Interspeech 2022
18-22 September 2022, Incheon, Korea

PercepNet+: A Phase and SNR Aware PercepNet for Real-Time Speech Enhancement

Xiaofeng Ge[1], *Jiangyu Han*[1], *Yanhua Long*[1,2*], *Haixin Guan*[2]

[1]Shanghai Normal University, Shanghai, China
[2]Unisound AI Technology Co., Ltd., Beijing, China
{xfge01,jyhan03}@163.com, yanhua@shnu.edu.cn, guanhaixin@unisound.com

Abstract

PercepNet, a recent extension of the RNNoise, an efficient, high-quality and real-time full-band speech enhancement technique, has shown promising performance in various public deep noise suppression tasks. This paper proposes a new approach, named PercepNet+, to further extend the PercepNet with four significant improvements. First, we introduce a phase-aware structure to leverage the phase information into PercepNet, by adding the complex features and complex subband gains as the deep network input and output respectively. Then, a signal-to-noise ratio (SNR) estimator and an SNR-switched post-processing are specially designed to alleviate the over attenuation (OA) that appears in high SNR conditions of the original PercepNet. Moreover, the GRU layer is replaced by TF-GRU to model both temporal and frequency dependencies. Finally, we propose to integrate the loss of complex subband gain, SNR, pitch filtering strength, and an OA loss in a multi-objective learning manner to further improve the speech enhancement performance. Experimental results show that, the proposed PercepNet+ outperforms the original PercepNet significantly in terms of both PESQ and STOI, without increasing the model size too much.

Index Terms: speech enhancement, phase-aware structure, SNR-switched post-processing, multi-objective learning

1. Introduction

Speech enhancement (SE) aims to improve speech perceptually quality and intelligibility under noisy condition.. Recently, deep learning based SE approaches [1, 2] have shown performance over most traditional methods, such as log-spectral amplitude estimation [3], spectral subtraction [4], etc. In many scenarios, such as telecommunication and online conference, SE systems are required to meet both good denoising performance and real-time constrains. For real-time SE, the current mainstream methods can be divided into two categories. One is end-to-end systems based on U-Net structure [5, 6], such as DCCRN [7], DCCRN+ [8], and DPCRN [9], etc. The other is perceptually-motivated, hybrid signal processing/deep learning approaches, such as RNNoise [10], and its extensions like PercepNet [11], Personalized PercepNet [12], etc. Our work focuses on improving PercepNet due to its excellent ability of improving speech perceptual quality and noise suppression.

PerceptNet [11] aims to enhance fullband (48 kHz sampled) noisy speech with low-complexity, and has been shown to deliver high-quality speech enhancement in real-time even when operating on less than 5% of a CPU core (1.8 GHz Intel i7-8565U CPU). Instead of performing on the Fourier transform bins in state-of-the-art end-to-end SE methods, PerceptNet features the speech short-time Fourier transform (STFT) spectrum from 0 to 20 kHz with only 34 bands, according to the human hearing equivalent rectangular bandwidth (ERB) scale [13], which greatly lowers the system computational complexity. Together with the design of pitch filter and envelope postfiltering, PercepNet can produce high quality enhanced speech.

However, we find that, compared with the enhanced noisy speech with a low SNR, PercepNet results in much heavier over attenuation (OA) when the input noisy speech with a relative high SNR. It significantly impairs the perceptual quality of enhanced speech in high SNR condition (even worse than the original noisy speech). This heavier quality impairing may be due to the inaccurate estimation of the frequency band gains, and the further speech enhancing by envelope postfiltering to remove residual noise, since the high SNR noisy speech is actually a clean speech from human perception. Furthermore, during the PercepNet pipeline, only the speech spectral envelope is enhanced, the phase of noisy speech is directly used to reconstruct the target clean speech. All of these mentioned issues might limit the PercepNet performance.

