Generative Data Augmentation Challenge: Zero-Shot Speech Synthesis for Personalized Speech Enhancement

Jae-Sung Bae[1,7], Anastasia Kuznetsova[2], Dinesh Manocha[3], John Hershey[4], Trausti Kristjansson[5,6], and Minje Kim[1,5,8]
[1]*University of Illinois Urbana-Champaign*, [2]*Indiana University*, [3]*University of Maryland*,
[4]*Google Research*, [5]*Amazon Lab126*, [6]*Reykjavik University*
[7]jb82@illinois.edu, [8]minje@illinois.edu

***Abstract*—This paper presents a new challenge that calls for zero-shot text-to-speech (TTS) systems to augment speech data for the downstream task, personalized speech enhancement (PSE), as part of the Generative Data Augmentation workshop at ICASSP 2025. Collecting high-quality personalized data is challenging due to privacy concerns and technical difficulties in recording audio from the test scene. To address these issues, synthetic data generation using generative models has gained significant attention. In this challenge, participants are tasked first with building zero-shot TTS systems to augment personalized data. Subsequently, PSE systems are asked to be trained with this augmented personalized dataset. Through this challenge, we aim to investigate how the quality of augmented data generated by zero-shot TTS models affects PSE model performance. We also provide baseline experiments using open-source zero-shot TTS models to encourage participation and benchmark advancements. Our baseline code implementation and checkpoints are available online[1].**

***Index Terms*—Zero-shot speech synthesis, personalized speech enhancement, generative data augmentation**

I. Introduction

In this paper, we introduce a new challenge that accompanies the Generative Data Augmentation workshop at ICASSP 2025. The main research problem the challenge addresses is the typical data shortage issues when a machine-learning model is developed for a particular target user so that the model respects and exploits individuals' specificity. This type of problem contrasts the remarkable advancements driven by deep neural networks (DNNs) in speech technology, including automatic speech recognition (ASR) [1], text-to-speech (TTS) [2]–[5], and speech enhancement (SE) [6], [7]. These models are typically large in size and trained on extensive datasets, designed to perform robustly across diverse inputs; we refer to such systems as *generalist* systems. However, due to their large size, these systems are challenging to run directly on user devices, necessitating the transfer of user data to servers, which can raise privacy concerns. To address this, smaller *personalized* models have recently been proposed [8], [9]. These models, being compact, can operate on user devices and achieve high performance by focusing on individual users. In theory, personalized models can be created by fine-tuning pre-trained models using individual-specific data. However, gathering enough personalized data for each individual remains a significant challenge due to privacy concerns and technical difficulties in recording clean voices at test time.

Data augmentation is an essential technique to improve the performance of many DNN models. In image processing, a variety of data augmentation techniques are employed, ranging from simple methods like rotating or flipping images [10] to generating new images using other state-of-the-art image generation models [11]. Similarly, numerous data augmentation techniques are also being explored in the field of speech [8], [9], [12], [13]. SpecAugment [12] is one of the most commonly used data augmentation methods in speech self-supervised learning and automatic speech recognition models. Recently, with advancements in TTS models enabling near-human-level speech generation, approaches using TTS for data augmentation have been actively attempted [8], [9], [14], [15].

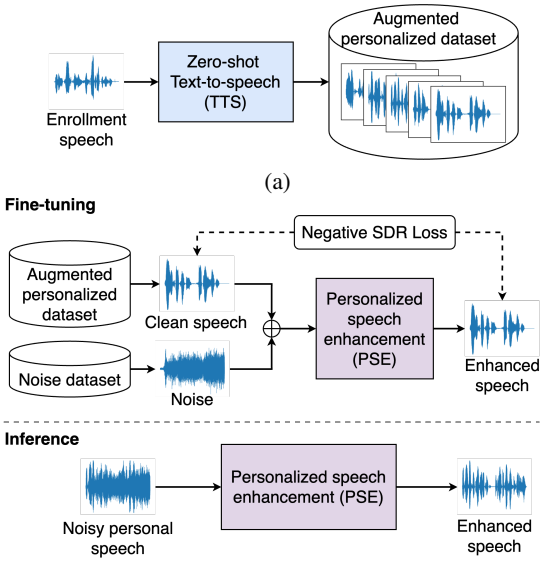

Fig. 1. Overall flow of our challenge. (a) First, participants augment the personalized data using a zero-shot TTS model. (b) Next, organizers ask participants to fine-tune the PSE model with this augmented personalized data, followed by an inference phase in which enhanced speech is generated for evaluation.

