Interspeech 2024
1-5 September 2024, Kos, Greece

GLOBE: A High-quality English Corpus with Global Accents for Zero-shot Speaker Adaptive Text-to-Speech

Wenbin Wang[1], *Yang Song*[1], *Sanjay Jha*[1]

[1]School of Computer Science and Engineering, University of New South Wales, Australia

wenbin.wang@unsw.edu.au, yang.song1@unsw.edu.au, sanjay.jha@unsw.edu.au

Abstract

This paper introduces GLOBE, a high-quality English corpus with worldwide accents, specifically designed to address the limitations of current zero-shot speaker adaptive Text-to-Speech (TTS) systems that exhibit poor generalizability in adapting to speakers with accents. Compared to commonly used English corpora, such as LibriTTS and VCTK, GLOBE is unique in its inclusion of utterances from 23,519 speakers and covers 164 accents worldwide, along with detailed metadata for these speakers. Compared to its original corpus, i.e., Common Voice, GLOBE significantly improves the quality of the speech data through rigorous filtering and enhancement processes, while also populating all missing speaker metadata. The final curated GLOBE corpus includes 535 hours of speech data at a 24 kHz sampling rate. Our benchmark results indicate that the speaker adaptive TTS model trained on the GLOBE corpus can synthesize speech with better speaker similarity and comparable naturalness than that trained on other popular corpora. We will release GLOBE publicly after acceptance. The GLOBE dataset is available at `https://globecorpus.github.io/`.

Index Terms: dataset, text-to-speech, speaker adaptation

1. Introduction

Recent advances in deep learning have significantly propelled TTS research forward. The latest generation of neural networks-based TTS models [1, 2, 3, 4] can now generate highly lifelike human speech. This advancement has shifted the TTS research focus toward more sophisticated and challenging tasks [5]. Among these emerging tasks, speaker adaptive TTS [5, 6], also known as voice cloning, especially in zero-shot scenarios [7, 8, 9], has emerged as an active area of interest. *Zero-shot speaker adaptive TTS* allows TTS models to swiftly adapt to new speaker voices, which are not included in the training dataset, using only seconds of speech samples. This technique significantly broadens TTS technology's acceptability.

In our previous works [8, 10], we found a significant challenge in current zero-shot speaker adaptive TTS research is models' limited generalizability to accented voices. Despite increasing model parameters and enlarging the training dataset, this challenge persists [11, 9]. Our analysis identifies one of the crucial factors contributing to this issue: the prevalent English TTS datasets contain a limited set of accents. For example, LibriTTS [12] and LibriSpeech [13] mainly consists of speakers with US English during the filtering process, and the VCTK [14] dataset encompasses speakers with only 11 accents. The Common Voice dataset [1] [15], which comprises more than 3,000 hours of speech and covers up to 337 accents, presents a potential solution. However, it also exhibits several undesirable characteristics for building TTS system [16]: 1) A significant number of speech samples contain noticeable background or electromagnetic noise; 2) Despite the audio file has a sample rate of 48 kHz, the actual signal bandwidth is limited; 3) Many samples feature mispronunciations or corrections where speakers repeat unfamiliar words upon realizing a mispronunciation; 4) Half of the speakers have missing metadata and the accent labels are confusing.

To address these issues, we introduce GLOBE, a hi**G**h-quality eng**L**ish c**O**rpus with glo**B**al acc**E**nts, based on the Common Voice dataset. To construct this data, we remove low-quality, bandwidth-limited audio samples and re-align the utterance and text. Then, we manually cleaned the accent labels and populated missing speaker metadata through our prediction model. Compared to other popular English TTS datasets [14, 12, 15], the GLOBE dataset has its unique features:

High Speech Quality: The GLOBE dataset contains 535 hours of high-quality speech filtered from over 3,000 hours, with signal-to-noise ratio, signal bandwidth, and transcription accuracy. Our experimental results on zero-shot speaker adaptive TTS indicate that speech samples in GLOBE surpass those in VCTK and LibriTTS in terms of objective and subjective naturalness in mean opinion score evaluations.

