Investigating the Role of Synthetic Data Augmentation and Training Strategies on Improving Low-Resource Language ASR

Yun Hao, Reihaneh Amooie, Wietse de Vries, Rik van Noord, Martijn Wieling
Center for Language and Cognition Groningen (CLCG)
University of Groningen
Groningen, The Netherlands
{yun.hao, r.amooie, wietse.de.vries, r.i.k.van.noord, m.b.wieling}@rug.nl

Abstract

Low-resource automatic speech recognition (ASR) is challenging due to a scarcity of annotated data. While synthetic data from text-to-speech (TTS) systems can augment ASR training, its efficacy for low-resource languages remains unclear. In this study, we investigate under which conditions TTS-based data augmentation is most effective for low-resource languages. Experiments on six low-resource languages in Common Voice show that synthetic data is most beneficial under extremely low-resource ASR conditions (i.e., less than one hour of available real speech data), or for languages with larger amounts of TTS data (i.e., more than 10 hours). Additionally, increasing the amount and diversity of synthetic data while keeping an appropriate ratio of synthetic-to-real data can further improve ASR performance.

Keywords: speech recognition, low-resource languages, data augmentation, text-to-speech

1. Introduction

Over the past few years, automatic speech recognition (ASR) systems have made remarkable progress, achieving significant improvements in accuracy and robustness. In particular, the development of multilingual self-supervised or weakly-supervised speech models such as XLS-R (Babu et al., 2022) has greatly advanced speech processing tasks for low-resource languages. However, ASR performance remains low for languages with very limited transcribed audio available for fine-tuning.

One widely adopted strategy to deal with data scarcity is using data augmentation techniques. Traditional augmentation techniques for speech data are typically designed to create speech samples by modifying speech in the time and frequency domains, such as speed and tempo perturbation (Ko et al., 2015), noise addition, and SpecAugmentation (Park et al., 2019), which warps the spectral features in frequency and time blocks. With the advancement of deep learning-based speech generation technologies, such as text-to-speech (TTS) and voice conversion (VC), these methods have been adopted to generate additional training data for ASR (Li et al., 2018; Rosenberg et al., 2019; Rossenbach et al., 2020; Du and Yu, 2020; Baas and Kamper, 2022). Unlike traditional acoustic augmentation, data augmentation using TTS can synthesize speech with novel content or speaker identities, thus expanding data diversity to potentially improve the robustness of ASR.

Text-to-speech systems aim to synthesize speech from textual input with close-to-human-level naturalness and clarity. Conventional TTS systems are typically designed for single-speaker voice synthesis, but recent advancements in neural TTS architectures have enabled high-quality multi-speaker voice generation. These state-of-the-art systems can produce natural speech for speakers not present in the training data using merely a few seconds of reference audio, which is referred to as zero-shot multi-speaker TTS (ZS-TTS) (Jia et al., 2018; Casanova et al., 2022, 2024). Some ZS-TTS models, such as YourTTS (Casanova et al., 2022), XTTS (Casanova et al., 2024), and CosyVoice (Du et al., 2024), leverage large-scale multilingual pre-training, making them well-suited for adaptation to low-resource languages.

Related work Many studies have demonstrated that speech synthesis can improve ASR performance, with early research primarily focusing on English as the target language (Li et al., 2018; Rosenberg et al., 2019; Rossenbach et al., 2020). While some investigations have attempted to simulate low-resource conditions using reduced English datasets (Laptev et al., 2020; Du and Yu, 2020), the generalizability of these findings to actual low-resource languages remains questionable. Unlike English, many low-resource languages lack sufficient data to train TTS systems. Moreover, for multilingual pre-trained models, the proportion of training data of each language during pre-training also impacts ASR performance (Rouditchenko et al., 2023), which makes low-resource conditions simulated using English not well comparable to actual low-resource languages.

A few recent studies have explored synthetic data augmentation in actual low-resource languages. Casanova et al. (2023) demonstrated that YourTTS-based augmentation improved ASR performance for low-resource Russian and Portuguese. However, their study showed the improvements solely

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5633–5639
11–16 May 2026.

on single-speaker datasets, leaving its effectiveness for multi-speaker corpora such as Common Voice uncertain. Yang et al. (2025) enhanced low-resource ASR performance using CosyVoice for data augmentation, but their datasets each contained at least 200 hours of speech, limiting the applicability of their findings to languages with fewer resources.