Furthermore, the performance of zero-shot TTS systems [3]–[5], which can generate speech that mimics a target speaker's characteristics from just a short speech signal, has significantly improved. These systems, with the ability to generate an unlimited number of speech samples that reflect a target speaker's characteristics from a single speech signal, are gaining attention as a data augmentation approach to address the data shortage problem of personalized models [9], [16]. Although prior research has demonstrated this potential, studies exploring the relationship between various zero-shot TTS models' performance and downstream tasks remain limited.

To conduct a more comprehensive investigation into how zero-shot TTS systems can benefit downstream personalized systems, we propose the **Zero-shot Speech Synthesis Challenge for Personalized Speech Enhancement (PSE)**. We hypothesize that the higher the quality of augmented speech samples generated by zero-shot TTS systems, the better the performance of downstream tasks fine-tuned

This material is in part based on work supported by the National Science Foundation under grant numbers, 2512987 and 1910940.

The authors appreciate Apoorv Vyas at Meta for his help in establishing the virtual speaker dataset.

[1]https://sites.google.com/view/genda2025

using those samples. However, we also believe different performance evaluation factors of augmented speech, such as the precise perceptual similarity between synthesized and original speakers, perceived naturalness of the speech, and the quality of speech signals may have different implications in different downstream tasks. With the proposed challenge, as illustrated in Fig. 1, we aim to highlight the potential difference between the method for evaluating TTS results and their usefulness in the particular downstream task, personalized speech enhancement. In the first phase, the organizers provide a set of enrollment speech samples for test speakers, and participants are tasked with generating speech samples with target speaker characteristics using their zero-shot TTS systems. The generated speech is then evaluated based on speaker similarity, perceptual quality, and intelligibility. In the second phase, participants are asked to train PSE systems using the augmented personalized speech dataset, with performance evaluated on enhanced speech quality.

We conduct baseline experiments for this challenge, providing a detailed description of the experiments and results in this paper. We open-source our baseline model, offering participants a reference for their experiments.

II. Challenge Description

There are two main technical components in this challenge: the generative models that the participants develop to synthesize personalized speech utterances (i.e., the zero-shot TTS systems) and the downstream PSE task where the synthesized personal speech signals are used as the training target of the PSE systems. The challenge organizers will evaluate the submissions mainly based on their usefulness on the PSE tasks, while some basic speech quality estimation will also be conducted for a comprehensive evaluation of the TTS systems.

A. Zero-Shot TTS Models

Participants are required to develop a zero-shot TTS system, which is supposed to synthesize new utterances from a short enrollment signal of the target speaker. The organizers will provide from 3 to 14 second-long enrollment signal per target speaker, who is one of ten randomly chosen speakers from the LibriTTS test-clean dataset [17] (Sec. III-A) or another ten virtual speakers we create for the challenge (Sec. III-B). Additionally, the organizers provide 50 text sentences per target speaker for the participating TTS systems to synthesize corresponding speech signals. This will be used to evaluate the overall quality of generated utterances from zero-shot TTS systems. For the evaluation details of the zero-shot TTS system, please refer to Section IV-A.

