INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising

Maximilian Strake[1], *Bruno Defraene*[2], *Kristoff Fluyt*[2], *Wouter Tirry*[2], *Tim Fingscheidt*[1]

[1]Technische Universität Braunschweig
Institute for Communications Technology, 38106 Braunschweig, Germany
[2]Goodix Technology (Belgium) BV, 3000 Leuven, Belgium

m.strake@tu-bs.de, bdefraene@goodix.com, kfluyt@goodix.com, wtirry@goodix.com, t.fingscheidt@tu-bs.de

Abstract

The Interspeech 2020 Deep Noise Suppression (DNS) Challenge focuses on evaluating low-latency single-channel speech enhancement algorithms under realistic test conditions. Our contribution to the challenge is a method for joint dereverberation and denoising based on complex spectral mask estimation using a fully convolutional recurrent network (FCRN) which relies on a convolutional LSTM layer for temporal modeling. Since the effects of reverberation and noise on perceived speech quality can differ notably, a multi-target loss for controlling the weight on desired dereverberation and denoising is proposed. In the crowdsourced subjective P.808 listening test conducted by the DNS Challenge organizers, the proposed method shows a significant overall improvement of 0.43 MOS points over the DNS Challenge baseline and ranks amongst the top-3 submissions for both realtime and non-realtime tracks of the challenge.
Index Terms: speech enhancement, denoising, dereverberation, convolutional recurrent neural networks, realistic data

1. Introduction

In real-world scenarios, speech signals are often affected by background noise and reverberation, which significantly degrades intelligibility and perceived speech quality in applications such as mobile speech communication or hearing aids. Also the performance of automatic speech recognition systems is affected negatively. Single-channel speech enhancement methods based on deep neural networks (DNNs) have been shown to effectively perform either dereverberation [1, 2] or denoising [3, 4], but only few studies consider both interference types jointly [5–8].

For denoising tasks, methods based on convolutional neural networks have recently shown very promising performance [9–16], which is credited to their ability to focus on local structural patterns of speech (e.g., spectral harmonics) that facilitate the discrimination from noise, see also [17, 18]. Park et al. [10] introduce a fully convolutional encoder-decoder network (FCN), which first maps the input features to a latent space representation in the encoder, before mapping back to the original input feature structure using a decoder that mirrors the encoder. Several studies extend these CNN architectures towards convolutional recurrent networks (CRNs) by introducing long short-term memory (LSTM) layers for temporal modeling [11, 12], which leads to improved denoising performance, but also comes with the need to discard the *fully* convolutional nature of the solely CNN-based models. Very recently, a fully convolutional recurrent network (FCRN) for denoising was proposed [15, 16], which replaces the LSTM layers in standard CRNs with a convolutional LSTM (ConvLSTM). This has been shown to preserve the local harmonic structure of speech spectra throughout the feature representations in the network, leading to improved denoising performance.

The difficult task of joint dereverberation and denoising in a single-channel scenario is addressed by Han et al. using a fully connected DNN performing spectral mapping, but they mostly restrict themselves to speaker-dependent processing [5]. In [6] the authors argue that reverberation only significantly affects intelligibility of clean speech when reverberation times (RT60) are longer than those in typical real-world scenarios. Therefore, they propose to use reverberated clean speech targets for a masking-based DNN approach, effectively performing reverberation-aware denoising. To alleviate the difficulties in training a model for joint dereverberation and denoising, a progressive learning framework is used in [7], where LSTM layers are stacked and each successive layer is trained with an intermediate target with increasingly higher SNR and lower RT60. A promising attempt at using FCNs for joint dereverberation and denoising has been made in [19], but explicit temporal modeling via recurrent neural network layers is not considered, although it could be especially advantageous in handling the strong temporal relations introduced by reverberation.

All of the discussed methods use a supervised training approach based on synthetic data, where clean speech and noise are recorded separately and reverberation (if considered at all) is applied by convolution with either simulated or separately recorded room impulse responses (RIRs). If also validation is conducted only on such synthetic data, the generalization to real-world conditions is not at all guaranteed. Furthermore, most studies only report results based on instrumental measures for speech quality [20–22], which have been shown to not correlate strongly to subjective ratings and therefore to not be fully reliable in predicting the quality of speech enhancement algorithms [23]. The Interspeech 2020 Deep Noise Suppression (DNS) Challenge [24] addresses these issues by providing test data recorded under real conditions and evaluating the speech enhancement performance in a crowdsourced subjective listening test setup recently standardized in ITU-T P.808 [25].