For research on low-resource languages with less than 10 hours of real training data and multiple speakers, Baas and Kamper (2022) found that VC-based augmentation improved ASR performance in four low-resource South African languages in extremely low-resource settings (i.e., 10 minutes of real speech data), but observed that there was no benefit when there was more real speech data. Similarly, Bartelds et al. (2023) found that combining 168 minutes of synthetic speech generated by a FastSpeech2-based TTS system with 24 minutes of real data improved the performance of ASR for Gronings, but that adding more synthetic data only yielded marginal improvements.

These studies suggest that synthetic data augmentation is promising in extremely low-resource settings, but its effectiveness across varying levels of resource scarcity and diverse languages remains underexplored. Additionally, as the data required for training TTS and ASR models differ regarding the minimal numbers of speaker, the available data per speaker and level of noise (Cooper et al., 2017; Ogun et al., 2023), the available amount of training data for both TTS and ASR in a given language could also substantially impact the effectiveness of data augmentation.

Contributions We aim to investigate the effectiveness of fine-tuning a multilingual model to generate synthetic data for low-resource languages. Specifically, we leverage the capability of the ZS-TTS model to synthesize speech for unseen speakers in the training data. We use the Common Voice dataset for training both TTS and ASR models for six low-resource languages: Estonian, Frisian, Romanian, Swedish, Uyghur, and Welsh. We fine-tune the multilingual XTTS-v2 model to generate synthetic speech and assess its impact on fine-tuning XLS-R for ASR. Specifically, we examine how synthetic data influences ASR performance across different languages, levels of data scarcity and varying ratios of synthetic to real data. Our main contributions are as follows:

1. We develop TTS systems and systematically evaluate the effectiveness of TTS-based data augmentation for ASR under truly low-resource conditions, rather than relying on simulated low-resource settings based on high-resource languages;
2. We show through a detailed analysis how ASR performance is influenced by the varying levels of resource scarcity when using TTS-based data augmentation;
3. We explore different training strategies for incorporating synthetic data, revealing how the balance between the synthetic and real data, as well as the diversity of synthetic data, affects ASR performance.

2. Methodology

The overall pipeline of the proposed data augmentation method is shown in Figure 1.

2.1. Dataset

We selected six low-resource languages from Common Voice 17.0 (Ardila et al., 2020): Estonian (et), Frisian (fy), Romanian (ro), Swedish (sv), Uyghur (ug), and Welsh (cy). The Common Voice dataset is a publicly available multilingual corpus of crowd-sourced speech clips with text transcripts, recorded by volunteers, validated through community voting, and subsequently split into official train / development / test splits. For ASR training, we adopted the official splits as train / development / test set for each language. To evaluate the impact of real data size on the effectiveness of synthetic data augmentation, we randomly sampled training subsets of 30 minutes, 1 hour, 2 hours, 3 hours, 4 hours, and 5 hours from the original training splits, simulating low-to-medium resource scenarios.

For TTS model training, which ideally requires as many speech samples as possible from each speaker, we selected speakers from the *full validated set* of each language whose total speech duration exceeded 20 minutes (following Ogun et al., 2023), while strictly excluding any speakers overlapping with the ASR validation or test splits to prevent data contamination. Due to variations in speaker composition across languages, the selected data for each language ranges from 4.2 to 146.4 hours. This provides an opportunity for us to examine the impact of TTS training data size on the quality of synthetic speech and its effect on ASR data augmentation. The detailed statistics of our ASR and TTS training datasets are summarized in Table 1.

2.2. Text-to-speech model: XTTS-v2

We fine-tuned the XTTS-v2 model to obtain a TTS model for each language. XTTS-v2 (Casanova et al., 2024) is a multilingual ZS-TTS model based on Tortoise (Betker, 2023) and includes modifications to enable multilingual training, enhanced ZS-TTS performance, while also allowing for faster training and inference. The multilingual model is trained on a total of 27,282 hours of

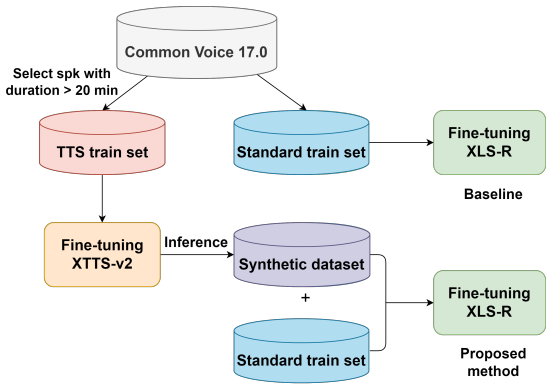

Figure 1: Pipeline of our baseline and proposed TTS-augmented ASR training method.

