SCALABLE AND EFFICIENT SPEECH ENHANCEMENT USING MODIFIED COLD DIFFUSION: A RESIDUAL LEARNING APPROACH

Minje Kim[*1,2] *and Trausti Kristjansson*[1]

[1]Amazon Lab126, Sunnyvale, CA, USA 94089
[2]University of Illinois at Urbana-Champaign, Department of Computer Science, IL, USA 61801

ABSTRACT

We introduce flexibility to the supervised learning-based speech enhancement framework to achieve scalable and efficient speech enhancement (SESE). To this end, SESE conducts a series of segmented speech enhancement inference routines, each of which incrementally improves the result of its preceding inference. The formulation is conceptually similar to cold diffusion, while we modify the sampling process so each step benefits from an easier milestone task rather than aggressively targeting the clean speech. In addition, the incremental enhancement steps are learned to recover the residual between the adjacent milestones, thus improving the overall enhancement performance. We show that the proposed method improves the baseline supervised model's performance, while it necessitates fewer diffusion steps to achieve the comparable performance with the more complex cold diffusion-based counterpart. Furthermore, SESE's scalability can be useful in applications where moderately suppressed non-speech interference is preferred to aggressive enhancement results, e.g., boosting dialog in movie soundtracks, speech enhancement on hearing aids, etc.

Index Terms— Speech enhancement, model compression, cold diffusion, scalability

1. INTRODUCTION

Recent advancements in the speech enhancement (SE) research benefited greatly from the innovations made in the deep neural network (DNN) architectures, the models' large capacity, and supervised learning done with a big training dataset. Numerous architectural innovations have been made in the literature, such as a better consideration of complex values [1], introduction of the novel gating mechanism [2], skip connections in the U-Net architecture [3], generative adversarial networks [4], autoregressive models, i.e., WaveNet [5], and more. In addition to the structural innovation, another predominant factor is the sheer complexity of the model that enables quality approximation of the nonlinear mapping between the noisy input signal and its clean version. Typically, such a large DNN model requires millions of parameters to accurately approximate the enhancement function, prohibiting its use in resource-constrained environments. Finally, the large model requires a big labeled dataset that represents the real-world acoustic scenes, such as the deep noise suppression challenge dataset [6], LibriMix [7], etc.

More recently, model compression became an emerging topic to reduce the SE models' complexity. Pruning and low-bit quantization have been popular approaches [8, 9, 10], including the single-bit quantization scheme, i.e., binarization [11]. Structural alteration is also popular in the speech separation literature, such as subband processing [12] and multi-resolution temporal features [13]. Finally, personalization of single-talker SE models showed superior compression performance [14, 15]. Yet, all these model compression techniques are based on typical supervised learning pipeline. Once trained, the SE model is "fixed." Then, it performs SE in one shot, i.e., by a single execution of the inference routine, in the test time.

We break down the one-shot inference process into a series of less complex inference runs, each of which is sufficient to solve a simpler sub-task. It can be seen as a process of gradually improving the observed speech quality by relaying intermediate SE results from one step to the next. A few such scalable systems have been proposed in the source separation literature. Deep nonnegative matrix factorization (NMF) [16] and deep unfolding network [17] proposed unfolded iterative optimization methods, whose intermediate results are used in the subsequent steps. They are scalable in that one can control the model complexity by the number of repeated inferences. Meanwhile, their loss measures the final unfolded result's quality, missing the milestone losses as we propose. Multi-view networks [18] are another branch that can handle an unknown number of observation channels of an audio scene, while the scalability is defined in terms of the channels rather than resource constraints. A more relevant approach is the block-wise optimization for masking-based SE networks, BLOOM-Net [19], which selectively uses a flexible number of ResNet-like blocks in its masking module for inference. However, once again, its optimization is missing the intermediate milestone targets we are proposing in this paper, which turns out to be useful. In addition, BLOOM-Net is designed for the masking-based architectures, while we pursue an architecture-agnostic method.

