Speech Enhancement based on cascaded two flows

Seonggyu Lee, Sein Cheong, Sangwook Han, Kihyuk Kim, Jong Won Shin

Gwangju Institute of Science and Technology, Korea

{lsqjin2022,seiinjung,swhan9873, kpaul073@gm.gist.ac.kr}@gm.gist.ac.kr,
jwshin@gist.ac.kr

Abstract

Speech enhancement (SE) based on diffusion probabilistic models has exhibited impressive performance, while requiring a relatively high number of function evaluations (NFE). Recently, SE based on flow matching has been proposed, which showed competitive performance with a small NFE. Early approaches adopted the noisy speech as the only conditioning variable. There have been other approaches which utilize speech enhanced with a predictive model as another conditioning variable and to sample an initial value, but they require a separate predictive model on top of the generative SE model. In this work, we propose to employ an identical model based on flow matching for both SE and generating enhanced speech used as an initial starting point and a conditioning variable. Experimental results showed that the proposed method required the same or fewer NFEs even with two cascaded generative methods while achieving equivalent or better performances to the previous baselines.[†]

Index Terms: flow matching, diffusion, conditioning, generative model, speech enhancement

1. Introduction

Speech enhancement (SE) aims to restore clean speech signals from those contaminated by environmental noises [1–15]. Traditional methods often utilize the statistical characteristics of clean speech signals and environmental noises [2, 3], but most of recent studies employ deep neural networks (DNNs) to estimate clean speech signals [4–6]. Generative approaches that focus on modeling the underlying distribution of clean speech signals have recently been introduced [7–15]. Among them, the SE based on diffusion probabilistic models which utilizes stochastic differential equations (SDEs) have demonstrated remarkable performance [10–15]. This class of approaches requires repeated evaluation of a DNN model that approximates the score function to estimate clean speech. The number of times that the DNN model is evaluated, called the number of function evaluations (NFE), typically exceeds 25 [10–14], which may limit the applicability of the diffusion model-based SE.

Flow matching (FM) to model continuous normalizing flows (CNFs) which converts a random vector following a simple distribution into another one with a complex distribution through invertible transformations has been proposed as an alternative to the diffusion probabilistic models [16, 17]. The FM with a conditional flow matching (CFM) loss and the optimal transport (OT) conditional vector field showed faster sampling

---

[†]Our codes are available at online :
https://github.com/seongq/
cascadingtwoflowmatching

and better performance than the previous diffusion models in several tasks [16, 17]. It has also been applied to speech processing such as speech separation [18], speech enhancement [18, 19], and audio-visual speech enhancement [20]. Among them, the FlowSE in [19] adopts the noisy speech as an additional condition and modifies the OT conditional vector field so that the mean of the conditional probability path moves linearly from noisy speech to clean speech and the standard deviation decreases linearly. It showed equivalent or better performance to the previous diffusion model-based SE with a fewer NFE.

The DNN models that estimate scores in the diffusion model-based SE or vector fields in the flow matching-based SE have conditioning variables other than the state and noise level at a specific time. Noisy speech is given to the model in [10–13, 18–20], while speech enhanced with an additional predictive SE model is used as a conditioning variable along with noisy speech in [14, 21–23]. The utilization of the enhanced speech brought about performance improvement, but it requires a separate predictive SE model.

In this paper, we propose a generative model that cascades two flows for SE where two flows are approximated by a single model. The first flow models the transformation from a random vector following a simple distribution centered on noisy speech to the one distributed by the probability distribution of clean speech given noisy speech just like FlowSE. The output of the first flow is used as an additional conditioning variable and the mean of the starting point for the second flow. Experimental results showed that the proposed method outperformed previously proposed approaches without increasing the total NFE.

