Synthetic Voice Data for Automatic Speech Recognition in African Languages

Brian DeRenzi
Dimagi
bderenzi@dimagi.com

Anna Dixon
Dimagi*
anna@datafoss.com

Mohamed Aymane Farhi
CLEAR Global
aymenfarhi.25@gmail.com

Christian Resch†
CLEAR Global
christian.resch@clearglobal.org

Abstract

Speech technology remains out of reach for most of the 2,300+ languages in Africa. We present the first systematic assessment of large-scale synthetic voice corpora for African ASR. We apply a three-step process: LLM-driven text creation, TTS voice synthesis, and ASR fine-tuning. Eight out of ten languages for which we create synthetic text achieved readability scores above 5 out of 7. We evaluated ASR improvement for three (Hausa, Dholuo, Chichewa) and created more than 2,500 hours of synthetic voice data at below 1% of the cost of real data. W2v-BERT 2.0 speech encoder fine-tuned on 250h real and 250h synthetic data in Hausa matched a 500h real-data-only baseline, while 579h real and 450h to 993h synthetic data created the best performance. We also present gender-disaggregated ASR performance evaluation. For very low-resource languages, gains varied: Chichewa WER improved by ∼6.5% with a 1:2 real-to-synthetic ratio; a 1:1 ratio for Dholuo showed similar improvements on some evaluation data, but not on others. Investigating intercoder reliability, ASR errors and evaluation datasets revealed the need for more robust reviewer protocols and more accurate evaluation data. All data and models are publicly released to invite further work to improve synthetic data for African languages.

1 Introduction

Africa is home to over 2,300 languages, the vast majority of which have neither functional automatic speech recognition to transcribe speech nor speech synthesis to generate it (Orife et al., 2020). Yet speech technology holds great promise in providing a more inclusive digital experience, especially for vulnerable groups.

Conventional approaches to creating speech technology rely on human data collection in as-yet unsupported languages, incurring substantial costs estimated at more than US$100–150 per hour, even in the best case[1]. As most speech recognition models need several hundred hours of training data to achieve performance sufficient for practical application, with current investments in AI for development, human data collection is prohibitively costly to cover the many languages that remain unsupported. We support further investment in African language technology, but even if it were to become available, those investments have opportunity costs, diverting funds that could otherwise be spent on other interventions.

Therefore, we are investigating synthetic voice data as a complementary approach to create and improve automatic speech recognition (ASR) in African languages. The principle hypothesis motivating our work is that we can leverage Large Language Models (LLMs) and Text-to-Speech (TTS) models to create synthetic voice data of sufficient quality to improve automatic speech recognition models. Our work shows that this synthetic voice data can be created for less than 1% of the cost of collecting real human data[2], while holding potential to complement this human data in creating and improving ASR models for African languages.

2 Synthetic Data: Risks and Rewards

Synthetic data—machine-generated text or speech—has become a well-researched topic in major languages like English in recent years (for an overview, see Liu et al. (2024)). Much of this research is motivated by concerns that all readily

*Current affiliation: Datafoss
†Authors are listed in alphabetical order. Corresponding author: Christian Resch

[1]Based on internal estimates. The authors of NaijaVoices report that the 'true cost' of the whole dataset is more than US$600,000. Assuming an equal split across the three languages in the dataset, this would imply a true cost per hour of Hausa data of US$ ∼345.42, see https://naijavoices.com/membership.

[2]This is excluding fixed costs in both cases, like setting up data collection platforms for real data or TTS model development.

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 152–186, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-016

available data for the development of AI models has been leveraged (Villalobos et al., 2024), and therefore further advancements spurred by data scaling will require new approaches to create data. Other concerns like privacy (Abay et al., 2018) or costs (Gilardi et al., 2023) also prompted research on synthetic data. Additionally, there is a growing body of research into synthetic data in low-resource languages which, save for several notable exceptions, so far only offers limited evidence for African languages.

For automatic speech recognition, Huang et al. (2023) have shown that for English, synthetic data generation using large-scale pre-trained neural networks in combination with TTS models, a process similar to ours, can reduce word error rate (WER) by between 9 and 15% (see also Moslem (2024) and Hu et al. (2021)). Gokay and Yalcin (2019) report reductions of around 15% in WER for Turkish, Yang et al. (2024) between 27.45 and 45.85% for Chinese dialects and Zevallos (2022) 8.73% for Quechua. Joshi et al. (2025) find that adding synthetic data in Hindi improves Bhojpuri ASR performance by 4.7 points WER on average. Wang et al. (2020) also demonstrate that an approximate 50/50 combination of human and synthetic data performed comparably to the same amount of human data alone. In a very successful application, Xu et al. (2020) achieve 17% WER for Lithuanian using only 1.3 hours of labeled and 12 hours of unlabeled data.[3]

As this research shows, there is a body of research on synthetic data for low-resource languages, but African languages have been rarely covered. There are notable exceptions for synthetic text (Abdulmumin et al., 2022; Kreutzer et al., 2022), linguistically informed data augmentation and synthetic data frameworks (Ajuzieogu, 2023) and synthetic text for language and topic classification models (Quinjica and Adelani, 2024; Adelani et al., 2024).

While previous research shows the potential of synthetic data to complement human data in the data-scarce situation that we face in many African languages, the research also highlights limitations and risks. Most limitations stem from the gap between real and synthetic data (Hu et al., 2021), as well as from synthetic data inheriting and potentially amplifying the same biases as the models used to create it (Wyllie et al., 2024; Wang et al., 2025). Certain tone and noise that are typically present in real-world data are often missing from synthetic data (Xue et al., 2022; Hu et al., 2021). Many commonly used LLMs have been shown to exhibit a bias towards Western, industrialized cultural norms and a lack of cultural understanding in other contexts (Rao et al. (2023), Magdy et al. (2025) for Arabic, Pranida et al. (2025) and Putri et al. (2024) for Sundanese and Indonesian). Creating synthetic text will likely aggravate this missing representation, especially for semantically meaningful tasks such as machine translation.

3 Methodology

Our process of creating and evaluating synthetic voice data has three key steps:

1. Generate and evaluate synthetic text using an LLM

2. Generate and evaluate synthetic voice data with a Text-to-Speech (TTS) model based on the synthetic text

3. Fine-tune an automatic speech recognition (ASR) model with different ratios of human and synthetic voice data and evaluate performance differences

We describe our methods and details for those steps separately:

3.1 Step 1: Synthetic Text Generation and Evaluation

We created and evaluated synthetic text for the following 10 African languages: Hausa, Northern Somali, Yoruba, Wolof, Dholuo, Kanuri, Chichewa, Twi, Kinande, and Bambara. Additionally, we included small-scale generations and evaluations for two further languages, Yemba and Ewondo, but with only very poor performance. Our language selection criteria aimed to select languages that represent African language diversity through the representation of different regions, language families, and speaker populations (see Appendix A). Beyond linguistic diversity, we also considered practical factors, such as the capacity of the Translators without Borders (TWB) linguist community

[3]We recommend not to compare WER between languages because of substantial morphological differences, i.e. depending on the language a single word might carry different semantic meaning while it is always counted individually in the WER. Furthermore, evaluation datasets such as FLORES and, by extension, FLEURS have known shortcomings which differ by language, see Abdulmumin et al. (2024) for an investigation for African languages.

for text evaluation and the availability of key publicly available datasets (e.g., open.bible, FLEURS, Common Voice) necessary for subsequent steps.

For the synthetic text generation, the system prompt (see Appendix B) instructs the LLM to generate simple short sentences and questions directly in the target language, as well as to return English translations for further evaluation of language understanding. We incorporated two-shot prompting for contextual guidance. The topic of synthetic text generation can be configured in the prompt. Our experiments sampled equally from 34 distinct themes with 17 themes covering the UN Sustainable Development Goals and 17 themes covering the most common topics covered in the FLORES/FLEURS dataset extracted through language topic modelling. We primarily evaluated LLMs provided by OpenAI and Anthropic, especially GPT-4o, GPT-4.5, o1, Claude 3.5 and Claude 3.7.

Automated evaluation of the synthetic text is not feasible for low-resource languages and we relied on human evaluation. For each language we generated 1,200 randomly shuffled sentences over two rounds for a few different configurations (usually three LLMs).[4] Linguists on the Translators without Borders (TWB) platform were sourced according to their native language and experience delivering linguistic tasks. TWB linguist reviewers rate each sentence on five key metrics intended to capture the quality of the sentence in the target language and understanding: 1) Readability and Naturalness [1–7], 2) Grammatical Correctness (Yes/No), 3) All Words Real (Yes/No), 4) Notable Error in Translation (Yes/No), and 5) Adequacy and Accuracy of Translation [1–7]. Based on the human evaluation, we selected the configuration (LLM) with the highest mean Readability and Naturalness.

For the subset of languages selected for subsequent synthetic voice generation and ASR fine-tuning, we generated a large corpora of between 650,000 and 674,000 sentences with the best performing LLM (see Appendix C for details). The text generation process was identical to that used for text evaluation, except that we utilized a batch processing API for reduced cost.

[4]For most experiments, we opted for a two-round sentence generation and evaluation approach, in which we first compared the readability and naturalness of 600 sentences generated by 3-4 LLMs and subsequently generated another 600 sentences using the best model to analyze the impact of theme. After discovering that our experiments were frequently not yielding significant differences between themes, we opted for more equal sampling among different LLMs.

3.2 Step 2: Synthetic Voice Data Generation and Evaluation

Based on the results of the evaluation of the synthetic text generation in Step 1, we selected three languages for the creation and evaluation of synthetic voice data: Hausa, Dholuo and Chichewa.

As necessary conditions, we required languages where at least one LLM was capable of generating synthetic text of sufficient quality. Due to limitations in time and resources, we also required that at least two of the three languages have available data for fine-tuning or training TTS models, as well as available ASR evaluation data.

Beyond those necessary conditions, our goal was to maximize the variance of the speaker populations, available ASR training data, language families, and geography.

We used the open.bible corpus (Global Bible Initiative, n.d.) of Bible recordings to fine-tune and evaluate different TTS models. We excluded this data from our ASR training data for this reason. The open.bible corpus only contains recordings of Bible recitations by male speakers, and given the training data which we used, our synthetic voice data is also exclusively male. This raises the risk that the resulting ASR models show gendered performance, e.g. that they perform worse for female speakers than for male speakers. To investigate this risk, we also evaluated gender bias in the ASR performance where the evaluation data allows this.

For each language, we fine-tuned both the XTTS-v2 model (Casanova et al., 2024) and VITS or one of its variants, specifically YourTTS (Casanova et al., 2023), using the Coqui TTS framework, and building on the BibleTTS project (Meyer et al., 2022). XTTS-v2 does not support any African languages, but has been fine-tuned for Wolof.[5] For fine-tuning YourTTS, we used the checkpoint trained on the CML-TTS dataset (Oliveira et al., 2023) that supports eight languages. We used the original BibleTTS model for Hausa, but also retrained the model based on a revised processing of the open.bible corpus, a different checkpoint and different hyperparameters ("Modified Bible TTS"). For Hausa and Chichewa, we also evaluated the available MMS TTS models (Pratap et al., 2023)[6].

