INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus

Yuma Koizumi[1], *Heiga Zen*[1], *Shigeki Karita*[1], *Yifan Ding*[1], *Kohei Yatabe*[2], *Nobuyuki Morioka*[1], *Michiel Bacchiani*[1], *Yu Zhang*[3], *Wei Han*[3], *Ankur Bapna*[3]

[1] Google, Japan, [2] Tokyo University of Agriculture & Technology, Japan, [3] Google, USA

{koizumiyuma,heigazen,karita}@google.com

Abstract

This paper introduces a new speech dataset called "LibriTTS-R" designed for text-to-speech (TTS) use. It is derived by applying speech restoration to the LibriTTS corpus, which consists of 585 hours of speech data at 24 kHz sampling rate from 2,456 speakers and the corresponding texts. The constituent samples of LibriTTS-R are identical to those of LibriTTS, with only the sound quality improved. Experimental results show that the LibriTTS-R ground-truth samples showed significantly improved sound quality compared to those in LibriTTS. In addition, neural end-to-end TTS trained with LibriTTS-R achieved speech naturalness on par with that of the ground-truth samples. The corpus is freely available for download from http://www.openslr.org/141/.

Index Terms: Text-to-speech, dataset, speech restoration

1. Introduction

Text-to-speech (TTS) technologies have been rapidly advanced along with the development of deep learning [1–6]. With studio-quality recorded speech data, one can train acoustic models [2, 3] and high-fidelity neural vocoders [7, 8]. These have enabled us to synthesize speech in a reading style almost as natural as human speech. In addition, many implementations of the latest TTS models have been published [9, 10], and the gateway to TTS research is certainly widening.

One of the remaining barriers to develop high-quality TTS systems is the lack of large and high-quality public dataset. Training of high-quality TTS models requires a large amount of studio-quality data. In several TTS papers, over 100 hours of studio-recorded data have been used [3, 8, 11]. Unfortunately, such studio-recorded datasets are not publicly available, and thus reproducing their results is difficult for others.

At the same time, speech restoration (SR) has advanced using speech generative models [12–18]. These state-of-the-art models can convert reverberated lecture and historical speech to studio-recorded quality [16–18]. Inspired by these results, we came up with an idea that the above-mentioned barrier can be removed by applying SR to public datasets.

With this paper, we publish *LibriTTS-R*, a quality-improved version of LibriTTS [19]. LibriTTS is a non-restrictive license multi-speaker TTS corpus consisting of 585 hours of speech data from 2,456 speakers and the corresponding texts. We cleaned LibriTTS by applying a text-informed SR model, *Miipher*, [20] that uses w2v-BERT [21] feature cleaner and WaveFit neural vocoder [8]. By subjective experiments, we show that the speech naturalness of a TTS model trained with LibriTTS-R is greatly improved from that trained with LibriTTS, and is comparable with that of the ground-truth.

LibriTTS-R is publicly available at http://www.openslr.org/141/, with the same non-restrictive license. Audio samples of the ground-truth and TTS generated samples are available at our demo page[1].

2. The LibriTTS corpus

The LibriTTS corpus is one of the largest multi-speaker speech datasets designed for TTS use. This dataset consists of 585 hours of speech data at 24 kHz sampling rate from 2,456 speakers and the corresponding texts. The audio and text materials are derived from the LibriSpeech corpus [22], which has been used for training and evaluating automatic speech recognition systems. Since the original LibriSpeech corpus has several undesired properties for TTS including sampling rate and text normalization issues, the samples in LibriTTS were re-derived from the original materials (MP3 from LibriVox and texts from Project Gutenberg) of LibriSpeech.