Global Accent Coverage: With 23,519 speakers representing 164 different English accents from more than 50 countries, GLOBE offers unparalleled accent diversity. Our experiments demonstrate that such diversity significantly improves the generalizability of zero-shot speaker adaptive TTS models to different accents.

Extra Speaker Information: In addition to speech audio and corresponding text, GLOBE provides detailed metadata for all 23,519 speakers, including accent, age, and gender. This additional information enables future research of more personalized TTS models and mitigation of bias.

2. Relevant English Multi-speaker Corpus

VCTK [14]. The VCTK dataset is a widely utilized corpus for developing TTS and voice cloning systems. It contains 44 hours of 48 kHz speech data from 109 English speakers and each speaker reads about 400 sentences from newspapers. Moreover, the VCTK dataset includes labels for each speaker's accent and gender, covering a total of 11 accents.

[1] In this paper, the "Common Voice" dataset specifically refers to its Common Voice Corpus 14.0 English subset at https://commonvoice.mozilla.org/en/datasets. The original Common Voice dataset includes multiple languages and versions.

10.21437/Interspeech.2024-70

Table 1: *Statistics on GLOBE and Relevant English Multi-speaker Corpus*

Corpus	Total Hours	Total Speakers	Sample Rate	Total Accent	Speaker Info	License
CSTR VCTK [14]	44	109	48 kHz	11	Accent, Gender	CC BY 4.0 [17]
LibriTTS [12]	586	2,456	24 kHz	-	-	CC BY 4.0 [17]
LibriTTS-R [18]	585	2,456	24 kHz	-	-	CC BY 4.0 [17]
Common Voice [15]	3,347	88,904	48 kHz	337	Accent, Age, Gender	CC-0 1.0 [19]
GLOBE	535	23,519	24 kHz	164	Accent, Age, Gender	CC-0 1.0 [19]

LibriTTS [12]. The LibriTTS dataset is a well-known multi-speaker dataset for training speaker adaptive TTS systems. It derives from the LibriSpeech dataset [13] and includes 585 hours of audio recordings at a 24 kHz sampling rate, contributed by 2,456 speakers. This dataset specifically targets and mitigates various limitations in the original LibriSpeech collection, making it suitable for TTS system applications.

LibriTTS-R [18]. The LibriTTS-R corpus, an enhanced version of the LibriTTS dataset, significantly improves audio quality by incorporating speech restoration techniques. These enhancements support the training of high-quality TTS models. This corpus is the same as LibriTTS, retaining 585 hours of speech data from 2,456 speakers.

Common Voice [15]. Common Voice is a dataset powered by the voices of volunteer contributors from all around the world. It contains 3,347 hours of audio from 88,904 speakers, recorded at a 48kHz sample rate. The Common Voice dataset is frequently utilized to build automatic speech recognition systems. Another study [16] demonstrates that despite Common Voice's extensive collection of audio, it is not suitable for building TTS models due to the prevalence of poor-quality audio.

The key information for these datasets, along with GLOBE, is summarized in Table 1.

3. Data Processing Pipeline

3.1. Speech Sample Pre-processing and Filtering

During the initial construction phase of the GLOBE corpus, low-quality speech samples are identified and removed to prevent adverse impacts on the performance of TTS models when utilized as training data. The primary metric used for assessing speech sample quality is the signal-to-noise ratio (SNR), estimated through waveform amplitude distribution analysis [20], as a crucial indicator of audio clarity by quantifying the ratio of unwanted noise to identifiable speech. In line with [12], utterances with an SNR below 0 dB are excluded from the GLOBE corpus. Furthermore, the actual signal bandwidth is determined by identifying the highest frequency that is at least -50 dB below the power spectrogram's peak average, following the methodology outlined in [21]. Utterances with actual signal bandwidths below 12 kHz are excluded. Additionally, utterances containing more than 930 milliseconds of continuous internal silence, attributed to abnormal pauses or hesitations, are removed to further avoid negative impacts on the duration predictors within TTS models. The internal silence is detected by a voice activity detection (VAD) tool [2] and the selection of the 930-millisecond threshold is derived from the maximum duration of internal silence observed within the LibriTTS clean subset.

