Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 579–584 DOI: 10.15439/2024F9637 ISSN 2300-5963 ACSIS, Vol. 39

Spoken Language Corpora Augmentation with Domain-Specific Voice-Cloned Speech

Mateusz Czyżnikiewicz, Łukasz Bondaruk,
Jakub Kubiak, Adam Wiącek, Łukasz Degórski
Samsung R&D Institute Poland
Plac Europejski 1
00-844 Warszawa, Poland
Email: {m.czyznikiew,l.bondaruk,j.kubiak3,
a.wiacek2,l.degorski}@samsung.com

Marek Kubis, Paweł Skórzewski
0000-0002-2016-2598
0000-0002-5056-2808
Adam Mickiewicz University, Poland
Faculty of Mathematics and Computer Science
ul. Uniwersytetu Poznanskiego 4
61-614 Poznan, Poland
Email: {mkubis, pawel.skorzewski}@amu.edu.pl

***Abstract*—In this paper we study the impact of augmenting spoken language corpora with domain-specific synthetic samples for the purpose of training a speech recognition system. Using both a conventional neural TTS system and a zero-shot one with voice cloning ability we generate speech corpora that vary in the number of voices. We compare speech recognition models trained with addition of different amounts of synthetic data generated using these two methods with a baseline model trained solely on voice recordings. We show that while the quality of voice-cloned dataset is lower, its increased multivoiceity makes it much more effective than the one with only a few voices synthesized with the use of a conventional neural TTS system. Furthermore, our experiments indicate that using low variability synthetic speech quickly leads to saturation in the quality of the ASR whereas high variability speech provides improvement even when increasing total amount of data used for training by 30%.**

I. Introduction

WITH THE development of better TTS systems in recent years, there has been an increasing number of research papers on using synthesized data for ASR training [1], [2], [3]. One could argue that, if synthesized samples covered a more diverse set of voice characteristics, even with decrease in speech quality, the data could be used more effectively for training ASR. Conventional neural TTS systems [4], like Tacotron2 [5] or FastSpeech [6], require large amount of high-quality paired text and speech data, which is not available for most languages, especially for multiple voices. Because of that, we cannot use them to produce output with more than a few to a dozen of voices, even for otherwise high-resource languages like German [4]. Recent advancements in speech synthesis brought zero-shot models that use neural codec encoding instead of mel-spectogram speech representation [7], [8], [9]. Thanks to their zero-shot voice cloning ability, they are able to generate high quality audio with any person's voice, having just a few seconds recording of it. This allows for generating synthetic corpora with hundreds of voices.

Our work examines the usefulness of having a synthetic corpora with a diverse set of voices. For comparison, we employ a zero-shot TTS and a conventional neural TTS to produce a domain-specific synthetic dataset with high and low number of speakers, respectively. We select a virtual assistant (VA) domain as our experiment target. Then, we examine the usefulness of both synthetic datasets in improving the ASR model's performance. We show that the high voice diversity of generated data makes it much more effective. Furthermore, our results indicate that the potential for using synthesized data to improve the ASR performance is limited by variability of the speech produced by a conventional neural TTS system.

This research was partially funded by the *CAIMAC: Conversational AI Multilingual Augmentation and Compression* project, a cooperation between Adam Mickiewicz University and Samsung Electronics Poland.

II. Related Work

Prior work has shown that using text-to-speech data can improve ASR performance. Rossenbach et al. [3] examined the impact of synthetic data for various ASR architectures. They showed that using TTS data pre-processing techniques can increase the robustness of ASR training. They reported 38% relative improvement after adding synthetic data to the attention encoder-decoder ASR system.

The addition of synthetic data can play an important role in a low-resource setting. Bartelds et al. [10] showed that adding synthetic data to the ASR training on such languages like Besemah and Nasal reduced relative WER up to 25.5%.

In some situations, all that is needed to build an ASR is a text corpus. Rossenbach et al. [11] demonstrated this strategy. They achieved relative improvement of up to 33% in WER over the baseline with data augmentation in a low-resource setting.

Another use for synthetic data can be to improve the recognition of out-of-vocabulary (OOV) words [12]. OOV is a prevalent issue encountered by real-world virtual assistants that must adapt to the ever-evolving environment. Augmentation using TTS-generated data for these specific OOV words can positively affect the robustness of the ASR model without significant degradation on the general dataset.

