Improving GANs for Speech Enhancement

Huy Phan*, Ian V. McLoughlin, Lam Pham, Oliver Y. Chén, Philipp Koch, Maarten De Vos, Alfred Mertins

Abstract—Generative adversarial networks (GAN) have recently been shown to be efficient for speech enhancement. However, most, if not all, existing speech enhancement GANs (SEGAN) make use of a single generator to perform one-stage enhancement mapping. In this work, we propose to use multiple generators that are chained to perform multi-stage enhancement mapping, which gradually refines the noisy input signals in a stage-wise fashion. Furthermore, we study two scenarios: (1) the generators share their parameters and (2) the generators' parameters are independent. The former constrains the generators to learn a common mapping that is iteratively applied at all enhancement stages and results in a small model footprint. On the contrary, the latter allows the generators to flexibly learn different enhancement mappings at different stages of the network at the cost of an increased model size. We demonstrate that the proposed multi-stage enhancement approach outperforms the one-stage SEGAN baseline, where the independent generators lead to more favorable results than the tied generators. The source code is available at http://github.com/pquochuy/idsegan.

Index Terms—speech enhancement, generative adversarial networks, SEGAN, ISEGAN, DSEGAN

I. INTRODUCTION

The goal of speech enhancement is to improve the quality and intelligibility of speech which are degraded by background noise [1], [2]. Speech enhancement can serve as a front-end to improve performance of an automatic speech recognition system [3]. It also plays an important role in applications like communication systems, hearing aids, and cochlear implants in which contaminated speech needs to be enhanced prior to signal amplification to reduce discomfort [2]. Significant progress on this research topic has been made with the involvement of deep learning paradigms. Deep neural networks (DNNs) [4], [5], convolutional neural networks (CNNs) [6], [7], and recurrent neural networks (RNNs) [3], [8] have been exploited either to produce the enhanced signal directly via a regression form [4], [6] or to estimate the contaminating noise, which is subtracted from the noisy signal to obtain the enhanced signal [7]. Significant improvements on speech enhancement performance have been reported by these deep-learning based methods over more conventional ones, such as Wiener filtering [9], spectral subtraction [10] or minimum mean square error (MMSE) estimation [11], [12].

There exists a class of generative methods relying on GANs [13], which have been demonstrated to be efficient for speech enhancement [14]–[19]. When GANs are used for this task, the enhancement mapping is accomplished by the generator G whereas the discriminator D, by discriminating between real
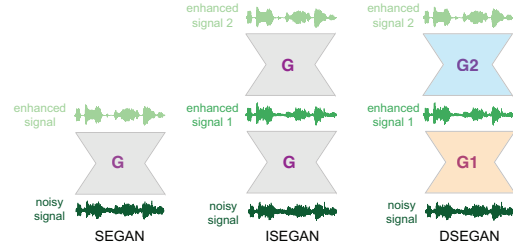
Fig. 1: Illustration of SEGAN with a single generator G, ISEGAN ($N = 2$) with the shared generators G, and DSEGAN ($N = 2$) with two independent generators $G1$ and $G2$.

and fake signals, transmits information to G so that G can learn to produce output that resembles the realistic distribution of the clean signals. Using GANs, speech enhancement has been done using either magnitude spectrum input [18] or raw waveform input [14], [15].

Existing speech enhancement GAN (SEGAN) systems share a common feature – the enhancement mapping is accomplished via a single stage by a single generator G [14], [15], [18], which may not be optimal. Here, we aim to divide the enhancement process into multiple stages and accomplish it via multiple enhancement mappings, one at each stage. Each of the mappings is realized by a generator, and the generators are chained to enhance a noisy input signal gradually, step by step, to yield an enhanced signal. By doing so, a generator is tasked to refine or correct the output produced by its predecessor. We hypothesize that it would be better to carry out multi-stage enhancement mapping rather than a single-stage one as in prior works [14], [15], [18]. We then propose two new SEGAN frameworks, namely iterated SEGAN (ISEGAN) and deep SEGAN (DSEGAN) as illustrated in Fig. 1, to study two scenarios: (1) using a common mapping for all the enhancement stages and (2) using independent mappings at different enhancement stages. In the former the generators' parameters are tied and parameter sharing constrains ISEGAN's generators to learn a common mapping (i.e. the generators apply the same mapping iteratively). The latter's generators have independent parameters, allowing them to learn different enhancement mappings flexibly. Note that, due to parameter sharing, ISEGAN's footprint is expected to be smaller than that of DSEGAN.

