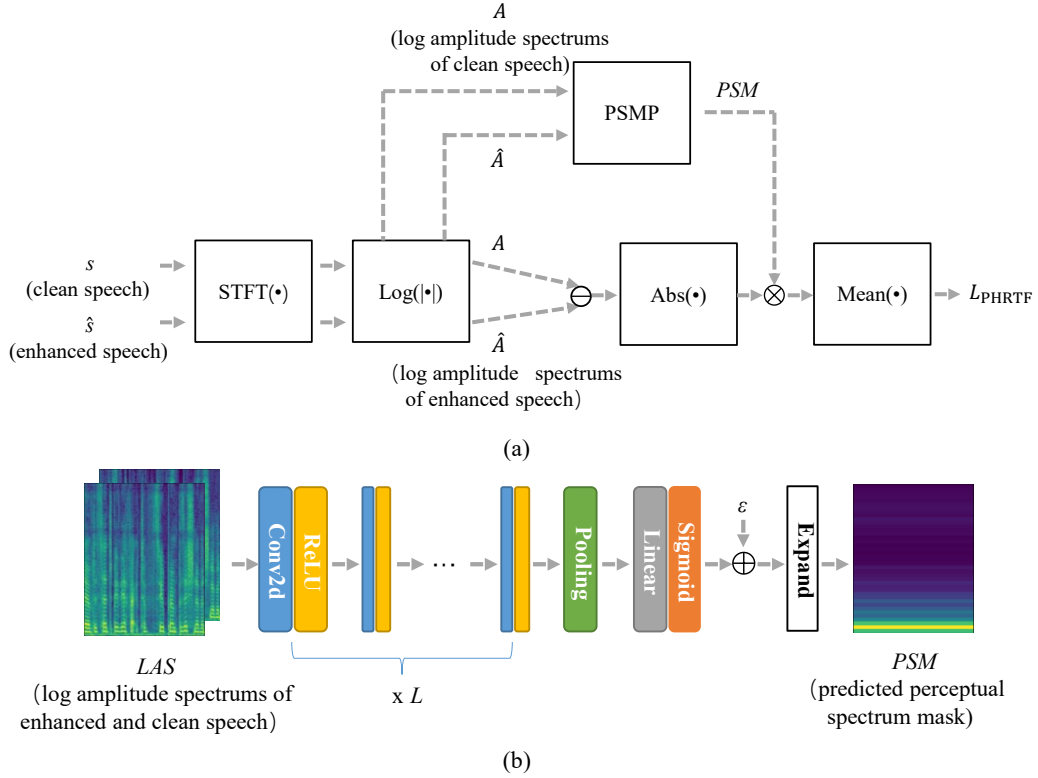


Figure 1: Description of the proposed PHRTF loss. (a) shows the calculation flow of PHRTF loss. (b) shows the detailed structure of the PSMP.

the speech quality. Compared with the existing losses, PHRTF loss is lightweight and more effective.

2. Method

The calculation flow of PHRTF loss is shown in Figure 1(a). Here s and \hat{s} represent clean and enhanced speech, respectively. A and \hat{A} represent corresponding log-wise amplitude spectrum (LAS). PHRTF loss first applies STFT transformation on the enhanced and clean speech and obtains the corresponding LASs. Then it subtracts A with \hat{A} and takes the absolute value to get the amplitude error spectrum (AES). Unlike STFT loss which directly averages the AES to get the final loss, the PHRTF loss feeds A and \hat{A} into PSMP to predict PSM. Then PSM is multiplied with the AES element by element to get the masked amplitude error spectrum (MAES), where perception-relevant loss components are emphasized. Finally, the MAES is averaged over frequency and time to obtain the final loss.

