

FlowSE: Flow Matching-based Speech Enhancement

Seonggyu Lee

Gwangju Institute of Science and Technology
Gwangju, Korea
lsqjin2022@gm.gist.ac.kr

Sangwook Han

Gwangju Institute of Science and Technology
Gwangju, Korea
swhan9873@gm.gist.ac.kr

Sein Cheong

Gwangju Institute of Science and Technology
Gwangju, Korea
seiinjung@gm.gist.ac.kr

Jong Won Shin

Gwangju Institute of Science and Technology
Gwangju, Korea
jwshin@gist.ac.kr

Abstract—Diffusion probabilistic models have shown impressive performance for speech enhancement, but they typically require 25 to 60 function evaluations in the inference phase, resulting in heavy computational complexity. Recently, a fine-tuning method was proposed to correct the reverse process, which significantly lowered the number of function evaluations (NFE). Flow matching is a method to train continuous normalizing flows which model probability paths from known distributions to unknown distributions including those described by diffusion processes. In this paper, we propose a speech enhancement based on conditional flow matching. The proposed method achieved the performance comparable to those for the diffusion-based speech enhancement with the NFE of 60 when the NFE was 5, and showed similar performance with the diffusion model correcting the reverse process at the same NFE from 1 to 5 without additional fine tuning procedure. We also have shown that the corresponding diffusion model derived from the conditional probability path with a modified optimal transport conditional vector field demonstrated similar performances with the NFE of 5 without any fine-tuning procedure.

Index Terms—speech enhancement, generative model, flow matching, diffusion model

I. INTRODUCTION

The objective of speech enhancement (SE) is to recover the clean speech signals from those corrupted by environmental noises [1], [2]. Classical approaches often utilize statistical properties of the clean speech signals and the environmental noises [3]–[6], but most of the recent studies utilize deep neural networks (DNNs) to estimate clean speech signals or masks to be multiplied to noisy signals or features [7]–[10]. Recently, generative approaches for SE learning the underlying distribution of clean speech signals have been proposed [11]–[20]. Among them, score-based generative models or diffusion models with stochastic differential equations (SDEs) have shown impressive results for SE [14]–[20]. An estimate of the clean speech signal is obtained by solving the reverse SDEs of diffusion models, which requires the estimation of the scores using DNNs repeatedly. The number of function evaluations (NFEs), *i.e.*, the number of running a DNN model to evaluate scores in the inference phase, were more than 25 in [14]–[19], which may limit the applications for the diffusion model-based SE. In [20], a fine-tuning method for the score network was proposed, which reduced the NFE to 5 with similar performance to the method in [17] with the NFE of 60 by correcting the reverse process (CRP).

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2021-0-01835) and Artificial Intelligence Graduate School Program (No.2019-0-01842) supervised by the IITP (Institute of Information Communications Technology Planning Evaluation).

Source codes are available online at: <https://github.com/seonggy/flowmse>

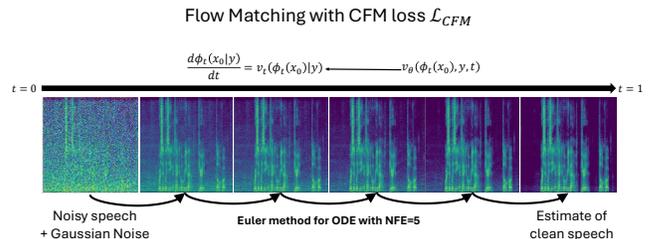


Fig. 1. Illustration of an exemplary trajectory of an ODE trained with CFM loss, which transforms noisy speech corrupted by Gaussian noise into clean speech using the Euler method with the NFE of 5.