We will demonstrate that the proposed method obtains more favorable results than the SEGAN baseline [14] on both objective and subjective evaluation metrics and that learning independent mappings with DSEGAN leads to better performance than learning a common one with ISEGAN.

II. SEGAN

Given a dataset $\mathcal{X} = \{(\mathbf{x}_1, \tilde{\mathbf{x}}_1), (\mathbf{x}_2, \tilde{\mathbf{x}}_2), \ldots, (\mathbf{x}_N, \tilde{\mathbf{x}}_N)\}$ consisting of N pairs of raw signals: clean speech signal $\mathbf{x}$ and noisy speech signal $\tilde{\mathbf{x}}$, speech enhancement is to find a mapping $f(\tilde{\mathbf{x}}) : \tilde{\mathbf{x}} \mapsto \mathbf{x}$ to map the noisy signal $\tilde{\mathbf{x}}$ to the clean signal

x. Conforming to GAN's principle [13], SEGAN proposed in [14] has its generator G tasked for the enhancement mapping. Presented with the noisy signal $\tilde{\mathbf{x}}$ together with the latent representation $\mathbf{z}$, G produces the enhanced signal $\hat{\mathbf{x}} = G(\mathbf{z}, \tilde{\mathbf{x}})$. The discriminator D of SEGAN receives a pair of signals as input. D learns to classify the pair $(\mathbf{x}, \tilde{\mathbf{x}})$ as real and the pair $(\hat{\mathbf{x}}, \tilde{\mathbf{x}})$ as fake while G tries to fool D such that D classifies the pair $(\hat{\mathbf{x}}, \tilde{\mathbf{x}})$ as real. The objective function of SEGAN reads

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim p_{\text{data}}(\mathbf{x}, \tilde{\mathbf{x}})} \log D(\mathbf{x}, \tilde{\mathbf{x}})$$
$$+ \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}), \tilde{\mathbf{x}} \sim p_{\text{data}}(\tilde{\mathbf{x}})} \log(1 - D(G(\mathbf{z}, \tilde{\mathbf{x}}), \tilde{\mathbf{x}})). \quad (1)$$

To improve the stability, SEGAN further employs least-squares GAN (LSGAN) [20] to replace the discriminator D's cross-entropy loss by the least-square loss. The least-squares objective functions of D and G are explicitly written as

$$\min_D V_{\text{LS}}(D) = \frac{1}{2} \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim p_{\text{data}}(\mathbf{x}, \tilde{\mathbf{x}})} (D(\mathbf{x}, \tilde{\mathbf{x}}) - 1)^2$$
$$+ \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}), \tilde{\mathbf{x}} \sim p_{\text{data}}(\tilde{\mathbf{x}})} D(G(\mathbf{z}, \tilde{\mathbf{x}}), \tilde{\mathbf{x}})^2, \quad (2)$$
$$\min_G V_{\text{LS}}(G) = \frac{1}{2} \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z}), \tilde{\mathbf{x}} \sim p_{\text{data}}(\tilde{\mathbf{x}})} (D(G(\mathbf{z}, \tilde{\mathbf{x}}), \tilde{\mathbf{x}}) - 1)^2$$
$$+ \lambda ||G(\mathbf{z}, \tilde{\mathbf{x}}) - \mathbf{x}||_1, \quad (3)$$

respectively. In (3), ℓ_1 distance between the clean sample $\mathbf{x}$ and the generated sample $G(\mathbf{z}, \tilde{\mathbf{x}})$ is included to encourage the generator G to generate more fine-grained and realistic results [14], [21], [22]. The influence of the ℓ_1-norm term is regulated by the hyper-parameter λ which was set to $\lambda = 100$ in [14].

