ASR data augmentation in low-resource settings using cross-lingual multi-speaker TTS and cross-lingual voice conversion

Edresson Casanova[1,2], Christopher Shulby[3], Alexander Korolev[4], Arnaldo Candido Junior[5], Anderson da Silva Soares[6], Sandra Aluísio[2], Moacir Antonelli Ponti[2,7]

[1] Coqui, Germany; [2] Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Brazil;
[3] QuintoAndar, Portugal; [4] Darmstadt University of Applied Sciences, Germany; [5] São Paulo State University, Brazil;
[6] Federal University of Goiás, Brazil; [7] Mercado Livre, Brazil.

edresson@coqui.ai

Abstract

We explore cross-lingual multi-speaker speech synthesis and cross-lingual voice conversion applied to data augmentation for automatic speech recognition (ASR) systems in low/medium-resource scenarios. Through extensive experiments, we show that our approach permits the application of speech synthesis and voice conversion to improve ASR systems using only one target-language speaker during model training. We also managed to close the gap between ASR models trained with synthesized versus human speech compared to other works that use many speakers. Finally, we show that it is possible to obtain promising ASR training results with our data augmentation method using only a single real speaker in a target language.

Index Terms: Speech Recognition, Speech Synthesis, Cross-lingual Zero-shot Voice Conversion, Cross-lingual Zero-shot Multi-speaker TTS, ASR Data Augmentation, Low-resource

1. Introduction

Text-to-Speech (TTS) systems have received a lot of attention in recent years due to the great advances in deep learning, which have allowed for massive use in applications such as virtual assistants. These advances have allowed TTS models to achieve naturalness similar to human speech [1, 2, 3, 4]. Still, most TTS systems are tailored for a single speaker, where many applications could benefit from new-speaker synthesis, i.e., not seen during training, employing only a few seconds of the target speech. This approach is called zero-shot multi-speaker TTS (ZS-TTS) [5, 6, 7]. Advances in TTS have motivated works that exploit it as a way to improve Automatic Speech Recognition (ASR). Researchers have explored two different approaches. The first is parallel training of ASR and TTS models; in this approach the TTS and ASR systems can improve themselves, as in [8, 9]. The second is the use of a pre-trained TTS model to generate new ASR training data, such as [10], [11] and [12]. In this work, we will focus on the latter approach.

Many studies that explore a pre-trained TTS model to generate ASR data used the LibriSpeech dataset [13] to train the ASR model. For the TTS model training, [10] used 3 speakers from the American English M-AILABS dataset [14], while [11] and [12] trained the TTS model with more than 251 speakers from LibriSpeech. In Table 1 we report the Word Error Rate (WER) of the best experiment from the related works in the test-other subset of the LibriSpeech dataset. Although the studies contain both test-clean and test-other results, we focus on the results of the most difficult sub-set. Also, [12] reported results using an external language model (LM); however, for fairness, we omit this LM in our comparison.

The ASR models trained with synthesized speech combined

Table 1: *Related works comparison in the test-other subset.*

Paper	TTS Model	ASR Model	Train data	WER
[10]	Tacotron GST	Wav2Letter	Human	16.21
			Synthesized	81.78
			Hum + Synt	15.47
[11]	ZS-Tacotron + VAE	LAS	Human	13.89
			Synthesized	66.10
			Hum + Synt	13.78
[12]	GMVAE Tacotron	LAS	Human	14.10
			Synthesized	–
			Hum + Synt	13.50

with human speech achieved relative improvement[1] of 4.56%, 0.79% and 4.25% compared to the models trained with human speech alone, respectively, for [10], [11] and [12]. A greater difference is observed between the model trained with only human speech and only synthesized speech, with a relative difference of 80.17% and 78.98%, respectively, for [10] and [11], which motivates further improvements and research.

In parallel with our work, [15] explored cross-lingual voice conversion (VC) for ASR data augmentation in low-resource settings. They showed that when using a sensible amount of voice conversion data augmentation, ASR performance is improved in all low-resource languages explored.

Although previous work shows the potential for multi-speaker TTS models for ASR data augmentation, these models still require high-quality datasets with many speakers and hours of speech to converge [12]. Generally, such models are trained on English with big datasets such as LibriSpeech [13] and LibriTTS [16], which is not suitable for low-resource languages that do not have an open multi-speaker TTS dataset.