In this paper, we propose a scalable and efficient speech enhancement (SESE) network. We formulate our problem similarly to the cold diffusion (CD) framework [20], which recently showed promising performance on the SE task [21]. In the CD for SE method (CDSE), the SE process follows the *deterministic* sampling process, whose job is to generate the clean speech from a frozen noise pattern (i.e., the noisy speech input). Instead of employing the CD model directly, we modify it (a) to tackle specific intermediate SE goals at every step (b) to benefit from residual learning (c) to turn the process into a more scalable and efficient SE process with no significant performance degradation. To summarize,

- SESE achieves similar SE performances to the best CDSE variant, while using essentially 0.58% of CDSE's effective computation.
- The intermediate results produced during the reverse diffusion process are useful as they provide different levels of noise suppression, which can be useful when aggressive SE results are not necessary (e.g., dialog boosting, music remixing, hearing aids, etc.)[1].
- We employ the residual learning concept to offload the SE models' burden, resulting in a better performance.

[*]Work done at Amazon.

[1]Sound examples: https://minjekim.com/research-projects/sese

2. BASELINE MODELS

2.1. Traditional SE models' one-shot inference

We begin by illustrating how the traditional SE model works (Fig. 1a). The SE model as a function takes noisy speech as input, which is a mixture of the target clean utterance $\boldsymbol{s}$ and an additive noise source $\boldsymbol{n}$: $\boldsymbol{x} = \boldsymbol{s} + \boldsymbol{n}$. They are time-domain signals with N samples, i.e., they reside in the N-dimensional vector space. Hence, the *restoration* function $R(\boldsymbol{x})$ is learned to estimate the clean speech via a single inference process: $\boldsymbol{s} \approx \hat{\boldsymbol{s}} \leftarrow R(\boldsymbol{x})$.

To learn the potentially complex mapping function between various pairs of $\boldsymbol{x}$ and $\boldsymbol{s}$, a large model architecture is commonly employed, which in turn increases $R(\boldsymbol{x})$'s computational complexity.

2.2. The Baseline Cold Diffusion Model for SE

The cold diffusion for SE (CDSE) method is based on a sophisticated sampling algorithm that generates a clean signal from the noisy input [21]. It differs from the other probabilistic diffusion methods in that the diffusion process (and its inverse) works with deterministic noise, which corresponds to the noisy speech utterance.

It is based on the two contradicting operations, *degradation* and *restoration*. The degradation process $D(\cdot)$ is a continuous contamination process whose severity is defined by t. In the SE context, the degradation starts from the clean speech utterance when $t = 0$, i.e., $\boldsymbol{s} = \boldsymbol{x}_0$, and increases its severity as t increases, until it reaches the maximum degradation at $t = T$. CDSE defines an interpolation function with the mixing ratio parameter α_t as follows:

$$\boldsymbol{x}_t \leftarrow D(\boldsymbol{x}_0, t) = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{x}_T, \tag{1}$$

where α_t is the same cosine schedule as in [20, 21]:

$$\alpha_t = \frac{f(t)}{f(0)}, \quad f(t) = cos\left(\frac{t/T + s}{1+s} \cdot \frac{\pi}{2}\right)^2, \quad s = 0.008. \tag{2}$$

The deterministic noise is defined by the observed noisy speech in the SE context, i.e., $\boldsymbol{x}_T = \boldsymbol{s} + \boldsymbol{n}$, where $\alpha_T = 0$.

Conversely, a parametric restoration function $R(\cdot)$, e.g., a DNN, is trained to convert a given contaminated speech $\boldsymbol{x}_t$ at any severity level t back to the clean speech utterance $\boldsymbol{x}_0$ as best as it can:

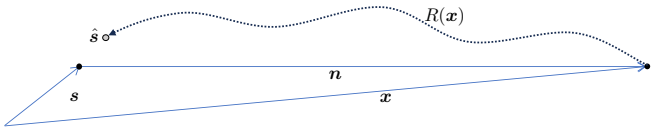
$$\boldsymbol{s} = \boldsymbol{x}_0 \approx \hat{\boldsymbol{x}}_0 \leftarrow R(\boldsymbol{x}_t, t), \tag{3}$$

where $R(\cdot)$ now takes the severity level t as conditioning input to inform the model of the required amount of restoration, while the earlier one-shot SE function does not need this conditioning input.