One issue is that the LibriTTS sound quality is not on par with smaller scale but higher quality TTS datasets such as LJspeech [23]. The quality of the TTS output is highly affected by that of the speech samples used in model training. Therefore, the quality of the generated samples of a TTS model trained on LibriTTS doesn't match those of the ground-truth samples [24, 25]. For example, Glow-TTS achieved 3.45 mean-opinion-score (MOS) on LibriTTS where the speech obtained from the ground-truth mel-spectrograms by a vocoder was 4.22 [24]. Note that MOSs on the LJspeech for generated and ground-truth were 4.01 and 4.19, respectively [24]. The results suggest that the quality of speech samples in LibriTTS are inadequate for training of high-quality TTS models.

3. Data processing pipeline

Although noisy TTS datasets are useful for advanced TTS model training [26–28], access to large scale high-quality datasets is as equally important for advancing TTS techniques. To provide a public large-scale and high-quality TTS dataset, we apply a SR model to LibriTTS.

3.1. Speech restoration model overview

One critical requirement of SR models for the purpose of cleaning datasets is robustness. If the SR model generates a large number of samples with artifacts, it will adversely impact the subsequent TTS model training. Therefore, for our purposes, we need to reduce as much as possible the number of samples that fail to be recovered.

[1] https://google.github.io/df-conformer/librittsr/

10.21437/Interspeech.2023-1584

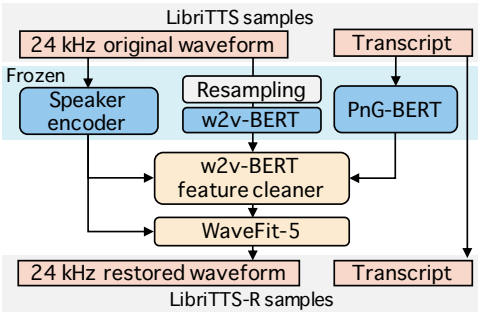

Figure 1: *Data processing pipeline overview. Speech samples in the LibriTTS corpus are restored using Miipher [20].*

To satisfy this requirement, we use a text-informed parametric re-synthesis-based SR model, *Miipher* [20], as shown in Fig. 1. In this model, first, w2v-BERT features are extracted by w2v-BERT [21] from the noisy waveform. Then, a DF-Conformer [29]-based feature-cleaner predicts the w2v-BERT features of the clean waveform. Finally, the restored waveform is synthesized using a WaveFit-5 neural vocoder [8].

The reason for selecting Miipher is that it adresses two particularly difficult to restore degradation patterns observed in LibriTTS samples. The first degradation is phoneme masking. Speech signals are sometimes masked by noise and/or reverberation, resulting in speech that is is difficult to discriminate from noise without additional information. The second degradation is phoneme deletion. Important frequency parts of some phonemes could be missing from the signal due to non-linear audio processing and/or down-sampling. To address these problems, Miipher introduced two techniques. (i) for the input feature, it uses w2v-BERT [21] features instead of log-mel spectrogram used in a conventional SR model [17], and (ii) to use linguistic features conditioning extracted by PnG-BERT [3] from the transcript corresponding to the noisy speech. Since w2v-BERT is trained on large amounts of degraded speech samples and it improves ASR performance, we expect its effectiveness on making SR models robust against speech degradation. In addition, the use of text information improving speech inpainting performance [30], we consider that it also improves speech restoration performance. For the detail, please see the original paper [20].

3.2. Speech restoration model training

We trained a Miipher model with a proprietary dataset that contains 2,680 hours of noisy and studio-quality speech pairs. The target speech dataset contains 670 hours of studio-recorded Australian, British, Indian, Nigerian, and North American English at 24 kHz sampling. For the noise dataset, we used the TAU Urban Audio-Visual Scenes 2021 dataset [31], internally collected noise snippets that simulate conditions like cafe, kitchen, and cars, and noise sources. The noisy utterances were generated by mixing randomly selected speech and noise samples from these datasets with signal-to-noise ratio (SNR) from 5 dB to 30 dB. In addition, we augmented the noisy dataset with 4 patterns depending on the presence or absence of reverberation and codec artifacts. A room impulse response (RIR) for each sample was generated by a stochastic RIR generator using the image method [32]. For simulating codec artifacts, we randomly applied one of MP3, Vorbis, A-law, Adaptive Multi-Rate Wideband (AMR-WB), and OPUS with a random bit-rate. The detailed simulation parameters were listed in [20].

