ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10095310

DIFFUSION-BASED GENERATIVE SPEECH SOURCE SEPARATION

Robin Scheibler[1], *Youna Ji*[2], *Soo-Whan Chung*[2], *Jaeuk Byun*[2], *Soyeon Choe*[2], *Min-Seok Choi*[2]

[1]LINE Corporation, Tokyo, Japan
[2]NAVER Cloud Corporation, Seongnam, South Korea

ABSTRACT

We propose DiffSep, a new single channel source separation method based on score-matching of a stochastic differential equation (SDE). We craft a tailored continuous time diffusion-mixing process starting from the separated sources and converging to a Gaussian distribution centered on their mixture. This formulation lets us apply the machinery of score-based generative modelling. First, we train a neural network to approximate the score function of the marginal probabilities of the diffusion-mixing process. Then, we use it to solve the reverse time SDE that progressively separates the sources starting from their mixture. We propose a modified training strategy to handle model mismatch and source permutation ambiguity. Experiments on the WSJ0_2mix dataset demonstrate the potential of the method. Furthermore, the method is also suitable for speech enhancement and shows performance competitive with prior work on the VoiceBank-DEMAND dataset.

Index Terms— source separation, stochastic differential equation, diffusion, score matching

1. INTRODUCTION

Source separation (SS) refers to a family of techniques that can be used to recover signals of interests from their mixtures. It has broad applications, but we focus on single channel audio and speech [1]. This is an underdetermined problem where several signals are recovered from a single mixture. Early success in this area was made with methods such as non-negative matrix factorization (NMF) [2]. Impressive progress was brought by data-driven methods and deep neural networks (DNN). We mention deep clustering where a DNN is trained to create an embedding space in which the time-frequency bins of the input spectrogram cluster by source [3]. Then, one can recover isolated sources by creating masks corresponding to clusters of bins and multiply the spectrogram with it. These methods then evolved to networks predicting mask values directly [4]. All these data-driven methods require to solve the fundamental source permutation ambiguity. Namely, there is no inherent preferred order for the sources. This is usually handled by permutation invariant training (PIT) [5]. In PIT, the objective function is computed for all permutations of the sources and the one with minimum value is chosen to compute the gradient. Recently very strong time-domain baselines such as Conv-TasNet [6] or dual-path transformer [7] have been proposed. We mention also all-attention based [8], multi-scale networks [9], WaveSplit [10], and the recent state-of-the-art TF-GridNet [11] back in the time-frequency domain. For a survey and benchmarks of recent methods see [11], [12]. All these recent approaches are trained with a discriminative objective such as the scale-invariant signal-to-distortion ratio (SI-SDR) [13]. Except the traditional methods, such as NMF, few approaches leverage generative models. One exception adversarially trains speaker-specific generative models and uses them for separation [14].

Unlike discriminative methods, generative modelling allows to approximate complex data distributions [15]. Recently, score-based generative modelling (SGM) [16], also known as diffusion-based modelling, has made rapid and impressive progress, in particular in the domain of image generation [17]. SGM defines a forward process that progressively transforms a sample of the target distribution into Gaussian noise. Given its *score-function*, i.e., the gradient of the log-probability, the target distribution can be sampled by running a reverse process starting from noise. Crucially, the score-function is (usually) unknown, but can be approximated by a DNN using a simple training strategy [18]. Recent approaches result from the combination of score-matching and Langevin sampling [16], [19], [20]. Two frameworks have been proposed to unify various approaches, one based on graphical models [21], and the other on stochastic differential equations (SDE) [22]. SGM has quickly found application in audio and speech processing. It has been successfully applied in speech synthesis [23]–[26], speech restoration [27], [28], and bandwidth extension [29]. Surprisingly, perhaps, several SGMs have been proposed for speech enhancement, a typically discriminative task [30]–[33]. Closest to this work, separation of images and sounds using deep generative priors and Langevin sampling has been demonstrated [34][35]. However, the sources in these works are from different distributions, e.g., voice and piano. A diffusion-based speech separation method has yet to be proposed.

