Generating Underwater Acoustic Communication Channel Impulse Responses Using A Diffusion Model

Yongjie Zhuang* yongjie.zhuang@stonybrook.edu
Dariush Kari† dkari2@illinois.edu
Zhengnan Li‡
Milica Stojanovic‡ millitsa@ece.neu.edu
Andrew C. Singer* acs@sunysb.edu

* *Dept. Electrical and Computer Engineering*, *Stony Brook University*, Stony Brook, New York 11794, USA
† *Dept. Electrical and Computer Engineering*, *University of Illinois Urbana-Champaign*, Urbana, Illinois 61801, USA
‡ *Dept. Electrical Engineering*, *Northeastern University*, Boston, Massachusetts 02115, USA

***Abstract*—Underwater acoustic communication data is expensive to collect, often yielding experimental data sets that are sufficient for proof-of-concept analysis, but are insufficient for applications that demand orders of magnitude more channel realizations, such as data-driven methods for outage-capacity analysis, network simulation, or statistical regression of new communication algorithms. For such applications, a realistic synthetic dataset is a necessity for the successful generalization of data-driven underwater acoustic (UWA) communication systems. This paper investigates the application of diffusion models for UWA data augmentation for such systems. Diffusion models, as opposed to physics-based models such as BELLHOP or other acoustic propagation tools, extract the essential characteristics of the data without explicit knowledge of environmental parameters. We demonstrate the capability of such models by generating data whose multipath structure and spatiotemporal correlation match those of the Kauai ACOMMS MURI 2011 (KAM11) experiment.**

***Index Terms*—Underwater acoustics, channel impulse responses, diffusion model, data augmentation, multipath**

I. Introduction

In underwater acoustic (UWA) communications, the complex ocean environment, with multipath propagation and large spatiotemporal variability, presents a host of challenges for designing robust and effective digital communication systems [1], [2]. Consequently, a realistic dataset that captures these effects is crucial for designing UWA communication algorithms.

As there is no standard model that accurately depicts the intricate nature of UWA channels for general environments, experimental demonstration of candidate communication schemes has become the de facto standard [1], [3]. However, acquiring extensive field data is both expensive and time-consuming. For applications where a prohibitively large dataset is required for designing or evaluating UWA communication algorithms, data augmentation methods that are faithful to the underlying distribution of field measurements are essential. To mimic the randomness of multipath effects and spatiotemporal variability using physics-based models, one may use a parameterized model and introduce randomness into parameters, such as the surface profile, bathymetry, or sound speed profile, though such methods are computationally expensive and have not proven adequate. This motivates the pursuit of data-driven generative models, which learn to generate channel realizations by sampling from measured data.

Recently, deep generative models, particularly diffusion models, have proven to be a powerful tool for generating realistic datasets in image and audio synthesis [4], [5], and have shown great promise for generating radio-frequency digital communication signals [6] and learning ocean sound speed field [7]. This paper exploits a diffusion model that is compatible with UWA channel modeling applications. Generative models such as diffusion models can learn a distribution over the sampled data that captures the randomness exhibited therein. By sampling from that distribution (via experimental measurements), we can realize new channels from the distribution for that environment. We show the efficacy of the proposed method by comparing the statistics of the multipath arrival times in the generated impulse responses to measurements.

II. Methods

Generative models assume that data is created by an underlying probability distribution and create new data consistent with that distribution [8]. Diffusion models have proven useful in a variety of domains [4]. Conceptually, given a data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, a forward Markov process with Gaussian transition kernel $q(\mathbf{x}_r|\mathbf{x}_{r-1})$ generates a random variable sequence $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_R$ can be generated from available training data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ by using a forward Markov process with Gaussian transition kernel $q(\mathbf{x}_r|\mathbf{x}_{r-1})$: [4]

$$q(\mathbf{x}_r|\mathbf{x}_{r-1}) = \mathcal{N}(\sqrt{1-\beta_r}\mathbf{x}_{r-1}, \beta_r\mathbf{I}), \tag{1}$$

where $\beta_r \in (0,1)$ is the hyperparameter. Letting $\alpha_r := 1-\beta_r$ and $\bar{\alpha}_r := \Pi_{s=0}^{r}\alpha_s$, it can be shown that [4], [9]

$$q(\mathbf{x}_r|\mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_r}\mathbf{x}_0, (1-\bar{\alpha}_r)\mathbf{I}), \tag{2}$$

which intuitively means we can obtain the noisy sample $\mathbf{x}_r$ by adding a Gaussian perturbation $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathbf{x}_0$ [4]

$$\mathbf{x}_r = \sqrt{\bar{\alpha}_r}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_r}\epsilon. \tag{3}$$

$\mathbf{x}_r$ is almost I.I.D. Gaussian when $\bar{\alpha}_r \approx 0$.