We first pre-trained the feature-cleaner and WaveFit neural vocoder 150k and 1M steps, respectively, where WaveFit was trained to reconstruct waveform from clean w2v-BERT features. Then, we fine-tuned the WaveFit neural vocoder 350k steps using cleaned w2v-BERT features by the pre-trained feature-cleaner.

Table 1: *MOS and SxS test results on the ground-truth samples with their 95% confidence intervals. A positive SxS score indicates that LibriTTS-R was preferred.*

Split	MOS (↑) LibriTTS	MOS (↑) LibriTTS-R	SxS
test-clean	4.36 ± 0.08	$\mathbf{4.41 \pm 0.07}$	0.80 ± 0.15
test-other	3.94 ± 0.10	$\mathbf{4.09 \pm 0.10}$	1.42 ± 0.14

3.3. Speech restoration pipeline

First, we calculated PnG-BERT [3] features from a transcript and a speaker embedding using the speaker encoder described in [20] from the original 24 kHz sampling waveform. Here, for speech samples with waveform lengths shorter than 2 seconds, the speaker embedding was calculated after repeating them to get a pseudo longer waveform. Since the w2v-BERT [21] model was trained on 16 kHz waveforms, we applied down-sampling to the LibriTTS sample for calculating w2v-BERT features. Finally, we synthesized restored 24 kHz sampling waveform using WaveFit [8].

4. Experiments

4.1. Subjective experiments for ground-truth samples

4.1.1. Experimental setups

We first compared the quality of ground-truth speech samples in LibriTTS-R with those in LibriTTS. We evaluated the sound quality using "test-clean" and "test-other" subsets. We randomly selected 620 samples from each subset. Since the "train-*" and "dev-*" subsets are also divided into "clean" and "other" according to the same word-error-rate (WER)-based criteria, the sound quality of the entire dataset can be predicted by evaluating the sound quality of these two subsets.

To evaluate subjective quality, we rated speech quality through mean-opinion-score (MOS) and side-by-side (SxS) preference tests. We asked to rate the naturalness in MOS test, and "*which sound quality is better?*" in SxS test. The scale of MOS was a 5-point scale (1: Bad, 2: Poor, 3: Fair, 4: Good, 5: Excellent) with rating increments of 0.5, and that of SxS was a 7-point scale (-3 to 3). Test stimuli were randomly chosen and each stimulus was evaluated by one subject. Each subject was allowed to evaluate up to six stimuli, that is, over 100 subjects participated in this experiment to evaluate 640 samples in each condition. The subjects were paid native English speakers in the United States. They were requested to use headphones in a quiet room. Audio samples are available in our demo page [1].

4.1.2. Results

Table 1 shows the MOS and SxS test results. In terms of speech naturalness, LibriTTS achieved high MOSs: 4.36 and 3.94 on test-clean and test-other, respectively. Although LibriTTS-R achieved better MOSs than LibriTTS in both splits, the differ-