2. Related Works

2.1. Diffusion model-based Speech Enhancement

In the diffusion model-based SE [7–13], a diffusion process describes gradual transformation of a clean speech sample x_0 onto the noisy speech y with additional Gaussian noise using a forward SDE

$$dx_t = f(x_t, y, t)dt + g(t)dw_t, \quad (1)$$

where $t \in [0, T]$, w_t is a Brownian motion, and f and g are called the drift and the diffusion coefficients, respectively. The reverse SDE governing the reverse process that transforms x_T following $\mathcal{N}(y, \sigma_T^2 \mathbf{I})$ onto x_0 is given by [24]

$$dx_t = \left[f(x_t, y, t) - g(t)^2 \nabla_{x_t} \log p_t(x_t|y) \right] dt + g(t)d\bar{w}_t \quad (2)$$

where $\nabla_{x_t} \log p_t(x_t|y)$ called a score function is the gradient of log for probability density function (pdf) of x_t given y, and $\bar{w}_t$ is a reverse Brownian motion. The score function needed to evaluate the SDE is approximated by a DNN called a score

model $s_\theta(x_t, y, t)$. The score model is trained using the denoising score matching (DSM) loss [24], $\mathcal{L}_{DSM}$, which is defined as

$$\mathcal{L}_{DSM} := \mathbf{E}\left\|s_\theta(x_t, y, t) - \nabla_{x_t} \log p_t(x_t|x_0, y)\right\|^2 \quad (3)$$

where t is randomly chosen from $U[0, T]$, a uniform distribution between 0 and T, and x_t is sampled from a distribution $p_t(x_t|x_0, y)$ called the perturbation kernel which is a Gaussian distribution determined by assumed f and g. It was reported [7–13] that NFEs greater than 25 were required to achieve their performances for those diffusion model-based SE models.

2.2. Flow Matching-based Speech Enhancement (FlowSE)

FlowSE [19] based on FM [16, 17] models a CNF which transforms a random vector following $p_1(x_1|y) := \mathcal{N}(x_1|y, \sigma^2\mathbf{I})$, where $\sigma \geq 0$ is a hyperparameter, into a clean speech distribution given a noisy speech y $q(x_0|y)$ described by an ODE:

$$\frac{d\psi_t(x_1|y)}{dt} := v_t(\psi_t(x_1|y)|y), \psi_1(x_1|y) = x_1, \quad (4)$$

where $\psi_t(x_1|y)$ and $v_t(x_t|y)$ conditioned on y are called a flow and a vector field, respectively, and $x_1 \sim p_1(x_1|y)$. The goal of this formulation is to find the vector field or the flow such that $x_t := \psi_t(x_1|y)$ has a pdf $p_t(x_t|y)$ satisfying $p_0(x_0|y) = q(x_0|y)$. It is noted that the time index t in (4) aligns with the diffusion model in the subsection 2.1, which is not same as that in [19]. FlowSE employed the modified OT conditional vector field with a conditional probability path $p_t(x_t|x_0, y) = \mathcal{N}(x_t|\mu_t(x_0, y), \sigma_t^2\mathbf{I})$ in which

$$\mu_t(x_0, y) = (1-t)x_0 + ty, \quad \sigma_t = t\sigma. \quad (5)$$

Then, the target vector field $v_t(x_t|x_0, y)$ becomes

$$v_t(x_t|x_0, y) = \frac{\frac{d}{dt}\sigma_t}{\sigma_t}(x_t - \mu_t(x_0, y)) + \frac{d}{dt}\mu_t(x_0, y). \quad (6)$$

The vector field model $v_\theta(x_t, y, t)$ is trained with the CFM loss $\mathcal{L}_{CFM}$ given by

$$\mathcal{L}_{CFM} := \mathbf{E}\|v_\theta(x_t, y, t) - v_t(x_t|x_0, y)\|^2, \quad (7)$$

where t is from $U[t_\delta, 1]$ with $0 < t_\delta < 1$ and x_t is from the conditional probability path $p_t(x_t|x_0, y)$.