Synthesized signals from zero-shot TTS systems are supposed to be used to train the PSE systems, whose SE performance is assumed to be associated with the TTS systems' performance, i.e., the better the synthesized speech preserves the target speaker identity and speech quality, the more useful it is to personalize a speech enhancement system. The organizers impose no restrictions on the number of utterances the zero-shot TTS system generates for training the PSE systems. While there are no other specific restrictions on the zero-shot TTS model in terms of its architecture or how it is trained, **the LibriTTS test-clean dataset must be excluded in the TTS model's training as it is used for testing.**

B. The Downstream Personalized Speech Enhancement Task

PSE is a category of SE methods designed to train specialized models for individual users at test time. Here, we assume that PSE systems are personalized not only to specific speakers but also to the noise environments that particular users frequently encounter. Unlike the speaker and noise-agnostic SE models, which are trained to generalize to any arbitrary test speakers and noises, PSE models focus on the specific speaker and noise types of interest. Ideally, personalization could be achieved by training an SE model by setting up the clean speech of the target speaker as the denoising target and noisy speech with specific noise types as the noisy input. However, in practice, such personal data are difficult to acquire due to privacy concerns and the technical difficulties of recording the clean voice samples from the test-time user.

Successful TTS systems can resolve the data shortage issue. They can synthesize virtually as many clean speech signals as a PSE model needs for its personalized training. Since the main goal of personalization is to narrow down its usage to the target user, it is also crucial for the downstream task to be able to use personality-preserving, high-quality speech signals. Likewise, PSE performance is sensitive to the quality of participants' TTS systems, making it a suitable task for evaluating the generative data augmentation systems.

Participants are asked to use their synthesized speech to develop 20 PSE models for 20 target speakers, respectively. In addition, the organizers provide a test set consisting of 45 noisy utterances for each speaker, which are mixtures of nine clean utterances and five speaker-specific types of noise sources (Sec. III-C). The input signal-to-noise ratio (SNR) is randomly selected from $\{-2.5, 0, 2.5\}$ dB. Participants are required to submit enhanced speech for these samples. For a fair comparison, participants are first requested to enhance the noisy signals with the baseline PSE model architecture that organizers provide. Then, participants can optionally provide results from their own model architectures. We strongly encourage participants to build lightweight PSE models as one of the main benefits of personalization is about being able to reduce model size. To this end, the complexity of the model must be reported as well.

III. Challenge Datasets

A. Real-World Speakers

From the LibriTTS [17] dataset's test-clean fold, five male and five female speakers are randomly selected as the personalization target. After excluding utterances that are either shorter than 3 seconds or longer than 16 seconds, we collect 60 utterances from each speaker that are then divided into three subsets: one utterance for enrollment, 50 for evaluating the generated speech directly, and nine for testing the PSE systems.

B. Virtual Speakers

In order to propose a solution and draw attention to the potential ethics and privacy issues that revolve around using the speech corpus as a seed to synthesize personalized speech, we additionally adopted ten virtual speakers as target speakers. Similar to the real-world speakers, these virtual speakers include five male and five female speakers. However, unlike real-world speakers, who primarily exhibit reading-style speech, we aimed to leverage the diversity of TTS models to generate virtual speakers with various accents and emotional expressions. Note that all the challenge configurations remain the same for these ten speakers as well. The significant difference is that these virtual speakers were generated using a state-of-the-art TTS system provided by Meta.

C. Noise Sources

Not only to personalize the PSE systems for speakers but also to adapt to noisy environments, we designate five specific noise types per target speaker, which are randomly chosen from the *sound-bible* subset of the MUSAN [18] dataset.

TABLE I
DESCRIPTION OF THE DATASETS USED IN OUR CHALLENGE AND BASELINE EXPERIMENTS

Duration	Quantity	Descriptoin	Corpus
1 utt/spkrs	20 spkrs	Enrollment speech of target speakers for zero-shot TTS system.	10 spkrs from LibriTTS [17] test-clean and another 10 spkrs from the virtual speaker dataset (Sec. III-B)
9 utts/spkrs		Test speech samples for PSE system evaluation.	
50 utts/spkrs		Test speech samples for zero-shot TTS system evaluation. Forty utterances were used to train the **GT-6min** model in the baseline experiments.	
5 noises/spkrs	88 noises	Injection noise used during PSE training; a unique set of 5 noise sources are used for each speaker.	MUSAN [18]
40 utts/sprks	20 spkrs	Synthesized speech from zero-shot TTS models, used to train **6min** models in the baseline experiments.	Three augmented personalized datasets from three baseline zero-shot TTS systems, respectively.
180 utts/sprks		Additional synthesized speech from zero-shot TTS models, used to train **30min** models in the baseline experiments.	

