ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447219

DPM-TSE: A DIFFUSION PROBABILISTIC MODEL FOR TARGET SOUND EXTRACTION

Jiarui Hai[1,†], *Helin Wang*[2,†], *Dongchao Yang*[3], *Karan Thakkar*[1], *Najim Dehak*[2], *Mounya Elhilali*[1,2]

[1]Laboratory for Computational Auditory Perception, Johns Hopkins University, Baltimore, USA
[2]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[3]The Chinese University of Hong Kong, Hong Kong SAR, China

ABSTRACT

Common target sound extraction (TSE) approaches primarily relied on discriminative approaches in order to separate the target sound while minimizing interference from the unwanted sources, with varying success in separating the target from the background. This study introduces DPM-TSE, a generative method based on diffusion probabilistic modeling (DPM) for Target Sound Extraction (TSE), to achieve both cleaner target renderings as well as improved separability from unwanted sounds. The technique also tackles the noise floor of DPM by introducing a correction method for noise schedules and sample steps. This approach is evaluated using both objective and subjective quality metrics on the FSD Kaggle 2018 dataset. The results show that DPM-TSE has a significant improvement in perceived quality in terms of target extraction and purity.

Index Terms— Target sound extraction, diffusion probabilistic model, generative model

1. INTRODUCTION

There are countless sounds in the world that offer crucial information about our environment, including the melody of a violin during a concert and sirens in the streets. Our daily lives could be significantly enhanced if we were able to create listening devices that could filter out unwanted sounds and focus on the sounds we want to hear. In recent years, machine hearing has studied target sound extraction and removal applications, which aim to identify specific speakers [1], musical instruments [2], and sound events [3, 4, 5]. Among them, the extraction of sound events is much more challenging than others because of a wide range of sounds, such as animal noises, baby cries, and telephone calls. This work addresses the problem of target sound extraction (TSE).

TSE aims to separate the sound of a specific sound event class from a mixed audio given a target sound [4, 6, 7]. Researchers have explored the challenges of new classes and weakly-labelled data, with some proposing solutions such as combining one-hot-based and enrollment-based target sound extraction [3], weakly-supervised sound separation [5], and random sound mixing [4]. These methods are based on discriminative models, which minimize the difference between estimated audio and target audio. They can produce good separation for non-overlapping regions but always suffer severe performance drops when addressing overlapping regions. Indeed, overlap often occurs in real-world scenarios, making it one of the key issues that needs to be addressed in TSE. Wang *et al.* [8] propose a TSE method utilizing timestamp information with a target-weighted loss function. However, this system requires an additional accurate detection network, and the discriminative model still struggles in separating overlaps.

Unlike discriminative methods, generative modelling that aims to match the distribution of signals allows to approximate complex data distributions, which have the potential to produce more natural audio. DPM-based generative models have recently become increasingly popular due to their remarkable performance and reliable training. In particular, the intersection of DPM and audio signal generation, such as neural vocoder [9], voice synthesis [10], and text-to-audio generation [11], has seen significant progress. Diffusion models have also been adopted in speech enhancement and speech separation. CDiffuSE [12] is a DPM-based speech enhancement model designed to directly remove the environmental noise during the reverse stage of DPM, which essentially performs a discriminative task. SGMSE [13] uses a purely generative strategy and demonstrates measurable advancements for speech enhancement. DiffSep [14] applies a score-based diffusion model on both speech separation and speech enhancement tasks. Diff-TSE [15] and DiffSpEx [16] successfully apply DPM to target speaker extraction. However, to the best of our knowledge, the application of DPM in target sound extraction has not been explored.

In this paper, we introduce a DPM-based generative method for TSE, called DPM-TSE[1]. This method applies a correction method for a diffusion noise scale and a different prediction target ”velocity” to deal with the noise floor issue caused by the original diffusion model. We conduct experiments on the FSD Kaggle 2018 dataset [17] and objec-

[†] Indicates equal contribution.