In the inference phase, $v_\theta(x_t, y, t)$ is numerically integrated starting with x_1 sampled from $p_1(x_1|y)$. Euler method is adopted as the numerical integrator in [19]. Given N time points $t_0 = 0 < t_1 = t_\delta < t_2 < ... < t_N = 1$ in $[0, 1]$, the clean speech estimate x_0 is generated from

$$x_{t_{i-1}} = x_{t_i} + (t_{i-1} - t_i)v_\theta(x_{t_i}, y, t_i). \quad (8)$$

FlowSE with the NFE of 5 achieved performance comparable to a diffusion-based model [13] with the NFE of 60 and the fine-tuning method for the diffusion-based model [25] with the NFE of 5 [19]. It was also shown that FlowSE can be interpreted as a diffusion model-based SE model with a specific SDE [19].

2.3. Diffusion-based Stochastic Regenration Model for SE (StoRM)

StoRM [14] consists of a predictive model and a generative model based on a diffusion model. The predictive model denoted as D_ϕ estimates a clean speech x_0 from noisy speech y,

and the estimated speech $D_\phi(y)$ is used as an additional input to the score model $s_\theta(x_t, y, D_\phi(y), t)$ and also utilized to sample the starting point of the reverse process. The loss function to train these models is the weighted summation of the mean squared error (MSE) loss $\mathcal{L}_1$ for D_ϕ and the DSM loss $\mathcal{L}_2$ for the score model, i.e.,

$$\mathcal{L}^{StoRM} = \alpha\mathcal{L}_1 + \mathcal{L}_2, \quad (9)$$

where $\alpha > 0$ is a hyperparameter and

$$\mathcal{L}_1 := \mathbf{E}\|D_\phi(y) - x_0\|^2, \quad (10)$$

$$\mathcal{L}_2 := \mathbf{E}\|s_\theta(x_t, y, D_\phi(y), t) - \nabla_{x_t} \log p_t(x_t|x_0, D_\phi(y))\|^2, \quad (11)$$

with t from $\mathcal{U}[0, T]$ and x_t sampled from a perturbation kernel $p_t(x_t|x_0, D_\phi(y))$.

During the inference stage of StoRM, $D_\phi(y)$ is evaluated first and then a clean speech is estimated by integrating the reverse SDE (2) starting from $\mathcal{N}(D_\phi(y), \sigma_T^2 I)$ using the score model $s_\theta(x_t, y, D_\phi(y), t)$ with conditioning variables y and $D_\phi(y)$. StoRM showed better performance than the early diffusion-based model [10] with only noisy speech as a conditioning variable [14].

3. Cascading Two Flows for Speech Enhancement (CTFSE)

We design the first flow to transform a random vector x_1 following $p_1(x_1|y) = \mathcal{N}(x_1|y, \sigma^2\mathbf{I})$ into a crude estimate of clean speech, $D_\theta(x_1, y)$, which is ideally distributed according to $q(x_0|y)$. And then, the second flow starts with a sample drawn from a Gaussian distribution centered on $D_\theta(x_1, y)$, $\mathcal{N}(D_\theta(x_1, y), \sigma^2\mathbf{I})$, and then transforms it using the ODE in (4) using the trained vector field model conditioned by both y and $D_\theta(x_1, y)$ to finally obtain $\tilde{x}_0$. We use a single vector field model v_θ for both of the flows by configuring the conditioning variable for the second flow to be a simple summation of y and $D_\theta(x_1, y)$, which is proven to work as a fusion method in many researches [26–28]. The loss for the first flow is the CFM loss in (7) :

$$\mathcal{L}_1 := \mathbf{E}\|v_\theta(x_t, y, t) - v_t(x_t|x_0, y)\|^2, \quad (12)$$

where t is from $\mathcal{U}[t_\delta, 1]$ and x_t follows $p_t(x_t|x_0, y)$. To control the computational complexity when two flows are adopted, we fix the number of time steps for the first flow to 1. Then, $D_\theta(x_1, y)$ is obtained by a simple equation if the Euler method is used for one time step:

$$D_\theta(x_1, y) = x_1 - v_\theta(x_1, y, 1). \quad (13)$$

The loss for the second flow is again the CFM loss, with different conditioning variables in the vector field model and the target conditional vector field:

$$\mathcal{L}_2 := \mathbf{E}\left\|v_\theta\left(\tilde{x}_t, \frac{D_\theta(x_1, y) + y}{2}, t\right) - v_t(\tilde{x}_t|x_0, D_\theta(x_1, y))\right\|^2, \quad (14)$$

where t is from $\mathcal{U}[t_\delta, 1]$ and $\tilde{x}_t$ is sampled from $p_t(\tilde{x}_t|x_0, D_\theta(x_1, y))$. Additionally, we use the CFM loss $\mathcal{L}_3$ when t is fixed to 1, to enforce $D_\theta(x_1, y)$ to be close to x_0:

$$\mathcal{L}_3 := \mathbf{E}\|v_\theta(x_1, y, 1) - v_1(x_1|x_0, y)\|^2, \quad (15)$$
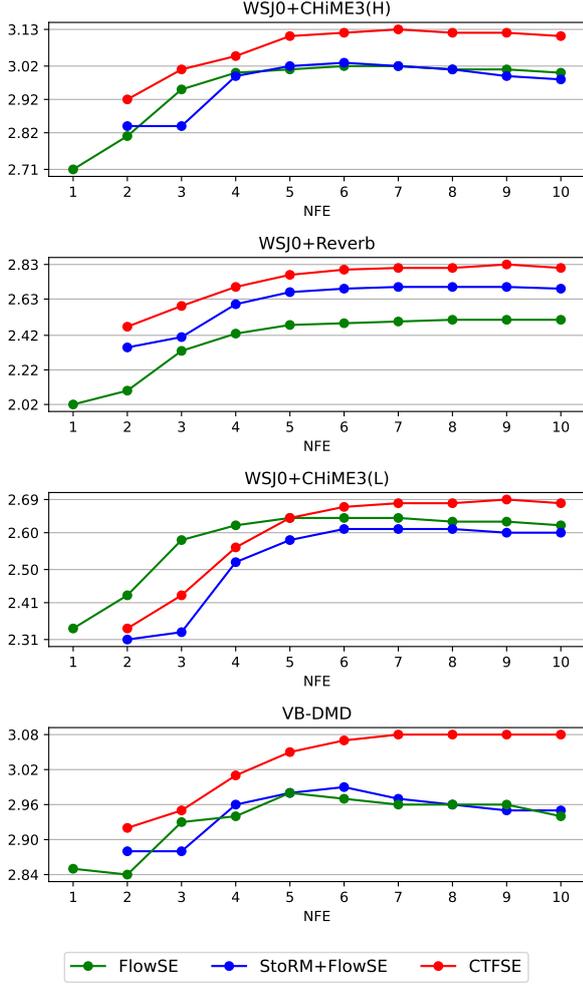
Figure 1: *Comparison of WB-PESQ scores as a function of the NFE for FlowSE, StoRM+FlowSE, and CTFSE.*

where x_1 follows $p_1(x_1|x_0, y)$. It can be shown that $\mathcal{L}_3$ becomes the MSE loss between $D_\theta(x_1, y)$ and a clean speech x_0. The total loss $\mathcal{L}_{CTF}$ for training is given as a weighted summation of $\mathcal{L}_1, \mathcal{L}_2$ and $\mathcal{L}_3$

$$\mathcal{L}_{CTF} := \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 \qquad (16)$$

where $\lambda_1, \lambda_2, \lambda_3 \geq 0$ are hyperparameters. It is noted that we use the identical model v_θ to produce $D_\theta(x_1, y)$ and to model the second flow, while StoRM used two different models.