Although some multilingual multi-speaker datasets were released in recent years [17, 18, 19], they just attend a small number of languages and for many applications, even these may not be sufficient to build a competitive ASR system. In addition, creating a high-quality multi-speaker dataset is hard, because it requires the effort of multiple target-language speakers. It is especially hard for languages with small populations, where recruiting participants is difficult or in more extreme scenarios with languages that are almost extinct and have just a few speakers (e.g. indigenous languages). In a range of scenarios creating a high-quality multi-speaker dataset is not viable.

In light of this, an approach that applies TTS/VC for ASR data augmentation that requires just a medium/low-quality single-speaker dataset could make the application of this technology viable for languages that really need it, helping to preserve/protect nearly extinct languages, for example. The objective of this paper is to improve upon such issues and make it

---

[1]We used relative improvement/difference metric to show the real improvement achieved by related works' approaches.

viable. Here, we seek to answer two questions: Is a TTS model trained with just one speaker in a given target language sufficient for ASR data augmentation? Is only one human speaker in the target language enough to get a reasonable ASR model via cross-lingual voice conversion and cross-lingual multi-speaker TTS ASR data augmentation? The contributions of this work are as follows:

- A novel approach for ASR data augmentation that explores cross-lingual voice conversion and a cross-lingual multi-speaker TTS model. For TTS and voice conversion. We used the YourTTS model [7], which was developed in a previous work to meet the requirements needed for this paper. Our novel approach requires just one speaker in the target language, making the application of this technology possible for low-resource languages;

- We are the first to combine multi-speaker TTS and voice conversion for ASR data augmentation. In addition, we are the first to explore cross-lingual multi-speaker TTS and cross-lingual voice conversion using speakers of other languages to fill the lack of speakers for low-resource ASR model training.

- We are the first to explore ASR data augmentation via TTS with a very limited number of speakers. To do so, we emulate a scenario where the ASR model and the TTS model would need to be trained with just one real speaker (low-resource language scenario) on two target languages. Our novel approach improves WER from 64-74% to 34-37%, approximately a 33% absolute improvement. Such results indicate the feasibility of applying this technology for low-resource languages.

2. Audio datasets

We used 3 languages/training datasets for the TTS model:

English: VCTK dataset [20], containing 44 hours of speech from 109 speakers, sampled at 48KHz. We divided the VCTK dataset into training, development and test subsets following [7]. To further increase the number of training speakers, we used the subsets *train-clean-100* and *train-clean-360* from LibriTTS [16]. In the end, our TTS model was trained with approximately 298 hours from 1,248 English speakers.

Portuguese: TTS-Portuguese Corpus [21], a single-speaker male dataset in Brazilian Portuguese (pt-BR) containing approximately 10 hours, sampled at 48KHz. As the authors did not use a soundproof studio, the dataset contains some environmental noise. Following [7], we resampled the audios to 16Khz and used FullSubNet [22] as a denoiser. For development, we randomly selected 500 samples, leaving the rest for training.

Russian: ru_RU set of the M-AILABS dataset based on LibriVox, consisting of 46 hours from 1 female and 2 male speakers. We used samples only from the female speaker for diversity, since we already used a male for pt-BR. For development, we randomly selected 500 samples, leaving the rest for training.

For all TTS datasets, pre-processing was carried out to normalize volume and to remove long silences, following [7]. After pre-processing, the datasets contained 8.38 hours for pt-BR and 14.94 hours for ru-RU (Russian). For ASR training, we used Common Voice version 7.0 [23] for pt-BR and ru-RU. In all experiments, we used the default train, development and test partitions. For pt-BR, these sets have 14.52, 8.9 and 9.5 hours, respectively; and ru-RU has 25.95, 13.06 and 13.75 hours, in the same order. In addition, the speaker distribution for train, development and test partitions are 103, 238 and 1252 speakers

for pt-BR; and ru-RU has 117, 210 and 1202 speakers.