3.2. Speech Text Alignment

Given the significantly higher word error rate of the Common Voice dataset compared to other popular TTS datasets [14, 12],

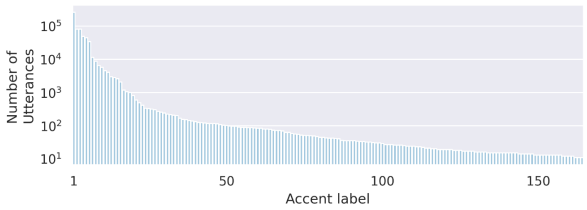

Figure 1: *Distribution of the number of utterances in different ground-truth accent labels in the GLOBE dataset.*

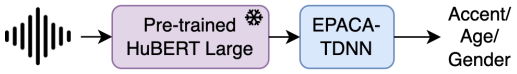

Figure 2: *Structure of speaker information prediction model, which is based on Hubert [24] and EPACA-TDNN [25].*

which seriously impacts the intelligibility of synthesized speech if trained on these data [16], the next phase in the development of the GLOBE corpus focuses on improving transcription accuracy. Following [12], Whisper [22] is firstly utilized to transcribe all utterances and a weighted finite-state transducer-based system [23] is then employed to normalize the corresponding text from the original dataset and the transcribed texts to their spoken forms. Following that, the word-level edit distance between them is computed by a publicly available toolkit [3]. Utterances exhibiting an edit distance greater than 1 are eliminated. Furthermore, clips that have an edit distance of 1 due to consecutive word repetitions are also discarded, as this often results from the repeated pronunciation of unfamiliar words according to our analysis.

3.3. Speaker Information Refinement and Speech Post-processing

The final development stage involves populating missing metadata for speakers including accent, age and gender, and applying speech sample post-processing. Initially, accents represented by fewer than five speakers or 10 utterances are merged with the most similar accent label or removed from the dataset, as we believe that such limited speech samples do not adequately capture the full scope of an accent's characteristics. Furthermore, meaningless accent labels, such as "not bad" and "A'lo," are also removed along with their corresponding speakers. Subsequently, for each metadata class, i.e., accent, age and gender, a training dataset, along with label-balanced validation and test sets, are developed from the refined speech data and speaker labels. Each training set includes utterances from 11,000 speakers, while each validation and test set features at least 1,000 speakers. Utilizing these subsets, three speaker information prediction models with the same structure, as depicted in Figure 2, are developed to predict the speaker's accent, age and

[2] https://github.com/snakers4/silero-vad

[3] https://github.com/roy-ht/editdistance

gender, respectively. Due to the long-tail distribution of the speaker's accent label, as shown in Figure 1, the square-root sampling method [26] is employed during model training to mitigate the negative impact. These models finally achieve accuracies of 97.22%, 99.55%, and 99.95% for accent, age, and gender prediction, respectively, across the test sets. The models are then leveraged to populate the missing speaker metadata. After that, post-processing is applied to all utterances, which includes the elimination of leading and trailing silences based on VAD results and the further suppression of the background noise through a speech enhancement tool [4].

4. Experiments

4.1. Experiments for Ground-Truth Speech Samples

4.1.1. Experimental setups

In the first experiment, we evaluated the quality of ground-truth speech samples within GLOBE and other popular English multi-speaker datasets [14, 27, 18, 15]. To objectively evaluate audio quality, we randomly selected 10,000 samples from the full set of VCTK [14], the training set of GLOBE and Common Voice [15] and the "train-clean" subsets of LibriTTS [27] and LibriTTS-R [18]. These subsets were selected because they represent the highest audio quality available in each dataset. For subjective evaluations, particularly the mean opinion score, we randomly chose 120 samples from those used in the objective evaluation for each dataset. The following evaluation metrics were utilized:

Naturalness Mean Opinion Score (NMOS). To evaluate the naturalness of speech samples, following [18, 28], we employed the Mean Opinion Score. Participants were asked to rate the naturalness of each utterance using a five-point Likert Scale [27]. Each speech sample was rated by five distinct participants, and we calculated the average score along with a 95% confidence interval by the official tool [5] for each evaluated dataset.

UTokyo-SaruLab Mean Opinion Score (UT-MOS) [29]. In line with prior studies [30, 31], predicted NMOS values were also provided for reference. The UT-MOS model was employed for NMOS prediction, which achieved state-of-the-art performance in 10 out of 16 metrics in the VoiceMOS Challenge [32].