IV. EVALUATION METRICS

A. Basic TTS Performance Evaluation

Although the challenge's main objective is to prove the generative models' capabilities in the downstream task, we can conduct some basic performance evaluation of the synthesized data themselves. To this end, the zero-shot TTS system's performance is evaluated across three aspects: speaker similarity, intelligibility, and perceptual quality. To measure speaker similarity between the generated speech and the reference speech, we calculate the cosine distance of speaker embeddings (SECS). X-vectors are extracted using a pre-trained speaker verification model from SpeechBrain [19] and utilized to calculate SECS. To assess the intelligibility of the generated speech, we employ the open-source Whisper model [1] for speech recognition and calculate the word error rate (WER).

Subjective evaluations are commonly conducted to assess the perceptual quality of TTS models. However, conducting these evaluations across numerous models poses practical challenges. Recently, neural network-based perceptual quality metrics have been widely studied [20], [21] and increasingly adopted [9], [22], [23]. In this work, we chose to use the UTMOS [24] metric—one of the best-performed MOS prediction networks in VoiceMOS Challenge 2024 [25]- to evaluate the perceptual quality and naturalness of the generated speech.

B. Evaluation metrics for PSE models

To evaluate the performance of the PSE models and examine the effectiveness of incorporating augmented speech from zero-shot TTS systems, we employ four metrics commonly used in SE: signal-to-distortion ratio improvement (SDRI), signal-to-distortion ratio (SDR) [26], extended short-time objective intelligibility (eSTOI) [27], and perceptual evaluation of speech quality (PESQ) [28]. All the metrics are based on the direct comparison between the participants' submissions, i.e., the enhanced versions of the noisy test signals and the held-out ground-truth clean speech signals. SDRI measures the improvement in SDR, indicating how effectively the model reduces distortion relative to the input. SDR evaluates the absolute quality of the enhanced signal by comparing the energy ratio between the target and distortion signals, reflecting the fidelity preservation of the model's output. eSTOI, a measure of intelligibility, assesses how well the model retains short-time temporal patterns. Lastly, PESQ evaluates the perceptual quality of generated speech by simulating human auditory perception.

V. BASELINE MODELS

In this section, we introduce the baseline zero-shot TTS and PSE models and report their performances as a reference for the participants.

A. Baseline zero-shot TTS models

We use three open-source zero-shot TTS models as the baseline models: YourTTS [3], SpeechT5-based zero-shot TTS model [29], and XTTS [4]. YourTTS is built on VITS [2] and conditions it via a speaker embedding extracted from an external pre-trained speaker verification model. SpeechT5 is a pre-trained encoder-decoder model for various spoken language processing tasks, including its TTS application, which we use as a baseline. Finally, XTTS is a zero-shot TTS model built on Tortoise [30], which incorporates a decoder-only Transformer with some modifications to improve voice cloning and enable faster training and inference.

B. Baseline PSE models

We adopt the ConvTasNet [31]-based architecture for our PSE models based on the original PSE model architecture proposed in [9], [32]. Following their recipe, the generalist models are first trained on LibriSpeech [33] and FSD50K [34], which have clean and noise-mixed speech samples, respectively. To introduce artificial noise addition, we utilize MUSAN [18] datasets. Given that reduced size is one of the main advantages of the personalized system, we focus on the medium, small, and tiny models from [9], [32], containing 437K, 224K, and 138.8K parameters, respectively.

Then, we fine-tune the generalist model into a test speaker-specific version for each test speaker, using the personalized speech datasets synthesized by the TTS models. To this end, we run the zero-shot TTS models to generate 40 new clean utterances per speaker. These synthesized signals work as the training target and are mixed with speaker-specific noise sources to create corresponding noisy input mixtures. In addition, we also provide an oracle performance by fine-tuning the generalist model with the ground-truth clean speech, i.e., 40 actual utterances from the same test speaker. Since the duration of these 40 utterances is about six minutes, we refer to these PSE models as **6min** models. To investigate the performance gains from using additional synthesized speech, we generate 180 additional utterances for fine-tuning, resulting in 220 training utterances (about 30 minutes). We refer to the PSE models with this extended dataset as **30min** models. Note that fine-tuning uses 10 and 30 validation utterances per speaker for the **6min** and **30min** models, respectively.