The diffusion model used in this study is based on the ν-objective diffusion model [10] for music [11], with $\sqrt{\bar{\alpha}_r}$ chosen to be $\cos(0.5\pi\sigma_r)$ and $\sqrt{1-\bar{\alpha}_r} = \sin(0.5\pi\sigma_r)$, where $\sigma_r \sim U[0,1]$. During training stage, the ν-objective diffusion model estimates a model for $\nu_{\sigma_r} = \frac{2}{\pi}\frac{\partial \mathbf{x}_r}{\partial \sigma_r} = \cos(0.5\pi\sigma_r)\epsilon - \sin(0.5\pi\sigma_r)\mathbf{x}_0$ by minimizing the loss function [11]:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{r,\sigma_r,\mathbf{x}_{\sigma_r}} \left[\| f_\theta(\mathbf{x}_r, \sigma_r) - \nu_{\sigma_r} \|_2^2 \right], \quad (4)$$

where f_θ is a model to be trained using a 1D U-Net architecture [11]. The denoising diffusion implicit models (DDIM) sampler denoises the noisy signal $\mathbf{x}_r$ to get estimated clean data $\hat{\mathbf{x}}_0$ by repeatedly applying [11], [12]

$$\begin{aligned} \hat{\nu}_{\sigma_r} &= f_{\theta^*}(\mathbf{x}_r, \sigma_r), \\ \hat{\mathbf{x}}_0 &= \cos(0.5\pi\sigma_r)\mathbf{x}_r - \sin(0.5\pi\sigma_r)\hat{\nu}_{\sigma_r}, \\ \hat{\epsilon}_r &= \sin(0.5\pi\sigma_r)\mathbf{x}_r + \cos(0.5\pi\sigma_r)\hat{\nu}_{\sigma_r}, \\ \hat{\mathbf{x}}_{r-1} &= \cos(0.5\pi\sigma_{r-1})\mathbf{x}_r + \sin(0.5\pi\sigma_{r-1})\hat{\epsilon}_r, \end{aligned} \quad (5)$$

to estimate initial data and noise at each step r. In the generation stage, a new data sample can be generated by replacing the $\mathbf{x}_r$ in (5) with a random Gaussian noise. The rest of this section explains how we prepare the training data for the diffusion model.

The UWA channel models considered here are the discrete-time baseband complex-valued time-domain impulse responses. The impulse response from one source to one of the receiver locations z_i with sampling rate F_s is denoted as a matrix $\mathbf{H}_{z_i}$, where $\mathbf{H}_{z_i}[t_j, \tau_k]$ denotes the element at jth row and kth column of $\mathbf{H}_{z_i}$ which represents the impulse response value at time t_j and delay τ_k. The available time range of the training dataset is from $t = 0$ to $\frac{D}{F_s}$, where D is the total number of time samples. The desired time duration captured in the generated impulse responses is $\frac{\Delta}{F_s}$, of maximum delay-spread τ_N (the discrete-time impulse response length is N). Then L impulse responses evenly spaced in time range from $t = t_j$ to $t_j + \frac{\Delta}{F_s}$, for any $j : 0 < t_j < \frac{D-\Delta}{F_s}$ in the training data can be chosen as one training sample. Choosing all M receivers, one vector of length NML can be constructed:

$$\begin{aligned} \mathbf{y}_0 = \Bigg[&\mathbf{H}_{z_0}[t_j, :], ..., \mathbf{H}_{z_{M-1}}[t_j, :], \mathbf{H}_{z_0}\left[t_j + \frac{\Delta}{LF_s}, :\right], \\ &..., \mathbf{H}_{z_{M-1}}\left[t_j + \frac{(L-1)\Delta}{LF_s}, :\right] \Bigg]^{\mathrm{T}}, \end{aligned} \quad (6)$$

where T denotes the transpose operation.