Table 1: Statistics of ASR and TTS training data for low-resource languages selected from Common Voice 17.0.

Language	Task	Duration (h)	# Speakers
Estonian	ASR	5.97	267
	TTS	4.21	5
Frisian	ASR	5.53	195
	TTS	36.90	38
Romanian	ASR	5.65	11
	TTS	7.59	8
Swedish	ASR	8.58	25
	TTS	26.92	16
Uyghur	ASR	15.74	168
	TTS	146.43	114
Welsh	ASR	11.44	39
	TTS	58.53	62

data in 16 languages and is publicly available.[1] For each language adaptation, we fine-tuned for five epochs with an effective batch size of 32, weight decay of 1e-2, and learning rate of 5e-6. Our scripts for TTS data splitting, training, and inference are available at: https://github.com/haoyunlf/LREC2026_augment_low_resource_ASR_with_TTS.

2.3. Synthetic data generation

For each language, every sentence in the training set was paired with a randomly selected speaker from the same set (excluding the original speaker), which served as the target voice for TTS inference. Here, we leverage the zero-shot TTS capability to synthesize speech for unseen speakers. This resulted in a synthetic training dataset with the same number of sentences as the original training set but spoken by different speakers.

For Frisian, we generated additional synthetic data to further investigate the effect of increasing the amount and diversity of synthetic data relative to real data. To do so, we use Frisian news texts from the Leipzig Corpora (Goldhahn et al., 2012), selecting sentences containing 4 to 16 words. The synthetic speech is then generated by randomly selecting a target voice from the train split of Common Voice Frisian dataset for each utterance. In the experiments described in Section 3.3, the additional synthetic speech required beyond the Frisian Common Voice dataset was randomly drawn from this generated dataset.

2.4. Automatic speech recognition model: XLS-R

We adopted the XLS-R 1B model[2] as our pre-trained base model. XLS-R (Babu et al., 2022) is a large-scale cross-lingual self-supervised speech model, which was pre-trained on approximately 436,000 hours of speech data spanning 128 languages. It builds on the wav2vec 2.0 (Baevski et al., 2020) structure, containing a convolutional feature encoder and a stack of Transformer-based (Vaswani et al., 2017) contextual encoder blocks. During pre-training, the model is optimized by solving a contrastive task over quantized masked latent audio representations. The extensive cross-lingual data enables XLS-R to excel across various languages and dialects, including those with limited resources (Rouditchenko et al., 2023).

For fine-tuning, we froze the feature extractor and only updated the Transformer encoder layers using a Connectionist Temporal Classification (CTC) loss. We fine-tuned each model with an effective batch size of 64, weight decay of 5e-3, and learning rate of 5e-5. To ensure the comparability of all models, we trained each model for 2,000 steps.

3. Experiments and Results

3.1. Validating synthetic data quality

We conducted two experiments to validate the performance of our TTS model on the six low-resource languages of interest. First, to assess the intelligibility of the synthetic data, we trained an ASR model following the method in Section 2.4 using five hours of real speech data for each language, and measured the word error rates (WERs) of the generated speech using the models. Next, to assess the potential of using synthetic data for improving ASR, we trained an ASR model for each language exclusively using five hours of the synthetic data, and evaluated the resulted models using the real

[1] https://github.com/coqui-ai/TTS

[2] https://huggingface.co/facebook/wav2vec2-xls-r-1b

test set of each language. These results are shown in Table 2.

Table 2: WERs (%) of ASR models trained with either only real or only synthetic data, evaluated on real and synthetic test sets. The percentages in parentheses indicate the relative WER increase compared to the baseline case of training and testing on real speech (*Train: real* → *Real test*). Lower WER indicates better ASR performance.