This work was supported by ONR N00014-23-1-2050 and N00014-23-1-2086 and NIH U01AG058532

[1] Demos and source code: https://jhu-lcap.github.io/DPM-TSE.

tive measures show that the perceptual quality of DPM-TSE is much better than the-state-of-art discriminative models. Subjective evaluations consistently show a preference among human listeners for the audio extracted via DPM-TSE, underscoring its heightened efficacy in extracting target sounds and eliminating irrelevant sounds.

2. METHODOLOGY

2.1. Diffusion Probabilistic Model

Diffusion probabilistic models include a forward and a backward process. The forward process gradually adds Gaussian noise to the data, commonly based on a manually-defined variance schedule $\beta_1, \ldots, \beta_T$.

$$q\left(x_{1:T} \mid x_0\right) := \prod_{t=1}^{T} q\left(x_t \mid x_{t-1}\right) \tag{1}$$

$$q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t \mathbf{I}\right) \tag{2}$$

The forward process allows sampling x_t at an arbitrary timestep t in a closed form:

$$q\left(x_t \mid x_0\right) := \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t} x_0, \left(1-\bar{\alpha}_t\right) \mathbf{I}\right) \tag{3}$$

Equivalently:

$$x_t := \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t} \epsilon, \quad \text{where } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$.

Diffusion models learn the reverse process to recover information step by step. In this way, DPM can generate new data from random Gaussian noises. When β_t is small, the reverse step is also found to be Gaussian:

$$p_\theta\left(x_{0:T}\right) := p\left(x_T\right) \prod_{t=1}^{T} p_\theta\left(x_{t-1} \mid x_t\right) \tag{5}$$

$$p_\theta\left(x_{t-1} \mid x_t\right) := \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t, \tilde{\beta}_t \mathbf{I}\right) \tag{6}$$

where variance $\tilde{\beta}_t$ can be calculated from the forward process posteriors: $\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$

In most previous DPMs, neural networks are used to predict noise ϵ, since:

$$\tilde{\mu}_t := \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon\right) \tag{7}$$

2.2. Corrected Noise Schedule and Sampling Steps

The original noise schedule commonly used in DPMs will lead to a non-zero Signal-to-noise ratio (SNR) at the last timestep T, where the SNR can be calculated as:

$$\operatorname{SNR}(t) := \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \tag{8}$$

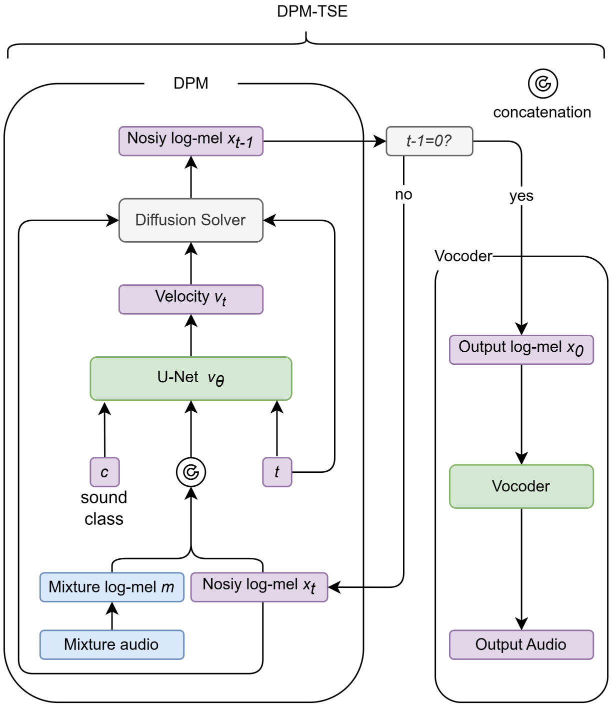

Fig. 1: The inference framework of Diff-TSE.

