DIFFUSION-BASED SPEECH ENHANCEMENT WITH JOINT GENERATIVE AND PREDICTIVE DECODERS

Hao Shi[1,2]*, Kazuki Shimada*[2]*, Masato Hirano*[2]*, Takashi Shibuya*[2]*, Yuichiro Koyama*[2]*, Zhi Zhong*[2]*, Shusuke Takahashi*[2]*, Tatsuya Kawahara*[1]*, Yuki Mitsufuji*[2]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Sony Group Corporation, Tokyo, Japan

ABSTRACT

Diffusion-based generative speech enhancement (SE) has recently received attention, but reverse diffusion remains time-consuming. One solution is to initialize the reverse diffusion process with enhanced features estimated by a predictive SE system. However, the pipeline structure currently does not consider for a combined use of generative and predictive decoders. The predictive decoder allows us to use the further complementarity between predictive and diffusion-based generative SE. In this paper, we propose a unified system that use jointly generative and predictive decoders across two levels. The encoder encodes both generative and predictive information at the shared encoding level. At the decoded feature level, we fuse the two decoded features by generative and predictive decoders. Specifically, the two SE modules are fused in the initial and final diffusion steps: the initial fusion initializes the diffusion process with the predictive SE to improve convergence, and the final fusion combines the two complementary SE outputs to enhance SE performance. Experiments conducted on the Voice-Bank dataset demonstrate that incorporating predictive information leads to faster decoding and higher PESQ scores compared with other score-based diffusion SE (StoRM and SGMSE+).

Index Terms— speech enhancement, diffusion model, generative model, predictive system

1. INTRODUCTION

Speech enhancement (SE) aims to recover clean speech from noisy signals. Since noise in real-world scenarios [1, 2, 3] significantly degrades the performance of speech applications, SE is an important front-end in speech processing applications, such as automatic speech recognition [4, 5, 6], speaker identification [7], and semantic communication [8, 9, 10, 11]. Supervised SE systems [12, 13] have more robust performance compared to traditional SE systems [14, 15]. Therefore, they have recently been intensively investigated [16, 17]. They can be classified into two types: predictive (also called discriminative) [18, 19, 20, 21] and generative [22, 23, 24, 25]. They adopt different paradigms. Predictive SE systems learn the single best deterministic mapping between noisy speech and its corresponding clean speech [25]. In contrast, the target distribution of clean speech is implicitly or explicitly learned in generative SE systems [25]. Generative SE models include variational auto-encoders [26], generative adversarial networks [22] and diffusion models [23, 24, 25]. Among them, diffusion models have recently attracted a significant attention due to their success in other fields [27, 28].

Diffusion models are inspired by non-equilibrium thermodynamics. The data are gradually turned into noise, and the neural network learns to invert the progressive noise-adding process. The conditional diffusion model [23] uses noisy spectrograms as the conditioner. However, its objective function assumes that the global distribution of the additive noise follows a standard white Gaussian distribution, which is inconsistent with real noise statistics. The score-based diffusion model [24] is based on stochastic differential equations (SDEs), which makes the training fully generative without any prior noise distribution assumptions. The reverse diffusion process (decoding) of the diffusion models is very time-consuming. To reduce the number of reverse diffusion steps, several studies [25, 29] have combined a predictive model with the generative model. A previous work use the predictive information as the initialization for the generative model [25]. The work also show that the generative and predictive models have different distortions [25]. Furthermore, UNIVERSE [30] shows that adding predictive information to the decoder of the diffusion model can help diffusion score estimation. However, the previous pipeline structures [25, 29] limit the systems further use the complementarity between the generative and predictive modules.

In this paper, we propose a unified speech enhancement (SE) system that integrates generative and predictive SE modules at the shared encoding and enhanced feature levels. At the shared encoding level, the model incorporates a shared encoder along with both predictive and generative decoders. The generative module is a score-based diffusion model adopted [31]. To leverage the complementarity between the two modules at the enhanced feature level, we fuse the enhanced generative-based and predictive-based features during the first and final diffusion steps. The first step fusion utilizes the predicted enhanced feature to initialize the subsequent diffusion processes. In order to maintain small changes in the feature distribution, the two features are fused instead of using the predicted spectra directly, although the enhanced predictive feature has higher performance in the first step than the enhanced generative feature. Since the two systems introduce different signal distortions, feature fusion is also adopted in the final step to leverage the complementarity between the generative and predictive SE modules.

