SCHRÖDINGER BRIDGES FOR SPEECH ENHANCEMENT WITH UNPAIRED DATA

Andreas Hansen Bagge *† *Michael Riis Andersen* * *Bjørn Sand Jensen* *

* DTU Compute, Technical University of Denmark. † WS Audiology.

ABSTRACT

Unpaired audio domain transfer is a promising yet underexplored direction for achieving robust speech enhancement models capable of handling real-world noisy and distorted speech. We propose solving the unpaired audio transfer problem by employing principled Schrödinger Bridges that provide a mapping between probability distributions using the Diffusion Schrödinger Bridge (DSB) algorithm. We employ state-of-the-art training methods for Schrödinger Bridges based on diffusion and provide an efficient implementation. The proposed approach is evaluated on three unpaired speech enhancement tasks; declipping, dereverberation, and denoising in both the Fourier domain and the Mel domain using both statistical and perceptual metrics. We compare against a Gaussian flow bridge model, which is an unpaired diffusion-based bridge model, as well as classical algorithms, and industry-standard deep learning models trained on paired data. We show that DSB trained on Log Mel spectrograms outperforms the diffusion baseline model across all metrics and surpasses classical algorithms and paired deep learning models on pMOS.

Index Terms**—** Unsupervised, Unpaired, Diffusion, Speech Enhancement, Schrödinger Bridge

1. INTRODUCTION

Speech enhancement problems, such as denoising, declipping, or dereverberation, are classic problems in audio processing. Traditionally, such problems are solved with supervised methods using collections of paired instances of clean and degraded speech, where the degraded speech is typically simulated by modifying the clean speech, e.g. by adding noise or clipping the amplitude. Such approaches are limited to training on simulated distortions, and cannot be applied to actual, in-the-wild degraded audio because large-scale paired data collection is infeasible or, in some instances, impossible. Enabling model training on unpaired in-the-wild audio therefore has the potential to improve speech enhancement models in real-world conditions.

Diffusion models constitute the current state of the art for image and audio synthesis [1, 2], and they have also shown great results in the field of paired speech enhancement [3]. Diffusion models generally consist of two processes, a forward and a backward process. In traditional diffusion models, the forward process gradually turns data samples into Gaussian noise. After learning the reverse process, the models can then synthesize high-quality samples from Gaussian noise. Though traditional diffusion methods yield high-quality samples, they do not allow for mapping between two arbitrary data distributions. In contrast, Schrödinger Bridges (SB) allow for high-quality sample generation and direct transfer between two arbitrary data distributions through the entropy-regularized optimal transport problem (SB Problem) of finding the most likely random evolution between two continuous probability distributions [4]. That is, Schrödinger Bridges allow learning stochastic maps between clean and degraded audio from unpaired data only, enabling training on in-the-wild data. SBs are formulated in terms of stochastic differential equations (SDEs), which can be simulated with SDE solvers, and solving the SB problem therefore allows for a diffusion-based generative approach, having the potential of also yielding high-quality samples. In this paper, we investigate the use of Diffusion Schrödinger Bridges (DSB) for speech enhancement using unpaired data. We show that DSB outperforms the Gaussian flow bridge method, which is a diffusion method for unpaired data, and that DSB outperforms paired deep learning models and signal processing approaches on perceptual metrics, demonstrating that DSB is a promising method for unpaired speech enhancement. Audio examples and code are available at kommodeskab.github.io/Latent-DSB/.

Related work: A^{2}SB (Audio-to-Audio Schrödinger Bridge) [5, 6] is a tractable, simulation-free method for learning maps between the distributions $p_{data}(\mathbf{X}_0)$ and $p_{prior}(\mathbf{X}_1|\mathbf{X}_0)$, usually representing clean and degraded samples respectively, i.e. a conditional probability distribution that requires paired samples. A^{2}ASB exhibits state-of-the-art results on paired audio restoration tasks such as bandwidth extension and inpainting. [7] propose using Gaussian Flow Bridges, a diffusion-based approach where samples are first mapped to a Gaussian, for unpaired domain transferring on audio samples. The approach suffers from the fact that the simulated trajectories first have to map to a third distribution, a Gaussian, therefore yielding non-optimal transport. RemixIT [8] is an example of a self-supervised method for training unpaired speech enhancement models. The RemixIT algorithm uses a teacher-student setup in which the teacher provides pseudo targets for the student. However, this approach requires the model to be pretrained on paired examples from a similar speech enhancement task and it is therefore not fully unpaired.