Our contribution is twofold and focuses on adapting the FCRN training from [16] to handle dereverberation in addition to denoising. First, we propose to employ the FCRN for the challenging task of joint dereverberation and denoising in strongly varying conditions of noise type, reverberation, recording devices, signal-to-noise ratio (SNR), and input signal level. Additionally, the model performance is evaluated in a subjective

http://dx.doi.org/10.21437/Interspeech.2020-2439

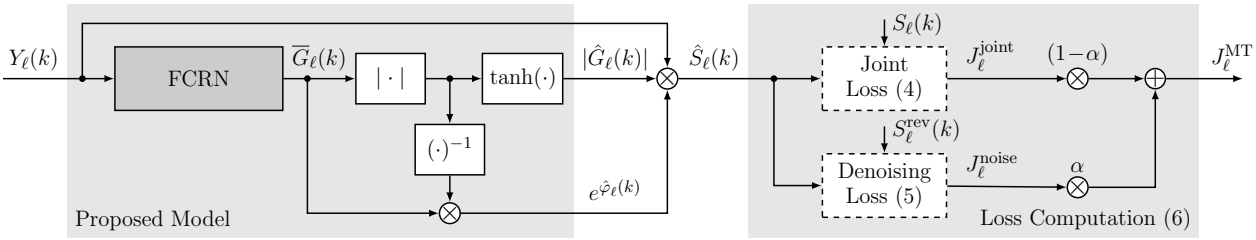

Figure 1: *Proposed model (training & test) and multi-target loss J_ℓ^{MT} computation (training only).*

test setup including real data, whereas in [16] only instrumental measures were employed. As a second contribution, we propose a multi-target loss function providing the possibility to balance between denoising and dereverberation performance by using weighted loss terms employing either non-reverberated clean speech or reverberated clean speech as targets. This loss is motivated by two works: The components loss by Xu et al. [13, 14] for denoising already allowed to separately control speech distortion, residual noise power, and residual noise quality. Secondly, from the findings of Zhao et al. [6] (and supported by audiological research [26]) we derive that a strong dereverberation does not necessarily improve perceptual quality of speech and the loss can be adapted towards a stronger focus on noise reduction, while still providing acceptable dereverberation.

The paper is structured as follows. Section 2 briefly discusses the processing framework, followed by the description of the proposed method in Section 3. The experimental evaluation including data setup, training details, and result discussion is presented in Section 4. We conclude the paper in Section 5.

2. Signal Model and Notations

A clean speech signal $s(n)$ disturbed by additive noise $d(n)$ and reverberation characterized by the room impulse response (RIR) $h(n)$ can be described as

$$y(n) = s(n) * h(n) + d(n) = s^{\text{rev}}(n) + d(n), \quad (1)$$

where $s^{\text{rev}}(n)$ is the reverberated speech component, n is the sample index, and $*$ denotes convolution. The proposed framework for dereverberation and denoising operates in the discrete Fourier transform (DFT) domain, where the noisy speech spectrum for a frame ℓ is given by

$$Y_\ell(k) = S_\ell(k) \cdot H_\ell(k) + D_\ell(k) = S_\ell^{\text{rev}}(k) + D_\ell(k), \quad (2)$$

with $k \in \mathcal{K} = \{0, \ldots, K-1\}$ being the frequency bin index and K the DFT size.

3. Proposed Method

The proposed method for joint dereverberation and denoising is composed of a fully convolutional recurrent network (FCRN) model topology and a training procedure for joint dereverberation and denoising that is adaptable to human quality perception by adjusting the focus of learning either more on denoising or on dereverberation. An illustrating system overview is given in Figure 1. The FCRN is trained to estimate a magnitude-bounded complex mask $\hat{G}_\ell(k) \in \mathbb{C}$ which is used to retain the enhanced speech spectrum $\hat{S}_\ell(k) = \hat{G}_\ell(k) \cdot Y_\ell(k)$. The bounding of the mask magnitude range to $|\hat{G}_\ell(k)| \in [0, 1]$ is achieved following [27] by

$$\hat{G}_\ell(k) = |\hat{G}_\ell(k)| \cdot e^{\hat{\varphi}_\ell(k)} = \tanh(|\overline{G}_\ell(k)|) \cdot \frac{\overline{G}_\ell(k)}{|\overline{G}_\ell(k)|}, \quad (3)$$

