 DOI:10.30420/456164035

# Low-complexity Real-time Single-channel Speech Enhancement Based on Skip-GRUs

*Ragini Sinha*[*], *Christian Rollwage*[*], *Simon Doclo*[*,†]

[*] Fraunhofer Institute for Digital Media Technology IDMT, Oldenburg Branch for Hearing, Speech and Audio Technology HSA, Germany
[†] Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
Email:ragini.sinha@idmt.fraunhofer.de, christian.rollwage@idmt.fraunhofer.de, simon.doclo@uni-oldenburg.de

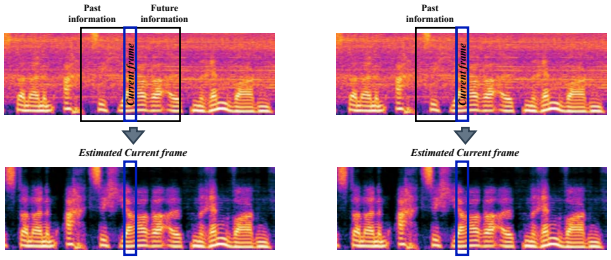
## Abstract

Recently, algorithms based on deep neural networks have led to a significant speech enhancement performance improvement in terms of speech quality and intelligibility both for offline as well as online processing. However, obtaining a low-complexity and resource-efficient system is still a challenge. In this paper, we focus on real-time single-channel speech enhancement systems that are both compact and resource-efficient during inference. We propose two systems, either applying a real-valued or a complex-valued mask. Both systems are based on the Skip-GRU architecture, which employs a skip connection between the GRU layers. Experimental results on reverberant noisy signals demonstrate significant advantages of using the Skip-GRU architecture vs. the GRU architecture and applying a complex-valued mask vs. a real-valued mask. Moreover, the proposed Skip-GRU system with complex-valued masking achieves a similar speech enhancement performance as the best-performing baseline system but with a significantly reduced number of parameters and computational complexity.

## 1 Introduction

Speech enhancement algorithms aim at improving the perceived quality and intelligibility of noisy speech signals. Recently, the use of deep neural networks (DNNs) has enabled significant progress in speech enhancement, both for offline processing [1–7] as well as for online processing [8–15]. Offline processing does not consider the causality in the processing chain and exploits both past as well as future information of the input signal for speech enhancement (see Figure 1(a)), whereas online processing maintains the causality in the processing chain and does not consider future information (see Figure 1(b)). Research on online speech enhancement mainly focuses on further improving the performance, often without considering the model size and overall computational complexity. However, running such systems on resource-limited mobile devices may be very challenging. In this paper, we focus on single-channel real-time speech enhancement systems that are resource-efficient and have low computational complexity.

Various architectures have been explored for single-channel real-time speech enhancement, either performing enhancement in the time-domain [8, 9] or in the spectral domain [10–14]. The systems in [8, 9] perform speech enhancement in the time-domain using an encoder-decoder architecture, either with a temporal convolutional neural (TCN) network or a long short-term memory (LSTM) network as the bottleneck. The systems in [10–12] perform speech enhancement in the spectral domain using a real-valued mask, where the mask is estimated using LSTM-based architectures. Instead of using real-valued masks, the systems in [13, 14] use a complex-valued mask, estimated using either an LSTM or a TCN-based architecture. Because of their temporal modeling capabilities and low computational complexity during inference, recurrent neural networks (RNNs) are a prevalent choice to perform real-time speech enhancement [10–13], especially using spectral features [16]. Therefore, in this paper, we focus on speech enhancement in the spectral domain based on RNNs, more specifically gated recurrent units (GRUs) [17].




(a) Offline processing (b) Online processing

Figure 1: An overview of offline and online processing for speech enhancement.

Inspired by the successful application of skip connections in DNN-based speech enhancement systems [18–20], we propose to utilize gated recurrent units with skip connection (Skip-GRU), where a skip connection with an identity weight matrix is applied between the GRU layers. The skip connection along the depth of the GRU layers ensures a smooth flow of information directly from the lower layer to the upper layer without adding significant additional complexity to the system. We propose two systems based on the Skip-GRU architecture, either estimating a real-valued mask or a complex-valued mask. Experimental results on reverberant noisy signals from the test set of the first DNS challenge [21] show that both proposed Skip-GRU systems achieve a significant performance improvement compared to the plain GRU systems without the skip connection. In addition, the Skip-GRU system estimating a complex-valued mask outperforms the Skip-GRU system estimating a real-valued mask. Moreover, the proposed complex-valued masking Skip-GRU system outperforms two baseline

The Oldenburg Branch for Hearing, Speech and Audio Technology HSA is funded in the program »Vorab« by the Lower Saxony Ministry of Science and Culture (MWK) and the Volkswagen Foundation for its further development.

