Multilingual TTS Accent Impressions for Accented ASR

Georgios Karakasidis[1,3 (*)], Nathaniel Robinson[2 (*)], Yaroslav Getman[1], Atieno Ogayo[2], Ragheb Al-Ghezi[1], Ananya Ayasi[2], Shinji Watanabe[2], David R. Mortensen[2], and Mikko Kurimo[1]

[1] Department of Signal Processing and Acoustics, Aalto University, Finland
{firstname.lastname}@aalto.fi
[2] Language Technologies Institute, Carnegie Mellon University, USA
{nrrobins,aogayo,aayasi,swatanab,dmortens}@cs.cmu.edu
[3] Institute for Language, Cognition and Communication, University of Edinburgh, UK
g.karakasidis@ed.ac.uk

Abstract. Automatic Speech Recognition (ASR) for high-resource languages like English is often considered a solved problem. However, most high-resource ASR systems favor socioeconomically advantaged dialects. In the case of English, this leaves behind many L2 speakers and speakers of low-resource accents (a majority of English speakers). One way to mitigate this is to fine-tune a pre-trained English ASR model for a desired low-resource accent. However, collecting transcribed accented audio is costly and time-consuming. In this work, we present a method to produce synthetic L2-English speech via pre-trained text-to-speech (TTS) in an L1 language (target accent). This can be produced at a much larger scale and lower cost than authentic speech collection. We present initial experiments applying this augmentation method. Our results suggest that success of TTS augmentation relies on access to more than one hour of authentic training data and a diversity of target-domain prompts for speech synthesis.

Keywords: accented speech recognition, data augmentation, low-resource speech technologies, speech synthesis

1 Introduction

English is one of the most widely spoken languages in the world [11]. Like many languages, it is diverse and multi-dialectal [3]. ASR systems for English and other high-resource languages are celebrated for high accuracy [19]. However, these ASR systems are often tailored for a small number of dialects, due to limited data diversity [5]. Studies have shown bias in English ASR systems against marginalized language varieties [15], an ethical concern since this bias can disproportionately affect marginalized groups [16] and immigrants [10]. Demonstrated ASR bias against non-native English accents [25] is particularly concerning, due to the large and growing number of L2 English speakers [8]. Similar trends exist for other high-resource languages [4], but we direct our focus to English.

One potential strategy to accommodate a greater number of English speakers is to adapt existing trained English ASR models to different accents [28]. This requires

(*) Equal contribution.

labeled accented English speech data. However, labeled data in specific English accents is scarce [5], and collecting human speech for a large number of English accents is costly and time-intensive.

We propose a novel method: *produce L2-accented English for ASR training via text-to-speech (TTS) pre-trained for another language*. Accented speech can be approximated by passing English inputs through TTS for a language corresponding to the target accent. For example, English text through Spanish TTS will approximate Spanish-accented English. This strategy is inspired by the success of applying TTS speech for low-resource language ASR [20], [21], [6], [27]. It is also inspired by the adaptability of commercial TTS systems such as Microsoft TTS, Google TTS and Amazon Polly to English accents. We chose Microsoft TTS because its online documentation[4] states that "All neural voices are multilingual and fluent in their own language and English" and indicates that English text prompts passed through another language's system will be rendered as accented English speech. In summary, we contribute:

- A novel method for accented ASR training by producing synthetic accented speech via a readily available foreign TTS system
- Reduced ASR error rates in some settings via our synthetic augmentation method
- Indications that synthetic accented speech augmentation relies on at least one hour of authentic data

2 Related Work

We are not the first researchers to investigate augmenting ASR training data via TTS. Multiple researchers have used TTS to extend ASR training data for a variety of languages including Mandarin [12] and low-resource languages in a variety of settings [27], including for languages with no TTS systems [20] and for children's ASR [9]. Others have leveraged TTS to replace a need for real speech features in training [24], [17]. These TTS-based methods show promising results for improving ASR in low-resource settings. Our work, however, is the first to apply this approach to adapt ASR models to low-resource accents.