2. METHODS

Our work builds on the Diffusion Schrödinger Bridge algorithm, an iterative approach for solving the SB problem using the iterative proportional fitting algorithm [9, 10]. Given i.i.d. observations from two data distributions, i.e. $\mathbf{X}_0 \sim p_{data}, \mathbf{X}_1 \sim p_{prior}$ where $\mathbf{X}_0, \mathbf{X}_1 \in \mathbb{R}^n$, we seek two stochastic processes, or a 'bridge', that transports samples from p_{data} to p_{prior} and vice versa. Having obtained the bridge, samples can be transported between the two domains, thereby applying the characteristics of said domain while preserving content. We parameterize the bridge using a forward and backward SDE capable of transporting samples to and from the respective marginal distributions:

$$\begin{aligned} \mathrm{d}\mathbf{X}_t &= f(\mathbf{X}_t, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{W}_t, \qquad \mathbf{X}_0 \sim p_{data}, \\ \mathrm{d}\mathbf{X}_t &= b(\mathbf{X}_t, t)\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{W}}_t, \qquad \mathbf{X}_1 \sim p_{prior}. \end{aligned}$$

Here f and b are the drift terms, $g(t)$ is the diffusion term, and $\mathbf{W}$ and $\bar{\mathbf{W}}$ are standard Wiener processes. We will use $g(t) = 2$ throughout and parameterize the forward and backward drift with a single neural network denoted v_θ such that:

$$f(\mathbf{X}_t, t) \approx v_\theta(\mathbf{X}_t, t, 1), \qquad b(\mathbf{X}_t, t) \approx v_\theta(\mathbf{X}_t, t, 0).$$

The goal is to learn $v_\theta(\mathbf{X}_t, t, s)$, where $s \in \{0, 1\}$ is a binary indicator variable, such that simulating the backward process on the interval from $t = 1$ to $t = 0$ starting from p_{data} yields p_{prior} at time $t = 0$ (and oppositely for the forward process). We can simulate (sample) the forward and backward processes, respectively, on the discrete time interval $0 = t_0 < t_1 \cdots < t_N = 1$ where $\Delta t_{k+1} = t_{k+1} - t_k$ using:

$$p(\mathbf{X}_{t_{k+1}}|\mathbf{X}_{t_k}) \approx p(\mathbf{X}_{t_{k+1}}|\mathbf{X}_{t_k}, \mathbf{X}_1) = \mathcal{N}(\mathbf{X}_{t_{k+1}}; \tilde{\mu}_k, \tilde{\sigma}_k^2\mathbf{I}), \tag{1}$$

$$p(\mathbf{X}_{t_k}|\mathbf{X}_{t_{k+1}}) \approx p(\mathbf{X}_{t_k}|\mathbf{X}_{t_{k+1}}, \mathbf{X}_0) = \mathcal{N}(\mathbf{X}_{t_k}; \mu_{k+1}, \sigma_{k+1}^2\mathbf{I}), \tag{2}$$

where:

$$\begin{aligned} \tilde{\mu}_k &\approx \mathbf{X}_{t_k} + \Delta t_{k+1}(\mathbf{X}_1 - \mathbf{X}_{t_k})/(1 - t_k) \\ &\approx \mathbf{X}_{t_k} + \Delta t_{k+1} v_\theta(\mathbf{X}_{t_k}, t_k, 1), \\ \mu_{k+1} &\approx \mathbf{X}_{t_{k+1}} + \Delta t_{k+1}(\mathbf{X}_0 - \mathbf{X}_{t_{k+1}})/t_{k+1} \\ &\approx \mathbf{X}_{t_{k+1}} + \Delta t_{k+1} v_\theta(\mathbf{X}_{t_{k+1}}, t_{k+1}, 0), \end{aligned}$$

$$\tilde{\sigma}_k^2 = \frac{2\Delta t_{k+1}(1 - t_{k+1})}{1 - t_k}, \qquad \sigma_{k+1}^2 = \frac{2\Delta t_{k+1} t_k}{t_{k+1}}.$$