In the field of image generation, this problem is assumed to limit the generated images to have plain medium brightness, making it difficult to generate completely dark or white image content [18]. When it comes to TSE, the extracted target sound often contains many silent regions. Therefore, a non-zero terminal SNR might prevent the model from generating completely silent frames, impairing the purity and overall performance of sound extraction. Following [18], we adjust existing noise schedules to enforce zero terminal SNR by keeping $\sqrt{\bar{\alpha}_1}$ unchanged, changing $\sqrt{\bar{\alpha}_T}$ to zero, and linearly rescaling $\sqrt{\bar{\alpha}_t}$ for intermediate $t \in [2, \ldots, T-1]$ respectively.

When SNR is zero at the terminal step, it becomes meaningless to predict noise ϵ, as the input and output become the same. Therefore, the neural network is switched to predict velocity v instead:

$$v_t := \sqrt{\bar{\alpha}_t} \epsilon - \sqrt{1-\bar{\alpha}_t} x_0 \tag{9}$$

$$\epsilon = \sqrt{\bar{\alpha}_t} v + \sqrt{1-\bar{\alpha}_t} x_t \tag{10}$$

According to (4) and (7), the backward process is then performed by the following functions:

$$x_0 := \sqrt{\bar{\alpha}_t} x_t - \sqrt{1-\bar{\alpha}_t} v_t \tag{11}$$

$$\tilde{\mu}_t := \frac{\sqrt{\bar{\alpha}_{t-1}} \beta_t}{1-\bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_{t-1}\right)}{1-\bar{\alpha}_t} x_t \tag{12}$$

At the terminal step, the neural network with v prediction now predicts the mean of the data distribution under the given conditions. Additionally, the diffusion sampler always starts from the last timestep during inference.

2.3. DPM-TSE Framework

As shown in Figure 1, DPM-TSE comprises two modules: a diffusion model for generating the log-mel spectrogram of the target sound conditioned on the mixture audio and the target sound class token, and a neural vocoder for waveform reconstruction. The neural network $v_\theta(x_t, m, c, t)$ with parameters θ in the diffusion model is used to predict velocity v_t given the noisy target sound x_t, the audio mixture m, the one-hot target sound token c, and the corresponding diffusion step t. The diffusion step t is encoded by sinusoidal position embedding [19]. The architecture of the diffusion network is based on U-Net [20] consisting of 4 downsampling blocks and 4 upsampling blocks, each of which includes 2 convolutional blocks and 2 self-attention blocks, so that the model will be able to capture both local and temporal features of sound events. The HiFi-GAN vocoder [21] trained on AudioSet [22] is employed as the neural vocoder for universal audio waveform reconstruction.

3. EXPERIMENTAL SETUPS

3.1. Dataset

Following [7, 8], we formulate datasets comprised of synthetic sound event mixtures using the Freesound Dataset Kaggle 2018 corpus (FSD) [17]. This corpus encompasses a wide variety of 41 sound event categories ranging from human-produced sounds to musical instruments and object noises. Audio clips in the FSD have durations varying from 0.3 to 30 seconds. We generate 10-second audio mixtures. Each mixture incorporates one target sound and 1-3 interfering sounds randomly selected from the FSD. These are then superimposed at arbitrary time points over a 10-second background noise, which we obtain from the DCASE 2019 Challenge's acoustic scene classification task [23]. The signal-to-noise ratio (SNR) for each foreground sound is randomly set within a range of -5 to 10 dB. To optimize computational efficiency, all audio clips are down-sampled to 16 kHz. The dataset is partitioned into training, validation, and testing sets, containing 47,356, 16,000, and 16,000 samples respectively.