We fill this gap by proposing what we believe is the first diffusion-based approach for the separation of speech signals. The proposed algorithm, DiffSep, is based on a carefully designed SDE that starts from the separated signals and converges in average to their mixture. Then, by solving the corresponding reverse-time SDE, it is possible to recover the individual sources from their mixture. See Fig. 1 for an illustration. We propose a modified training strategy for the score-matching network that addresses ambiguity in the assignment of the sources at the beginning of the inference process. We demonstrate the method on the widely used *WSJ0_2mix* benchmark dataset [3]. While it does not outperform, yet, discriminative baselines in terms of separation metrics, it achieves high non-intrusive DNSMOS P.835 [36] score, indicating naturalness. In addition, our approach is also suitable for speech enhancement by considering the noise as just an extra source, and performs comparably to prior work [32] on the *VoiceBank-DEMAND* dataset [37].

2. BACKGROUND

2.1. Notation and Signal Model

Vectors and matrices are represented by bold lower and upper case letters, respectively. The norm of vector $\boldsymbol{x}$ is written $\|\boldsymbol{x}\| =$

Code: `https://github.com/fakufaku/diffusion-separation`.

$(\boldsymbol{x}^\top \boldsymbol{x})^{1/2}$. The $N \times N$ identity matrix is denoted by $\boldsymbol{I}_N$.

We represent audio signals with N samples by real valued vectors from $\mathbb{R}^N$. We introduce a simple mixture model with K sources,

$$\boldsymbol{y} = \sum_{k=1}^{K} \boldsymbol{s}_k, \tag{1}$$

where $\boldsymbol{s}_k \in \mathbb{R}^N$ for all k. We remark that this model can account for environmental noise, reverberation, and other degradations by adding an extra source. By concatenating all the source signals, we obtain vectors in $\mathbb{R}^{KN}$. We define in this notation the vector of separated sources and their average value,

$$\boldsymbol{s} = [\boldsymbol{s}_1^\top, \ldots, \boldsymbol{s}_K^\top]^\top, \quad \bar{\boldsymbol{s}} = K^{-1}[\boldsymbol{y}^\top, \ldots, \boldsymbol{y}^\top]^\top. \tag{2}$$

We can define the time-invariant mixing operation as a multiplication by $(\boldsymbol{A} \otimes \boldsymbol{I}_N)$ where $\boldsymbol{A}$ is a $K \times K$ matrix, and $\otimes$ is the Kronecker product. Multiplying a vector $\boldsymbol{v} \in \mathbb{R}^{KN}$ by such a matrix amounts to multiplying by $\boldsymbol{A}$ all sub-vectors of length K with elements taken from $\boldsymbol{v}$ at interval N, i.e.,

$$((\boldsymbol{A} \otimes \boldsymbol{I}_N)\boldsymbol{v})_{kN+n} = \sum_{\ell=1}^{K} A_{k\ell}\, v_{\ell N+n}, \qquad \begin{array}{l} k = 1, \ldots, K, \\ n = 1, \ldots, N. \end{array} \tag{3}$$

To lighten the notation, we overload the regular matrix product as $\boldsymbol{A}\boldsymbol{v} \triangleq (\boldsymbol{A} \otimes \boldsymbol{I}_N)\boldsymbol{v}$ in the rest of this paper. Next, we define the projection matrices

投影矩阵

$$\boldsymbol{P} = K^{-1}\mathbb{1}\mathbb{1}^\top, \quad \bar{\boldsymbol{P}} = \boldsymbol{I}_K - \boldsymbol{P}. \tag{4}$$

where $\mathbb{1}$ is the all one vector, of size K here. The matrix $\boldsymbol{P}$ projects onto the subspace of average values, and $\bar{\boldsymbol{P}}$ its orthogonal complement. For example, a compact alternative to (1) is $\boldsymbol{P}\boldsymbol{s} = \bar{\boldsymbol{s}}$.

2.2. SDE for Score-based Generative Modelling

Score-based generative modelling (SGM) via SDE is a principled approach to model complex data distributions [22]. First, a diffusion process starting from samples from the target distribution and converging to a Gaussian distribution can be described by the following stochastic differential equation

正向

$$d\boldsymbol{x}_t = f(\boldsymbol{x}_t, t)\, dt + g(t)\, d\boldsymbol{w}_t \tag{5}$$

where $\boldsymbol{x}_t : \mathbb{R}^N \to \mathbb{R}^N$ is a vector-valued function of the time $t \in \mathbb{R}$, $d\boldsymbol{x}_t$ is its derivative with respect to t, and dt an infinitesimal time step. Functions $f : \mathbb{R}^N \to \mathbb{R}^N$ and $g : \mathbb{R} \to \mathbb{R}$ are called the *drift* and *diffusion* coefficients of $\boldsymbol{x}_t$. The term $d\boldsymbol{w}_t$ is a standard Wiener process. For more details on SDEs, see [38]. We insist that the time t in the diffusion process is unrelated to the time of the audio signals that will appear later in this paper.

