Concatenated Identical DNN (CI-DNN) to Reduce Noise-Type Dependence in DNN-Based Speech Enhancement

Ziyi Xu
Institute for Communications Technology
Technische Universität Braunschweig
Braunschweig, Germany
ziyi.xu@tu-bs.de

Maximilian Strake
Institute for Communications Technology
Technische Universität Braunschweig
Braunschweig, Germany
m.strake@tu-bs.de

Tim Fingscheidt
Institute for Communications Technology
Technische Universität Braunschweig
Braunschweig, Germany
t.fingscheidt@tu-bs.de

***Abstract*—Estimating time-frequency domain masks for speech enhancement using deep learning approaches has recently become a popular field of research. In this paper, we propose a mask-based speech enhancement framework by using *concatenated identical* deep neural networks (CI-DNNs). The idea is that a single DNN is trained under multiple input and output signal-to-noise power ratio (SNR) conditions, using targets that provide a moderate SNR gain with respect to the input and therefore achieve a balance between speech component quality and noise suppression. We concatenate this single DNN several times without any retraining to provide enough noise attenuation. Simulation results show that our proposed CI-DNN outperforms enhancement methods using classical spectral weighting rules w.r.t. total speech quality and speech intelligibility. Moreover, our approach shows similar or even a little bit better performance with much fewer trainable parameters compared with a noisy-target single DNN approach of the same size. A comparison to the conventional clean-target single DNN approach shows that our proposed CI-DNN is better in speech component quality and much better in residual noise component quality. Most importantly, our new CI-DNN generalized best to an unseen noise type, if compared to the other tested deep learning approaches.**

***Index Terms*—Speech enhancement, noise reduction, DNN, noisy speech target**

I. Introduction

Speech enhancement aims at improving the perceived quality and intelligibility of a speech signal degraded by additive noise. This task can be very challenging when only a single-channel mixture signal is available. The classical method to perform single-channel speech enhancement is to estimate the *a priori* signal-to-noise ratio (SNR), which can subsequently be used by way of a spectral weighting rule [1]–[8]. The decision-directed (DD) method proposed by Ephraim and Malah [1], or [5], [6], are widespread *a priori* SNR estimation approaches, that can be combined with spectral weighting rules such as the well known Wiener filter (WF) [3], the MMSE log-spectral amplitude estimator (LSA) [2], and the super-Gaussian joint maximum *a posteriori* estimator (SG) [4]. Nevertheless, using these classical approaches often still leads to poor performance in non-stationary noise environments [9]. Goldstein et al. [10] proposed a multistage representation of the Wiener filter which is achieved by successively decomposing the observed noisy signal onto orthogonal, lower dimensional subspaces. Tinston and Ephraim [11] showed that estimating clean speech using the multistage Wiener filter from the subspace of the obtained noisy speech signal outperforms the conventional Wiener filter. However, the multistage Wiener filter approach still faces the same problem in the presence of non-stationary noise.

Deep learning methods used in speech enhancement tasks have shown excellent results, even in non-stationary noise, and have become state of the art [12]–[21]. A regression-based speech enhancement method using deep neural networks is proposed in [18]. Du et al. [12] proposed a DNN architecture with dual outputs to estimate the speech features belonging to the target and interfering speakers from the input mixture signal, which achieves a better generalization to the unseen interfering speaker. In contrast to use DNNs for a direct regression task, a time-frequency domain mask can be estimated to perform speech enhancement making the ideal estimate independent of the absolute signal level [13], [14], [17]. A complex ratio mask that can enhance the amplitude spectrogram and estimate the right phase information is proposed in [17]. Comparing to other mask estimation methods, using DNNs to directly predict the clean speech signal while estimating the mask representation implicitly is shown to outperform the direct estimation of masks for speech separation [15], [16]. For these deep learning based speech enhancement algorithms, a common problem is the degradation of performance in unseen noise conditions [18]–[20]. One method to address this mismatched noise condition problem is to include many different noise types in the training data [19], [20]. A drawback of this method is that a very large training set is needed, e.g., Xu et al. [20] used 104 types of noise, and a total amount of 100 hours of training data for most experiments to improve generalization capabilities.

Another challenge in DNN-based speech enhancement is to find a good tradeoff between speech distortion and noise reduction, especially for low SNR conditions. Gao et al. [22] proposed to use progressive learning with SNR-based targets to address this problem. A novel progressive deep neural network (PDNN) with a parallel structure of neural networks with less and less noisy targets for each network and horizontal connections between the layers towards less-noisy trained DNNs is proposed in [23]. Training of this PDNN is done one-by-one while freezing the weights of the higher-noise target networks, until the last parallel network, which is trained with clean targets. The total spectral estimate output is the average of all these networks.