As a potential alternative to the diffusion models, flow matching (FM) to train continuous normalizing flows (CNFs) has been proposed [21], [22]. FM has been utilized in applications such as generative models to downstream tasks in speech processing [23] and audio-visual speech enhancement [24]. A flow is an invertible mapping which gradually transforms a random vector following a simple distribution into another random vector with a complex distribution, and a CNF describes a flow with an ordinary differential equation (ODE) [25]. The FM tries to estimate the time-dependent vector field, which is the time derivative of the flow, but the ground-truth value for the vector field is often intractable. In [21], the conditional flow matching (CFM) which estimates conditional vector fields given the training sample following the complex distribution was proposed, which is proven to be equivalent to the FM. As an example, optimal transport (OT) conditional vector field was suggested, providing faster sampling and better performance than previous diffusion models [21]. Furthermore, it was shown that existing diffusion models can be trained with FM, leading faster sampling compared to training with score matching (SM) used for the diffusion models [21].

In this paper, we propose FlowSE: flow matching-based speech enhancement as an alternative to previous diffusion-based speech enhancement models, which is illustrated in Fig. 1. In the FlowSE, the noisy speech is used as a conditioning variable along with the clean speech for the conditional flow and the conditional vector field, and the OT conditional vector field is modified to incorporate the noisy speech. Our experimental results demonstrate that the performance of FlowSE with the NFE of 5 is comparable to that of the previous diffusion model with the NFE of 60, and the fine-tuned CRP model. We also present an SDE to explain how FlowSE can be considered as a diffusion model and train this diffusion model using SM, achieving similar performance to FlowSE trained with CFM when the NFE was 5 without the fine-tuning procedure in CRP.

II. BACKGROUND

A. Speech Enhancement using Diffusion Models

SE is the process of estimating a clean speech s from a noisy speech $y = s + n$, suppressing an additive environmental noise n . In the diffusion model-based SE [14]–[20], a diffusion process which gradually transforms a clean speech x_0 into a mixture of the noisy speech and Gaussian noise x_T is defined as a following 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, f and g are called the drift and the diffusion coefficients, respectively. The reverse process to transform \tilde{x}_T back to \tilde{x}_0 can be described by a reverse SDE [26]:

$$d\tilde{x}_t = [f(\tilde{x}_t, y, t) - g(t)^2 \nabla_{\tilde{x}_t} \log p_t(\tilde{x}_t|y)] dt + g(t)d\tilde{w}_t, \quad (2)$$

where $\nabla_{x_t} \log p_t(x_t|y)$, called a score function, is the gradient of log of the probability density function (pdf) of x_t given y , and \tilde{w}_t is a reverse Brownian motion. As the f and g are pre-defined functions, the estimate of the clean speech can be obtained by solving the reverse SDE (2) from $t = T$ to 0 with a starting point \tilde{x}_T sampled from $\mathcal{N}(y, \sigma_T^2 \mathbf{I})$, once the score function $\nabla_{x_t} \log p_t(x_t|y)$ is estimated. The score function is approximated by a neural network called a score network $s_\theta(x_t, y, t)$ parametrized by θ , which is trained with the denoising score matching (DSM) loss [26] $\mathcal{L}_{DSM}(\theta)$ defined as

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

where t follows a uniform distribution between 0 and T , $\mathcal{U}[0, T]$, and x_t is sampled from a perturbation kernel $p_t(x_t|x_0, y) = \mathcal{N}(\mu_t(x_0, y), \sigma_t^2 \mathbf{I})$, in which μ_t and σ_t^2 depend on f and g .

B. Brownian Bridge with Exponential Diffusion Coefficient (BBED)

In [17], a diffusion model with the SDE called Brownian Bridge with Exponential Diffusion Coefficient (BBED) is proposed, which has shown the best performance on the WSJ0-CHiME3 dataset to the best our knowledge. The drift and diffusion coefficients for the forward BBED SDE are given by

$$f(x_t, y, t) = \frac{y - x_t}{1 - t}, \quad (4)$$

$$g(t) = ck^t, \text{ where } c, k > 0, \quad (5)$$

for $0 \leq t \leq T \leq 1$. It is reported that the best result of BBED was achieved with the NFE of 60 when the time point at which the reverse SDE started, t_{rsp} , equals to $T = 0.999$, but it is also shown that a slightly lower performance was achieved with the NFE of 30 when t_{rsp} was 0.5 [17].