In the inference phase of CTFSE, x_1 is sampled from $p_1(x_1|y)$ first and using the Euler method with only one time point, $D_\theta(x_1, y)$ is generated by (13). Given N time points $t_0 = 0 < t_1 = t_\delta < t_2 < ... < t_N = 1$, a clean speech estimate $\tilde{x}_0$ is generated by the Euler method

$$\tilde{x}_{t_N} \sim \mathcal{N}(D_\theta(x_1, y), \sigma^2 \mathbf{I})$$

$$\tilde{x}_{t_{i-1}} = \tilde{x}_{t_i} + (t_{i-1} - t_i) v_\theta \left(\tilde{x}_{t_i}, \frac{D_\theta(x_1, y) + y}{2}, t_i \right). \qquad (17)$$

4. Experimental settings

We evaluated the performance of the proposed and compared methods for two versions of WSJ0+CHiME3 dataset,

WSJ0+Reverb dataset, and VoiceBank-DEMAND (VB-DMD) dataset. WSJ0+CHiME3 (H) and WSJ0+CHiME (L) datasets were constructed by mixing clean speech utterances from the Wall Street Journal (WSJ0) dataset [29] and environmental noises from the CHiME3 [30] dataset with the signal-to-noise ratio (SNR) between 0 and 20 dB for WSJ0+CHiME3 (H), and -4 and 16 dB for WSJ0+CHiME3 (L). WSJ0+Reverb was constructed convolving each utterance from the WSJ0 dataset with a simulated room impulse response (RIR) as in [14]. The dimension of the room was randomly chosen from [5,15], [5,15], and [2,6] m for the length, width, and height, respectively, and the T_{60} was selected in [0.4,1.0] s. Then anechoic target speech is generated by simulating the same room with an absorption coefficient of 0.99. These three datasets, WSJ0+CHiME3 (H), WSJ0+CHiME3 (L) and WSJ0+Reverb, were created using the source codes[1] provided by the authors of [14]. The VB-DMD dataset [31], which is publicly available, is generated by mixing clean speech from the VCTK dataset [32] with eight real-recorded noise samples from the DEMAND database [33] and two artificially generated noise samples (babble and speech shaped) at SNRs of 0,5,10, and 15 dB. The SNRs for the test set are 2.5, 7.5, 12.5 and 17.5 dB.

The clean and noisy speech signals, x_0, y, are the magnitude-compressed complex-valued spectrograms in $\mathbb{C}^{K \times F}$ as in [14]. We used a lighter configuration of the NCSN++ architecutre denoted NCSN++M as in [14, 34] for the neural network v_θ. NCSN++M has 27.8 M parameters, while NCSN++ has 65.0 M parameters. We trained the neural network v_θ using Adam optimizer [35] with a learning rate of 0.0001 and a batch size of 4. An exponential moving average with a decay of 0.999 was utilized. t_δ and σ were set to 0.03 and 0.5, respectively. We set $\lambda_1 = \lambda_2 = \lambda_3 = 1$ in (16). We trained the model for a maximum of 1,000 epochs with early stopping based on the validation loss with a patience of 50 epochs. For the generation process in the subsection 3.2, the time points $0 = t_0 < t_1 = t_\delta < t_2 < ... < t_N = 1$ were chosen so that $t_{i-1} - t_i$ has the same value for $i \in \{2, ..., N\}$ for $N \geq 2$. In the case of $N = 1$, we set $t_0 = 0$ and $t_1 = 1$.

The diffusion model-based SE method SGMSE+M [14], StoRM [14] with a predictive model and a score model, and a flow matching-based model FlowSE [19] were compared with the proposed CTFSE. Additionally, we have implemented the StoRM system with a flow matching-based second stage, StoRM+FlowSE, and compared it with other systems. All architectures for the neural networks were NCSN++M and only the number of conditioning variables differ from each other. For numerical intergration for SDEs, we adopted the predictor-corrector scheme [24] with the Euler-Maruyama method as a predictor and one step of annealed Langevin dynamics correction for SGMSE+M and StoRM, and the Euler method for FlowSE and StoRM+FlowSE for ODEs. We utilized the pre-trained SGMSE+M, StoRM models from checkpoints[1] shared by the authors of [14], and implemented the remaining baseline models.