In this paper, however, we propose a *serial* concatenation of networks in so-called stages, with the specific property that the network in each stage is *identical*, hence the name *concatenated identical* DNNs (CI-DNNs). The idea is to train a basic DNN module which can yield some moderate enhancement of the input, in our case a 5 dB target signal-to-noise power ratio (SNR) improvement, implemented by using additive noise and a respectively configured target signal. This network is then, e.g., concatenated three times (i.e., 3 stages), in order to provide a sufficient amount of noise attenuation. An important aspect is that such a stage DNN must be trained for multiple input and (enhanced by 5 dB SNR improvement, respectively) output SNRs to operate well both in the first stage and in all subsequent stages. The idea and major advantage vs. [22], [23]

978-9-0827-9703-9/19/$31.00

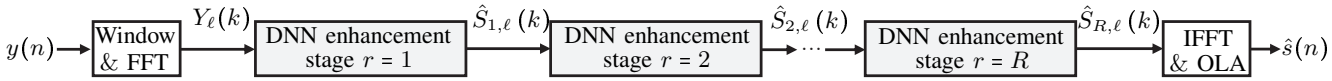

Fig. 1. Diagram of the speech enhancement using CI-DNNs; for details of the DNN enhancement stage see Fig. 2

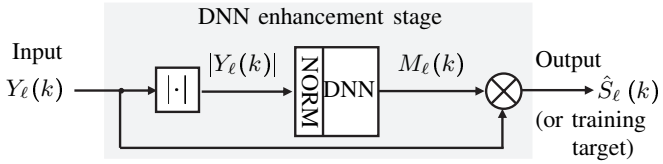

Fig. 2. Basic DNN module for speech spectrum enhancement

is to save free (trainable) parameters compared to (deeper) DNNs with the same number of weights, and thereby to provide better generalization properties particularly for unseen noise types, which so far is oftentimes reported to be a major issue in noise reduction by neural networks.

This paper is structured as follows: Section 2 describes our speech enhancement system and our novel CI-DNN architecture, along with training and testing aspects of the CI-DNN. The experimental setup as well as the results and discussion are presented in Section 3. We conclude the paper in Section 4.

II. Concatenated Identical DNN

A. Basic DNN Module and New CI-DNNs

We assume the single-channel mixture $y(n) = s(n) + d(n)$ of the clean speech signal $s(n)$ and the added noise signal $d(n)$ with n being the discrete-time sample index. Our speech enhancement system operates in the discrete Fourier transform (DFT) domain. Therefore, let $Y_\ell(k)$, $S_\ell(k)$, and $D_\ell(k)$ be the respective DFTs, and $|Y_\ell(k)|$, $|S_\ell(k)|$, and $|D_\ell(k)|$ be their DFT magnitudes, with frame index $\ell \in \mathcal{L} = \{1, 2, \ldots, L\}$ and frequency bin index $k \in \mathcal{K} = \{0, 1, \ldots, K-1\}$ with DFT size K. In this paper, we only enhance the magnitude spectrogram of the noisy speech and use the unaltered noisy speech phase for reconstruction. Then, we can write our task as

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot M_\ell(k), \tag{1}$$

with $M_\ell(k) \in [0, 1]$ and $\hat{S}_\ell(k)$ being the real-valued spectral mask and the estimated enhanced speech spectrum, respectively. As proposed in [15], [16], we predict the unknown enhanced amplitude speech spectrum $|\hat{S}_\ell(k)|$, while estimating the spectral mask $M_\ell(k)$ implicitly as shown in Fig. 2. The "NORM" operation in Fig. 2 represents a zero-mean and unit-variance normalization over the frame direction based on statistics collected on the training set, which are also used in the test phase.