C. Correcting the Reverse Process (CRP)

To reduce the NFE of diffusion model BBED for SE, a fine-tuning method called CRP [20] was proposed to correct the error in discretizing reverse SDE too coarsely by actually running the reverse SDE for each training sample and minimizing the estimation error [20]. Let $s_{\theta, BBED}(x_t, y, t)$ be the pre-trained score model of BBED, n_{rev} be a fixed positive integer denoting the number of time points used in the discretized reverse process, and $t_\delta \approx 0$ be a sufficiently small positive real number. Let $t_{n_{rev}} = t_{rsp} > t_{n_{rev}-1} > \dots > t_1 = t_\delta > t_0 = 0$ and $\Delta t(i) = t_{i-1} - t_i$. The estimate of the clean speech, \tilde{x}_0 , can

be obtained by solving the discretized reverse SDE using the Euler-Maruyama (EuM) method [26] starting from $\tilde{x}_{t_{n_{rev}}}$ sampled from $\mathcal{N}(y, \sigma_{t_{n_{rev}}}^2 \mathbf{I})$, which is given by

$$\tilde{x}_{t_{i-1}} = [f(\tilde{x}_{t_i}, y, t_i) - g(t_i)^2 s_{\theta, BBED}(\tilde{x}_{t_i}, y, t_i)] \Delta t(i) + g(t_i) \sqrt{-\Delta t(i)} \epsilon_{t_i} + \tilde{x}_{t_i}, \quad (6)$$

where ϵ_{t_i} is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The parameters of $s_{\theta, BBED}(\cdot, \cdot, t_1)$ is fine-tuned with the CRP loss, $\mathcal{L}_{CRP}(\theta)$, which is the mean square error between \tilde{x}_0 and x_0 , i.e.,

$$\mathcal{L}_{CRP}(\theta) = \mathbf{E}_{y, (x_0, y), \tilde{x}_0 | (x_0, y)} \|\tilde{x}_0 - x_0\|^2. \quad (7)$$

As reported in [20], it was shown that CRP with $n_{rev} = 5$ evaluated with the NFE of 5 achieved a nearly equivalent performance to BBED with the NFE of 60, with the proposed fine-tuning procedure.

D. Flow matching

FM [21], [22] is a method to train CNFs [25]. CNFs are continuous time versions of normalizing flows [27], which models an invertible mapping from a simple space with a known distribution $p(x_0)$, e.g., $\mathcal{N}(\mathbf{0}, \mathbf{I})$, to another space distributed by an unknown distribution $q(x_1)$, when only samples from $q(x_1)$ are accessible. The invertible mapping $\phi_t : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$, called a flow, is defined by a time-dependent vector field $v_t : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ with an ODE with an initial condition at $t = 0$:

$$\frac{d}{dt} \phi_t(x_0) = v_t(\phi_t(x_0)), \quad \phi_0(x_0) = x_0 \quad (8)$$

where d is a dimension of a data sample. It is noted that the pdf of $x_t = \phi_t(x_0)$, $p_t(x_t)$, can be represented as

$$p_t(x_t) = p_0(\phi_t^{-1}(x_t)) \det \left[\frac{\partial \phi_t^{-1}(x_t)}{\partial x_t} \right]. \quad (9)$$

$p_t(x_t)$ for $t \in [0, 1]$, $p_t : [0, 1] \times \mathbb{R}^d \rightarrow [0, \infty)$, is called a probability density path or a probability path generated by a flow ϕ_t or a time-dependent vector field v_t .

Training CNFs aims to estimate a flow ϕ_t , or equivalently a time-dependent vector field v_t such that $p_0(x_0) := p(x_0)$ and $p_1(x_1) \approx q(x_1)$. The vector field $v_t(x)$ generating the probability density path p_t can be learned by a neural network $v_\theta(x, t)$ [25]. If a target probability density path p_t and a vector field v_t generating p_t are given, the target vector field can be learned by optimizing the FM loss $\mathcal{L}_{FM}(\theta)$, which is defined by

$$\mathcal{L}_{FM}(\theta) := \mathbf{E}_{t, p_t(x_t)} \|v_\theta(x_t, t) - v_t(x_t)\|^2, \quad (10)$$