Kubis et al. [13] use synthesized data to study the impact of speech recognition errors on the performance of natural language understanding models. In [14] text-to-speech models are used in conjunction with an automatic speech recognition system to produce a dataset for improving the robustness of

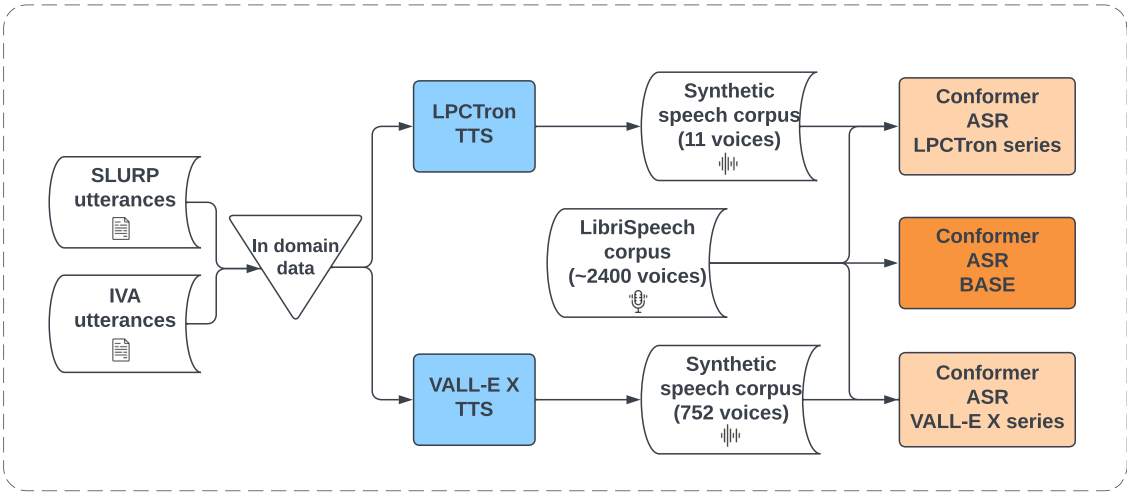

Fig. 1. Experimental workflow.

natural language understanding models to speech recognition errors.

Furthermore, synthetic data might be useful in ASR personalization [15]. The aforementioned study shows high effectives in ASR personalization using synthetic data, in particular when there are few recordings of a speaker in the dataset.

Previous works also addressed the problem of imperfections in data produced by TTS. Synthetic data differs from the real one in terms of naturalness and because of the presence of artifacts. Hu et al. [16] proposed two techniques for ASR training to alleviate the issues arising from the problems mentioned above. They observed up to 13% relative error reduction in ASR task.

The authors of VoiceBox [17] investigate the performance of ASR models trained on real and synthetic data. For training the ASR model on real data they use LibriSpeech 100h and 960h datasets. The synthetic data are generated from the texts collected in the LibriSpeech training set. The evaluation is performed with respect to *test-clean* and *test-other* subsets of LibriSpeech which do not contain conversational speech. Le et al. [17] show that their best performing TTS models lead to the absolute WER increase of 0.4% on *test-clean* and 1.7% on *test-other*, if compared to the models trained no real data. Contrary to [17], we investigate the impact of using voice-cloned speech on domain-specific adaptation of ASR in the conversational setting and use for this purpose datasets that contain conversational speech (SLURP and IVA).

III. Data

Measuring the impact of synthesized data on the performance of the ASR model requires careful selection of speech resources to be used for training and evaluation. We decided to use LibriSpeech [18] as a resource for training baseline ASR model and as a target corpus for augmentation. LibriSpeech is a corpus of approximately $1,000$ hours of read English speech, recorded by more than $2,400$ speakers. It is derived from the LibriVox project, which features audiobooks read by volunteers.

For training speech synthesizers we used LJ Speech Dataset [19] and Hi-Fi TTS Dataset [20]. LJSpeech is a dataset of about 24 hours of audio from a single speaker reading book passages, specifically from Project Gutenberg. Hi-Fi TTS Dataset is also based on Project Gutenberg texts and LibriVox audiobooks and contains about 292 hours of speech from 10 speakers with at least 17 hours per speaker. Both of these datasets were designed for training models for speech-based applications, with the main focus on speech synthesis.