2. SCORE-BASED DIFFUSION MODEL

2.1. Stochastic process

The linear stochastic differential equation (SDE) relies on a stochastic diffusion process $\{x_t\}_{t=0}^{T}$ [31]:

$$\mathrm{d}x_t = \underbrace{\gamma(y - x_t)\mathrm{d}t}_{:=\mathrm{f}(x_t, y)} + \underbrace{\left[\sigma_{min}\left(\frac{\sigma_{max}}{\sigma_{min}}\right)^t \sqrt{2\log\left(\frac{\sigma_{max}}{\sigma_{min}}\right)}\right]}_{:=g(t)} \mathrm{dw} \quad (1)$$

where x_t is the current state of the process indexed by a continuous time variable $t \in [0, T]$ [31]. x_0 is the clean speech, which represents the initial condition, y is the noisy speech, and w denotes a

standard Wiener process. The vector-valued function $\mathrm{f}(x_t, y)$ is referred to as the drift coefficient, $g(t)$ is the diffusion coefficient of x_t, and γ is the stiffness. σ_{min} and σ_{max} control the amount of Gaussian white noise at each diffusion timestep. The SDE in (1) has an associated reverse SDE [31, 32],

$$\mathrm{d}x_t = \left[-\mathrm{f}(x_t, y) + g(t)^2 \bigtriangledown_{x_t} \log p_t(x_t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathrm{w}} \quad (2)$$

Practically, the score $\bigtriangledown_{x_t} \log p_t(x_t)$ is estimated by a score model. The score model can be denoted as $s_\theta(x_t, y, t)$, which is parameterized by a set of DNN parameters θ. It receives the current state of the process x_t, the noisy speech y, and the current timestep t as inputs. Finally, by substituting the score model into the reverse SDE in (2) [33], we obtain

$$\mathrm{d}x_t = \left[-\mathrm{f}(x_t, y) + g(t)^2 s_\theta(x_t, y, t)\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathrm{w}}] \quad (3)$$

which can be solved with various solver procedures.

2.2. Training objective

The mean and variance of the process state x_t can be derived when its initial conditions are known [34]. Since the feature used in this paper is a complex spectrogram, at an arbitrary timestep t, x_t can be directly sampled by x_0 and y with the perturbation kernel:

$$p_{0t}(x_t|x_0, y) = \mathcal{CN}(x_t; \mu(x_0, y, t), \sigma(t)^2 \mathrm{I}) \quad (4)$$

where $\mathcal{CN}$ denotes the circularly symmetric complex normal distribution. I is the identiry matrix. The mean and variance can be estimated as follows [34]:

$$\mu\left(x_0, y, t\right) = e^{-\gamma t} x_0 + (1 - e^{-\gamma t}) y \quad (5)$$

$$\sigma(t)^2 = \frac{\sigma_{min}^2((\sigma_{max}/\sigma_{min})^{2t} - e^{-2\gamma t})\log(\sigma_{max}/\sigma_{min})}{\gamma + \log(\sigma_{max}/\sigma_{min})} \quad (6)$$

The denoising score matching is described as follows [24]:

$$\begin{aligned} \bigtriangledown_{x_t} \log p_{0t}(x_t|x_0, y) &= \bigtriangledown_{x_t} \log \left[|2\pi\sigma\mathrm{I}|^{-\frac{1}{2}} e^{-\frac{||x_t - \mu||_2^2}{2\sigma^2}}\right] \\ &= \bigtriangledown_{x_t} \log |2\pi\sigma(t)\mathrm{I}|^{-\frac{1}{2}} - \bigtriangledown_{x_t} \frac{||x_t - \mu(x_0, y, t)||_2^2}{2\sigma(t)^2} \quad (7) \\ &= -\frac{x_t - \mu(x_0, y, t)}{\sigma(t)^2} \end{aligned}$$

At each training step, the following four steps are executed [24]:

① Sample a random $t \sim \mathcal{U}\left[t_\varepsilon, T\right]$;

② Sample (x_0, y) from the dataset;

③ Sample $\mathrm{z} \sim \mathcal{CN}(\mathrm{z}; 0, \mathrm{I})$;