	Train: real		Train: synth
Lang	Real test	Synth test	Real test
et	21.1	68.7 (↑ 226%)	46.5 (↑ 120%)
fy	14.5	21.2 (↑ 46%)	22.8 (↑ 57%)
ro	10.3	35.6 (↑ 246%)	15.2 (↑ 48%)
sv	20.7	35.2 (↑ 70%)	28.6 (↑ 38%)
ug	30.1	31.8 (↑ 6%)	39.2 (↑ 30%)
cy	26.3	31.7 (↑ 21%)	32.4 (↑ 23%)

Compared to real test data, the WERs of synthetic data increase to varying degrees for different languages. Among them, Uyghur's synthetic data is the closest to real data, followed by Welsh, Frisian, and Swedish, while Estonian and Romanian show the worst WERs, with a relative increase of over 200% compared to real data. This discrepancy can be largely attributed to the amount of training data available for the TTS systems. Notably, Estonian and Romanian had the least training data available for the TTS system (under 6 hours), while Uyghur had the most, with about 146 hours from 114 speakers (see Table 1).

Although the quality of TTS-generated data varies, the ASR models trained on them exhibit relatively stable performance. For lower-quality synthetic data, such as Estonian and Romanian, the resulting ASR models still achieve better WER than the intelligibility of the synthetic speech itself (et: 46.5% < 68.7%; ro: 15.2% < 35.6%). Models trained on higher-quality synthetic data perform closer to those trained on real data, even though a performance gap remains. Interestingly, these findings suggest that while TTS quality does influence ASR performance, its impact is not as substantial as one might expect. In case of extreme data scarcity, even imperfect TTS-generated data can serve as a useful resource for ASR training.

3.2. Effect of available real data size

In this experiment, we evaluate the performance of synthetic data augmentation on the six languages. For each language, we fine-tuned XLS-R using $\{0.5, 1, 2, 3, 4, 5\}$ hours of real data and compared its performance against a setting where the same amount of synthetic data was added to the real data. Note that the number of steps remained the same to ensure a fair comparison.

Table 3: WERs (%) of ASR models trained with real data and synthetic augmented data for each language. Green numbers indicate WER improvements due to data augmentation, while red numbers represent WER degradation. The durations shown denote real training data only; in the *Real + Synt* setting, an equal amount of synthetic data is added (i.e., the total training data is doubled).

		Durations of real training data					
Lang	Train	30 min	1 h	2 h	3 h	4 h	5 h
et	Real	34.9	28.9	26.3	22.7	22.6	21.1
	Real + Synt	**-1.0**	**+0.9**	**+0.3**	**+1.7**	**+0.7**	**+2.0**
fy	Real	29.2	23.6	18.5	16.5	15.5	14.5
	Real + Synt	**-4.0**	**-2.4**	**-0.8**	**-0.2**	**+0.1**	**+0.1**
ro	Real	15.9	13.8	11.8	11.2	10.6	10.3
	Real + Synt	**-0.6**	**-0.1**	**+0.4**	**+0.2**	**+0.6**	**+0.5**
sv	Real	34.8	28.3	24.7	22.7	21.7	20.7
	Real + Synt	**-4.1**	**-1.6**	**-0.8**	**+1.0**	**+0.2**	**+1.1**
ug	Real	47.3	41.7	36.3	33.0	30.7	30.1
	Real + Synt	**-4.9**	**-2.8**	**-2.6**	**-2.2**	**-0.3**	**-0.5**
cy	Real	40.1	33.3	29.7	26.0	26.9	26.3
	Real + Synt	**-1.8**	**-1.1**	**-0.2**	**+0.3**	**-1.1**	**-1.3**

Our results indicate that when the amount of real training data is limited to 30 minutes, data augmentation consistently improves ASR performance across all languages, but as the data increases, its effect either diminishes or even harms performance. For Estonian, synthetic data augmentation ceases to be effective once real data exceeds 1 hour, while for Romanian the benefit diminishes beyond 2 hours. In contrast, for Uyghur and Welsh, the positive effect of augmentation remains relatively stable even with up to 5 hours of real data.

