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/ICASSP49357.2023.10095979

SEPDIFF: SPEECH SEPARATION BASED ON DENOISING DIFFUSION MODEL

Bo Chen, Chao Wu, Wenbin Zhao

Huawei Technologies, Hangzhou, China

ABSTRACT

Speech separation aims to extract multiple speech sources from mixed signals. In this paper, we propose SepDiff - a monaural speech separation method based on the denoising diffusion model (diffusion model). By modifying the diffusion and reverse process, we show that the diffusion model achieves an impressive performance on speech separation. To generate speech sources, we use mel spectrogram of the mixture as a condition in the training procedure and insert it in every step of the sampling procedure. We propose a novel DNN structure to leverage local and global speech information through successive feature channel attention and dilated 2-D convolution blocks on multi-resolution time-frequency features. We use a neural vocoder to get waveform from the generated mel spectrogram. We evaluate SepDiff on LibriMix datasets. Compared to SepFormer approach, SepDiff yields a higher mean opinion score (MOS) of 0.11.

Index Terms— speech separation, diffusion model, generative model, deep learning

1. INTRODUCTION

Many attempts based on deep neural networks (DNN) have been made in previous works, which brought dramatic progress in speech separation. Early works use short-time Fourier transform (STFT) of the mixture as the input and calculate a mask on each source. The waveform is calculated using the inverse STFT (iSTFT) of the estimated magnitude [1, 2]. Since TasNet [3] was proposed in 2019, more and more time domain architectures have been studied [4]. These architectures usually consist of an encoder, separator, and decoder. Encoder and decoder are applied to replace STFT and ISTFT, providing a better transformation for mixed signal and overcoming phase reconstruction problems. The separator between the encoder and decoder is designed in different ways to generate masks, indicating the contribution of each source on latent feature [5, 6]. Among these time domain architectures, SepFormer [7] with dynamic mixing achieves the best SI-SNRi of 22.3 dB on WSJ0-2mix dataset. Recently, SFSRNet [8] and TF-GridNet [9] further improve the performance of DNN approach by introducing the super-resolution (SR) network into existing models and operating in the complex T-F domain, respectively.

The generative method, which aims at learning the distribution of clean speech as prior knowledge, is another approach to speech separation. Several works have utilized generative adversarial networks (GANs) [10, 11], flow-based models [12] and variational autoencoders (VAEs) [13, 14] for speech separation. Recently, diffusion models [15], which achieve state-of-the-art performance in image generation, have been introduced to the research of speech enhancement [16, 17, 18]. The diffusion model consists of the diffusion and reverse processes, as shown in Fig. 1. In the diffusion process, Gaussian noise is added to clean speech gradually (from right to left). In the reverse process, a DNN is learned to sample clean speech from Gaussian noise iteratively (from left to right).

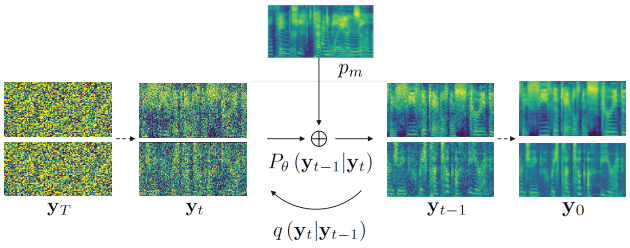

Fig. 1. The diffusion and reverse process of SepDiff. In every step t, the mixture is concatenated with noisy speech sources to guide the reverse process.

This paper introduces SepDiff, a method for speech separation based on the diffusion model. The diffusion process is applied to mel spectrogram of clean sources, and the reverse process is guided by mel spectrogram of the mixture. We start from Gaussian noise and sample certain steps until we get clean mel spectrogram. Then we use a neural vocoder to convert the mel spectrogram to a waveform. The contributions of this work are listed as follows:

(1) We propose SepDiff, which is the first speech separation method based on the diffusion model.