We use the negative SDR loss function as in [9]. For the optimization, Adam [35] is used with a low learning rate of 10^{-6}. The batch size is 8. We stop fine-tuning if the validation loss does not improve after 20 epochs. The input mixtures are with a randomly selected SNR value from the range of $[-5, 5]$.

C. Results

Basic TTS performance: The performance of zero-shot TTS models for real-world and virtual speakers is detailed in Table II and Table III, respectively. SpeechT5 achieves the best speaker similarity (SECS) and intelligibility (WER) scores, despite having the lowest perceptual

TABLE II
ZERO-SHOT TTS RESULTS FOR REAL-WORLD SPEAKERS WITH 95% CONFIDENCE INTERVALS (CIS).

Model	SECS	UTMOS	WER (%)
GT	0.973 ± 0.020	2.814 ± 1.011	3.152
YourTTS	0.940 ± 0.041	2.645 ± 0.709	5.313
SpeechT5	$\mathbf{0.969 \pm 0.013}$	2.064 ± 0.706	**3.352**
XTTS	0.950 ± 0.031	$\mathbf{2.746 \pm 0.857}$	8.030

TABLE III
ZERO-SHOT TTS RESULTS FOR VIRTUAL SPEAKERS WITH 95% CIS.

Model	SECS	UTMOS	WER (%)
Virtual-GT	0.977 ± 0.010	2.773 ± 0.716	2.740
YourTTS	0.938 ± 0.035	$\mathbf{2.755 \pm 0.627}$	5.298
SpeechT5	$\mathbf{0.961 \pm 0.011}$	2.221 ± 0.684	**2.195**
XTTS	0.952 ± 0.022	2.696 ± 0.708	2.551

quality (UTMOS) scores in both cases. For real-world speakers, XTTS achieves the highest UTMOS score but the worst WER score, while for virtual speakers, it ranks second in both the UTMOS score and WER score. We assume that the diversity of accents and emotions among the virtual speakers, along with some artificial noise already present in the reference speech, influenced these performance outcomes. Overall, each TTS system has its own unique property, which can be measured in different ways. Next, we examine how these properties affect the downstream task performance when used as a data augmentation method.

PSE performance: The PSE results for real-world and virtual speakers are detailed in Table IV and Table V, respectively. Across all model sizes, the generalist models performed the worst on every evaluation metric. This implies that even PSE models built using the lowest-performing zero-shot TTS systems achieved significantly better performance than the generalist SE model. This demonstrates the effectiveness of personalized data augmentation using external generative models. For all sizes, the **GT-6min** model achieved the highest scores across all metrics, outperforming all **30min** models. This suggests that the data quality is crucial for PSE performance; even with a larger dataset of lower quality, performances were inferior to those achieved with a smaller amount of high-quality data. Compared to [9], where some TTS models did not introduce PSE improvement, this time, the adaptation to the noise sources could have contributed to better PSE performance.

When comparing the **6min** and **30min** models fine-tuned with augmented data from various zero-shot TTS models, we observed that in most cases, PSE performance improved as the amount of the augmented data increased, although the improvement is marginal. For both real-world and virtual speakers, the **SpeechT5-30min** model achieved the best performance in SDRI, SDR, and eSTOI. Given the high speaker similarity of the **SpeechT5** model, we believe that speaker similarity is an important factor in building an effective PSE model. The **XTTS-30min** model achieved the highest PESQ scores for real-world speakers and for medium-sized PSE model for virtual speakers, while the **SpeechT5-30min** model performed best for small and tiny-sized models for virtual speakers. Since PESQ focuses on perpetual quality, the **XTTS** model's high perceptual quality likely contributed to its strong PESQ performance.