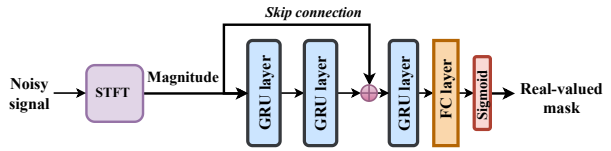



Figure 2: Proposed Skip-GRU for real-valued masking.

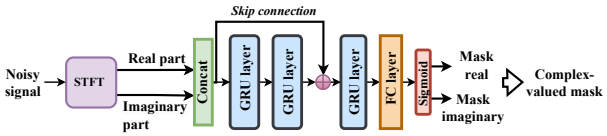



Figure 3: Proposed Skip-GRU for complex-valued masking.

systems for real-time speech enhancement [11, 14] in terms of the scale-invariant signal-to-distortion ratio (SI-SDR) [22], while achieving a similar performance as the best-performing baseline system [9] with a significantly reduced number of parameters and complexity.

The remainder of this paper is organized as follows. In Section 2 we present the proposed Skip-GRU systems for real-valued and complex-valued masking. Section 3 discusses the utilized datasets, architectures, and parameters used for training and evaluation. Section 4 presents the experimental results for the proposed systems and the baseline systems.

# 2 Methods

We consider a scenario where a single microphone records a mixture of the desired speech signal and background noise. In the short-time Fourier transform (STFT) domain, the microphone signal $Y(k,l)$ is equal to

$$Y(k,l) = X(k,l) + V(k,l), \tag{1}$$

where $X(k,l)$ and $V(k,l)$ denote the desired speech component and the noise component, respectively, and $k$ and $l$ denote the frequency bin index and the time frame index, respectively. In this paper, we consider masking-based approaches [10–14, 23], where the enhanced speech signal $\hat{X}(k,l)$ in the STFT-domain is estimated by applying a mask to the noisy signal, i.e.

$$\hat{X}(k,l) = M(k,l) \cdot Y(k,l), \tag{2}$$

where $M(k,l)$ represents either a real-valued mask, only enhancing the magnitude, or a complex-valued mask, enhancing both magnitude as well as phase.

In the next subsections we will discuss how these masks are estimated using the proposed Skip-GRU architecture. The enhanced speech signal in the time-domain is computed by applying an inverse STFT to $\hat{X}(k,l)$ and using an overlapp-add procedure.

## 2.1 Skip-GRU for Real-valued Masking

Figure 2 depicts the proposed Skip-GRU system for real-valued masking. The system consists of 3 GRU layers with a skip connection between the input of the first GRU layer and the output of the second GRU layer. The output of the second GRU layer is utilized as the input of the third GRU layer to ensure a smooth flow of information from the lower GRU layer to the higher GRU layer. The real-valued mask is estimated using a fully connected (FC) layer with a sigmoid activation. The input feature to the first GRU layer is the magnitude of the noisy STFT coefficients $|Y(k,l)|$.

## 2.2 Skip-GRU for Complex-valued Masking

Figure 3 depicts the proposed Skip-GRU system for complex-valued masking, where the architecture with 3 GRU layers and a skip connection very similar to Figure 2 is utilized. The real and imaginary parts of the mask are estimated using an FC layer with sigmoid activation. The input feature to the first GRU layer is the concatenation of the real and imaginary part of the noisy STFT coefficients $Y(k,l)$.

# 3 Experiments

In this section, we discuss the dataset used for training, and validation along with the network architectures and their training procedures.