(2) We propose a novel DNN structure in the reverse process. This structure is based on U-Net architecture. In each block, feature channel attention and stacked dilated 2-D convolution are employed to leverage local and global information.

(3) The experiment result shows the effectiveness of SepDiff. MOS achieved by SepDiff is 0.11 higher than SepFormer.

2. SEPDIFF

2.1. Diffusion Model

The diffusion and reverse process of SepDiff is shown in Fig. 1. Two-channel Gaussian noise is used as input, and a mixture guides the reverse process to get speech sources. Define mel spectrogram of the mixture as $\mathbf{y}_0$, and is of dimension $(2, F, L)$, where F denotes the features dimension and L denotes the time sequence length. We get $\mathbf{y}_t$ by adding Gaussian noise on $\mathbf{y}_0$ at time step t, and $\mathbf{y}_T$ contains maximum noise. When T is sufficiently large, $\mathbf{y}_T$ correspond to Gaussian distribution. Diffusion process is a Markov chain, and $\mathbf{y}_t$ is obtained by adding a small amount of Gaussian noise to $\mathbf{y}_{t-1}$ [15]:

$$q\left(\mathbf{y}_t|\mathbf{y}_{t-1}\right) := \mathcal{N}\left(\mathbf{y}_t; \sqrt{1-\beta_t}\mathbf{y}_{t-1}, \beta_t\mathbf{I}\right), \quad (1)$$

where $\mathbf{I}$ denotes the identity matrix, and $(\beta_1, \ldots, \beta_T)$ is a variance schedule. When Gaussian noise is added for t steps, $\mathbf{y}_t$ can be written as:

$$q\left(\mathbf{y}_t|\mathbf{y}_0\right) := \mathcal{N}\left(\mathbf{y}_t; \sqrt{\overline{\alpha}_t}\mathbf{y}_0, \left(1-\overline{\alpha}_t\right)\mathbf{I}\right), \quad (2)$$

where $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, then $\mathbf{y}_t$ can be directly written as:

$$\mathbf{y}_t = \sqrt{\overline{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\overline{\alpha}_t}\boldsymbol{\epsilon}, \quad (3)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Reverse process P_θ is also regarded as a Markov chain from $\mathbf{y}_T$ to $\mathbf{y}_0$:

$$P_\theta\left(\mathbf{y}_{t-1}|\mathbf{y}_t\right) \sim \mathcal{N}(\mathbf{y}_{t-1}; \mu_\theta\left(\mathbf{y}_t, t\right), \sum_\theta\left(\mathbf{y}_t, t\right)). \quad (4)$$

In this paper, we modify original idea in DDPM [15] to use diffusion model for speech separation. Define mel spectrogram of single-channel mixture as p_m, which has dimension $(1, F, L)$. In the training of diffusion model, we expect mel spectrogram of clean sources, for a given p_m. To achieve this, we concatenate p_m and $\mathbf{y}_t$ along channel dimension to get $\mathbf{Y}_t := \mathbf{y}_t \oplus p_m$. $\mathbf{Y}_t$ has dimension $(3, F, L)$. During training, Gaussian noise is only added to first two channels. Then $\mathbf{y}_{t-1}$ is predicted from $\mathbf{y}_t$ by learnable parameters θ. We set $\sigma_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$ [15] and $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$\mathbf{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{Y}_t, t)) + \sigma_t\mathbf{z}. \quad (5)$$

Since the distribution $q\left(\mathbf{y}_{t-1}|\mathbf{Y}_t\right)$ is intractable, we train a DNN - $\boldsymbol{\epsilon}_\theta$ to approximate it. In training procedure, DNN is trained with ground truth $\mathbf{y}_0$. In sampling procedure, we get speech sources from Gaussian noise $\mathbf{y}_T$ by sampling for T steps. The training and sampling procedure of SepDiff are described in Algorithm 1 and Algorithm 2.