Based on this single-stage feedforward basic DNN module, we build a speech enhancement system using the newly proposed serially concatenated identical DNN (CI-DNN) structure as shown in Fig. 1. Both topology and weights of each DNN enhancement stage are the same, in some more detail depicted as shown in Fig. 2. The idea is to train a single basic DNN module which can offer some moderate enhancement of the input, in our case a 5 dB SNR improvement. Then we concatenate the same module several times in so-called *stages* with stage index $r \in \mathcal{R} = \{1, 2, \ldots, R\}$ without any additional re-training. During inference, the DNN enhancement stage will enhance the input of stage r, and the output of stage r serves as respective input for stage $r+1$. Hence, we divide our speech enhancement task into multiple sub-tasks, where the total number of stages can be decided by the total target SNR improvement. In this work, each stage is designed to improve the SNR by 5 dB, so the 2-stage and 3-stage CI-DNNs will ideally offer 10 dB and 15 dB SNR improvement, respectively. Another factor that can be influenced by the number of stages is the tradeoff between noise reduction and speech distortion. Being able to decide on the number of stages based on development set performance, without the need for retraining, makes our proposed CI-DNN very flexible to adapt for tasks with different requirements. The maximum number of stages we used for this work is $R = 3$. The final stage output $\hat{S}_{R,\ell}(k)$ is transformed to the time domain by IFFT and overlap add (OLA) as shown in Fig. 1.

B. New Approach Training

The most important aspect to train a basic DNN module that can be concatenated as shown in Fig. 1, is to make sure the basic DNN is trained under multiple input and output SNR conditions to enable it to operate well both in the first stage and in all subsequent stages.

To train our basic module shown in Fig. 2, we use the input noisy spectrogram $Y_\ell(k)$ in six SNR levels ranging from −5 dB to 20 dB with a step size of 5 dB. The corresponding enhanced noisy targets $\hat{S}_\ell^{\text{target}}(k)$ have 5 dB higher SNR, which means the corresponding SNR levels range from 0 dB to 25 dB with the same 5 dB step size. The SNR level is measured according to ITU P.56 [24]. We define the loss function for each frame ℓ as

$$J_\ell = \frac{1}{K} \sum_{k \in \mathcal{K}} \left(|\hat{S}_\ell(k)| - |\hat{S}_\ell^{\text{target}}(k)| \right)^2, \tag{2}$$

with $|\hat{S}_\ell(k)|$ being calculated using (1). Given a DFT length of $K = 256$, the input size of the basic DNN module is $5 \times 129 = 645$, which includes 2 left and 2 right context frames. There are 5 hidden layers for each basic DNN module with a succeeding size of 1024 – 512 – 512 – 512 – 256. All these hidden layers use leaky rectified linear units (ReLU) as activation function and a dropout rate of $p = 0.2$. The size of the output layer is $\frac{K}{2} + 1 = 129$, determined by our target dimensionality. We use a sigmoid activation function for the output layer to make sure the value of the mask $M_\ell(k)$ is between 0 and 1. All possible forward residual skip connections are added to the layers with matched dimensions, which results to 3 bypasses in total to ease the vanishing gradient problem during training [25]. Batch normalization is used for each layer except for the input layer, and we use a minibatch size of 128 for all trainings.

C. New Approach Test

As shown in Fig. 1, the input noisy speech spectrum $Y_\ell(k)$ will be enhanced progressively as

$$\hat{S}_{R,\ell}(k) = Y_\ell(k) \cdot \prod_{r=1}^{R} M_{r,\ell}(k), \tag{3}$$

with $M_{r,\ell}(k)$ being the estimated mask in stage r. Since the identical basic DNN module uses the same number of context frames in each stage, an additional amount of context is needed the more stages are employed. As an example, the 2-stage CI-DNN needs 9 frames at the input of the first stage, which includes 4 left and 4 right context frames to produce one output frame at the final stage. This number will increase to 13 frames including 6 left and 6 right context frames for a 3-stage CI-DNN.

III. Experimental Validation

A. Database and Measures

The clean speech data for our training and test is taken from the Grid Corpus [26]. To make our CI-DNNs *speaker-independent*,

we randomly select 16 speakers, containing 8 male and 8 female speakers, and use 160 sentences per speaker for the basic DNN module training. For evaluation, four different speakers are chosen, two male and two female, with 10 sentences each.

Three types of superimposed noise are used to construct the training data: Pedestrian noise (PED), café noise (CAFE), and street noise (STR) which are obtained from the CHiME-3 dataset [27]. We train the basic DNN module under multiple input and output SNR conditions containing 6 different SNR levels. Thus, the training material consists of $16 \times 160 \times 3 \times 6 = 46080$ sentences. From the overall training material, 20% of the data is used for validation and 80% is used for actual training. All the speech and noise signals have a sampling rate of 16 kHz and are transferred to the DFT domain with $K = 256$ using a periodic Hann window with 50% overlap.