## 3.1 Datasets

To generate the training and validation data, we consider 3 different datasets with a sampling rate of 16 kHz, namely the second DNS challenge dataset [24], the MUSAN dataset [25], and the GlobalPhone dataset [26]. The second DNS challenge dataset consists of clean speech samples categorized as read speech, singing voice, emotional speech, and non-English speech, while the noise samples come from 150 different categories. The MUSAN dataset consists of clean speech, noise, and music, while the GlobalPhone dataset consists of only clean speech samples from 20 different languages. The second DNS challenge dataset also consists of room impulse responses (RIRs) collected from the SLR26 and SLR28 datasets [27].

For the training, we generated both noisy input signals with and without reverberation. To generate noisy signals without reverberation, we randomly selected a clean speech sample either from the second DNS challenge dataset or the MUSAN dataset or the GlobalPhone dataset and a noise sample either from the second DNS challenge dataset or from the MUSAN dataset (noise or music) and mixed them together at an SNR between $-10$ and 30 dB. When considering reverberation, we first convolved the speech signal with a randomly chosen room impulse response from the SLR26 and SLR28 datasets. All samples have a fixed length of 4 seconds. In total, we generated 700 hours of training data containing 300 hours with reverberation and 400 hours without reverberation. The data were split into an 80 : 20 ratio for training and validation. It should be noted that we utilized all clean speech categories from the second DNS challenge dataset including singing voice, emotional speech, and non-English speech.

## 3.2 Network Architectures and Training Settings

Both proposed Skip-GRU systems utilize 3 GRU layers along with an FC layer and sigmoid activation, where a skip connection with an identity weight matrix is employed between the input of the first GRU layer and the output of

| | (Layers,Nodes) | FC Nodes |
|---|---|---|
| GRU (real) | 3, 256, 256, 256 | 257 |
| Skip-GRU (real) | 3, 256, 256, 256 | 257 |
| GRU (complex) | 3, 512, 512, 256 | 514 |
| Skip-GRU (complex) | 3, 512, 512, 256 | 514 |

Table 1: Parameters of the proposed Skip-GRU systems with real-valued and complex-valued masks.

the second GRU layer. For both systems, the STFT input features are computed using an FFT size of 512, a square-root Hann window with a frame length of 512, and a frame shift of 256 samples. The zeroth frequency is not considered for the input features for both Skip-GRU systems.

The Skip-GRU system for real-valued masking using the noisy magnitude $|Y(k,l)|$ as input feature utilizes 256 nodes in each GRU layer and 257 nodes for the FC layer. The Skip-GRU system for complex-valued masking uses the concatenation of real and imaginary parts of $Y(k,l)$ as input feature and hence, doubling the input size compared to the Skip-GRU system for real-valued masking. The Skip-GRU system for complex-valued masking utilizes $2 \cdot 256$ nodes for the first two GRU layers to efficiently employ a skip connection between the input of the first GRU layer and the output of the second GRU layer concerning real and imaginary parts, whereas it utilizes only 256 nodes for the last GRU layer. The last GRU layer is followed by an FC layer having $2 \cdot 257$ nodes (257 for the real part and 257 for the imaginary part). A dropout of 50% is utilized between the last GRU layer and the FC layer for both Skip-GRU systems. The number of nodes in the last GRU layer in the Skip-GRU system for complex-valued masking is kept the same as the Skip-GRU for real-valued masking to reduce the number of parameters and keep it as small as possible while comparing with the Skip-GRU system for real-valued masking.

To investigate the advantage of employing a skip connection between GRU layers, we have also considered plain GRU systems without a skip connection, one for real-valued masking and one for complex-valued masking. Table 1 summarizes the parameters utilized for plain GRU systems and the proposed Skip-GRU systems. To train all systems, we have utilized the scale-invariant signal-to-noise ratio (SI-SNR) loss function [28] and a max-normalization scheme on the input signals. It should be noted that the max-normalization is only utilized during the training of the systems. Each system has been trained for 200 epochs with an early stopping criterion of 10 epochs using the ADAM optimizer [29] with a learning rate of 0.0002.