We have evaluated the wideband extension to perceptual evaluation of speech quality (WB-PESQ) scores [36], Deep Noise Supression mean opinion score (DNSMOS) [37], Extended Short-Time Objective Intelligibility (ESTOI) [38], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) [39], wav2vec MOS (WVMOS) [40], and DNSMOS P.835 including SIG, BAK and OVRL [41].

---

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

Table 1: *Speech enhancement performances for proposed and compared methods on the WSJ0+CHiME3 (H), WSJ0+CHiME3 (L), WSJ0+REVERB and VB-DMD datasets. * indicates that the results of the model come from checkpoints shared by the authors of [14].*

METHOD	NFE	WB-PESQ	ESTOI	SI-SDR	WVMOS	DNSMOS	SIG	BAK	OVRL
Tranined and Tested on WSJ0+CHiME3 (H)									
SGMSE+M	100	2.82±0.04	0.92±0.00	17.44±0.34	3.77±0.02	3.93±0.02	3.55±0.01	**4.19±0.00**	3.33±0.01
StoRM	101	2.91±0.04	0.92±0.00	17.73±0.33	3.77±0.03	4.00±0.01	**3.60±0.01**	4.12±0.01	3.32±0.01
FlowSE	6	3.02±0.04	0.93±0.00	18.71±0.34	3.86±0.03	4.03±0.01	**3.60±0.01**	4.16±0.00	3.35±0.01
StoRM+FlowSE	6	3.03±0.04	0.93±0.00	18.83±0.34	3.86±0.03	4.04±0.01	**3.60±0.01**	**4.19±0.00**	**3.37±0.01**
CTFSE	6	**3.12±0.04**	**0.94±0.00**	**19.37±0.33**	**4.02±0.02**	**4.05±0.01**	**3.60±0.01**	**4.19±0.00**	**3.37±0.01**
CTFSE	7	**3.13±0.04**	**0.94±0.00**	19.20±0.33	3.96±0.02	**4.05±0.01**	**3.60±0.01**	**4.19±0.00**	**3.37±0.01**

METHOD	NFE	PESQ	ESTOI	SI-SDR	WVMOS	DNSMOS	SIG	BAK	OVRL
Tranined and Tested on WSJ0+CHiME3 (L)									
SGMSE+M*	100	2.30±0.05	0.85±0.01	13.19±0.38	3.65±0.03	3.83±0.02	3.51±0.01	4.19±0.00	3.28±0.01
StoRM*	101	2.55±0.05	0.88±0.01	14.91±0.33	3.73±0.03	4.00±0.01	3.57±0.01	4.05±0.01	3.26±0.01
FlowSE	5	2.64±0.05	**0.89±0.01**	15.34±0.33	3.72±0.03	4.01±0.01	3.57±0.01	4.19±0.00	3.34±0.01
StoRM+FlowSE	6	2.61±0.05	**0.89±0.01**	15.46±0.33	3.69±0.03	4.02±0.01	3.59±0.01	4.18±0.00	3.35±0.01
CTFSE	5	2.64±0.05	**0.89±0.01**	**15.96±0.33**	**3.85±0.03**	4.01±0.02	3.57±0.01	**4.20±0.00**	3.34±0.01
CTFSE	9	**2.69±0.05**	**0.89±0.01**	15.34±0.33	3.71±0.03	**4.03±0.01**	**3.60±0.01**	4.19±0.00	**3.36±0.01**

METHOD	NFE	PESQ	ESTOI	SI-SDR	WVMOS	DNSMOS	SIG	BAK	OVRL
Tranined and Tested on WSJ0+Reverb									
SGMSE+M*	100	2.33±0.03	0.82±0.01	-0.21±0.67	3.43±0.03	3.89±0.02	3.20±0.02	4.05±0.01	2.84±0.02
StoRM*	101	2.52±0.03	0.85±0.00	5.54±0.32	3.61±0.03	3.97±0.01	**3.29±0.02**	4.10±0.01	**2.98±0.02**
FlowSE	9	2.51±0.03	0.85±0.00	4.01±0.35	3.59±0.03	3.99±0.01	3.24±0.02	**4.13±0.00**	2.91±0.02
StoRM+FlowSE	9	2.70±0.03	0.87±0.00	6.37±0.29	3.74±0.02	**4.01±0.01**	3.25±0.02	**4.13±0.00**	2.92±0.02
CTFSE	9	**2.83±0.03**	**0.89±0.00**	**7.25±0.30**	**3.77±0.02**	**4.01±0.01**	3.27±0.02	**4.13±0.00**	2.95±0.02