To develop a real-time SE system with better and more robust performance,in this study, we focus on improving the PercepNet to further enhance its speech denoising ability and achieve better speech perceptual quality. Four main contributions are: 1) A phase-aware structure is introduced to leverage the phase information, by adding the complex subband features as additional deep network input, and replacing the original energy gains with subband real and imaginary part gains for target clean speech construction; 2) To handle the over attenuation problem and alleviate perceptual quality impairing of enhanced high SNR noisy speech, we design an SNR estimator and an SNR-switched post-processing to control the degree of residual noise removal; 3) We replace the first two GRU [14] layers in PercepNet with TF-GRU structure to well learn both the time-scale temporal and frequency dependencies; 4) Based on the above revisions, we finally propose to learn the complex gains, SNR, the original pitch filtering strength, together with an OA loss in a multi-objective training manner to further improve the SE performance. Compared with PercepNet, our proposed PercepNet+ achieves absolute 0.19 PESQ [15] and 2.25% STOI [16] gains on the public VCTK [17] test set, and 0.15 PESQ and 2.93% STOI gains on our simulated test set.

2. PercepNet

PercepNet is a perceptually-motivated approach for low-complexity, real-time enhancement of full-band speech [11]. It extracts various hand-crafted ERB-based subband acoustic features from 34 triangular spectral bands as the model input. The model outputs energy gain, $\hat{g}_b$ is then multiplied with the pitch

Yanhua Long is the corresponding author. The work is supported by the National Natural Science Foundation of China (Grant No.62071302).

Copyright © 2022 ISCA

10.21437/Interspeech.2022-43

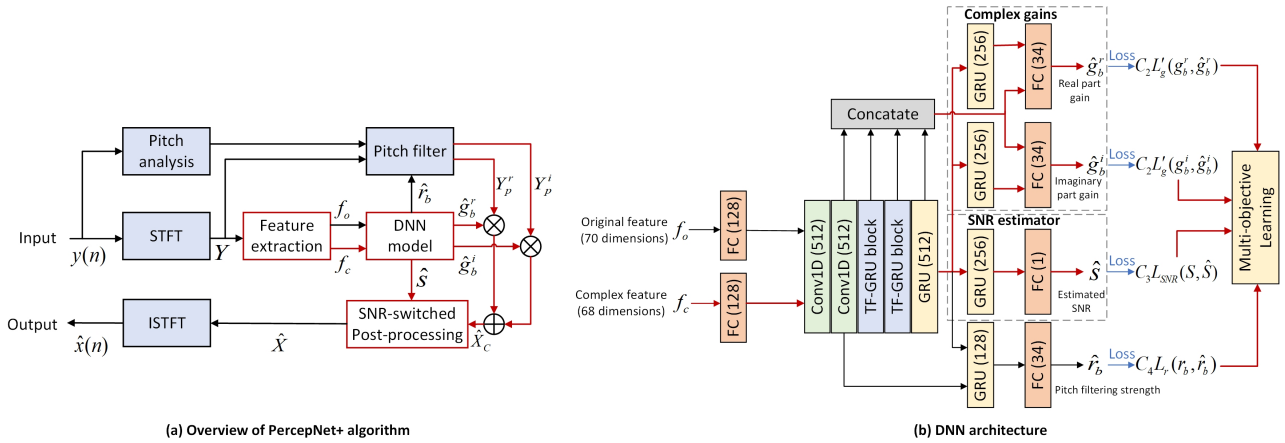

Figure 1: *The framework of proposed PercepNet+ algorithm. All dark red blocks and lines are our improvements over the original PercepNet [11]. (a) Overview of PercepNet+; (b) DNN model architecture of PercepNet+.*

filtered spectrum of noisy speech to remove the background noise, where the pitch filter is a comb filter that designed to remove the noise between pitch harmonics [18]. The effects of pitch filter in each ERB band is controlled by a pitch filter strength $\hat{r_b}$. Both $\hat{g_b}$ and $\hat{r_b}$ are automatically learnt by a deep neural network (DNN), which is mainly composed of two convolution layers and five GRU layers. The DNN model utilizes the features of current and three extra future frames to compute its outputs, which makes the PercepNet achieve 30ms look-ahead. With the help of envelope post-filtering, the denoised speech is further enhanced. More details can be found in [11].

3. Proposed PercepNet+

As mentioned in Section 1, we extend PercepNet to PercepNet+ in four new aspects: the phase-aware structure, SNR estimator and SNR-switched post-processing, the multi-objective loss function and updated TF-GRU blocks. Fig.1 illustrates the whole framework of PercepNet+, and all the dark red blocks and lines are our improvements over original PercepNet [11].

