







# The Effect of Training Dataset Size on Discriminative and Diffusion-Based Speech Enhancement Systems

Philippe Gonzalez , Associate Member, IEEE, Zheng-Hua Tan , Senior Member, IEEE, Jan Østergaard , Senior Member, IEEE, Jesper Jensen , Tommy Sonne Alstrøm , Member, IEEE, and Tobias May 

**Abstract**—The performance of deep neural network-based speech enhancement systems typically increases with the training dataset size. However, studies that investigated the effect of training dataset size on speech enhancement performance did not consider recent approaches, such as diffusion-based generative models. Diffusion models are typically trained with massive datasets for image generation tasks, but whether this is also required for speech enhancement is unknown. Moreover, studies that investigated the effect of training dataset size did not control for the data diversity. It is thus unclear whether the performance improvement was due to the increased dataset size or diversity. Therefore, we systematically investigate the effect of training dataset size on the performance of popular state-of-the-art discriminative and diffusion-based speech enhancement systems in matched conditions. We control for the data diversity by using a fixed set of speech utterances, noise segments and binaural room impulse responses to generate datasets of different sizes. We find that the diffusion-based systems perform the best relative to the discriminative systems in terms of objective metrics with datasets of 10 h or less. However, their objective metrics performance does not improve when increasing the training dataset size as much as the discriminative systems, and they are outperformed by the discriminative systems with datasets of 100 h or more.

**Index Terms**—Speech enhancement, training data, discriminative models, diffusion models.

## I. INTRODUCTION

UNDERSTANDING speech in noisy and reverberant environments can be challenging for both normal-hearing and hearing-impaired listeners [1], [2]. Therefore, speech enhancement, which aims to improve the intelligibility and quality of speech signals corrupted by noise and reverberation, is an integral part of many technical applications, such as hearing aids and communication systems. The majority of newly-proposed speech enhancement systems are based on deep neural networks

(DNNs) due to their superior performance over traditional approaches [3], [4]. These systems are commonly trained in a supervised manner with a large number of noisy and clean speech signals. As the number of trainable parameters increases, DNNs have the potential to capture more details in the probability distribution of the training data, but this requires training them with larger datasets. For example, diffusion models [5], [6], [7], which have been recently applied to speech enhancement [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], are typically trained with huge datasets in image generation literature [21]. However, whether this is also required for speech enhancement is unknown.

Few studies have investigated the effect of training dataset size on the performance of state-of-the-art speech enhancement systems in a systematic way. In [22], a feedforward neural network (FFNN) was trained with an increasing number of noises and utterances, which improved the classification performance of individual time-frequency units into speech and noise. However, the dataset size increased with the number of noises and utterances used to generate the mixtures. As a consequence, the effects of the size and diversity of the training data were entangled. In [3], a FFNN was trained with datasets of different sizes generated from a fixed number of noises and utterances, i.e. fixed diversity. Speech quality results increased with the training dataset size, but since the utterances were selected from TIMIT [23], whose training split is only 4 h-long, the performance saturated for datasets larger than 100 h due to the increased redundancy in the training data. In [24], a FFNN was trained with a fixed number of mixtures, i.e. fixed dataset size, using either 100 or 10000 noises. The system strongly benefited from the increased number of noises and matched the performance of a noise-specific system, but the effect of training dataset size was not investigated. In [25], a bidirectional gated recurrent unit (GRU) network was trained with datasets of different sizes generated using an increasing range of signal-to-noise ratios (SNRs), and the performance of a downstream speech recognition system improved consequently. However, by increasing the range of SNRs seen during training, the acoustic mismatch between training and testing was reduced. The performance improvement was thus attributed to the reduced mismatch rather than the increased dataset size. In [26], a FFNN and a recurrent long short-term memory (LSTM) network were trained with 100 h-long datasets generated from different noise databases. The systems performed the best when trained with datasets generated from large and diverse noise databases. While the study differentiated between the size and diversity of the data, it referred to the size of the noise databases used to generate the mixtures, rather than the amount of mixtures

Received 24 June 2024; revised 9 August 2024; accepted 13 August 2024. Date of publication 23 August 2024; date of current version 4 September 2024. The associate editor coordinating the review of this article and approving it for publication was Dr. Jeremy Thomas Reed. (Corresponding author: Philippe Gonzalez.)