where $t \sim \mathcal{U}[0, 1]$. However, the FM loss can not be used in practice, because the vector field v_t and the target probability path p_t are often intractable. In [21], a CFM loss is proposed, in which mathematically tractable $p_t(x_t|x_1)$ and $v_t(x_t|x_1)$ are considered instead of $p_t(x_t)$ and $v_t(x_t)$. The conditional flow $\phi_t(x_0|x_1)$ and the conditional vector field $v_t(x_t|x_1)$ generate the conditional probability path $p_t(x_t|x_1)$ and follow the same ODE in (8). The CFM loss given by

$$\mathcal{L}_{CFM}(\theta) := \mathbf{E}_{t, x_1, p_t(x_t|x_1)} \|v_\theta(x_t, t) - v_t(x_t|x_1)\|^2. \quad (11)$$

satisfies $\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{CFM}(\theta)$ under mild regularity conditions [21]. For a Gaussian conditional probability path $p_t(x_t|x_1) = \mathcal{N}(\mu_t(x_1), \sigma_t^2(x_1) \mathbf{I})$ in which μ_0 and σ_0 do not depend on x_1 , i.e., $p_0(x_0|x_1) = p_0(x_0) = p(x_0)$, the conditional flow and the conditional vector field generating $p_t(x_t|x_1)$ have closed forms:

$$\phi_t(x_0|x_1) = \frac{\sigma_t(x_1)}{\sigma_0} (x_0 - \mu_0) + \mu_t(x_1), \quad (12)$$

$$v_t(x_t|x_1) = \frac{\sigma'_t(x_1)}{\sigma_t(x_1)}(x_t - \mu_t(x_1)) + \mu'_t(x_1), \quad (13)$$

where $\sigma'_t = \frac{d}{dt}\sigma_t$, $\mu'_t = \frac{d}{dt}\mu_t$. Since $x_t = \phi_t(x_0|x_1)$, the CFM loss in (11) can be alternatively written as

$$\mathcal{L}_{CFM}(\theta) := \mathbf{E}_{t,x_1,p(x_0)} \|v_\theta(\phi_t(x_0|x_1), t) - v_t(\phi_t(x_0|x_1)|x_1)\|^2. \quad (14)$$

It is noted that μ_t and σ_t should be pre-determined to define the Gaussian conditional path $p_t(x_t|x_1) = \mathcal{N}(\mu_t(x_1), \sigma_t^2(x_1)\mathbf{I})$. As a useful example, the mean and the standard deviation corresponding to OT conditional vector field were introduced in [21]:

$$\mu_t(x_1) = tx_1, \quad \sigma_t = 1 - (1 - \sigma)t. \quad (15)$$

III. METHOD

A. FlowSE: Flow Matching-based Speech Enhancement

To adopt the flow matching for speech enhancement in which noisy speech y is given, the ODE of the flow in (8) is modified to incorporate y in a similar way to the SDE in (1):

$$\frac{d\phi_t(x_0|y)}{dt} = v_t(\phi_t(x_0|y)|y), \quad \phi_0(x_0|y) = x_0 \quad (16)$$

in which the flow $\phi_t(x_0|y)$ and vector field $v_t(x_t|y)$ conditioned on y generate the probability path $p_t(x_t|y)$ from $p_0(x_0|y) = \mathcal{N}(y, \sigma^2\mathbf{I})$ to $p_1(x_1|y) \approx q(x_1|y)$, where $q(s|y)$ is the distribution of clean speech given the noisy speech and σ is a hyperparameter satisfying $\sigma \geq 0$. Then, we estimate the vector field $v_t(x_t|y)$ by a neural network $v_\theta(x_t, y, t)$ using the CFM loss so that x_1 can be found by solving the ODE. The CFM loss for SE is given by

$$\begin{aligned} \mathcal{L}_{CFM}(\theta) &:= \\ &\mathbf{E}_{t,(x_1,y),p(x_0|y)} \|v_\theta(\phi_t(x_0|x_1, y), y, t) - v_t(\phi_t(x_0|x_1, y)|x_1, y)\|^2. \end{aligned} \quad (17)$$