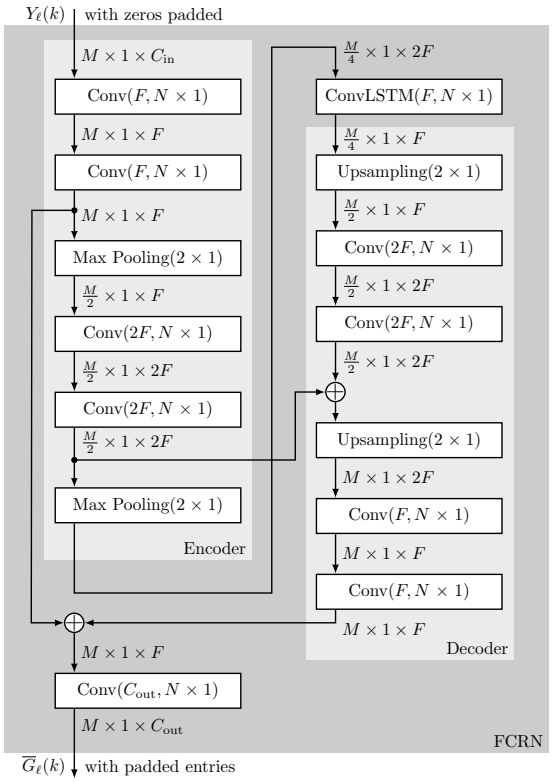

Figure 2: *FCRN topology used by the proposed method.*

where $\overline{G}_\ell(k) \in \mathbb{C}$ is the network output representing the unbounded complex mask and $\hat{\varphi}_\ell(k) = \arg(\overline{G}_\ell(k))$ is the respective estimated phase.

3.1. FCRN Model Topology

The FCRN topology is based on the convolutional encoder-decoder topology proposed for enhancement of coded speech in [28] and adapted by including a convolutional LSTM (ConvLSTM) layer [29] for temporal modeling. This keeps the model *fully* convolutional and was shown to be effective for denoising in [16]. The FCRN topology is depicted in Figure 2, where the dimensionality of the feature representations before and after each layer is given in the form *feature axis size* $\times$ *time frame axis size* $\times$ *number of feature maps*. The input size on the feature axis is computed as $M = K/2 + 1 + P$, where P refers to an amount of padded zeros (see Section 4.2 for details). Standard convolutional layers in the encoder, decoder, and the output layer are denoted by $\text{Conv}(F, N \times 1)$, where F determines the amount of filter kernels and N the size of these kernels along the feature axis for the respective layer. We denote the kernel size as $N \times 1$ to emphasize that convolutions are only performed

along the feature axis and not along the time frame axis. The encoder part of the FCRN compresses the feature axis from size M to $M/4$ by applying maximum pooling exclusively along this axis. The encoded feature representation is input to the convolutional LSTM layer denoted by $\mathrm{ConvLSTM}(F, N \times 1)$ with F and N, as above, determining the number of filter kernels and their size along the feature axis for all convolutional mappings in the ConvLSTM. Next, the decoder part mirrors the encoder and reconstructs the original feature axis size M by applying upsampling layers. All convolutional layers use leaky ReLU activations with a negative slope coefficient of 0.2, except for the output layer, which employs a linear activation and maps back to the desired amount of output feature maps C_{out}. The topology uses skip connections between encoder and decoder layers with matching feature representation sizes to give the decoder access to high-resolution features and to ease the gradient flow during backpropagation training.

3.2. Multi-Target Loss for Dereverberation and Denoising

The proposed multi-target (MT) loss, as shown in Figure 1 on the right, is based on two separate loss terms operating in the complex spectral domain. The *joint* loss term

$$J_\ell^{\mathrm{joint}} = \frac{1}{K} \sum_{k \in \mathcal{K}} \left| \hat{S}_\ell(k) - S_\ell(k) \right|^2 \tag{4}$$

aims at joint dereverberation and denoising by employing clean speech spectra $S_\ell(k)$ as targets. The *denoising* loss term

$$J_\ell^{\mathrm{noise}} = \frac{1}{K} \sum_{k \in \mathcal{K}} \left| \hat{S}_\ell(k) - S_\ell^{\mathrm{rev}}(k) \right|^2 \tag{5}$$