Given $\mathbf{X}_0$ and $\mathbf{X}_1$, intermediate points are sampled from:

$$p(\mathbf{X}_{t_k}|\mathbf{X}_0, \mathbf{X}_1) = \mathcal{N}((1 - t_k)\mathbf{X}_0 + t_k\mathbf{X}_1, 2t_k(1 - t_k)\mathbf{I}).$$

The algorithm has a pre-training and a fine-tuning phase. During pre-training, we draw a random pair of points $(\mathbf{X}_0, \mathbf{X}_1)$ from p_{data} and p_{prior} independently. During fine-tuning, pairs are generated by simulating the SDEs in either direction using v_θ such that each pair consists of the start- and endpoint of the simulated trajectory. In both stages, v_θ is trained by sampling an intermediate point using $p(\mathbf{X}_{t_k}|\mathbf{X}_0, \mathbf{X}_1)$, and then learning to estimate the forward flow $(\mathbf{X}_1 - \mathbf{X}_{t_k})/(1 - t_k)$ and backward flow $(\mathbf{X}_0 - \mathbf{X}_{t_k})/t_k$. Once the forward and backward flows are learned, we can finally simulate the forward and backward SDEs. Simulating SDEs using v_θ requires multiple forward passes and is computationally expensive. Therefore, we also keep a cache of simulated endpoints such that these endpoints can be pseudo-sampled by drawing a previously simulated endpoint instead of simulating a new one. The algorithm is outlined in 1. As mentioned, the DSB algorithm is computationally demanding due to the simulation of the SDEs, which requires N forward passes per full simulation. This problem can be mitigated by reducing the dimensionality of the training data. Additionally, lower-dimensional samples allow for storing more samples in the cache. We therefore propose training on logarithmic Mel spectrograms instead of raw waveforms or spectrograms. During inference, the Mel spectrograms can be decoded using a vocoder [11]. Unlike other popular encoders, generating logarithmic Mel spectrograms does not require expensive forward passes.

Algorithm 1 Training algorithm

```
Initialize neural network $v_\theta$ and batch size $2B$
Let $i \in \{1, 2 \ldots B\}$
while not converged do
    $\mathbf{X}_0^{1:B} \overset{\text{iid}}{\sim} p_{data}, \tilde{\mathbf{X}}_1^{1:B} \overset{\text{iid}}{\sim} p_{prior}$
    if pre-training then
        $\mathbf{X}_1^{1:B} \overset{\text{iid}}{\sim} p_{prior}, \tilde{\mathbf{X}}_0^{1:B} \overset{\text{iid}}{\sim} p_{data}$
    else
        Sample $\mathbf{X}_1^i$ using (1) and $v_\theta(\cdot, \cdot, 1)$ starting from $\mathbf{X}_0^i$
        Sample $\tilde{\mathbf{X}}_0^i$ using (2) and $v_\theta(\cdot, \cdot, 0)$ starting from $\tilde{\mathbf{X}}_1^i$
        Optional: Save pairs $(\mathbf{X}_0^i, \mathbf{X}_1^i)$ and $(\tilde{\mathbf{X}}_0^i, \tilde{\mathbf{X}}_1^i)$ to a
        cache for fast future pseudo sampling.
    end if
    $t^{1:B}, \tilde{t}^{1:B} \overset{\text{iid}}{\sim} \mathcal{U}(0, 1)$
    $\mathbf{X}_t^i \sim \mathcal{N}((1 - t^i)\mathbf{X}_0^i + t^i\mathbf{X}_1^i, 2t^i(1 - t^i)\mathbf{I})$
    $\tilde{\mathbf{X}}_t^i \sim \mathcal{N}((1 - \tilde{t}^i)\tilde{\mathbf{X}}_0^i + \tilde{t}^i\tilde{\mathbf{X}}_1^i, 2\tilde{t}^i(1 - \tilde{t}^i)\mathbf{I})$
    $\mathcal{L}_b = \frac{1}{B}\sum_{i=1}^B \left\| v_\theta(\mathbf{X}_t^i, t^i, 0) - \frac{\mathbf{X}_0^i - \mathbf{X}_t^i}{t^i} \right\|^2$
    $\mathcal{L}_f = \frac{1}{B}\sum_{i=1}^B \left\| v_\theta(\tilde{\mathbf{X}}_t^i, \tilde{t}^i, 1) - \frac{\tilde{\mathbf{X}}_1^i - \tilde{\mathbf{X}}_t^i}{1 - \tilde{t}^i} \right\|^2$
    Take gradient step $\nabla_\theta \frac{1}{2}(\mathcal{L}_b + \mathcal{L}_f)$
end while
```