We found that the diffusion model captures the UWA channel responses spatiotemporal correlation better if the model is trained on the Fourier transformed $\mathbf{y}_0$, i.e., $\mathbf{x}_0 = \mathcal{F}\mathbf{y}_0$. We treat the real and imaginary parts of $\mathbf{x}_0$ as two input channels for the diffusion model to yield an overall input shape of $(N_b, 2, NML)$, where N_b is the batch size (the number of different training samples $\mathbf{x}_0$). By applying the Inverse Fourier transform to a generated sample, we can recover the time-domain impulse responses.

III. Results

We used extracted channels derived from the Kauai Acomms MURI 2011 (KAM11) dataset [13], [14] in a repository of standard channel models that is currently under development to train the diffusion model. We used $M = 2$ receivers (the first and seventh hydrophones) to investigate the generative model's multichannel learning capability. All baseband impulse responses have a sampling rate of $F_s = 6250$ Hz and length $N = 64$ ($\tau_N = 10.24$ ms). The time duration captured in the generated impulse responses is set to $\frac{\Delta}{F_s} = 3.2$ s, which approximates the decorrelation time of impulse responses in the training data. Time resolution $\frac{\Delta}{LF_s} = 0.4$ s and $L = 8$. The total measurement time range for the training data used is around $\frac{D}{F_s} = 20$ s.

The diffusion model architecture used in this paper is the "audio-diffusion-pytorch"[1] [5], [10], [11], [15] with no pre-training, and 7 nested U-Net blocks with 32 channels in each block ([32, 32, 32, 32, 32, 32, 32]) for which we downsampled each time by 2 except the first two blocks ([1, 1, 2, 2, 2, 2, 2]). This downsampling process provides a 32 times compression factor that accelerates the training speed. ResNet and modulation items with the repetitions [1, 2, 2, 2, 2, 2, 2] were used and no attention was used to allow the learning of possibly long latent representations [11]. An NVIDIA RTX A6000 was used to train the diffusion model. Only 20 minutes of training (18,000 iterations) was used for the model that generates the results in this section. Each generated impulse response from the trained diffusion model is less than 1 s. Considering the offline nature of the current application, computational complexity is not a major issue in using the current diffusion model for UWA channel response regeneration.

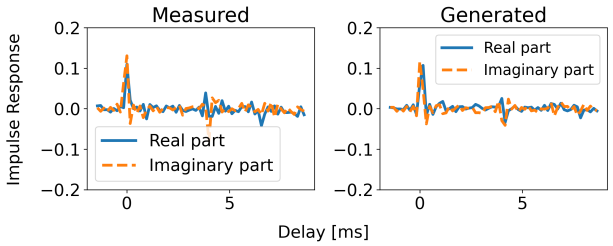

Fig. 1: One of the measured and generated impulse responses.

To illustrate the impulse responses extracted from KAM11 dataset used in this work, Figure 1 shows one realization of the measured and generated impulse responses at a random time instance. Figures 2 and 3 show the magnitude of one realization of the measured and generated UWA channel impulse responses respectively. Each subplot shows the time-varying impulse responses across $\frac{\Delta}{F_s} = 3.2$ s (time resolution $\frac{\Delta}{LF_s} = 0.4$ s). It clearly shows the first two peaks around 0 ms and 4 ms. The arrival time of both peaks changes slightly. Next, we show that the model successfully learns the

[1] https://github.com/archinetai/audio-diffusion-pytorch.git

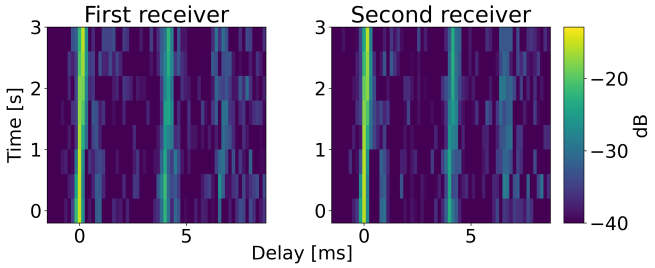

Fig. 2: The magnitude of measured impulse responses.