The test data is constructed using PED and CAFE noise, however, extracted from different files. Speech material is from unseen speakers. To additionally perform a *noise-type independent* test, we also create test data using bus noise (BUS), taken also from the CHiME-3 data, with this noise type not being seen during training. All the test data sets contain SNR levels from -5 dB to 20 dB with a step size of 5 dB. The evaluation is based on both the *filtered* clean speech component $\tilde{s}(n)$, the *filtered* noise component $\tilde{d}(n)$, and also the enhanced speech signal $\hat{s}(n)$. Using (3), $\tilde{S}(\ell,k)$ and $\tilde{D}(\ell,k)$ are obtained, replacing $Y_\ell(k)$ by $S_\ell(k)$ and $D_\ell(k)$, respectively.

In this paper, we use the following measures [28]:

1) SNR improvement: $\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}$, measured in dB

2) Speech quality (PESQ MOS-LQO) is measured using $s(n)$ as reference signal and either the *filtered* clean speech component $\tilde{s}(n)$ or the enhanced speech $\hat{s}(n)$ as test signal according to [29], [30], being referred to as $\text{PESQ}(\tilde{s})$ and $\text{PESQ}(\hat{s})$, respectively.

3) Segmental speech-to-speech-distortion ratio (SSDR):

$$\text{SSDR} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \text{SSDR}(\ell) \qquad [\text{dB}]$$

with $\mathcal{L}_1 \subset \mathcal{L}$, denoting the set of speech-active frames [28], and using $\text{SSDR}(\ell) = \max\{\min\{\text{SSDR}'(\ell), 30\,\text{dB}\}, -10\,\text{dB}\}$, with $\text{SSDR}'(\ell) = 10\log_{10}\Big(\big(\sum_{n\in\mathcal{N}_\ell} s^2(n)\big)/\big(\sum_{n\in\mathcal{N}_\ell} [\tilde{s}(n+\Delta) - s(n)]^2\big)\Big)$, with $\mathcal{N}_\ell$ denoting the sample indices n in frame ℓ, and Δ being used to perform time alignment for the filtered signal $\tilde{s}(n)$.

4) The weighted log-average kurtosis ratio (WLAKR) measures the noise distortion (especially for musical tones) using $d(n)$ as reference signal and the *filtered* noise component $\tilde{d}(n)$ as test signal according to ITU P.1130 [31], [32]. A WLAKR score that is closer to zero indicates less noise distortion, whereas being far away (+ or -) from zero indicates strong noise distortion [32]. In our analysis we will show averaged *absolute* WLAKR values.

5) Short-time objective intelligibility (STOI) measures the intelligibility of the enhanced speech as proposed in [33].

We group these measurements to noise component measures (ΔSNR and WLAKR), speech component ones (SSDR and $\text{PESQ}(\tilde{s})$), and total performance measures ($\text{PESQ}(\hat{s})$ and STOI).

B. Baseline Methods

The baseline methods include the classical LSA, SG, and WF spectral weighting rules combined with the DD approach for *a priori* SNR estimation and minimum statistics (MS) [34] for noise power estimation as mentioned before. To compare with the conventional mask estimation methods, we also train a so-called single DNN for speech enhancement similar to Fig. 2, but potentially deeper. We construct three single DNNs named 1 stage (same as CI-DNN, 1.23M weights), "2 stage" (2.43M weights), and "3 stage" (4.55M weights). These three single DNNs have a similar number of weights compared to the 1-stage, 2-stage, and 3-stage CI-DNNs, respectively, but all the weights in these baseline single DNNs are free weights that are trainable.

As mentioned before, the number of input frames required to obtain one output frame is different for CI-DNNs with different numbers of stages. To allow a fair comparison, the input size of the "2 stage" and "3 stage" single large DNNs are $9 \times 129 = 1161$ with 4 left and 4 right context frames, and $13 \times 129 = 1677$ with 6 left and 6 right context frames, respectively.

There are 6 hidden layers for the "2 stage" single large DNN with sizes of $1400-800-512-512-512-256$. The "3 stage" single large DNN contains 7 hidden layers with sizes of $1800-750-512-512-512-512-256$. Both networks have an output layer size of 129 using sigmoid activation functions. Except for the output layer, all layers use the same leaky ReLU activation functions. The same dropout rate $p = 0.2$ is used in all hidden layers and batch normalization is used except for the input layer. All possible bypasses are employed, which results in a total of 3 and 6 bypasses for "2 stage" and "3 stage", respectively. The 1 stage network has the same structure as our basic DNN module with identical weights.

We train these three baseline single DNNs using either clean speech or noisy speech as targets, separately. For the noisy target training, the target noisy speech provides 5 dB, 10 dB, and 15 dB SNR improvement for the 1 stage, "2 stage", and "3 stage" single DNNs, respectively.