Transformer-based TTS models like XTTS have

[5]https://huggingface.co/galsenai/xTTS-v2-wolof

[6]The MMS TTS model language coverage is available at https://dl.fbaipublicfiles.com/mms/misc/language_coverage_mms.html

the problem of hallucinating, especially at the end of audio files. To remedy this issue, we re-transcribed the synthetic audio files with an existing ASR model and then calculated the ratio of the length of the transcript to the length of the original synthetic text. We only generate synthetic data with XTTS for Hausa and used the MMS-1B (Pratap et al., 2023) for this analysis. This method does not rely on the accuracy of the ASR model used but assumes a basic performance to ensure that the length of the re-transcription is meaningful.[7] We finally removed outliers of this ratio of less than 0.85 and more than 1.06 which indicate that the TTS model had hallucinated additional words not present in the original synthetic text or omitted words that were. We picked the cut-off points by manually expecting ∼100 samples. This process removed ∼26.9% of the synthetic audio created by XTTS.[8]

After training a total of ten TTS models across all three languages, we evaluated 337 synthetic audio samples per model with the help of two native speakers from the TWB Community per language. As commonly applied, our evaluation included intelligibility and naturalness on five-point scales.

We then selected YourTTS, the best-performing TTS model across all three languages, to create synthetic voice data corpora of 993h for Hausa, 775h for Dholuo and 550h for Chichewa (see Appendix C for details). For Hausa, we also created 450h of synthetic data with XTTS, the only transformer-based model. We make these synthetic data corpora openly available on CLEAR Global's Hugging Face page[9].[10]

To improve the robustness of our models to noisy acoustic environments, we augmented the synthetic data by adding noise. We mixed the clean synthetic data with noise samples drawn from the Room Impulse Response and Noise Database[11]. For each utterance, we randomly sampled the signal-to-noise ratio (SNR) from a normal distribution with mean 50dB and standard deviation 15dB. Similarly, we randomized the audio amplitude using a normal distribution ($\mu = -20\,\mathrm{dB},\ \sigma = 5\,\mathrm{dB}$).

3.3 Step 3: ASR Model Fine-tuning and Evaluation

Given the substantial differences in available ASR training data between the three languages for which we created synthetic voice data, we conducted our ASR evaluation based on two scenarios: a medium data scenario with Hausa as the representative language and a low data scenario with Dholuo and Chichewa as the representative languages.

3.3.1 Medium Data Scenario: Hausa

Through the NaijaVoices project (Emezue et al., 2025), we had over 500 hours of human Hausa voice data available. Only a few other African languages like Igbo and Yoruba (NaijaVoices) or Swahili, Kinyarwanda, Kabyle, and Luganda (Common Voice) have available datasets of comparable size. This led us to investigate whether synthetic voice data can substitute human data at this training corpus size, therefore allowing languages with smaller corpora to achieve similar ASR performance. In this scenario, we keep the *total size of the training data corpus constant*, but *vary the ratio between real and synthetic data*.

As a result, we investigated ASR performance for training with 500h of real data, a 1:1 ratio of 250h of real and 250h of synthetic data, and a 1:4 ratio of 100h of real data and 400h of synthetic data. With 100h and 250h of real data, this also covers scenarios that, while not currently the case, are realistically achievable for many African languages. We trained models for all data ratios with synthetic data created with YourTTS and XTTS separately.

We also needed to rule out the case that ASR models might saturate at a given amount of human training data of one single source, meaning that comparable performance at different ratios between real and synthetic stems from the model being saturated (e.g. showing no or only very low marginal improvements beyond 100h of real data). We therefore also trained the same models on only 100h and 250h of real data.

[7] In most cases in which this method would be applied, a minimum of real data to train a basic but not highly capable ASR model should be available (e.g. for the complementary human data in later ASR training or from the training data of the TTS model). In addition, models like MMS cover a large number of languages albeit often with only poor performance which should still be sufficient for this approach.

[8] After filtering the synthetic voice data, the dataset created with XTTS consists of a substantially larger share of questions (∼40%), indicating that XTTS hallucinates less for questions than for normal sentences. To avoid bias in our synthetic voice training data, we sampled a subset of these questions to create a dataset with the original share of questions (25%), resulting in a smaller dataset of 450 hours and removal of ∼42.7% of the original data.

[9] https://huggingface.co/CLEAR-Global

[10] We also created a large Chichewa text corpus with Claude 3.7 as part of our investigation of duplicates. We created the synthetic voice data based on the Claude 3.5 corpus which we had evaluated before but also make the Claude 3.7 text corpus available on CLEAR Global's HuggingFace page.

[11] https://www.openslr.org/28/

Finally, we trained ASR models on all data available to us: one model on 579h of real human data mixed with 993h of synthetic data created with YourTTS and one model with 579h of real data mixed with 450h of synthetic data created with XTTS.

We evaluated the ASR performance on our NaijaVoices test set split, the FLEURS test set (Conneau et al., 2022), and the Common Voice test set (Mozilla Foundation, 2024a). We conducted the analysis of gender bias in ASR performance on the NaijaVoices and Common Voice test sets.[12]

Since the NaijaVoices dataset did not provide splits at the time, we performed a split to generate train, validation, and test sets that contain 579.1, 3.6, and 3.4 hours of data, respectively. We ensured the per-split sets of speakers and transcriptions are mutually exclusive.[13]

3.3.2 Low Data Scenario: Dholuo and Chichewa

In contrast to Hausa, we only had 19 and 34 hours of usable human data available for Dholuo and Chichewa, respectively.[14] Although this is generally insufficient data to train general purpose ASR ready for practical application, this is representative of many African languages. As the human data available is itself insufficient, we kept the total amount of human data constant in this scenario and *added increasing amounts of synthetic data to the training corpus*. The total size of the training corpus consequently increases in this scenario.

For both languages, we trained ASR models on just the human data available, and 1:1, 1:2, and 1:4 ratios of human and synthetic data. Given some indications of improvement for Chichewa, we also trained a 1:9 ratio of 34h of human and 307h of synthetic data.

We evaluated the Dholuo ASR models on the FLEURS test set (Conneau et al., 2022) and the Common Voice test set (Mozilla Foundation, 2024b), and the Chichewa ASR models on the FLEURS test set and the Zambezi Voice test set (Sikasote et al., 2023).

[12]The Hausa FLEUR test set only includes a single male speaker.

[13]Splits are now available on their Huggingface website `https://huggingface.co/datasets/naijavoices/naijavoices-dataset/tree/main/split`

[14]We use 10 hours from Common Voice (Mozilla Foundation, 2025; Ardila et al., 2020) and 9 hours from the FLEURS train set (Conneau et al., 2022) for Dholuo. We use 10 hours from the FLEURS train set and 24 hours from Zambezi Voice (Sikasote et al., 2023) for Chichewa.

3.3.3 ASR Model Selection and Evaluation

For step 3, we fine-tuned the W2v-BERT 2.0 speech encoder (Communication et al., 2023), for which our results indicated continued improvement for fine-tuning with 100h and 250h of real data.[15] This model was pre-trained on 4.5M hours of unlabelled audio data covering more than 143 languages. The pre-training crucially includes Hausa but not Dholuo and Chichewa. Details on our hyperparameters can be found in Appendix O.

We estimated the confidence intervals for WER and CER by performing bootstrap resampling on each evaluation set (Raschka, 2020; Efron, 1992). For each of 1,000 iterations[16], we randomly drew m samples with replacement—where m equals the size of the original test set—and computed WER and CER of each resampled set. We then calculated the mean and standard deviation of WER and CER across all bootstrapped samples.

4 Results and Discussion

4.1 Synthetic Text Generation

For 8 of 10 languages, at least one LLM generated sentences with a Readability and Naturalness rating mean greater than 5.0 on a seven-point scale (see Figure 1). In general, we found that Claude 3.5 Sonnet performed the best for the languages studied here, outperforming OpenAI's GPT-4o and o1 models for 8 of 10 languages. Summary statistics aggregated by language and LLM for all metrics examined are provided in Appendix D.

Of the languages studied here, Kanuri and Kinande are classified as category 0 (lowest resource, "The Left-Behinds") according to the taxonomy established by Joshi et al. (2021).[17] This data scarcity directly impacts the effectiveness of LLMs in these languages, as evidenced by our findings:

[15]We also fine-tuned the MMS-1B model (Pratap et al., 2023) on the Hausa subset of the NaijaVoices dataset using adapters. We found that the model's performance doesn't improve or only improves marginally by adding real data beyond 50 hours, and the addition of synthetic data consistently degrades performance (see Appendix I for results). This aligns with research by Nabende et al. (unpublished) who compare different ASR architectures for African languages and their data scaling behavior.

[16]This is five times more than the usual number of iterations between 50 and 200 recommended by Efron and Tibshirani (1994) and in line with what Koehn (2004) proposes for similar applications in machine translation.

[17]While this categorization is partially outdated, only limited data collection has taken place in those low-resourced languages. Of the languages we studied, Dholuo was not classified by Joshi et al. (2021), but would probably be classified as category 0 or 1.

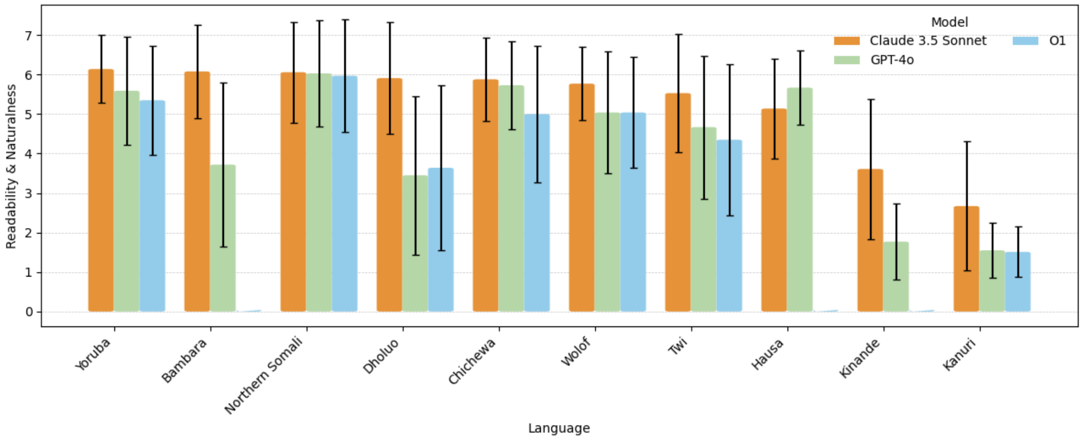


Figure 1: Comparison of Readability and Naturalness [1..7] scores for synthetic text generated by various LLMs for 10 African languages. Errors bars show standard deviation.

Synthetic text generation in these languages consistently demonstrates the poorest performance, with mean Readability and Naturalness ratings falling below 4.0 on a seven-point scale.[18] During the large-scale Chichewa text generation, we observed large amount of sentence duplication: Claude 3.5 Sonnet generated only 37% unique sentences out of 700,000, while Claude 3.7 Sonnet exhibited significantly less duplication, with 86% of 530,400 sentences being unique. Subsequent analysis revealed that unique sentence generation decreases with batch size (see Appendix G for details).