④ Sample x_t from (4) by computing:

$$x_t = \mu(x_0, y, t) + \sigma(t)\mathrm{z} \quad (8)$$

The training loss is computed between the model output and the score of the perturbation kernel. By substituting (8) into (7), the overall training objective is described as follows [24]:

$$\arg\min_\theta \mathbb{E}_{t,(x_0,y),\mathrm{z},x_t|(x_0,y)} \left[||\mathrm{s}_\theta(x_t, y, t) + \frac{\mathrm{z}}{\sigma(t)}||_2^2\right] \quad (9)$$

2.3. Inference

For inference, a trained score model s_θ approximates the true score for all $t \in [0, T]$. The noisy speech y is conditioned to estimate clean speech x_0 by solving the plug-in reverse SDE in (3). The initial condition of the reverse process at $t = T$ can be determined as follows [24]:

$$x_T \sim \mathcal{N}_{\mathbb{C}}(x_T; y, \sigma(T)^2 \mathrm{I}) \quad (10)$$

The denoising process through the reverse process starts at $t = T$ and ends at $t = 0$ iteratively. PC samplers combine single-step methods with numerical optimization approaches for the reverse SDE [31]. PC samplers consist of a predictor and a corrector. The predictor solves the reverse process by iterating through the reverse SDE [31]. The corrector refines the current state after each iteration step of the predictor [31].

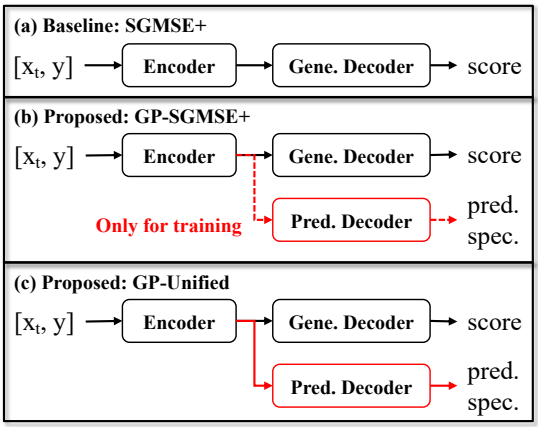

Fig. 1. The score model (used in PC samplers) structure: (a) the baseline SGMSE+; (b) the proposed Generative and Predictive based SGMSE+ (GP-SGMSE+); (c) the proposed Unified Generative and Predictive model (GP-Unified). **Note that the skip connection exists between the encoder and decoders (generative and predictive).**

3. PROPOSED METHOD

Score-based diffusion models have already achieved performance comparable to that of the predictive model. However, the predictive model calls the neural network only once, while the diffusion model needs to call the neural network several times, which significantly increases the decoding time. Some multi-stage models incorporate enhanced predictive features into the diffusion model to significantly reduce the number of diffusion steps. However, these systems are pipeline structures, which limit to further utilize the complementarity between the predictive and generative models, since the generative models distort signals much differently from how the predictive models do [25]. Besides, introducing predictive information into generative models can help improve the performance of the diffusion process [30]. Therefore, we implement these two different SE systems in the unified system by fusing them. The flowchart of the proposed method is shown in Fig. 2.

3.1. GP-SGMSE+

The structure of the generative- and predictive-based model (**GP-SGMSE+**) to estimate a score $\bigtriangledown_{x_t} \log p_t(x_t)$ is shown in Fig. 1(b). The model contains a shared encoder and two decoders. The original SGMSE+ neural network [24] contains an encoder and a decoder. And the encoder only focuses on encoding generative information. In GP-SGMSE+, the generative and predictive decoders share an encoder. In this model, the predictive decoder is only used during training to introduce predictive enhancement information into the model. When reverse diffusion process, no additional parameters are introduced to the baseline model shown in Fig. 1(a). The mean square error is used for computing the predictive loss:

$$L_{pred} = ||x_{pred} - x_0||^2 \quad (11)$$

where x_{pred} is the output of the predictive decoder. The final loss combines both the predictive and generative loss in (9) and (11). Because the two tasks are equally important, the weights of the losses of the two parts are 0.5 during training. The weights of the losses will not affect the two separate decoders, but will affect the shared encoder information, which will be explored in the future. During reverse diffusion process, only the generative decoder is used. The inference process is the same as in Section 2.3.