III. ITERATED SEGAN AND DEEP SEGAN

Quan *et al.* [23] showed that using an additional generator chained to the generator of a GAN leads to better image-reconstruction performance. In light of this, instead of using the single-stage enhancement mapping with one generator as in SEGAN, we propose to learn multiple mappings with a chain of N generators $\mathfrak{G} = G_1 \rightarrow G_2 \rightarrow \ldots \rightarrow G_N$ with $N > 1$ to perform multi-stage enhancement. We study both the cases when a common mapping is learned and shared by all the stages (i.e. ISEGAN) and when independent mappings are learned at different stages (i.e. DSEGAN). In ISEGAN, the generators share their parameters (i.e. they are realized by a common generator G) and can be viewed as an iterated generator with the *number of iterations* of N. In contrast, DSEGAN's generators are independent and can be viewed as a deep generator with the *depth* of N. ISEGAN and DSEGAN with $N = 2$ are illustrated alongside SEGAN in Fig. 1. Both ISEGAN and DSEGAN reduce to SEGAN when $N = 1$.

At the enhancement stage n, $1 \le n \le N$, the generator G_n receives the output $\hat{\mathbf{x}}_{n-1}$ of its predecessor G_{n-1} together with the latent representation $\mathbf{z}_n$ and is expected to produce a better enhanced signal $\hat{\mathbf{x}}_n$:

$$\hat{\mathbf{x}}_n = G_n(\mathbf{z}_n, \hat{\mathbf{x}}_{n-1}), \quad 1 \le n \le N. \quad (4)$$

Note that $\hat{\mathbf{x}}_0 \equiv \tilde{\mathbf{x}}$. The output of the last generator G_N is considered as the final enhanced signal, i.e. $\hat{\mathbf{x}} \equiv \hat{\mathbf{x}}_N$, which is expected to be of better quality than all the intermediate enhanced versions. The outputs of the generators can be interpreted as different checkpoints and by forcing the ground-truth between the checkpoints, we encourage the chained generators to produce gradually better enhancement results.
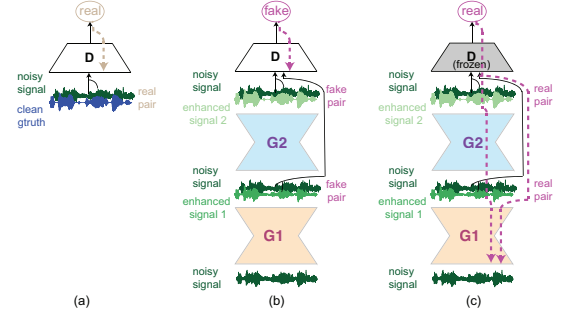

Fig. 2: Adversarial training with two generators. The discriminator D is learned to classify the pair $(\mathbf{x}, \tilde{\mathbf{x}})$ as real (a), and all the pairs $(\hat{\mathbf{x}}_1, \tilde{\mathbf{x}})$, $(\hat{\mathbf{x}}_2, \tilde{\mathbf{x}})$, $\ldots$, $(\hat{\mathbf{x}}_N, \tilde{\mathbf{x}})$ as fake (b). The generators G_1 and G_2 are learned to fool D so that D classifies the pairs $(\hat{\mathbf{x}}_1, \tilde{\mathbf{x}})$, $(\hat{\mathbf{x}}_2, \tilde{\mathbf{x}})$, $\ldots$, $(\hat{\mathbf{x}}_N, \tilde{\mathbf{x}})$ as real (c). Dashed lines represent the flow of gradient backdrop.