4.1.1 Inter-coder Reliability Investigation

A persistent challenge in low-resource language evaluation is the limited availability of expert linguist reviewers, a constraint that significantly impacts the reliability of assessments. This study was no exception, with most text samples being evaluated by two to three linguists per language. A two-way analysis of variance (ANOVA) demonstrates that not just model choice, but also linguist identity significantly affected readability ratings ($p < 0.05$), with linguist identity explaining a larger proportion of variance than the model itself for Chichewa, Kanuri, Northern Somali, and Wolof.

In a supplementary analysis for Kanuri, ten linguists independently rated 100 sentences each generated by Claude 3.5 Sonnet, Claude 3.7 Sonnet, GPT-4o, and GPT-4.5. We performed a bootstrap analysis, resampling varying numbers of raters. Increasing the number of raters consistently narrowed the 95% confidence intervals across all models, indicating improved rating stability. Especially GPT-4o benefitted from additional raters with the 95% confidence interval reducing from 5.86 to 3.91 when increasing the number of raters from two to four (further details are available in Appendix F).

4.2 Synthetic Voice Generation

Our findings indicate that the TTS model, albeit not its architecture, has a substantial impact on synthetic voice data quality, with the best model outperforming the worst by up to 2.03 points for intelligibility and 1.72 points for naturalness on a five-point scale. The VITS-based YourTTS models generally performed best (Intelligibility: Hausa: 4.5, Dhuluo: 4.71, Chichewa: 4.45), although VITS-based MMS performed better in naturalness for Chichewa (MMS: 4.03 versus YourTTS: 3.82) where we prioritized intelligibility and therefore YourTSS (detailed results are available in Appendix H). We find that the quality of the TTS model probably does not matter beyond a certain threshold. The Hausa ASR models trained on synthetic data generated using YourTTS don't outperform those trained on data generated by XTTS, even though the former TTS model outperforms the latter on both intelligibility and naturalness.

4.3 ASR Model Performance with Synthetic Data

4.3.1 Medium Data Scenario: Hausa

While the performance differs between the evaluation datasets, in general, replacing half of the human data with synthetic data, i.e. 250h:250h, in model training yields performance equally or marginally better than a model trained on 500h of

[18] We explored fine-tuning and a separate language quality classifier to retain only high quality out put but both did not yield improved results. Details are available in Appendix E.

Data ratios	FLEURS		NaijaVoices		Common Voice	
	WER	CER	WER	CER	WER	CER
500h constant						
100h:400h	28.58 (28.63 ± 0.86)	10.64 (10.04 ± 0.47)	24.57 (24.57 ± 0.33)	6.26 (6.26 ± 0.12)	17.95 (17.94 ± 0.67)	3.73 (3.73 ± 0.16)
250h:250h	26.17 (26.23 ± 0.65)	9.01 (9.03 ± 0.42)	22.91 (22.92 ± 0.31)	5.84 (5.84 ± 0.17)	18.69 (18.68 ± 0.67)	3.67 (3.73 ± 0.16)
500h:0h	26.91 (26.9 ± 0.67)	9.64 (9.63 ± 0.47)	22.49 (22.5 ± 0.34)	5.71 (5.72 ± 0.11)	17.91 (17.9 ± 0.67)	3.60 (3.61 ± 0.15)
Full data						
579h:450h XTTS	**25.73** (25.75 ± 0.63)	**8.96** (8.97 ± 0.44)	22.43 (22.42 ± 0.33)	5.74 (5.74 ± 0.11)	18.16 (18.17 ± 0.67)	**3.44** (3.44 ± 0.15)
579h:993h YourTTS	28.42 (28.47 ± 0.98)	11.22 (11.27 ± 0.79)	**22.06** (22.06 ± 0.3)	**5.64** (5.64 ± 0.12)	**17.45** (17.42 ± 0.66)	3.45 (3.45 ± 0.15)

Table 1: WER and CER for Wav2Vec-Bert 2.0 Hausa models trained on different ratios of real and synthetic XTTS-generated data. In parentheses, we present bootstrapped mean and standard deviation WER and CER.

real data for models trained with XTTS-generated data (see Table 1).[19] On the Common Voice test set, the model trained on a 1:4 ratio performs equally to the model trained on 500h of real data. The best performing model across most evaluation sets and metrics resulted from training on all data available, albeit with only minor improvements.

The data ablation study indicated that the ASR models does not saturate at 100h or 250h of data as the models trained with synthetic data still showed the same slight improvements as with adding real data.

Investigating gender bias, we found that on average the fine-tuned models perform slightly worse for male voices than for female voices (see the Appendix M for detailed results), despite our synthetic voice data being exclusively male. The gender-disaggregated performance on the NaijaVoices and Common Voice test sets showed an average difference in WER/ CER of $\sim 1.82/ \sim -0.17$ and $\sim 1.29/ \sim 0.57$ percentage points, respectively (positive numbers indicate worse performance for male speakers).

4.3.2 Low Data Scenario: Dholuo and Chichewa

For Dholuo and Chichewa, we added increasing amounts of synthetic data to increase the total training corpus. Therefore, we would expect improvements in performance as the training corpus size increases.

For Dholuo, the results depended on the test set. On FLEURS, no amount of synthetic data improved the WER and improvements in CER were not statistically significant (e.g. the improvement 6.06 to 6.00 CER when adding 77h of synthetic data is well within the standard deviation of 0.29 and 0.23, respectively). On Common Voice, adding 19h of synthetic data for a 1:1 ratio improved performance from 30.64 ± 0.51 to 28.76 ± 0.46 WER and from 6.99 ± 0.22 to 6.09 ± 0.15 CER. Adding further synthetic data did not yield further improvements. Full results are available in Appendix K.

In contrast, for Chichewa (see Table L.1), we found consistent improvements when adding synthetic data. Adding 68 hours of synthetic data for a 1:2 ratio between real and synthetic data and adding 307h of synthetic data for a 1:9 ratio resulted in the best performing models. On Zambezi Voice, the 1:2 ratio yielded the best absolute performance of 18.54 ± 0.71 WER and 4.41 ± 0.38 CER although this is not statistically better than the performance of a 1:9 ratio (18.71 ± 0.70 WER and 4.48 ± 0.38 CER, with the 34h:0h baseline resulting in 19.76 ± 0.70 WER and 4.52 ± 0.40 CER). Evaluation on the FLEURS test set confirms these results, only that the 1:9 exhibited best absolute WER performance (34h:0h: 35.39 ± 0.59 WER and 7.67 ± 0.40 CER, 34h:68h: 33.4 ± 0.59 WER and 7.15 ± 0.36 CER, 34h:307h: 32.95 ± 0.61 WER and 7.25 ± 0.38 CER, full results are available in Appendix L).

Unfortunately, the available evaluation data did not allow an analysis of performance by gender as all speakers were either female (Dholuo), male (Chichewa FLEURS test set) or metadata wasn't available (Zambezi Voice).

[19]More details on models trained on YourTTS-generated data and on the data ablation are available in Appendix J.

4.3.3 Evaluation Challenges Due to Non-standardized Scripts and Potential Errors in Evaluation Data

Spot checks of the errors by different models indicated that part of the word error rate might be due to non-standardized scripts and diacritics in Hausa, Dholuo, and Chichewa, where different but equally legitimate ways of writing the same word are counted as errors and where diacritics are not consistently used or transcribed. This aligns with work on potential limitations of WER (Aksënova et al., 2021) and benchmarking for Indic languages (Watts et al., 2024).

Although a thorough analysis of this issue is beyond the scope of this paper, we extracted the words from the evaluation transcripts that were most often incorrectly transcribed. Native speakers than evaluated these errors. This analysis indicates that we potentially underestimate the ASR performance. Of 19 to 20 errors per language evaluated, the evaluators labeled all 19 as no errors for Dholuo, 20 as no errors for Hausa (with 4 errors in the evaluation transcript), and 2 out of 20 as no errors for Chichewa (see Appendix N for examples). Those wrongly labeled errors often stem from varying use of special characters or different, but equally legitimate, spellings. Those results also imply that comparisons of WER between languages are not robust, as those issues differ between languages in kind and number. Text normalization as a potential remedy is an ongoing field of research. However, some research has shown that current techniques might not be appropriate for low-resource languages with non-Latin scripts (Manohar et al., 2024).

5 Conclusion

We investigated the creation of synthetic text and voice data for 10 African languages. Our results show that synthetic text generation with LLMs is feasible for various languages, except for the lowest-resource languages such as Kanuri or Kinande. Our results also show promising utility of synthetic voice data in complementing human data when training ASR models. But our results also indicate that a minimum of human data is needed. For Hausa, we show that the use of synthetic data either worsens the performance for male voices or does not increase gender bias in ASR performance, depending on the evaluation dataset and in spite of our synthetic data only including male voices. Further investigations also illustrated the challenges of working with human evaluators in low-resource languages where code-mixing and non-standardized scripts are common, as well as the limitations and shortcomings of existing evaluation datasets and resulting metrics.

Limitations

Our work is limited to the selected languages, and future research would need to expand the language coverage of studies on synthetic data for African languages. In addition, we could only explore a certain set of parameters for our data generation pipeline and model training. As illustrated in the previous sections, our results are also limited by potential issues in the evaluation datasets that we used, despite their common usage. Furthermore, we present our findings on challenges in human evaluation for low-resource languages, which we think require further investigation.

Future Work

Our research could show the utility of synthetic voice data in a controlled setting on commonly used evaluation sets. Future research should further investigate the robustness of synthetic data for use in practical applications. This might include investigating the utility of multiple-speaker TTS and voice cloning based on very small voice samples to create more diverse or targeted synthetic data (Ogun et al., 2024; Yang et al., 2024). Future work should also investigate methods and effectiveness of increasing text diversity and options to target text generation to specific domains and use cases (see Yang et al. (2024), Chen et al. (2024) and Finch and Choi (2024)). Our work illustrates the challenges in human evaluation. Future work to improve intercoder reliability should include better evaluation guidelines and identification of key metrics indicating downstream performance. The examples presented illustrate that beyond synthetic data, ASR evaluation in low-resource languages requires further investigation to handle non-standardized scripts, either through semantic measures, measures robust to plurality in spellings or language-appropriate normalizers, and work on improved evaluation datasets. Lastly, further work should be undertaken to investigate other uses of synthetic data in African languages.

Acknowledgements

CLEAR Global is grateful for the support of the Gates Foundation that enabled this work and the sub-award and partnership with Dimagi. The authors are indebted to Polly Harlow and Arisha Siddiqui, whose organizational skills we could not replace. We are also grateful to Daniel Wilson at XRI Global, Muhammad Abdul-Mageed and his team at the University of British Columbia, and Howard Lakougna at the Gates Foundation, who provided trusted partnership and valuable feedback. We thank Alp Öktem at CLEAR Global for review and feedback and Joyce Nabende, Alvin Nahabwe, and the team at Makerere University for the close collaboration and exchange around their related project on ASR in African languages, which informed and strengthened our work. Lastly, we would like to thank the evaluators from the TWB Community, without whom we could not have implemented this research.

This work was supported by the Gates Foundation (Grant number INV-076358). The conclusions and opinions expressed in this work are those of the authors alone and shall not be attributed to the Foundation.

References

Nazmiye Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2018. Privacy preserving synthetic data release using deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 510–526.