A remarkable property of (5) is that under some mild conditions, there exists a corresponding *reverse-time SDE*,

逆向

$$d\boldsymbol{x}_t = -[f(\boldsymbol{x}_t, t) - g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)]\, dt + g(t)\, d\bar{\boldsymbol{w}}, \tag{6}$$

going from $t = T$ to $t = 0$ [39]. Here, $d\bar{\boldsymbol{w}}$ is a reverse Brownian process, and $p_t(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{x}_t$.

During the forward process, $\boldsymbol{x}_0$ being known, it is usually possible to have a closed-form formula for $p_t(\boldsymbol{x})$. For an affine $f(\boldsymbol{x}, t)$, it is Gaussian and its parameters, i.e., mean and covariance matrix, usually tractable [38]. However, in the reverse process, $p_t(\boldsymbol{x})$ is unknown, and thus (6) cannot be solved directly. The key idea of SGM [16], [19], [22] is to train a neural network $\boldsymbol{q}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$ so that $\boldsymbol{q}_{\boldsymbol{\theta}}(\boldsymbol{x}, t) \approx \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$. Provided that the approximation is good enough, we can generate samples from the distribution of $\boldsymbol{x}_0$ by numerically solving (6), replacing $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$ by $\boldsymbol{q}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)$.

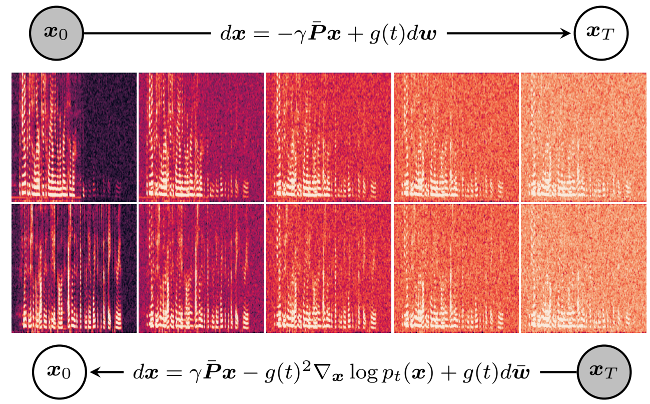

Fig. 1: Illustration of the diffusion-mixing process. From left to right, the sources start separated and progressively mix as more and more noise is added. The reverse process starts from the mixture and ends with separated sources. The spectrograms were produced by actual separation of the mixture on the right using the proposed method.

2.3. Related Work in Diffusion-based Speech Processing

SGM first found widespread adoption for the generation of very high quality synthetic images, e.g. [17]. Speech synthesis is thus a logical candidate for its adoption [23], [24]. In this area, several approaches that shape the noise adaptively for the target speech have been proposed. One uses the local signal energy [25], while the other introduces correlation in the noise according to the target spectrogram [26]. Bandwidth extension and restoration also require to generate missing parts of the audio and high quality diffusion-based approaches have been proposed [27]–[29]. Diffuse [30] and CDiffuse [31] use conditional probabilistic diffusion to enhance a noisy speech signal. Also for speech enhancement, Richter et al. have proposed another approach [32], [33] based on solving the reverse-time SDE corresponding to the forward model

前向SDE:

$$d\boldsymbol{x}_t = \gamma(\boldsymbol{y} - \boldsymbol{x}_t)\, dt + g(t)\, d\boldsymbol{w}, \quad \boldsymbol{x}_0 = \boldsymbol{s}. \tag{7}$$

This is the *Ornstein-Uhlenbeck* SDE that, in the forward direction, takes the clean speech $\boldsymbol{s}$ progressively towards the noisy one $\boldsymbol{y}$.