Word Error Rate (WER). Following [18, 27], we employed the WER metric to measure the average misalignment in speech transcripts relative to the ground-truth text. A lower WER indicates more accurate alignment. Speech transcription was conducted using a pre-trained Conformer-based automatic speech recognition model [6] [33].

Speaker Embedding Cosine Similarity (SMCS). Consistent with [18, 7], speaker embedding cosine similarity was used to evaluate the similarity between two speeches from the same speaker. The speaker embeddings were extracted using the TitaNet-L speaker verification model [34], which achieves the state-of-the-art equal error rate on the VoxCeleb1 [35].

Speaker Embedding Vendi Score (SEVS). Considering that the number of speakers in a dataset does not necessarily reflect its speaker or accent diversities, we introduced the speaker embedding-based Vendi Score [36] to evaluate the diversity of accents contained in each dataset. The Vendi Score is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix and has been used in both computer vision [37] and natural language processing research [38].

[4] https://podcast.adobe.com/enhance

[5] https://github.com/luferrer/ConfidenceIntervals

[6] https://huggingface.co/nvidia/parakeet-rnnt-1.1b

Table 2: *Evaluation Results of GT Speech Samples*

Corpus	NMOS↑	UT-MOS↑	WER(%)↓	$SMCS$↑	SEVS↑
VCTK [14]	4.23 ± 0.06	4.01	**2.1**	0.887	71.16
LibriTTS [27]	4.20 ± 0.07	4.00	3.8	0.893	94.71
LibriTTS-R [18]	4.27 ± 0.07	**4.12**	3.8	0.895	94.18
Common Voice [15]	3.54 ± 0.10	3.68	7.3	0.887	**112.21**
GLOBE	4.25 ± 0.06	4.09	3.9	**0.903**	110.08

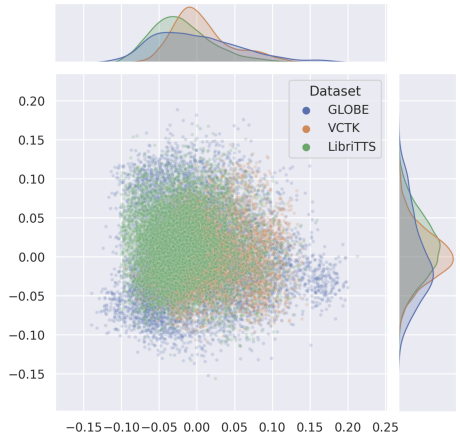

Figure 3: *Speaker embedding distributions and their marginal distributions KDE of different datasets after dimension reduction via PCA.*

4.1.2. Experimental results

Table 2 presents the experiment results for ground-truth speech samples. Concerning speech naturalness, all datasets, except for the Common Voice dataset, which recorded a lower NMOS of 3.54 due to prevalent low-quality audio, achieved similar NMOS scores. We also conducted the Mann–Whitney–Wilcoxon (MWW) test [39] to determine whether there are statistically significant differences in NMOS scores between any two datasets. It was found that the p-value between LibriTTS and GLOBE was $0.91 > 0.05$, indicating that the speech samples from GLOBE do not show a statistically significant difference from LibriTTS in naturalness. In contrast, the p-value between Common Voice and GLOBE is $4.3e^{-6} < 0.05$, denoting that the naturalness of speech from GLOBE is statistically better than that of Common Voice. These findings are also corroborated by UT-MOS results. Regarding the WER, the VCTK dataset had the lowest WER, with LibriTTS, LibriTTS-R, and GLOBE showing comparably low WER levels, contrasting with Common Voice's higher WER. For SMCS, all datasets displayed similar scores, suggesting well-defined speaker characteristics within each dataset, given that TitaNet-L's threshold for determining utterances from the same speaker is 0.7. SEVS results indicate that LibriTTS and LibriTTS-R have wider speaker diversity than VCTK, with both Common Voice and GLOBE also showing improvements in speaker diversity compared to LibriTTS and LibriTTS-R. In summary, these results demonstrate that GLOBE achieves speech naturalness comparable to other popular TTS datasets by filtering out low-quality speech samples from Common Voice while maintaining richer speaker diversity compared to other datasets. We also visualized the speaker embedding distributions and their marginal distributions' kernel density estimates (KDE) of the GLOBE, VCTK, and LibriTTS datasets, as shown in Figure 3. To do this, we extracted 10,000 speaker embeddings from the