Idris Abdulmumin, Michael Beukman, Jesujoba Alabi, Chris Chinenye Emezue, Everlyn Chimoto, Tosin Adewumi, Shamsuddeen Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Gwadabe. 2022. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1001–1014, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Uchechukwu Ajuzieogu. 2023. *Ethical Data Augmentation Techniques for Low-Resource Language AI: A Framework for African Languages*. Ph.D. thesis, University of Nigera.

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34, Online. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

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.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. 2023. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I. Abdin. 2024. On the diversity of synthetic data and its impact on training large language models.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual expressive and streaming speech translation.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech.

Anna Dixon. 2024. Talk at openai devday 2024, community spotlight: Dimagi. `https://www.youtube.com/watch?v=Cj4MU_CJhwU`. Accessed: May 2025.

Bradley Efron. 1992. Bootstrap methods: Another look at the jackknife. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 569–593. Springer New York, New York, NY.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages.

James D. Finch and Jinho D. Choi. 2024. Diverse and effective synthetic data generation for adaptable zero-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12527–12544, Miami, Florida, USA. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Global Bible Initiative. n.d. Open.bible translation corpus [data set]. `https://open.bible`. Accessed May 2025.

Ramazan Gokay and Hulya Yalcin. 2019. Improving low resource turkish speech recognition with data augmentation and tts. In *2019 16th International Multi-Conference on Systems, Signals and Devices (SSD)*, pages 357–360.

Ting-Yao Hu, Mohammadreza Armandpour, Ashish Shrivastava, Jen-Hao Rick Chang, Hema Koppula, and Oncel Tuzel. 2021. Synt++: Utilizing imperfect synthetic data to improve speech recognition.

Zhuangqun Huang, Gil Keren, Ziran Jiang, Shashank Jain, David Goss-Grubbs, Nelson Cheng, Farnaz Abtahi, Duc Le, David Zhang, Antony D'Avirro, Ethan Campbell-Taylor, Jessie Salas, Irina-Elena Veliche, and Xi Chen. 2023. Text generation with speech synthesis for asr data augmentation.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world.

Sakshi Joshi, Eldho Ittan George, Tahir Javed, Kaushal Bhogale, Nikhil Narasimhan, and Mitesh M. Khapra. 2025. Recognizing Every Voice: Towards Inclusive ASR for Rural Bhojpuri Women. In *Interspeech 2025*, pages 4243–4247.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Tai K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163. Erratum in: J Chiropr Med. 2017 Dec;16(4):346. doi: 10.1016/j.jcm.2017.10.001.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12320–12341, Albuquerque, New Mexico. Association for Computational Linguistics.

Kavya Manohar, Leena G Pillai, and Elizabeth Sherly. 2024. What is lost in normalization? exploring pitfalls in multilingual asr model evaluations.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. 2022. Bibletts: a large, high-fidelity, multilingual, and uniquely african speech corpus.

Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation.

Mozilla Foundation. 2024a. Common voice dataset version 17.0 [data set]. `https://commonvoice.mozilla.org/en/datasets`. Accessed May 2025.

Mozilla Foundation. 2024b. Common voice dataset version 19.0 [data set]. `https://commonvoice.mozilla.org/en/datasets`. Accessed May 2025.

Mozilla Foundation. 2025. Common voice dataset version 21.0 [data set]. `https://commonvoice.mozilla.org/en/datasets`. Accessed May 2025.

Sewade Ogun, Abraham T. Owodunni, Tobi Olatunji, Eniola Alese, Babatunde Oladimeji, Tejumade Afonja, Kayode Olaleye, Naome A. Etori, and Tosin Adewumi. 2024. 1000 african voices: Advancing inclusive multi-speaker multi-accent speech synthesis.

Frederico S. Oliveira, Edresson Casanova, Arnaldo Cândido Júnior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. Cml-tts a multilingual dataset for speech synthesis in low-resource languages.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane – machine translation for africa.

Salsabila Zahirah Pranida, Rifo Ahmad Genadi, and Fajri Koto. 2025. Synthetic data generation for culturally nuanced commonsense reasoning in low-resource languages.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can llm generate culturally relevant commonsense qa data? case study in indonesian and sundanese.

Osvaldo Luamba Quinjica and David Ifeoluwa Adelani. 2024. Angofa: Leveraging ofa embedding initialization and synthetic data for angolan language model.

Abhinav Sukumar Rao, Aditi Khandelwal, Kumar Tanmay, Utkarsh Agarwal, and Monojit Choudhury. 2023. Ethical reasoning over moral alignment: A case and framework for in-context ethical policies in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13370–13388, Singapore. Association for Computational Linguistics.

Sebastian Raschka. 2020. Model evaluation, model selection, and algorithm selection in machine learning.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

C. Sikasote, K. Siaminwe, S. Mwape, B. Zulu, M. Phiri, M. Phiri, D. Zulu, M. Nyirenda, and A. Anastasopoulos. 2023. Zambezi voice: A multilingual speech corpus for zambian languages. In *Proceedings of Interspeech 2023*, pages 3984–3988.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Will we run out of data? limits of llm scaling based on human-generated data.

Gary Wang, Andrew Rosenberg, Zhehuai Chen, Yu Zhang, Bhuvana Ramabhadran, Yonghui Wu, and Pedro Moreno. 2020. Improving speech recognition using consistent predictions on synthesized speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7029–7033.

Ze Wang, Zekun Wu, Jeremy Zhang, Xin Guan, Navya Jain, Skylar Lu, Saloni Gupta, and Adriano Koshiyama. 2025. Bias amplification: Large language models as increasingly biased media.

Ishaan Watts, Varun Gumma, Aditya Yadavalli, Vivek Seshadri, Manohar Swaminathan, and Sunayana Sitaram. 2024. Pariksha: A large-scale investigation of human-llm evaluator agreement on multilingual and multi-cultural data.

Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: Training on synthetic data amplifies bias.

Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition.

Shaofei Xue, Jian Tang, and Yazhu Liu. 2022. Improving speech recognition with augmented synthesized data and conditional model training. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 443–447.

Guanrou Yang, Fan Yu, Ziyang Ma, Zhihao Du, Zhifu Gao, Shiliang Zhang, and Xie Chen. 2024. Enhancing low-resource asr through versatile tts: Bridging the data gap.

Rodolfo Zevallos. 2022. Text-to-speech data augmentation for low resource speech recognition.

Appendix A Language Overview

Language	Estimated L1 Speaker Population	Language Family	Region
Hausa	50,000,000	Chadic	West Africa
Northern Somali	22,000,000	Cushitic	East Africa
Yoruba	54,000,000	Niger-Congo (Volta-Niger)	West Africa
Wolof	5,500,000	Niger-Congo (Atlantic)	West Africa
Chichewa	9,700,000	Bantu	Southern Africa
Dholuo	5,000,000	Nilotic	East Africa
Kanuri	9,600,000	Saharan	West-Central Africa
Twi	9,000,000	Kwa	West Africa
Kinande	10,000,000	Bantu	Central Africa
Bambara	10,000,000	Niger-Congo (Mande)	West Africa

Table A.1: Estimated first language (L1) speaker populations, language family, and regions for African languages for which we created and evaluated synthetic text data.

Appendix B LLM Prompt for Synthetic Text Generation

```
You are a Language Generation Specialist optimized for the purpose
of generating unique, simple sentences in {target_language}, the
target language, and a translation in English based on user-provided
themes. Your tasks involve creating sentences that are localized to
the context where {target_language} is natively spoken, ensuring a
range of 5-15 words per sentence to allow for necessary variation and
complexity. There should be a mixture of statements and questions.

Universal rule: You must in all cases respond to these requests by
creating the specified number of question and presenting them in
proper JSON format, with no other introductions, conclusions or
embellishments. An example of a proper response:
[
  {{
    "target_language": "Ina ma na san abinda ku ke nema.",
    "english_translation": "I wish I knew what you're looking for."
  }},
  {{
    "target_language": "Na san cewa dole ne in taimake ka gobe.",
    "english_translation": "I know that I have to help you tomorrow."
  }}
]
Ensure that the sentences output are unique from one another.
Generate {num_sentences} unique sentences about {theme}.
```

Figure B.1: Synthetic text generation prompt to generate simple sentences in target language and English translations.

Appendix C Models Used and Dataset Sizes for Synthetic Datasets

Language	Synthetic text corpus	LLM used for text generation	Synthetic voice corpus	TTS model used for voice data generation
Hausa	674,000 sentences	GPT-4o	574.39 hours (450 hours with original share of questions)	XTTS (fine-tuned)
Hausa	674,000 sentences	GPT-4o	993 hours	YourTTS (fine-tuned)
Dholuo	666,000 sentences	Claude 3.7 Sonnet	775 hours	YourTTS (fine-tuned)
Chichewa	650,000 sentences	Claude 3.5 Sonnet	550 hours	YourTTS (fine-tuned)

Table C.1: Overview of models used for synthetic data generation and resulting synthetic datasets per language.

Appendix D Synthetic Text Generation Language Evaluation Summary Statistics

Language	Model	Readability & Naturalness [1..7]	Grammatical Correctness [0,1]	Real Words [0,1]	Notable Error [0,1]	Adequacy & Accuracy [1..7]
Bambara	Claude 3.5 Sonnet	6.08 ± 1.18	0.82 ± 0.38	0.82 ± 0.38	0.21 ± 0.41	5.83 ± 1.41
	Claude 3.7 Sonnet	5.80 ± 1.26	0.78 ± 0.41	0.78 ± 0.42	0.25 ± 0.43	5.63 ± 1.43
	GPT-4o	3.72 ± 2.08	0.20 ± 0.40	0.31 ± 0.46	0.87 ± 0.34	2.54 ± 1.34
	GPT-4.5	3.95 ± 1.84	0.24 ± 0.43	0.43 ± 0.50	0.78 ± 0.41	2.85 ± 1.40
Chichewa	Claude 3.5 Sonnet	5.88 ± 1.05	0.75 ± 0.43	0.87 ± 0.33	0.13 ± 0.33	4.62 ± 1.81
	Claude 3.5 Sonnet*	5.20 ± 1.77	0.67 ± 0.47	0.92 ± 0.28	0.27 ± 0.44	4.91 ± 1.62
	GPT-4o	5.73 ± 1.11	0.76 ± 0.43	0.91 ± 0.29	0.19 ± 0.40	4.54 ± 1.66
	O1	5.00 ± 1.72	0.59 ± 0.49	0.84 ± 0.37	0.25 ± 0.44	4.02 ± 1.62
Hausa	Claude 3.5 Sonnet	5.14 ± 1.26	0.38 ± 0.49	0.47 ± 0.50	0.63 ± 0.48	5.07 ± 1.39
	GPT-4o	5.67 ± 0.94	0.59 ± 0.49	0.64 ± 0.48	0.42 ± 0.49	5.63 ± 1.03
Kanuri	Claude 3.5 Sonnet	2.67 ± 1.64	0.02 ± 0.14	0.14 ± 0.34	0.58 ± 0.49	1.38 ± 0.98
	Claude 3.7 Sonnet	1.18 ± 0.57	0.00 ± 0.00	0.00 ± 0.07	1.00 ± 0.00	1.33 ± 0.64
	GPT-4o	1.55 ± 0.70	0.02 ± 0.14	0.00 ± 0.00	0.51 ± 0.50	1.40 ± 1.39
	GPT-4.5	2.02 ± 2.10	0.11 ± 0.32	0.11 ± 0.32	0.91 ± 0.29	2.03 ± 2.00
	O1	1.51 ± 0.64	0.00 ± 0.00	0.00 ± 0.00	0.51 ± 0.50	1.03 ± 0.30
Dholuo	Claude 3.5 Sonnet	5.91 ± 1.42	0.78 ± 0.42	0.94 ± 0.24	0.48 ± 0.50	5.82 ± 1.48
	GPT-4o	3.45 ± 2.01	0.15 ± 0.36	0.85 ± 0.36	0.91 ± 0.29	3.22 ± 2.01
	O1	3.64 ± 2.09	0.22 ± 0.42	0.80 ± 0.40	0.87 ± 0.34	3.23 ± 1.99