3. DIFFUSION-BASED SOURCE SEPARATION

We propose a framework for applying SGM to the source separation problem. Essentially, we design a forward SDE where both diffusion and mixing across sources happen over time. Each step of the process can be described as adding an infinitesimal amount of noise and doing an infinitesimal amount of mixing. This can be formalized as the following SDE

可调超参数

源扩散SDE:

$$d\boldsymbol{x}_t = -\gamma\bar{\boldsymbol{P}}\boldsymbol{x}_t\, dt + g(t)d\boldsymbol{w}, \quad \boldsymbol{x}_0 = \boldsymbol{s}, \tag{8}$$

where $\bar{\boldsymbol{P}}$ is defined in (4) and $\boldsymbol{s}$ in (2). We use the diffusion coefficient of the *Variance Exploding SDE* of Song et al. [22],

方差爆炸SDE系数

$$g(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^t \sqrt{2\log\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)}. \tag{9}$$

The SDE (8) has the interesting property that the mean of the marginal $\boldsymbol{\mu}_t$ goes from the vector of separated signals at $t = 0$ to the mixture vector $\bar{\boldsymbol{s}}$ as t grows large. This is formalized in the following theorem proved in Appendix 6.

t↑, 边际分布的均值 μ_t 从分离信号(t=0) 逐渐向混合信号转变

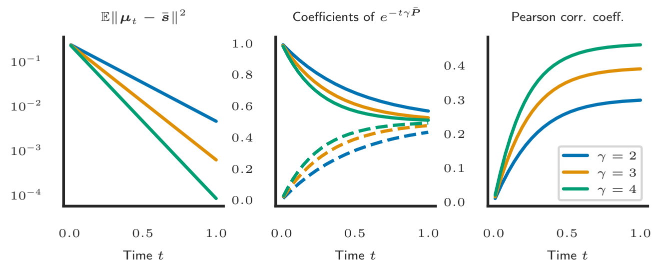

Fig. 2: Evolution of some of the parameters of the SDE (8) over time.

Theorem 1. *The marginal distribution of $\boldsymbol{x}_t$ is Gaussian with mean and covariance matrix*

在边际分布是高斯 均值:
$$\boldsymbol{\mu}_t = (1 - e^{-\gamma t})\bar{\boldsymbol{s}} + e^{-\gamma t}\boldsymbol{s} \tag{10}$$
方差:
$$\boldsymbol{\Sigma}_t = \lambda_1(t)\boldsymbol{P} + \lambda_2(t)\bar{\boldsymbol{P}} \tag{11}$$

where if we let 设 *$\xi_1 = 0$, $\xi_2 = \gamma$, and $\rho = \frac{\sigma_{\max}}{\sigma_{\min}}$, then,*

$$\lambda_k(t) = \frac{\sigma_{\min}^2 \left(\rho^{2t} - e^{-2\xi_k t}\right) \log \rho}{\xi_k + \log \rho}. \tag{12}$$

Fig. 1 shows an example of the process on a real mixture. We now have an explicit expression for a sample $\boldsymbol{x}_t$,

$$\boldsymbol{x}_t = \boldsymbol{\mu}_t + \boldsymbol{L}_t \boldsymbol{z}, \quad \boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{KN}), \tag{13}$$

where $\boldsymbol{L}_t = \lambda_1(t)^{1/2}\boldsymbol{P} + \lambda_2(t)^{1/2}\bar{\boldsymbol{P}}$. Fig. 2 shows the evolution of some of the parameters of the distribution of $\boldsymbol{x}_t$ over time. We can make the difference of $\boldsymbol{\mu}_T$ and $\bar{\boldsymbol{s}}$ arbitrarily small by tuning the value of γ. In addition, we observe that the correlation of the noise added onto both sources increases over time due to the mixing process.

Previous work on speech enhancement has successfully applied the diffusion process in the short-time Fourier transform (STFT) domain with a non-linear transform [32], [33]. The approach we propose could be equally applied in the time-domain or in any linearly transformed domain such as the STFT. However, since the process of (8) models the linear mixing of the sources, we cannot apply a non-linear transformation. Instead, we perform the diffusion process in the time-domain, but operate our network in the non-linear STFT domain of [33], as described in Section 4.2.

3.1. Inference

At inference time, the separation is done by solving (6) with the help of a score-matching network $\boldsymbol{q}_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{y}) \approx \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x})$. The initial value of the process is sampled from the distribution,

$$\bar{\boldsymbol{x}}_T \sim \mathcal{N}(\bar{\boldsymbol{s}}, \boldsymbol{\Sigma}_T \boldsymbol{I}_{KN}). \tag{14}$$

Then, we adopt the predictor-corrector approach described in [22] to solve (6). The prediction step is done by *reverse diffuse sampling*. The correction is done by annealed Langevin sampling [20].