Philippe Gonzalez and Tobias May are with the Department of Health Technology, Technical University of Denmark, 2800 Lyngby, Denmark (e-mail: phigon@dtu.dk; tobmay@dtu.dk).

Zheng-Hua Tan, Jan Østergaard, and Jesper Jensen are with the Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark (e-mail: zt@es.aau.dk; jo@es.aau.dk; jje@es.aau.dk).

Tommy Sonne Alstrøm is with the Department of Applied Mathematics and Computer Science, Technical University of Denmark, 2800 Lyngby, Denmark (e-mail: tsal@dtu.dk).

Digital Object Identifier 10.1109/LSP.2024.3449221

generated to train the systems, which was kept constant. In summary, most studies have not investigated the effect of the training dataset size independently of its diversity, and have considered outdated FFNNs or recurrent neural networks.

In this study, we systematically investigate the effect of training dataset size on the performance of popular state-of-the-art speech enhancement systems in matched conditions. We consider three discriminative systems, namely Conv-TasNet [27], DCCRN [28] and MANNER [29], and three diffusion-based approaches, namely SGMSE+ [11], SGMSE+M [30] and the system from [17], [18]. To control for the data diversity, we generate mixture datasets of different sizes using a fixed set of speech utterances, noise segments and binaural room impulse responses (BRIRs). This way, the training distribution is fixed, and only the number of training examples is changed. Code and audio examples are available online.<sup>1</sup>

## II. SIGNAL MODEL

Let  $s$  denote a clean speech signal and  $\{n_i\}_{i=1}^N$  a set of noise signals where  $N$  is the number of noise sources in the acoustic scene. The mixture  $x_L$  at the left ear of a binaural receiver in the acoustic scene can be expressed as follows,

$$x_L = s * h_{s,L} + \sum_{i=1}^N n_i * h_{n_i,L}, \quad (1)$$

where  $h_{s,L}$  is the left channel of the BRIR between the receiver and the speech source, and  $h_{n_i,L}$  is the left channel of the BRIR between the receiver and the  $i$ -th noise source. Defining the target signal for the speech enhancement system requires choosing a reflection boundary  $b$  beyond which speech reflections are considered detrimental to speech intelligibility [31], [32]. Let  $\mathbb{1}_{[a,b]}$  denote the indicator function of a time interval  $[a, b]$ . By denoting  $h_{s,L}^{\text{early}} = h_{s,L} \cdot \mathbb{1}_{[0,b]}$  the part of  $h_{s,L}$  up to time instant  $b$  and  $h_{s,L}^{\text{late}} = h_{s,L} \cdot \mathbb{1}_{[b,\infty]}$  the part of  $h_{s,L}$  beyond time instant  $b$ , the target signal  $y_L$  for the speech enhancement system at the left ear is

$$y_L = s * h_{s,L}^{\text{early}}, \quad (2)$$

while the interfering signal  $n_L$  at the left ear includes the late speech reflections and the reverberant noise,

$$n_L = s * h_{s,L}^{\text{late}} + \sum_{i=1}^N n_i * h_{n_i,L}. \quad (3)$$

Note that since  $h_{s,L} = h_{s,L}^{\text{early}} + h_{s,L}^{\text{late}}$ , we have  $x_L = y_L + n_L$  due to the linearity of the convolution operator. Similar expressions are obtained for the right ear signals  $x_R$ ,  $y_R$  and  $n_R$ .

As the speech enhancement systems considered in this study are single-channel, we average the left and right channels of the different signals. We denote the single-channel mixture, target and interfering signals as  $x = \frac{x_L + x_R}{2}$ ,  $y = \frac{y_L + y_R}{2}$  and  $n = \frac{n_L + n_R}{2}$  respectively. Even though the channels are averaged, a binaural model allows to simulate realistic mixtures where the position of the sources contributes to the acoustic diversity.