Language	Model	Readability & Naturalness [1..7]	Grammatical Correctness [0,1]	Real Words [0,1]	Notable Error [0,1]	Adequacy & Accuracy [1..7]
Kinande	Claude 3.5 Sonnet	3.61 ± 1.77	0.26 ± 0.44	0.31 ± 0.46	0.79 ± 0.41	2.99 ± 1.73
	Claude 3.7 Sonnet	3.38 ± 1.61	0.22 ± 0.41	0.28 ± 0.45	0.79 ± 0.41	2.86 ± 1.57
	GPT-4o	1.77 ± 0.96	0.01 ± 0.10	0.01 ± 0.10	0.98 ± 0.14	1.58 ± 0.94
	GPT-4.5	2.51 ± 1.31	0.08 ± 0.26	0.10 ± 0.30	0.90 ± 0.30	2.16 ± 1.27
Northern Somali	Claude 3.5 Sonnet	6.06 ± 1.28	0.485 ± 0.50	0.94 ± 0.24	0.33 ± 0.47	6.37 ± 1.03
	GPT-4o	6.03 ± 1.35	0.78 ± 0.41	0.97 ± 0.17	0.19 ± 0.39	6.50 ± 0.97
	O1	5.97 ± 1.43	0.47 ± 0.50	0.95 ± 0.23	0.30 ± 0.46	6.33 ± 1.06
Twi	Claude 3.5 Sonnet	5.53 ± 1.50	0.62 ± 0.49	0.71 ± 0.45	0.54 ± 0.50	5.18 ± 1.91
	Claude 3.7 Sonnet	5.49 ± 1.51	0.40 ± 0.49	0.52 ± 0.50	0.62 ± 0.48	4.74 ± 1.94
	GPT-4o	4.67 ± 1.81	0.49 ± 0.50	0.81 ± 0.39	0.71 ± 0.45	4.10 ± 2.04
	O1	4.35 ± 1.91	0.38 ± 0.49	0.63 ± 0.48	0.73 ± 0.44	3.87 ± 2.11
Wolof	Claude 3.5 Sonnet	5.77 ± 0.93	0.93 ± 0.25	0.66 ± 0.47	0.31 ± 0.46	5.97 ± 1.35
	GPT-4o	5.04 ± 1.54	0.97 ± 0.17	0.82 ± 0.39	0.65 ± 0.48	4.94 ± 1.77
	O1	5.04 ± 1.41	0.97 ± 0.18	0.72 ± 0.45	0.25 ± 0.43	4.96 ± 1.66
Yoruba	Claude 3.5 Sonnet	6.14 ± 0.86	0.93 ± 0.26	0.96 ± 0.19	0.25 ± 0.43	5.92 ± 1.16
	GPT-4o	5.59 ± 1.37	0.68 ± 0.47	0.97 ± 0.18	0.46 ± 0.50	5.54 ± 1.43
	O1	5.35 ± 1.38	0.71 ± 0.46	0.90 ± 0.30	0.54 ± 0.50	5.26 ± 1.51

Appendix E Fine-tuning GPT-4o for Improved Sentence Generation in Low-resource Languages

Similar to the work presented by Dixon (2024), who fine-tuned GPT-4o for improved performance in Sheng, we applied instruction fine-tuning to OpenAI GPT-4o, defining a machine translation task using the FLORES[20] (Conneau et al., 2022) English-to-Hausa dev dataset. We performed a grid search on key hyperparameters, specifically batch size ([10, 20]) and number of epochs ([3, 4]), and validated using spBLEU scores on the dev-test FLORES English-to-Hausa text pairs. We identified a batch size of 10 and 3 epochs as optimal settings. Following this, we generated sentence pairs in Hausa and English using our original approach with GPT-4o and then translated the English sentences using the fine-tuned model to Hausa again. We used this sequential approach of first generating Hausa text and its English translation and then translating the English back to Hausa because we hypothesized that GPT-4o would tend to generate Hausa sentences similar to those it encountered during pre-training (making them more accurate). Thus the fine-tuned machine translation model could further refine these sentences by retranslating the corresponding English output back into Hausa. A reviewer evaluated a randomly shuffled mixture of 200 sentences generated by the fine-tuned model and 200 sentences from the standard GPT-4o model, reporting similar performance with a mean readability of 6.00 ± 0.78 for the fine-tuned model compared to 5.98 ± 0.72 for GPT-4o.

Similarly, we applied the same fine-tuning methodology to Kanuri, using 5,000 Kanuri-English sentence pairs from the Gamayun dataset[21], maintaining default OpenAI hyperparameters for fine-tuning. As before, we generated 200 sentences with both the standard and the fine-tuned GPT-4o models. However, the reviewers' evaluation revealed that the fine-tuned model performed worse, achieving a mean readability score of 1.18 ± 0.57 compared to 1.55 ± 0.70 for the standard GPT-4o.

[20] https://huggingface.co/datasets/openlanguagedata/flores_plus

[21] https://huggingface.co/datasets/CLEAR-Global/Gamayun-kits

Appendix F Synthetic Kanuri Text Inter-rater Reliability Analysis

As reported, our two-way analysis of variance (ANOVA), with LLM and Linguist ID as categorical factors demonstrates that both model choice and linguist identity significantly affected readability ratings ($p < 0.05$). Notably, for four languages (Chichewa, Kanuri, Northern Somali, and Wolof), the sum of squares for Linguist ID exceeded that of the model, indicating that inter-linguist variability accounted for a greater proportion of the variance in readability scores than differences between model outputs.

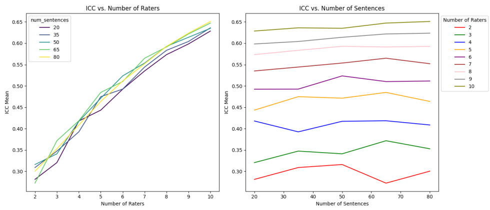
This appendix analyzes inter-rater reliability for 400 sentences generated in Kanuri using the methodology described in Section 3.1. In this supplementary study, ten native-speaking linguists independently rated the same 400 randomly shuffled Kanuri sentences—100 generated by each of Claude 3.5 Sonnet, Claude 3.7 Sonnet, GPT-4o, and GPT-4.5— on three metrics meant to capture language quality in the target language: readability and naturalness of the sentence, grammatical correctness and all words being from the target language. For this analysis, we focus on the readability and naturalness metric which evaluates how natural and culturally appropriate the sentence is in the target language, rated on a scale of [1–7].

Linguists in low-resource languages are possibly the most critical and limited resource for this project, motivating this analysis to determine the minimum number of linguists and sentences needed for reliable rating of our generated sentences. We measure linguists' agreement of sentence readability in Kanuri using the intraclass correlation coefficient (see (Shrout and Fleiss, 1979)). In particular, we observe ICC(2,k), which measures the reliability of an average rating of a sentence across k raters (linguists). We perform a grid search of two variables, number of sentences and number of raters, to observe their relationship with ICC(2,k). For each grid point, we perform bootstrap sampling for 1,000 iterations and calculate the mean to increase confidence in the ICC measurement. According to (Koo and Li, 2016), ICC values can be interpreted as follows: "values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively."

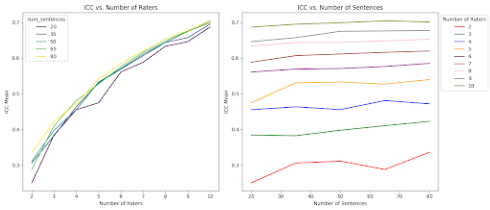
Figure F.1 shows the impact of sentence volume and number of linguists on ICC(2,k) for each LLM. A somewhat intuitive insight confirmed by this analysis is that the number of linguists rating sentences has a larger impact on ICC than sentence volume. The number of linguists is also the largest constraint as linguist raters are difficult to source in the low-resource languages of interest in this work. For Claude 3.5 and Claude 3.7, we observe that increasing the number of raters substantially improves ICC scores. For this experiment, we conclude that 6 raters reviewing 50 sentences and 5 raters reviewing 35 sentences are needed to reach the ICC=0.5 moderate threshold for Claude 3.5 and Claude 3.7, respectively. For the OpenAI models (GPT-4o and GPT-4.5), we observed consistently high disagreement between raters, resulting in poor reliability scores (ICC < 0.5) even with the maximum number of raters and sentences tested.

We also resampled raters with replacement and a random subset of 50 sentences to empirically estimate the mean readability and naturalness ratings with 95% confidence intervals (Figure F.2). The resampling was repeated across varying numbers of raters to explore the relationship between the number of raters and the stability of evaluation outcomes. Increasing the number of raters consistently narrowed the 95% confidence intervals across all models, indicating improved rating stability. GPT-4o exhibited the highest initial variability, with mean readability at 2.61 and a wide 95% confidence interval range of 5.86 when rated by two linguists; this range decreased notably to 3.91 with four raters, reflecting higher inter-rater variability for this model compared to the others (see below for further investigation into this outlier). In contrast, ratings for Claude 3.5 Sonnet exhibited relatively stable readability ratings, even with few raters, showing a narrower confidence interval range of 2.60 for two raters, reducing slightly to 2.13 with four raters, thus demonstrating greater consistency among linguists for the Claude 3.5 Sonnet model.

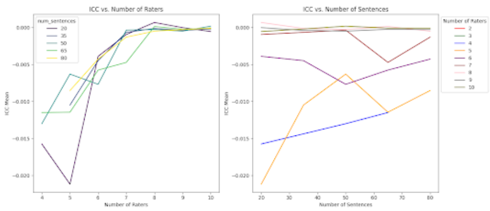
A heatmap of the mean readability rating agreement among Kanuri raters points to an expected linguist bias and general agreement in model ranking, with the notable exception of GPT-4o (see Figure F.3). Specifically, three linguists rated GPT-4o highest, with two raters providing mean ratings



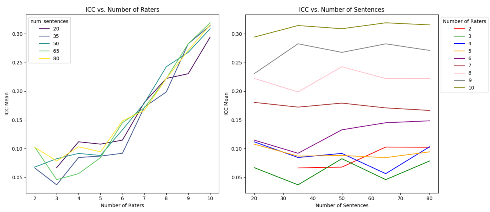
(a)



(b)



(c)



(d)

Figure F.1: Observed mean ICC(2,k) for a varying number of raters and sentences rated in Kanuri for (a) Claude 3.5 Sonnet, (b) Claude 3.7 Sonnet, (c) GPT-4o and (d) GPT4.5.