We also utilize open-sourced VALL-E X model[1] that was trained on LibriTTS [21], AISHELL-1 [22], AISHELL-3 [23] and Japanese subset of CommonVoice dataset [24]. The authors also used some self-gathered data that was not described. In total they used about 704 hours of speech for English, 598 hours for Chinese and 437 hours for Japanese.

We evaluate ASR models using three general-purpose and two domain-specific ASR datasets. The general-purpose datasets include two test splits of LibriSpeech, *test-clean* and *test-other*. The *test-clean* split has higher quality of samples compared to *test-other* [18]. As a third general-purpose dataset, we use the test split of FLEURS [25] which provides natural speech recordings for many languages, out of which we use an English subset only.

As for the testsets in the domain of virtual assistants, we chose to use the test split of SLURP [26] and our internal virtual assistant (IVA) dataset. The SLURP testset has 13078 recordings totalling 10.3 hours of audio, while the IVA dataset contains 14094 recordings and 12.5 hours of speech. IVA has a broader set of domains and intents (55 and 223 respectively) compared to SLURP (18 and 94). Table I describes the language resources used for evaluation.

For prompting VALL-E X, we randomly chose one record-

[1] https://github.com/Plachtaa/VALL-E-X

TABLE I
RESOURCES USED FOR EVALUATION.

Dataset	Samples	Hours	Speakers
LS-clean	2620	5.4	40
LS-other	2939	5.1	33
FLEURS	647	1.8	–
SLURP	13078	10.3	–
IVA	14094	12.5	–

ing for each of the speakers. As sources of prompts we used LibriSpeech, HiFi TTS Dataset and LJ Speech Dataset described above and VCTK dataset [27] which contains high quality speech data recorded by 110 English speakers.

IV. Models

A. Speech Recognition

For our experiments we chose the Conformer on-device ASR model [28]. It is based on a RNN-Transducer architecture and has been commercialized on edge devices, which proves its high quality. This makes it a compelling target for our experiments on improving the ASR performance.

The model provides real time ASR performance on edge devices. Although the authors used two pass model for better quality, we limited ourselves to the first pass. Our main goal was to observe the difference between both augmentation approaches so we did not find improving ASR by ensembling relevant. In our single pass approach the transcription network encodes acoustic features of speech, while the predictor network acts as language model and tries to predict the next token based on the previous ones. These two, the acoustic features and language features are joined together in the joint network that outputs the final label.

B. Speech Synthesis

As a conventional neural approach to speech synthesis we decided to use a two-stage end-to-end TTS, consisting of an acoustic model mapping phonetic labels to acoustic features and a vocoder mapping these features to audio samples.

The set of phonetic labels contained symbols for phonemes, word delimiters and end of sentence marks (affirmative sentences, questions and exclamations). Acoustic features were derived from F0 (interpolated in unvoiced regions), mel-spectra and band-aperiodicity in a manner of the WORLD vocoder [29]. We utilized vocoder architecture that follows LPCNet [30] and an acoustic model based on Tacotron and [31] Tacotron2 [5], as described in [32]. For simplicity, later we refer to this system as a whole by the name LPCTron.

C. Voice Cloning

VALL-E X [8] is a zero-shot TTS that offers state of the art quality of cloning a sample voice, having only a 3-second recording of it. Instead of regarding speech synthesis as a continuous regression task, it adopts conditional language modelling approach, where the synthesis is conditioned on the input text and audio. It also ceases to use mel-spectogram in favor of acoustic tokens that are generated by neural codec LM.

The output speech is modeled at two stages with a total of 8 quantizers. In the first stage, the autoregressive language model generates codec codes of the first quantizer. During the second stage, the non-autoregressive language model generates codes for the rest of the quantizers, it is conditioned not on previously generated tokens but on all the tokens from previous quantizers. This makes the second stage much faster, because codes from previous quantizers are known at the start. The intention is that each next quantizer encodes the details that were not captured by previous ones.