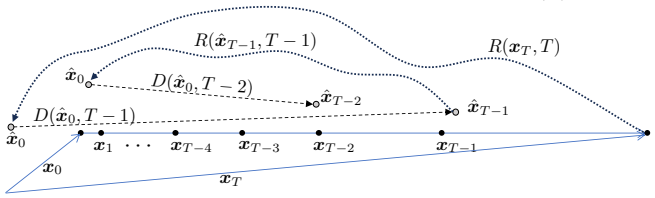
CDSE's sampling process refines the SE result by performing multiple back-and-forth runs of the degradation and restoration processes, as illustrated in Fig. 1b. First, the restoration function $R(\boldsymbol{x}_T, T)$ takes the most noisy utterance $\boldsymbol{x}_T$ as input and recovers the clean speech. The initial result $\hat{\boldsymbol{x}}_0$ may not be optimal, i.e., a compromised solution. Then, the degradation function estimates the second-most noisy degradation $\hat{\boldsymbol{x}}_{T-1}$. It is only an indirect estimation of $\boldsymbol{x}_{T-1}$, as the result deterministically relies on the quality of the first estimation of the clean speech $\hat{\boldsymbol{x}}_0$. This concludes a pair of restoration and degradation runs. In the next step at $T-1$, the restoration begins with the estimated intermediate degradation $\hat{\boldsymbol{x}}_{T-1}$ to recover a potentially better $\hat{\boldsymbol{x}}_0$. Then, the full iterative sampling process can be written: $\hat{\boldsymbol{x}}_0 \leftarrow R(\cdots D(R(\boldsymbol{x}_T, T), T-1), \cdots, 1)$.

In practice, a more sophisticated degradation function replaces (1) to improve the stability of the model,

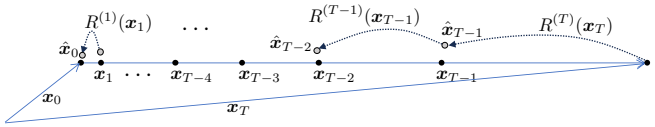
$$\hat{\boldsymbol{x}}_{t-1} \leftarrow \sqrt{\alpha_{t-1}}\hat{\boldsymbol{x}}_0 + \frac{\sqrt{1-\alpha_{t-1}}}{\sqrt{1-\alpha_t}}\left(\hat{\boldsymbol{x}}_t - \sqrt{\alpha_t}\hat{\boldsymbol{x}}_0\right). \tag{4}$$



(a) An ordinary one-shot inference of an SE model $R(\boldsymbol{x})$.



(b) The sampling process of a CD model that iteratively calls restoration $R(\boldsymbol{x}, t)$ and degradation $D(\hat{\boldsymbol{x}}_0, t-1)$ operations. In practice, an improved degradation function eq. (1) is used.



(c) The proposed SESE's sampling method with modified cold diffusion. It runs a series of step-wise dedicated models $R^{(t)}(\hat{\boldsymbol{x}}_t)$ to gradually convert the most noisy input $\boldsymbol{x}_T$ into its less noisy versions $\boldsymbol{x}_{t<T}$, which work as intermediate milestone targets of the gradual enhancement process.

Fig. 1: Inference of the traditional SE, CDSE, and SESE models.

In addition, the CDSE method proposes a more robust "unfolded" training method that encompasses degradations stemmed from both the clean speech $\boldsymbol{x}_0$ and its reconstruction $\hat{\boldsymbol{x}}_0$ as opposed to the original CD's training algorithm that only uses the former.