As in [21], we let $p_t(x_t|x_1, y)$ be a Gaussian conditional probability path. Then, the conditional flow $\phi_t(x_0|x_1, y)$ and the conditional target vector field $v_t(x_t|x_1, y)$ becomes

$$\phi_t(x_0|x_1, y) = \frac{\sigma_t}{\sigma_0}(x_0 - \mu_0(x_1, y)) + \mu_t(x_1, y), \quad (18)$$

$$v_t(x_t|x_1, y) = \frac{\sigma'_t}{\sigma_t}(x_t - \mu_t(x_1, y)) + \mu'_t(x_1, y), \quad (19)$$

in which the mean $\mu_t(x_1, y)$ and variance σ_t^2 for $p_t(x_t|x_1, y) = \mathcal{N}(\mu_t(x_1, y), \sigma_t^2\mathbf{I})$ are configured in a similar way to the OT conditional vector field in (15) as

$$\mu_t(x_1, y) = tx_1 + (1 - t)y, \quad (20)$$

$$\sigma_t = (1 - t)\sigma. \quad (21)$$

In (20), the mean moves linearly from the noisy speech y to the clean speech x_1 from $t = 0$ to 1, while μ_t in (15) starts at 0 and goes linearly to x_1 . The standard deviation in (21) decreases linearly from σ at $t = 0$ to 0 at $t = 1$. In contrast, the standard deviation σ_t in (15) decreases from 1 at $t = 0$ to a small positive value $\sigma < 1$. In the training procedure, t is sampled from $\mathcal{U}[0, 1 - t_\delta]$ for each training sample pair (x_1, y) , and then x_t is directly sampled from $p_t(x_t|x_1, y)$, which is used to construct the training target in (17).

In the inference phase, the ODE in (16) is solved numerically using Euler method with trained $v_\theta(x_t, y, t)$ starting at x_0 sampled from $p_0(x_0|y)$. N time points are chosen on $[0, 1]$ and denoted as $0 = t_0 < t_1 < \dots < t_{N-1} = 1 - t_\delta < t_N = 1$. Starting from

the initial point x_0 from $p_0(x_0|y)$, the clean speech estimate x_{t_N} is generated by applying the following equation repeatedly:

$$x_{t_i} = v_\theta(x_{t_{i-1}}, y, t_{i-1})\Delta t(i) + x_{t_{i-1}}, \quad (22)$$

where $\Delta t(i) = t_i - t_{i-1}$.

B. FlowSE as a Diffusion model

The proposed FlowSE with a Gaussian conditional probability path can also be described as a diffusion model with an SDE. We can find the drift coefficient f and the diffusion coefficient g in (1) which make the perturbation kernel $p_t(x_t|x_0, y)$ match the conditional probability path $p_t(x_t|x_1, y)$ in the FM with reversed time, *i.e.*, $p_t(x_t|x_0, y) = \mathcal{N}(\mu_{1-t}(x_0, y), \sigma_{1-t}^2\mathbf{I})$. By utilizing the ODEs that the mean and the covariance for the perturbation kernel $p_t(x_t|x_0, y)$ of a diffusion model satisfy, which is (5.50), (5.53) in [28], the drift and diffusion coefficients for the equivalent diffusion model are given by

$$f(x_t, y, t) = \frac{y - x_t}{1 - t}, \quad g(t) = \sqrt{\frac{2t\sigma^2}{1 - t}}. \quad (23)$$

It is noted that the drift coefficient f is the same as the f in the BBED SDE given in (4), while the diffusion coefficient is different. One of the method to find x_0 from the diffusion model in (1) is to solve the probability flow ODE [26] given by

$$\frac{dx_t}{dt} = \left[f(x_t, y, t) - \frac{1}{2}g(t)^2\nabla_{x_t} \log p_t(x_t|y) \right], \quad (24)$$

which has the form of the ODE in (16) [21], [22]. It can be solved using the Euler method, moving from $t_N \in (0, 1)$ to $t_0 = 0$, with $s_\theta(x_t, y, t)$ trained with the DSM loss, instead of $\nabla_{x_t} \log p_t(x_t|y)$.