C. Experimental Results and Discussion

We report on PED noise, CAFE noise, and on unseen BUS noise separately, both for all baselines and our proposed CI-DNNs. The measures are averaged over all speakers and all SNR levels, and are shown in Tables I, II, and III. Additionally, we report the performance at SNR = -5 dB, where results are shown in Tables IV, V, and VI. In each column, the **two best results** are **greyshaded**.

First, we look at PED noise and CAFE noise which are the *seen* noise types (not seen noises!) in CI-DNN and single DNN training.

The classical spectral weighting rule baselines LSA, SG, WF are strong in speech component quality, particularly at low SNR, which is reflected in very good $\text{PESQ}(\tilde{s})$ performance as shown in Table IV and V. However, they are all very bad in terms of residual noise quality (the by far highest WLAKR values for this measurement in all tables), and in terms of speech intelligibility, which shows an almost 0.1 points lower STOI score compared to the other methods.

Inspecting the single DNN trained with clean speech target, we find that this method always shows strong performance in ΔSNR and in total quality measures (see Tables I, II, IV, and V), but low fidelity of the speech component measures (SSDR and $\text{PESQ}(\tilde{s})$), interestingly. Although $\text{PESQ}(\hat{s})$ accepts this due to the high noise suppression, the clean target training seems to provide unbalanced results. Moreover, the residual noise quality, particularly at low SNRs, suffers audibly, which is also reflected by more than ten times higher WLAKR scores compared to other DNN-based or CI-DNN-based methods.

Now, we discuss the single DNN trained with noisy speech targets and our proposed CI-DNN (comparisons between DNNs and CI-DNNs with a similar number of weights). The single DNN trained with noisy target shows very good STOI, speech and noise component quality — as is the case with our new CI-DNN. Regarding the speech component measures, the noisy-target single DNN is a bit better in SSDR, while CI-DNN is a bit better in $\text{PESQ}(\tilde{s})$. The noisy-target single DNN provides a bit more ΔSNR in very noisy condition and on average, but not consistently so. Finally, concerning total

TABLE I
PERFORMANCE FOR PED NOISE, **ALL SNRS** AVERAGED; ΔSNR AND SSDR ARE MEASURED IN DB. **TWO BEST ARE GREYSHADED**.

Method		Noise Component		Speech Component		Total	
		ΔSNR	WLAKR	SSDR	PESQ($\hat{s}$)	PESQ($\hat{s}$)	STOI
LSA		3.08	0.66	**15.05**	3.40	2.09	0.62
SG		2.73	0.70	**15.33**	3.37	2.06	0.62
WF		3.85	0.75	13.52	3.45	2.09	0.61
Single DNN clean target	1 stage	6.67	0.11	13.24	3.20	2.45	**0.72**
	"2 stage"	**6.71**	0.28	13.33	3.20	**2.55**	**0.72**
	"3 stage"	**6.78**	0.20	13.46	3.21	**2.51**	**0.72**
Single DNN noisy target	1 stage	2.99	**0.02**	14.24	**3.52**	2.11	0.70
	"2 stage"	5.15	**0.02**	13.87	**3.48**	2.25	**0.71**
	"3 stage"	6.40	0.05	13.73	3.40	2.37	**0.72**
New CI-DNN ΔSNR target: +5dB for each stage	1 stage	2.99	**0.02**	14.24	**3.52**	2.11	0.70
	2 stage	5.07	**0.02**	13.58	**3.52**	2.28	**0.71**
	3 stage	6.03	**0.03**	12.74	3.47	2.43	**0.71**

TABLE II
PERFORMANCE FOR CAFE NOISE, **ALL SNRS** AVERAGED; ΔSNR AND SSDR ARE MEASURED IN DB. **TWO BEST ARE GREYSHADED**.

Method		Noise Component		Speech Component		Total	
		ΔSNR	WLAKR	SSDR	PESQ($\hat{s}$)	PESQ($\hat{s}$)	STOI
LSA		3.48	0.61	**14.43**	3.40	2.17	0.63
SG		3.29	0.65	**14.65**	3.38	2.14	0.62
WF		4.30	0.68	12.90	**3.48**	2.18	0.62
Single DNN clean target	1 stage	5.73	0.11	12.87	3.22	2.52	**0.71**
	"2 stage"	**5.99**	0.13	12.94	3.19	**2.58**	**0.71**
	"3 stage"	**6.01**	0.11	13.02	3.19	**2.55**	**0.72**
Single DNN noisy target	1 stage	2.72	**0.02**	13.71	**3.48**	2.19	**0.71**
	"2 stage"	4.62	**0.02**	13.39	**3.45**	2.33	**0.71**
	"3 stage"	5.78	0.04	13.29	3.37	2.44	**0.72**
New CI-DNN ΔSNR target: +5dB for each stage	1 stage	2.72	**0.02**	13.71	**3.48**	2.19	**0.71**
	2 stage	4.54	**0.03**	13.04	**3.48**	2.36	**0.71**
	3 stage	5.97	**0.03**	12.17	3.41	2.50	**0.71**