Algorithm 1 Training procedure

for $i = 1, 2, \ldots, N_{iter}$ **do**
  Sample $(\mathbf{y}_0, p_m) \sim q_{data}$, with $\epsilon \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$
  $t \sim$ Uniform $(1, \ldots, T)$
  $\mathbf{Y}_t := \left(\sqrt{\overline{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon\right) \oplus p_m$
  Take gradient descent step on
    $\nabla_\theta || \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\mathbf{Y}_t, t\right) ||$
end for

Algorithm 2 Sampling procedure

Input: p_m, mel spectrogram of mixture
Sample $\mathbf{y}_T \sim (\mathbf{0}, \boldsymbol{I})$
for $t = T, \ldots, 1$ **do**
  $\mathbf{Y}_t := \mathbf{y}_t \oplus p_m$
  $\mathbf{z} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{I}\right)$, if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{y}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{y}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{Y}_t, t\right)\right) + \sigma_t\mathbf{z}$
end for
return $\mathbf{y}_0$

2.2. DNN Structures

In this paper, we propose a U-Net architecture based on down/up blocks shown in Fig. 2 to approximate the distribution $q\left(\mathbf{y}_{t-1}|\mathbf{Y}_t\right)$.

Motivated by Restormer [19], we introduce multi-Dconv head transposed attention (MDTA) module in each block to aggregate local and global speech information. We adapt the U-Net in several ways for our task: (1) We use the sinusoidal position embedding followed by two linear and self-gated activation layers (LS) to get time-step embedding. In each block, time-step embedding passes through another LS layer and is added to each channel of the block to provide information of time-step t for the model. (2) We insert stacked 2-D dilated convolutional blocks (2-D DilConv) into each block to ensure a sufficiently receptive field on both time and frequency dimensions. The dilation factors increase exponentially and are different on two dimensions. 2-D DilConv module helps the DNN to take advantage of the long-term correlations of the speech signal. (3) We add a pre-processing stage with the same structure shown in Fig. 2 at the beginning of the U-Net to increase the capacity of the DNN at high time-frequency resolution.

2.3. Neural Vocoder

In this paper, we use BigVGAN [20] as a neural vocoder to generate waveform from mel spectrogram. BigVGAN introduces periodic nonlinearities and anti-aliased representation into the generator to improve audio quality. In practical applications, speech separation methods are often employed to separate signals mixed by unseen speakers. BigVGAN achieves state-of-the-art performance under various unseen

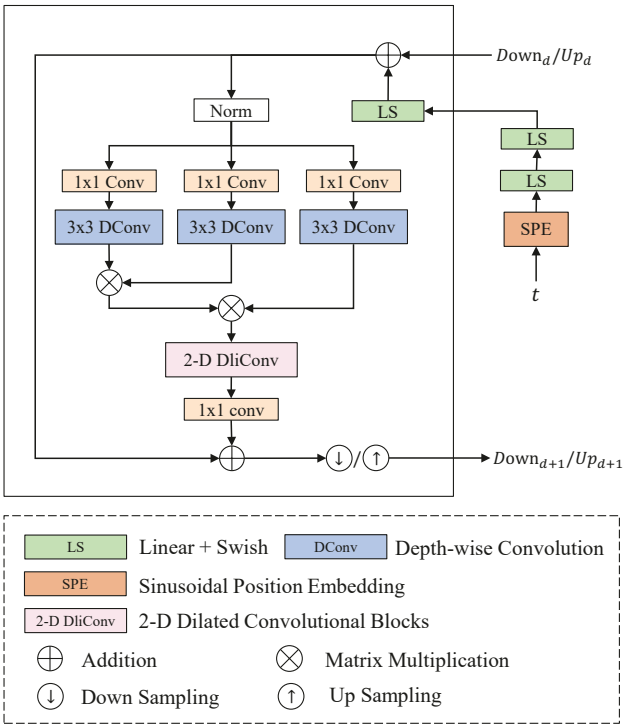

Fig. 2. Architecture of each block in proposed DNN.

conditions, including new speakers. We follow the configuration of BigVGAN-base with 14M parameters. The number of mel filters and STFT parameters are modified to match SepDiff.