3.2. DPM-TSE Setups

The default U-Net model in DPM-TSE has 4 downsampling and 4 upsampling blocks configured with 128, 256, 512, and 512 channels respectively, totaling 106.40M parameters. The larger model variant has channel configurations of 194, 384, 768, and 768, with 239.30M total parameters. One-hot vector is applied for each target event with an embedding of 256 hidden units. Mel-spectrogram is used as our training target of U-Net since it can provide compact acoustic features and has been successfully used in many audio tasks [24, 25]. In our experiments, we use 64-dimensional mel-spectrograms with a window size of 64 ms and a hop size of 10 ms, and we zero-pad mel-spectrograms if the number of frames is not a multiple of 4. We use randomly segmented mel-spectrogram clips containing part of target sounds for training. The model is trained using the Adam optimizer with a learning rate of 0.0001, a weight decay of 0.0001, batch size of 24 and 150 epochs. The default DPM-TSE model uses corrected schedule and sampling steps and is trained with the v prediction. The diffusion steps and inference steps for the default DPM-TSE are 1000 and 50, and the corresponding variance β is set from 0.0001 to 0.02. For the DPM-TSE with 100 diffusion steps and 30 inference steps, the variance β is set from 0.0001 to 0.06.

3.3. Baselines

We utilize two latest TSE models, WaveFormer and Tim-TSENet with the same settings of their original implementations, as our baselines. WaveFormer and Tim-TSENet both use masking-based discriminative strategies for TSE. WaveFormer [26] is a time-domain TSE model which incorporates transformer blocks. Tim-TSENet [8] proposes an STFT-based TSE. For fair comparison, we also tried mel-spectrogram-based Tim-TSNet and STFT-based DPM-TSE. However, the Mel-spectrogram-based Tim-TSENet exhibited a performance degradation due to the difficulty of introducing a time-domain loss function using inverse STFT as in the original Tim-TSENet. Meanwhile, the STFT-based DPM-TSE suffered from significant performance degradation and excessive computational complexity.

3.4. Evaluation Metrics

The primary objective of this study is to enhance the auditory quality of the output generated through TSE. As such, we opt for a several perceptual evaluations and subjective assessment to gauge the performance of the target sound extraction models. We steer away from relying solely on objective measures such as Signal-to-Distortion Ratio (SDR) [27], since existing objective metrics for source separation are imperfect proxies for human auditory perception, as highlighted in previous research [28, 29].

Objective metrics: We use two automatic evaluation functions: (1) **ViSQOL** [30] is an algorithm originally designed to predict the quality of speech signals, and has since been adapted to assess the quality of audio signals by approximating human perceptual responses based on five-scaled mean opinion scores. (2) **CDPAM** [31] is a perceptual audio metric based on deep neural network that correlates well with human subjective ratings across sound quality assessment tasks, measuring audio similarity by distance of deep features.

Human evaluation: For subjective evaluation, 15 participants with recording or music production experiences were recruited to evaluate the listening perceptual quality of audios predicted by different TSE models. We randomly selected 1 sample from 41 sound categories from the test set. Each subject was asked to evaluate 20 randomly assigned audio pairs for each model, and each audio pair contains both a ground

Table 1: Objective and subjective scores with their 95% confidence intervals. ViSQOL-T and CDPAM-T are calculated with the target sound regions, while other scores are calculated with the whole audio.

Method	ViSQOL ↑	CDPAM ↓	ViSQOL-T ↑	CDPAM-T ↓	Extraction ↑	Purity ↑
WaveFormer [26]	1.96 ± 0.05	0.38 ± 0.02	1.78 ± 0.05	0.50 ± 0.02	3.38 ± 0.17	2.61 ± 0.19
Tim-TSENet [8]	2.32 ± 0.05	0.31 ± 0.02	2.04 ± 0.05	0.42 ± 0.02	3.80 ± 0.18	3.19 ± 0.21
DPM-TSE	$\mathbf{2.53 \pm 0.05}$	$\mathbf{0.22 \pm 0.01}$	$\mathbf{2.18 \pm 0.05}$	$\mathbf{0.38 \pm 0.03}$	$\mathbf{4.19 \pm 0.14}$	$\mathbf{3.74 \pm 0.18}$