TABLE III
PERFORMANCE FOR **UNSEEN** BUS NOISE, **ALL SNRS** AVERAGED; ΔSNR AND SSDR ARE MEASURED IN DB. **TWO BEST ARE GREYSHADED**.

Method		Noise Component		Speech Component		Total	
		ΔSNR	WLAKR	SSDR	PESQ($\hat{s}$)	PESQ($\hat{s}$)	STOI
LSA		4.68	0.56	**16.08**	3.57	**2.71**	0.72
SG		3.57	0.63	**16.21**	**3.62**	**2.69**	0.72
WF		**6.24**	0.68	14.62	**3.76**	**2.71**	0.71
Single DNN clean target	1 stage	5.01	0.06	15.42	3.36	2.58	**0.77**
	"2 stage"	3.95	0.17	15.45	3.33	2.57	**0.77**
	"3 stage"	3.78	0.13	15.54	3.34	2.55	**0.77**
Single DNN noisy target	1 stage	1.90	**0.03**	15.96	3.47	2.35	**0.77**
	"2 stage"	2.90	0.05	15.78	3.46	2.48	**0.77**
	"3 stage"	3.67	0.08	15.71	3.41	2.54	**0.77**
New CI-DNN ΔSNR target: +5dB for each stage	1 stage	1.90	**0.03**	15.96	3.47	2.35	**0.77**
	2 stage	4.02	**0.04**	15.18	3.49	2.56	**0.77**
	3 stage	**5.86**	0.05	14.14	3.44	**2.71**	**0.76**

TABLE IV
PERFORMANCE FOR PED NOISE AT **SNR**= -5 **DB**; ΔSNR AND SSDR ARE MEASURED IN DB. **TWO BEST ARE GREYSHADED**.

Method		Noise Component		Speech Component		Total	
		ΔSNR	WLAKR	SSDR	PESQ($\hat{s}$)	PESQ($\hat{s}$)	STOI
LSA		2.22	0.66	**5.86**	2.54	1.34	0.41
SG		2.02	0.66	**5.47**	**2.72**	1.31	0.40
WF		2.41	0.75	4.42	**2.87**	1.33	0.41
Single DNN clean target	1 stage	6.16	0.15	3.96	2.21	1.57	**0.54**
	"2 stage"	**7.00**	0.27	4.02	2.20	**1.56**	**0.54**
	"3 stage"	**7.21**	0.30	4.06	2.18	**1.57**	**0.55**
Single DNN noisy target	1 stage	2.18	**0.02**	4.88	2.51	1.41	**0.54**
	"2 stage"	4.57	**0.02**	4.53	2.46	1.45	**0.55**
	"3 stage"	6.26	**0.02**	4.26	2.39	1.49	**0.55**
New CI-DNN ΔSNR target: +5dB for each stage	1 stage	2.18	**0.02**	4.88	2.51	1.41	**0.54**
	2 stage	4.40	**0.03**	4.51	2.58	1.48	**0.55**
	3 stage	6.26	0.05	4.02	2.58	1.54	**0.54**

TABLE V
PERFORMANCE FOR CAFE NOISE AT **SNR**= -5 **DB**; ΔSNR AND SSDR ARE MEASURED IN DB. **TWO BEST ARE GREYSHADED**.

Method		Noise Component		Speech Component		Total	
		ΔSNR	WLAKR	SSDR	PESQ($\hat{s}$)	PESQ($\hat{s}$)	STOI
LSA		2.84	0.59	**5.55**	2.60	1.43	0.42
SG		3.17	0.60	**5.08**	**2.75**	1.40	0.41
WF		3.14	0.65	4.19	**2.99**	1.41	0.41
Single DNN clean target	1 stage	5.87	0.13	3.68	2.20	1.59	0.53
	"2 stage"	**6.28**	0.20	3.75	2.15	**1.58**	0.54
	"3 stage"	**6.50**	0.24	3.75	2.15	**1.59**	0.54
Single DNN noisy target	1 stage	1.99	**0.02**	4.38	2.45	1.43	**0.55**
	"2 stage"	4.11	**0.02**	4.18	2.40	1.47	**0.56**
	"3 stage"	5.76	**0.01**	4.02	2.33	1.52	**0.56**
New CI-DNN ΔSNR target: +5dB for each stage	1 stage	1.99	**0.02**	4.38	2.45	1.43	**0.55**
	2 stage	3.76	0.04	4.11	2.50	1.50	**0.55**
	3 stage	5.35	0.04	3.73	2.51	1.57	**0.55**