This pattern mirrors the trend observed in Section 3.1, where we found that the quality of synthetic data varies across languages, depending on the availability of TTS training data. Together, these findings suggest that while synthetic data can help improve ASR performance in cases of extreme data scarcity, the effectiveness of synthetic data augmentation is influenced by the balance between TTS and ASR resources for each language. For languages with abundant TTS resources, synthetic data can substantially enhance ASR performance, even for increasing amounts of real ASR training data. However, when TTS resources are limited, the impact of synthetic data diminishes quickly, especially as the amount of real data increases.

3.3. Effect of increasing amount of synthetic data

To further investigate the impact of increasing the amount of synthetic data, we conducted additional

experiments on Frisian. Specifically, we kept the amount of real training data fixed at one hour and investigated the impact of increasing synthetic data in two different settings.

Setting 1: We fixed the amount of real training data to one hour and varied the synthetic data from 1 to 10 hours, shifting the synthetic-to-real ratio from 1:1 to 10:1, thereby progressively increasing the proportion of synthetic data within the combined training set.

Setting 2: We increased the ratio of synthetic to real data from 1 to 10, in line with Setting 1. However, in this setting, the model never encounters each single synthetic sample more than once. For example, at a synthetic-to-real ratio of 1:1, instead of training for three epochs on one hour of synthetic data as in Setting 1, we train for one epoch on three hours of synthetic data (illustrated in Figure 2). The intuition is straightforward: the model potentially benefits from a larger amount of unique synthetic data, while the ratio of synthetic versus real data encountered in one epoch remains fixed.

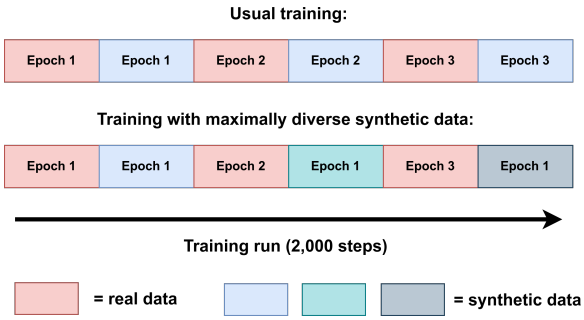

Figure 2: Illustration of the training procedure in Setting 1 (top part) vs. Setting 2 (bottom part), where each synthetic training example is seen only once. The visualization illustrates the situation in which the synthetic-to-real ratio is 1:1, with the same principle applying to other ratios. The label "Epoch 1" for synthetic data in the bottom part indicates that each synthetic example is encountered only once per experiment.

Figure 3 illustrates the WERs across various training conditions, where the x-axis represents the duration of synthetic data and the y-axis denotes the WER in percentage. The baseline (23.6%) corresponds to the WER without synthetic data augmentation (see Table 3). All models were trained with 2000 steps to ensure a fair comparison.

For Setting 1, increasing the proportion of synthetic data leads to a rapid decrease in WER initially. However, after reaching a ratio of 5:1, the WER stabilizes, and the lowest WER of 17.9% is achieved at a ratio of 7:1. For Setting 2, we observe that when the synthetic-to-real ratio is low, it

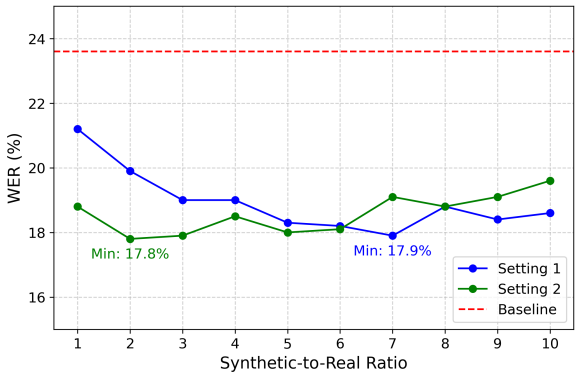

Figure 3: WER (%) based on the Frisian test set for Setting 1 and Setting 2 with varying synthetic-to-real ratios, using one hour of real data.

performs substantially better than Setting 1. At a ratio of 1:1, the WER already drops to 18.8%, and it reaches the lowest value of 17.8% at a ratio of 2:1. Beyond this point, there is either no meaningful improvement or the performance is worse than in Setting 1. This suggests that maximizing the diversity of synthetic data is only beneficial at lower synthetic-to-real ratios, but as the synthetic-to-real ratio increases, the imbalance in data quantity may cause the synthetic data to overshadow the real data, thereby increasing the risk of overfitting, ultimately diminishing performance gains. Overall, while the optimal synthetic-to-real ratio differs between the two settings, the lowest WERs achieved are nearly identical.