Table 3: *Evaluation Results of the modified YourTTS Trained on Different Datasets.*

Evaluation Corpus	$\text{LibriTTS}_{\text{test}}$					$\text{GLOBE}_{\text{test}}$				
Metrics	NMOS↑	SMOS↑	SMCS↑	WER(%)↓	UT-MOS↑	NMOS↑	SMOS↑	SMCS↑	WER(%)↓	UT-MOS↑
Ground Truth	4.16 ± 0.06	4.02 ± 0.05	0.772	3.82	3.93	4.19 ± 0.06	4.08 ± 0.08	0.774	3.91	4.07
VCTK [14]	3.91 ± 0.07	3.72 ± 0.08	0.716	6.4	3.76	3.89 ± 0.08	3.52 ± 0.09	0.698	6.6	3.78
LibriTTS [27]	4.03 ± 0.07	3.83 ± 0.07	**0.740**	**6.2**	3.88	4.01 ± 0.07	3.63 ± 0.09	0.715	6.4	3.83
LibriTTS-R [18]	4.09 ± 0.07	3.80 ± 0.08	0.736	6.3	**3.91**	4.02 ± 0.06	3.62 ± 0.09	0.712	6.4	3.85
Common Voice [15]	2.98 ± 0.07	3.79 ± 0.09	0.733	10.6	2.86	3.07 ± 0.07	3.72 ± 0.08	0.726	10.3	2.97
GLOBE	4.03 ± 0.06	3.84 ± 0.07	0.738	6.3	3.85	4.05 ± 0.08	3.81 ± 0.08	**0.732**	**6.3**	**3.86**

samples that were utilized for objective evaluation and reduced the dimension of all speaker embeddings to 2 via principal component analysis (PCA). As illustrated in the figure, the GLOBE dataset, represented in blue, displays a broader distribution of speaker embeddings across both dimensions compared to the other datasets, indicating a wider variety of speaker characteristics contained within the GLOBE corpus.

4.2. Experiments for Speaker Adaptive TTS Synthesized Samples

4.2.1. Model details

To investigate the influence of training datasets on the synthesized speech quality of zero-shot speaker-adaptive TTS models, we employed YourTTS [7], a widely used zero-shot speaker-adaptive TTS approach, as the baseline model with three modifications: firstly, the language embedding was removed from the model, given our focus solely on English; secondly, to thoroughly assess the datasets' influence on model performance and avoid bias introduced by pre-trained models, we replaced the pre-trained speaker encoder with a trainable encoder, i.e., EPACA-TDNN [25]; thirdly, some model parameters were adjusted to facilitate training with 24 kHz audio data.

4.2.2. Experimental setups

We trained the modified YourTTS models on all datasets outlined in Section 2. For the VCTK dataset, training encompassed the entire dataset. For both the LibriTTS and LibriTTS-R datasets, training was performed on the "train-clean" and "train-other" subsets. In terms of the Common Voice and GLOBE datasets, models were trained on the training subset. Throughout the training phase, we downsampled all speech samples to a 24 kHz sampling rate. Both phoneme sequences used for training and evaluation were generated from the ground-truth text using Phonemizer [40]. Training for both models was executed end-to-end and each training session was conducted on two NVIDIA V100 GPUs. The YourTTS model was trained for $1.8m$ iterations with a total batch size of 48. All training utilized AdamW optimizer [41], featuring $\beta_1 = 0.8$, $\beta_2 = 0.99$, and a weight decay parameter of 0.01. The initial learning rate was set at 2×10^{-4} and followed a decay factor of $\gamma = 0.9999$.