The reason that VALL-E X is useful for our task is that it has in-context learning ability, which means that it can synthesize high-quality output on previously unseen inputs. While conventional neural TTS systems needed fine-tuning for unseen speakers, VALL-E X does not.

Open-source VALL-E X implementation follows the original paper [7] and uses G2P tool for converting the input sentence to phonemes and EnCodec [33] as a neural codec.

V. Experiments

The goal of our study is to investigate how does the multivoiceity of synthesized, domain-specific training data impact the performance of the resulting ASR model. For this purpose we conduct experiments with ASR models trained on speech recordings, speech recordings combined with data synthesized with LPCTron and speech recordings combined with data synthesized with VALL-E X.

For synthesis, we created a text corpus consisting of $129,000$ user commands directed to a task-oriented virtual assistant which includes $81,500$ utterances from our internal dataset, and $47,500$ utterances obtained in the process of augmenting the training split of the SLURP dataset.

The augmentation employed to enrich SLURP consisted of two steps. First, we used RoBERTa [34] and BART [35] models to randomly substitute words in the user commands with their counterparts supplied by the language models. Second, the sentences were transcribed from English to French, German, Italian and Spanish and backwards with the use of OPUS-MT models [36].

The text corpus was split into 3 equal parts and synthesized using both LPCTron and VALL-E X. For LPCTron we selected voices randomly from 11 available options and for VALL-E X from 752. The audio prompts for VALL-E X were collected from 4 datasets in a manner described in section III. The first part of the text corpus was synthesized with 2 voices per sentence, second part with 3 voices and the last part with 10 voices. This way we obtained three sets of 40 hours, 60 hours and 200 hours of synthesized speech. We combined these sets into splits: 40 hours, 60 hours, 100 hours 200 hours and 300 hours, which were later utilized for experiments.

We used 960 hour subset of LibriSpeech corpus for training along with splits of synthetic data. The *Lxxx* models combine LibriSpeech recordings with LPCTron synthesized dataset with *xxx* hours, e.g. *L060* used 60 hours split mentioned above. Analogically, the *Vxxx* models combine LibriSpeech data with *xxx* hours of spoken commands generated with the use of

TABLE II
LIBRISPEECH 960H ASR MODELS WER.

Dataset	BASE	L040	L060	L100	L200	L300	V040	V060	V100	V200	V300
LS-clean	8.08	7.88	7.62	7.91	8.07	7.80	7.97	9.60	10.74	8.29	8.10
LS-other	20.57	19.84	20.17	20.23	20.51	20.58	20.47	21.43	22.17	20.79	20.75
FLEURS	34.31	34.90	34.02	34.04	34.83	34.44	33.39	33.72	36.28	35.03	33.24
SLURP	74.89	70.02	69.22	68.37	69.56	68.83	66.67	64.64	65.56	63.39	62.58
IVA	75.14	66.82	64.75	62.13	64.09	62.54	50.62	54.01	52.91	47.82	44.61

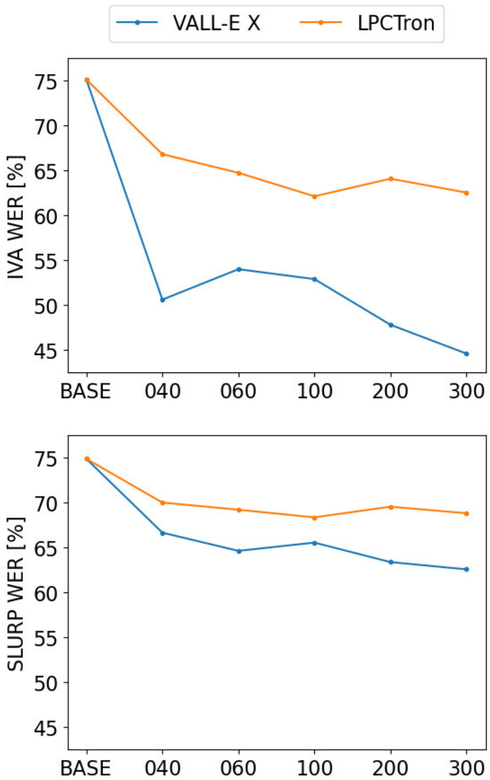

Fig. 2. WER obtained on SLURP and IVA.