3. TTS model setup

In our previous work, we presented YourTTS model [7], a multilingual zero-shot multi-speaker TTS model, which achieved state-of-the-art (SOTA) results using only a single-speaker dataset in the target language. Although the focus of the model is on TTS it can also do zero-shot voice conversion. This model was developed to meet the requirements needed for this paper. In the original work, we trained YourTTS using English (LibriTTS and VCTK datasets), French (M-AILABS dataset) and pt-BR (TTS-Portuguese Corpus). The model was trained using only one male pt-BR speaker, but still produced strong results in zero-shot multi-speaker TTS and zero-shot voice conversion for pt-BR. Furthermore, it was able to produce female voices even without being trained on female voices, making it adequate for the objective of this study. Here, we fine-tuned the YourTTS model in English, pt-BR and ru-RU. For this we used transfer learning from the original checkpoints made publicly available. The dataset in English and pt-BR were the same dataset used in our previous work, but we replaced the French M-AILABS dataset with a female speaker from the ru-RU M-AILABS dataset due to experiment requirements as explained in Section 4. We trained the YourTTS model for 140k steps with the same parameters used in [7]. In summary, we trained YourTTS with 1,248 speakers in English, 1 male pt-BR speaker and 1 female ru-RU speaker. After the training, this model checkpoint was used as TTS/VC model for all experiments in this paper.

YourTTS can synthesize different audios for the same input sentence. During inference, the latent variable, predicted by the text encoder, is added with a random sample of the standard normal distribution multiplied by a temperature T. In this way, diversity can be controlled by the temperature T. As shown by [3], the manipulation of T allows for generating different pitches; for more details see [3, 4, 7]. Furthermore, YourTTS is trained with the stochastic flow-based duration predictor proposed in [4], which can produce several different speech rhythms for the same sentence. To do so, during inference, a random sample of the standard normal distribution is multiplied by a temperature T_{dp} and added to the latent variable before being inverted by the flow. In this way, it is possible to control the variety of rhythms with the temperature T_{dp} [4]. Finally, it is possible to control the speaking rates by multiplying the predicted durations by a positive scalar L, thus making the pronunciation faster when L is smaller and slower when L is bigger.

4. Is a multi-speaker TTS model trained with just one speaker in the target languages enough for ASR data augmentation?

Previous works have shown a large gap between ASR models trained with human and synthesized speech [11, 12], where researchers have used a large number of speakers and hours in the target language during the TTS training. To apply this method to languages with low/medium resources, we need an approach that only requires a single speaker in the target language. In this section, we propose to employ YourTTS trained with only one speaker in the target language to do data augmentation for ASR. It is important to note that this TTS is also trained in English and it is exposed to embeddings from English speakers. We follow related works comparing ASR models trained with synthesized and human speech, to verify if our approach can be used as data

augmentation for ASR.

For fair comparison between human and synthesized speech, we employ YourTTS model to generate a synthesized version of pt-BR and ru-RU Common Voice datasets. For each sentence in Common Voice, we generate its pronunciation for the same speaker, using the sentence's pronunciation as reference for speaker embedding extraction. So, we have used speaker embeddings from the target language's native speakers. The idea being that if the zero-shot multi-speaker TTS model is good enough, it will generate the same speaker's voice as in the original audio, additionally, the synthesized and human dataset will have the same vocabulary. During the generation, as explained previously in Section 3, diversity is achieved by randomly choosing L, T and T_{dp}: for L, a value between 0.7 and 2, while for temperatures (T and T_{dp}) a value between 0 and 0.667.

To increase the diversity in ASR training, in some experiments, we also explored three popular augmentation methods in speech processing – additive noise, pitch shifting and room impulse response (RIR) simulation. For additive noise and RIR filters we have used the same approach and dataset from [24]. For pitch shift, we randomly chose a semitone from -4 to 4. All augmentations are randomly selected with a 25% probability of being chosen for each audio in every training step. For all methods, we used the implementations available in the Python Audiomentations[2] package. We will refer to it as audio augmentations (AA).

As for the ASR model, we use Wav2vec 2.0 [25], a large self-supervised model trained on the VoxPopuli dataset [26]. We used the model checkpoint provided by the authors[3] which was trained on 100k hours of speech in the following 23 languages: Bulgarian (Bg), Czech (Cs), Croatian (Hr), Danish (Da), Dutch (Nl), English (En), Estonian (Et), Finnish (Fi), French (Fr), German (De), Greek (El), Hungarian (Hu), Italian (It), Latvian (Lv), Lithuanian (Lt), Maltese (Mt), Polish (Pl), Portuguese (Pt), Romanian (Ro), Slovak (Sk), Slovene (Sl), Spanish (Es) and Swedish (Sv). We chose Pt and Ru because Pt was used in the self-supervised pre-training and Ru was not, presenting realistic results for both scenarios and are from different language families. We carried out four experiments:

- **Experiment 1:** ASR models trained in pt-BR and ru-RU with Common Voice using the standard training and development subsets. For pt-BR, the model was trained for 140 epochs and ru-RU for 100;
- **Experiment 1.1:** Transfer learning from experiment 1, adding audio augmentations (AA) in the training. In this experiment, the ASR is trained with half the number of epochs used in experiment 1.
- **Experiment 2:** Similar to experiment 1, but the model is trained using the synthesized version of Common Voice in pt-BR and ru-RU. For model training and development, we used synthesized speech.
- **Experiment 2.1:** Transfer learning from experiment 2 plus the AA. The ASR is trained with half the number of epochs used in experiment 2.