For evaluation, we employed several metrics introduced in Section 4.1.1. Specifically, we assessed the naturalness of the synthesized speech for each model using NMOS and UTMOS. The intelligibility of the synthesized speech was evaluated using WER. Additionally, SMCS was used to measure the similarity between the synthesized speech and the ground-truth speech. Furthermore, we introduced an additional evaluation metric: **Speaker Similarity Mean Opinion Score (SMOS) [28].** Parallel to the NMOS, this metric evaluates the speaker similarity between the synthesized utterance and a random utterance from the same speaker. Assessments were conducted on a five-point Likert Scale [27] by the same participants of the NMOS. The mean score and confidence interval were also calculated.

4.2.3. Experimental results

Table 3 presents the evaluation results for the modified YourTTS model trained on different datasets. When evaluated on the LibriTTS test set, the model trained on LibriTTS-R achieved the highest NMOS, closely followed by models trained on GLOBE and LibriTTS datasets, which exhibited no statistically significant difference in NMOS, as indicated by a p-value > 0.05 in the MWW test. The model trained on the Common Voice dataset underperformed, exhibiting a statistically significant decrease in NMOS, with the p-value in the MWW test at $3.2e^{-4}$, which is < 0.05. In terms of SMOS, models trained on LibriTTS, LibriTTS-R, Common Voice, and GLOBE yielded comparable outcomes. However, the model trained on VCTK performed slightly worse. This discrepancy can be attributed to the dataset's limited speaker diversity, which also adversely affected the SMCS scores. In the evaluation of the GLOBE test set, NMOS results mirrored those obtained on the LibriTTS test set. However, a significant decline in SMOS was observed across all models, attributed to the GLOBE set's broader diversity of speaker accents, presenting a notable challenge to the models' generalization capabilities. Specifically, the SMOS for the baseline model trained on LibriTTS dropped by 0.20 compared to its performance on the LibriTTS test set, with statistically significant p-values at $9e^{-4} < 0.05$ from the MWW test. In contrast, models trained on GLOBE exhibited the smallest decline in speaker similarity, and the MWW test indicated that these declines were not statistically significant, with p-values of $0.54 > 0.05$, demonstrating that the broad accent coverage by GLOBE enhanced TTS models' adaptability to diverse accents. In summary, these results demonstrate that the speaker-adaptive TTS model trained on GLOBE exhibits better generalization compared to models trained on other TTS datasets, enabling it to effectively adapt to various accents.

5. Conclusions

This paper introduces GLOBE, a high-quality English corpus featuring worldwide accents originating from Common Voice, aimed at addressing the poor generalizability issue of current zero-shot speaker-adaptive TTS models. GLOBE not only matches the audio quality of popular TTS datasets like LibriTTS but also surpasses them by covering a broader range of worldwide accents and offering metadata for an extensive array of over 20,000 speakers. Our experiments demonstrate that speaker-adaptive TTS models trained on GLOBE achieve better generalizability than those trained on other datasets. We hope that the release of GLOBE will contribute to advancements in TTS research.

6. References

[1] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," in Proc. of ICLR, 2021.

[2] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. A. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," in Proc. of ICML, vol. 139, 2021, pp. 8599–8608.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in Proc. of ICML, vol. 139, 2021, pp. 5530–5540.

[4] Z. Ye, W. Xue, X. Tan, J. Chen, Q. Liu, and Y. Guo, "Comospeech: One-step speech and singing voice synthesis via consistency model," in Proc. of ACM MM, 2023, pp. 1831–1839.

[5] X. Tan, T. Qin, F. K. Soong, and T. Liu, "A survey on neural speech synthesis," CoRR, vol. abs/2106.15561, 2021.

[6] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T. Liu, "Adaspeech: Adaptive text to speech for custom voice," in Proc. of ICLR, 2021.

[7] E. Casanova, J. Weber, C. D. Shulby, A. C. Júnior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone," in Proc. of ICML, vol. 162, 2022, pp. 2709–2720.

[8] W. Wang, Y. Song, and S. Jha, "Generalizable zero-shot speaker adaptive speech synthesis with disentangled representations," in Proc. of INTERSPEECH, 2023, pp. 4454–4458.

[9] Z. Jiang, Y. Ren, Z. Ye, J. Liu, C. Zhang, Q. Yang, S. Ji, R. Huang, C. Wang, X. Yin, Z. Ma, and Z. Zhao, "Mega-tts: Zero-shot text-to-speech at scale with intrinsic inductive bias," CoRR, vol. abs/2306.03509, 2023.