VALL-E X model. The *BASE* model is a baseline trained on LibriSpecch 960h without addition of synthetic data.

The results presented in Table II indicate significant improvement in performance of the augmented models on domain-specific testsets (SLURP and IVA). We can also observe no significant performance drop on general-purpose test sets (LS-clean, LS-other and FLEURS) meaning that ASR models maintained generalization capability. The *V300* performs the best out of all trained models and results in absolute WER reduction, with regard to the *BASE*, of $30.53pp$ and $12.31pp$ in comparison to $12.60pp$ and $6.06pp$ obtained by *L300* on the IVA dataset and SLURP, respectively.

To investigate how the amount of synthetic data used for training impacts the ASR, we compared WER obtained using different data splits of IVA and SLURP. As shown in Figure 2, models trained with addition of VALL-E X data outperform their counterparts augmented with LPCTron data. There is also a noticeable improvement in WER with addition of more voice-cloned data, whereas the results plateau for models trained with the usage of LPCTron data.

To verify the quality of the audio data produced by VALL-E X and LPCTron we used Whisper [37] ASR model. We computed WER on the subset of 40 hours data. We got 37.55% and 20.38% WER on VALL-E X and LPCTron datasets, respectively.

VI. DISCUSSION

The choice of LPCTron as the baseline for conducting experiments can be questioned as there are several other more recent, conventional neural TTS models that can be used for the task. However, when comparing ratio between MOS for synthesized speech and MOS measured for ground truth across different architectures the results for LPCTron [32] 93% $(= 4.2/4.5)$ are on par with 89% $(= 3.83/4.3)$ achieved for FastSpeech2 [38], 98% $(= 4.36/4.45)$ for HiFiGAN [39] and 93% $(= 3.961/4.274)$ for WaveGlow [40]. Taking into account that HiFiGAN and WaveGlow are vocoders, not the full TTS systems, only FastSpeech2 would be a direct replacement for LPCTron in our experimental setting. Still, FastSpeech2 model presents similar quality to Tacotron2-based TTS models as shown in [38]. Furthermore, as we reported in Section V, the transcriptions of the audio samples produced by LPCTron obtained with the use of Whisper [37] had significantly lower WER than their VALL-E X counterparts. This shows that the quality of generated speech was higher in the case of LPCTron making our study sound, even if the LPCTron model is outperformed by some other conventional neural TTS model that can be potentially used as a baseline for experiments.

Taking into consideration that the compared TTS models are trained in a different manner with VALL-E X being trained for zero-shot (voice cloning) synthesis and LPCTron being trained for a conventional synthesis, there are differences in the model architecture that we cannot control in the experimental setting. However, it should be noted that although VALL-E X is a decoder-only model and Tacotron is an encoder-decoder model both of them are autoregressive, thus we do not consider the differences in the architecture to have a significant impact on the results.

Before VALL-E X, other approaches to zero-shot voice-cloning speech synthesis were considered. They were mainly based on providing the acoustic model with speaker embeddings extracted from speech sample with speaker verification models [41]. This approach still relies on the availability of high quality data for multiple speakers to train acoustic model to utilize speaker embedding space properly. On the other hand, conditional language modelling approach allows for utilizing lower quality data which makes it more suitable to our study.

VII. Conclusions

In this study we investigated the efficacy of using voice-cloned speech for augmenting spoken language with the goal of improving the performance of an ASR system. In this setting, we compared a baseline dataset that contains solely voice recordings, the dataset with addition of voice-cloned samples and the dataset expanded with samples synthesized by a conventional neural TTS system.

The conducted experiments show that the use of voice cloning to generate data with multiple voices and pronunciations improves the ASR performance significantly, compared to data from a conventional TTS speaking in just one or a few voices. The lower quality of voice-cloned speech, showed in terms of intelligibility, does not prevent the mentioned improvement.

We also showed that improvements gained by adding more synthetic data to the speech corpus plateau quickly for data generated using conventional neural TTS, but adding even 300 hours of synthetic speech generated using VALL-E X does not seem to saturate the results of ASR model.