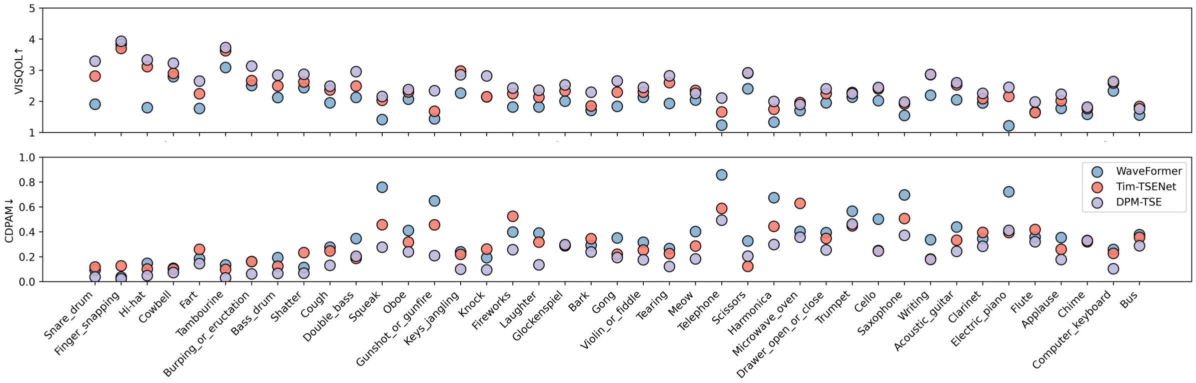

Fig. 2: Distribution of objective performance by sound category in ascending order of average sound event duration.

Table 2: Results of ablation study on DPM-TSE based on objective scores with their 95% confidence intervals. Steps refer to the diffusion and inference steps of the diffusion model.

Model	Schedule	Steps	ViSQOL ↑	CDPAM ↓
Base	Default	1000/50	2.39 ± 0.06	0.34 ± 0.02
Base	Corrected	100/30	2.43 ± 0.05	0.25 ± 0.01
Base	**Corrected**	**1000/50**	$\mathbf{2.53 \pm 0.05}$	$\mathbf{0.22 \pm 0.01}$
Large	Corrected	1000/50	2.38 ± 0.05	0.24 ± 0.01

truth and a model prediction for the extracted sound. They were given two questions for each audio pair: (1) **Extraction: Does the generated audio contain everything from the reference audio?** Rating from 1 to 5, where 1 means that the content of the reference audio cannot be heard at all in the generated audio, and 5 means that the generated audio completely contains everything from the reference audio. (2) **Purity: Does the generated audio only have the sound from the reference audio?** Rating from 1 to 5, where 1 means that it is pretty obvious that the generated audio has a lot of sounds that the reference audio doesn't have, and 5 means that the generated audio only has the sound corresponding to the reference audio and other sounds cannot be detected.

4. RESULTS

The results[2] in Table 1 demonstrate that DPM-TSE achieves the best performance in both subjective and objective experiments. The key observations include: (1) DPM-TSE has a promising performance in localizing and recovering target sound. (2) DPM-TSE shows a significant advantage of producing cleaner target sound, while Tim-TSENet and WaveFormer fail to remove non-target sound very well, especially in regions where the target sound overlaps with other sounds.

In Fig. 2, we explore the performance of target sound extraction in different sounds categories based on objective metrics. The three models simultaneously show good results for short-duration events (like finger snapping, tambourine, cowbell and hi-hat) while performance drops for long-duration complex events (like bus, saxophone, chime and flute). CDPAM and ViSQOL have similar distributions across the majority of classes. With that, we clearly note that DPM-TSE demonstrates pronounced advantages in most categories.

In addition, we conduct ablation study on noise schedule methods, number of training and inference steps, and model scales. As shown in Table 2, the proposed corrected noise schedule significantly improves the model performance. We find that the DPM-TSE using the original noise schedule produces additional noise, which is prominently noticeable in non-target sound regions. The DPM-TSE with larger model shows a performance degradation, which may be due to overfitting. Comparing 100 training steps with 1000 training step, we find that the DPM-TSE model with fewer diffusion and inference steps still achieves relatively good performance and can be used in situations where faster inference is preferred.