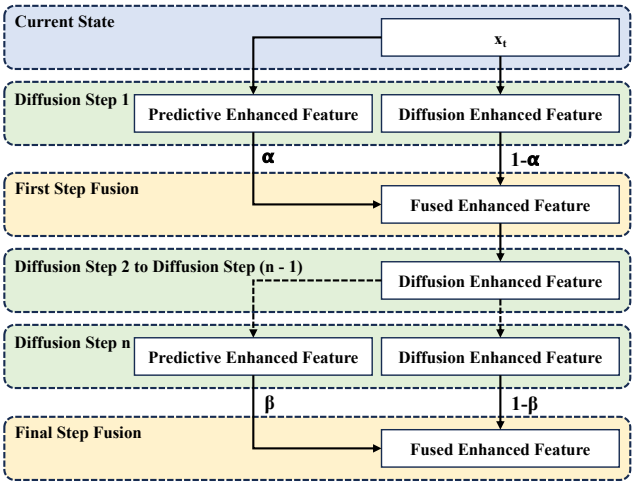

Fig. 2. A flowchart of the proposed method. Generative and predictive SE systems are fused in the first and final diffusion steps.

3.2. GP-Unified

The structure of the unified generative and predictive model (**GP-Unified**) is shown in Fig. 1(c). Unlike in "GP-SGMSE+", the predictive decoder is also used in the reverse diffusion process. During the reverse diffusion process, the enhanced generative and predictive features are fused in the first and final diffusion steps. The first step fusion is to use the predictive enhanced spectrogram to initialize the follow-up processes of diffusion:

$$\widehat{x_1} = \alpha * x_1 + (1-\alpha) * x_1^{pre} \tag{12}$$

where $\widehat{x_1}$ is the first diffusion-enhanced complex spectrogram, which will be used for subsequent diffusion steps, and x_1^{pre} is the predictive enhanced complex spectrogram in the first diffusion step. The reason for not using the predicted complex spectrogram directly is to maintain the feature distribution of the diffusion complex spectrograms. The two enhanced complex spectrograms are fused in the final step to exploit the complementary information:

$$\widehat{x_n} = \beta * x_n + (1-\beta) * x_n^{pre} \tag{13}$$

where $\widehat{x_n}$ is the final enhanced feature, and x_n^{pre} is the predictive enhanced feature in the final diffusion step. The first and final fusion are only used for reverse diffusion process. The final loss combines both predictive and generative loss in (9) and (11).

4. EXPERIMENTAL EVALUATIONS

The Noise Conditional Score Network (NCSN++[1]) architecture was used for the score model in both the baseline model (SGMSE+) and the proposed models (GP-SGMSE+, GP-Unified). The generative and predictive decoders had the same structure. The real and imaginary parts of the complex spectrograms were used as inputs. The residual blocks in upsampling and downsampling layers were based on the BigGAN architecture. Each upsampling layer consisted of three residual blocks, and each downsampling layer consisted of two blocks with the last block performing the upsampling or downsampling. Global attention was added at a resolution of 16×16 and in the bottleneck layer. All models were trained for 100 epochs.

The experiments were based on the public Voicebank-DEMAND [35]. The dataset can be accessed from this URL[2]. All speech data were sampled at 16 kHz.

To evaluate the proposed method, perceptual evaluation of speech quality (PESQ) [36], extended short-time objective intelligibility (ESTOI) [37], scale-invariant signal-to-distortion ratio (SI-SDR) [38], scale-invariant signal-to-interference ratio (SI-SIR) [38], and scale-invariant signal-to-artifact ratio (SI-SAR) [38] are used as the evaluation metric. Besides, the real-time factor (RTF) is used to evaluate the efficiency of different systems.

We set the hyperparameter α for the first step fusion from 0.1 to 0.9, and finally found that 0.2 was the best. The hyperparameter β for the final step fusion was 0.1. Because the two tasks are equally important, the weights of the training losses of the two parts are 0.5. We also tried fusing complex spectrograms at each step, but did not obtain improved performance.