3. EXPERIMENTS AND RESULTS

We evaluate the proposed DSB approach on several unpaired speech enhancement tasks: declipping, dereverberation, and denoising. These tasks are traditionally solved using paired data, where the degraded samples are created synthetically. As in paired setups, we synthetically degrade clean samples

			pMOS ↑	WER ↓	SR-CS ↑	KAD ↓
Declipping	U	DSB Mel (1)	$\mathbf{3.19 \pm 0.01}$	0.23 ± 0.01	$\mathbf{0.52 \pm 0.01}$	5.06
	U	DSB Mel (50)	3.11 ± 0.02	0.42 ± 0.02	0.46 ± 0.01	**2.05**
	U	DSB Mel Deterministic (50)	2.80 ± 0.03	0.32 ± 0.02	0.43 ± 0.01	4.88
	U	GFB (2)	1.74 ± 0.01	0.99 ± 0.00	0.05 ± 0.00	45.1
	U	GFB (50)	2.74 ± 0.02	0.58 ± 0.02	0.49 ± 0.01	5.44
	SP	SPADE	2.51 ± 0.02	$\mathbf{0.12 \pm 0.01}$	$\mathbf{0.52 \pm 0.01}$	6.25
		Baseline	2.56 ± 0.02	0.13 ± 0.01	0.47 ± 0.01	6.02
Dereverberation	U	DSB Mel (1)	$\mathbf{3.22 \pm 0.02}$	0.43 ± 0.11	0.53 ± 0.01	5.58
	U	DSB Mel (50)	3.16 ± 0.02	0.46 ± 0.02	0.45 ± 0.01	**3.87**
	U	DSB Mel Deterministic (50)	3.02 ± 0.02	0.47 ± 0.02	0.41 ± 0.01	6.82
	U	GFB (2)	1.64 ± 0.01	0.99 ± 0.00	0.01 ± 0.00	27.4
	U	GFB (50)	2.79 ± 0.02	1.25 ± 0.24	0.29 ± 0.01	6.94
	SP	WPE	2.05 ± 0.04	$\mathbf{0.15 \pm 0.01}$	$\mathbf{0.54 \pm 0.01}$	4.18
		Baseline	1.95 ± 0.04	0.15 ± 0.01	0.52 ± 0.01	4.41

Table 1: Evaluation metrics for different methods with 95% confidence interval. Number of diffusion steps is denoted in parentheses when relevant. U = unpaired training, SP = classical signal processing. Best unpaired model is underlined; best overall (including U and SP) is **bold**.

			pMOS ↑	SI-SDRi ↑
Denoising	U	DSB Mel (1)	$\mathbf{3.22 \pm 0.01}$	$-31.22 \pm 0.63^*$
	U	DSB Mel (50)	3.21 ± 0.01	$-31.86 \pm 0.63^*$
	U	DSB Mel Det. (50)	3.01 ± 0.03	$-32.61 \pm 0.67^*$
	U	DSB STFT (1)	2.74 ± 0.02	5.39 ± 0.11
	U	DSB STFT (50)	2.93 ± 0.02	2.72 ± 0.10
	P	SepFormer	3.04 ± 0.02	8.16 ± 0.21
	P	ConvTasNet	2.96 ± 0.02	$\mathbf{8.38 \pm 0.11}$
		Baseline	1.44 ± 0.02	0.00 ± 0.00