VI. DISCUSSION AND FUTURE WORK

In our baseline experiments, we demonstrated the potential of zero-shot TTS models for data augmentation in PSE applications. We also highlighted the importance of adaptation data quality for PSE model performance. Speaker similarity and intelligibility emerged as the most relevant factors, with perceptual quality also influencing PSE outcomes. However, as the number of zero-shot TTS models in our baseline experiments was limited, we anticipate that this challenge will enable a more in-depth exploration of the relationship between TTS model performance and PSE outcomes through a broader variety of zero-shot TTS systems. Data augmentation stands to benefit significantly from advances in generative AI, though its application requires careful consideration due to the complex nature of synthetic data usability. We also explored virtual speakers as a privacy-preserving alternative. A possible application is to build a PSE model that reflects the target speaker's characteristics using virtual speakers, thereby addressing privacy concerns associated with collecting target speaker data. In the future, the organizers plan to expand this challenge to additional downstream tasks.

TABLE IV
PSE RESULTS FOR THE REAL-WORLD SPEAKERS. M, S, AND T IN SIZE INDICATE MEDIUM, SMALL, AND TINY SIZES, RESPECTIVELY.

Model	Size	SDRI	SDR	eSTOI	PESQ
Generalist	M	9.495	9.997	0.708	1.487
	S	9.069	9.572	0.693	1.446
	T	8.267	8.770	0.670	1.382
GT-6min	M	13.890	14.393	0.824	2.108
	S	12.991	13.494	0.799	1.952
	T	12.247	12.750	0.776	1.810
YourTTS-6min	M	12.110	12.613	0.786	1.950
	S	11.217	11.720	0.765	1.835
	T	10.521	11.024	0.746	1.721
YourTTS-30min	M	12.347	12.850	0.789	1.968
	S	11.427	11.931	0.770	1.857
	T	10.714	11.217	0.749	1.728
SpeechT5-6min	M	12.390	12.893	0.798	1.991
	S	11.479	11.982	0.774	1.847
	T	10.710	11.213	0.751	1.712
SpeechT5-30min	M	**12.519**	**13.022**	**0.801**	2.004
	S	**11.626**	**12.129**	**0.778**	1.862
	T	**10.842**	**11.345**	**0.754**	1.724
XTTS-6min	M	12.302	12.805	0.795	**2.013**
	S	11.413	11.916	0.770	1.870
	T	10.514	11.017	0.743	1.735
XTTS-30min	M	12.341	12.844	0.795	**2.013**
	S	11.453	11.956	0.771	**1.877**
	T	10.593	11.097	0.744	**1.742**

TABLE V
PSE RESULTS FOR THE VIRTUAL SPEAKERS.

Model	Size	SDRI	SDR	eSTOI	PESQ
Generalist	M	10.291	10.749	0.814	1.467
	S	10.298	10.756	0.807	1.440
	T	9.455	9.912	0.788	1.368
Virtual-GT-6min	M	14.907	15.365	0.890	1.986
	S	14.123	14.580	0.870	1.828
	T	13.396	13.854	0.850	1.679
YourTTS-6min	M	12.831	13.288	0.856	1.903
	S	12.567	13.025	0.841	1.776
	T	11.737	12.195	0.817	1.632
YourTTS-30min	M	12.960	13.418	0.859	1.908
	S	12.624	13.082	0.842	1.773
	T	11.768	12.226	0.818	1.625
SpeechT5-6min	M	13.682	14.140	0.873	1.915
	S	12.959	13.417	0.851	1.784
	T	12.188	12.646	0.832	1.636
SpeechT5-30min	M	**13.713**	**14.171**	**0.874**	1.918
	S	**12.979**	**13.437**	**0.852**	**1.784**
	T	**12.199**	**12.657**	**0.832**	**1.641**
XTTS-6min	M	13.288	13.746	0.866	1.910
	S	12.480	12.938	0.840	1.761
	T	11.824	12.282	0.821	1.627
XTTS-30min	M	13.276	13.734	0.864	**1.927**
	S	12.479	12.937	0.839	1.758
	T	11.765	12.223	0.818	1.626

REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. on Machine Learning (ICLR)*, 2023.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *Proc. Int. Conf. on Machine Learning (ICLR)*, 2021.