3. EXPERIMENTAL SETUP

3.1. Data

LibriMix [21] dataset is an open source dataset for speech separation and has been shown to achieve better generalization under various conditions. LibriMix dataset takes speech utterances from two or three speakers of LibriSpeech and mixes them together. In this paper, we evaluate SepDiff on Libri2Mix, which consists of 270 hours of training data, 11 hours of validation data, and 11 hours of evaluation data. The mixed data comes from around 470 hours of speech from 1252 speakers.

3.2. Model Details

We extract mel spectrogram from raw waveform as speech representation. Hann window is applied as the analysis window; the window size is 50 ms; the hop size is 12.5 ms. Firstly a 2048-point discrete Fourier transform is applied to extract the spectrum. Then a mel filter bank is used to obtain an 80-dimensional mel spectrogram, and finally, mel spectrogram is coverted to logarithmic form. In the DNN structure, define D as the depth of the U-Net. At depth-0, define X as the number of convolutional blocks, and C as the number of channels. Starting from the high time-frequency resolution input, the down block halves the feature size, while doubling the number of channels hierarchically.

3.3. Training and Sampling Procedure

During training, we set maximum time step T to 1000 and variance schedule β from to $1e^{-4}$ to 0.02. During sampling, to balance the computational effort and the performance of SepDiff, we evaluate the performance with different reverse steps T_{infer}, which will be analyzed in Section 4.3. We train the model for 1M steps with L1 loss, Adam optimizer with a learning rate of 1e-4, and a batch size of 4.

3.4. Metrics

We used MOS as the subjective evaluation metric, which varies from 1 (bad quality, serious interference) to 5 (excellent quality, no interference). We invited 15 qualified listeners to score separated sources in a quiet environment and average these values to get the final result.

We use perceptual objective listening quality analysis (POLQA) [22] and prediction for generative neural speech codecs (WARP-Q) [23] as the objective evaluation metric. POLQA is an ITU-T standard with a perceptual model for predicting speech quality. WARP-Q is proposed to evaluate generative neural vocoder based codecs. WARP-Q is more robust to slight misalignment between generative and target signals in the time-domain than traditional objective metrics. We set the maximum frequency to 8kHz and the number of mel frequency cepstrum coefficients (MFCCS) to 12 for WARP-Q calculation.

4. RESULTS AND DISCUSSIONS

4.1. Result on Libri2Mix

Table 1. MOS, POLAQ and WARP-Q results achieved by different method on Libri2Mix. GT as the reference signal for POLQA and WARP-Q calculation. Best values in each column are bold.

Method	MOS ↑	POLQA ↑	WARP-Q ↓
GT	4.39	–	–
GT-Mel	4.21	3.90	0.40
ConvTasnet	3.51	2.78	0.89
DPRNN	3.70	3.05	0.84
SepFormer	3.79	**3.21**	0.83
SepDiff(proposed)	**3.90**	3.12	**0.82**

In Table 1, we report the performance of our proposed method SepDiff with Ground Truth (GT), GT-mel, ConvTasnet, DPRNN, and SepFormer on Libri2Mix test set. In the subjective test, utterances with same content but from different methods are shuffled, and listeners do not know the order of these utterances. GT-Mel uses mel spectrogram of GT as input to the neural vocoder. Thus, GT-Mel can be regarded as the upper limit of the SepDiff performance. We observe that GT-Mel achieves a MOS of 4.21, which is close to the MOS of GT. Compared with other speech separation methods, SepDiff achieves the best MOS score of 3.90 and WARP-Q of 0.82, indicating the effectiveness of our proposed method. In the meanwhile, we observe that SepDiff performs worse on POLQA than SepFormer. As mentioned in 3.4, there is always misalignment between the waveform generated by neural vocoder and the GT waveform, even with the same perceptual quality. This misalignment degrades the objective metrics considerably, which is common in generative models [24].