<sup>1</sup>[Online]. Available: <https://github.com/philgzl/brevert> and <https://philgzl.github.io/lst>

TABLE I  
CORPORA AND DATABASES DETAILS

Corpus	Speakers	Utterances	Hours	Avg. len.	Min. len.	Max. len.
TIMIT	630	6 300	5.4	3.1 s	0.9 s	7.8 s
Libri.	251	28 539	100.6	12.7 s	1.4 s	24.5 s
WSJ	131	34 738	69.5	7.2 s	0.9 s	44.8 s
Clarity	40	11 352	8.9	2.8 s	1.2 s	7.7 s
VCTK	110	44 455	41.6	3.4 s	1.2 s	16.6 s

(a) Speech corpora

Database	Types	Hours	Database	Rooms	BRIRs
TAU	10	40.0	Surrey	4	148
NOISEX	15	1.0	ASH	35	538
ICRA	10	1.1	BRAS	4	180
DEMAND	18	1.5	CATT	11	407
ARTE	13	0.5	AVIL	4	96

(b) Noise databases

(c) BRIR databases

## III. DATASET GENERATION

We generate noisy and reverberant mixtures at 16 kHz using speech utterances, noise segments and BRIRs from multiple corpora and databases. The speech utterances are selected from TIMIT [23], LibriSpeech (100-hour version) [33], WSJ SI-84 [34], Clarity [35] and VCTK [36]. The noises are selected from TAU [37], NOISEX [38], ICRA [39], DEMAND [40] and ARTE [41]. The BRIRs are selected from Surrey [42], ASH [43], BRAS [44], CATT [45] and AVIL [46]. Details about each corpus and database are provided in Table I. Each acoustic scene is simulated by placing one speech source and  $N \in \{1, 2, 3\}$  noise sources at random spatial locations in the same room, uniformly distributed in  $[-90^\circ, 90^\circ]$  in front of the receiver. The reflection boundary is set to  $b = 50$  ms to include early reflections in the target signal  $y$  as suggested in [31], [32]. The target and interfering signals  $y$  and  $n$  are mixed at a random SNR uniformly distributed in  $[-5$  dB, 10 dB].

To generate a training dataset, we randomly select speech utterances, noise segments and BRIRs from a subset of each speech corpus, noise database and BRIR database respectively. For the speech, this subset is constructed by randomly selecting 80 % of the utterances from each corpus. For the noise, we sample segments within 80 % of the length of each file. For the BRIRs, we select every other BRIR in each room. The validation dataset uses the same subset of speech utterances, noise segments and BRIRs as the training dataset, but consists of different random mixture realizations. The test dataset is generated using the remaining set of speech utterances, noise segments and BRIRs. Note that utterances from the same speaker, segments from the same noise file and BRIRs from the same room can be used for training and testing. This is deliberate, as we are interested in evaluating the system in matched conditions, such that distribution shifts between training and testing are minimized and only the effect of the number of training examples is captured. We generate five training datasets with sizes of 3 h, 10 h, 30 h, 100 h and 300 h respectively. The number of mixtures in each dataset is 2595, 8660, 25930, 86074 and 258259 respectively. The validation and test datasets are both fixed to 30 min. They consist of 305 and 310 mixtures respectively.

When randomly selecting a corpus from which to pick a utterance to form a mixture, one option is to use equal probabilities for each corpus. While this would result in a similar number of

TABLE II  
PERCENTAGE OF MIXTURES USING REPEATED SPEECH UTTERANCES IN EACH TRAINING DATASET IN TERMS OF DURATION

Dataset	TIMIT	WSJ	Clarity	Libri.	VCTK	Total
3 h	2%	1%	2%	0%	0%	5%
10 h	8%	1%	5%	0%	1%	15%
30 h	15%	3%	12%	1%	3%	34%
100 h	20%	9%	19%	5%	9%	61%
300 h	20%	15%	20%	11%	17%	83%

utterances from each corpus, it would result in a large imbalance in terms of duration, since the five considered speech corpora have very different utterance length distributions as shown in Table I. To avoid this, we weight the probability of selecting a corpus by the inverse of its average utterance length. This is only done for the training dataset, as the systems are evaluated on the validation and test datasets on a per-mixture basis. This is not an issue when selecting noise segments and BRIRs, since the mixture length is defined by the speech utterance length. We thus use equal probabilities for selecting the noise and BRIR databases.

The speech utterances, noise segments and BRIRs are randomly drawn with replacement. This means that the same speech utterance can be used to generate multiple mixtures in the same dataset, and the number of repetitions increases with the dataset size. Moreover, as the number of utterances in each corpus varies substantially, utterances are more or less likely to be repeated depending on which corpus they are selected from. Note that the mixtures are still unique, as the chances of selecting the same speech utterance, noise segment, BRIR, SNR and spatial locations twice are very low. Table II shows the percentage of mixtures using repeated speech utterances in each training dataset, and the corpus they are selected from. A similar analysis can be made for the noise segments and the BRIRs, but is not shown here for brevity.