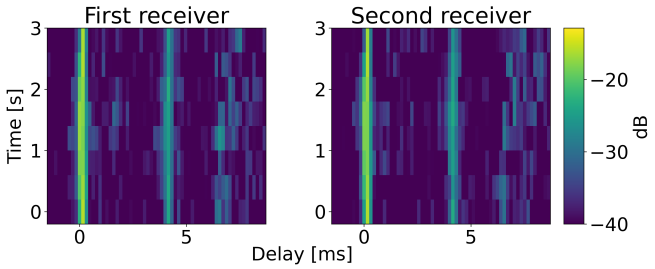

Fig. 3: The magnitude of generated impulse responses.

distribution of the arrival time of peaks and spatiotemporal correlation.

A. Distribution of Peak Arrival Times

One important characteristic of UWA channels is the multipath effects, which impact the communication algorithm design [1], [2]. From Fig. 2, there are two time-varying peaks in the impulse responses. Thus, estimates of the probability mass function (PMF) of the arrival times for each of the two peaks for both measured and generated impulse responses can be compared. The arrival times of peaks in this section are determined by finding the delay with the locally-maximum time-domain magnitude.

Figure 4 shows an estimate of the PMF (histogram) of the arrival time of the first peak τ_{p_1} using the diffusion model trained over 80,000 iterations. Since the diffusion model generation starts from a randomly sampled Gaussian white sequence, τ_{p_1} will be uniformly distributed in the beginning. The PMF of τ_{p_1} gradually converges to that of the measured data with more training iterations. In fact, after only 1800 iterations (around 2 minutes into the training process), the generated impulse responses have a PMF of the arrival time of the first peak τ_{p_1} close to that of the measured data. To measure the similarity numerically, the Kullback-Leibler (KL) divergence is also plotted in Fig. 4.

Figure 5 shows a comparison of the estimated PMF of the arrival time of the first two peaks (τ_{p_1} and τ_{p_2}) in both receivers between the measured and generated impulse responses. The arrival time of the generated peaks in the impulse responses has almost the same PMF compared with the measurement data. The PMF of the difference of the arrival time of the first and second peak (i.e., the PMF of

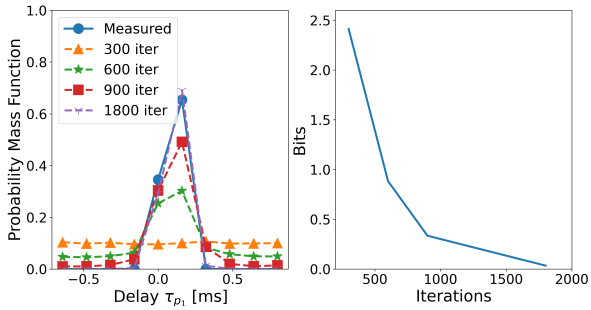

Fig. 4: Left plot shows the PMF of the arrival time for the first peak in the first receiver. Generated responses use models at different training iterations (and training time) are shown. The KL divergence between the generated distribution and the measured distribution is shown on the right.

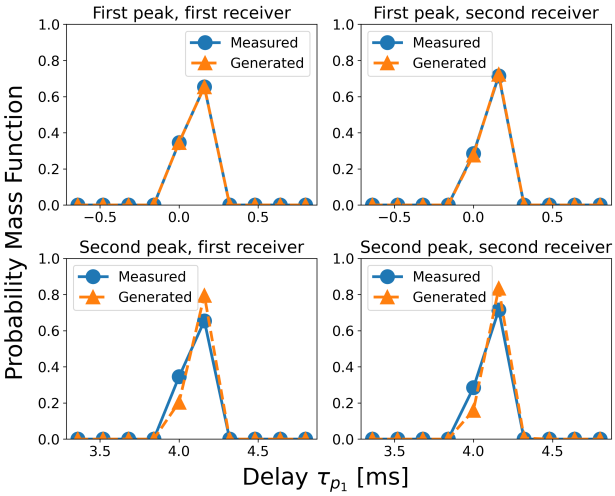

Fig. 5: Comparison of the PMF of arrival times of two peaks in the measured and generated UWA channel impulse responses. The KL divergence between measured and generated distributions are 0.0002 and 0.001 bits for the first row and 0.091 bits for both plots in the second row.

the $\delta\tau_{p_{1,2}} = \tau_{p_2} - \tau_{p_1}$) is shown in Fig. 6. From Fig. 6, it is clear that in the field measurement, although the arrival time of both peaks is slightly changing across time t, the difference $\delta\tau_{1,2}$ remains fairly constant. This illustrates how the trained diffusion model learns the relationship between the different peak arrivals exhibited in the training dataset.