We are also not the first researchers to approach improving accented ASR. The Accented English Speech Recognition Challenge (AESRC2020) [22] garnered developments in the area from a variety of researchers, including accent embeddings and model layers [13], [2]. [23] ranked first in AESRC2020 with 10.1% word error rate (WER) by data augmentation and ensembling accoustic models. [5] improved ASR by 33% in multiple accents by leveraging as little as 105 minutes of unannotated speech in a target accent with an adversarial transfer learning approach. Like these methods, our approach incorporates data augmentation and is largely unsupervised, incorporating a small amount of optional labeled data. However, we are the first researchers to take a multilingual TTS-based approach to accented ASR.

[4] https://learn.microsoft.com/en-us/azure/cognitive-services/speech-service/get-started-text-to-speech

Table 1. Data statistics for authentic sets. $n_{\mathbf{a}}$ represents the number of TTS voices for synthetic data production.

Accent	train mins.	dev mins.	test mins.	$n_{\mathbf{a}}$
Common Voice				
German	3.5K	396	438	18
Malaysian	66	6	18	4
Filipino	264	30	30	2
Arctic				
Arabic	37	9	55	32
Chinese	40	10	60	36
Hindi	37	10	42	2
Korean	44	12	51	8
Spanish	43	11	58	68
Vietnamese	45	11	56	2

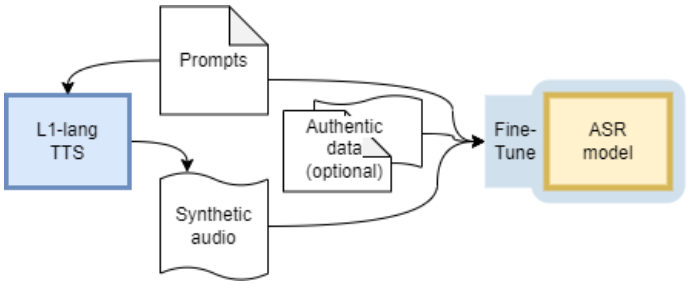

Fig. 1. Accented ASR via synthesized audio

3 Methodology

Our method of accented English ASR via synthetic dataset curation is illustrated in Figure 1. It requires (1) a generic pre-trained English ASR model; (2) a trained TTS system in the L1 language corresponding to the target L2-English accent; (3) a corpus of English sentences to use as TTS prompts; and, (4) optionally, a small amount of authentic accented speech data with transcriptions (which may serve as the English corpus). Accented ASR adaptation involves two steps: the data synthesis step consists of passing TTS prompts as input to the pre-trained TTS system (using a variety of TTS-voices as speakers if available) to produce automatically annotated synthetic audio. The training step involves fine-tuning the pre-trained English ASR model in the synthetic accented speech, along with the small amount of authentic accented speech, if available. We assume any authentic data set would be small, since this method is intended for low-resource language varieties.

In our experiments we explore the following methodological variations: fine-tuning on a small authentic dataset; fine-tuning on a large synthetic dataset combined with

a small authentic dataset, with the authentic data up-sampled; and fine-tuning in two steps, first with a large synthetic dataset and subsequently with a small authentic dataset. (Details in § 4.)

3.1 Data

We tested our hypothesis on a total of nine accents, with authentic accented train, validation (dev), and test sets taken from the publicly available Common Voice (CV) [1] and L2-Arctic [26] databases. Table 1 contains statistics about the train/dev/test splits for all nine accents from both sources. More detailed specifications regarding the data preparation can be found on our github repository [5].

L2-Arctic is a corpus originally designed for the development of TTS systems for non-native English speakers. The small size of the dataset represents extremely low-resource settings in our experiments. It consists of six accents corresponding to L1 languages Arabic, Chinese, Hindi, Korean, Spanish, and Vietnamese, each of them represented by four speakers (two males and two females) with audio recordings spoken in clean environments. We split this data into separate train, dev, and test sets. We took great care to ensure that there was no overlap of voices or text prompts between the test set and the train and dev sets. Because L2-Arctic uses largely the same text prompts for all four speakers of a given accent, this meant we had to discard nearly half of the available data. For each accent we designated one male and one female speaker as test speakers and the remaining male and female speaker as train/dev speakers. (This also ensured we would train and test on both male and female voices.) In our main experiments we designated 40% of the utterances from the test speakers as our test set. We then constructed the train/dev sets by splitting the train/dev speaker files with an 80/20 ratio and afterward removing any files that had prompts contained in the test set. This allowed for a sizeable test set but resulted in small training amounts (see Table 1). Due to our concerns that the small train set may inhibit performance, we conducted some experiments where we allowed the train/dev sets to be as large as possible, though this resulted in very small test sets since we wanted to keep the sets of test prompts and train/dev prompts disjoint.

