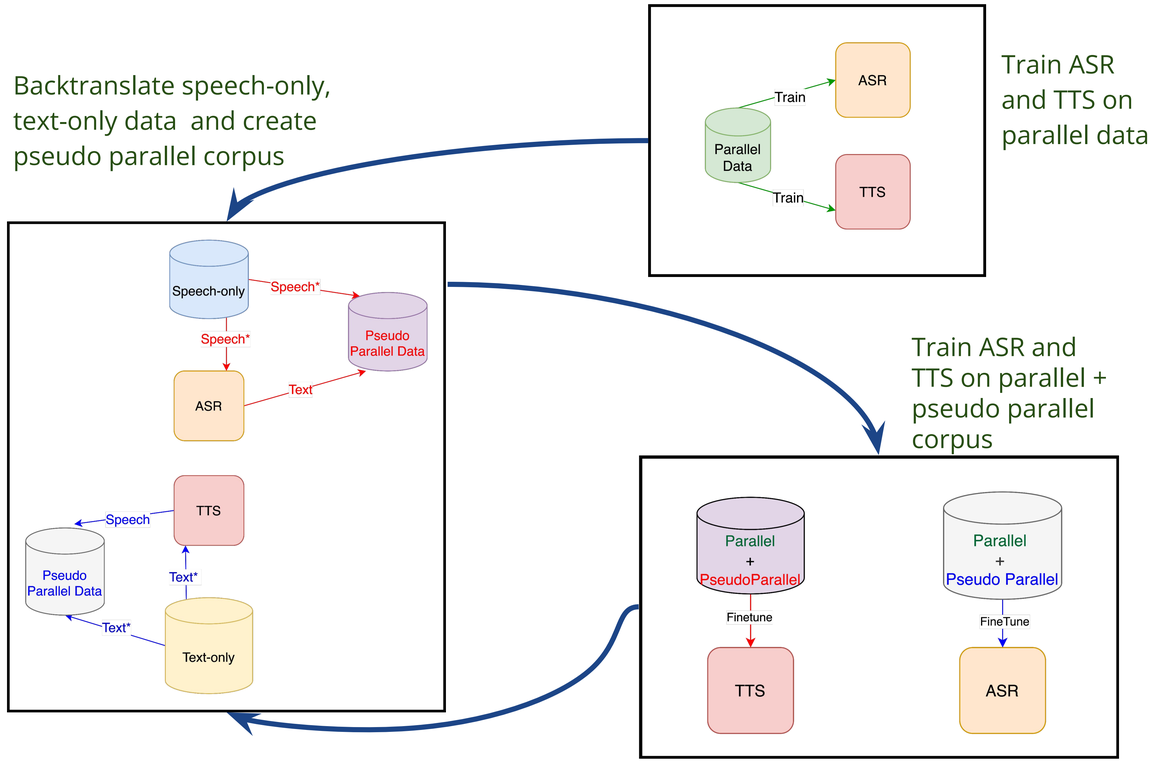
Iterative Back-Translation-Style Data Augmentation for Low Resource ASR and TTS

Bharani Ujjaini Kempaiah Preksha Patel Ruben John Mampilli
{buk, pup, rmampill}@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University

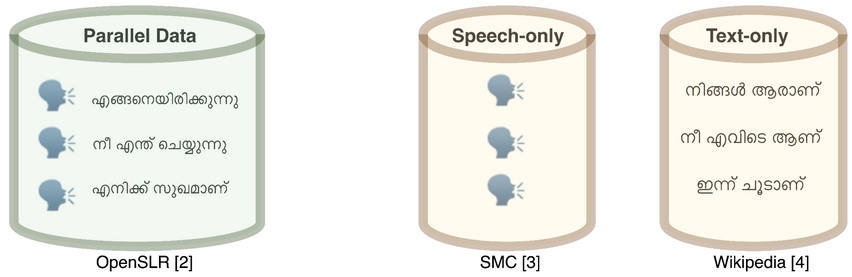
Introduction

Often automatic speech recognition (**ASR**) and text-to-speech (**TTS**) synthesis models are trained independently. However, previous research has shown that training both the models in parallel would allow us to integrate speech perception and speech production to improve the performance of both ASR and TTS (Andros et al., 2020). Authors of this paper observe that training ASR and TTS **together** is more beneficial when the amount of data is lower indicating that it might generalize well to **low-resource** languages. Moreover, **text-only** and **speech-only** data could also be leveraged while training the ASR and TTS in parallel. We hypothesize that training the ASR and TTS models in parallel for **Malayalam**, a low resource language with only ~6 hours of parallel speech and text data, would improve the performance of both the models. We aim to leverage text-only and speech-only Malayalam data by utilizing the trained models to **generate** pseudo parallel speech and text data which can be used for fine tuning the models.

Methodology



Data



Results

We report preliminary results after 1 iteration of data augmentation. We measure the performance of ASR using **word error rate** (WER) and **character error rate** (CER). The performance of TTS is measured using **Log-F0 root mean square error** (f0-RMSE) and **Mel-cepstral distortion** (MCD). The reported results are on the same test data from the original parallel corpus.

ASR	WER	CER
Original Parallel Data	36.5	7.9
With TTS Generated data	38.3	8.3

TTS	f0-RMSE	MCD
Pretrained model	0.253 ± 0.060	12.06 ± 0.80
Original Parallel Data	0.249 ± 0.069	10.64 ± 1.34
With ASR Generated Data	0.249 ± 0.067	10.33 ± 1.04

Next Steps

- ➔ **Fine-tune** on pseudo parallel data
- ➔ **Iterative** back translation style data augmentation
- ➔ Model loss propagation between **ASR ⇌ TTS**

References

[1] Tjandra, Andros, Sakriani Sakti, and Satoshi Nakamura. "Machine speech chain." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 976-989.
[2] He et al., Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems LREC 2020
[3] Swathanthra Malayalam Computing Malayalam Speech Corpus
[4] Malayalam Wikipedia Articles [https://www.kaggle.com/datasets/disisbig/malayalam-wikipedia-articles]
[5] Gupta, Anirudh, et al. "CLSRIL-23: cross lingual speech representations for indic languages." arXiv preprint arXiv:2107.07402 (2021).
[6] Watanabe, Shinji, et al. "Espnet: End-to-end speech processing toolkit." arXiv preprint arXiv:1804.00015 (2018).