The main issues in the traditional one-shot inference and CDSE that we try to overcome in the proposed models are as follows:

- CDSE outperforms the traditional SE baseline. Furthermore, as in universal SE [22] and the score-based generative models [23], CDSE requires relatively fewer diffusion steps than the other diffusion models for SE [24, 25]. However, the complexity still linearly increases by the number of iterations (up to 50). A model with even fewer iterations is beneficial. The proposed SESE method reduces the number of steps down to 5 or 10 at no cost of performance drop.
- In CDSE, the same restoration model $R(\cdot)$ is supposed to handle all the different levels of degradations, conditioned on the degradation level t. Consequently, a large model is suitable for the maximum generalization power. SESE employs 17 times smaller models for each iteration, again with no performance drop.
- The advanced unfolding method for CDSE's training closes the gap, yet the degradation process does not directly aim at the original degradation $\boldsymbol{x}_t$, once it stems from an estimated clean speech $\hat{\boldsymbol{x}}_0$, not the ground truth $\boldsymbol{x}_0$. We postulate a more accurate reconstruction of the intermediate degradations is important for the sampling process to stay in the right course of denoising path, i.e., the interpolation line between $\boldsymbol{x}_0$ and $\boldsymbol{x}_T$. SESE tackles this issue by employing milestone goals .

3. THE PROPOSED MODELS

3.1. The Proposed Scalable and Efficient Speech Enhancement

Starting from the input noisy utterance $\boldsymbol{x}_T$, the proposed SESE model directly estimates $[\boldsymbol{x}_{T-1}, \ldots, \boldsymbol{x}_0]$, i.e, the intermediate degradations, during its iterative SE process. To this end, we break

down the traditional SE model's restoration function $R(\cdot)$ into a series of severity-specific models that run one after another:

$$\boldsymbol{x}_0 \approx \hat{\boldsymbol{x}}_0 \leftarrow R(\boldsymbol{x}_T) = R^{(1)} \circ R^{(2)} \circ \cdots R^{(T-1)} \circ R^{(T)}(\boldsymbol{x}_T), \quad (5)$$

where each $R^{(t)}(\boldsymbol{x}_t)$ achieves the t-th milestone enhancement goal,

$$\boldsymbol{x}_{t-1} \approx \hat{\boldsymbol{x}}_{t-1} \leftarrow R^{(t)}(\boldsymbol{x}_t). \quad (6)$$

Our assumption is that the difference between the adjacent degradations $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ is smaller, thus easier to learn than a bigger jump from $\boldsymbol{x}_t$ to $\boldsymbol{x}_0$ that the CDSE has to handle, not to mention the traditional SE model's biggest leap $\boldsymbol{x}_0 \approx R(\boldsymbol{x}_T)$ (Fig. 1a). The assumption leads us to our main efficiency-related argument that small models suffice to learn these simpler SE problems. In our experiments we employ a 17 times smaller model architecture for each of the incremental models $R^{(t)}(\cdot)$ than the default model $R(\cdot)$ CDSE uses, or 11 times smaller than the one-shot SE model.

An issue with the proposed chain of SESE operations is the propagation of the reconstruction error. As a remedy, we employ a two-step training method that first pretrains the models on the ground-truth degradations as the input and target of the models, and then finetunes them using estimated degradations as the input, with their corresponding loss functions, $\mathcal{L}_{\text{PT}}$ and $\mathcal{L}_{\text{FT}}$, respectively:

$$\mathcal{L}_{\text{PT}} = \sum_{t=0}^{T-1} \mathcal{L}_{\text{PT}}^{(t)}, \quad \text{where } \mathcal{L}_{\text{PT}}^{(t)} = \mathcal{D}\big(\boldsymbol{x}_t || R^{(t+1)}(\boldsymbol{x}_{t+1})\big), \quad (7)$$

$$\mathcal{L}_{\text{FT}} = \sum_{t=0}^{T-1} \mathcal{L}_{\text{FT}}^{(t)}, \quad \text{where } \mathcal{L}_{\text{FT}}^{(t)} = \mathcal{D}\big(\boldsymbol{x}_t || R^{(t+1)}(\hat{\boldsymbol{x}}_{t+1})\big), \quad (8)$$

with a choice of error metric, $\mathcal{D}(\cdot)$.