We used the eighth version of CV, a crowd-sourced dataset with messier audio than L2-Arctic. CV annotation only lists the speakers' country of origin, not L1 language. We selected German, Malaysian, and Filipino accents for our experiments because they mapped straightforwardly to L1 languages supported by Microsoft TTS.[6] Due to limitations in the number of speakers and their gender distribution for each language, splitting the CV dataset was less straightforward than L2-Arctic. We sampled 20% of the speakers for each accent, and used them for the accent's test set. From the utterances of the remaining 80% of speakers, we used 90% as the training set and 10% as the validation set. The German accent did not contain detailed speaker information (all of the prompts were uttered by the same client who seemingly corresponded to the same male speaker), so we used a random train/dev/test split where each subset consisted of 81%, 9%, and 10% of the whole set, respectively.

[5] https://github.com/geoph9/accent-adaptation-through-tts

[6] More details about the TTS voices can be found on our repository.

3.2 Model

Our baseline model is wav2vec 2.0 [2], which is an end-to-end neural network that consists of a convolutional feature encoder, a transformer, and a quantizer. In particular, we use the publicly available *wav2vec2-base-960h* model[7] which is pre-trained and fine-tuned on 960 hours of transcribed audio from the Librispeech data set [18]. We followed the same training setup and hyperparameters for all fine-tuning experiments, with some small variations in batch size[8]. Our models were fine-tuned for 20 epochs with a learning rate of 1e-4. This procedure was done by first freezing the CNN feature encoder and updating the rest of the weights while training.

Table 2. Performance on controlled comparison. Underlined results in the **Synth.** column outperformed **Before Adapt.**. Best results across all experiments (including those in Tables 3 and 4) are **bold**.

Accent	Before Adapt.		Auth.		Synth.		Combined	
	WER%	CER%	WER%	CER%	WER%	CER%	WER%	CER%
CommonVoice								
German	32.77	11.61	8.42	2.12	60.58	27.72	**8.07**	2.05
Malaysian	44.59	18.84	**30.72**	**12.55**	42.12	17.76	34.81	13.98
Filipino	27.53	9.41	18.92	6.23	26.49	9.32	19.06	6.33
Arctic								
Arabic	19.85	7.75	**15.49**	**5.84**	23.71	8.33	17.47	6.68
Chinese	34.78	15.37	26.29	11.31	34.85	14.62	**25.69**	**11.10**
Hindi	17.26	6.73	**11.34**	**3.77**	17.49	5.56	12.30	4.17
Korean	19.51	7.65	15.92	6.10	26.98	10.84	**15.26**	**5.94**
Spanish	25.69	10.50	**21.06**	**8.21**	38.59	12.88	22.23	8.67
Vietnamese	42.25	19.43	**31.93**	14.30	47.27	20.60	33.50	14.73

4 Experiments and Results

We conducted a set of experiments to compare the effectiveness of synthetic to authentic accented audio, the results of which are displayed in Table 2. We used these same test sets for all experiments. **Before Adapt.** (baseline): We tested the wav2vec 2.0 model off the shelf on the test set for each accent $\mathbf{a}$. **Auth.**: Next, for each accent, we fine-tuned our model using the authentic train and dev sets $\mathcal{A}_{\mathbf{a}}$ detailed in Table 1. **Synth.**: We then generated synthetic audio through Microsoft TTS, using the exact prompts from the authentic train and dev sets to produce new synthetic train and dev sets $\mathcal{S}_{\mathbf{a}}$ for fine-tuning . We produced exactly one TTS file for each prompt, by uniformly sampling one of the $n_{\mathbf{a}}$ available Microsoft TTS voices.[9] (See Table 1 for $n_{\mathbf{a}}$ values.) **Combined**:

[7] https://huggingface.co/facebook/wav2vec2-base-960h

[8] We initially opted for batch size of 128, which we used to produce results for German **Auth.** and all Filipino fine-tuned results in Table 2. However in subsequent experiments, this exceeded memory constraints. Accordingly we used a batch size of 96 for all other experiments.

[9] We included voices for multiple TTS dialects corresponding to the L1 language for each accent (and more than one L1 language in the case of Malaysian) and sampled voices assigned to each accent uniformly without regard for TTS dialect.