4.2. Ablation Study

Table 2. Ablation study of SepDiff.

Model	D	C	X	POLQA ↑	WARP-Q ↓
SepDiff-ϵ_θ	4	24	7	3.12	0.82
	4	12	7	2.65	0.92
	3	36	7	3.13	0.82
	4	24	4	3.05	0.84
	4	24	8	3.11	0.82
	4	36	7	**3.14**	**0.82**
DDPM-ϵ_θ	–	–	–	2.96	0.85
Restormer	–	–	–	3.04	0.83

In table 2, we evaluate the performance of SepDiff with different configurations. All the models are trained and tested on Libri2Mix datasets. From rows 1-6, we compare the results of DNN structures with different parameters. We observe that the number of channels is a crucial benefit for the model (rows 1-3); with sufficiently large channels ($C = 36$), using a smaller depth ($D = 3$) does not degrade the performance. The results in rows 1, 4, and 5 indicate that $X = 7$ provides enough receptive field for the model. From rows 6-8, we compare the performance of SepDiff with different DNN structures, SepDiff-ϵ_θ achieves better objective metrics than DDPM-ϵ_θ and Restormer.

4.3. Reverse Steps

During reverse process, because $\overline{\alpha}_t$ (solid line in Fig. 3(b)) is almost zero at start of procedure, latent $\mathbf{y}_t$ is too noisy and do not contribute much for final results as shown in Fig. 3(a). We use a fast reverse method by skipping most of the reverse steps at the start and decreasing the skipped step gradually (asterisk line in Fig. 3b) to get $\overline{\alpha}_{infer}$. Let $T_{infer} < T$ be the number of reverse steps. At reverse step s, we get t_s by $t_s = (s/T_{infer})^k \cdot T$, $s \in (1, T_{infer})$. This is an empirical formula, in which we set $k = 3$ to approximate the mapping from t_s to t. Then we use $\boldsymbol{\epsilon_\theta}(\cdot, t_s)$ to estimate noise. We compare WARP-Q, and POLQA results of the proposed method with different T_{infer} on Libri2Mix test set and generate speech sources with the same model configuration. The final result in Fig. 3(c) shows that WARP-Q and POLQA start to stagnate almost at the same T_{infer} of 143, which is selected as the optimal reverse step.

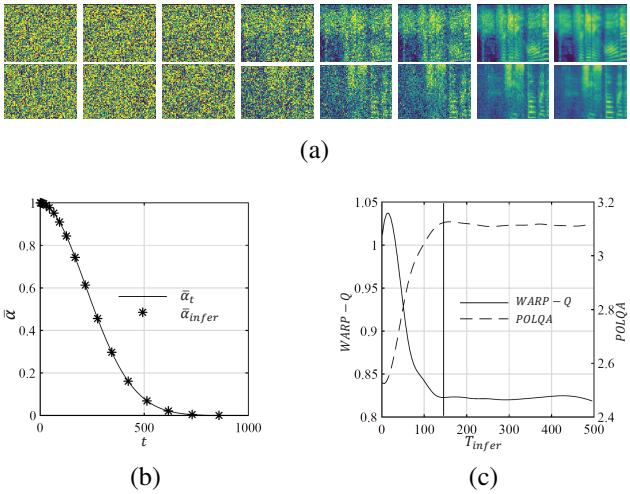

Fig. 3. (a) Latent $\mathbf{y}_t$ with t from T to 0 in reverse process (from left to right); (b) An example of $\overline{\alpha}_{infer}$ with $T_{infer} = 20$; (c) WARP-Q and POLQA results with different T_{infer}.