exclusively aims at denoising by employing reverberated clean speech spectra $S_\ell^{\mathrm{rev}}(k)$ as targets. The combination of these terms to a total *multi-target* (MT) loss

$$J_\ell^{\mathrm{MT}} = (1-\alpha) J_\ell^{\mathrm{joint}} + \alpha J_\ell^{\mathrm{noise}} \tag{6}$$

can be adapted towards a weaker or stronger desired dereverberation using the weighting factor $\alpha \geqslant 0$, where $\alpha > 0$ lets the model put an additional focus on denoising and implicitly puts less weight on dereverberation. In addition, the combination of clean and reverberated targets implicitly provides information about the two distinct disturbances of noise and reverberation, which is not the case when exclusively using J_ℓ^{joint} ($\alpha = 0$) or J_ℓ^{noise} ($\alpha = 1$).

4. Experimental Evaluation

4.1. Datasets and Preprocessing

We train our models using a two-step training approach, first performing a pretraining using a dataset based on WSJ0 speech [30] (denoted as $\mathcal{D}_{\mathrm{WSJ0}}$) and second, finetuning the pretrained models using a subset of the training data provided from the DNS Challenge [24] (denoted as $\mathcal{D}_{\mathrm{DNS}}$). This approach allows us to limit the amount of training time by testing multiple hyperparameter settings only for finetuning on $\mathcal{D}_{\mathrm{DNS}}$. In addition, we believe that it can be advantageous for learning to first train the model on an easier task ($\mathcal{D}_{\mathrm{WSJ0}}$), where, e.g., a good feature representation of speech can be found, and only in a second step adapt this model to a more difficult task ($\mathcal{D}_{\mathrm{DNS}}$).

The pretraining dataset $\mathcal{D}_{\mathrm{WSJ0}}$ uses 15 hours of clean speech from WSJ0 `SI-84` for training and 2.5 hours of clean speech from WSJ0 `si_dt_05` for validation. The clean speech is mixed with noise material from the DEMAND [31] and QUT [32] databases (35 different noise files shared in training and validation) using SNR conditions of 0, 5 and 10 dB. We mix different random subsets each using 1 hour of the total speech material with each of the conditions, resulting in 105 hours of training and 18 hours of validation material. The active speech level of clean speech is set to $-26\,\mathrm{dBov}$ using ITU-T P.56 [33] for all files of $\mathcal{D}_{\mathrm{WSJ0}}$ before mixing.

Finetuning is carried out based on an 100 hour subset (randomly chosen files) of the official DNS Challenge training material [24], where SNRs are sampled uniformly between 0 and $40\,\mathrm{dB}$ and the RMS level of the resulting signal is set to a value uniformly sampled between -38 and $-18\,\mathrm{dBov}$. To reflect a large variety of real-world conditions, we include reverberation to 50% of the files in $\mathcal{D}_{\mathrm{DNS}}$. This is achieved by convolving the clean speech component of the mixture with simulated RIRs generated with the mirror method using [34]. For RIR generation we use room sizes uniformly drawn from $(l, w, h) \in ([3, 10]\,\mathrm{m}, [3, 10]\,\mathrm{m}, [2.5, 3.5]\,\mathrm{m})$ and an absorbtion coefficient uniformly drawn from $\alpha \in [0.1, 0.3]$ for all room surfaces. The microphone is assumed in the center of the room and the source distance in the l-w-plane is drawn from $d \in [0.1, 1]\,\mathrm{m}$, placing the source on a randomly chosen point in the sampled distance. This configuration leads to estimated RT60s between 0.28 and $1.66\,\mathrm{s}$, calculated with Sabine's equation. The development test set of the DNS Challenge described in [24] is randomly split into two halves, where one is used for validation during finetuning and the other serves as our preliminary test set for model evaluation. The final evaluation using the subjective test framework P.808 with crowdsourcing [25] is carried out by the DNS Challenge organizers with the DNS Challenge blind test set, which has not been used to perform any training or optimization in the development process.