Table 2: SI-SDRi on denoising experiment comparing DSB with industry-standard denoising models. U = unpaired training, P = paired training. *The Mel vocoder does not preserve the phase, and hence, subsequent evaluations of SI-SDRi measures are less informative.

for evaluation, but critically, we do not pair the samples at any point during training. For the clean speech, p_{data}, we use the VCTK dataset [12] with speakers `p225` to `p230` used for validation. The degraded data, p_{prior}, is created from VCTK as follows:

- **Clipped** speech is generated as in [7] by drawing and applying a random gain between 5 and 30 dB, clipping the resulting signal at full scale, and then dividing by the gain.
- **Reverberated** speech is generated similar to [7] by applying a room impulse response (RIR) from [13, 14, 15, 16].
- **Noisy speech** is generated by mixing clean speech with WHAM noise [17] at an SNR level randomly sampled between -2 and 18 dB.

We test two different audio representations, namely raw STFT spectrograms and logarithmic Mel spectrograms. Training samples are 4.096 s long for the STFT approach and 4.47 s long for the Mel approach (to accommodate model architecture) and sampled at 16 kHz. The STFT spectrograms are calculated using an FFT size and window length of $N_{FFT} = L_{win} = 510$ and a hop length of $H_{hop} = 128$ samples, and the resulting real and imaginary components are stacked along the channel dimension. The Mel spectrograms are calculated using $N_{FFT} = L_{win} = 1024$, $H_{hop} = 160$, and $n_{mels} = 64$. The Mel spectrograms are decoded using a SpeechT5 vocoder [11]. Each model was trained for a total of $300k$ training steps, whereof $150k$ steps were used for pre-training. We use a conditional UNet architecture [1] (≈ 60 million parameters) to parameterize $v_\theta(\mathbf{X}_t, t, s)$. Throughout training we use a batch size of 8 for the STFT approach and batch size 64 for the Mel approach, and the AdamW optimizer with 10^{-4} learning rate. During fine-tuning, cached samples are generated with $N = 30$ and a symmetric cosine schedule, i.e., $t_k = 0.5\left(1 - \cos\left(\frac{k\pi}{N}\right)\right)$. The cache contains 3840 samples and is refreshed every 19200 training steps for the STFT approach, and it contains 10240 samples and is refreshed every 2500 training steps for the Mel approach. During inference, we use an exponential moving average (EMA) of the training weights with decay $\alpha = 0.999$. The models are trained using a single NVIDIA A100 GPU.

Test setup: We focus on evaluating the backward process, i.e., generating clean speech samples. We use 256 samples from the clean LibriSpeech dataset [18] as test set. The clipped test data is fixed at a signal-to-distortion (SDR) level of 2 dB, the reverberated test data is generated using an unseen RIR dataset [19], and the noisy speech test data is mixed at a random SNR between -2 and 5 dB. We compare with an unpaired, diffusion-based approach called Gaussian Flow Bridge [7] on declipping and dereverberation. GFB differs from our approach since it encodes and decodes audio samples to and from a Gaussian distribution by simulating ordinary differential equations. GFB and DSB are evaluated at 1 step to

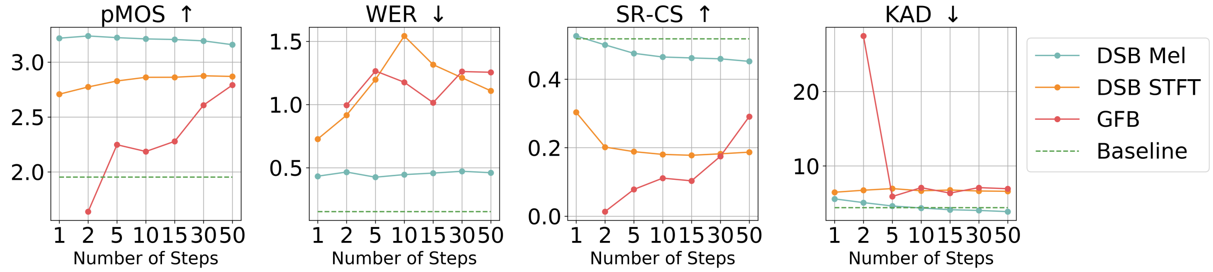

Fig. 1: Results on dereverberation experiment for different diffusion steps.