3.2. Permutation and Mismatch-aware Training Procedure

To train the score network $\boldsymbol{q}_{\boldsymbol{\theta}}(\boldsymbol{x}, t, \boldsymbol{y})$, we propose a slightly modified score-matching procedure. But, first, we follow the usual procedure. Thanks to Theorem 1, the score function has a closed form expression. The computation is similar to that in [32], and we obtain

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) = -\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_t) = -\boldsymbol{L}_t^{-1}\boldsymbol{z}. \tag{15}$$

The training loss is

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}, t} \left\| \boldsymbol{q}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}) + \boldsymbol{L}_t^{-1}\boldsymbol{z} \right\|_{\boldsymbol{\Sigma}_t}^2 \tag{16}$$
$$= \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}, t} \left\| \boldsymbol{L}_t \boldsymbol{q}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y}) + \boldsymbol{z} \right\|^2, \tag{17}$$

where $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{KN})$, $t \sim \mathcal{U}(t_\epsilon, T)$, and $\boldsymbol{x}_0$ is chosen at random from the dataset. We use the norm induced by $\boldsymbol{\Sigma}_t$ which is equivalent to the weighting scheme proposed by Song et al. [16]. This scheme makes the order of magnitude of the cost independent of $\boldsymbol{\Sigma}_t$ and lets us avoid the computation of the inverse of $\boldsymbol{L}_t$.

However, when training only with the procedure just described, we found poor performance at inference time. We identified two problems. First, there is a discrepancy between $\mathbb{E}[\bar{\boldsymbol{x}}_T] = \bar{\boldsymbol{s}}$ from (14) and $\boldsymbol{\mu}_T$. At $t = T$, the true score function does not simplify as (15), but is instead $\nabla_{\boldsymbol{x}_T} \log p(\bar{\boldsymbol{x}}_T) = -\boldsymbol{\Sigma}_T^{-1}(\bar{\boldsymbol{s}} + \boldsymbol{L}_T\boldsymbol{z} - \boldsymbol{\mu}_T)$. Second, the network needs to decide in which order to output the sources. To solve these challenges, we propose to train the network with a PIT objective [5] that includes the model mismatch. For each sample, with probability $1 - p_T$, $p_T \in [0, 1]$, we apply the regular score-matching procedure minimizing (17). With probability p_T, we set $t = T$ and minimize the following alternative loss,

$$\mathcal{L}_T = \mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{z}} \min_{\pi \in \mathcal{P}} \left\| \boldsymbol{L}_T \boldsymbol{q}_\theta(\bar{\boldsymbol{x}}_T, T, \boldsymbol{y}) + \boldsymbol{z} + \boldsymbol{L}_T^{-1}(\bar{\boldsymbol{s}} - \boldsymbol{\mu}_T(\pi)) \right\|^2,$$

where $\bar{\boldsymbol{x}}_T$ is sampled from (14), $\mathcal{P}$ is the set of permutations of the sources, and $\boldsymbol{\mu}_T(\pi)$ is $\boldsymbol{\mu}_T$ computed for the permutation π. The rational for this objective is that the score network will learn to remove both the noise and the model mismatch. We note that for speech enhancement, there is no ambiguity in the order of sources and the minimization over permutations can be removed.

4. EXPERIMENTS

4.1. Datasets

We use the *WSJ0_2mix* dataset [3] for the speech separation experiment. It contains 20000 (30 h), 5000 (10 h), and 3000 (5 h) two speakers mixtures whose individual speech samples were taken from the WSJ0 dataset. The utterances are fully overlapped and speakers do not overlap between training and test datasets. The relative signal-to-noise ratio (SNR) of the mixtures is from -5 dB to 5 dB. The sampling frequency is 8 kHz. In addition, we use the clean test set of Libri2Mix [40] to test mismatched conditions.

For the speech enhancement task, we use the *VoiceBank-DEMAND* dataset [37]. It is a collection of clean speech samples of 30 speakers from the VoiceBank corpus degraded by noise, recorded as well as artificial. The dataset is split into train (8.6 h), validation (0.7 h), and test (0.6 h). The SNR of the mixtures ranges from 0 dB to 15 dB for training and validation, and from 2.5 dB to 17.5 dB for testing. The sampling rate of the mixtures is 16 kHz.