We consider 3 different speech enhancement baseline systems, namely: DEMUCS [9], DTLN [11], and FullSubNet+ [14]. DEMUCS performs the speech enhancement in the time-domain utilizing an encoder-LSTM-decoder architecture, while DTLN and FullSubNet+ perform the speech enhancement in the spectral domain utilizing an LSTM architecture to compute a real-valued mask and a TCN architecture to compute a complex-valued mask, respectively. In particular, DEMUCS and FullSubNet+ have shown impressive speech enhancement results. However, they consist of a large number of parameters and require high computational complexity. On the other hand, DTLN is a system having less parameters and requires low computational complexity. All these systems are current state-of-the-art and can perform speech enhancement in real time. As the goal of this work is to achieve a resource-efficient low-complexity real-time speech enhancement system, to make a fair comparison of the performance of proposed systems we chose two baseline systems with relatively large complexity having a large number of parameters and one system with low complexity having less number of parameters. It should be noted that all considered baseline systems have been retrained on the same data as the proposed systems without using any look-ahead. It should also be noted that all considered systems have the same latency of 32 ms.

# 4 Results and Discussion

The performance of all considered speech enhancement systems was evaluated on the publicly available test set of the first DNS challenge dataset [21], where we only considered the reverberant subset (150 signals). As performance measures for speech enhancement, we have used the scale-invariant signal-to-distortion ratio (SI-SDR) [22], the wideband perceptual evaluation of speech quality (PESQ) metric [30], and the short-time objective intelligibility (STOI) metric [31]. For all performance measures, the clean speech signal was used as the reference signal. Furthermore, we have also utilized a DNN-based MOS predictor, namely: DNSMOS [24] to evaluate the performance of all considered systems. As additional metrics for complexity and computational costs, we also computed the total number of multiplications and additions (MACs) per second and the total number of parameters (#Param) using the Torchinfo library of PyTorch.

Table 2 shows the mean SI-SDR (in dB), the mean PESQ, the mean STOI, the DNSMOS, the number of MACs, and the number of parameters obtained for the considered baseline systems, plain GRU systems, and the proposed Skip-GRU systems. First, it can be observed that the proposed Skip-GRU system for real-valued masking achieves a significant performance improvement in terms of all performance measures compared to the input signal and the plain GRU system for real-valued masking. By employing the skip connection, a significant performance improvement can hence be achieved for the same system complexity and number of parameters. Second, it can be observed that both the plain GRU system and the proposed Skip-GRU system for complex-valued masking improve the performance compared to their corresponding systems for real-valued masking, however at a larger system complexity and number of parameters. It can also be observed that the proposed Skip-GRU system for complex-valued masking achieves a better performance in terms of both SI-SDR as well as DNSMOS compared to the plain GRU system for complex-valued masking.

When comparing the proposed Skip-GRU systems to the baseline systems [9, 11, 14], it can be observed that both Skip-GRU systems have significantly less parameters and number of MACs than the FullSubNet+ and DEMUCS systems, while the DTLN system has even less parameters and number of MACs than the Skip-GRU systems. In terms of all performance measures, the best-performing Skip-GRU system for complex-valued masking achieves a similar performance as the best-performing DEMUCS baseline system, but with a significantly lower computational complexity and number of parameters. In addition,

| Systems | SI-SDR (dB) | WB-PESQ | STOI | DNSMOS | MACs (G/s) | #Param |
|---|---|---|---|---|---|---|
| Input Noisy Signals | 9.2 | 1.8 | 0.87 | 2.73 | - | - |
| FullSubNet+ (ret) [14] | 13.5 | 2.6 | 0.89 | 2.96 | 31.81 | 8.7 M |
| DEMUCS-48 (ret) [9] | 14.5 | 2.5 | 0.90 | 3.08 | 1.51 | 18.9 M |
| DTLN (ret) [11] | 12.2 | 2.1 | 0.88 | 2.84 | 0.12 | 1.0 M |
| GRU (real) | 13.2 | 2.2 | 0.89 | 2.92 | 0.21 | 1.8 M |
| Skip-GRU (real) | 13.9 | 2.3 | 0.90 | 2.95 | 0.21 | 1.8 M |
| GRU (complex) | 14.1 | 2.4 | 0.90 | 2.96 | 0.39 | 4.4 M |
| Skip-GRU (complex) | **14.4** | **2.4** | **0.90** | **2.98** | 0.39 | 4.4 M |

Table 2: SI-SDR (dB), wideband PESQ, STOI, DNSMOS, MACs, and number of parameters for the baseline systems, plain GRU systems without skip connections, and proposed skip-GRU systems for real-valued and complex-valued masking. The performance measures were averaged over all signals of the test set of the first DNS challenge dataset.

the same Skip-GRU system achieves a significant performance improvement of 0.9 dB in terms of SI-SDR compared to the FullSubNet+ system and 2.2 dB compared to the DTLN system, while achieving a comparable performance in terms of PESQ and STOI.