TABLE VI
PERFORMANCE FOR **UNSEEN** BUS NOISE AT **SNR**= -5 **DB**; ΔSNR AND SSDR ARE MEASURED IN DB. **TWO BEST ARE GREYSHADED**.

Method		Noise Component		Speech Component		Total	
		ΔSNR	WLAKR	SSDR	PESQ($\hat{s}$)	PESQ($\hat{s}$)	STOI
LSA		5.10	0.59	**8.20**	3.03	**1.68**	0.56
SG		5.09	0.60	**7.75**	**3.22**	1.64	0.56
WF		**6.42**	0.69	6.63	**3.37**	**1.69**	0.55
Single DNN clean target	1 stage	6.07	**0.01**	7.09	2.55	1.64	**0.65**
	"2 stage"	4.47	0.15	7.11	2.43	1.57	0.64
	"3 stage"	4.73	0.11	7.19	2.43	1.57	0.64
Single DNN noisy target	1 stage	2.14	**0.01**	7.42	2.59	1.40	**0.65**
	"2 stage"	3.48	**0.03**	7.36	2.60	1.48	**0.65**
	"3 stage"	4.57	0.07	7.30	2.56	1.55	**0.65**
New CI-DNN ΔSNR target: +5dB for each stage	1 stage	2.14	**0.01**	7.42	2.59	1.40	**0.65**
	2 stage	4.46	**0.03**	7.34	2.72	1.51	**0.66**
	3 stage	**6.50**	0.04	7.00	2.75	1.66	**0.65**

PESQ($\hat{s}$), CI-DNN is always (at low SNR, and average over SNR conditions) *slightly ahead* of the DNN with noisy-target training (for two or more stages, of course). Our proposed CI-DNNs only have $1/2$ or $1/3$ of the trainable weights compared to the noisy-target single DNN to achieve this performance. In summary, *for noise types that have been seen in training, the CI-DNN is overall slightly ahead of any here investigated baseline single DNN in terms of total speech quality.*

Secondly, we look at the measurement results for the *unseen* noise type BUS. As expected, the classical spectral weighting rule baselines perform similarly bad in BUS noise concerning background noise quality (WLAKR scores) and total speech intelligibility (STOI). For the neural network approaches, the BUS noise type has not been seen in training. They all perform nicely and equally well in STOI as shown in Tables III and VI. The relative performance of the DNN trained with noisy targets vs. the DNN trained with clean targets is the same, whether we test with seen or unseen noise types, again disqualifying the clean target training due to its bad residual noise quality performance.

Comparing the noisy-target single DNN to the new CI-DNN in the unseen BUS noise type, however, we make surprising observations: The speech component quality is roughly comparable as before (SSDR better for DNN with noisy target, PESQ($\hat{s}$) better with CI-DNN), so is also the noise component quality (WLAKR). *However, the 3-stage CI-DNN clearly excels the respective single DNN in ΔSNR. In the 3-stage case, the total PESQ($\hat{s}$) of the CI-DNN is on average over all SNRs by 0.17 points better than that for the single DNN.*

IV. CONCLUSIONS

In this paper, we proposed serially concatenated identical DNNs (CI-DNNs), where each basic DNN module (stage) can offer some moderate enhancement of the input. Our proposed CI-DNNs outperform the classical spectral weighting rules both in total speech quality and speech intelligibility. The CI-DNN also shows more

balanced performance than the conventional clean-target single DNN. Comparing with the noisy-target single DNN, our proposed CI-DNN offers quite similar or even a bit better performance concerning total PESQ($\hat{s}$), but with only $1/2$ or $1/3$ of the trainable weights. Under a comparable noise and speech *component* quality, our proposed CI-DNNs also generalize better to an unseen noise type by offering higher total PESQ($\hat{s}$) and SNR improvement.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE T-ASSP*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE T-ASSP*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[3] P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP*, Atlanta, GA, USA, May 1996, pp. 629–632.