#### IV. SYSTEMS, TRAINING AND OBJECTIVE METRICS

We evaluate three different discriminative speech enhancement systems, namely Conv-TasNet [27], DCCRN [28] and MANNER [29]. They have 4.9 M, 3.7 M and 21.2 M parameters respectively. We also evaluate three diffusion-based systems, namely SGMSE+ [11], SGMSE+M [30] and the system from [17], [18], which we denote as SGMSE+M<sub>Heun</sub><sup>cos</sup>. These three systems all use the NCSN++ architecture from [6] for the score network. SGMSE+M and SGMSE+M<sub>Heun</sub><sup>cos</sup> use a smaller version of NCSN++ (27.8 M parameters) compared to SGMSE+ (66.1 M parameters), as this was reported to reduce the computational cost without degrading performance [30]. Compared to SGMSE+M, SGMSE+M<sub>Heun</sub><sup>cos</sup> uses a cosine noise schedule, a Heun-based sampler and a different preconditioning [17], [18]. The number of sampling steps is fixed to 64. Technical details about the implementation of both discriminative and diffusion-based systems can be found in [17].

The systems are trained with the different datasets for the same number of neural network parameter updates. That is, as the dataset size increases, the number of training epochs is proportionally reduced. This allows for a fair comparison, since the systems are trained for the same amount of time. The number of epochs is set to 1000 for the 3 h dataset, 300 for the 10 h dataset, 100 for the 30 h dataset, 30 for the 100 h dataset and

10 for the 300 h dataset. The experiment is repeated three times with different random neural network parameter initializations, and the metrics are averaged across repetitions.

The systems are evaluated in terms of perceptual evaluation of speech quality (PESQ) [47], extended short-term objective intelligibility (ESTOI) [48] and SNR. The results are reported in terms of average objective metric improvement from the input mixture to the enhanced output. The improvements are denoted as  $\Delta$ PESQ,  $\Delta$ ESTOI and  $\Delta$ SNR respectively.

#### V. RESULTS

Fig. 1 shows the performance on the test dataset as training progresses for each system and each dataset size. It can be seen that the discriminative systems, i.e. Conv-TasNet, DCCRN and MANNER, strongly benefit from increasing the training dataset size, despite the increasing amount of utterance repetitions reported in Table II. More specifically, these systems heavily overfit when training with the 3 h dataset for too many epochs. As the training dataset size increases, they stop overfitting and benefit from longer training. Performance improvements with the dataset size can be observed until 300 h. Meanwhile, the diffusion-based systems, i.e. SGMSE+, SGMSE+M and SGMSE+M<sub>Heun</sub><sup>cos</sup>, show a very different behavior. With the 3 h dataset, they overfit less severely and show substantially higher results compared to the discriminative systems. While the performance of SGMSE+M<sub>Heun</sub><sup>cos</sup> improves when increasing the dataset size to 10 h, they show very similar curves when further increasing the dataset size beyond 10 h. Despite having more parameters, SGMSE+ shows similar performance to SGMSE+M, which is in line with [30].

The best model for each training run is selected based on the validation loss. The performance of the selected model on the test dataset is plotted as a function of the training dataset size in Fig. 2. Similar to the previous results, the performance of the discriminative systems substantially improves with the dataset size. However, while the diffusion-based systems outperform the discriminative systems with the 3 h and 10 h datasets, they do not improve when further increasing the dataset size. With the 100 h and 300 h datasets, they are outperformed by Conv-TasNet and MANNER in terms of  $\Delta$ PESQ, and by Conv-TasNet and DCCRN in terms of  $\Delta$ SNR. Since SGMSE+ shows the same behavior as SGMSE+M, these results cannot be explained by a lack of model capacity.