3.1. Phase-aware Structure

Because the DNN input features of PercepNet are tied to 34 ERB bands, as shown in Fig.1, the original 70-dimensional acoustic features f_o, are composed of 68-dimensional band-related features (34 for spectrum energy, 34 for pitch coherence), a pitch period [19] and a pitch correlation [20]. For each band b, the DNN model outputs two elements: the energy gain $\hat{g_b}$ and pitch filter strength $\hat{r_b}$. These features only focus on enhancing the noisy spectral envelop and pitch harmonics, the importance of phase information [21] that can affect the human perception significantly is ignored.

To exploit the phase information in PercepNet+, we concatenate the real and imaginary part of complex STFT for noisy speech $y(n)$ in each ERB band directly, to form a total 68-dimensional complex feature f_c. Then, as in Fig.1(b), the linear transformed (FC layer) f_o and f_c are finally concatenated to train the improved DNN model. Besides adding the complex features, we also replace the original energy gain with "complex gains" to pay more attention on phase as in Fig.1(b). Specifically, we propose to guide the network to learn real and imaginary part gains, g_b^r and g_b^i, for re-constructing both target clean speech magnitude and phase spectrum, and define them as:

$$g_b^r(t) = \frac{\|X_b^r(t)\|_2}{\|Y_b^r(t)\|_2}, \quad g_b^i(t) = \frac{\|X_b^i(t)\|_2}{\|Y_b^i(t)\|_2} \tag{1}$$

where $X_b(t)$ and $Y_b(t)$ are the complex-valued spectrum of the clean signal $x(n)$ and its noisy signal $y(n)$ for ERB band b in frame t, and $\|\cdot\|_2$ means the L_2-norm operation.

3.2. SNR Estimator and SNR-switched Post-processing

Speech distortion is easy to occur during the process of removing noise, and it may seriously impair the speech perceptual quality [22]. In PercepNet, this distortion may be due to the inaccurate estimation of energy gain and the inappropriate design of envelope post-filtering. In PercepNet+, we propose an SNR estimator and design an SNR-switched post-processing to alleviate the speech distortion in PercepNet.

SNR estimator: this estimator is inspired by works in [8, 23]. As shown in Fig.1(b), it is composed of one GRU and one fully-connected (FC) layer with sigmoid activation function, and predicts frame-level SNR under a multi-objective learning framework to maintain good speech quality. The normalized ground-truth SNR $S(t) \in [0, 1]$ for frame t of $y(n)$ is defined as:

$$\begin{aligned} S(t) &= \frac{Q(t) - \mu}{\sigma} \quad with \\ Q(t) &= 20\log_{10}(X_m(t)/N_m(t)) \end{aligned} \tag{2}$$

where μ and σ are the mean and standard deviation of SNR $Q(t)$ for the whole noisy speech, $X_m(t)$ and $N_m(t)$ represent the magnitude spectrum of clean speech and noise respectively.

SNR-switched MMSE-LSA post-processing: Although post-processing module is proved very effective in removing residual noise [24, 25], it may impair the perceptual quality of test samples with almost no noise as we found in our experiments. Therefore, in our PercepNet+, as shown in Fig.1(a), the predicted SNR $\hat{S}$ of each frame is used to control whether the post-processing module should be performed. We call this strategy as SNR-switched post-processing. If $\hat{S}$ is greater than a predefined threshold, the spectrum $\hat{X}_c$ enhanced by $\hat{g}_b^r$ and $\hat{g}_b^i$ will directly be the final output. Otherwise, $\hat{X}_c$ will be further enhanced by post-processing to remove residual noise.

In addition, we find that, the conventional MMSE-LSA [3] based post-processing has achieved remarkable results in recent end-to-end SE systems [8, 26]. Therefore, in PercepNet+, we also replace the original envelop postfiltering with MMSE-LSA in our SNR-switched post-processing module as follows:

$$G(t) = \text{MMSE-LSA}(\xi(t), \gamma(t)) \tag{3}$$

$$\hat{X}(t) = G(t) * \hat{X}_c(t) \tag{4}$$

where $G(t)$ is the MMSE-LSA frame-level gain, $\hat{X}_c(t)$ is the spectrum enhanced by complex gains of frame t, $\hat{X}(t)$ is the final enhanced clean speech as in Fig.1(a), and $\xi(t)$, $\gamma(t)$ are priori and posterior frame-level SNR that defined as in [3].