2.1. Perceptual Spectrum Mask Predictor

The detailed structure of PSMP is shown in Figure 1(b). The predictor consists of L cascaded convolution modules, one adaptive average pooling layer, and one linear layer followed by Sigmoid activation. Every convolution module is composed of one 2-D CNN and ReLU activation. For all CNNs, the convolution stride is set to S , and the kernel size is set to K . Given the input LASs, PSMP first utilizes CNNs to extract the perception-relevant feature. Then adaptive average pooling is applied to every channel to merge the information within the channel. After that, the linear layer is applied to merge the information among channels and produce the

perceptual mask vector with the dimension of d_{PM} . Sigmoid activation is used to limit the mask value ranging from 0 to 1. Here we add a small constant ϵ to the mask to avoid extremely small values. Finally, the vector is duplicated in time and interpolated in frequency to obtain PSM. To make PSMP be a smooth function and insensitive to small perturbations in the input, we apply spectral normalization as in [20] to PSMP, constraining it to be 1-Lipschitz continuous.

2.2. Training of PHRTF Loss

The training goal for PHRTF loss is to improve its correlation to the perceptual quality of speech. Here we use the PESQ score to represent the perceptual quality of speech and adopt the Pearson correlation coefficient (PCC) to measure relevance. Since the calculation of the PCC is derivable, we use it as the loss function here directly, which is represented by L_{PCC} . The specific calculation process is shown in formula (1)-(4), where N represents the number of batch samples, L_{PHRTF}^n represents the PHRTF loss of the n th sample, and PESQ^n represents the PESQ value of the n th sample. In (4), we add a minimal constant δ to the denominator to prevent division by zero.

$$\bar{L} = \frac{1}{N} \sum_{n=1}^N L_{\text{PHRTF}}^n, \bar{S} = \frac{1}{N} \sum_{n=1}^N \text{PESQ}^n \quad (1)$$

$$\sigma_L = \sqrt{\frac{1}{N-1} \sum_{n=1}^N (L_{\text{PHRTF}}^n - \bar{L})^2} \quad (2)$$

$$\sigma_S = \sqrt{\frac{1}{N-1} \sum_{n=1}^N (\text{PESQ}^n - \bar{S})^2} \quad (3)$$

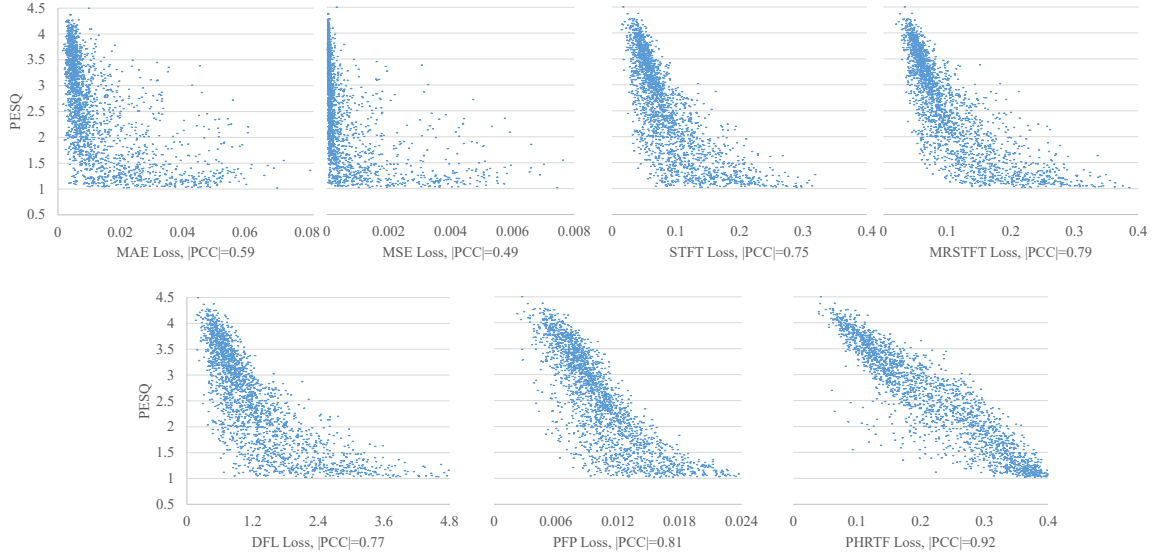


Figure 2: Illustration of the correlations of different losses with PESQ. Every point in the figure represents a sample from the VBD test dataset.