To enforce the generators in the chain $\mathfrak{G}$ to learn a proper mapping for signal enhancement, the discriminator D is tasked to classify the pair $(\mathbf{x}, \tilde{\mathbf{x}})$ as real while all N pairs $(\hat{\mathbf{x}}_1, \tilde{\mathbf{x}})$, $(\hat{\mathbf{x}}_2, \tilde{\mathbf{x}})$, $\ldots$, $(\hat{\mathbf{x}}_N, \tilde{\mathbf{x}})$ as fake, as illustrated in Fig. 2 for the case of $N = 2$. The least-squares objective functions of D and $\mathfrak{G}$ are given as

$$\min_D V_{\text{LS}}(D) = \frac{1}{2} \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}} \sim p_{\text{data}}(\mathbf{x}, \tilde{\mathbf{x}})} (D(\mathbf{x}, \tilde{\mathbf{x}}) - 1)^2$$
$$+ \sum_{n=1}^N \frac{1}{2N} \mathbb{E}_{\mathbf{z}_n \sim p_{\mathbf{z}}(\mathbf{z}), \tilde{\mathbf{x}} \sim p_{\text{data}}(\tilde{\mathbf{x}})} D(G_n(\mathbf{z}_n, \hat{\mathbf{x}}_{n-1}), \tilde{\mathbf{x}})^2, \quad (5)$$

$$\min_{\mathfrak{G}} V_{\text{LS}}(\mathfrak{G}) = \sum_{n=1}^N \frac{1}{2N} \mathbb{E}_{\mathbf{z}_n \sim p_{\mathbf{z}}(\mathbf{z}), \tilde{\mathbf{x}} \sim p_{\text{data}}(\tilde{\mathbf{x}})} (D(G_n(\mathbf{z}_n, \hat{\mathbf{x}}_{n-1}), \tilde{\mathbf{x}}) - 1)^2$$
$$+ \sum_{n=1}^N \lambda_n ||G_n(\mathbf{z}_n, \hat{\mathbf{x}}_{n-1}) - \mathbf{x}||_1. \quad (6)$$

Unlike SEGAN, the discriminator D in cases of ISEGAN and DSEGAN needs to handle imbalanced data as there are N fake examples generated with respect to every real example. Therefore, it is necessary to divide the second term in (5) by N to balance out penalization for real and fake examples misclassification. In addition, the first term in (6) is also divided by N to level its magnitude with that of the ℓ_1-norm term [14]. To regulate the enhancement curriculum in multiple stages, we set $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ to $(\frac{100}{2^{N-1}}, \ldots, \frac{100}{2^1}, \frac{100}{2^0})$. That is, λ_n is set to double λ_{n-1} while the last λ_N is fixed to 100 as in case of SEGAN. With this curriculum, we expect the enhanced output of a generator to be twice as good as that of its preceding generator in terms of ℓ_1-norm. As a result, the enhancement mapping learned by a generator in the chain doesn't need to be perfect as in single-stage enhancement since its output will be refined by its successor.

IV. NETWORK ARCHITECTURE

A. Generators G_n

The architecture of the generators G_n, $1 \le n \le N$, used in ISEGAN and DSEGAN is illustrated in Fig. 3. They make use of an encoder-decoder architecture with fully-convolutional layers [24], which is similar to that used in SEGAN. Each generator receives a segment of raw signal with a length of $L = 16384$ samples (approximately one second at 16 kHz) as input. The generators' encoder is composed of 11 one-dimensional strided convolutional layers with a common filter width of 31 and a stride length of 2, followed by parametric rectified linear units (PReLUs) [25]. The number of filters is designed to
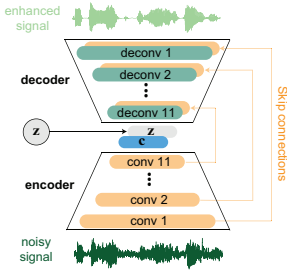
Fig. 3: The generator architecture used in ISEGAN and DSEGAN, featuring 11 strided convolutional layers in the encoder and 11 deconvolutional layers in the decoder.

increase along the encoder's depth to compensate for the smaller and smaller convolutional output, resulting in output sizes of 8192×16, 4096×32, 2048×32, 1024×64, 512×64, 256×128, 128×128, 64×256, 32×256, 16×512, 8×1024 at the 11 convolutional layers, respectively. At the end of the encoder, the encoding vector $\mathbf{c} \in \mathbb{R}^{8 \times 1024}$ is concatenated with the noise sample $\mathbf{z} \in \mathbb{R}^{8 \times 1024}$ sampled from the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and presented to the decoder. The generator's decoder mirrors the encoder architecture with the same number of filters and filter width (see Fig. 3) to reverse the encoding process by means of deconvolutions (i.e. fractional-strided transposed convolution). Note that each deconvolutional layer is again followed by a PReLU. The skip connections are employed to connect an encoding layer to its corresponding decoding layer to allow the information of the waveform to flow into the decoding stage [14].