IV. EXPERIMENTAL SETTINGS

We use two different datasets, WSJ0-CHiME3 and VoiceBank-DEMAND (VB-DMD), to demonstrate the performance of FlowSE. The WSJ0-CHiME3 dataset was constructed by mixing clean speech utterances from the Wall Street Journal (WSJ0) dataset [29] and environmental noises from the CHiME3 dataset [30] with the signal-to-noise ratio (SNR) sampled from a uniform distribution between 0 and 20 dB. We created the WSJ0-CHiME3 dataset using the source codes¹ provided by the authors of [14]. The dataset consists of a train (12777 files), validation (1206 files) and test sets (615 files). The VB-DMD dataset [31] is a publicly available dataset 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 shape) 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 training dataset is split into a training and validation dataset, using speakers ‘‘p226’’ and ‘‘p287’’ for validation, as in [14], [16], [20].

In this work, the speech signal from the datasets is represented in the complex-valued Short-Time Fourier Transform (STFT) domain after the amplitude compression as in [17], [20], so $s, y \in \mathbb{C}^{K \times F}$, where K is the number of frames and F is the number of frequency bins. All configurations of STFT and amplitude compression followed those in [17], [20].

We adopted the Noise Conditional Score Network (NCSN++) as the architecture of the neural network $v_\theta(x_t, y, t)$ with the same configuration as the compared BBED [17] and CRP [20] models. t_δ was set to be 0.03. We trained the neural network $v_\theta(x_t, y, t)$, using Adam optimizer [34] with a learning rate of 10^{-4} and a batch

¹ <https://github.com/sp-uhh/sgmse>

TABLE I
SPEECH ENHANCEMENT PERFORMANCES FOR PROPOSED AND COMPARED METHODS ON THE WSJ0-CHiME3 AND VB-DMD DATASETS.

Trained and Tested on WSJ0-CHiME3											
Method	NFE	PESQ	DNSMOS	WVMOS	ESTOI	SIG	BAK	OVRL	SI-SDR	SI-SIR	SI-SAR
BBED	60	3.00±0.05	4.04±0.01	3.93±0.03	0.93±0.00	3.59±0.01	4.18±0.00	3.35±0.01	18.98±0.35	31.48±0.39	19.31±0.36
BBED	5	2.90±0.04	3.86±0.02	3.75±0.03	0.92±0.00	3.53±0.01	4.05±0.01	3.23±0.01	18.33±0.35	26.82±0.38	19.04±0.36
CRP	5	3.02±0.05	4.01±0.01	3.90±0.03	0.94±0.00	3.55±0.01	4.17±0.00	3.31±0.01	19.71±0.34	30.77±0.39	20.14±0.34
Ours	5	3.03±0.04	4.04±0.01	3.96±0.03	0.93±0.00	3.60±0.01	4.20±0.00	3.37±0.01	18.95±0.33	32.43±0.41	19.21±0.34
Trained and Tested on VB-DMD											
Method	NFE	PESQ	DNSMOS	WVMOS	ESTOI	SIG	BAK	OVRL	SI-SDR	SI-SIR	SI-SAR
BBED	60	3.09±0.05	3.57±0.02	4.29±0.02	0.88±0.01	3.48±0.01	4.04±0.01	3.2±0.01	18.75±0.23	30.11±0.41	19.43±0.24
BBED	5	2.43±0.04	3.3±0.02	3.93±0.04	0.86±0.01	3.48±0.01	3.68±0.02	3.02±0.02	16.84±0.29	21.67±0.33	18.97±0.28
CRP	5	3.08±0.05	3.57±0.02	4.29±0.03	0.88±0.01	3.47±0.01	4.03±0.01	3.18±0.02	19.31±0.3	28.3±0.47	20.57±0.26
Ours	5	3.12±0.05	3.58±0.02	4.34±0.02	0.88±0.01	3.49±0.01	4.05±0.01	3.21±0.01	18.95±0.23	32.2±0.43	19.4±0.23

TABLE II
PERFORMANCES OF THE FLOWSE WITH THE NFE OF 5 TRAINED WITH FLOW MATCHING AND SCORE MATCHING ON WSJ0-CHiME3 DATASET.