To run the experiments, we use the HuggingFace Transformers framework[4]. The models were trained with a NVIDIA TITAN RTX 24GB GPU using a batch size of 8 and gradient accumulation over 24 steps. We used the AdamW optimizer with linear learning rate warm-up from 0 to 3e-05 in the first 8

epochs and after using linear decay to zero. During training, the best checkpoint was chosen, using the loss in the development set and early stopping after the development loss had not improved for 10 consecutive epochs. The code used for all of the experiments, as well as model checkpoints are publicly available at: https://github.com/Edresson/Wav2Vec-Wrapper.

Table 2 presents the WER results for our experiments on the original pt-BR and ru-RU test subsets of Common Voice.

Table 2: *Human and synthesized speech comparison (WER).*

Exp.	PT	RU
1. Human Speech	23.50	25.47
1.1 Human Speech + AA	21.54	22.27
2. Synthesized speech	56.84	65.85
2.1 Synthesized speech + AA	43.99	50.46

The model trained only with human speech (Experiment 1) reached a WER of 23.50% and 25.47%, respectively for pt-BR and ru-RU. The model trained only with synthesized speech (Experiment 2) reached a WER of 56.84% and 65.85%, for pt-BR and ru-RU. Therefore, without AA, the relative difference between the model trained with only human speech and synthesized speech is of 58.65% and 61.32% for those two languages.

As expected, fine-tuning the models with AA (Experiment 1.1 and Experiment 2.1) improved performance. The model trained with human speech only improved its result by 1.96% and 3.20% WER for pt-BR and ru-RU after the addition of AA. The model trained only with synthesized speech improved by 12.85% and 15.39% WER. Therefore, using AA benefits the model trained with synthesized speech much more than the one with human speech. This can be explained by the absence of noise diversity in the synthesized speech. Common Voice is a dataset composed of a lot of environmental noises, whereas the synthesized speech just has some artifacts, but not environmental noises. Therefore, with the use of AA, the gap between models trained only with human and synthesized speech is reduced to a relative difference of 51.03% and 55.86% for those two languages. Our results are interesting because, despite using only a single-speaker dataset for training the TTS model for pt-BR and ru-RU, our ASR model trained only with synthesized speech achieves a comparable result to related works. Considering [11], the relative difference between models trained only with human and synthesized speech was of 78.98%; and ours around 60% for two non-English languages. Even though this work explores the English language in a setting with many available speakers and it is not directly comparable, we believe that our results indicate that our approach requiring just 1 speaker in the target language, can be used for low-resource languages.

In this way, we believe that the YourTTS model proposed previously meets the requirements of this paper, trained with just one speaker in the target languages is enough to be used for ASR data augmentation.

5. Is only one human speaker in the target language enough to get a reasonable ASR model via cross-lingual voice conversion and cross-lingual multi-speaker TTS ASR data augmentation?

In Section 4 we showed that YourTTS model trained with only a single speaker in pt-BR and ru-RU was effective as ASR data augmentation, achieving similar results as previous works that explored multi-speaker datasets. Although the TTS model was trained with only one target-language speaker, we have used all

---

[2]https://github.com/iver56/audiomentations
[3]https://huggingface.co/facebook/wav2vec2-large-100k-voxpopuli
[4]https://github.com/huggingface/transformers

Common Voice dataset speakers' embeddings to create the synthetic version of this dataset. So, the speaker embeddings were extracted from target languages' native speakers. This approach has shown good results, but many low-resources languages do not have datasets with many available speakers. So, the results reported are not realistic for extreme low-resource languages. For this reason, in this section we explored the use of a single-speaker in the target language, for training both TTS and ASR. To make up for the lack of speakers during the creation of the synthetic dataset, we have used English speakers, rather than cloning speakers from the target language. We created two synthetic datasets using YourTTS model:

GEN_TTS: Was created by synthesizing all the sentences in pt-BR and ru-RU Common Voice using English speaker embeddings, chosen at random from the 1,248 speakers available in the training set. Differently from Section 4, no target-language speaker embeddings are used because we focus on a more restricted, extremely low resource scenario. Each sentence was synthesized using one English speaker embedding. We would like to note that, in preliminary experiments, we explored increasing the number of speakers per sentence from this dataset; however, this did not bring significant improvements. During the generation of this dataset, as explained previously in Section 3, diversity is achieved by randomly choosing L, T and T_{dp}: for L, a value between 0.7 and 2, while for temperatures (T and T_{dp}) a value between 0 and 0.667.