4.2. Model

We use the noise conditioned score-matching network (NCSN++) of Song et al. [22]. An STFT and its inverse layer sandwich the network and, in addition, we apply the non-linear $c(x)$ transform proposed in [33] after the STFT, and its inverse before the iSTFT,

$$c(x) = \beta^{-1}|x|^{\alpha} e^{j\angle x}, \quad c^{-1}(x) = \beta|x|^{1/\alpha} e^{j\angle x}. \tag{18}$$

The real and imaginary parts are concatenated. The network has $K + 1$ inputs, one for each of the target sources plus one for the mixture on which we condition, and K outputs.

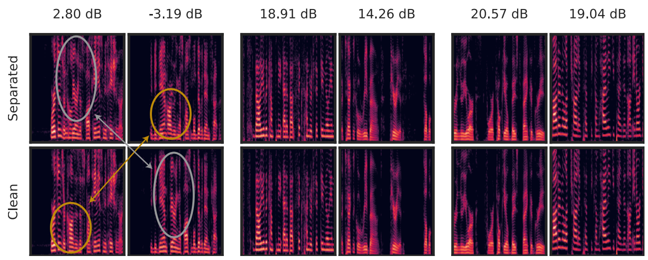

Fig. 3: Spectrograms of the clean sources (bottom row) and the separated speech obtained by applying the proposed method to their mixture (top row). The three samples, from left to right, were chosen randomly from the bottom $10\,\%$, mid $80\,\%$, and top $10\,\%$ samples sorted by SI-SDR. The SI-SDR of the samples is indicated at the top. In the low quality sample, the spectrogram looks good, but we can see block permutations of the sources occurring.

Table 1: Separation results on WSJ0_2mix and Libri2Mix (clean) datasets.

Dataset	Model	SI-SDR	PESQ	ESTOI	OVRL
WSJ0_2mix (matched)	Conv-TasNet [6]	16.0	3.29	0.91	3.21
	DiffSep (proposed)	14.3	3.14	0.90	3.29
Libri2Mix (mismatched)	Conv-TasNet [6]	10.3	2.60	0.80	2.95
	DiffSep (proposed)	9.6	2.58	0.78	3.12

4.3. Hyperparameters for Inference and Training

For inference, we use 30 prediction steps. For each one, we do one correction step with step size $r = 0.5$. The networks are trained using Pytorch with the Adam optimizer. We use exponential averaging of the network weights as in [32].

Separation We use $\gamma = 2$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.5$. For the transformation (18), we use $\alpha = 0.5$ and $\beta = 0.15$. The probability of selecting a sample for modified training is $p_T = 0.1$. The effective batch size is 48 and the learning rate 0.0005. We trained for 1000 epochs and we chose the one with largest SI-SDR on a sentinel validation batch.

Enhancement The SDE parameters are $\gamma = 2$, $\sigma_{\min} = 0.05$, $\sigma_{\max} = 0.5$. The transformation (18) uses $\alpha = 0.5$ and $\beta = 0.15$. We found it beneficial to use PriorGrad [25] for the noise generation. The variance of the process noise for a given sample is chosen as the average power of the 500 closest samples in the mixture. In the enhancement setting, there is no permutation ambiguity so we always order the sources with speech first and noise second. We use $p_T = 0.03$. The effective batch size was 16 and the learning rate 0.0001. The network was trained for around 160 epochs.

4.4. Results

We measure the performance in terms of SI-SDR [13], perceptual evaluation of speech quality (PESQ) [41], extended short-time objective intelligibility (ESTOI) [42], and the overall (OVRL) metric of the deep noise challenge mean opinion score (DNSMOS) P.835 [36].

Separation Table 1 shows the results of the speech separation experiments. We also train a baseline Conv-TasNet [6] network. On the intrusive metrics, SI-SDR, PESQ, and STOI, the proposed method scores lower than the baseline. However, it scores higher on the non-intrusive OVRL metric. Informal listening tests further reveal that low separation metrics are often due to very natural sounding permutations of the speakers. This problem seems to happen for similar sounding speakers. Example spectrograms of poor, fair, and high quality separation results are shown in Fig. 3.

Table 2: Enhancement results on the VoiceBank-DEMAND dataset.

Model	SI-SDR	PESQ	ESTOI	OVRL
Conv-TasNet [6]	18.3	2.88	0.86	3.20
CDiffuse† [31]	12.6	2.46	0.79	—
SGMSE+† [32]	17.3	2.93	0.87	—
DiffSep (proposed)	17.5	2.56	0.84	3.09

† results reported in [32].