Table 3. Performances on the larger synthesized sets.

Accent	Before Adapt.		Gutenberg synth.		Domain synth.	
	WER%	CER%	WER%	CER%	WER%	CER%
CommonVoice						
German	32.77	11.61	61.10	27.22	56.67	24.64
Malaysian	44.59	18.84	66.86	33.99	63.92	30.80
Filipino	27.53	9.41	34.27	13.29	39.32	15.18
Arctic						
Arabic	19.85	7.75	50.71	26.57	31.09	11.91
Chinese	34.78	15.37	41.53	19.11	32.57	14.38
Hindi	17.26	6.73	25.85	10.10	18.57	5.90
Korean	19.51	7.65	61.37	31.17	33.80	14.33
Spanish	25.69	10.50	35.94	15.16	34.41	12.39
Vietnamese	42.25	19.43	76.01	40.21	52.76	23.49

Table 4. Performances on the larger synthesized sets. Underlined results outperformed **Auth.**. Best results across all experiments (including those in Tables 2 and 3) are **bold**.

Accent	Before Adapt.		Comb. Up-samp.		Two-stage FT	
	WER%	CER%	WER%	CER%	WER%	CER%
CommonVoice						
German	32.77	11.61	8.78	2.20	8.09	**2.02**
Malaysian	44.59	18.84	36.13	14.93	36.08	15.20
Filipino	27.53	9.41	18.51	6.30	**18.00**	**5.93**
Arctic						
Arabic	19.85	7.75	18.66	7.18	17.89	6.90
Chinese	34.78	15.37	27.16	11.93	26.44	11.49
Hindi	17.26	6.73	11.91	3.98	11.74	3.93
Korean	19.51	7.65	18.91	7.30	17.49	6.56
Spanish	25.69	10.50	22.98	9.16	22.72	8.92
Vietnamese	42.25	19.43	32.66	14.55	31.95	**14.25**

Finally, we experimented fine-tuning on $\mathcal{A}_{\mathbf{a}}$ and $\mathcal{S}_{\mathbf{a}}$ combined. As expected, when using otherwise identical train and dev sets, authentic data was more effective than synthetic. However, for three of nine accents, combining the two was more effective.

Next, using the same test sets from Table 2, we experimented with a large amount of synthetic data. Initially, we used 28,104 prompts from the Gutenberg literature corpus[10] [7] to synthesize off-domain speech (**Gutenberg synth.** in Table 3). Then, due to the drastically different text domain of this corpus (compared to our small authentic test set), we constructed large TTS audio sets out of prompts corresponding to the authentic files from our data sources (**Domain synth.** in Table 3). For CV we sampled 25,000 prompts from the original dataset (excluding German, Malaysian, and Filipino accents) and produced TTS files as before to create $\mathcal{C}_{\mathbf{a}}$. For L2-Arctic accents, the largest set of prompts we could create from combining all of the clean L2-Arctic prompts was 1853, resulting in only $\sim$700 train and dev sentences per L1-language once we removed

[10] https://github.com/geoph9/accent-adaptation-through-tts#synthesised-data-tts

prompts appearing in the respective test sets. We made up for the scarcity by changing our TTS approach: instead of uniformly sampling a TTS voice for each prompt, we used up to 6 TTS voices for each prompt to produce a larger set $\mathcal{L}_{\mathbf{a}}$.[11] Hence, this strategy resulted in training repeatedly on the same relatively small set of ∼700 prompts.

Next, we incorporated the large in-domain synthetic sets, $\mathcal{C}_{\mathbf{a}}$ for CV and $\mathcal{L}_{\mathbf{a}}$ for L2-Arctic, with the small authentic sets $\mathcal{A}_{\mathbf{a}}$ to fine-tune in two ways. First, we combined synthetic and authentic data and then up-sampled (i.e. duplicated) the authentic data to be as close to equal as possible to the synthetic data amount (**Comb. Up-samp.** in Table 4). Next, we kept synthetic and authentic sets separate, fine-tuning first on the synthetic, and then again on the authentic data (**Two-stage FT** in Table 4).