GEN_VC: consists of the single-speaker dataset used for training the TTS model in the target language converted to a multi-speaker dataset using cross-lingual voice conversion with English speakers. Each sample used in the TTS training was converted to the voice of 5 speakers, chosen at random from the 1,248 English speakers. The value of 5 transfers per sample was chosen in preliminary experiments.

We carried out three experiments that used AA and the same training parameters used in Section 4:

- **Baseline:** ASR models trained with the single-speaker dataset used during the TTS model training on pt-BR and ru-RU.
- **Upper Bound:** ASR models trained on pt-BR and ru-RU with Common Voice plus the single-speaker TTS dataset.
- **Baseline + DA:** human speech from a single-speaker in the target language (TTS dataset), with data augmentation being accomplished by YourTTS. For data augmentation we merge the *GEN_TTS* and *GEN_VC* datasets, detailed above. Figure 1 presents a full pipeline diagram of this experiment.

Table 3 presents our experiments' results on the original test subsets of the pt-BR and ru-RU Common Voice datasets.

Table 3: *Results on the pt-BR and ru-RU CV testsets (WER)*

Experiment	Train data	PT	RU
Baseline	TTS dataset (single-speaker)	63.90	74.02
Upper Bound	Common Voice + TTS dataset	20.39	24.80
Baseline + DA	TTS dataset + *GEN_TTS* + *GEN_VC*	33.96	36.59

The model trained with just 1 target language speaker (Baseline) achieved a WER of 63.90% and 74.02% for pt-BR and ru-RU. The model trained with only 1 real target language speaker (TTS dataset) with data augmentation using voice conversion and speech synthesis (Baseline + DA), achieved a WER of 33.96% and 36.59%. Therefore, our data augmentation approach in scenarios with just 1 real speaker available improves the WER by 29.94% and 37.43% for pt-BR and ru-RU.

Comparing the results with the SOTA English ASR system on Common Voice (7.7% achieved by [27]), these results do not look so remarkable; however, [28] used approximately 158
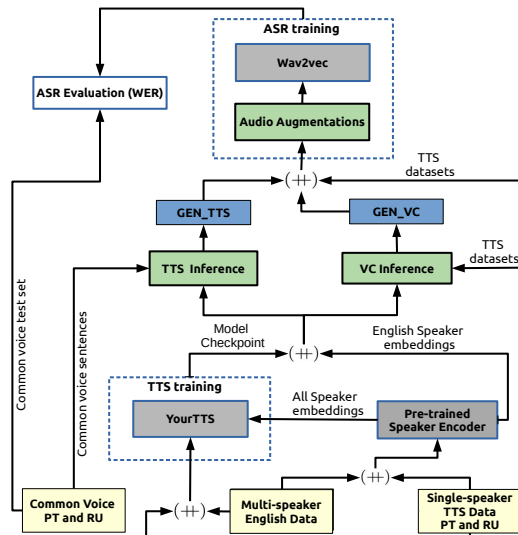

Figure 1: *Full pipeline diagram for Baseline + DA experiment*

hours of pt-BR speech and a non-self-supervised model without an external LM and achieved a WER of 47.41% on the test set of BRSD v2 dataset. Despite using a different dataset, [29] showed that the Common Voice test set is more challenging than the test set of BRSD v2, and for this reason, the model proposed by these authors reached a higher WER on the Common Voice dataset. In ru-RU, [30] used transfer learning from 5 large English datasets, trained the QuartzNet model [31] on Common Voice, obtaining a WER of 32.20% on the test set. Therefore, 33.96% WER achieved by our model is probably superior to the SOTA for pt-BR, before the introduction of self-supervised learning approaches. Also, the WER of 36.59% achieved in ru-RU is comparable with the SOTA. Comparing the results of the Baseline + DA experiment with the Upper Bound (20.39% and 24.80% for pt-BR and ru-RU), our results still miss the Upper Bound. However, the results are remarkable since the TTS and ASR models were trained with just **one real speaker** in the target language, and the model was able to recognize the voice of over a thousand speakers from the Common voice test set.