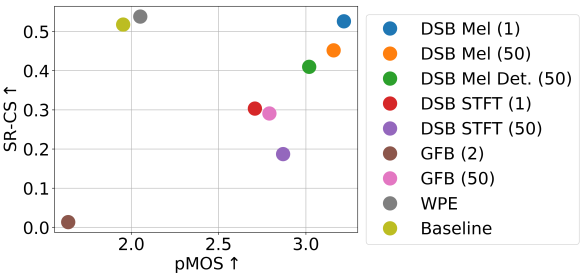

Fig. 2: SR-CS vs. pMOS on dereverberation experiment.

50 diffusion steps, though GFB cannot do one step diffusion since it needs to both encode and decode at least once, yieldng a minimum of two steps in total. We use the exact GFB model (≈ 40 million parameters) provided in [7], which was trained for $300k$ iterations. Since GFB is deterministic, i.e., no noise is added during inference, we also conduct the experiments with a deterministic version of DSB, where $\sigma^2_{k+1} = 0$. Following [7], we compare our approach with the performance on unprocessed degraded samples (baseline), and two classical approaches, namely SPADE [20], a synthesis-based sparse audio declipper, and WPE [21], a weighted prediction error dereverberation approach. We compare our denoising model with industry-standard speech enhancement models which use the traditional paired SI-SNR loss, namely a Sepformer [22][1] (≈ 25 million parameters) and a ConvTasNet [23][2] (≈ 5 million parameters) both trained on variations of WHAM.

Metrics: The generated samples are evaluated using WER (word error rate), pMOS (predicted mean opinion score) [24], SR-CS (speaker recognition cosine similarity), and KAD (kernel audio distance) [25]. WER and SR-CS can be interpreted as content fidelity measures capturing how well content is preserved, while pMOS and KAD can be interpreted as perceptual quality metrics. WER is calculated using the small Whisper model [26] to obtain ground-truth transcriptions. It is important to note that the transcription model is robust towards distorted speech and is therefore not necessarily a good measure of how well the audio is reconstructed. Nevertheless, WER serves as a useful indicator of whether the semantic content of the speech is preserved. SR-CS is calculated using a speaker embedding model [27] to obtain speaker embeddings of the ground truth signal and the generated signal. The metric is then calculated as the cosine similarity between these embeddings. The samples used in the KAD metric are first encoded using a CLAP model [28]. For the denoising task only, we also report SI-SDRi [29].

Results: Results are shown in tables 1-2 and in figures 1- 2. It is seen that DSB is uniformly superior on the pMOS metric, even outperforming the denoising models trained using paired data. Figure 1 shows that, on declipping and dereverberation, DSB using Mel Spectrograms outperforms GFB across diffusion steps while DSB using STFT yields comparable results. DSB is also more consistent across the number of diffusion steps, enabling faster inference. Figure 2 shows that DSB Mel is superior in terms of generating high-quality samples while preserving the speaker identity on the dereverberation task. Table 1 shows that DSB consistently beats GFB across all metrics. Additionally, DSB outperforms the traditional signal processing methods on perceptual metrics, but is outperformed on content fidelity metrics. Interestingly, Table 2 shows that the DSB model using STFT is comparable in terms of SI-SDRi compared to paired denoising models, while the SI-SDRi is low for the DSB Mel method as expected due to the fact that the vocoder is not phase-preserving.