$$L_{PCC} = \frac{1}{N-1} \sum_{n=1}^N \frac{(L_{\text{PHRTF}}^n - \bar{L})(\text{PESQ}^n - \bar{S})}{\sigma_L \sigma_S + \delta} \quad (4)$$

3. Experiments

3.1. Dataset

We evaluate the proposed method on the widely used dataset VoiceBank-DEMAND (VBD) [23], which is created by the speech dataset voicebank[24] and noise dataset DEMAND [25]. A total of thirty speakers are selected, of which twenty-eight are used to build the training dataset, and the remaining two are used to build the test dataset. There are ten types of noise in the training dataset, of which eight are selected from the DEMAND, and the left are self-made noises. The test dataset contains five other types of noises selected from DEMAND. The signal-to-noise ratio (SNR) in training dataset includes 15dB, 10dB, 5dB and 0dB, and the SNR in test dataset includes 17.5db, 12.5db, 7.5db and 2.5dB. The test dataset is speaker and noise independent from the training dataset. Before training and testing, all samples were downsampled to 16KHz.

To train the PHRTF loss, we use the data in the VBD to construct training samples and test samples without introducing additional datasets. Noisy speech and clean speech are required as inputs, and the corresponding PESQ score is needed to calculate L_{PCC} . To construct training samples, we decompose all noisy and clean speech in the training dataset of VBD into segments with the length of 32768 and calculate the corresponding PESQ scores. We also enhanced the noisy speech with a previously trained speech enhancement model[26] to augment the training set. The test samples are constructed in the same way as the training samples with the test dataset of VBD.

3.2. Evaluation Metrics

We utilize four widely used objective metrics to evaluate the enhanced speech, namely PESQ, CSIG [27], CBAK [27], and

COVL [27]. PESQ stands for perceptual evaluation of speech quality and ranges from -0.5 to 4.5. CSIG predicts the signal distortion mean opinion score (MOS), CBAK measures background intrusiveness, and COVL measures speech quality. CSIG, CBAK, and COVL are ranged from 1 to 5. Each metric is computed by comparing the enhanced speech with the corresponding clean speech, and the higher value means better SE performance.

3.3. Experimental Setup

For the proposed PHRTF loss, we set $L=5$, $K=(5,5)$, $S=(2,2)$, $d_{PM}=40$, $\varepsilon=0.1$, $\delta=1e-8$. The output channels of CNNs in PSMP are set to 36, 72, 144, 288, 288 in order. We test PHRTF loss on the classical SE model Wave-U-Net [3] and a more advanced SE model SASP [26]. For Wave-U-Net, we set the number of layers to 12 and the number of extra filters per layer to 32. For SASP, we use the same model setting as the original paper [26].

When training SE models, we set the length of audio clips to 32768. Random crop is applied for audios longer than 32768, and zero padding is applied for audios shorter than 32768.

Batch size is set to 512 for the training of PHRTF loss, and 144 for the others. We use the Adam optimizer [28] with a learning rate (LR) of $1e-4$. When there is no performance improvement for P consecutive epochs, we reduce the LR by 20%. P is set to 8 for the training of PHRTF loss and 25 for the others.