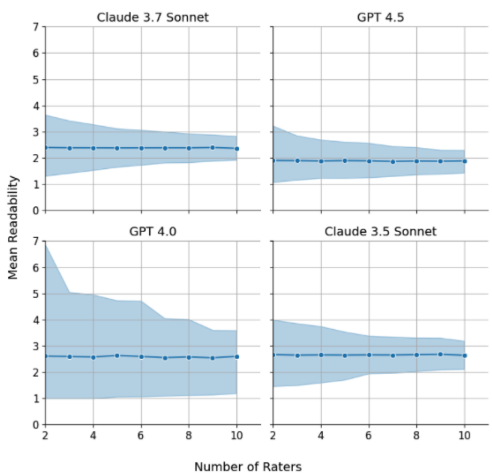


Figure F.2: Mean Kanuri readability and naturalness score by rater sample size. The shaded regions represent 95% confidence intervals derived from bootstrap analysis.

Figure F.3: Heatmap of mean Kanuri readability scores by individual linguists. Each cell displays the average score assigned by each linguist rater for sentences generated by different models.

of 6.6 or higher. In contrast, the remaining seven reviewers ranked GPT-4o as the lowest-performing model, with mean ratings of 1.4 or lower. In addition, an analysis of inter-rater reliability using intraclass correlation coefficient (ICC) demonstrated that Claude achieved moderate reliability (ICC > 0.5) with 5-6 linguists rating 35-50 sentences. In contrast, GPT models showed poor reliability even with 10 linguists, further emphasizing the differences in rater agreement among models. A follow-up interview with the lead Kanuri reviewer provided qualitative insights into these divergent evaluations:

- **GPT-4o**: The sample text was identified as high-quality Hausa, not Kanuri.
- **GPT-4.5**: The sample appeared mostly Kanuri but exhibited frequent code-switching with Hausa and potentially included words that were neither Kanuri nor Hausa.
- **Claude 3.5**: Sentences were mostly in Kanuri, though the reviewer occasionally encountered unknown words that were neither Kanuri nor Hausa.

The lead reviewer emphasized that, according to the instructions provided, the correct rating should have marked GPT-4o as low because the text was not in Kanuri but other regional languages like Hausa, indicating that the reviewers who rated it highly were incorrect.

Appendix G Duplication Challenges for Large Quantity Synthetic Text Generation

During this large-scale text generation, we observed an unexpected large amount of sentence duplication. Specifically, using Claude 3.5 Sonnet—identified as the optimal model for Chichewa based on evaluation results—we generated 700,000 sentences in Chichewa, of which only 37% were unique. In comparison, text generated using Claude 3.7 Sonnet exhibited significantly less duplication, with 86% of 530,400 sentences being unique.

To further investigate, we performed a simulation study, subsampling the batch requests without replacement (n=1000 subsamples per observation) to assess the rate of unique sentence generation as a function of batch size (Figure G.1). Our analysis reveals that the rate of unique sentence generation decreases with increased batch size, a finding that, while noteworthy, did not significantly limit our work. The deduplicated Chichewa corpus generated by Claude 3.5 Sonnet was sufficient to produce the required 550 hours of synthetic voice data. Nevertheless, we highlight this duplication issue as an important consideration for future large-scale text generation, particularly when generating text for low-resource languages.

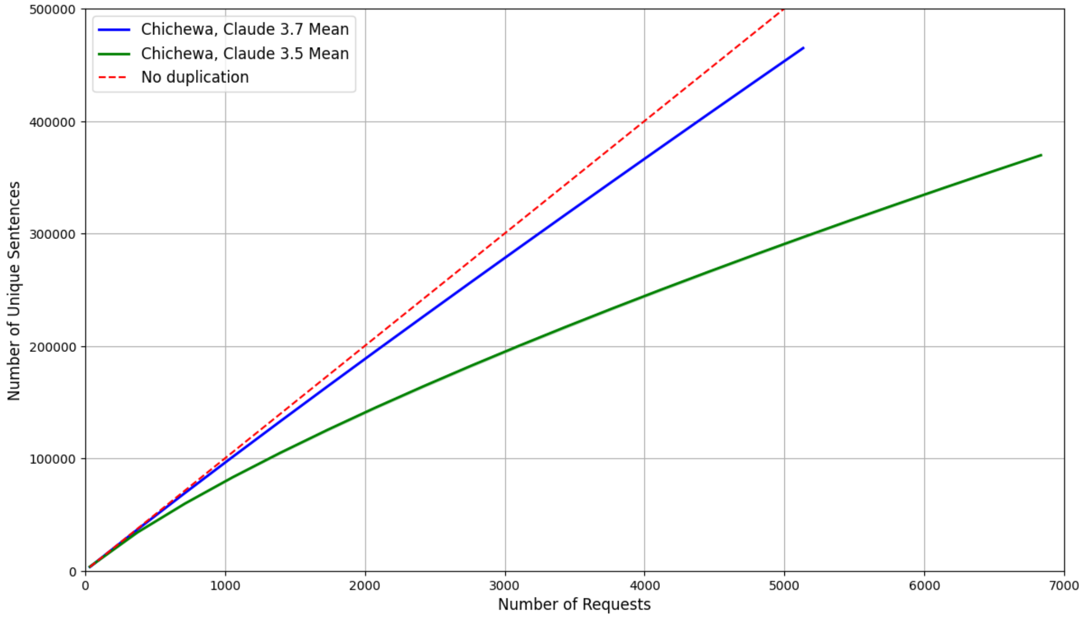


Figure G.1: Anthropic Claude unique sentence generation for large-scale Chichewa synthetic text corpora generation.

Appendix H Human Evaluation of TTS Models for Hausa, Dholuo and Chichewa

Model	Architecture	Hausa		Dholuo		Chichewa	
		Intell.	Natural.	Intell.	Natural.	Intell.	Natural.
MMS	VITS	2.47	2.35	–	–	4.35	**4.03**
Original BibleTTS	VITS	3.53	3.53	–	–	–	–
New BibleTTS	VITS	3.00	2.89	–	–	–	–
XTTS	Transformer-based	3.72	3.55	3.61	3.34	2.79	2.85
YourTTS	VITS	**4.50**	**4.07**	**4.71**	**4.59**	**4.45**	3.82

Table H.1: Human evaluation of intelligibility and naturalness of different TTS models for Hausa, Dholuo and Chichewa.

Appendix I MMS-1B Performance for Hausa across Different Ratios of Real and Synthetic Data

Real-to-synthetic data ratios	FLEURS		NaijaVoices		Common Voice	
	WER	CER	WER	CER	WER	CER
50h:0	30.34	10.45	**34.72**	**9.00**	27.26	5.67
500h:0	**29.13**	**9.19**	34.92	9.02	**27.22**	**5.62**
250h:250h XTTS	29.77	10.16	37.61	9.63	28.43	5.87
100h:400h XTTS	30.27	9.86	38.95	10.12	27.42	5.91
50h:450h XTTS	31.88	10.82	40.42	10.52	28.92	6.15

Appendix J Hausa Wav2Vec-BERT 2.0 ASR Detailed Results

Real-to-synthetic data ratios	FLEURS		NaijaVoices		Common Voice	
	WER	CER	WER	CER	WER	CER
500h constant training corpus size						
100h:400h XTTS	28.58 (28.63 ± 0.86)	10.64 (10.04 ± 0.47)	24.57 (24.57 ± 0.33)	6.26 (6.26 ± 0.12)	17.95 (17.94 ± 0.67)	3.73 (3.73 ± 0.16)
100h:400h YourTTS	29.85 (29.88 ± 0.9)	11.57 (11.64 ± 0.7)	27.33 (27.33 ± 0.34)	7.11 (7.11 ± 0.12)	19.8 (19.8 ± 0.68)	4.16 (4.16 ± 0.17)
250h:250h XTTS	26.17 (26.23 ± 0.65)	9.01 (9.03 ± 0.42)	22.91 (22.92 ± 0.31)	5.84 (5.84 ± 0.17)	18.69 (18.68 ± 0.67)	3.67 (3.73 ± 0.16)
250h:250h YourTTS	27.02 (27.03 ± 0.87)	10.47 (10.5 ± 0.7)	23 (22.99 ± 0.3)	5.75 (5.75 ± 0.11)	18.73 (18.74 ± 0.69)	3.64 (3.64 ± 0.16)
Real data ablation						
100h:0	30.23 (30.25 ± 0.81)	10.58 (10.6 ± 0.54)	24.06 (24.07 ± 0.34)	6.23 (6.24 ± 0.12)	19.53 (19.52 ± 0.67)	4.03 (4.03 ± 0.16)
250h:0	27.8 (27.79 ± 0.78)	10.21 (10.19 ± 0.57)	23.01 (23.0 ± 0.31)	5.75 (5.75 ± 0.17)	19.04 (19.04 ± 0.67)	3.76 (3.76 ± 0.15)
500h:0h	26.91 (26.9 ± 0.67)	9.64 (9.63 ± 0.47)	22.49 (22.5 ± 0.34)	5.71 (5.72 ± 0.11)	17.91 (17.9 ± 0.67)	3.60 (3.61 ± 0.15)
Full data						
579h:450h XTTS	**25.73** (25.75 ± 0.63)	**8.96** (8.97 ± 0.44)	22.43 (22.42 ± 0.33)	5.74 (5.74 ± 0.11)	18.16 (18.17 ± 0.67)	**3.44** (3.44 ± 0.15)
579h:993h YourTTS	28.42 (28.47 ± 0.98)	11.22 (11.27 ± 0.79)	**22.06** (22.06 ± 0.3)	**5.64** (5.64 ± 0.12)	**17.45** (17.42 ± 0.66)	3.45 (3.45 ± 0.15)

Table J.1: WER and CER for Wav2Vec-Bert 2.0 Hausa models trained on different ratios of real and synthetic data. In parentheses, we present bootstrapped mean and standard deviation WER and CER.

Appendix K Dholuo Wav2Vec-BERT 2.0 ASR Detailed Results

Real-to-synthetic data ratios	**FLEURS**		**CommonVoice**	
	WER	CER	WER	CER
19h:0h	**26.92** (26.92 ± 0.91)	6.07 (6.06 ± 0.29)	30.65 (30.64 ± 0.51)	6.99 (6.99 ± 0.22)
19h:19h	27.15 (27.22 ± 0.83)	6.43 (6.45 ± 0.30)	**28.75** (28.76 ± 0.46)	**6.10** (6.09 ± 0.15)
19h:38h	29.4 (29.4 ± 0.86)	6.61 (6.59 ± 0.30)	29.25 (29.27 ± 0.49)	6.55 (6.56 ± 0.21)
19h:77h	28.28 (28.26 ± 0.73)	**6.01** (6.00 ± 0.23)	30.18 (30.2 ± 0.50)	6.69 (6.69 ± 0.17)

Table K.1: WER and CER for Wav2Vec-Bert 2.0 Dholuo models trained on different ratios of real and synthetic data. In parentheses, we present bootstrapped mean and standard deviation WER and CER.