4. CONCLUSION

We proposed Diffusion Schrödinger Bridges for speech enhancement using unpaired data only and evaluated the algorithm using both content and perceptual metrics. We demonstrate that DSB outperforms the GFB method by generating high-quality audio samples that resemble the target distribution while preserving content in the audio samples. Furthermore, DSB can outperform both traditional audio processing methods and deep learning models trained on paired data on perceptual metrics. This work demonstrates the potential of training fully unpaired speech enhancement models using Schrödinger Bridges, thereby enabling adaptation to naturally degraded, in-the-wild audio and improving the overall performance of speech enhancement systems.

[1] https://huggingface.co/speechbrain/sepformer-wham16k-enhancement

[2] https://huggingface.co/cankeles/ConvTasNet_WHAMR_enhsingle_16k

5. REFERENCES

[1] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” 2021.

[2] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, “Diffwave: A versatile diffusion model for audio synthesis,” 2021.

[3] J. Richter, S. Welker, J. Lemercier, B. Lay, and T. Gerkmann, “Speech enhancement and dereverberation with diffusion-based generative models,” 2023.

[4] E. Schrödinger, “Sur la théorie relativiste de l’électron et l’interprétation de la mécanique quantique,” *Annales de l’institut Henri Poincaré*, vol. 2, no. 4, pp. 269–310, 1932.

[5] Z. Kong, K. J Shih, W. Nie, A. Vahdat, S. Lee, J. F. Santos, A. Jukic, R. Valle, and B. Catanzaro, “A2SB: Audio-to-Audio Schrodinger bridges,” 2025.

[6] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, “Schrödinger bridge for generative speech enhancement,” 2024.

[7] E. Moliner, S. Braun, and H. Gamper, “Gaussian flow bridges for audio domain transfer with unpaired data,” 2024.

[8] E. Tzinis, Y. Adi, V. K. Ithapu, B. Xu, P. Smaragdis, and A. Kumar, “Remixit: Continual self-training of speech enhancement models via bootstrapped remixing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1329–1341, Oct. 2022.

[9] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, “Diffusion schrödinger bridge with applications to score-based generative modeling,” 2023.

[10] V. De Bortoli, I. Korshunova, A. Mnih, and A. Doucet, “Schrodinger bridge flow for unpaired data translation,” in *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 103384–103441.

[11] H. Liu, y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, “AudioLDM 2: Learning holistic audio generation with self-supervised pretraining,” 2024.

[12] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92),” 2019.

[13] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černocký, “Building and evaluation of a real room impulse response dataset,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 863–876, 2019.

[14] R. Stewart and M. Sandler, “Database of omnidirectional and b-format impulse responses,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, 2010.

[15] Real World Computing Partnership, “RWCP (RWCP-SSD),” 2007.

[16] D. Murphy and F. Stevens, “Open Air Library 2025,” 2025.

[17] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, “Wham!: Extending speech separation to noisy environments,” 2019.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An ASR corpus based on public domain audio books,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[19] J. Traer and J. H. McDermott, “Statistics of natural reverberation enable perceptual separation of sound and space,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[20] Pavel Zaviska, Pavel Rajmic, Ondrej Mokry, and Zdenek Prusa, “A proper version of synthesis-based sparse audio declipper,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[21] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, “NARA-WPE: A python package for weighted prediction error dereverberation in numpy and tensorflow for online and offline processing,” in *Speech Communication; 13th ITG-Symposium*, 2018.

[22] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” 2021.

[23] Y. Luo and N. Mesgarani, “Conv-Tasnet: Surpassing ideal time–frequency magnitude masking for speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[24] C. K A Reddy, V. Gopal, and R. Cutler, “Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors,” 2021.

[25] Y. Chung, P. Eu, J. Lee, K. Choi, J. Nam, and B. S. Chon, “KAD: No more FAD! An effective and efficient evaluation metric for audio generation,” 2025.

[26] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.

[27] N. R. Koluguri, T. Park, and B. Ginsburg, “TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[28] Y. Wu, K. Chen, T. Zhang, Y. Hui, M. Nezhurina, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” 2024.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “Sdr - half-baked or well done?,” 2018.

Compliance with Ethical Standards

This is a study of stochastic differential equations applied to audio for which no ethical approval was required.