A sampling frequency of $16\,\mathrm{kHz}$ is used for all audio material and frames are extracted using a frame length of $32\,\mathrm{ms}$ (compliant with the maximum frame size of $40\,\mathrm{ms}$ allowed for the DNS Challenge), a frame shift of $16\,\mathrm{ms}$, square-root Hann windowing, and a DFT-size of $K = 512$. Only non-redundant bins of the spectra are used and real and imaginary parts are organized in separate feature maps for input features and targets, resulting in a feature axis size of $M = 257+3 = 260$ including zero-padding, which guarantees the devisibility by four as required for the FCRN topology in Figure 2. In accordance to the challenge rules, two frames of future context ($32\,\mathrm{ms}$ lookahead in total) are concatenated to the inputs as separate feature maps, resulting in $C_{\mathrm{in}} = 3 \cdot 2 = 6$ input and $C_{\mathrm{out}} = 2$ output channels, the factor 2 reflecting real and imaginary parts.

4.2. Training Details and Reference Models

For the propsed FCRN models, the amount of filter kernels is set to $F = 88$ and the kernel size is chosen as $N = 24$. Both pretraining and finetuning of these models use truncated backpropagation-through-time training with a sequence length of 100 frames and a batchsize of 16, employing the Adam optimizer with standard parameter settings as given in [35], except for the learning rate. For pretraining, we use a starting learning rate of 0.0001, which is reduced by a factor of 5 once the validation loss does not improve further for a consecutive four epochs. We stop training, once the learning rate falls below 0.00001. Since no reverberated data is included in the pretraining, we set $\alpha = 0$ for the loss computation (6). For finetuning, we train for 30 epochs with a fixed learning rate of 0.00002 and choose the model with the best validation set performance in

Table 1: ***Instrumental quality results*** *on the* ***preliminary test set****, evaluated separately for* ***synthetic data*** *without and with reverberation. Best results are in* ***bold font****.*

	Method	Without Reverb				With Reverb				
		PESQ	POLQA	STOI	$\Delta\mathrm{SNR}_{\mathrm{seg}}$ [dB]	PESQ	POLQA	STOI	$\Delta\mathrm{SNR}_{\mathrm{seg}}$ [dB]	SRMR
REF	Noisy	2.21	2.51	0.91	-	1.57	1.54	0.56	-	-
REF	DNS Baseline [36]	2.68	2.40	0.77	7.36	1.69	1.29	0.41	**8.54**	5.34
REF	FCRN-cSA	3.19	3.48	**0.96**	7.87	2.07	1.94	0.68	8.46	8.64
NEW	FCRN-MT, $\alpha = 0$	3.33	3.59	**0.96**	8.12	**2.16**	1.93	**0.70**	8.18	**9.10**
NEW	FCRN-MT, $\alpha = 0.1$	3.32	3.59	**0.96**	8.16	2.14	**1.95**	0.69	7.67	8.79
NEW	FCRN-MT, $\alpha = 1$	**3.40**	**3.70**	**0.96**	**8.71**	1.81	1.67	0.58	5.37	4.82

Table 2: ***Subjective quality results*** *in terms of* ***MOS scores*** *according to ITU-T P.808 on the* ***blind test set****. Extract from the test with all 28 submissions to the DNS Challenge's* ***realtime track (RT)*** *and* ***non-realtime track (NRT)****. Best results are in* ***bold font****.*

Submission	**Challenge Track**	Synth. Without Reverb	Synth. With Reverb	Real Recordings	**Overall**
Noisy	-	3.32	2.78	2.97	3.01
DNS baseline [36]	RT	3.49	2.64	3.00	3.03
FCRN-MT, $\alpha = 0.1$ (Ours)	RT	**3.86**	3.21	**3.39**	**3.46**
FCRN-MT, $\alpha = 0.1$ (Ours)	NRT	3.85	**3.23**	**3.39**	**3.46**

terms of the PESQ metric [20]. The MT loss (6) and respective weighting with α is effective for finetuning and is used as a tuning parameter for the proposed model which we refer to as **FCRN-MT**. The proposed model has 5.2 million trainable parameters and takes an average computation time of $10.71\,\mathrm{ms}$ (measured on an Intel Core i5 quad core machine with $3.4\,\mathrm{GHz}$ clock) for processing one frame. Considering the frame shift of $16\,\mathrm{ms}$, this results in a realtime factor of $r = 0.67$, which is below the limit of $r = 1.0$ as given by the realtime DNS Challenge track.

We compare our proposed model with the DNS Challenge baseline [36], which employs a model based on gated recurrent units (GRUs) and fully connected layers as well as a component-based loss formulation [13, 14]. Furthermore, we compare our masking-based FCRN-MT approach with the approach from [16] (**FCRN-cSA**), where the model directly estimates the complex spectrum of clean speech and a standard complex MSE loss is employed.