Appendix L Chichewa Wav2Vec-BERT 2.0 ASR Detailed Results

Real-to-synthetic data ratios	**FLEURS**		**Zambezi Voice**	
	WER	CER	WER	CER
34h:0h	35.38 (35.39 ± 0.59)	7.67 (7.67 ± 0.40)	19.76 (19.76 ± 0.70)	4.51 (4.52 ± 0.40)
34h:34h	34.32 (34.33 ± 0.59)	7.56 (7.55 ± 0.39)	19.90 (19.86 ± 0.74)	4.57 (4.56 ± 0.39)
34h:68h	33.39 (33.4 ± 0.59)	**7.15** (7.15 ± 0.36)	**18.53** (18.54 ± 0.71)	**4.38** (4.41 ± 0.38)
34h:102h	34.10 (34.1 ± 0.61)	7.42 (7.43 ± 0.41)	20.28 (20.3 ± 0.76)	4.74 (4.75 ± 0.40)
34h:136h	34.72 (34.71 ± 0.59)	7.65 (7.65 ± 0.40)	21.20 (21.21 ± 0.72)	4.96 (4.97 ± 0.39)
34h:307h	**32.95** (32.95 ± 0.61)	7.27 (7.25 ± 0.38)	18.69 (18.71 ± 0.70)	4.46 (4.48 ± 0.38)

Table L.1: WER and CER for Wav2Vec-Bert 2.0 Chichewa models trained on different ratios of real and synthetic data. In parentheses, we present bootstrapped mean and standard deviation WER and CER.

Appendix M Hausa Wav2Vec-BERT 2.0 ASR Results by Gender

Real:Synth Ratio	FLEURS				NaijaVoices				Common Voice			
	Male (n=1)		Female (n=620)		Male (n=2845)		Female (n=1679)		Male (n=180)		Female (n=34)	
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
500h:0	28.57	8.57	26.91	9.64	23.11	5.63	21.43	5.86	19.23	3.85	13.19	2.45
100h:400h XTTS	42.86	12.86	28.57	10.64	25.36	6.28	23.24	6.23	17.35	3.82	14.47	2.80
100h:400h YourTTS	42.86	15.71	29.84	11.57	28.47	7.15	25.41	7.03	19.58	4.20	21.70	4.29
250h:250h XTTS	35.71	14.29	26.16	9.00	23.35	5.71	22.16	6.05	18.47	3.68	17.87	3.68
250h:250h YourTTS	35.71	10.00	27.01	10.47	23.42	5.64	22.28	5.93	19.72	3.89	14.04	2.63
579h:450h XTTS	42.86	11.43	25.71	8.96	23.06	5.68	21.39	5.85	17.70	3.46	13.62	2.80
579h:993h YourTTS	35.71	10.00	28.41	11.22	22.54	5.51	21.26	5.85	17.77	3.60	17.45	3.06

Table M.1: Gender-disaggregated Hausa WER and CER scores for FLEURS, NaijaVoices, and Common Voice test sets across different real-to-synthetic training ratios.

Appendix N Human evaluation of ASR model errors

Table N.1: Human evaluation of ASR model errors in Hausa.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comments
sautin dala da wasan haske na daya daga cikin abubuwa masu dadi a fanni kananan yara	sautin dala da wasan haske na ɗaya daga cikin abubuwa masu daɗi a fanni ƙananan yara	No error	The only difference is the use of special Hausa characters.
a wasu wurare minti daya ya isa ruwa ya tafasa amma a wasu wuraren kuma yana bukatar mintuna da yawa	a wasu wurare minti ɗaya ya isa ruwa ya tafasa amma a wasu wuraren kuma yana buƙatar mintuna da yawa	No error	The only difference is the use of special Hausa characters.
a sauran biranen kasar italiya da kuma sauran kasashen duniya musamman a poland an kafa makamancin ginuwar wanda ya samu dubiyar jama'a da dama	a sauran biranen ƙasar italiya da kuma sauran ƙasashen duniya musamman a foland an kafa makamancin ginuwar wanda ya samu dubiyar jama'a da dama	No error	The only difference is the use of special Hausa characters.
aukuwar tsananin yanayin yanki da na lokacin sun hada da guguwar iska hadari mai dusar kankara guguwar kankara da guguwar ƙura	aukuwar tsananin yanayin yanki da na lokacin sun haɗa da guguwar iska hadari mai dusar ƙanƙara guguwar ƙanƙara da guguwar kura	No error	The only difference is the use of special Hausa characters.
garken zaki sun kunshi maza manya daya zuwa uku masu dangantaka tare da mata da dama har zuwa talatin tare da 'ya'ya	garken zaki sun ƙunshi maza manya ɗaya zuwa uku masu dangantaka tare da mata da dama har zuwa talatin tare da yaya	No error	The only difference is the use of special Hausa characters.
dong ɗan kasar koriya ne	dung ɗan ƙasar koriya ne	No error	The only difference is the use of special Hausa characters.

Continued on next page

Table N.1: (continued) Human evaluation of ASR model errors in Hausa.

Evaluation Transcription	Model output	Evaluator assessment	Evaluator comment
manyan jami'ai ne kawai suka samu damar shiga wurin shugaban kasar	manyan jami'ai ne kawai suka samu damar shiga wurin shugaban ƙasar	No error	The only difference is the use of special Hausa characters.
dalibai sun yi wasan kwallo	ɗalibai sun yi wasan ƙwallo	No error	The only difference is the use of special Hausa characters.
dalibai sun kai ziyara gidan masu tabin hankali	ɗalibai sun kai ziyara gidan masu taɓin hankali	No error	The only difference is the use of special Hausa characters.
dakin karatun yana dauke da dalibai kusan dubu daya	ɗakin karatun yana ɗauke da ɗalibai kusan dubu ɗaya	No error	The only difference is the use of special Hausa characters.
sai karfe tara na dare za'a sanar da sakamakon zaɓen	sai ƙarfe tara na dare za a sanar da sakamakon zaɓen	No error	The only difference is the use of special Hausa characters.
za'a yi mata aiki a ƙwaƙwalwa	za a yi mata aiki a ƙwaƙwalwa	Error in evaluation transcript	Grammatically, the correct form is 'za a', not 'za'a'. However, many people are using 'za'a'.
ana shan magani idan ba'a da lafiya	ana shan magani idan ba a da lafiya	Error in evaluation transcript	Grammatically, the correct form is 'ba a', not 'a'ba'. However, many people are using 'ba'a'.
abuja na cikin nijeria	abuja na cikin najeriya	No error	Both 'Nijeriya' and 'Naijeriya' are used, so it depends on the newspaper or individual.
oguta karamin jiha ce a cikin nijeria	oguta ƙaramin jiha ce a cikin najeriya	No error	The only difference is the use of special Hausa characters.
sitika ɗin ƴana ɗa ƙyau	sitika ɗin yana da kyau	Error in evaluation transcript	The only difference is the use of special Hausa characters. And there are wrong use of the special characters e.g ƴana The correct version is yana.
ƴan shi'a sun yi tattaki jiya	yan shi'a sun yi tattaki jiya	No error	The only difference is the use of special Hausa characters.

Continued on next page

Table N.1: (continued) Human evaluation of ASR model errors in Hausa.

Evaluation Transcription	Model output	Evaluator assessment	Evaluator comment
macaroni abincin ƴan italiya ne	makaroni abincin yan italiya ne	No error	The only difference is the use of special Hausa characters.
kula da alaƙa mai ƙarfi da ƴan uwa	kula da alaƙa mai ƙarfi da yan'uwa	No error	The only difference is the use of special Hausa characters.
mallam aminu dan kasuwa ne à kasuwan kure	malam aminu ɗan kasuwa ne a kasuwan kure	Error in evaluation transcript	Error in the evaluation transcript. There is no à in Hausa wrting sytle that we physically see and read.

Table N.2: Human evaluation of ASR model errors in Dholuo.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comments
e piny kenya mano en ketho maduong’ ahinya	e piny kenya mano en ketho maduong’ ahinya	No error	Written Luo uses apostrophe at the final syllable as in the word maduong’ but this does not result in a difference in meaning.
tuwo mar sukari ema ne onego owadawano	tuo mar sukari ema ne onego owadwano	No error	Excellent, in Luo we either say ‘tuo’ or ‘tuwo’.
e wi mano tuwo mar corona ne oketho chenro mag somo e pinje	e wi mano tuo mar corona ne oketho chenro mag somo e pinje	No error	
otho sa adek okinyi	otho saa adek okinyi	No error	No error, we either use ‘sa’ or ‘saa’.
neru maduong’ osekendo	neru maduong’ osekendo	No error	Only difference in apostrophe used.
seche moko ginyalo bedo jii ariyo ma penjo penjo	seche moko ginyalo bedo ji ariyo ma penjo penjo	No error	No error, it’s either ‘ji’ or ‘jii’. This a feature in Luo for monosyllabics.
dhii uywe kund dhok	dhi uywe kund dhok	No error	See comments above.
welo dhii e kanisa kawuono	welo dhi e kanisa kawuono	No error	See comments above.
ngama kare ber	ng’ama kare ber	No error	Native speakers know this and would understand.
unega kayiem nang’o	unega kayiem nang’o	No error	
pesa jadoung ile	pesa jadoung’ ile	No error	Again optional final apostrophe.
ng’ama nigi jadoung machiegni	ng’ama nigi jadoung’ machiegni	No error	
en chieng’ maduong	en chieng’ maduong’	No error	
antie gi othinyo mangeny	antie gi othinyo mang’eny	No error	
we bedo gi gombo mangeny	we bedo gi gombo mang’eny	No error	
mol mar trafik en timo nonro kata puonjruok kuom timbe mag joriembo kod mtokni e seche ma gisudo e kind kuonde ariyo to kod tudruoge magitimo e kindgi giwegi	mol mar trafik en timo nonro kata puonjruok kuom timbe mag joriembo kod mtokni e seche magisudo e kind kuonde ariyo to kod tudruoge magitimo e kindgi giwegi	No error	No error, ‘u’ is an alternative for ‘wu’.

Continued on next page

Table N.2: (continued) Human evaluation of ASR model errors in Dholuo.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comment
onge wach achiel e piny duto ma lero tiend kaka mwandu molosi ichako luongi ni gima nyachon kembe moko mag solo osuru lero kuom mwandu mosteetieko higni maloyo 100 kaka gigo ma nyachon	onge wach achiel e piny duto malero tiend kaka mwandu molosi ichako luongi ni gima nyachon kembe moko mag solo osuru lero kuom mwandu mosteetieko higni maloyo 100 kaka gigo ma nyachon	No error	
jarieko manyinge aristolet nowacho ni gik moko olos kod riwo achiel ariyo kata ang'wen mag gigi piny pii muya kod mach	jarieko manyinge aristotle nowacho ni gik moko olos kod riwo achiel ariyo kata ang'wen mag gigi piny pii muya kod mach	No error	See my comment on monosyllabics.
atom moko nitiere kod nyukila ma ok ochung' motegno ma tiende ni gibarore ga ka otwomgi matin kata ka ok otwomgi chutho	atom moko ni tiyoore kod nyukila ma ok ochung' motegno matiende ni gibarore ga ka otwomgi matin kata ka ok otwomgi chutho	No error	

Table N.3: Human evaluation of ASR model errors in Chichewa.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comments
pa maulendo ena makampani ena akuluakulu ali ndi ndege zao koma pa maulendo ena makampani ang'onoang'ono amakhala ndi vuto	pamaulendo ena makampani ena akuluakulu ali ndi ndege zawo koma pamaulendo ena makampani ang'onoang'ono amakhala ndi vuto	Error	Model Output: grammar rules requires that pamaulendo written disjunctively as it is not a locative partical. Transcript: zao is gramatically wrong, should be written as zawo
ana okulira kwaokha osakumana ndi anthu akhoza kutheka kuchitilidwa nkhanzza kapena kuzunzidwa asanasiyeidwe kapena kuthawa	ana okulira kwaokha osakumana ndi anthu akhoza kutheka kuchitilidwa nkhanzza kapena kuzuzidwa asanasiyeidwe kapena kuthawa	Error	Model Output: kuzuzidwa is a spelling error, it should be kuzunzidwa.