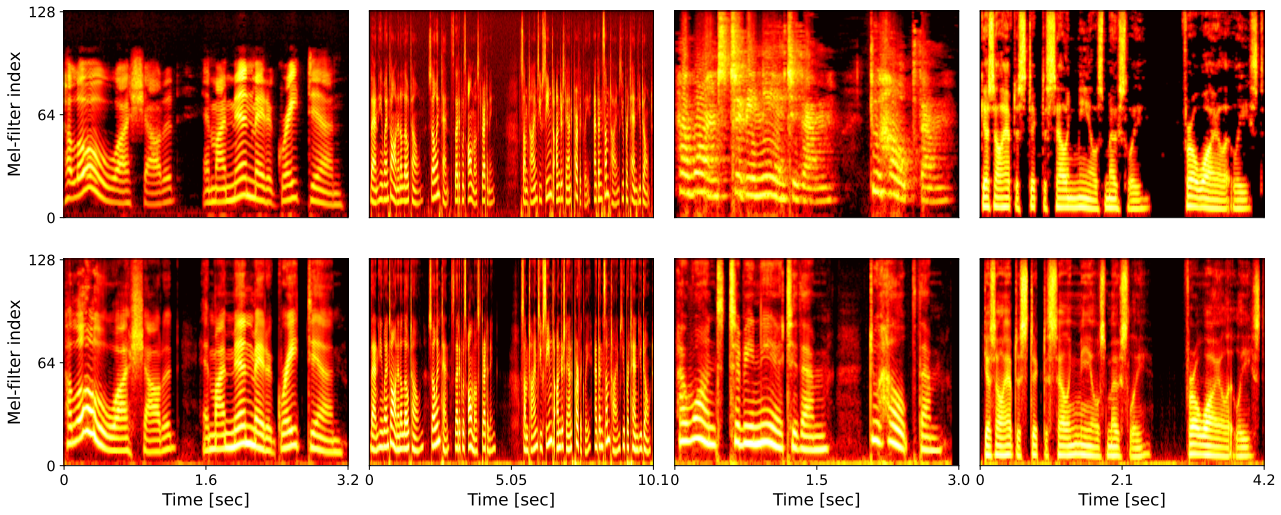

Figure 2: *Log-mel spectrograms of ground-truth waveforms from (top) LibriTTS and (bottom) LibriTTS-R. The left two and right two examples are from "test-clean" and "test-other" splits, respectively.*

ence was not significant. The reason of small difference in naturalness might be because ground-truth samples are real speech spoken by humans. In contrast, in terms of sound quality rated by SxS tests, significant differences were observed on both split.

To confirm whether the text-content and speaker in the restored speech samples are maintained, we evaluated the WER and speaker similarity. To compute WER, we used "Pre-trained Conformer XXL" model proposed in [33]. WER of "test-clean" and "test-other" splits of LibriTTS were 3.4 and 5.1, whereas those of LibriTTS-R were 3.2 and 5.1, respectively[2]. Therefore, the text contents are considered to be not changed. To evaluate speaker similarity, we used the cosine similarity of speaker embedding [34, 35]. We calculated the similarity between the different utterances spoken by the same speaker in the same dataset. This is because the samples in LibriTTS are distorted, even if the similarity between corresponding samples in LibriTTS and LibriTTS-R is small, this does not necessarily indicate speaker similarity. The cosine similarity of LibriTTS "test-clean" and "test-other" splits were 0.784 and 0.755, and those of LibriTTS-R were 0.762 and 0.745. Since the similarity calculated from the samples in LibriTTS spoken by different speakers was 0.302, the speech characteristics of each speaker is considered to be consistent.

Figure 2 shows the 128-dim log-mel spectrogram of speech samples from LibriTTS and LibriTTS-R. We can see that the LibriTTS samples are degraded by a variety of factors even if these are from the test-clean split: from left to right, it can be considered that speech samples were degraded by down-sampling, environmental noise, reverberation, and non-linear speech enhancement, respectively. As we can see spectrograms of LibriTTS-R samples, the SR model restored these speech samples into high-quality ones. This could be the reason of the significant differences in the SxS tests.

Note that we have found a few examples that LibriTTS speech sample achieved a better score in SxS comparison. By listening these examples, two of 640 LibriTTS-R speech samples were distorted due to the failure of SR. Since it is difficult to manually check all samples, we have not checked all speech samples in LibriTTS-R. Therefore, the samples in training splits may also contain a small number of distorted samples.

[2]WER were a bit higher than those reported in the original paper [33], because the ASR model was trained on noisy speech and transcripts normalized by a different text-normalizer.