It was reported that diffusion-based enhancement performs well in mismatched condition [32]. We thus evaluated the model trained on WSJ0_2mix on the Libri2Mix [40] clean test set. The proposed method still scores lower than Conv-TasNet for SI-SDR, PESQ, and STOI, but the metric degradation is smaller, e.g. $1\,\mathrm{dB}$ less reduction of SI-SDR. On the other hand, the proposed method still enjoys fairly high OVRL score, indicating good perceptual quality.

Enhancement Table 2 shows the results of the evaluation of the noisy speech enhancement task. The proposed method performs very similarly to other diffusion methods CDiffuse [31] and SGMSE+ [32]. We obtain an SI-SDR higher by $0.2\,\mathrm{dB}$, but PESQ value closer to that of CDiffuse. ESTOI is somewhere in the middle. The OVRL metric suggests good perceptual quality, but was not provided for the other methods. In terms of SI-SDR particularly, a gap remains with discriminative methods such as Conv-TasNet.

5. CONCLUSION

We proposed DiffSep, the first speech source separation method based on diffusion process. We successfully separated speech sources, and demonstrated that our approach extends to noisy speech enhancement. For separation, there was a substantial gap in performance with state-of-the-art discriminative methods. We believe this may be due to both outputs being from the same distribution, thus confusing the generative model. We have done initial work to alleviate the permutation ambiguity, using a modified training strategy for the score-matching network, but further work is likely necessary to robustify the inference procedure against permutation mistakes. Bringing in proven architectures from the separation literature is an interesting area of future work. We have also not used yet any of the theoretical properties of the distribution of mixtures, which could be elegantly integrated in the score network architecture, as in [34].

6. APPENDIX: PROOF OF THEOREM 1

Because the stochastic differential equation is linear, $\boldsymbol{x}_t$ follows a Gaussian distribution [39]. Following [38], eq. (5.50) and (5.53), the mean and covariance matrices of $\boldsymbol{x}_t$ are the solutions of

$$d\boldsymbol{\mu}_t = -\bar{\boldsymbol{P}}\boldsymbol{\mu}_t, \quad d\boldsymbol{\Sigma}_t = -\widehat{\boldsymbol{A}}\boldsymbol{\Sigma}_t - \boldsymbol{\Sigma}_t\widehat{\boldsymbol{A}} + g(t)^2\boldsymbol{I}, \tag{19}$$

respectively, with the initial conditions $\boldsymbol{\mu}_0 = \boldsymbol{x}_0$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{0}$. The solution for $d\boldsymbol{\mu}_t$ is given by the the matrix exponential operator,

$$\boldsymbol{\mu}_t = e^{-\gamma t\bar{\boldsymbol{P}}}\boldsymbol{x}_0 = exp(0)\boldsymbol{P}\boldsymbol{x}_0 + exp(-\gamma t)(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{x}_0. \tag{20}$$

Substituting $\boldsymbol{x}_0 = \boldsymbol{s}$, $\boldsymbol{P}\boldsymbol{x}_0 = \bar{\boldsymbol{s}}$, and rearranging terms yields (10). For $\boldsymbol{\Sigma}_t$, we work in the eigenbasis of $\gamma\bar{\boldsymbol{P}}$ and solve differential equations with respect to its distinct eigenvalues, 0 and γ. $\square$

7. REFERENCES

[1] S. Makino, Ed., *Audio Source Separation* (Signals and Communication Technology). Cham: Springer International Publishing, Jan. 2018.

[2] P. Smaragdis *et al.*, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66–75, Apr. 2014.

[3] J. R. Hershey *et al.*, "Deep clustering: Discriminative embeddings for segmentation and separation," in *ICASSP*, Mar. 2016, pp. 31–35.

[4] H. Erdogan *et al.*, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *MMSP*, Brisbane, AUD, Apr. 2015, pp. 708–712.

[5] M. Kolbaek *et al.*, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, Aug. 2017.

[6] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[7] J. Chen *et al.*, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *INTERSPEECH*, Oct. 2020, pp. 2642–2646.

[8] C. Subakan *et al.*, "Attention is all you need in speech separation," in *ICASSP*, Jun. 2021, pp. 21–25.