3.2. Residual Learning with SESE (ResSESE)

A variation can further improve SESE's enhancement performance by learning the residual between degradations. Our main efficiency argument comes from the premise that the difference of adjacent degaradations $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t-1}$ is smooth enough for small models to learn. Hence, we can also assume that their difference, or *residual*, requires even less modeling effort as shown in [26]. The restoration function,

$$\boldsymbol{x}_{t-1} \approx \hat{\boldsymbol{x}}_{t-1} \leftarrow R^{(t)}(\boldsymbol{x}_t) + \boldsymbol{x}_t, \quad (9)$$

now essentially estimates the difference $\boldsymbol{x}_{t-1} - \boldsymbol{x}_t$ rather than trying to estimate $\boldsymbol{x}_{t-1}$ directly. Our experiments in Sec. 4 show that the residual learning version of SESE (ResSESE) improves the plain SESE models that use eq. (6). Note that the loss functions eq. (7) and (8) need to be adjusted accordingly, i.e., $\mathcal{D}\big(\boldsymbol{x}_t || R^{(t+1)}(\boldsymbol{x}_{t+1}) + \boldsymbol{x}_{t+1}\big)$ and $\mathcal{D}\big(\boldsymbol{x}_t || R^{(t+1)}(\hat{\boldsymbol{x}}_{t+1}) + \hat{\boldsymbol{x}}_{t+1}\big)$.

4. EXPERIMENTS

4.1. Experimental Setup

We follow the experimental setup provided in the CDSE paper, which is based on the VoiceBank+DEMAND dataset [27], where 30 speakers and 10 noise sources are mixed up at four signal-to-noise ratio (SNR) settings $[0, 5, 10, 15]$ dB for training. Two speakers are set aside for testing, whose mixture SNRs are $[2.5, 7.5, 12.5, 17.5]$ dB. Out of DiffWave [28] and deep complex convolution recurrent network (DCCRN) [1] baselines, we chose the latter due to its superior performance reported in the CDSE paper.

Model Architecture: We inherit DCCRN's "CL" variation, which has $[16, 32, 64, 128, 256, 256]$ kernels in its convolutional-layer encoder and 256 hidden units in its two LSTM layers. We call this baseline architecture DCCRN-L-BL. Meanwhile, CDSE's severity conditioning module introduces additional complexity, which we denote DCCRN-L-CDSE. To prove our model compression argument, SESE uses a smaller architecture DCCRN-S with $[32, 64, 64, 64, 64]$ convolutional kernels and 64 LSTM units. Their numbers of trainable parameters are reported in Table 1.

Systems in Comparison: We compare the proposed SESE and ResSESE models to the baseline DCCRN's performance as well as its CDSE variants. We append SESE, ResSESE, BL, and CDSE to the model architecture's code name, respectively. Table 1 summarizes those models, where * indicates the models' original performance reported in the literature.

The Training Configuration: We follow the training recipe provided in [21] with the following variation. A randomly chosen 1 second-long segments with a sample rate of 16 kHz are used to form batches during training. The learning rate of Adam optimizer [29] is 1×10^{-3} for pretraining and 1×10^{-4} for finetuning. The loss function $\mathcal{D}(\cdot)$ is defined with the negative scale-dependent signal-to-noise ratio (SDSNR) [30], which gives the best validation performance than other candidates. Validation is performed on two selected training speakers (p282 and p287) via the wide-band perceptual evaluation of speech quality (PESQ) metric.

Evaluation Criteria: Following [21], the models' performance on the test signals (of speaker p232 and p257) is measured using PESQ and other objective metrics that simulate mean opinion scores, i.e., prediction of the signal distortion (CSIG), prediction of the background intrusiveness (CBAK), and prediction of the overall speech quality (COVL). In addition, for an in-depth analysis of the intermediate SE results of the proposed models, we also report the BSS_EVAL scores, source-to-distortion, source-to-interference, and source-to-artifact ratios (SDR, SIR, and SAR, respectively) [31].