B. Discriminator D

The discriminator D has similar architecture to the encoder part of the generators described in Section IV-A, except that it has two-channel input and uses virtual batch-norm [26] before LeakyReLU activation with $\alpha = 0.3$. In addition, D is topped up with a one-dimensional convolutional layer with one filter of width one (i.e. 1×1 convolution) to reduce the last convolutional output size from 8×1024 to 8 features before classification takes place with a softmax layer.

V. EXPERIMENTS

A. Dataset

To assess the performance of the proposed ISEGAN and DSEGAN and demonstrate their advantages over SEGAN, we conducted experiments on the database in [28] which was used to evaluate SEGAN in [14]. The database is originated from the Voice Bank corpus [29] and consists of data from 30 speakers. Following the database's original split, data from 28 speakers was used for training and data from two remaining speakers was used for testing.

A total of 40 noisy conditions was made in the training data by combining ten types of noises (two artificial and eight stemmed from the Demand database [30]) with four signal-to-noise ratios (SNRs) each: 15, 10, 5, and 0 dB. For the test data, 20 noisy conditions were created, combining five types of noise from the Demand database with four SNRs each: 17.5, 12.5, 7.5, and 2.5 dB. There are about 10 and 20 utterances for each noisy condition per speaker in the training and test set, respectively. All utterances were downsampled to 16 kHz.

B. Baseline system

SEGAN was used as a baseline for comparison. We repeated training SEGAN to ensure a similar experimental setting across systems. In addition, to shed some light on how generative models like ISEGAN and DSEGAN perform on the speech enhancement task in relation to discriminative models, we also compared the proposed method to two discriminative deep learning methods: (1) the popular DNN proposed in [4] and (2) the two-stage network (TSN) recently proposed in [27]. The DNN baseline was implemented based on [4], but with three main modifications: (a) wideband operation (16 kHz, leading to doubling of the feature dimension), (b) smaller frame size and shift (25 ms and 10 ms, respectively), and (c) use of the Adam optimizer [31] and simplified training (i.e. without unsupervised pre-training). In addition, early stopping was carried out during training via a leave-out validation set (10% of the training data). While these modifications may lead to a better baseline, they also allow a fair comparison with the SEGAN-based systems. The TSN baseline was configured based on [27], except for the use of wideband speech. For both the baselines, the features (log-power spectra) were normalized at utterance level to zero mean and unit standard deviation. De-normalization was then performed before waveform reconstruction.

C. Network parameters

The implementation was based on Tensorflow framework [32]. The networks were trained for 100 epochs with RMSprop optimizer [33] and a learning rate of 0.0002. The SEGAN baseline was trained with a minibatch size of 100 while it was reduced to 50 to train ISEGAN and DSEGAN to cope with their larger memory footprints. We experimented with different values for $N = \{2, 3, 4\}$ to investigate the influence of the number of iterations of ISEGAN and the depth of DSEGAN.

As in [14], during training, raw speech segments of length 16384 samples were extracted from the training utterances with 50% overlap. A high-frequency preemphasis filter of coefficient 0.95 was applied to each signal segment before presenting to the networks. During testing, raw speech segments were extracted from a test utterance without overlap. They were processed by a trained network, deemphasized, and eventually concatenated to produce the enhanced utterance.