4.2. Subjective experiments for TTS generated samples

4.2.1. Experimental setups

We trained multi-speaker TTS models with the same architecture and the same hyper-parameters using either the LibriTTS or LibriTTS-R corpus. The TTS model was build by concatenating the following acoustic model and neural vocoder without joint fine-tuning.

Acoustic model: We used a duration unsupervised Non-Attentive Tacotron (NAT) with a fine-grained variational auto-encoder (FVAE) [11]. We used the same hyper-parameters and training parameters listed in the original paper [11]. We trained this model for 150k steps with a batch size of 1,024.

Neural vocoder: We used a WaveRNN [36] which consisted of a single long short-term memory layer with 512 hidden units, 5 convolutional layers with 512 channels as the conditioning stack to process the mel-spectrogram features, and a 10-component mixture of logistic distributions as its output layer. The learning rate was linearly increased to 10^{-4} in the first 100 steps then exponentially decayed to 10^{-6} from 200k to 300k steps. We trained this model using the Adam optimizer [37] for 500k steps with a batch size of 512.

The TTS model was trained on two types of training datasets: Train-460 and Train-960. Train-460 consists of the "train-clean-100" and "train-clean-360" subsets, and Train-960 indicates using "train-other-500" in addition to Train-460.

For the test sentences, we randomly selected 620 evaluation sentences from the test-clean split. We synthesized waveforms with 6 speakers (three female and three male) those are used in the LibriTTS baseline experiments [19]. The female and male reader IDs were (19, 103, 1841) and (204, 1121, 5717), respectively. To evaluate subjective quality, we rated speech naturalness through MOS and side-by-side (SxS) preference tests. The listening test setting was the same as Sec. 4.1 Audio samples of generated speech are available in our demo page [1].

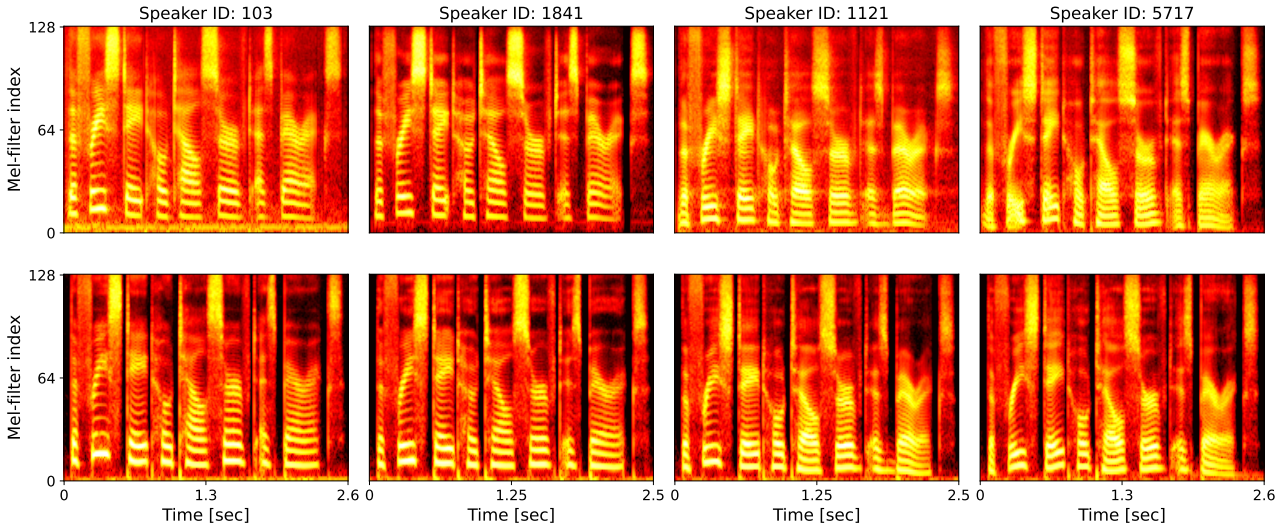

Figure 3: *Log-mel spectrograms of TTS generated waveforms where the multi-speaker TTS model was trained on (top) LibriTTS and (bottom) LibriTTS-R, respectively. The input text was "The Free State Hotel served as barracks".*