We test six other losses on Wave-U-Net, namely MAE loss, mean square error (MSE) loss, STFT loss, multi-resolution STFT (MRSTFT) loss [29], DFL loss [13], and PFP loss [14], as comparisons. We set FFT size=512, hop size=256, window size=512 for STFT loss and PHRTF loss. For MRSTFT loss, we set FFT sizes={512,1024,2048}, hop sizes={50,120,240}, window sizes={240,600,1200}. VGG16 is used as the feature extractor for DFL loss, and L1 distance is used in PFP loss. When combining MAE loss with other losses, we balance the two losses numerically by multiplying

Table 1: Comparison of performances of the proposed PHRTF loss with other losses in terms of PESQ, CSIG, CBAK, COVL on Wave-U-Net, based on the VBD dataset.

Loss	PESQ	CSIG	CBAK	COVL	Extral Data
Noisy	1.97	3.35	2.44	2.63	\
L_{MAE}	2.48	3.73	3.25	3.10	No
L_{MSE}	2.50	3.70	3.23	3.09	No
$L_{MAE} \& L_{STFT}$	2.61	4.01	3.27	3.31	No
$L_{MAE} \& L_{MRSTFT}$	2.60	4.01	3.25	3.30	No
$L_{MAE} \& L_{DFL}$	2.60	4.01	3.21	3.31	Yes
$L_{MAE} \& L_{PPF}$	2.68	3.99	3.34	3.33	Yes
$L_{MAE} \& L_{PHRTF}$	2.80	3.98	3.39	3.39	No

the other loss with the constant factor a . We set a to 1 for PFP loss and 0.1 for the others.

For the training of SASP, we replace the originally used STFT loss with PHRTF loss and modify the loss of noise decode as (5), where \hat{n} represents the predicted noise, x represents the noisy speech. The other training settings are consistent with the original paper [26].

$$Loss_n = L_{MAE}(x - \hat{n}, s) + 0.1 \times L_{PHRTF}(x - \hat{n}, s) \quad (5)$$

3.4. Experimental Results

Table 2: Comparison of performances of the proposed method with other competitive baselines based on the VBD dataset. PHRTF represents the SASP model trained with PHRTF loss.

Methods	PESQ	CSIG	CBAK	COVL
Noisy	1.97	3.35	2.44	2.63
Wiener [30]	2.22	3.23	2.68	2.67
TSTNN [7]	2.96	4.33	3.53	3.67
T-GSA [8]	3.06	4.18	3.59	3.62
SA-TCN [31]	3.02	4.29	3.50	3.67
DEMUCS [5]	3.07	4.31	3.40	3.63
SASP [26]	3.07	4.35	3.59	3.73
DFL [13]	-	3.86	3.33	3.22
MetricGAN [18]	2.86	3.99	3.18	3.42
HiFi-GAN [21]	2.94	4.07	3.07	3.49
SDR-PESQ [17]	3.01	4.09	3.54	3.55
PPF [14]	3.15	4.18	3.60	3.67
MetricGAN+ [19]	3.15	4.14	3.16	3.64
PHRTF	3.16	4.35	3.62	3.78

3.4.1. Training result of PHRTF loss

After training, the PCC of PHRTF loss and PESQ reached 0.92, which indicates that PHRTF loss is highly correlated with the speech quality. Figure 2 shows the correlations of different losses with PESQ. We can see that MAE loss and MSE loss get the lowest correlation with PESQ. STFT loss, MRSTFT loss, and DFL loss are slightly better with close PCC. PFP loss achieves the second highest correlation with PESQ, but it is still much lower than PHRTF loss. The high PCC of PHRTF loss indicates the effectiveness of PSMP and proves that with proper masking, STFT loss can achieve a high correlation with PESQ.

3.4.2. Training results on Wave-U-Net

The training results of Wave-U-Net under different losses are shown in Table 1. We can see that PHRTF loss achieves the highest score in terms of PESQ, CBAK, and COVL¹. Especially, PHRTF loss improves PESQ by 0.32 over MAE loss and 0.19 over STFT loss. Meanwhile, we notice that from the perspective of PESQ, the SE training results of each loss are closely related to its correlation with PESQ. The experiment result proves the effectiveness of PHRTF loss and indicates that the loss with a higher correlation to PESQ tends to obtain better SE training results under PESQ.