D. Objective evaluation

We quantified the quality of the enhanced signals based on five objective signal-quality metrics, including PESQ, CSIG, CBAK, COVL, and SSNR, as suggested in [1] and the speech-intelligibility measure STOI [34]. The tool used for computing the first five metrics is based on the implementation in [1]. This is also the one used in [14]. The metrics were computed for each system by averaging over all 824 files of the test set. Since we found that the performance may vary with different network checkpoints, the mean and standard deviation of each metric over the 5 latest network checkpoints are reported.

The objective evaluation results are shown in Table I. As expected, SEGAN enhances the noisy signals to result in speech signals with better quality and intelligibility, evidenced by its better results across the objective metrics compared to those measured from the noisy signals. In comparison to SEGAN, on the one hand, ISEGAN performs comparably in terms of speech-quality metrics, slightly surpassing the baseline in PESQ, CBAK, and SSNR (i.e. with $N = 2$ and $N = 4$) but marginally underperforming in CSIG and COVL. On the other hand, DSEGAN obtains the best results, consistently outperforming both SEGAN and ISEGAN across all the speech quality metrics. For example, with $N = 2$, DSEGAN leads to relative improvements of 7.3%, 4.7%, 6.9%, 6.2%, and

TABLE I: Results obtained by the studied speech enhancement systems on the objective evaluation metrics.

| Metric	Noisy	DNN [4]	TSN [27]	SEGAN	ISEGAN			DSEGAN		
					$N = 2$	$N = 3$	$N = 4$	$N = 2$	$N = 3$	$N = 4$
PESQ	1.97	2.45	2.68	2.19 ± 0.04	$\mathbf{2.24 \pm 0.05}$	2.19 ± 0.04	$\mathbf{2.21 \pm 0.06}$	$\mathbf{2.35 \pm 0.06}$	$\mathbf{2.39 \pm 0.02}$	$\mathbf{2.37 \pm 0.05}$
CSIG	3.35	3.73	3.96	3.39 ± 0.03	3.23 ± 0.10	2.96 ± 0.08	3.00 ± 0.14	$\mathbf{3.55 \pm 0.06}$	$\mathbf{3.46 \pm 0.05}$	$\mathbf{3.50 \pm 0.01}$
CBAK	2.44	2.89	2.94	2.90 ± 0.07	$\mathbf{2.95 \pm 0.07}$	2.88 ± 0.12	$\mathbf{2.92 \pm 0.06}$	$\mathbf{3.10 \pm 0.02}$	$\mathbf{3.11 \pm 0.05}$	$\mathbf{3.10 \pm 0.04}$
COVL	2.63	3.09	3.32	2.76 ± 0.03	2.69 ± 0.05	2.52 ± 0.04	2.55 ± 0.09	$\mathbf{2.93 \pm 0.05}$	$\mathbf{2.90 \pm 0.03}$	$\mathbf{2.92 \pm 0.02}$
SSNR	1.68	3.64	2.89	7.36 ± 0.72	$\mathbf{8.17 \pm 0.69}$	8.11 ± 1.43	$\mathbf{8.86 \pm 0.42}$	$\mathbf{8.70 \pm 0.34}$	$\mathbf{8.72 \pm 0.64}$	$\mathbf{8.59 \pm 0.49}$
STOI	92.10	89.14	92.52	93.12 ± 0.17	$\mathbf{93.29 \pm 0.16}$	93.35 ± 0.08	$\mathbf{93.29 \pm 0.19}$	$\mathbf{93.25 \pm 0.17}$	$\mathbf{93.28 \pm 0.17}$	$\mathbf{93.49 \pm 0.09}$
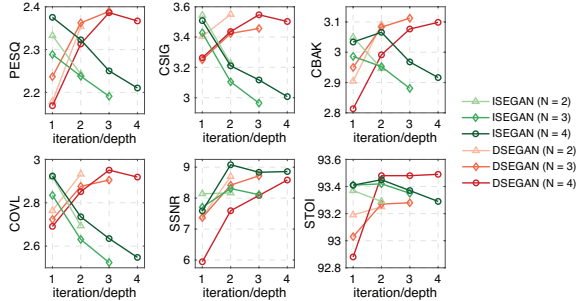
Fig. 4: Evolution of the evaluation metrics along the depth and iteration of DSEGAN and ISEGAN, respectively.