[9] X. Hu *et al.*, "Speech separation using an asynchronous fully recurrent convolutional neural network," in *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 22 509–22 522.

[10] N. Zeghidour and D. Grangier, "Wavesplit: End-to-end speech separation by speaker clustering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 2840–2849, 2021.

[11] Z.-Q. Wang *et al.*, *TF-GridNet: Making time-frequency domain models great again for monaural speaker separation*, arXiv:2209.03952, Sep. 2022.

[12] Y.-J. Lu *et al.*, "ESPnet-SE++: Speech enhancement for robust speech recognition, translation, and understanding," in *INTERSPEECH*, Sep. 2022, pp. 5458–5462.

[13] J. Le Roux *et al.*, "SDR–half-baked or well done?" In *ICASSP*, May 2019, pp. 626–630.

[14] Y. C. Subakan and P. Smaragdis, "Generative adversarial source separation," in *ICASSP*, Apr. 2018, pp. 26–30.

[15] S. Bond-Taylor *et al.*, "Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7327–7347, Nov. 2022.

[16] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[17] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *CVPR*, Jun. 2022, pp. 10 684–10 695.

[18] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, pp. 695–709, Apr. 2005.

[19] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Adv. neural inf. process. syst.*, vol. 33, Jun. 2020, pp. 12 438–12 448.

[20] A. Jolicoeur-Martineau *et al.*, "Adversarial score matching and improved sampling for image generation," in *ICLR*, May 2021.

[21] J. Ho *et al.*, "Denoising Diffusion Probabilistic Models," in *Adv. Neural Inf. Process. Syst.*, vol. 33, Curran Associates, Inc., 2020, pp. 6840–6851.

[22] Y. Song *et al.*, "Score-based generative modeling through stochastic differential equations," in *ICLR*, May 2021.

[23] Z. Kong *et al.*, "DiffWave: A versatile diffusion model for audio synthesis," in *ICLR*, May 2021.

[24] N. Chen *et al.*, "WaveGrad: Estimating gradients for waveform generation," in *ICLR*, May 2021.

[25] S.-g. Lee *et al.*, "PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior," in *ICLR*, Apr. 2022.

[26] Y. Koizumi *et al.*, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in *INTERSPEECH*, Sep. 2022, pp. 803–807.

[27] J. Zhang *et al.*, "Restoring degraded speech via a modified diffusion model," in *INTERSPEECH*, Apr. 2021, pp. 221–225.

[28] J. Serrà *et al.*, *Universal speech enhancement with score-based diffusion*, Arxiv:2206.03065, Jun. 2022.

[29] S. Han and J. Lee, "NU-Wave 2: A general neural audio upsampling model for various sampling rates," in *INTERSPEECH*, Incheon, KR, Sep. 2022, pp. 4401–4405.

[30] Y.-J. Lu *et al.*, "A study on speech enhancement based on diffusion probabilistic model," in *APSIPA*, Dec. 2021, pp. 659–666.

[31] Y.-J. Lu *et al.*, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP*, May 2022, pp. 7402–7406.

[32] J. Richter *et al.*, *Speech enhancement and dereverberation with diffusion-based generative models*, arXiv:2208.05830, Aug. 2022.

[33] S. Welker *et al.*, "Speech enhancement with score-based generative models in the complex STFT domain," in *INTERSPEECH*, Sep. 2022, pp. 2928–2932.

[34] V. Jayaram and J. Thickstun, "Source separation with deep generative priors," in *ICML*, PMLR, Nov. 2020, pp. 4724–4735.

[35] V. Jayaram and J. Thickstun, "Parallel and flexible sampling from autoregressive models via Langevin dynamics," in *ICML*, Jul. 2021, pp. 4807–4818.

[36] C. K. A. Reddy *et al.*, "Dnsmos P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, May 2022, pp. 886–890.

[37] C. Valentini-Botinhao *et al.*, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *INTERSPEECH*, Sep. 2016, pp. 352–356.

[38] S. Särkkä and A. Solin, *Applied Stochastic Differential Equations*, 1st ed. Cambridge University Press, Apr. 2019.

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

[40] J. Cosentino *et al.*, *LibriMix: An open-source dataset for generalizable speech separation*, arXiv:2005.11262 [eess], May 2020.

[41] A. W. Rix *et al.*, "Perceptual evaluation of speech quality (PESQ)," in *ICASSP*, Jan. 2001, pp. 749–752.

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