# 5 Conclusions

In this paper, we investigated the advantages of incorporating a skip connection between GRU layers for real-time masking-based speech enhancement in the spectral domain. Aiming at achieving a compact enhancement system with low complexity, we proposed two Skip-GRU systems, which utilize a skip connection between the input of the first GRU layer and the output of the second GRU layer. We evaluated the proposed systems on synthetic reverberant noisy signals from the first DNS challenge dataset. Experimental results demonstrate the advantage of the skip connection both for real-valued masking as well as for complex-valued masking. The best-performing Skip-GRU system achieves a similar performance as the best-performing baseline DEMUCS system but with about 4 times less parameters and computational complexity. In future work, we will focus on further reducing the complexity as well as the latency.

# References


[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2013.

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

[3] S. R. Park and J. W. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, (Stockholm, Sweden), pp. 1993–1997, Aug. 2017.

[4] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, (Stockholm, Sweden), pp. 3642–3646, Aug. 2017.

[5] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Calgary, Canada), pp. 5069–5073, Apr. 2018.

[6] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully convolutional recurrent networks for speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 6674–6678, May 2020.

[7] H. Fang, G. Carbajal, S. Wermter, and T. Gerkmann, "Variational autoencoder for speech enhancement with a noise-aware encoder," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 676–680, June 2021.

[8] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brighton, UK), pp. 6875–6879, May 2019.

[9] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, (Shanghai, China), pp. 3291–3295, Oct. 2020.

[10] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement.," in *Proc. Interspeech*, (Hyderabad, India), pp. 3229–3233, Sept. 2018.

[11] N. L. Westhausen and B. T. Meyer, "Dual-signal transformation LSTM network for real-time noise suppression," in *Proc. Interspeech*, (Shanghai, China), pp. 2477–2488, Oct. 2020.

[12] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Real-time speech enhancement using equilibriated RNN," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Barcelona, Spain), pp. 851–855, May 2020.

[13] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 6633–6637, June 2021.

[14] J. Chen, Z. Wang, D. Tuo, Z. Wu, S. Kang, and H. Meng, "Fullsubnet+: Channel attention fullsubnet with complex spectrograms for speech enhancement," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Singapore), pp. 7857–7861, May 2022.

[15] Z.-Q. Wang, G. Wichern, S. Watanabe, and J. Le Roux, "STFT-domain neural speech enhancement with very low algorithmic latency," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 31, pp. 397–410, 2022.

[16] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "Phasen: A phase-and-harmonics-aware speech enhancement network," in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 34, (New York, USA), pp. 9458–9465, 2020.

[17] K. Cho, B. van Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, A meeting of SIGDAT, a Special Interest Group of the ACL* (A. Moschitti, B. Pang, and W. Daelemans, eds.), (Doha, Qatar), pp. 1724–1734, ACL, Oct. 2014.

[18] M. Tu and X. Zhang, "Speech enhancement based on deep neural networks with skip connections," in *Proc. IEEE*

*International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New Orleans, USA), pp. 5565–5569, Mar. 2017.

[19] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 656–660, June 2021.

[20] O. K. Oyedotun, K. Al Ismaeil, and D. Aouada, "Why is everyone training very deep neural network with skip connections?," *IEEE Trans. on Neural Networks and Learning Systems*, 2022.

[21] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, and S. Braun, "The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results," in *Proc. Interspeech*, (Shanghai, China), pp. 2492–2496, Oct. 2020.

[22] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Brighton, UK), pp. 626–630, 2019.

[23] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[24] C. K. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 6623–6627, June 2021.

[25] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[26] T. Schultz, N. T. Vu, and T. Schlippe, "Globalphone: A multilingual text & speech database in 20 languages," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Vancouver, Canada), pp. 8126–8130, May 2013.

[27] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (New Orleans, USA), pp. 5220–5224, Mar. 2017.

[28] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd International Conference for Learning Representations*, (San Diego, USA), pp. 1–15, July 2015.

[30] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs P.862," tech. rep., International Telecommunications Union (ITU-T) Recommendation, Feb. 2001.

[31] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.