Table 2: *MOSs for the baseline multi-speaker TTS model outputs with their 95% confidence intervals.*

Training dataset	**Speaker ID**					
	19	103	1841	204	1121	5717
LibriTTS Train-460	2.49 ± 0.10	2.94 ± 0.10	3.40 ± 0.09	2.88 ± 0.10	2.72 ± 0.10	2.86 ± 0.09
LibriTTS Train-960	2.59 ± 0.10	2.75 ± 0.10	3.35 ± 0.10	2.74 ± 0.09	2.83 ± 0.10	2.97 ± 0.10
LibriTTS-R Train-460	$\mathbf{4.11 \pm 0.08}$	4.09 ± 0.08	3.88 ± 0.09	3.67 ± 0.09	3.92 ± 0.09	3.67 ± 0.08
LibriTTS-R Train-960	4.06 ± 0.08	$\mathbf{4.31 \pm 0.08}$	$\mathbf{4.20 \pm 0.08}$	$\mathbf{4.11 \pm 0.08}$	$\mathbf{4.23 \pm 0.07}$	$\mathbf{4.08 \pm 0.08}$

Table 3: *SxS test results on the baseline multi-speaker TTS model outputs with their 95% confidence intervals. A positive score indicates that training on LibriTTS-R was preferred.*

Speaker ID	**Training dataset**	
	Train-460	Train-960
19	2.38 ± 0.11	2.51 ± 0.10
204	1.84 ± 0.14	2.20 ± 0.12

4.2.2. Results

Table 2 shows the MOS results. In all speaker IDs except for ID 19, the TTS model using LibriTTS-R Train-960 as the training dataset achieved the highest MOSs. For Speaker ID 19, the model using LibriTTS-R Train-460 achieved the highest MOS, which was not significantly different from that using LibriTTS-R Train-960. In other speaker IDs, MOSs of LibriTTS-R Train-960 were significantly better than that of LibriTTS-R Train-460. This trend was not observed in LibriTTS, rather in some cases, MOS was decreased by using LibriTTS Train-960. The reason for this degradation might be because that the "train-other-500" split contains a lot of degraded speech samples. This result suggests that the use of LibriTTS "train-other-500" split rather degrades the output quality of the TTS. In contrast, speech samples in LibriTTS-R "train-other-500" split are restored to high-quality speech samples, and resulting in that enables us to use a large amount of high-quality training data and improved the naturalness of the TTS outputs. In addition, the TTS model trained on LibriTTS-R Train-960 achieved MOSs on a par with human spoken speech samples in LibriTTS, effects of a few distorted speech samples in the training can be considered as not significant.

Table 3 shows the SxS results. We observed that the use of LibriTTS-R also improve not only naturalness but also the sound quality of TTS outputs. Figure 3 shows 128-dim log-mel spectrograms of TTS outputs. We can see the harmonic structure is broken in the ID 5717 output of the TTS model trained on LibriTTS (top right). The presence of such a sample could be the reason for the lower naturalness scores on the MOS test. Also, from ID 103 and 1121 examples, we can observe background noise in the output of TTS model trained on LibriTTS. Such background noise does not exist in the outputs of TTS model trained on LibriTTS-R. From these results, we conclude that the LibriTTS-R corpus is a better TTS corpus that the LibriTTS corpus, and enables us to train a high-quality TTS model.

5. Conclusions

This paper introduced LibriTTS-R, a sound quality improved version of LibriTTS [19]. We cleaned speech samples in the LibriTTS corpus by applying an SR model [20]. By subjective experiments, we show that the speech naturalness of a TTS model trained with LibriTTS-R is improved from that trained with LibriTTS, and is comparable with that of the ground-truth. This corpus is released online, and it is freely available for download from `http://www.openslr.org/141/`. We hope that the release of this corpus accelerates TTS research.

6. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv:1609.03499*, 2016.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018.

[3] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS," in *Proc. Interspeech*, 2021.

[4] I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable TTS," in *Proc. ICASSP*, 2021.

[5] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019.

[6] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[7] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.

[8] Y. Koizumi, K. Yatabe, H. Zen, and M. Bacchiani, "WaveFit: An iterative and non-autoregressive neural vocoder based on fixed-point iteration," in *Proc. IEEE SLT*, 2023.

[9] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *Proc. ICASSP*, 2020.

[10] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. De Mori, and Y. Bengio, "SpeechBrain: A general-purpose speech toolkit," *arXiv:2106.04624*, 2021.

[11] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling," *arXiv:2010.04301*, 2020.

[12] S. Maiti and M. I. Mandel, "Parametric resynthesis with neural vocoders," in *Proc. IEEE WASPAA*, 2019.

[13] ——, "Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement," in *Proc. ICASSP*, 2020.

[14] T. Saeki, S. Takamichi, T. Nakamura, N. Tanji, and H. Saruwatari, "SelfRemaster: Self-supervised speech restoration with analysis-by-synthesis approach using channel modeling," in *Proc Interspeech*, 2022.

[15] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. Interspeech*, 2020.

[16] ——, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. IEEE WASPAA*, 2021.

[17] H. Liu, X. Liu, Q. Kong, Q. Tian, Y. Zhao, D. Wang, C. Huang, and Y. Wang, "VoiceFixer: A unified framework for high-fidelity speech restoration," in *Proc Interspeech*, 2022.

[18] J. Serrà, S. Pascual, J. Pons, R. O. Araz, and D. Scaini, "Universal speech enhancement with score-based diffusion," *arXiv:2206.03065*, 2022.

[19] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019.

[20] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, Y. Zhang, W. Han, A. Bapna, and M. Bacchiani, "Miipher: A robust speech restoration model integrating self-supervised speech and text representations," *arXiv:2303.01664*, 2023.

[21] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. IEEE ASRU*, 2021.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. ICASSP*, 2015.

[23] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[24] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[25] R. Valle, K. J. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[26] E. Casanova, J. Weber, C. Shulby, A. C. Junior, E. Gölge, and M. Antonelli Ponti, "YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," *arXiv:2112.02418*, 2021.

[27] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv:2301.02111*, 2023.

[28] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier, S. Girgin, O. Pietquin, M. Sharifi, M. Tagliasacchi, and N. Zeghidour, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *arXiv:2302.03540*, 2023.

[29] Y. Koizumi, S. Karita, S. Wisdom, H. Erdogan, J. R. Hershey, L. Jones, and M. Bacchiani, "DF-Conformer: Integrated architecture of Conv-TasNet and Conformer using linear complexity self-attention for speech enhancement," in *Proc. IEEE WASPAA*, 2021.

[30] Z. Borsos, M. Sharifi, and M. Tagliasacchi, "SpeechPainter: Text-conditioned speech inpainting," in *Proc. Interspeech*, 2022.

[31] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *Proc. ICASSP*, 2021.

[32] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, 1979.

[33] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," in *Proc. NeurIPS SAS 2020 Workshop*, 2020.

[34] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Proc. NeurIPS*, 2018.

[35] Y. Chen, Y. Assael, B. Shillingford, D. Budden, S. Reed, H. Zen, Q. Wang, L. C. Cobo, A. Trask, B. Laurie, C. Gulcehre, A. van den Oord, O. Vinyals, and N. de Freitas, "Sample efficient adaptive text-to-speech," in *Proc. ICLR*, 2019.

[36] N. Kalchbrenner, W. Elsen, K. Simonyan, S. Noury, N. Casagrande, W. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, 2018.

[37] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.