[4] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[5] I. Cohen, "Speech Enhancement Using Super-Gaussian Speech Models and Noncausal A Priori SNR Estimation," *Speech Commun.*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[6] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved A Posteriori Speech Presence Probability Estimation Based on a Likelihood Ratio with Fixed Priors," *IEEE T-ASLP*, vol. 16, no. 5, pp. 910–919, July 2008.

[7] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE T-ASLP*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[8] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 1740–1744.

[9] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking Speech-presence Uncertainty to Improve Speech Enhancement in Non-stationary Noise Environments," in *Proc. of ICASSP*, Phoenix, AZ, USA, Mar. 1999, pp. 789–792.

[10] J. S. Goldstein, I. S. Reed, and L. L. Scharf, "A Multistage Representation of the Wiener Filter Based on Orthogonal Projections," *IEEE Trans. on Inf. Theory*, vol. 44, no. 7, pp. 2943–2959, Nov. 1998.

[11] M. Tinston and Y. Ephraim, "Speech Enhancement Using the Multistage Wiener Filter," in *Proc. of IEEE Conf. on Information Sciences and Systems*, Baltimore, MD, USA, Mar. 2009, pp. 55–60.

[12] J. Du, Y. Tu, L. R. Dai, and C. H. Lee, "A Regression Approach to Single-Channel Speech Separation via High-Resolution Deep Neural Networks," *IEEE/ACM T-ASLP*, vol. 24, no. 8, pp. 1424–1437, Apr. 2016.

[13] Y. Wang, A. Narayanan, and D. L. Wang, "On Training Targets for Supervised Speech Separation," *IEEE/ACM T-ASLP*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[14] Y. Wang and D. L. Wang, "A Deep Neural Network for Time-Domain Signal Reconstruction," in *Proc. of ICASSP*, Brisbane, QLD, Australia, Aug. 2015, pp. 4390–4394.

[15] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-Sensitive and Recognition-Boosted Speech Separation Using Deep Recurrent Neural Networks," in *Proc. of ICASSP*, Brisbane, QLD, Australia, Aug. 2015, pp. 708–712.

[16] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively Trained Recurrent Neural Networks for Single-Channel Speech Separation," in *Proc. of 2nd IEEE GlobalSIP*, Atlanta, GA, USA, May 2014, pp. 577–581.

[17] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex Ratio Masking for Monaural Speech Separation," *IEEE/ACM T-ASLP*, vol. 24, no. 3, pp. 483–492, Mar. 2016.

[18] 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, Nov. 2014.

[19] Y. Wang and D. L. Wang, "Towards Scaling up Classification-Based Speech Separation," *IEEE T-ASLP*, vol. 21, no. 7, pp. 1381–1390, Mar. 2013.

[20] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," *IEEE/ACM T-ASLP*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

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

[22] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement.," in *Proc. of Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 3713–3717.

[23] X. Shu, Y. Zhou, and Y. Cao, "A New Speech Enhancement Approach Based on Progressive Deep Neural Networks," in *Proc. of IWAENC*, Tokyo, Japan, Sep. 2018, pp. 191–195.

[24] ITU, *Objective Measurement of Active Speech Level*, International Telecommunication Standardization Sector (ITU-T), Rec. P.56, Dec. 2011.

[25] A. Veit, M. J. Wilber, and S. Belongie, "Residual Networks Behave Like Ensembles of Relatively Shallow Networks," in *Proc. of NIPS*, Barcelona, Spain, Dec. 2016, pp. 550–558.

[26] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, Jun. 2006.

[27] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The Third 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines," in *Proc. of ASRU*, Scottsdale, AZ, USA, Feb. 2015, pp. 504–511.

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

[29] ITU, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, International Telecommunication Standardization Sector (ITU-T), Rec. P.862, Feb. 2001.

[30] ITU, *Wideband Hands-Free Communication in Motor Vehicles*, International Telecommunication Standardization Sector (ITU-T), Rec. P.1110, Jan. 2015.

[31] ITU, *Subsystem Requirements for Automotive Speech Services*, International Telecommunication Standardization Sector (ITU-T), Rec. P.1130, Jun. 2015.

[32] H. Yu and T. Fingscheidt, "A Figure of Merit for Instrumental Optimization of Noise Reduction Algorithms," in *Proc. of 5th Biennial Workshop on DSP for In-Vehicle Systems*, Kiel, Germany, Sep. 2011, pp. 1–8.

[33] 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, Jun. 2010, pp. 4214–4217.

[34] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE T-SAP*, vol. 9, no. 5, pp. 504–512, Jul. 2001.