METHOD	PESQ	DNSMOS	WVMOS	ESTOI	SI-SDR
Flow matching	3.03	4.04	4.00	0.93	18.95
Score matching	3.01	4.03	3.92	0.93	18.57

size of 8. An exponential moving average with a decay of 0.999 was applied to the network parameters. σ was determined empirically to 0.487 based on the wideband - perceptual evaluation of speech quality (WB-PESQ) [35] scores for the validation set. We trained the model for 250 epochs and logged the average WB-PESQ scores for 10 random files from the validation set for each epoch to find the best model for speech enhancement. The N and $\Delta t(i)$ in the Euler method in (22) or probability flow ODE in (24) were set to be $N \in \{1, 2, 3, 4, 5\}$, $\Delta t(i) = \frac{t_N - t_\delta}{N-1}$, when $i \in \{1, 2, \dots, N-1\}$ and $\Delta t(N) = t_\delta$ for $N \in \{2, 3, 4, 5\}$. In the case of $N = 1$, $\Delta t(1) = t_N$.

We selected two diffusion-based SE models BBED [17] and CRP [20] as baselines. We utilized the pre-trained BBED model available at the model checkpoint² shared by the authors of [20] and fine-tuned that model for CRP with $n_{rev} \in \{1, 2, 3, 4, 5\}$ with the same configurations in [20].

To assess the performance of the proposed method, we have evaluated the WB-PESQ scores [35], Deep Noise Suppression MOS (DNSMOS) [36], wav2vec MOS (WVMOS) [37], Extended Short-Time Objective Intelligibility (ESTOI) [38], DNSMOS P.835 including SIG, BAK, and OVRL [39], Scale-Invariant Signal-to-Distortion Ratio (SI-SDR), Signal-to-Inference Ratio (SI-SIR), and Signal-to-Artifact Ratio (SI-SAR) [40].

V. RESULTS

Table I summarizes the performances and NFEs for the proposed and compared methods on the WSJ0-CHiME3 and VB-DMD datasets, where the CRP was fine-tuned with $n_{rev} = 5$. The mean of each measure is reported, followed by a 95% confidence interval. We can see that the performances of our method with the NFE of 5 were nearly equivalent to BBED with the NFE of 60 and better than BBED with the NFE of 5. When compared with CRP, our method performed similarly to CRP in all measurements without any fine-tuning procedure.

For further analysis, we have shown the WB-PESQ scores, DNSMOS, and WVMOS with the NFE from 1 to 5 for the proposed model and the five CRP models fine-tuned with the same n_{rev} with

²https://github.com/sp-uhh/sgmse_crp

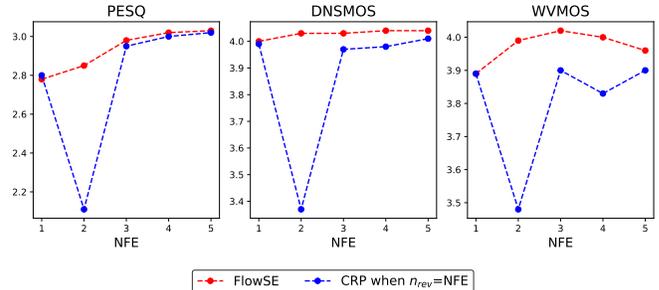


Fig. 2. Performances of the proposed FlowSE and the CRP models fine-tuned with matched n_{rev} for the NFE from 1 to 5 on the WSJ0-CHiME3 dataset.

the NFE in the inference phase in Fig. 2. Although the proposed FlowSE model was trained without the knowledge of the number of steps in the inference stage, the performances were comparable to the CRP models, each fine-tuned with the matched number of n_{rev} .

Table II demonstrates the performances of the proposed FlowSE trained with flow matching and score matching. Flow matching indicates that the model $v_\theta(x_t, y, t)$ was trained with the CFM loss in (17), while score matching denotes that the score model $s_\theta(x_t, y, t)$ was trained with the DSM loss in (3) when the drift and diffusion coefficients were given in (23). It is shown that the performances of two cases with the NFE of 5 were nearly the same. It demonstrated that the performance of the BBED with the NFE of 60 could be achieved with the diffusion model with the drift and diffusion coefficients designed from the probability path with linearly moving mean and standard deviation as in the OT conditional vector field, with the NFE of 5 without the need for any fine-tuning procedure.