4.2. Experimental Results and Discussion

The Baseline Models: Table 1 reports the test-time performance of the systems in comparison. First, we see that the baseline DCCRN model exhibits a performance drop when smaller architecture is used (DCCRN-L-BL vs. DCCRN-S-BL). For a fairer comparison, we also present the original performance reported in [1] (DCCRN-L-BL*) whose PESQ value is slightly worse than our trained model, while its CSIG, CBAK, and COVL scores are slightly better.

The CDSE Models: The two CDSE models (DCCRN-L-CDSE* and DiffWave-CDSE*) improve the baseline as reported in [21], especially when the maximum $\tau = T = 50$ steps are fully applied. Here, τ denotes the number of actual reverse diffusion steps applied. DiffWave-CDSE*'s suboptimal performance at $\tau = 1$ justifies the necessity of more sampling steps for a better SE performance. We use the notion of *effective complexity* to describe the run-time complexity of the model, which counts the total number of parameters involved in when an iterative model goes through repeated inferences by τ times. As for DCCRN-L-CDSE*, it goes up to $\tau \times 5.6M = 280M$. Hence, compared to the DCCRN-L-BL's $3.7M$ parameters that are involved only once, CDSE's test-time inference is considered much more computationally complex.

The Proposed SESE Models: When a relatively short diffusion process is used ($\tau = T = 5$), we see that the DCCRN-S-SESE variation starts to outperform the DCCRN-S-BL. In comparison to the larger baselines, it outperforms our own trained version DCCRN-L-BL, while DCCRN-L-BL*'s COVL is still slightly better. Al-

System	T	τ	PESQ	CSIG	CBAK	COVL	SDR	SIR	SAR	Trainable Params.	Effective Complexity	Compression Ratio
Unprocessed	-	-	1.97	3.37	2.45	2.65	8.66	8.66	Inf.	-	-	-
DCCRN-L-BL* [1]	-	-	2.59	3.71	3.23	3.13	-	-	-	$3.7M$	$3.7M$	1.32%
DCCRN-L-BL	-	-	2.64	3.45	3.32	3.03	20.53	26.10	22.17	$3.7M$	$3.7M$	1.32%
DCCRN-S-BL	-	-	2.61	3.38	3.30	2.98	20.58	25.91	22.40	$0.3M$	$0.3M$	0.12%
DCCRN-L-CDSE* [21]	50	50	**2.77**	3.91	3.32	**3.33**	-	-	-	$5.6M$	$50 \times 5.6M$	100.00%
DiffWave-CDSE* [21]	50	50	2.60	3.79	3.21	3.19	-	-	-	$2.3M$	$50 \times 2.3M$	41.07%
DiffWave-CDSE* [21]	50	1	2.50	3.59	3.21	3.04	-	-	-	$2.3M$	$2.3M$	0.82%
DCCRN-S-SESE	5	5	2.69	3.54	**3.38**	3.11	21.50	27.06	23.09	$5 \times 0.3M$	$5 \times 0.3M$	0.58%
DCCRN-S-SESE	5	1	2.06	3.45	2.64	2.73	12.25	12.48	26.53	$0.3M$	$0.3M$	0.12%
DCCRN-S-SESE	10	10	2.73	3.93	3.41	3.32	21.89	28.60	23.08	$10 \times 0.3M$	$10 \times 0.3M$	1.16%
DCCRN-S-SESE	10	1	1.97	3.34	2.58	2.63	12.05	12.29	26.04	$0.3M$	$0.3M$	0.12%
DCCRN-S-ResSESE	5	5	2.73	**3.96**	3.34	**3.33**	19.84	23.03	23.16	$5 \times 0.3M$	$5 \times 0.3M$	0.58%
DCCRN-S-ResSESE	5	1	2.20	3.59	2.40	2.87	12.12	12.20	31.14	$0.3M$	$0.3M$	0.12%
DCCRN-S-ResSESE	10	10	2.74	3.95	**3.38**	**3.33**	20.56	24.32	23.31	$10 \times 0.3M$	$10 \times 0.3M$	1.16%
DCCRN-S-ResSESE	10	1	2.11	3.49	2.37	2.77	11.22	11.26	33.93	$0.3M$	$0.3M$	0.12%
DCCRN-S-ResSESE	20	20	2.72	3.89	3.38	3.29	21.11	25.81	23.15	$20 \times 0.3M$	$20 \times 0.3M$	2.33%