18.2% over the baseline on PESQ, CSIG, CBAK, COVL, and SSNR, respectively. In terms of speech intelligibility, ISEGAN and DSEGAN obtain similar STOI results and both of them outperform SEGAN on this metric. The results in the table also suggest marginal impact of ISEGAN's number of iterations and DSEGAN's depth larger than $N = 2$ since no significant performance improvements are seen.

Interestingly, quite opposite results are seen between the discriminative baselines (DNN and TSN) and the generative models (ISEGAN and DSEGAN). In terms of speech quality, the discriminative models outperform the generative counterparts on PESQ, CSIG, COVL but underperform on CBAK and especially on SSNR. In addition, both DNN and TSN perform poorly on speech intelligibility. Degradation on STOI metric is even seen by DNN while TSN brings up modest improvement. On the contrary, both ISEGAN and DSEGAN obtain far better results on speech intelligibility. These results suggest that the discriminative models may alter the noisy input more aggressively than the generative ones and, as a result, introduce more artifacts to the enhanced signals.

To shed light on how the performance evolves during the enhancement process of DSEGAN and ISEGAN, we extracted and evaluated the output signals after each of their generators. The results are shown in Fig. 4. One can observe diverging patterns between DSEGAN and ISEGAN. With DSEGAN, overall, the enhancement performance is gradually improved when the signal is passed though the generators one after another. On the contrary, ISEGAN exposes a downward trend on most of the metrics with further enhancement iterations, except for SSNR. The rationale behind the SSNR improvement is that this measure best reflects the least-squares loss that was used to train the network. However, the improved SSNR does not properly reflect other metrics such as human perception and intelligibility represented by PESQ and STOI, which rely on frame-wise weighted frequency domain. This result tends to agree with the finding in psychoacoustics [35]. We speculate that parameter independency/sharing is the key. With independent parameters, each DSEGAN's generators is tasked for enhancement with one condition of noise and has full freedom to adapt to it. On the other hand, parameter sharing

forces the common generator of ISEGAN to deal with all conditions of noise, which is hard to achieve. Of note, instead of using all generators as a whole (i.e. the results in Table I), output of any generators can be used for inferencing. For ISEGAN, using the outputs of earlier generators for this purpose is apparently reasonable as suggested in Fig. 4.

E. Subjective evaluation

To validate the objective evaluation, we conducted a small-scale subjective evaluation of four conditions: noisy signals, SEGAN, ISEGAN and DSEGAN signals (with $N = 2$). Twenty volunteers aged 18–52 (F=6, M=14), with self-reported normal hearing, were asked to provide forced binary quality assessments between pairs of 20 randomly presented sentences, balanced in terms of speakers and noise types, i.e. each comparison varied only in the type of system. Following a familiarization session, tests were run individually using MAT-LAB, with listeners wearing Philips SHM1900 headphones in a low-noise environment. For each pair of utterances, the selected higher quality one was rewarded 1.0 while the lower quality received no reward. A *preference* score was obtained for each system by dividing its accumulated reward by the count of its occurrences in the test. Due to the small sample size, we assessed statistical significance of results using t-test. Results confirm that the three SEGAN signals are perceived as higher quality than the noisy signals (0.55 to 0.45, with $p < 0.05$). DSEGAN and ISEGAN together significantly outperform SEGAN (0.67 to 0.33, $p < 0.001$). However, DSEGAN and ISEGAN qualities were not significantly different (0.48 to 0.52) in this small test. Results support the detailed objective evaluation in which DSEGAN performs much better than either SEGAN or noise, however we find that ISEGAN also performs well in subjective tests.