3.4.3. Training results on SASP

To further verify the effectiveness of PHRTF loss, we apply it to the more advanced SE model SASP and compare it with other state-of-the-art methods. As shown in Table 2, after using PHRTF loss, SASP achieves better performance and outperforms all other competitive methods in terms of PESQ, CSIG, CBAK, and COVL.

4. Conclusions

In this paper, we proposed a new SE training loss named PHRTF loss, which modifies STFT loss by adding the PSMP module. PHRTF loss utilizes PSM predicted by PSMP to emphasize perception-relevant loss components to obtain a high correlation with the speech quality. Experimental results on the VBD dataset strongly demonstrate the effectiveness of PHRTF loss. The success of PHRTF loss shows that it is feasible to obtain better SE model training results by improving the correlation between loss and the speech quality. In future work, we will conduct further experiments to determine the optimal values of d_{PM} and ϵ . We will also test PHRTF loss on more models and larger datasets.

¹ Audio samples of the experimental results are shown in https://kenzd.github.io/phrtf_demo.github.io

5. References

- [1] T. Yoshioka *et al.*, “The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec. 2015, pp. 436–443. doi: 10.1109/ASRU.2015.7404828.
- [2] C. Karadagur Ananda Reddy, N. Shankar, G. Shreedhar Bhat, R. Charan, and I. Panahi, “An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device,” *IEEE Signal Processing Letters*, vol. 24, no. 11, pp. 1601–1605, Nov. 2017, doi: 10.1109/LSP.2017.2750979.
- [3] C. Macartney and T. Weyde, “Improved Speech Enhancement with the Wave-U-Net,” *arXiv:1811.11307 [cs, eess]*, Nov. 2018, Accessed: Oct. 23, 2021.
- [4] T.-A. Hsieh, H.-M. Wang, X. Lu, and Y. Tsao, “WaveCRN: An Efficient Convolutional Recurrent Neural Network for End-to-end Speech Enhancement,” *IEEE Signal Process. Lett.*, vol. 27, pp. 2149–2153, 2020, doi: 10.1109/LSP.2020.3040693.
- [5] A. Défossez, G. Synnaeve, and Y. Adi, “Real Time Speech Enhancement in the Waveform Domain,” in *Interspeech 2020*, Oct. 2020, pp. 3291–3295. doi: 10.21437/Interspeech.2020-2409.
- [6] L. Zhang and M. Wang, “Multi-Scale TCN: Exploring Better Temporal DNN Model for Causal Speech Enhancement,” in *Interspeech 2020*, Oct. 2020, pp. 2672–2676. doi: 10.21437/Interspeech.2020-1104.
- [7] K. Wang, B. He, and W.-P. Zhu, “TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 7098–7102. doi: 10.1109/ICASSP39728.2021.9413740.
- [8] J. Kim, M. El-Khary, and J. Lee, “T-GSA: Transformer with Gaussian-Weighted Self-Attention for Speech Enhancement,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, May 2020, pp. 6649–6653. doi: 10.1109/ICASSP40776.2020.9053591.
- [9] E. Kim and H. Seo, “SE-Conformer: Time-Domain Speech Enhancement Using Conformer,” in *Interspeech 2021*, Aug. 2021, pp. 2736–2740. doi: 10.21437/Interspeech.2021-2207.
- [10] 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,” p. 20, 2019.
- [11] Y. Hu *et al.*, “DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement,” *arXiv:2008.00264 [cs, eess]*, Sep. 2020, Accessed: Nov. 10, 2021.
- [12] S. Lv, Y. Hu, S. Zhang, and L. Xie, “DCCRN+: Channel-wise Subband DCCRN with SNR Estimation for Speech Enhancement,” *arXiv:2106.08672 [cs, eess]*, Jun. 2021, Accessed: Nov. 12, 2021.