Table 1: Summary of the test-time SE results of systems in comparison.

though the short diffusion process's effective complexity must take into account the five steps $\tau = 5$, DCCRN-S-SESE's building-block architecture is very small ($0.326M$), thus guaranteeing a certain level of compression compared to DCCRN-L-BL ($3.7M$ vs. $1.63M$). When T and τ increase to 10, we see the performance surely improves the large DCCRN baselines in all metrics. Compared to DCCRN-L-CDSE*, CSIG and CBAK is better, while PESQ and COVL are only slightly worse. This $\tau = 10$ version of SESE uses $325,579 \times 10 = 3.26M$ effective parameters, which is still much smaller than DCCRN-L-CDSE*'s $280M$ parameters, showcasing a significant 1.16% of compression ratio.

The Proposed ResSESE Models: Residual learning introduces additional performance improvement to SESE, free of computational cost. DCCRN-S-ResSESE outperforms DCCRN-S-SESE when $\tau = T = 5$, competing with SESE's $\tau = T = 10$ variation and DCCRN-L-CDSE*'s $\tau = T = 50$ version. However, the gain coming from the residual learning saturates when $\tau = T = 10$. If DCCRN-S-ResSESE uses $\tau = 5$ steps, its effective model parameters are $1.63M$, which is only 0.58% of DCCRN-L-CDSE*'s $280M$ parameters. Since these two models' performances are similar, 0.58% is the best compression ratio achieved by our proposed models.

Analysis of the Intermediate Results: We claim that the proposed SESE and ResSESE models produce useful intermediate solutions for some applications. The one-shot SE model or the CDSE models aim at the clean speech target whenever the restoration function $R(\cdot)$ is called, regardless of the severity level t. Conversely, SESE or ResSESE's intermediate results $\hat{\boldsymbol{x}}_t$ are trained to approximate their corresponding ground-truth $\boldsymbol{x}_t$. We use $\boldsymbol{x}_t$ as our milestone goals by varying t from $T-1$ to 0, i.e., by gradually attenuating the contribution of the noise in the target mixture. Our hypothesis is that by targeting these milestone goals the restoration function $R^{(t)}(\cdot)$ can refrain from performing too aggressive denoising, which can lead to perceptually suboptimal results, e.g., bleeding of the speech component (i.e., noise oversuppression), introduction of artifacts, etc. We observe that SESE and ResSESE's CSIG values at $\tau = 1$ are generally less deteriorated than the other metrics, implying the sustained speech quality in those compromised solutions. Another interesting observation is that the ResSESE models' $\tau = 1$ results are significantly better than the corresponding SESE models', showcasing the importance of the residual learning in the early stage of the reverse diffusion process. Although the BSS_EVAL metrics should only reflect a trend as the models' validation for early stopping was based on PESQ, the trend in the SAR values is similar: The SAR values of the $\tau = 1$ reconstructions are very high, meaning the level of artifacts contained in the restored speech is low. Hence, the overall distortion computed in the SDR values are largely correlated with SIR, i.e., whether the noise sources are sufficiently suppressed. Likewise, SESE and ResSESE can choose to provide moderate SE results when necessary, e.g., a mild volume adjustment of the speech source is sufficient for the application. It is also a cost-saving compromise as τ is small.

Failure Modes: SESE requires more trainable parameters as τ grows, because it consists of τ individual models rather than sharing a big model at all t stages as in CDSE. This could make optimization more difficult, as shown in Table 1's last row, where the $\tau = 20$ results are saturated. Reusing a model in multiple steps (not in all steps though) could be a potential solution. In addition, since cold diffusion uses deterministic noise, its generation power is limited. As a result, SESE might not explore creative solutions, which other generative models are expected to do.