[1]https://github.com/sp-uhh/sgmse

[2]https://datashare.ed.ac.uk/handle/10283/1942

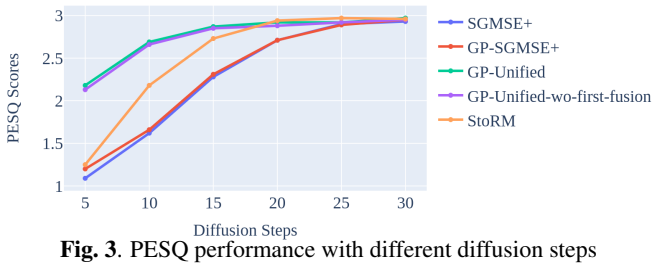

Fig. 3. PESQ performance with different diffusion steps

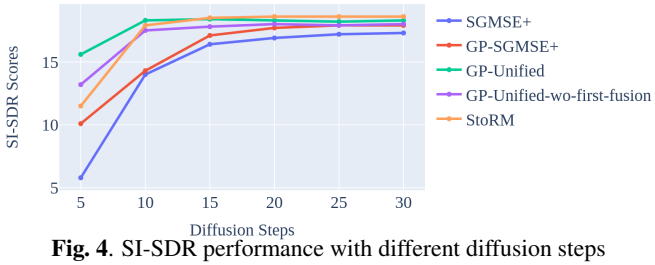

Fig. 4. SI-SDR performance with different diffusion steps

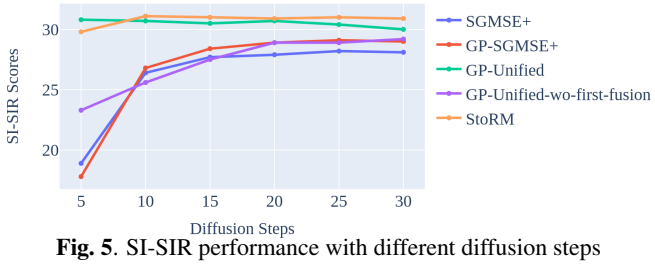

Fig. 5. SI-SIR performance with different diffusion steps

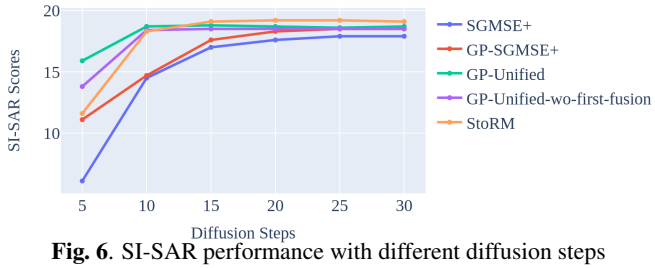

Fig. 6. SI-SAR performance with different diffusion steps

4.1. Effect of incorporating predictive loss function

As shown by comparison of "SGMSE+" and "GP-SGMSE+" in Figure 4 to Figure 6, introducing a predictive loss function into the diffusion model can effectively reduce speech distortion, reduce noise, and reduce artificial noise. However, the predictive information has a minor effect on the PESQ, as shown in Figure 3.

4.2. Effect of the first and final fusion

"GP-Unified-w/o-first-fusion" adopts only the final fusion without the first diffusion-iteration fusion. As shown by the comparison of "GP-Unified" and "GP-Unified-w/o-first-fusion" in Fig. 4 and 6,

Table 1. PESQ performance of the predictive output of "GP-Unified" in different diffusion steps.

Diffusion steps	5	10	15	20	25	30
PESQ	2.16	2.67	2.85	2.91	2.93	2.95

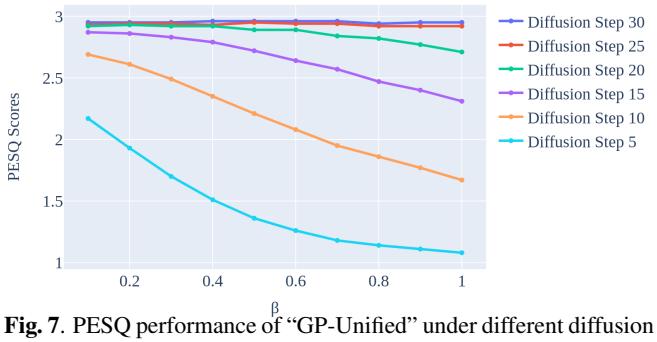

Fig. 7. PESQ performance of "GP-Unified" under different diffusion steps with different β.

the first step fusion mainly affects the diffusion speed. Besides, Fig. 5 shows that the "GP-Unified" outperforms "GP-Unified-w/o-first-fusion" even though the model already has large diffusion steps. This implies that the first step fusion not only plays the role of initialization but also compensates for some speech distortion caused by diffusion.