As discussed in § 3.1, our primary splitting method left very few train and dev data for L2-Arctic accents (∼40 minutes, as shown in Table 1). This could have a negative impact on both authentic and synthetic fine-tuning, since all of our synthetic augmentation methods for L2-Arcitc accents relied on the set of prompts present in the authentic train/dev sets. We ran additional experiments for three accents (Chinese, Korean, and Spanish), where we used all the prompts available with train and dev speakers for the train/dev data (again with an 80/20 split). This left only a small amount of viable test data that did not share any speakers or prompts with the train/dev data. (See **Test utts.** in Table 5 for the number of test utterances.) WER scores for three of our fine-tuning methods are in Table 5.

Table 5. WER for some L2-Arctic accents on small test sets with maximized train sets

Accent	Auth	Comb. Up-samp	Two-st. FT	train hrs	test utts.
Chinese	26.1	**17.6**	21.9	1.8	49
Korean	17.5	**13.1**	15.6	1.9	15
Spanish	23.4	22.4	**19.6**	1.8	10

5 Discussion and Analysis

Table 2 demonstrates that augmentation with authentic data is preferable to synthetic data, though combining the two yielded slightly improved results for three of the nine accents (German, Chinese, and Korean). Synthesized audio files, even in the same small quantities as authentic data, improved over baseline CER for four accents (Malaysian, Filipino, Chinese, and Hindi). Table 3 shows that increasing the amount of synthetic data alone, whether using prompts in the target domain (**Domain synth.**) or out of it (**Gutenberg synth.**), was ineffective across accents. This strategy likely caused the model to overfit on synthetic speech. The **Comb. Up-samp.** and **Two-stage FT** methods, combining synthetic and authentic data, consistently improved error rates over the

[11] Thus $|\mathcal{L}_{\mathbf{a}}| = \min(n_{\mathbf{a}}, 6) * N_{\mathbf{a}}$, where $N_{\mathbf{a}} \approx 700$ is the number of train/dev prompts available after removing prompts from the test set for $\mathbf{a}$, $n_{\mathbf{a}} = 2$ for Hindi and Vietnamese, and $n_{\mathbf{a}} \geq 6$ for other L2-Arctic accents. See Table 1.

baseline but only improved over simple authentic fine-tuning by small amounts for two accents (German and Filipino). Results in Table 5 represent small test sizes, however they suggest that settings with more training data may be conducive to greater success in the **Comb. Up-samp.** and **Two-stage FT** methods. The three accents displayed demonstrate WERR[12] of 32.6% (Chinese), 33.6% (Korean), and 16.2% (Spanish) compared to **Auth.** fine-tuning.

We acknowledge a limitation of our experimental setup. Scarcity of authentic accented data made it difficult to find diverse, representative test sets. This highlights the significance of synthetic augumentation improving on simple authentic fine-tuning in some cases. Each authentic set alone was advantaged, since it came from the same source as the test set. One potential advantage of synthetic augmentation is the expansion of model capabilities to more general settings. We hope future researchers will explore the benefits of our augmentation methods with more diverse test sets.

We explore the possible effect of TTS audio characteristics on suitability for augmentation. In Table 6 we show the effectiveness (***eff.***) of synthetic data augmentation, represented as the WERR% of our best-performing method involving TTS audio, compared to the best-performing method without TTS audio, from Tables 2, 3, and 4. We also show measures of TTS quality: average intelligibility (*intel.*) measured by ASR WER% using our wav2vec2.0 model to recognize TTS audio; average naturalness (***nat.***) measured by MOS score[13] [14]; and faithfulness in approximating the target accent (***accen.***). For this last characteristic we hired two proficient English speakers to rate an audio segment from each TTS voice on a five-point scale, where 5 corresponded to such a strong accent as to render the audio unintelligible and 1 corresponded to no accent at all. From these human annotations we calculated two accent scores. To measure accent excess, we counted ratings of 4 as one point and ratings of 5 as two points, then divided an accent's total points by its number of TTS voices $n_{\mathbf{a}}$. We calculated accent absence the same way, where a rating of 2 equaled one point, and a rating of 1 equaled two points. Table 6 shows average scores from the two evaluators, in the form: excess score / absence score.

Our analysis in Table 6 does not highlight any clear trends. Some accents with highly intelligible TTS and desirable accentedness (Filipino and Chinese) were more effective, but so were Korean (with poor **intel.** and **accen.** scores) and German. Interestingly, naturalness seems inversely correlated with effectiveness. And both accents displaying excessive accentedness (German and Korean) were more effective.