B. Spatiotemporal Correlation

To investigate the similarities between the field measurement and diffusion model generated impulse responses in terms of the spatiotemporal correlation, their complex correlation is computed. The complex correlation between the impulse response of the i-th receiver location and j-th receiver

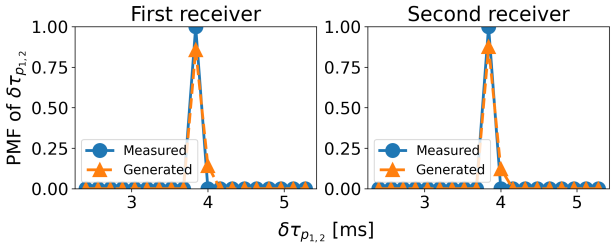

Fig. 6: Comparison of PMF of $\delta\tau_{1,2}$ in the measured and generated underwater acoustic channel impulse responses. The KL divergence between measured and generated distributions are 0.222 and 0.185 bits respectively.

location is denoted as matrices $\mathbf{C}_{z_i,z_j}$, the ℓth row and mth column of the matrix is denoted as $\mathbf{C}_{z_i,z_j}[\eta_\ell, \gamma_m]$ and defined

$$\mathbf{C}_{z_i,z_j}[\eta,\gamma] = \mathbb{E}_t\left[\sum_\tau \mathbf{H}_{z_i}[t,\tau]\mathbf{H}^*_{z_j}[t-\eta,\tau-\gamma]\right], \qquad (7)$$

where * denotes complex conjugate, η denotes time difference, and γ denotes the delay difference. When computing the correlation function using the measured data, 10000 different t_k are randomly selected to estimate the expected correlation value. When computing the correlation function using the generated data, 10000 sets of impulse responses are randomly generated to estimate the expected value. Each set of impulse responses contains 2 receiver locations and 8 impulse responses for each receiver location spanning 3.2 s in time and 10.24 ms in delay which matches the dimension of measured data.

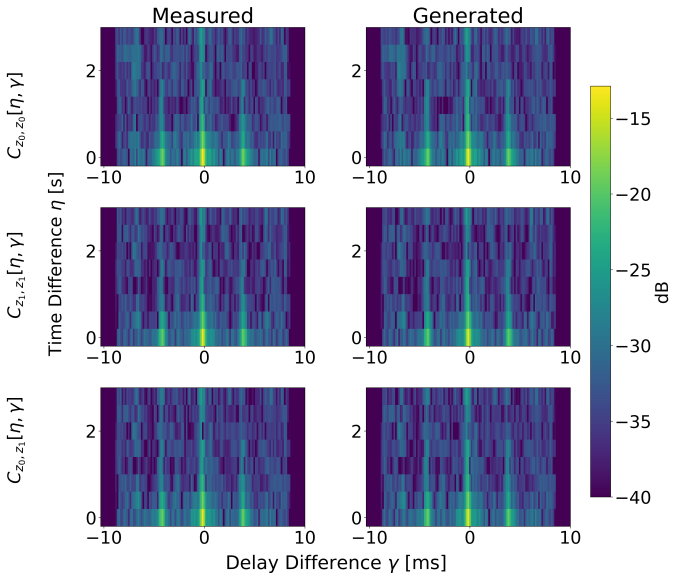

Fig. 7: The comparison of the magnitude of complex correlation functions of the measured and generated impulse responses.

Since we have two receiver locations z_0 and z_1, the comparison of the magnitude of correlation functions $\mathbf{C}_{z_0,z_0}[\eta,\gamma]$, $\mathbf{C}_{z_1,z_1}[\eta,\gamma]$, and $\mathbf{C}_{z_0,z_1}[\eta,\gamma]$ are plotted as the first, second, and third row respectively in Fig. 7, illustrating that the generated responses have a similar decrease in the correlation function magnitude when the time difference η becomes larger. Because of the first two peak arrivals, the correlation magnitude is larger when the delay difference γ is around 0 and ± 4 ms. Comparing the plots in Fig. 7 shows the trained diffusion model learns the spatiotemporal correlation function of the measured data. To better compare the correlation of measured and generated responses, correlations $\mathbf{C}_{z_0,z_0}[\eta,\gamma]$ and $\mathbf{C}_{z_0,z_1}[\eta,\gamma]$ when $\eta = 0, 0.4$, and 3.2 s respectively are shown in Fig. 8. The generated responses have qualitatively similar correlations compared with the measured responses.