METHOD	NFE	PESQ	ESTOI	SI-SDR	WVMOS	DNSMOS	SIG	BAK	OVRL
Tranined and Tested on VB-DMD									
SGMSE+M	100	2.80±0.04	0.86±0.01	16.19±0.39	4.27±0.02	3.54±0.02	3.48±0.01	3.95±0.02	3.15±0.02
StoRM*	101	2.90±0.04	0.87±0.01	18.48±0.23	4.29±0.02	3.56±0.02	**3.50±0.01**	4.02±0.01	**3.20±0.01**
FlowSE	5	2.98±0.05	0.87±0.01	18.97±0.23	**4.30±0.02**	**3.58±0.02**	3.48±0.01	**4.05±0.01**	**3.20±0.01**
StoRM+FlowSE	6	2.99±0.05	0.87±0.01	18.65±0.24	4.26±0.03	**3.58±0.02**	3.49±0.01	4.02±0.01	**3.20±0.01**
CTFSE	5	3.05±0.05	**0.88±0.01**	**19.13±0.24**	4.26±0.03	**3.58±0.02**	3.48±0.01	4.03±0.01	**3.20±0.01**
CTFSE	7	**3.08±0.05**	0.87±0.01	18.97±0.24	4.26±0.02	**3.58±0.02**	3.49±0.01	4.03±0.01	**3.20±0.01**

5. Results

Table 1 summarizes the performances with 95% confidence intervals and NFEs for the proposed and compared methods on the four datasets, WSJ0+CHiME3 (H), WSJ0+CHiME3 (L), WSJ0+Reverb, and VB-DMD. The NFEs for SGMSE+M and StoRM were set to 100 and 101 as in [14], and those for FlowSE, StoRM+FlowSE and CTFSE were selected to maximize the average WB-PESQ scores. For fair comparison, the performances for the CTFSE with the NFE of 5 or 6 are also shown when the compared methods have lower NFE of 5 or 6. On average, the proposed CTFSE showed the best performance for most of the mesures with the NFE less than 10. Compared with StoRM+FlowSE which had the same NFE but twice the parameters, the proposed CTFSE exhibited similar or better performances and the performance improvement was bigger for WSJ0+CHiME3 (H) and WSJ0+Reverb.

Figure 1 shows the WB-PESQ scores as a function of NFE for FlowSE, StoRM+FlowSE, and CTFSE. We can see that CTFSE outperformed other methods at the same NFE except for the WSJ0+CHiME3 (L) with the NFE less than 5.

6. Conclusion

In this work, we proposed a speech enhancement cascading two flows with the same vector field model. The first flow produces the crude estimate of clean speech which is used as the mean of the starting point of the second flow and summed up to noisy speech to be used as a conditioning variable for the second flow. The loss function to train the vector field model is the weighted combination of the CFM losses for two flows. Experimental results demonstrated that the proposed method showed comparable or better performance to the previously proposed diffusion model- or flow matching-based SE methods for four datasets.

7. Acklowedgements

8. References

[1] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2007.

[2] M. Kim and J. W. Shin, "Improved speech enhancement considering speech psd uncertainty," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 30, pp. 1939–1951, 2022.

[3] S. Cheong, M. Kim, and J. W. Shin, "Postfilter for dual channel speech enhancement using coherence and statistical model-based noise estimation," *Sensors*, vol. 24, no. 12, 2024.

[4] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *TASLP*, vol. 28, pp. 380–390, 2020.