#### VI. DISCUSSION

The results suggest that diffusion-based speech enhancement systems are remarkably suited when a small amount of training data is available, as SGMSE+, SGMSE+M and SGMSE+M<sub>Heun</sub><sup>cos</sup> perform the best relative to the discriminative systems when trained with the 3 h and 10 h datasets. Our results are in contrast to image generation literature, where diffusion models are typically trained with datasets of billions of images [21]. We hypothesize on possible explanations for our results:

- The speech enhancement task is very different from the image generation task. In image generation, the model is tasked with generating coherent images from scratch given a text prompt, and multiple valid yet very different images can be generated from the same prompt. In speech enhancement, the model is provided with a mixture that has the same modality and dimensionality as the clean speech, and can thus leverage a lot of information from the

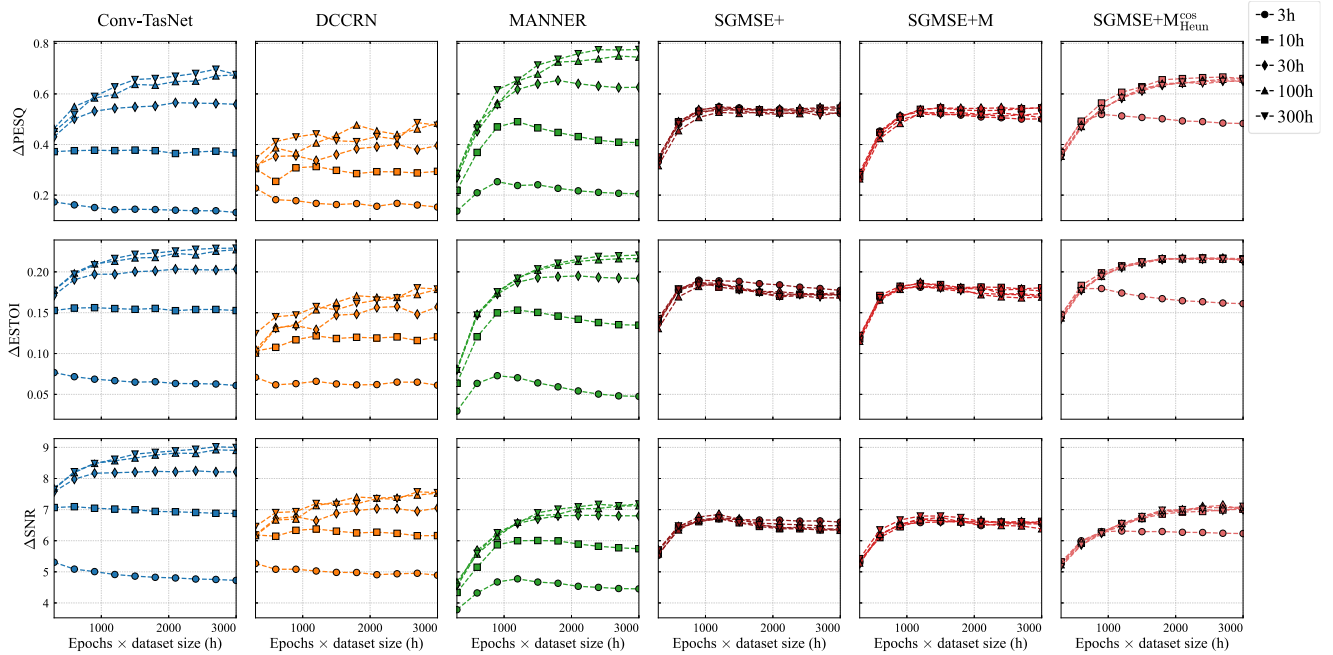


Fig. 1.  $\Delta$ PESQ,  $\Delta$ ESTOI and  $\Delta$ SNR results on the test dataset as a function of the number of epochs  $\times$  the training dataset size in hours.

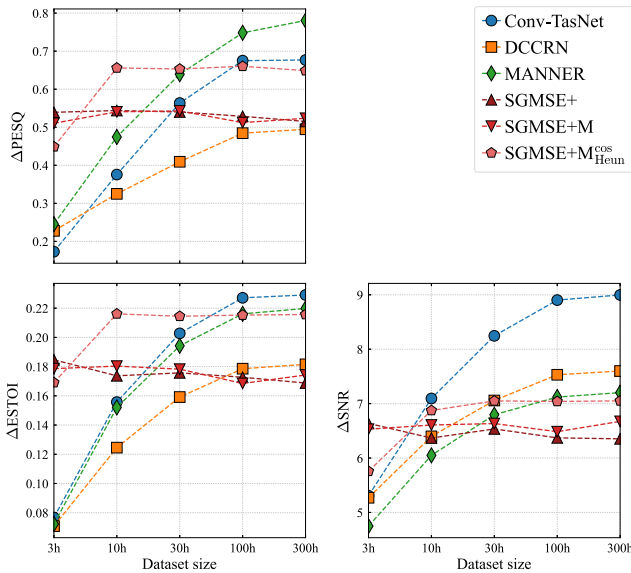


Fig. 2.  $\Delta$ PESQ,  $\Delta$ ESTOI and  $\Delta$ SNR results as a function of the training dataset size.

input to generate the output. While multiple versions of the clean speech can be generated from the same mixture, these versions should not be very different from each other, as the output speech should be coherent with the mixture.