5. CONCLUSION

In this paper, we propose a DPM-based generative method for TSE, which is quite effective at extracting target sounds and removing irrelevant sounds. In future works, our focus will pivot towards (1) enhancing the sampling speed of DPM-TSE and (2) delving into innovative avenues including zero-shot TSE and text-guided TSE and audio editing techniques.

[2] Audio samples for comparison: https://jhu-lcap.github.io/DPM-TSE.

6. REFERENCES

[1] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech 2019*, 2019, pp. 2728–2732.

[2] Olga Slizovskaia, Gloria Haro, and Emilia Gómez, "Conditioned source separation for musical instrument performances," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2083–2095, 2021.

[3] Marc Delcroix, Jorge Bennasar Vázquez, Tsubasa Ochiai, Keisuke Kinoshita, and Shoko Araki, "Few-Shot Learning of New Sound Classes for Target Sound Extraction," in *Proc. Interspeech 2021*, 2021, pp. 3500–3504.

[4] Beat Gfeller, Dominik Roblek, and Marco Tagliasacchi, "One-shot conditional audio filtering of arbitrary sounds," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 501–505.

[5] Fatemeh Pishdadian, Gordon Wichern, and Jonathan Le Roux, "Learning to separate sounds from weakly labeled scenes," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 91–95.

[6] Qiuqiang Kong, Yuxuan Wang, Xuchen Song, Yin Cao, Wenwu Wang, and Mark D Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 101–105.

[7] Tsubasa Ochiai, Marc Delcroix, Yuma Koizumi, Hiroaki Ito, Keisuke Kinoshita, and Shoko Araki, "Listen to What You Want: Neural Network-Based Universal Sound Selector," in *Proc. Interspeech 2020*, 2020, pp. 1441–1445.

[8] Helin Wang, Dongchao Yang, Chao Weng, Jianwei Yu, and Yuexian Zou, "Improving Target Sound Extraction with Timestamp Information," in *Proc. Interspeech 2022*, 2022, pp. 1526–1530.

[9] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan, "Wavegrad: Estimating gradients for waveform generation," in *International Conference on Learning Representations*, 2020.

[10] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao, "Diffsinger: Singing voice synthesis via shallow diffusion mechanism," in *AAAI Conference on Artificial Intelligence*, 2021.

[11] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *Proceedings of the International Conference on Machine Learning*, 2023.

[12] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7402–7406.

[13] Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo 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.

[14] Robin Scheibler, Youna Ji, Soo-Whan Chung, Jaeuk Byun, Soyeon Choe, and Min-Seok Choi, "Diffusion-based generative speech source separation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[15] Naoyuki Kamo, Marc Delcroix, and Tomohiro Nakatani, "Target Speech Extraction with Conditional Diffusion Model," in *Proc. INTERSPEECH 2023*, 2023, pp. 176–180.

[16] Theodor Nguyen, Guangzhi Sun, Xianrui Zheng, Chao Zhang, and Philip C Woodland, "Conditional diffusion model for target speaker extraction," *arXiv preprint arXiv:2310.04791*, 2023.

[17] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 486–93.

[18] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang, "Common diffusion noise schedules and sample steps are flawed," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5404–5411.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[21] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[22] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[23] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Scenes and Events 2018 Workshop (DCASE2018)*, 2018, p. 9.

[24] Helin Wang, Yuexian Zou, Dading Chong, and Wenwu Wang, "Environmental Sound Classification with Parallel Temporal-Spectral Attention," in *Proc. Interspeech 2020*, 2020, pp. 821–825.

[25] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[26] Bandhav Veluri, Justin Chan, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota, "Real-time target sound extraction," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[27] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[28] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John 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.

[29] Mark Cartwright, Bryan Pardo, and Gautham J Mysore, "Crowdsourced pairwise-comparison for source separation evaluation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 606–610.

[30] Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 twelfth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[31] Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 196–200.