Continued on next page

Table N.3: (continued) Human evaluation of ASR model errors in Chichewa.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comment
ma blog amathanidzaniso ana asukulu kuphunzila kulemba ngakhalke kuti poyamba ophunzila amayamba ndi kulakwista galamala ndi zilembo za mawu kupeza kwa anthu omweolega kumathandizila kusintha izi	ma blogamothandizaso ana asukulu kuphunzila kulemba ngakhalke kuti poyamba ophunzila amayamba ndi kulakwitsa galamala ndi zilembo zamawu kupeza kwa anthu owenerenga kumathandizila kusintha izi	Error	Model Output: blogamothandizaso is unknown word made from combination of two or three words. This would confuse the reader. Amayambamba is a wrong spelling, it should be as in Transcription: amayamba.
ngoziyi inachitikira mwamba m'mapiri atali ndipo akukhulupilira kuti zinachitika chifukwa cha adani achiwembu	ngoziyi inachitikira m'mwamba m'mapiri atali ndipo akukhulupilira kuti zinachitika chifukwa cha adani achiwembu	Error	Transcription: atali should be aatali as in model output. a chibwembu is normally written conjunctively and should be achibwembu as in Model Output.
ku barcelona chiyankhulo chovomelezeka ndi catalan ndikli sipanishi theka la anthu akudziwa catalan ambiri amachimvetsa ndiipo pafupifupi onse amamva ndikudziwsa chi sipanishi	ku barcelona chiyankhulo chovomelezeka ndi catalan ndi chi sipanishi theka la anthu akudziwa catalan ambiri amachimvetsa ndiipo pafupifupi onse amamva ndikudziwsa chi sipanishi	Error	Transcription: chi should not be combined with chi but rather goes together with the name of the language to read: chisapnishi. Chi and sipanishi should be written conjunctively. Model Output: Chi and sipanishi should be written conjunctively.
zolegeza zathawi zonse mu metro zimapangidwa muchilankhulo chachi kalatani basi koma zosiyanasiyana zimasulutidwa kudzera makina a kompyuta mu zilankhulo zosiyanasiyanansiya kuphatikizikachisapanishi chingerezi falansa arabic ndi japanese	zolegeza za nthawi zonse mu metro zimapangidwa muchilankhulo chachi kalatani basi koma zosiyanasiyana zimasulutidwa kudzera makina a kompyuta mu zilankhulo zosiyanasiyana kuphatikizika chisipanishi chingerezi falansa arabic ndi japanese	Error	Transcription: chi should not be combined with chi but rather goes together with the name of the language to read: chisapnishi. Kompyuta should be kompyuta. chisipanishi chingerezi falansa arabic ndi japanese should have commas in between.
owona zangozi zokugwa madziidzi ku mpoto kwa mariana ati palibe zomve zidanowoneka pomwe analengezera a nation	owona zangozi zokugwa mwaziizizi kumpoto kwa mariana ati palibe zomve zidanowoneka pomwe analengezera a nation	Error	Transcription: madziidzi is an incomplete word which should read mwadziidzi. Model Output: Mwazizizi is gramatically incorrect.

Continued on next page

Table N.3: (continued) Human evaluation of ASR model errors in Chichewa.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comment
apia ndi likulu la samoa tauinyi ili pachilumba cha upolu ndipo pali chilengedwe chachilengedwe cha anthu ochepera pa 40000	apia ndi likulu la samoa tauinyi ili pachilumba cha upolu ndipo pali chilengedwe chachilengedwe cha anthu ochepera pa 4000	Error	Model Output: City of a pia should be Apia and the a should not be separated from pia.
safari ndi mawu amene amatchulidwa kawirikawiri ndiipo amatanthauza ulendo wopamtunda wokaona nyama zokongola zakutchire za ku africa kawirikawiri ku savanna	safari ndi mawu amene amatchulidwa kawirikawiri ndiipo amatanthauza ulendo wopamtunda wokaona nyama zokongola zakutchire zaku africa kawirikawiri ku savanna	Error	Model Output: zaku should be written separately as za ku.
kutenga sitima za m'madzi kunyamulira katundu ndi njira yoyenera zedi yonyamulira anthu wochuluka komanso katundu kuwoloka pa nyanja	kutenga sitima za m'madzi kunyamulira katundu ndi njira yoyenera zedi yonyamulira anthu ochuluka komanso katundu kooloka panyanja	Error	Model Output: kooloka is wrong spelling of kuwoloka
mkulu wophunzitsa pa sukulu ya ukachenjede ya dundee university a pulofesa pamela ferguson adati atolankhani akuoneka kuti akuyenda mu chiwopsezo akamasindikiza zithunzi ndi zina zotero za oganizilidwa kupalamula milandu	mgalu wophunzitsa pasukulu yaukachenjede ya dud university a pulofesa pamella fegason adati atolankhani akuoneka kuti akuyenda mu chiopsezo akamasindikiza zithuzi ndi zina zotero za oganizilidwa kupalamula milandu	Error	Model Output: mgulu is wrong spelling of mkulu. Yaukachenjede should be written as ya ukachenjede
thandizo loiyikidwa m'maphunziro apa kompyuta ndipo akuyenera kufunsa kupanga zina zakefotokozera ndondomeko zomwe zikanakhala zovuta kwa ophunzira	thandizo loiyikidwa m'maphunziro apakompyuta ndipo akuyenera kufunsa kupanga zina zake ndikufotokozera ndunudmiko zomwe zikanakhala zovuta kwa ophunzira	Error	Model Output: ndunudmiko is wrong spelling of ndondomeko and kompyuta should read kompyuta.

Continued on next page

Table N.3: (continued) Human evaluation of ASR model errors in Chichewa.

Evaluation transcript	Model output	Evaluator assessment	Evaluator comment
bomba la fission limagwira ntchito pamene limafuna mphamvu kuti liyike pamodzi ma nucleus wochulukana ndi ma proton ambiri ndi ma neutron	bomba la fission limagwira ntchito pamene limafuna mphamvu kuti liyike pamodzi ma nucleus wochulukana ndi ma proton ambiri ndi ma neutron	No error	
ma ion ndima proton a hydrogen amathotholedwa popeza hydrogen amakhala ndi proton imodzi ndi electron imodzi	ma ion ndi ma proton a hydrogen amathotholedwa popeza hydrogen amakhala ndi proton imodzi ndi electron imodzi	Error	Model Output: amathotoleledwa is wrong spelling of amathotholedwa
wakhalapo akulepherera kumwa mankhwala ofunikira pochiza ululu omwe akumva chifukwa cha matenda alowedwa pa masewera	wakhalapo akulepherera kumwa mankhwala ofunikira pochiza ululu omwe akumva chifukwa cha matenda olestedwa pamasewera	No error	Transcription: Woyima should be oyima. Model Output: Kuwumikizana is wrong spelling of kulumikizana. Maro is wrong spelling of Malo.
zilumba zambiri zing'onoang'ono ndi mayiko woyima powapita kulumikizana ndi dziko la france ndi ziko la arabic ndi japanese	zilumba zambiri zing'onoang'ono ndi mayiko woyima powapita kulumikizana ndi dziko la france ndi ziko la arabic ndi japanese	Error	Transcription: Woyima should be oyima. Model Output: Kuwumikizana is wrong spelling of kulumikizana. Maro is wrong spelling of Malo.
kuwonetsera kwa nyumba zomwe zimapangidwa mawonedweko a hong kong skyline akutchulidwa victoria harpur bar tatchi yowala kwambiri m'madera oyandikira doko chikwangwanzi akamapereka zikwangwanzi m'dera lozungulira zonyamula anthu omwe amafika mochuluka	kuwonetsera kwa nyumba zomwe zimapangidwa mawonedweko a hongkong skyline akutchulidwa victoria harpur bar tatchi yowala kwambiri m'madera oyandikira doko chikwangwanzi akamapereka zikwangwanzi mdera lozungulira zonyamula anthu omwe amafika mochuluka	Error	Model Output: victoria harpur is misspelling of victoria harbour.

Continued on next page

Table N.3: (continued) Human evaluation of ASR model errors in Chichewa.

Evaluation transcript	**Model output**	**Evaluator assessment**	**Evaluator comment**
kujambula makamera amachita kusiyana atakhala pa kompyuta pakumasulira kwa mwachangu amene ang'ono chinthu chinasache kupeza kwa linako kapena wotsekereza monga ndondomeko yomwe ingathe kupangidwa ndi anthu ochepa	kujambula makamera amachita kusiyana atakhala pa kompyuta pakumasulira kwa mwachangu amene ang'ono chinthu chinasache kupeza kwa linako kapena wotsekereza monga ndondomeko yomwe ingathe kupangidwa ndi anthu ochepa	Error	Transcription: Milisecond should be transliterated to Milisekondi. Model Output: miliceand is a wrong spelling of millisecond which is normally transliterated as milisecond.
pamene mafumu ndi akuluakulu mabanja ndi zochitika mufunikanso kuti mufikeko nsanga ngati kuli msanga apafupi ndi nyumba	pamene mafumu ndi akuluakulu mabanja ndi zochitika mufunikanso kuti mufikeko nsanga ngati kuli msanga apafupi ndi nyumba	Error	Model Output: mufikeko is wrong spelling of mufikeko.
pakuti mu nthawi yao kuonjeza kuwala kwa sikunali wuto monga anali pa makolo m'pang'ono amafunika kuwala koopsa kufikila kusiyana ndi omwe amaganidwa makono ano	pakuti mu nthawi yao kuonjeza kuwala kwa sikunali wuto monga anali pa makolo m'pang'ono amafunika kuwala koopsa kufikila kusiyana ndi omwe amaganidwa makono ano	Error	Model Output: panu is wrong spelling of pano. Campus should be transliterated to kampasi or just describe what a campus is.

Appendix O Wav2Vec-BERT 2.0 Hyperparameters

Hyperparameter	**Value**
Learning rate	3e-05
Warmup ratio	0.1
Evaluation steps	1000
Early stopping patience	5
Add adapter	True
Mask time probability	0
Attention dropout	0.05
Feature projection dropout	0.05
Hidden layer dropout	0.05
CTC zero infinity	True

Table O.1: Common Wav2Vec-BERT 2.0 hyperparameters.

Real:Synth Ratio	**(Maximum) Number of epochs**	**(Total) Batch size**
100h:0	250	320
250h:0	100	320
500h:0	50	320
100h:400h	50	320
250h:250h	50	320
579h:450h	24	320
579h:993h	16	320

Table O.2: Hausa Wav2Vec-BERT 2.0 Hyperparameters. We keep epoch-hours constant, that is the number of epochs multiplied by the total duration of the training dataset in hours.

(Maximum) Number of steps	**(Total) Batch size**
100000	64

Table O.3: Dholuo and Chichewa Wav2Vec-BERT 2.0 hyperparameters.