[3] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge *et al.*, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2022.

[4] E. Casanova, K. Davis, E. Gölge, G. Göknar, I. Gulea *et al.*, "XTTS: A massively multilingual zero-shot text-to-speech model," in *Proc. Interspeech*, 2024.

[5] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu *et al.*, "Neural codec language models are zero-shot text to speech synthesizers," *CoRR*, vol. abs/2301.02111, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2301.02111

[6] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli *et al.*, "MetricGAN+: An improved version of MetricGAN for speech enhancement," in *Proc. Interspeech*, 2021.

[7] H. J. Park, B. H. Kang, W. Shin, J. S. Kim, and S. W. Han, "MANNER: Multi-view attention netwfork for noise erasure," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[8] K. Yang, T.-Y. Hu, J.-H. R. Chang, H. Swetha Koppula, and O. Tuzel, "Text is all you need: Personalizing ASR models using controllable speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[9] A. Kuznetsova, A. Sivaraman, and M. Kim, "The potential of neural speech synthesis-based data augmentation for personalized speech enhancement," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[10] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, p. 60, 2019.

[11] B. Trabucco, K. Doherty, M. A. Gurinas, and R. Salakhutdinov, "Effective data augmentation with diffusion models," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2024.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph *et al.*, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.

[13] J.-S. Bae, J. Y. Lee, J.-H. Lee, S. Mun, T. Kang *et al.*, "Latent filling: Latent space data augmentation for zero-shot speech synthesis," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[14] Q. Chen, Z. Ma, T. Liu, X. Tan, Q. Lu *et al.*, "Improving few-shot learning for talking face system with TTS data augmentation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[15] H. J. Park, D. Agarwal, N. Chen, R. Sun, K. Partridge *et al.*, "Utilizing TTS synthesized data for efficient development of keyword spotting model," *CoRR*, vol. abs/2407.18879, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2407.18879

[16] F. Nespoli, D. Barreda, and P. A. Naylor, "Zero shot text to speech augmentation for automatic speech recognition on low-resource accented speech corpora," in *57th Asilomar Conference on Signals, Systems, and Computers*, 2023.

[17] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia *et al.*, "LibriTTS: A corpus derived from librispeech for text-to-speech," in *Proc. Interspeech*, 2019.

[18] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015. [Online]. Available: http://arxiv.org/abs/1510.08484

[19] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell *et al.*, "SpeechBrain: A general-purpose speech toolkit," *CoRR*, vol. abs/2106.04624, 2021. [Online]. Available: https://arxiv.org/abs/2106.04624

[20] Y. Choi, Y. Jung, and H. Kim, "Deep MOS predictor for synthetic speech using cluster-based modeling," in *Proc. Interspeech*, 2020.

[21] G. Mittag, B. Naderi, A. Chehadi, and S. Moller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Proc. Interspeech*, 2021.

[22] W.-C. Huang, L. P. Violeta, S. Liu, J. Shi, and T. Toda, "The singing voice conversion challenge 2023," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[23] Z. Ju, Y. Wang, K. Shen, X. Tan, D. Xin *et al.*, "NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models," in *Proc. Int. Conf. on Machine Learning (ICML)*, 2024.

[24] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The T05 system for the VoiceMOS challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech," *CoRR*, vol. abs/2409.09305, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2409.09305

[25] W. Huang, S. Fu, E. Cooper, R. E. Zezario, T. Toda, H. Wang, J. Yamagishi, and Y. Tsao, "The VoiceMOS challenge 2024: Beyond speech quality prediction," *CoRR*, vol. abs/2409.07001, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2409.07001

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.

[28] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. (ICASSP)*, 2001.

[29] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren *et al.*, "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing," in *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[30] J. Betker, "Better speech synthesis through scaling," *CoRR*, vol. abs/2305.07243, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.07243

[31] 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, 2019.

[32] A. Sivaraman and M. Kim, "Efficient personalized speech enhancement through self-supervised learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1342–1356, 2022.

[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. on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[34] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6980