3.3. Multi-objective Loss Function

The original loss function L_P of DNN model in PercepNet has two parts: the loss of energy gain L_g and pitch filter strength L_r as defined:

$$\begin{aligned} L_g &= \sum_b (g_b^\lambda - \hat{g}_b^\lambda)^2 + C_1 \sum_b (g_b^\lambda - \hat{g}_b^\lambda)^4 \\ L_r &= \sum_b ((1 - r_b)^\lambda + (1 - \hat{r}_b)^\lambda)^2 \\ L_P &= \alpha L_g + \beta L_r \end{aligned} \tag{5}$$

where $g_b, \hat{g}_b, r_b, \hat{r_b}$ are the ground-truth and DNN predicted ERB-based band energy gain and pitch filter strength, respectively. $C_1, \lambda, \alpha, \beta$ are tuning parameters.

Besides the SNR-switched post-processing, results in [27] showed an asymmetric loss L_{OA} that proposed in [28] is effective in alleviating the over attenuation issue. Therefore, we adapt it to L_g to address the quality impairing in high SNR condition as follows:

$$h(x) = \begin{cases} 0, & if\ x \leq 0, \\ x, & if\ x > 0, \end{cases} \tag{6}$$

$$L_{OA}(g_b, \hat{g}_b) = |h(g_b - \hat{g}_b)|^2 \tag{7}$$

$$L'_g = \delta L_g + (1 - \delta) L_{OA}. \tag{8}$$

In PercepNet+, instead of using L_P, we use Eq.(8) to measure the difference between estimated $\hat{g}_b^r$, $\hat{g}_b^i$ and their ground-truth respectively. Together considering the original L_r, and mean-square-error (MSE) loss of SNR $L_{SNR}(S, \hat{S})$, the DNN model of PercepNet+ is finally joint trained using the following overall multi-objective loss function L_{P+}:

$$\begin{aligned} L_{P+} = C_2 L'_g(g_b^r, \hat{g}_b^r) + C_2 L'_g(g_b^i, \hat{g}_b^i) \\ + C_3 L_{SNR}(S, \hat{S}) + C_4 L_r(r_b, \hat{r}_b). \end{aligned} \tag{9}$$

where C_2, C_3, C_4 are tuning loss weight parameters.

3.4. TF-GRU Block

PercepNet models the temporal dependency at time-scale with GRU layers. Inspired by [29], we employ another GRU layer to model the frequency-wise evolution of spectral patterns. Specifically, as shown in Fig.1(b), we replace two GRU layers in PercepNet with two proposed TF-GRU blocks, each TF-GRU is composed of one time-GRU (TGRU) layer and one frequency-GRU (FGRU) layer. The FGRU is used to browse the frequency band features so that the frequency-wise dependency is summarized, and the TGRU is used to summarize time-wise dependency. The outputs of both TGRU and FGRU are then concatenated to form the final TF-GRU output.The parameter quantity of one TF-GRU is adjust to be consistent with one GRU layer in original PercepNet.

4. Experimental Setup

4.1. Datasets

The training data used in original PercepNet [11] is not publicly available, so we use the public dataset that used to train RNNoise model in [10] as our training set. The clean speech data is from McGill TSP speech database[1] and the NTT Multi-Lingual Speech Database for Telephonometry [30]. Various sources of noise are used[2], including computer fans, office, crowd, airplane, car, train, construction. Totally, we have 6 hours clean speech and 4 hours noise data, which is far less than the 120 hours speech plus 80 hours noise data used in PercepNet [11]. The training pairs are simulated by dynamically mixing noise and speech with a random SNR ranges from -5 to 20dB. Half of the speech are convoluted with room impulse response (RIR) coming from RIR_NOISES set [31].

Two evaluation sets are used to examine the proposed techniques, one is the public noisy VCTK test set [17] with 824 samples from 8 speakers. The other is a test set named D-NOISE, which is simulated by ourselves, with SNR ranging from -5 to 20dB. D-NOISE consists of 108 samples with the speech data from WSJ0 [32] dataset and noise data comes from RNNoise demo website, including the office, kitchen, cars, street and babble noises.