[5] H. Kim, K. Kang, and J. W. Shin, "Factorized mvdr deep beamforming for multi-channel speech enhancement," *IEEE Signal Processing Letters*, vol. 29, pp. 1898–1902, 2022.

[6] H. Kim and J. W. Shin, "Target exaggeration for deep learning-based speech enhancement," *Digital Signal Processing*, vol. 116, pp. 103–109, 2021.

[7] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 106–110.

[8] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *TASLP*, vol. 28, pp. 1104–1117, 2020.

[9] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *TASLP*, vol. 30, pp. 2993–3007, 2022.

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

[11] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP*, 2023.

[12] Z. Guo, Q. Wang, J. Du, J. Pan, Q.-F. Liu, and C.-H. Lee, "A variance-preserving interpolation approach for diffusion models with applications to single channel speech enhancement and recognition," *TASLP*, vol. 32, pp. 3025–3038, 2024.

[13] B. Lay, S. Welker, J. Richter, and T. Gerkmann, "Reducing the prior mismatch of stochastic differential equations for diffusion-based speech enhancement," in *Interspeech*, 2023.

[14] J.-M. Lermercier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *TASLP*, 2023.

[15] T. Trachu, C. Piansaddhayanon, and E. Chuangsuwanich, "Unified regression-diffusion speech enhancement with a single reverse step using brownian bridge," in *interpseech*, 2024.

[16] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations (ICLR)*, 2023.

[17] A. Tong, K. FATRAS, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio, "Improving and generalizing flow-based generative models with minibatch optimal transport," *Transactions on Machine Learning Research*, 2024.

[18] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *ICLR*, 2024.

[19] S. Lee, S. Cheong, S. Han, and J. W. Shin, "Flowse: Flow matching-based speech enhancement," in *ICASSP*, 2025.

[20] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, "Flowavse: Efficient audio-visual speech enhancement with conditional flow matching," in *Interspeech*, 2024, pp. 2210–2214.

[21] W. Tai, F. Zhou, G. Trajcevski, and T. Zhong, "Revisiting denoising diffusion probabilistic models for speech enhancement: Condition collapse, efficiency and refinement," in *AAAI*, vol. 37, no. 11, 2023, pp. 13 627–13 635.

[22] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv preprint*, vol. arXiv:2206.03065, 2022.

[23] D. Kim, D. H. Yang, D. Kim, J. H. Chang, J. Yang, J. Choi, M. Lee, and H. g. Moon, "Guided conditioning with predictive network on score-based diffusion model for speech enhancement," in *Interspeech*, 2024, pp. 1190–1194.

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

[25] B. Lay, J.-M. Lermercier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement," in *ICASSP*, 2024, pp. 626–630.

[26] J. Byun and J. W. Shin, "Monaural speech separation using speaker embedding from preliminary separation," *TASLP*, vol. 29, pp. 2753–2763, 2021.

[27] A. Li, W. Liu, C. Zheng, C. Fan, and X. Li, "Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement," *TASLP*, vol. 29, 2021.

[28] X. Hao, X. Su, S. Wen, Z. Wang, Y. Pan, F. Bao, and W. Chen, "Masking and inpainting: A two-stage speech enhancement approach for low snr and non-stationary noise," in *ICASSP*, 2020, pp. 6959–6963.

[29] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," 1993.

[30] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

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

[32] 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. Int. Conf. Oriental COCOSDA Held Jointly With Conf. Asian Spoken Lang. Res. Eval. (O-COCOSDA/CASLRE),*. IEEE, 2013, pp. 1–4.

[33] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19. AIP Publishing, 2013.

[34] J.-M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP*, 2023.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] *Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codec*, International Telecommunication Union, Geneva, 2007, iTU-T Recommendation P.862.2.

[37] O. F. Salas, V. Adzic, and H. Kalva, "Subjective quality evaluations using crowdsourcing," in *2013 Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 418–421.

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

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

[40] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *ICASSP*, 2023.

[41] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2022, pp. 886–890.