VI. CONCLUSIONS

This paper presented a GAN method with multiple generators to tackle speech enhancement. Using multiple chained generators, the method aims to learn multiple enhancement mappings, each corresponding to a generator in the chain, to accomplish a multi-stage enhancement process. Two new architectures, ISEGAN and DSEGAN, were proposed. ISEGAN's generators share their parameters and, as a result, are constrained to learn a common mapping for all the enhancement stages. DSEGAN, in contrast, has independent generators that allow them to learn different mappings at different stages. Objective tests demonstrated that the proposed ISEGAN and DSEGAN perform comparably and are better than SEGAN on speech-quality metrics and that learning independent mappings leads to better performance than a common mapping. In addition, both the proposed systems achieve more favourable results than SEGAN on the speech-intelligibility metric as well as the subjective perceptual test.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, Inc., 2 edition, 2013.

[2] L.-P. Yang and Q.-J. Fu, "Spectral subtraction-based speech enhancementfor cochlear implant patients in background noise," *Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001–1004, 2005.

[3] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," *Proc. Intl. Conf. on Latent Variable Analysis and Signal Separation*, pp. 91–99, 2015.

[4] 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 (TASLP)*, vol. 23, no. 1, pp. 7–19, 2015.

[5] A. Kumar and D. Florencio, "Speech enhancement in multiple-noise conditions using deep neural networks," in *Interspeech*, 2016, pp. 3738–3742.

[6] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Proc. Interspeech*, 2017.

[7] N. Mamun, S. Khorram, and J. H. L. Hansen, "Convolutional neural network-based speech enhancement for cochlear implant recipients," *arXiv Preprint arXiv:1907.02526*, 2019.

[8] 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. ICASSP*, 2015, pp. 708–712.

[9] J. Lim and A. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 26, no. 3, pp. 197–210, 1978.

[10] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

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

[12] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. on Audio, Speech, and Language Processing*, pp. 1383–1393, 2011.

[13] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 2672–2680.

[14] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech enhancement generative adversarial network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[15] S. Pascual, J. Serrà, and A. Bonafonte, "Towards generalized speech enhancement with generative adversarial networks," *CoRR abs/1904.03418*, 2019.

[16] T. Higuchi, K. Kinoshita, M. Delcroix, and T. Nakatani, "Adversarial training for data-driven speech enhancement without parallel corpus," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 40–47.

[17] S. Qin and T. Jiang, "Improved wasserstein conditional generative adversarial network speech enhancement," *EURASIP Journal on Wireless Communications and Networking*, vol. 2018, no. 1, pp. 181, 2018.

[18] Z. X. Li, L. R. Dai, Y. Song, and I. McLoughlin, "A conditional generative model for speech enhancement," *Circuits, Systems, and Signal Processing*, vol. 37, no. 11, pp. 5005–5022, 2018.

[19] B. Li C. Donahue and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *Proc. ICASSP,*, 2018, pp. 5024–5028.

[20] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. ICCV*, 2017, pp. 2813–2821.

[21] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. CVPR*, 2017, pp. 5967–5976.

[22] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. CVPR*, 2016, pp. 2536–2544.

[23] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing mri reconstructionusing a generative adversarial network with a cyclic loss," *IEEE Trans. on Medical Imaging*, vol. 37, no. 6, pp. 1488–1497, 2018.

[24] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. ICLR*, 2016.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. ICCV*, 2015, pp. 1026–1034.

[26] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, 2016, pp. 2226–2234.

[27] J. Kim and M. Hahn, "Speech enhancement using a two-stage network for an efficient boosting strategy," *IEEE Signal Processing Letters*, vol. 26, no. 5, pp. 770–774, 2019.

[28] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, pp. 146–152.

[29] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: design, collection and data analysis of a large regional accent speech database," in *Proc. 2013 International Conference Oriental COCOSDA*, 2013, pp. 1–4.

[30] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *The Journal of the AcousticalSociety of America*, vol. 133, no. 5, pp. 3591–3591, 2013.

[31] D. P. Kingma and J. L. Ba, "Adam: a method for stochastic optimization," in *Proc. ICLR*, 2015, number 1-13.

[32] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[33] T. Tieleman and G. Hinton, "Lecture 6.5 - RMSprop: divide the gradient by a running average of its recent magnitude," *Coursera: Neural Networks for Machine Learning*, 2012.

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

[35] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ - the ITU standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.