One avenue for further research is to investigate upper limits of augmenting speech corpora using voice-cloned samples. Other dimension worth experimenting with is voice characteristics variability and its impact on the ASR results. There is also noticeable gap in quality of synthesized speech in terms of intelligibility between conventional neural TTS and LM-based TTS which should be decreased.

References

[1] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "Synthasr: Unlocking synthetic data for speech recognition," 2021.

[2] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Data augmentation for asr using tts via a discrete representation," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 68–75.

[3] N. Rossenbach, M. Zeineldeen, B. Hilmes, R. Schlüter, and H. Ney, "Comparing the benefit of synthetic training data for various automatic speech recognition architectures," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021, pp. 788–795.

[4] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," 2021.

[5] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, pp. 4779–4783.

[6] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," 2019.

[7] 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," 2023.

[8] Z. Zhang, L. Zhou, C. Wang *et al.*, "Speak foreign languages with your own voice: Cross-lingual neural codec language modeling," 2023.

[9] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," 2023.

[10] M. Bartelds, N. San, B. McDonnell *et al.*, "Making more of little data: Improving low-resource automatic speech recognition using data augmentation," in *Proc.Annual Meeting of the Association for Computational Linguistics*, 2023, pp. 715–729.

[11] N. Rossenbach, A. Zeyer, R. Schlüter, and H. Ney, "Generating synthetic audio data for attention-based speech recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7069–7073.

[12] X. Zheng, Y. Liu, D. Gunceler, and D. Willett, "Using synthetic audio to improve the recognition of out-of-vocabulary words in end-to-end asr systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5674–5678.

[13] M. Kubis, P. Skórzewski, M. Sowański, and T. Ziętkiewicz, "Back Transcription as a Method for Evaluating Robustness of Natural Language Understanding Models to Speech Recognition Errors," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, December 2023, pp. 11 824–11 835.

[14] ——, "Center for Artificial Intelligence Challenge on Conversational AI Correctness," in *Proceedings of the 18th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 35. IEEE, 2023, pp. 1319–1324.

[15] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[16] T.-Y. Hu, M. Armandpour, A. Shrivastava, J.-H. R. Chang, H. Koppula, and O. Tuzel, "Synt++: Utilizing imperfect synthetic data to improve speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 7682–7686.

[17] M. Le, A. Vyas, B. Shi *et al.*, "Voicebox: Text-guided multilingual universal speech generation at scale," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 14 005–14 034.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5206–5210.

[19] K. Ito and L. Johnson, "The lj speech dataset," 2017.

[20] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-Fi Multi-Speaker English TTS Dataset," in *Interspeech*, 2021, pp. 2776–2780.

[21] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *CoRR*, vol. abs/1904.02882, 2019.

[22] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*, 2017, p. Submitted.

[23] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "AISHELL-3: A multi-speaker mandarin TTS corpus and the baselines," *CoRR*, vol. abs/2010.11567, 2020.

[24] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proc. 12th Conference on Language Resources and Evaluation*, 2020, pp. 4211–4215.

[25] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," 2022.

[26] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser, "SLURP: A spoken language understanding resource package," in *Proc. 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 7252–7262.

[27] C. Veaux, J. Yamagishi, and K. MacDonald, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2019.

[28] J. Park, S. Jin, J. Park *et al.*, "Conformer-based on-device streaming speech recognition with kd compression and two-pass architecture," in *IEEE Spoken Language Technology Workshop*, 2023, pp. 92–99.

[29] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[30] J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6 kb/s using LPCNet," in *Interspeech*, 2019.

[31] Y. Wang, R. J. Skerry-Ryan, D. Stanton *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017.

[32] N. Ellinas, G. Vamvoukakis, K. Markopoulos *et al.*, "High quality streaming speech synthesis with low, sentence-length-independent latency," in *Interspeech*, 2020, pp. 2022–2026.

[33] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," 2022.
[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.
[35] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.
[36] J. Tiedemann and S. Thottingal, "OPUS-MT - Building open translation services for the World," in *Proc. 22nd Annual Conferenec of the European Association for Machine Translation*, 2020.
[37] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.
[38] 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," 2022.
[39] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.
[40] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019.
[41] Y. Jia, Y. Zhang, R. Weiss *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.