5. CONCLUSIONS

In this work, we proposed SepDiff, a monaural speech separation method based on the diffusion model. We used the mel spectrogram of the mixture to guide the reverse process by concatenating it with multi-channel Gaussian noise. Moreover, we introduced a novel DNN structure to estimate noise at different time steps t. The proposed DNN is capable of leveraging local and global information by successive feature channel attention and dilated 2-D convolution. Our proposed method achieved better MOS and WARP-Q results than traditional DNN approaches.

Instead of learning a mapping from mixture to single speech, SepDiff follows a diffusion approach to learn the prior distribution of speech. The proposed method has been proven to be an effective way to improve speech quality further. In future work, we plan to apply SepDiff on universal sources separation and directly generate waveform without the neural vocoder.

6. REFERENCES

[1] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 803–806.

[2] Z.-H. Tan M. Kolbæk, D. Yu and J. Jensen, "Multi-talker 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, 2017.

[3] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. ICASSP*, 2017, p. 697–700.

[4] 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, 2019.

[5] Z. Wang E. Tzinis and P. Smaragdis, "Sudo rm -rf: Efficient networks for universal audio source separation," in *Proc. MLSP*, 2020.

[6] 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.

[7] S. Cornell C. Subakan, M. Ravanelli and M. Bronzi, "Attention is all you need in speech separation," in *Proc. ICASSP*, 2021, pp. 21–25.

[8] J. Rixen and M. Renz, "Sfsrnet: Super-resolution for single-channel audio source separation," in *Proceedings of AAAI*, 2022.

[9] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation," *arXiv preprint arXiv:2209.03952*, 2022.

[10] Chenxing Li, Lei Zhu, Shuang Xu, Peng Gao, and Bo Xu, "Cbldnn-based speaker-independent speech separation via generative adversarial training," in *Proc. ICASSP*, 2018, pp. 711–715.

[11] Y Cem Subakan and Paris Smaragdis, "Generative adversarial source separation," in *Proc. ICASSP*, 2018, pp. 26–30.

[12] Aditya Arie Nugraha, Kouhei Sekiguchi, Mathieu Fontaine, Yoshiaki Bando, and Kazuyoshi Yoshii, "Flow-based independent vector analysis for blind source separation," *IEEE Signal Processing Letters*, vol. 27, pp. 2173–2177, 2020.

[13] Katerina Zmolikova, Marc Delcroix, Lukáš Burget, Tomohiro Nakatani, and Jan Honza Černocky, "Integration of variational autoencoder and spatial clustering for adaptive multi-channel neural speech separation," in *Proc. SLT*, 2021, pp. 889–896.

[14] Hao Duc Do, Son Thai Tran, and Duc Thanh Chau, "Speech source separation using variational autoencoder and bandpass filter," *IEEE Access*, vol. 8, pp. 156219–156231, 2020.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[16] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. APSIPA ASC*, 2021, pp. 659–666.

[17] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. ICASSP*, 2022, pp. 7402–7406.

[18] Simon Welker, Julius Richter, and Timo Gerkmann, "Speech enhancement with score-based generative models in the complex stft domain," *arXiv preprint arXiv:2203.17004*, 2022.

[19] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. CVPR*, 2022, pp. 5728–5739.

[20] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[21] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent, "Librimix: An open-source dataset for generalizable speech separation," *arXiv preprint arXiv:2005.11262*, 2020.

[22] ITU-T Rec. P.863, "Perceptual objective listening quality prediction," *Int. Telecom. Union (ITU)*, 2018.

[23] Wissam A Jassim, Jan Skoglund, Michael Chinen, and Andrew Hines, "Warp-q: Quality prediction for generative neural speech codecs," in *Proc. ICASSP*, 2021, pp. 401–405.

[24] Rithesh Kumar, Kundan Kumar, Vicki Anand, Yoshua Bengio, and Aaron Courville, "Nu-gan: High resolution neural upsampling with gan," *arXiv preprint arXiv:2010.11362*, 2020.