6. Conclusions and future work

We presented a novel data augmentation approach for ASR training by using cross-lingual multi-speaker speech synthesis and voice conversion. We show that it is possible to achieve promising results for ASR model training with just a single-speaker dataset in a target language, making it viable for low-resource scenarios. Finally, our approach works both in a language (pt-BR) present in the Wav2Vec 2.0 self-supervised pre-training, as well as for a completely unseen language (ru-RU). In future work, we intend to explore the use of a self-supervised model feature extractor as a discriminator during the training of the YourTTS model. In this way, the YourTTS model may produce even more human-like speech. In addition, we intend to do ablation studies using other SOTA ASR models like Whisper [32] and WavLM [33]. Finally, we intend to apply our method to Brazilian indigenous languages that have a few or even only one single-speaker data available.

7. Acknowledgements

8. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[2] R. Valle, K. J. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," in *International Conference on Learning Representations*, 2020.

[3] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *arXiv preprint arXiv:2005.11129*, 2020.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[5] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480–4490.

[6] S. Choi, S. Han, D. Kim, and S. Ha, "Attentron: Few-shot text-to-speech utilizing attention-based variable-length embedding," *arXiv preprint arXiv:2005.08484*, 2020.

[7] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *International Conference on Machine Learning*. PMLR, 2022, pp. 2709–2720.

[8] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain with one-shot speaker adaptation," *Proc. Interspeech 2018*, pp. 887–891, 2018.

[9] ——, "Machine speech chain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976–989, 2020.

[10] J. Li, R. Gadde, B. Ginsburg, and V. Lavrukhin, "Training neural speech recognition systems with synthetic speech augmentation," *arXiv preprint arXiv:1811.00707*, 2018.

[11] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 996–1002.

[12] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. Rybin, "You do not need more data: improving end-to-end speech recognition by text-to-speech data augmentation," in *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2020, pp. 439–444.

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

[14] I. Solak, "The m-ailabs speech dataset," 2019.

[15] M. Baas and H. Kamper, "Voice Conversion Can Improve ASR in Very Low-Resource Settings," in *Proc. Interspeech 2022*, 2022, pp. 3513–3517.

[16] 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. Interspeech 2019*, 2019, pp. 1526–1530. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2441

[17] K. Park and T. Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," *Proc. Interspeech 2019*, pp. 1566–1570, 2019.

[18] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," *Proc. Interspeech 2020*, pp. 2757–2761, 2020.

[19] S. Elizabeth, W. Matthew, B. Jacob, R. Cattoni, M. Negri, M. Turchi, D. W. Oard, and P. Matt, "The multilingual tedx corpus for speech recognition and translation," in *Proceedings of Interspeech 2021*, 2021, pp. 3655–3659.

[20] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2016.

[21] E. Casanova, A. C. Junior, C. Shulby, F. S. d. Oliveira, J. P. Teixeira, M. A. Ponti, and S. Aluísio, "Tts-portuguese corpus: a corpus for speech synthesis in brazilian portuguese," *Language Resources and Evaluation*, pp. 1–13, 2022.

[22] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun 2021. [Online]. Available: http://dx.doi.org/10.1109/ICASSP39728.2021.9414177

[23] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4218–4222.

[24] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, "Clova baseline system for the voxceleb speaker recognition challenge 2020," *arXiv preprint arXiv:2009.14153*, 2020.

[25] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[26] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *arXiv preprint arXiv:2101.00390*, 2021.

[27] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang *et al.*, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[28] I. M. Quintanilha, S. L. Netto, and L. W. P. Biscainho, "An open-source end-to-end asr system for brazilian portuguese using dnns built from newly assembled corpora," *Journal of Communication and Information Systems*, vol. 35, no. 1, pp. 230–242, 2020.

[29] L. R. Stefanel Gris, E. Casanova, F. Santos de Oliveira, A. da Silva Soares, and A. C. Junior, "Brazilian portuguese speech recognition using wav2vec 2.0," *arXiv e-prints*, pp. arXiv–2107, 2021.

[30] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," *arXiv preprint arXiv:2005.04290*, 2020.

[31] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6124–6128.

[32] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[33] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.