4.2. Configurations

All training and test data are 48kHz sampled. The frames are extracted using a Vorbis window [33] of size 20 ms and an overlap of 10 ms. Batch size is set to 32. Adam optimizer [34] is used, and the initial learning rate is set to 0.001. We set the loss function weights, $\lambda, \alpha, \beta, \delta$ to 0.5, 4.0, 1.0 and 0.7, respectively, and C_1, C_2, C_3, C_4 to 10, 4, 1 and 1 respectively. The DNN network layer parameters of PercepNet+ are presented in Fig.1(b). To make results comparable, all other configurations are the same as the original PercepNet in [11] and RNNoise in [10]. Both PESQ and STOI are used to measure the speech quality and intelligibility repecetively.

5. Results and discussion

5.1. Baseline

Table 1: *PESQ and STOI(%) on VCTK test set. The training data of model 2, 4 are 190 hours less than that of model 1,3.*

ID	Model	PESQ	STOI
-	Noisy	1.97	92.12
1	RNNoise [10]	2.29	-
2	RNNoise	2.23	92.74
3	PercepNet [11]	2.54	-
4	PercepNet	2.46	93.43

Both RNNoise (open-sourced) and its extension-PercepNet (not open-sourced) are taken as our baselines. Table 1 presents the comparison results on VCTK test set. Model 1 and 3 are the published results in [11], in which the models were trained on non-public 120 hours speech plus 80 hours noise data, while model 2 and 4 are our implemented RNNoise and PercepNet models that trained on only the 6 hours speech and 4 hours noise data. It's clear that, PercepNet significantly outperforms RNNoise, and the PESQ scores of our models are only slightly worse than the ones in [11], even there is an extremely big training data size gap (190 hours) between our models and model 1,3. Therefore, we think that our implementation of PercepNet is correct and can be taken as the baseline of our PercepNet+.

Moreover, a detailed baseline results analysis are presented in Fig.2 and 3, in which the whole VCTK test set is divided into 4-level SNR ranges to observe the PercepNet denoising

[1] http://www-mmsp.ece.mcgill.ca/Documents/Data/

[2] https://jmvalin.ca/demo/rnnoise/

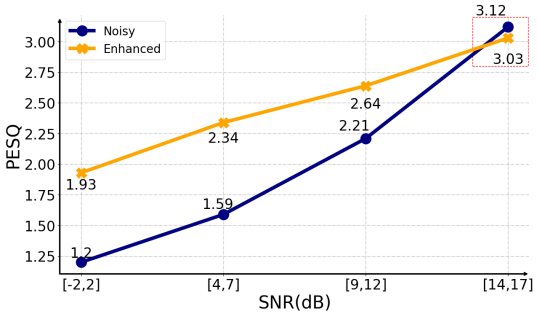

Figure 2: *PESQ of noisy and PercepNet enhanced VCTK test samples with different SNR ranges.*

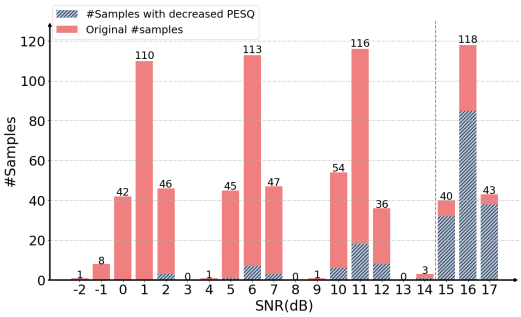

Figure 3: *Detailed samples distribution with decreased PESQ after enhancement by PercepNet in different SNR conditions of VCTK test set.*

performance and behaviors in different SNR conditions. From Fig.2, we find that the PESQ of samples with SNR>14dB is decreased after enhancement. Meanwhile, in Fig.3, comparing the histogram of red and its corresponding light blue parts, it's clear that most PESQ decrease occurs in higher SNR conditions, specifically, there are total 202 samples with decreased PESQ and 76.35% of them are with SNR greater than 14dB. It indicates that the original PercepNet has heavier OA or fails to perform well under high SNR conditions. Therefore, in this study, we take the 14dB as our proposed SNR-switched post-processing threshold.