4.3. Results and Discussion

Results on the synthetic data of the preliminary test set are reported in Table 1 in terms of perceptual evaluation of speech quality (PESQ) [20], perceptual objective listening quality analysis (POLQA) [22], the short-time objective intelligibility (STOI) metric [21], and the segmental SNR improvement $\Delta\mathrm{SNR}_{\mathrm{seg}}$, where the segmental SNR is computed following [37]. All of these metrics use the clean speech signal $s(n)$ as reference. For preliminary test data with reverberation, the dereverberation performance is measured using the speech-to-reverberation modulation energy ratio (SRMR) [38]. In terms of almost all instrumental measures the proposed FCRN-MT ($\alpha = 0$) model outperforms the reference models including FCRN-cSA, which suggests that a masking based approach is better suited for the joint denoising and dereverberation task. The comparison of FCRN-MT with different weighting parameters α shows that for data without reverb, performance in terms of instrumental measures is best for $\alpha = 1$, whereas for data with reverb the best performance for all metrics except POLQA is reached with $\alpha = 0$. In this case, the instrumental measures only partly reflect the observations we made by informal subjective listening, where $\alpha = 0.1$ showed comparable performance for data without reverberation and the best performance for data with reverberation. The latter can be credited to speech component distortions introduced by aiming at complete dereverberation with $\alpha = 0$. Taking into account the limitations of the instrumental measures addressed in Section 1, which is also one of the main problems addressed by the DNS Challenge, we decided to choose FCRN-MT ($\alpha = 0.1$) as our final submission.

In Table 2, the results in terms of MOS scores of the first subjective P.808 test for our submissions to the realtime track (RT) and non-realtime track (NRT) are shown. Please note that both of our submissions for RT and NRT use the exact same FCRN-MT ($\alpha = 0.1$) model which fulfills the more strict RT requirements. Our method significantly outperforms the DNS baseline by overall 0.43 MOS points. In the mixed RT and NRT ranking of all submitted methods, both of our submissions were amongst the top-ranked in a field of in total 28 submissions. For the top-scoring submissions, a second ITU-T P.808 test was conducted for which our method secured the third rank in the RT and the second rank in the NRT of the challenge (for details see [39], team #17).

5. Conclusions

This paper presents a fully convolutional recurrent network (FCRN) for joint dereverberation and denoising as a contribution to the Interspeech 2020 Deep Noise Suppression (DNS) Challenge. We propose to train the FCRN with a multi-target loss accounting for differences in quality perception of noisy or reverberated speech by controlling the weight on desired dereverberation and denoising. Our method is evaluated in a preliminary test based on instrumental measures and in the realistic test setup of the DNS Challenge including real test recordings and evaluation by a crowdsourced subjective listening test. The proposed method outperforms all reference methods of the preliminary test and ranks third for the realtime and second for the non-realtime track amongst all submissions to the challenge.

6. References

[1] B. Wu, K. Li, M. Yang, and C. Lee, "A Reverberation-Time-Aware Approach to Speech Dereverberation Based on Deep Neural Networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 1, pp. 98–107, 2017.

[2] J. F. Santos and T. H. Falk, "Speech Dereverberation With Context-Aware Recurrent Neural Networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 7, pp. 1232–1242, 2018.

[3] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[4] Y. Wang, A. Narayanan, and D. L. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[5] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning Spectral Mapping for Speech Dereverberation and Denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 982–992, 2015.

[6] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-Based Enhancement of Noisy and Reverberant Speech," in *Proc. of ICASSP*, Shanghai, China, Mar. 2016, pp. 6525–6529.

[7] X. Tang, J. Du, L. Chai, Y. Wang, Q. Wang, and C.-H. Lee, "A LSTM-Based Joint Progressive Learning Framework for Simultaneous Speech Dereverberation and Denoising," in *Proc. of APSIPA ASC*, Lanzhou, China, Nov. 2019, pp. 274–278.

[8] D. Ribas, J. Llombart, A. Miguel, and L. Vicente, "Deep Speech Enhancement for Reverberated and Noisy Signals Using Wide Residual Networks," *arXiv:1901.00660*, 2019.