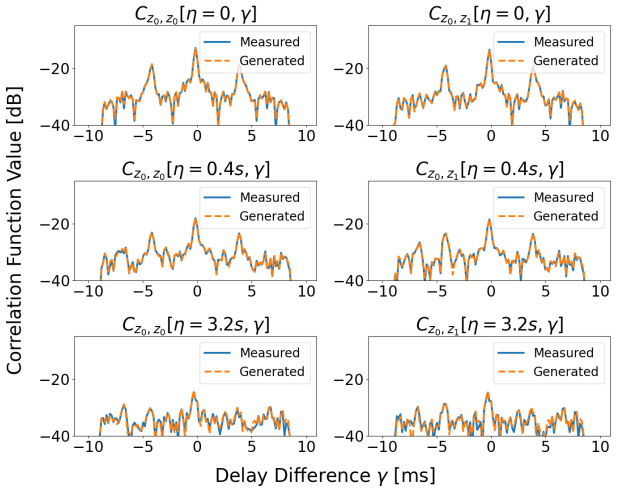

Fig. 8: Comparison of the magnitude of some correlation functions of the measured and generated underwater acoustic channel impulse responses.

IV. Conclusion

In this article, a diffusion model trained on a UWA channel impulse response dataset was demonstrated to generate realistic UWA channel impulse responses. The PMF of the dominant arrival time peaks in generated responses was shown to match the field measurement. The spatiotemporal correlation functions also show a good match between the field measurement and generated responses. The proposed method can be used to augment field measurements for designing underwater communication algorithms. Moreover, diffusion-based generative models can serve as deep priors for tasks such as channel estimation, equalization, or blind deconvolution.

Acknowledgments

This work was supported by the Office of Naval Research under Grant N00014-23-1-2852. The authors would like to thank James Preisig, Mandar Chitre, and Paul van Walree for their helpful discussions.

REFERENCES

[1] M. Stojanovic and J. Preisig, "Underwater acoustic communication channels: Propagation models and statistical characterization," *IEEE communications magazine*, vol. 47, no. 1, pp. 84–89, 2009.

[2] A. C. Singer, J. K. Nelson, and S. S. Kozat, "Signal processing for underwater acoustic communications," *IEEE Communications Magazine*, vol. 47, no. 1, pp. 90–96, 2009.

[3] S. Yang, G. B. Deane, J. C. Preisig, N. C. Sevüktekin, J. W. Choi, and A. C. Singer, "On the reusability of postexperimental field data for underwater acoustic communications r&d," *IEEE Journal of Oceanic Engineering*, vol. 44, no. 4, pp. 912–931, 2019.

[4] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851.

[6] T. Jayashankar, G. C. Lee, A. Lancho, A. Weiss, Y. Polyanskiy, and G. Wornell, "Score-based source separation with applications to digital communication signals," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[7] S. Li, L. Cheng, J. Li, Z. Wang, and J. Li, "Learning data distribution of three-dimensional ocean sound speed fields via diffusion models," *The Journal of the Acoustical Society of America*, vol. 155, no. 5, pp. 3410–3425, 2024.

[8] G. Harshvardhan, M. K. Gourisaria, M. Pandey, and S. S. Rautaray, "A comprehensive survey and analysis of generative models in machine learning," *Computer Science Review*, vol. 38, p. 100285, 2020.

[9] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[10] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations*, 2021.

[11] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," *arXiv preprint arXiv:2301.11757*, 2023.

[12] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2020.

[13] W. Hodgkiss and J. Preisig, "Kauai acomms MURI 2011 (KAM11) experiment," in *Proc. Eur. Conf. Underwater Acoust*, 2012, pp. 993–1000.

[14] P. Qarabaqi and M. Stojanovic, "Statistical characterization and computationally efficient modeling of a class of underwater acoustic communication channels," *IEEE Journal of Oceanic Engineering*, vol. 38, no. 4, pp. 701–717, 2013.

[15] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022.