5.2. Results of PercepNet+

Table 2 shows the performance comparison of all the proposed techniques in PercepNet+ on both VCTK and our simulated D-NOISE test sets. Compared with PercepNet, the proposed PercepNet+ significantly improves the PESQ from 2.46 to 2.65, and STOI from 93.43% to 95.68% on VCTK test set. Specifically, the additional complex features and gains lead to an absolute increase of 0.08 and 1.11% in terms of PESQ and STOI. With the help of SNR estimator, we obtain 0.04 PESQ and 0.27% STOI improvements. When the SNR-switched post-processing (PP), and over attenuation loss are applied, the PESQ and STOI reach to 2.62 and 95.49%. Finally, we see the updated TF-GRU results further performance gains. Moreover, on D-NOISE test set, we achieve consistent performance gains from all the proposed techniques of PercepNet+, an overall 0.15 PESQ and 2.93% STOI gains are obtained. In addition, the proposed PercepNet+ has 8.5M trainable parameters, an increase of 0.5M compared with PercepNet, and Real Time Factor (RTF) is equal to 0.351, which is tested on a machine with an Intel(R) Xeon(R) CPU E5-2650 v2@2.60GHz in single thread. Therefore, we can draw the conclusion that, PercepNet+ has greatly surpassed PercepNet without significantly increasing the parameters of neural network.

Table 2: *Various models' PESQ and STOI(%) on both VCTK and D-NOISE test set.*

Model	VCTK		D-NOISE	
	PESQ	STOI	PESQ	STOI
Noisy	1.97	92.12	2.10	86.53
RNNoise	2.23	92.74	2.33	88.27
PercepNet	2.46	93.43	2.57	90.53
+Complex Features	2.51	93.88	2.60	91.28
+Complex Gains	2.54	94.54	2.63	92.19
+SNR Estimator	2.58	94.87	2.65	92.41
+SNR-switched PP	2.61	95.40	2.69	93.11
+OA Loss	2.62	95.49	2.70	93.12
+TF-GRU (PercepNet+)	**2.65**	**95.68**	**2.72**	**93.46**

5.3. Performance of SNR-sensitive techniques

We further investigate the effectiveness of proposed OA loss and SNR-switched PP in solving the speech perceptual quality impairing issue after enhancement in high SNR condition. Results of both two VCTK sub test sets (">14dB" and "≤14dB") are shown in Table 3. Comparing the PESQs in first two lines, we see the PercepNet does impair the perceptual quality with high SNR, which is consistent as observed in Fig.2. However, in PercepNet+, we see this issue is effectively alleviated by either the proposed OA loss or the SNR-switched PP. When both two techniques are applied, the performance is further slightly improved, without impairing the speech perceptual quality in low SNR condition.

Table 3: *PESQ scores of PercepNet+ with (or without) over attenuation(OA) loss and SNR switch on VCTK sub test sets*

Model	≤14dB	>14dB	Overall
Noisy	1.59	3.12	1.97
PercepNet	2.28	3.03	2.46
PercepNet+	**2.49**	**3.14**	**2.65**
−SNR Switch	**2.49**	3.11	2.64
−OA Loss	2.48	3.13	2.64
−SNR Switch	2.48	3.03	2.62

6. Conclusions

In this paper, we propose PercepNet+, by extending the high quality, real-time and low-complexity PercepNet in several aspects to further improve speech enhancement, including the phase-aware structure to learn the phase information, two SNR-sensitive improvements to maintain speech perceptual quality while removing noise, the updated TF-GRU to model time and frequency scale dependencies simultaneously, and the multi-objective learning for further system performance improvements. Importantly, the heavy speech perceptual quality impairing by the original PercepNet in high SNR conditions has been well solved due to the proposed OA loss and SNR-switched post-processing. Experimental results show that the proposed PercepNet+ has significantly outperformed the original PercepNet in both PESQ and STOI. Some noisy and enhanced samples, including the D-NOISE test set can be found from `https://github.com/orcan369/PercepNet-plus-samples`.

7. References

[1] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," in *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, 2015, pp. 7–19.

[2] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," in *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, no. 12, 2014, pp. 1849–1858.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, 1985, pp. 443–445.