5. CONCLUSION

We proposed a scalable and efficient speech enhancement (SESE) framework, which employs a series of small DNN models that learn to achieve easier milestone goals, i.e., the interpolation between the clean and noisy utterances. We showed that SESE can learn to perform incremental speech enhancement in contrast to the cold diffusion-based speech enhancement model that has to employ a large DNN model to approximate the final clean speech target at every step. With the help of residual learning, the proposed method provided low-artifact intermediate results that stay in the smooth enhancement path, while the end result is on par with its 172 times more complex cold diffusion method's. In the future, we will verify that the proposed method extends to other model architectures.

6. REFERENCES

[1] Y. Hu *et al.*, "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. Interspeech*, 2020.

[2] K. Tan, J. Chen, and D. Wang, "Gated residual networks with dilated convolutions for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 189–198, 2019.

[3] C. Macartney and T. Weyde, "Improved Speech Enhancement with the Wave-U-Net," *arXiv preprint arXiv:1811.11307*, 2018.

[4] S. Pascual, A. Bonafonte, and J. Serrà, "SEGAN: Speech Enhancement Generative Adversarial Network," in *Proc. Interspeech*, 2017, pp. 3642–3646.

[5] D. Rethage, J. Pons, and X. Serrà, "A WaveNet for speech denoising," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5069–5073.

[6] H. Dubey *et al.*, "Deep speech enhancement challenge at ICASSP 2023," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[7] J. Cosentino *et al.*, "LibriMix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[8] J.-Y. Wu *et al.*, "Increasing compactness of deep learning based speech enhancement models with parameter pruning and quantization techniques," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1887–1891, 2019.

[9] I. Fedorov *et al.*, "TinyLSTMs: Efficient neural speech enhancement for hearing aids," in *Proc. Interspeech*, 2020.

[10] K. Tan and D. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1785–1794, 2021.

[11] S. Kim, M. Maity, and M. Kim, "Incremental binarization on recurrent neural networks for single-channel source separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[12] J. Yu and Y. Luo, "Efficient monaural speech enhancement with universal sample rate band-split RNN," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[13] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo RM -RF: Efficient Networks for Universal Audio Source Separation," in *Proc. of the IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.

[14] A. Sivaraman and M. Kim, "Efficient Personalized Speech Enhancement Through Self-Supervised Learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1342–1356, 2022.

[15] S. Kim and M. Kim, "Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2021.

[16] J. Le Roux, J. R. Hershey, and F. Weninger, "Deep NMF for speech separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 66–70.

[17] S. Wisdom, J. Hershey, J. Le Roux, and S. Watanabe, "Deep unfolding for multichannel source separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 121–125.

[18] J. Casebeer, B. Luc, and P. Smaragdis, "Multi-view networks for denoising of arbitrary numbers of channels," in *Proc. of the International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 496–500.

[19] S. Kim and M. Kim, "BLOOM-Net: Blockwise optimization for masking networks toward scalable and efficient speech enhancement," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022.

[20] A. Bansal *et al.*, "Cold diffusion: Inverting arbitrary image transforms without noise," *arXiv preprint arXiv:2208.09392*, 2022.

[21] H. Yen, F. G. Germain, G. Wichern, and J. Le Roux, "Cold diffusion for speech enhancement," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[22] J. Serrà *et al.*, "Universal speech enhancement with score-based diffusion," *arXiv preprint arXiv:2206.03065*, 2022.

[23] S. Welker, J. Richter, and T. Gerkmann, "Speech Enhancement with Score-Based Generative Models in the Complex STFT Domain," in *Proc. Interspeech*, 2022, pp. 2928–2932.

[24] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 659–666.

[25] Y.-J. Lu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 7402–7406.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[27] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," *https://doi.org/10.7488/ds/2117*.

[28] Z. Kong *et al.*, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. of the International Conference on Learning Representations (ICLR)*, 2021.

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

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

[31] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.