[10] W. Wang, Y. Song, and S. K. Jha, "USAT: A universal speaker-adaptive text-to-speech approach," IEEE ACM Trans. Audio Speech Lang. Process., vol. 32, pp. 2590–2604, 2024.

[11] 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," CoRR, vol. abs/2301.02111, 2023.

[12] 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," in Proc. of INTERSPEECH, 2019, pp. 1526–1530.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in Proc. of ICASSP, 2015, pp. 5206–5210.

[14] V. Christophe, Y. Junichi, M. Kirsten, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," in University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2016.

[15] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in Proc. of LREC, 2020, pp. 4218–4222.

[16] S. Ogun, V. Colotte, and E. Vincent, "Can we use common voice to train a multi-speaker TTS system?" in Proc. of IEEE SLT, 2022, pp. 900–905.

[17] "Creative commons attribution 4.0 license (cc-by 4.0)," https://creativecommons.org/licenses/by/4.0/.

[18] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, "Libritts-r: A restored multi-speaker text-to-speech corpus," CoRR, vol. abs/2305.18802, 2023.

[19] "Cc0 1.0 universal public domain dedication," https://creativecommons.org/publicdomain/zero/1.0/.

[20] C. Kim and R. M. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in Proc. of INTERSPEECH, 2008.

[21] E. Bakhturina, V. Lavrukhin, B. Ginsburg, and Y. Zhang, "Hi-fi multi-speaker english TTS dataset," in Proc. of INTERSPEECH, 2021, pp. 2776–2780.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in Proc. of ICML, 2023.

[23] Y. Zhang, E. Bakhturina, and B. Ginsburg, "Nemo (inverse) text normalization: From development to production," in Proc. of INTERSPEECH, 2021, pp. 4857–4859.

[24] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," IEEE ACM Trans. Audio Speech Lang. Process., vol. 29, pp. 3451–3460, 2021.

[25] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in Proc. of INTERSPEECH, 2020, pp. 3830–3834.

[26] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in Proc. of ICLR, 2020.

[27] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-stylespeech : Multi-speaker adaptive text-to-speech generation," in Proc. of ICML, vol. 139, 2021, pp. 7748–7759.

[28] S. Ö. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in Proc. of NeurIPS, 2018, pp. 10 040–10 050.

[29] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: utokyo-sarulab system for voicemos challenge 2022," in Proc. of INTERSPEECH, 2022, pp. 4521–4525.

[30] S. Huang, C. Lin, D. Liu, Y. Chen, and H. Lee, "Meta-tts: Meta-learning for few-shot speaker adaptive text-to-speech," IEEE ACM Trans. Audio Speech Lang. Process., vol. 30, pp. 1558–1571, 2022.

[31] J. Xue, Y. Deng, Y. Han, Y. Li, J. Sun, and J. Liang, "ECAPA-TDNN for multi-speaker text-to-speech synthesis," in Proc. of ISCSLP, 2022, pp. 230–234.

[32] W. Huang, E. Cooper, Y. Tsao, H. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2022," in Proc. of INTERSPEECH, 2022, pp. 4536–4540.

[33] D. Rekesh, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, A. Kumar, and B. Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," CoRR, vol. abs/2305.05084, 2023.

[34] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in Proc. of ICASSP, 2022.

[35] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in Proc. of Interspeech, 2017.

[36] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," Transactions on Machine Learning Research, 2023.

[37] R. Burgert, X. Li, A. Leite, K. Ranasinghe, and M. S. Ryoo, "Diffusion illusions: Hiding images in plain sight," CoRR, vol. abs/2312.03817, 2023.

[38] Y. Chen, B. Xu, Q. Wang, Y. Liu, and Z. Mao, "Benchmarking large language models on controllable generation under diversified instructions," CoRR, vol. abs/2401.00690, 2024.

[39] F. Wilcoxon, "Individual comparisons by ranking methods," in Breakthroughs in Statistics: Methodology and Distribution, 1992, pp. 196–202.

[40] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," Journal of Open Source Software, vol. 6, no. 68, p. 3958, 2021.

[41] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. of ICLR, 2019.