- The stochastic nature of the diffusion process acts as a strong regularizer, which allows the model to perform well despite being trained with a small amount of data. Indeed, the neural network is presented with training examples mixed with random Gaussian noise realizations at different levels during training. This is in line with [49], which showed that the diffusion model objective is equivalent to the evidence lower bound (ELBO) with data augmentation

consisting of Gaussian noise perturbation. This also explains the similar curves in Fig. 1 for the 10 h, 30 h, 100 h and 300 h datasets.

- While SGMSE+M<sub>Heun</sub><sup>cos</sup> shows superior performance compared to SGMSE+M thanks to the updated noise schedule, sampler and preconditioning, there might be other design aspects that prevent the system from scaling. E.g. the NCSN++ architecture, which was borrowed from image generation literature, might not be optimal for speech processing.

## VII. CONCLUSION

We investigated the effect of training dataset size on the performance of three discriminative and three diffusion-based speech enhancement systems in matched conditions. We found that the diffusion-based systems performed the best relative to the discriminative systems in terms of objective metrics with datasets of 10 h or less, but they were outperformed by the discriminative systems with datasets of 100 h or more. This suggests that diffusion-based approaches are remarkably suited when a small amount of training data is available. However, this also suggests that they do not benefit from increasing the training dataset size as much as discriminative systems. Future work should investigate if the conclusions generalize to unseen speaker, noise and room conditions. In addition, a formal listening test should be conducted to investigate if the reported differences in terms of objective metrics are perceptually relevant.

## REFERENCES

- [1] A. K. Nabelek and J. M. Pickett, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, vol. 17, pp. 724–739, 1974.