[9] S. Fu, T. Hu, Y. Tsao, and X. Lu, "Complex Spectrogram Enhancement by Convolutional Neural Network With Multi-Metrics Learning," in *Proc. of MLSP*, Tokyo, Japan, Sep. 2017, pp. 1–6.

[10] S. Park and J. Lee, "A Fully Convolutional Neural Network for Speech Enhancement," in *Proc. of Interspeech*, Stockholm, Sweden, Aug. 2017, pp. 1993–1997.

[11] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-Recurrent Neural Networks for Speech Enhancement," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 2401–2405.

[12] K. Tan and D. L. Wang, "Complex Spectral Mapping With a Convolutional Recurrent Network for Monaural Speech Enhancement," in *Proc. of ICASSP*, Brighton, UK, May 2019, pp. 6865–6869.

[13] Z. Xu, S. Elshamy, Z. Zhao, and T. Fingscheidt, "Components Loss for Neural Networks in Mask-Based Speech Enhancement," *arXiv:1908.05087*, 2019.

[14] Z. Xu, S. Elshamy, and T. Fingscheidt, "Using Separate Losses for Speech and Noise in Mask-Based Speech Enhancement," in *Proc. of ICASSP*, Barcelona, Spain, May 2020, pp. 7519–7523.

[15] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated Noise Suppression and Speech Restoration: LSTM-Based Speech Enhancement in Two Stages," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 239–243.

[16] ——, "Fully Convolutional Recurrent Networks for Speech Enhancement," in *Proc. of ICASSP*, Barcelona, Spain, May 2020, pp. 6674–6678.

[17] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[18] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018.

[19] O. Ernst, S. E. Chazan, S. Gannot, and J. Goldberger, "Speech Dereverberation Using Fully Convolutional Networks," in *Proc. of EUSIPCO*, Rome, Italy, Sep. 2018, pp. 390–394.

[20] ITU-T, *Rec. P.862.2: Wideband Extension to Recommendation P862 for the Assesment of Wideband Telephone Networks and Speech Codecs*, Feb. 2001.

[21] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech," in *Proc. of ICASSP*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217.

[22] ITU-T, *Rec. P.863: Perceptual Objective Listening Quality Prediction (POLQA)*, Feb. 2018.

[23] C. K. A. Reddy, E. Beyrami, J. Pool, R. Cutler, S. Srinivasan, and J. Gehrke, "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," in *Proc. of Interspeech*, Graz, Austria, Sep. 2019, pp. 1816–1820.

[24] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Speech Quality and Testing Framework," *arXiv:2001.08662*, 2020.

[25] ITU-T, *Rec. P.808: Subjective Evaluation of Speech Quality With a Crowdsourcing Approach*, Feb. 2018.

[26] R. W. Harris and D. W. Swenson, "Effects of Reverberation and Noise on Speech Recognition by Adults With Various Amounts of Sensorineural Hearing Impairment," *Audiology*, vol. 29, no. 6, pp. 314–321, 1990.

[27] 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," *arXiv:1903.03107*, 2019.

[28] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional Neural Networks to Enhance Coded Speech," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 4, pp. 663–678, Apr. 2019.

[29] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Proc. of NIPS*, Montreal, QC, Canada, Dec. 2015, pp. 802–810.

[30] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[31] J. Thiemann, N. Ito, and E. Vincent, "The Diverse Environments Multi-Channel Acoustic Noise Database: A Database of Multichannel Environmental Noise Recordings," *J. Acoust. Soc. Am*, vol. 133, no. 5, pp. 3591–3591, 2013.

[32] D. B. Dean, S. Sridharan, R. J. Vogt, and M. W. Mason, "The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms," in *Proc. of Interspeech*, Makuhari, Japan, Sep. 2010, pp. 3110–3113.

[33] ITU-T, *Rec. P.56: Objective Measurement of Active Speech Level*, Dec. 2011.

[34] R. Scheibler, E. Bezzam, and I. Dokmanic, "Pyroomacoustics: A Python Package for Audio Room Simulation and Array Processing Algorithms," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 351–355.

[35] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of ICLR*, San Diego, CA, USA, May 2015, pp. 1–15.

[36] Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *Proc. of ICASSP*, Barcelona, Spain, May 2020, pp. 871–875.

[37] Philipos C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[38] T. H. Falk, C. Zheng, and W. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.

[39] C. K. A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. of Interspeech*, Shanghai, China, Oct. 2020, pp. 1–5.