VI. CONCLUSION

In this work, we propose FlowSE, a speech enhancement model based on conditional flow matching. To apply the CFM for SE, we add noisy speech as conditioning variable for the conditional flow and vector field, and the mean and the covariance for the conditional probability path are configured to move linearly from noisy speech to clean speech and decrease linearly, respectively, in a similar way to the OT conditional vector field. Experimental results showed that FlowSE with the NFE of 5 achieved performances comparable to the BBED with the NFE of 60 and the CRP using additional fine-tuning procedure with the NFE of 5. We also derived the diffusion model with the perturbation kernel the same as the conditional probability path for the FlowSE, and demonstrated that the diffusion model with SM achieved similar performances to FlowSE with the NFE of 5 without any fine-tuning procedure.

REFERENCES

- [1] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2007.
- [2] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-domain based single-microphone noise reduction for speech enhancement*. Springer Nature, 2022.
- [3] T. Gerkmann and E. Vincent, "Spectral masking and filtering," *Audio source separation and speech enhancement*, pp. 65–85, 2018.
- [4] M. Kim and J. W. Shin, "Improved speech enhancement considering speech psd uncertainty," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1939–1951, 2022.
- [5] 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.
- [6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [7] H. Song and J. W. Shin, "Speech enhancement using mlp-based architecture with convolutional token mixing module and squeeze-and-excitation network," *IEEE Access*, vol. 10, pp. 119 283–119 289, 2022.
- [8] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *arXiv preprint arXiv:2008.00264*, 2020.
- [9] K. Tan and D. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.
- [10] 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.
- [11] D. Baby and S. Verhulst, "Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 106–110.
- [12] A. A. Nugraha, K. Sekiguchi, and K. Yoshii, "A flow-based deep latent variable model for speech spectrogram modeling and enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1104–1117, 2020.
- [13] X. Bie, S. Leglaive, X. Alameda-Pineda, and L. Girin, "Unsupervised speech enhancement using dynamical variational autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2993–3007, 2022.
- [14] J. Richter, S. Welker, J.-M. Lemerrier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2351–2364, 2023.
- [15] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [16] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, "Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.
- [17] 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.
- [18] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Diffusion-based speech enhancement in matched and mismatched conditions using a heun-based sampler," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 431–10 435.
- [19] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3025–3038, 2024.
- [20] B. Lay, J.-M. Lemerrier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 626–630.
- [21] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *The Eleventh International Conference on Learning Representations*, 2023.
- [22] A. Tong, K. FATRAS, N. Malkin, G. Huguët, 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.
- [23] A. H. Liu, M. Le, A. Vyas, B. Shi, A. Tjandra, and W.-N. Hsu, "Generative pre-training for speech with flow matching," in *The Twelfth International Conference on Learning Representations*, 2024.
- [24] C. Jung, S. Lee, J.-H. Kim, and J. S. Chung, "Flowavse: Efficient audio-visual speech enhancement with conditional flow matching," in *Interspeech 2024*, 2024, pp. 2210–2214.
- [25] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," *Advances in neural information processing systems*, vol. 31, 2018.
- [26] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021.
- [27] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1530–1538.
- [28] S. Särkkä and A. Solin, *Applied stochastic differential equations*. Cambridge University Press, 2019, vol. 10.
- [29] J. S. Garofolo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," Online, 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 *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–4.
- [33] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multichannel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19. AIP Publishing, 2013.
- [34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [35] *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.
- [36] O. F. Salas, V. Adzic, and H. Kalva, "Subjective quality evaluations using crowdsourcing," in *2013 Picture Coding Symposium (PCS)*. IEEE, 2013, pp. 418–421.
- [37] P. Andreev, A. Alanov, O. Ivanov, and D. Vetrov, "Hifi++: A unified framework for bandwidth extension and speech enhancement," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Jun. 2023.
- [38] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.
- [39] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2022, pp. 886–890.
- [40] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr-half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.