- [2] A. K. Nabelek and D. Mason, "Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms," *J. Speech Lang. Hear. Res.*, vol. 24, pp. 375–383, 1981.
- [3] Y. Xu, J. Du, L. R. Dai, and C. H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
- [4] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, Oct. 2018.
- [5] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.
- [6] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [7] J. Ho, A. Jain, and P. Abbeel, "Denosing diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.
- [8] Y.-J. Lu, Y. Tsao, and S. Watanabe, "A study on speech enhancement based on diffusion probabilistic model," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, 2021, pp. 659–666.
- [9] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7402–7406.
- [10] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," in *Proc. INTERSPEECH*, 2022, pp. 2928–2932.
- [11] J. Richter, S. Welker, J.-M. Lemerrier, B. Lay, and T. Gerkmann, "Speech enhancement and dereverberation with diffusion-based generative models," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2351–2364, 2023.
- [12] H. Yen, F. G. Germain, G. Wichern, and J. L. Roux, "Cold diffusion for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
- [13] C. Chen, Y. Hu, W. Weng, and E. S. Chng, "Metric-oriented speech enhancement using diffusion probabilistic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
- [14] Z. Qiu, M. Fu, Y. Yu, L. Yin, F. Sun, and H. Huang, "SRTNET: Time domain speech enhancement via stochastic refinement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
- [15] W. Tai, Y. Lei, F. Zhou, G. Trajcevski, and T. Zhong, "DOSE: Diffusion dropout with adaptive prior for speech enhancement," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 40272–40293.
- [16] H. Shi et al., "Diffusion-based speech enhancement with joint generative and predictive decoders," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1–5.
- [17] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Investigating the design space of diffusion models for speech enhancement," 2023, *arXiv:2312.04370*.
- [18] P. Gonzalez, Z.-H. Tan, J. Østergaard, J. Jensen, T. S. Alstrøm, and T. May, "Diffusion-based speech enhancement in matched and mismatched conditions using a Heun-based sampler," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1–5.
- [19] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised speech enhancement with diffusion-based generative models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 12481–12485.
- [20] J.-E. Ayilo, M. Sadeghi, and R. Serizel, "Diffusion-based speech enhancement with a weighted generative-supervised learning loss," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2024, pp. 1–5.
- [21] Z. Wang et al., "Patch Diffusion: Faster and more data-efficient training of diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 72137–72154.
- [22] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
- [23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM, NIST speech disc 1-1.1," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, 1993.
- [24] J. Chen, Y. Wang, S. E. Yoho, D. Wang, and E. W. Healy, "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," *J. Acoust. Soc. Amer.*, vol. 139, pp. 2604–2612, 2016.
- [25] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "A cross-entropy-guided measure (CEGM) for assessing speech recognition performance and optimizing DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 106–117, 2021.
- [26] R. Rehr and T. Gerkmann, "SNR-based features and diverse training data for robust DNN-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1937–1949, 2021.
- [27] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.
- [28] Y. Hu et al., "DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Proc. INTERSPEECH*, 2020, pp. 2472–2476.
- [29] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "MANNER: Multi-view attention network for noise erasure," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 7842–7846.
- [30] J.-M. Lemerrier, J. Richter, S. Welker, and T. Gerkmann, "Analysing diffusion-based generative approaches versus discriminative approaches for speech restoration," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2023, pp. 1–5.
- [31] J. S. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *J. Acoust. Soc. Amer.*, vol. 113, pp. 3233–3244, 2003.
- [32] N. Roman and J. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold," *J. Acoust. Soc. Amer.*, vol. 133, pp. 1707–1717, 2013.
- [33] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 5206–5210.
- [34] D. B. Paul and J. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362.
- [35] S. Graetzer et al., "Dataset of British English speech recordings for psychoacoustics and speech processing research: The clarity speech corpus," *Data Brief*, vol. 41, 2022, Art. no. 107951.
- [36] C. Veaux, J. Yamagishi, and S. King, "The Voice Bank corpus: Design, collection and data analysis of a large regional accent speech database," in *Proc. Int. Conf. Oriental COCODA Held Jointly With 2013 Conf. Asian Spoken Lang. Res. Eval.*, 2013, pp. 1–4.
- [37] T. Heittola, A. Mesaros, and T. Virtanen, "TAU urban acoustic scenes 2019, development dataset," *Zenodo*, 2019, doi: [10.5281/zenodo.2589280](https://doi.org/10.5281/zenodo.2589280).
- [38] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.
- [39] W. A. Dreschler, H. Verschuere, C. Ludvigsen, and S. Westermann, "ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment," *Audiology*, vol. 40, pp. 148–157, 2001.
- [40] J. Thiemeann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," in *Proc. Meetings Acoust.*, 2013, Art. no. 035081.
- [41] A. Weisser et al., "The ambisonic recordings of typical environments (ARTE) database," *Acta Acust. United Acust.*, vol. 105, pp. 695–713, 2019.
- [42] C. Hummersone, R. Mason, and T. Brookes, "Dynamic precedence effect modeling for source separation in reverberant environments," *IEEE Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1867–1871, Sep. 2010.
- [43] S. Pearce, "Audio spatialisation for headphones—impulse response dataset," *Zenodo*, 2021, doi: [10.5281/zenodo.4780815](https://doi.org/10.5281/zenodo.4780815).
- [44] F. Brinkmann, L. Aspöck, D. Ackermann, R. Opdam, M. Vorländer, and S. Weinzierl, "A benchmark for room acoustical simulation. Concept and database," *Appl. Acoust.*, vol. 176, 2021, Art. no. 107867.
- [45] "Simulated room impulse responses," Institute of Sound Recording, University of Surrey. [Online]. Available: [http://iosr.surrey.ac.uk/software/index.php#CATT\\_RIRs](http://iosr.surrey.ac.uk/software/index.php#CATT_RIRs)
- [46] M. Marschall and A. Ahrens, "Database of simulated impulse responses and a scene constructor framework for the Audiovisual Immersion Lab (AVIL) at DTU," *Zenodo*, 2017, doi: [10.5281/zenodo.10834971](https://doi.org/10.5281/zenodo.10834971).
- [47] Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, Int. Telecommun. Union Rec. ITU-T P.862, 2001.
- [48] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.
- [49] D. P. Kingma and R. Gao, "Understanding diffusion objectives as the ELBO with simple data augmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 65484–65516.