4. Conclusion

In this work, we systematically assessed the augmentation of TTS-based data for ASR in six truly low-resource languages, rather than relying on a simulated low-resource scenario using English data. Our results thereby better reflect real-world conditions for the low-resource languages. We investigated the feasibility and impact of using fine-tuned multilingual ZS-TTS models to generate synthetic speech for unseen speakers for these low-resource languages.

Data augmentation Our results show that data augmentation yields clear gains when real ASR data is extremely limited (less than 1 hour). As the availability of ASR training data increases, the relative effectiveness of synthetic data augmentation decreases or reverses, especially when the quality of the generated speech is suboptimal. Differences across languages are largely explained by the amount of the TTS training data available (and thus effectively by the quality of the TTS model).

Diversity Additionally, we found that increasing the amount and diversity of synthetic speech can substantially enhance ASR performance. However,

the benefits plateau when synthetic data exceeds approximately twice the amount of real data. Based on these findings, we recommend maintaining a moderate synthetic-to-real ratio, favoring diverse synthetic utterances over repeated ones, and monitoring WER of the synthetic speech as an indicator of its effectiveness. Overall, our study highlights the potential of multilingual ZS-TTS models in supporting underrepresented languages.

Future work First, since our experiments were conducted with a single ASR backbone and a single TTS architecture, it would be valuable to examine whether the observed trends generalize to alternative architectures. Second, it would be beneficial to explore the impact of using more diverse speakers and text content for TTS augmentation. Third, a more detailed analysis of synthetic speech quality, including objective acoustic metrics and perceptual evaluations, may provide additional insight into when and why synthetic data benefits or harms downstream ASR performance. Finally, future work could explore controlled settings in which the amount of TTS training data is balanced across languages, allowing for a clearer analysis of language-specific effects independent of data availability.

5. Acknowledgements

This work was partly supported by the China Scholarship Council (CSC). We thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high performance computing cluster.

6. Bibliographical References

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common Voice: A massively multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Matthew Baas and Herman Kamper. 2022. Voice conversion can improve ASR in very low-resource settings. In *Interspeech 2022*, pages 3513–3517.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12449–12460.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–729, Toronto, Canada. Association for Computational Linguistics.

James Betker. 2023. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: A massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pages 4978–4982.

Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson da Silva Soares, Sandra Aluísio, and Moacir Antonelli Ponti. 2023. ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion. In *Interspeech 2023*, pages 1244–1248.

Edresson Casanova, Julian Weber, Christopher D. Shulby, Arnaldo Candido Junior, Eren Gölge, and

Moacir A. Ponti. 2022. YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR.

Erica Cooper, Xinyue Wang, Alison Chang, Yocheved Levitan, and Julia Hirschberg. 2017. Utterance selection for optimizing intelligibility of TTS voices trained on ASR data. In *Interspeech 2017*, pages 3971–3975.

Chenpeng Du and Kai Yu. 2020. Speaker augmentation for low-resource speech recognition. In *ICASSP 2020*, pages 7719–7723.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, et al. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC 2012*, pages 31–43.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Interspeech 2015*, pages 3586–3590.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. 2020. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *CISP-BMEI 2020: 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pages 439–444.

Jason Li, Ravi Gadde, Boris Ginsburg, and Vitaly Lavrukhin. 2018. Training neural speech recognition systems with synthetic speech augmentation. *arXiv preprint arXiv:1811.00707*.

Sewade Ogun, Vincent Colotte, and Emmanuel Vincent. 2023. Can we use common voice to train a multi-speaker TTS system? In *IEEE Spoken Language Technology Workshop (SLT) 2022*, pages 900–905.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, pages 2613–2617.

Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro Moreno, Yonghui Wu, and Zelin Wu. 2019. Speech recognition with augmented synthesized speech. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002.

Nick Rossenbach, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2020. Generating synthetic audio data for attention-based speech recognition systems. In *ICASSP 2020*, pages 7069–7073.

Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, and James Glass. 2023. Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages. In *Interspeech 2023*, pages 2268–2272.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2025. Enhancing low-resource ASR through versatile TTS: Bridging the data gap. In *ICASSP 2025*. To appear.