- [13] F. G. Germain, Q. Chen, and V. Koltun, “Speech Denoising with Deep Feature Losses,” *arXiv:1806.10522 [cs, eess]*, Sep. 2018, Accessed: Nov. 12, 2021.
- [14] T.-A. Hsieh, C. Yu, S.-W. Fu, X. Lu, and Y. Tsao, “Improving Perceptual Quality by Phone-Fortified Perceptual Loss using Wasserstein Distance for Speech Enhancement,” *arXiv:2010.15174 [cs, eess]*, Apr. 2021, Accessed: Sep. 25, 2021.
- [15] S. Kataria, J. Villalba, and N. Dehak, “Perceptual Loss based Speech Denoising with an ensemble of Audio Pattern Recognition and Self-Supervised Models,” *arXiv:2010.11860 [cs, eess]*, Oct. 2020, Accessed: Nov. 17, 2021.
- [16] P. Plantinga, D. Bagchi, and E. Fosler-Lussier, “Perceptual Loss with Recognition Model for Single-Channel Enhancement and Robust ASR,” *arXiv:2112.06068 [cs, eess]*, Dec. 2021, Accessed: Mar. 08, 2022.
- [17] J. Kim, M. El-Kharmy, and J. Lee, “End-to-End Multi-Task Denoising for joint SDR and PESQ Optimization,” *arXiv:1901.09146 [cs, eess, stat]*, Jan. 2019, Accessed: Sep. 25, 2021.
- [18] 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,” *arXiv:1905.04874 [cs, eess]*, May 2019, Accessed: Sep. 25, 2021.
- [19] S.-W. Fu *et al.*, “MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement,” in *Interspeech 2021*, Aug. 2021, pp. 201–205. doi: 10.21437/Interspeech.2021-599.
- [20] 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, 2020, doi: 10.1109/LSP.2019.2953810.
- [21] J. Su, Z. Jin, and A. Finkelstein, “HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks,” *arXiv:2006.05694 [cs, eess]*, Sep. 2020, Accessed: Nov. 12, 2021.
- [22] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, May 2001, vol. 2, pp. 749–752 vol.2. doi: 10.1109/ICASSP.2001.941023.
- [23] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, “Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech,” in *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, Sep. 2016, pp. 146–152. doi: 10.21437/SSW.2016-24.
- [24] C. Veaux, J. Yamagishi, and S. King, “The voice bank corpus: Design, collection and data analysis of a large regional accent speech database,” in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4. doi: 10.1109/ICSDA.2013.6709856.
- [25] J. Thiemann, N. Ito, and E. Vincent, “The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings,” *Proc. Mtgs. Acoust.*, vol. 19, no. 1, p. 035081, 2013, doi: 10.1121/1.4799597.
- [26] D. Zhao, Z. Zhang, B. Yu, and Y. Wang, “Time Domain Speech Enhancement using Self-Attention-Based Subspace Projection,” in *2021 7th International Conference on Computer and Communications (ICCC)*, 2021, pp. 530–534. doi: 10.1109/ICCC54389.2021.9674447.
- [27] Y. Hu and P. C. Loizou, “Evaluation of Objective Quality Measures for Speech Enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008, doi: 10.1109/TASL.2007.911054.
- [28] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, Accessed: Mar. 11, 2022.
- [29] R. Yamamoto, E. Song, and J.-M. Kim, “Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203. doi: 10.1109/ICASSP40776.2020.9053795.
- [30] J. Lim and A. Oppenheim, “All-pole modeling of degraded speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, Jun. 1978, doi: 10.1109/TASSP.1978.1163086.
- [31] J. Lin, A. J. van Wijngaarden, K.-C. Wang, and M. C. Smith, “Speech Enhancement Using Multi-Stage Self-Attentive Temporal Convolutional Networks,” *arXiv:2102.12078 [cs, eess]*, Feb. 2021, Accessed: Oct. 26, 2021.