"GP-SGMSE+" and "GP-Unified-w/o-first-fusion" were trained by the same manner; the difference lies in whether they are fused in the final diffusion step. The final fusion step gives a significant PESQ improvement, especially when the number of diffusion steps is small, as shown in Fig. 3. Furthermore, the final step fusion can significantly reduce speech distortion (Fig. 4) and artificial noise (Fig. 6) when the number of steps is small. However, the effect is not obvious for noise suppression (Fig. 5). Table 1 shows the performance of the predictive output at different diffusion steps. The performance improves as the iteration steps increases. The proposed system can combine the characteristics of predictive and generative SE and fully use the predictive complex spectrograms even when the number of diffusion steps is small. When the diffusion steps increase, the complementarity between the generative and predicted complex spectrograms is manifested. Through the fusion, PESQ and SI-SDR can be further improved, as shown in Fig. 3 and 4.

4.3. Effect of fusion hyperparameter β

Figure 7 shows the PESQ performance of "GP-Unified" under several diffusion steps with different values of β. β has a more significant impact on smaller diffusion steps, and the advantages of the fusion are mainly reflected for smaller diffusion steps, especially in steps 5, 10, and 15. In step 5, the main performance improvement comes from predictive information. In steps 10, 15, and 20, it mainly comes from the feature fusion: the performance of "GP-Unified" was improved compared to the generative ("GP-SGMSE+") and the predictive model (results shown in the Table 2). Their difference lies in the feature fusion. The system performs better when β is smaller, which means that the generative information is incorporated into the predictive complex spectrogram.

[3]Note that we implemented UNIVERSE ourselves because the code is not publicly available. The network was trained on VB, and we only considered mel bands for feature NLLs.

Table 2. Performance of different speech enhancement systems in VB dataset: "Type" denotes the type of the system, "P" represents the predictive model, "G" represents the generative model.

System	Type	PESQ	ESTOI	SI-SDR
Mixture	-	1.97	0.79	8.4
Conv-Tasnet [20]	P	2.84	0.85	19.1
MetricGAN+ [19]	P	3.13	0.83	8.5
GaGNet [21]	P	2.94	0.86	**19.9**
SEGAN [22]	G	2.16	-	-
CDiffuSE [23]	G	2.46	0.79	12.6
SGMSE+ [24]	G	2.93	0.87	17.3
StoRM [25]	G	2.93	**0.88**	**18.8**
UNIVERSE[3]	G	2.91	0.84	10.1
GP-SGMSE+	G	2.95	0.87	17.9
GP-Unified	G	**2.97**	0.87	18.3

Table 3. RTF with different evaluation metrics: "**Score**" represents the corresponding evaluation score.

Evaluation Metrics	Model	Steps	Score	RTF
PESQ	SGMSE+	30	2.93	1.68
	StoRM	25	2.93	1.4
	GP-Unified	15	2.93	1.35
SI-SDR	SGMSE+	30	17.3	1.68
	StoRM	20	18.6	1.12
	GP-Unified	10	18.3	0.91

4.4. Comparasion with other methods

Comparison with other methods, including "StoRM" and "UNIVERSE", are listed in Table 2. Compared with the proposed method, "UNIVERSE" significantly degraded SI-SDR. This suggests that incorporating predictive information by the form of "UNIVERSE" is not beneficial to improving segment-level performance. Compared with "StoRM", the proposed method had better PESQ performance. This suggests that the proposed method can achieve better frequency-domain enhancement performance, since PESQ depends on the frequency-domain performance. Table 3 presents the real-time factor (RTF) values for "SGMSE+" and "GP-Unified". In comparison to "SGMSE+", "GP-Unified" achieves a substantial reduction in diffusion steps, resulting in improved RTF performance. Specifically, for PESQ, 15 diffusion steps are sufficient for convergence, saving 50% of steps (RTF 1.35). For SI-SDR, 10 diffusion steps are adequate to achieve a comparable score of SGMSE+ and StoRM., resulting in a 66% reduction in steps (RTF 0.91). Compared with "StoRM", the RTF is still improved.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we have proposed a unified generative and predictive speech enhancement model (GP-Unified). The model encodes both generative and predictive information and applies the generative and predictive decoders separately, whose results are fused. The predictive information helps the model to reduce speech distortion, noise, and artifacts. The two systems are fused in the first and final steps. The information fusion can speed up the diffusion process by reducing the number of diffusion steps by about 50%, which leads better RTF. Besides, information fusion can lead to better performance with the complementary between the predictive and generative SE.

6. REFERENCES

[1] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An Overview of Noise-Robust Automatic Speech Recognition," *IEEE/ACM TASLP*, vol. 22, no. 4, pp. 745–777, 2014.