[4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, 1979, pp. 113–120.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[6] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes," in *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, 2018, pp. 2663–2674.

[7] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement," in *Proceedings of INTERSPEECH*, 2020, pp. 2472–2476.

[8] S. Lv, Y. Hu, S. Zhang, and L. Xie, "DCCRN+: Channel-Wise Subband DCCRN with SNR Estimation for Speech Enhancement," in *Proceedings of INTERSPEECH*, 2021, pp. 2816–2820.

[9] X. Le, H. Chen, K. Chen, and J. Lu, " DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement," in *Proceedings of INTERSPEECH*, 2021, pp. 2811–2815.

[10] J.-M. Valin, "A Hybrid DSP/Deep Learning Approach to Real-Time Full-Band Speech Enhancement," in *Proceedings of IEEE Multimedia Signal Processing (MMSP)*, 2018, pp. 1–5.

[11] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, K. Helwani, and A. Krishnaswamy, "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *Proceedings of INTERSPEECH*, 2020, pp. 2482–2486.

[12] R. Giri, S. Venkataramani, J.-M. Valin, U. Isik, and A. Krishnaswamy, "Personalized PercepNet: Real-Time, Low-Complexity Target Voice Separation and Enhancement," in *Proceedings of INTERSPEECH*, 2021, pp. 1124–1128.

[13] B. Moore, "An introduction to the psychology of hearing," *Brill*, 2012.

[14] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proceedings of Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014, pp. 103–111.

[15] I. Rec, "P.862.2: Wideband extension to recommendation p.862 for the assessment of wideband telephone networks and speech codecs," *International Telecommunication Union,CH–Geneva*, 2005.

[16] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," in *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 19, no. 7, 2011, pp. 2125–2136.

[17] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech," in *Proceedings of ISCA Speech Synthesis Workshop (SSW)*, 2016, pp. 146–152.

[18] J. H. Chen, Gersho, and A., "Adaptive postfiltering for quality enhancement of coded speech," in *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing*, vol. 3, no. 1, 1995, pp. 59–71.

[19] D. Talkin., "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, 1995, pp. 495–518.

[20] K. Vos, K. V. Sorensen, S. S. Jensen, and J.-M. Valin., "Voice coding with opus," in *Proceedings of AES Convention*, 2013.

[21] K. K. Paliwal, K. K. Wójcicki, and B. J. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, 2011.

[22] C. Zheng, X. Peng, Y. Zhang, S. Srinivasan, and Y. Lu, "Interactive speech and noise modeling for speech enhancement," in *AAAI*, 2021, pp. 14 549–14 557.

[23] A. Nicolson and K. K. Paliwal, "Masked multi-head self-attention for causal speech enhancement," *Speech Communication*, vol. 125, no. 3, pp. 80–96, 2020.

[24] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "ICASSP 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6628–6632.

[25] Y.-H. Tu, J. Du, L. Sun, and C.-H. Lee, "Lstm-based iterative mask estimation and post-processing for multi-channel speech enhancement," in *Proceedings of Asia-Pacific Signal and Information Processing Association (APSIPA)*, 2017, pp. 488–491.

[26] A. Li, W. Liu, X. Luo, G. Yu, C. Zheng, and X. Li, "A Simultaneous Denoising and Dereverberation Framework with Target Decoupling," in *Proceedings of INTERSPEECH*, 2021, pp. 2801–2805.

[27] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Personalized speech enhancement: New models and comprehensive evaluation," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 356–360.

[28] Q. Wang, I. L. Moreno, M. Saglam, K. Wilson, A. Chiao, R. Liu, Y. He, W. Li, J. Pelecanos, M. Nika, and A. Gruenstein, "VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition," in *Proceedings of INTERSPEECH*, 2020, pp. 2677–2681.

[29] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "Lstm time and frequency recurrence for automatic speech recognition," in *Proceedings of Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 187–191.

[30] "Multi-lingual speech database for telephonometry," *NTT Advanced Technology Corporation (NTT-AT)*, 1996.

[31] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer, and S. Khudanpur, "A study on data augmentation of reverberant speech for robust speech recognition," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5220–5224.

[32] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of Second International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 357–362.

[33] X. O. Foundation, "Vorbis I specification," 2004.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2015.