In summary, results from Tables 2, 3, and 4 suggest that augmentation by synthetic accented speech should be accompanied by a small authentic dataset to prevent overfitting on synthetic speech. In Tables 2, 3, 4, and 5 we find that synthetic data augmentation was only effective when authentic train data exceeded one hour (German and Filipino in Table 4 and experiments in Table 5). This may be in part so that authentic speech can give a strong signal in fine-tuning and not be drowned out by synthetic speech. A related factor is the diversity of prompts for TTS. **Comb. Up-samp.** and **Two-stage FT** models for L2-Arctic accents trained repeatedly on the same $\sim$700 prompts

[12] Calculated as $\frac{rate_{old} - rate_{new}}{rate_{old}}$

[13] We took both intelligibility and naturalness measurements over the dev set used for **Domain synth.**, with maximalized dev sets for L2-Arctic accents.

Table 6. TTS quality analysis. ***eff.***=effectiveness, ***intel.***=intelligibility, ***nat.***=naturalness, ***accen.***=accentedness, shown as excess/absence of an accent

Accent	*eff.* (↑)	*intel.* (↓)	*nat.* (↑)	*accen.* (↓/↓)
German	3.92	64.5	3.14	1.36/0.31
Malaysian	-13.3	68.6	3.41	0.88/0.13
Filipino	**4.86**	17.6	2.71	0.0/1.0
Arabic	-12.8	64.6	3.39	0.91/0.09
Chinese	2.28	**13.7**	3.16	0.07/1.01
Hindi	-3.53	20.6	**3.43**	0.25/0.50
Korean	4.15	93.2	3.07	1.94/0.0
Spanish	-5.56	51.8	3.22	0.80/0.43
Vietnamese	-0.06	89.8	3.17	2.0/0.0

and may have implicitly overtrained on them, rendering them ill-equipped to predict other prompts. This could explain why these two methods were ineffective in such settings but performed better for CV accents and in Table 5 (where larger training sets afforded larger prompt sets for augmentation).

6 Conclusion

The failure of many English ASR systems to accommodate non-native accents has a negative impact on the world's millions of L2-English speakers. We present a novel approach to assist in this problem, utilizing multilingual TTS systems with English prompts to approximate L2-accented speech and produce scalable augmentation data. We evaluated multiple realizations of this approach for ASR of 9 non-native English accents. Given our experiments and analysis, we find that TTS-based augmentation for accented ASR is best realized, and assists in error rate reductions for multiple accents, when accompanied by more than one hour of authentic speech and when sufficiently diverse target-domain TTS prompts are available.

References

1. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
3. Bhatt, R.M.: World englishes. Annual review of anthropology **30**(1), 527–550 (2001)
4. Cumbal, R., Moell, B., Águas Lopes, J.D., Engwall, O.: "you don't understand me!": Comparing asr results for l1 and l2 speakers of swedish. In: Interspeech 2021 (2021)
5. Das, N., Bodapati, S., Sunkara, M., Srinivasan, S., Chau, D.H.: Best of Both Worlds: Robust Accented Speech Recognition with Adversarial Transfer Learning. In: Interspeech 2021. pp. 1314–1318. ISCA (Aug 2021). https://doi.org/10.21437/Interspeech.