[2] K. Saito, S. Uhlich, G. Fabbro, and Y. Mitsufuji, "Training speech enhancement systems with noisy speech datasets," *arXiv*, 2021.

[3] S. Uhlich and Y. Mitsufuji, "Open-unmix for speech enhancement (umx se)," *May*, 2020.

[4] A. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," in *Proc. INTERSPEECH*, 2012.

[5] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[6] M. Mimura, S. Sakai, and T. Kawahara, "Exploring deep neural networks and deep autoencoders in reverberant speech recognition," in *Proc. HSCMA*, 2014, pp. 197–201.

[7] Z. Bai and X.-L. Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[8] X. Luo, H.-H. Chen, and Q. Guo, "Semantic communications: Overview, open issues, and future research directions," *IEEE Wireless Communications*, vol. 29, no. 1, pp. 210–219, 2022.

[9] E. Grassucci, S. Barbarossa, and D. Comminiello, "Generative semantic communication: Diffusion models beyond bit recovery," 2023.

[10] F. Jiang, Y. Peng, L. Dong, K. Wang, K. Yang, C. Pan, and X. You, "Large ai model-based semantic communications," 2023.

[11] E. Grassucci, Y. Mitsufuji, P. Zhang, and D. Comminiello, "Enhancing Semantic Communication with Deep Generative Models – An ICASSP Special Session Overview," *arXiv*, 2023.

[12] Z.-Q. Wang, P. Wang, and D. Wang, "Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR," *IEEE/ACM TASLP*, vol. 28, pp. 1778–1787, 2020.

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

[14] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. ICASSP*, vol. 4, 2002, pp. IV–4164–IV–4164.

[15] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[16] D. Yin, C. Luo, Z. Xiong, and W. Zeng, "PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network," *Proc. AAAI*, vol. 34, no. 05, pp. 9458–9465, 2020.

[17] M. Kolbæk, Z.-H. Tan, S. H. Jensen, and J. Jensen, "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE/ACM TASLP*, vol. 28, pp. 825–838, 2020.

[18] H. Shi, L. Wang, M. Ge, S. Li, and J. Dang, "Spectrograms Fusion with Minimum Difference Masks Estimation for Monaural Speech Dereverberation," in *Proc. ICASSP*, 2020, pp. 7544–7548.

[19] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "Metricgan+: An improved version of metricgan for speech enhancement," *arXiv*, 2021.

[20] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[21] A. Li, C. Zheng, L. Zhang, and X. Li, "Glance and gaze: A collaborative learning framework for single-channel speech enhancement," *Applied Acoustics*, vol. 187, p. 108499, 2022.

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

[23] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional Diffusion Probabilistic Model for Speech Enhancement," in *Proc. ICASSP*, 2022, pp. 7402–7406.

[24] J. Richter, S. Welker, J.-M. Lemercier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM TASLP*, vol. 31, pp. 2351–2364, 2023.

[25] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM TASLP*, vol. 31, pp. 2724–2737, 2023.

[26] S. Leglaive, L. Girin, and R. Horaud, "A variance modeling framework based on variational autoencoders for speech enhancement," in *Proc. MLSP*, 2018, pp. 1–6.

[27] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Proc. NIPS*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019.

[28] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," in *Proc. NIPS*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 21 696–21 707.

[29] R. Sawata, N. Murata, Y. Takida, T. Uesaka, T. Shibuya, S. Takahashi, and Y. Mitsufuji, "Diffiner: A Versatile Diffusion-based Generative Refiner for Speech Enhancement," in *Proc. INTERSPEECH*, 2023, pp. 3824–3828.

[30] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal Speech Enhancement with Score-based Diffusion." arXiv, 2022.

[31] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021.

[32] B. D. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, vol. 12, no. 3, pp. 313–326, 1982.

[33] C.-W. Huang, J. H. Lim, and A. C. Courville, "A variational perspective on diffusion-based generative models and score matching," in *Proc. NIPS*, vol. 34, 2021, pp. 22 863–22 876.

[34] B. Øksendal, *Stochastic Differential Equations*. Springer Berlin Heidelberg, 2003, pp. 65–84.

[35] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech." in *Proc. SSW*, 2016, pp. 146–152.

[36] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, vol. 2, 2001, pp. 749–752 vol.2.

[37] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM TASLP*, vol. 24, no. 11, pp. 2009–2022, 2016.

[38] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?" in *Proc. ICASSP*, 2019, pp. 626–630.