2021-1888, https://www.isca-speech.org/archive/interspeech_2021/das21b_interspeech.html
6. Du, C., Yu, K.: Speaker augmentation for low resource speech recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7719–7723 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053139
7. Gerlach, M., Font-Clos, F.: A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. Entropy **22**(1), 126 (2020)
8. Graddol, D.: The decline of the native speaker. Translation today: Trends and perspectives pp. 152–167 (2003)
9. Kadyan, V., Kathania, H., Govil, P., Kurimo, M.: Synthesis Speech Based Data Augmentation for Low Resource Children ASR. In: Karpov, A., Potapova, R. (eds.) Speech and Computer. pp. 317–326. Lecture Notes in Computer Science, Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_29
10. Kulkarni, K., Sengupta, S., Ramasubramanian, V., Bauer, J.G., Stemmer, G.: Accented indian english asr: Some early results. In: 2008 IEEE Spoken Language Technology Workshop. pp. 225–228 (2008). https://doi.org/10.1109/SLT.2008.4777881
11. Kuo, I.C.: Addressing the issue of teaching english as a lingua franca. ELT journal **60**(3), 213–221 (2006)
12. Laptev, A., Korostik, R., Svischev, A., Andrusenko, A., Medennikov, I., Rybin, S.: 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). pp. 439–444 (Oct 2020). https://doi.org/10.1109/CISP-BMEI51763.2020.9263564
13. Li, S., Ouyang, B., Liao, D., Xia, S., Li, L., Hong, Q.: End-To-End Multi-Accent Speech Recognition with Unsupervised Accent Modelling. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6418–6422 (Jun 2021). https://doi.org/10.1109/ICASSP39728.2021.9414833, iSSN: 2379-190X
14. Lo, C.C., Fu, S.W., Huang, W.C., Wang, X., Yamagishi, J., Tsao, Y., Wang, H.M.: MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion. In: Proc. Interspeech 2019. pp. 1541–1545 (2019). https://doi.org/10.21437/Interspeech.2019-2003
15. Markl, N., McNulty, S.J.: Language technology practitioners as language managers: arbitrating data bias and predictive bias in asr. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 6328–6339 (2022)
16. Martin, J.L.: Spoken corpora data, automatic speech recognition, and bias against african american language: The case of habitual'be'. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 284–284 (2021)
17. Mimura, M., Ueno, S., Inaguma, H., Sakai, S., Kawahara, T.: Leveraging Sequence-to-Sequence Speech Synthesis for Enhancing Acoustic-to-Word Speech Recognition. In: 2018 IEEE Spoken Language Technology Workshop (SLT). pp. 477–484 (Dec 2018). https://doi.org/10.1109/SLT.2018.8639589
18. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
19. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022)
20. Robinson, N.R., Ogayo, P., Gangu, S.R., Mortensen, D.R., Watanabe, S.: When Is TTS Augmentation Through a Pivot Language Useful? In: Proc. Interspeech 2022. pp. 3538–3542 (2022). https://doi.org/10.21437/Interspeech.2022-11203

21. Rossenbach, N., Zeyer, A., Schlüter, R., Ney, H.: Generating Synthetic Audio Data for Attention-Based Speech Recognition Systems. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7069–7073 (May 2020). https://doi.org/10.1109/ICASSP40776.2020.9053008, iSSN: 2379-190X
22. Shi, X., Yu, F., Lu, Y., Liang, Y., Feng, Q., Wang, D., Qian, Y., Xie, L.: The accented english speech recognition challenge 2020: Open datasets, tracks, baselines, results and methods. CoRR **abs/2102.10233** (2021), https://arxiv.org/abs/2102.10233
23. Tan, T., Lu, Y., Ma, R., Zhu, S., Guo, J., Qian, Y.: AISpeech-SJTU ASR System for the Accented English Speech Recognition Challenge. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6413–6417 (Jun 2021). https://doi.org/10.1109/ICASSP39728.2021.9414471, iSSN: 2379-190X
24. Ueno, S., Mimura, M., Sakai, S., Kawahara, T.: Data Augmentation for ASR Using TTS Via a Discrete Representation. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 68–75. IEEE, Cartagena, Colombia (Dec 2021). https://doi.org/10.1109/ASRU51503.2021.9688218, https://ieeexplore.ieee.org/document/9688218/
25. Zhang, Y., Zhang, Y., Halpern, B.M., Patel, T., Scharenborg, O.: Mitigating bias against non-native accents. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. vol. 2022, pp. 3168–3172 (2022)
26. Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J., Gutierrez-Osuna, R.: L2-arctic: A non-native english speech corpus. In: Proc. Interspeech. p. 2783–2787 (2018). https://doi.org/10.21437/Interspeech.2018-1110, http://dx.doi.org/10.21437/Interspeech.2018-1110
27. Zheng, X., Liu, Y., Gunceler, D., Willett, D.: Using Synthetic Audio to Improve the Recognition of Out-of-Vocabulary Words in End-to-End Asr Systems. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5674–5678 (Jun 2021). https://doi.org/10.1109/ICASSP39728.2021.9414778, iSSN: 2379-190X
28. Zhu, H., Wang, L., Zhang, P., Yan, Y.: Multi-Accent Adaptation Based on Gate Mechanism. In: Interspeech 2019. pp. 744–748. ISCA (Sep 2019). https://doi.org/10.21437/Interspeech.2019-3155, https://www.isca-speech.org/archive/interspeech_2019/zhu19_interspeech.html