Learning from Scarcity: Building and Benchmarking Speech Technology for Sukuma

Macton Mgonzo[1], Kezia Oketch[2], Naome A. Etori[3], Winnie Mang'eni[4], Elizabeth Nyaki[4,5], Michael S. Mollel[4,5],

[1]Brown University, [2]University of Notre Dame, [3]University of Minnesota - Twin Cities, [4]Pawa AI, [5]Sartify Company Limited.

macton_mgonzo@brown.edu

Abstract

Automatic Speech Recognition (ASR) systems are gaining increasing attention in both academia and industry. Despite having remarkable performance in high-resource languages, their efficacy is less pronounced in low-resource settings. We present the first ASR system for Sukuma, one of the most severely under-resourced Tanzanian languages, and provide an open-source Sukuma speech corpus comprising 7.47 hours of carefully transcribed audio. The data, sourced primarily from Bible readings, was rigorously annotated to ensure phonetic and orthographic consistency, making it the most linguistically reliable resource currently available for the Sukuma language. To establish baselines, we train lightweight ASR and Text-to-Speech (TTS) models that demonstrate the feasibility of building end-to-end speech systems for this underrepresented language. This work addresses the challenges of developing language and communication tools for speakers of less-represented languages, particularly the scarcity of representative datasets and benchmarks, and highlights future research directions for linguistically challenging languages, such as Sukuma. We make our data and code publicly available to facilitate reproducibility and further research.[1]

1 Introduction

Recent breakthroughs in speech technology have brought human–machine communication closer to the natural flow of conversation. However, these advances overwhelmingly benefit a small set of well-resourced languages, such as English, Mandarin, and Spanish, driven by progress in large-scale modeling and self-supervised pre-training (Baevski et al., 2020b; Radford et al., 2023), effectively hindering the visibility of the vast majority of the world's linguistic diversity from AI. Fewer than 1% of the approximately 7,000 languages in the world have publicly available speech corpora (Joshi et al., 2020), and African languages, spoken by more than a billion people, represent only a tiny fraction of existing resources (Adelani et al., 2021; Etori et al., 2025; Bayes et al., 2024; Wang et al., 2024). These systems power everyday applications, such as voice assistants, tools for automating transcription services, and tools for people with special needs, including those with limited language understanding (Ogun et al., 2024; Afonja et al., 2024; Sanni et al., 2025). However, their benefits remain concentrated in only a handful of languages. Hundreds of widely spoken languages, particularly in Africa, still lack digital corpora and computational infrastructure (Adelani et al., 2021; Olatunji et al., 2023; Oketch et al., 2025; Etori and Gini, 2024). The persistent imbalance risks reinforcing digital inequality and limiting the cultural and technological participation of millions of speakers.

In Tanzania, although Swahili and English serve as official languages, most citizens speak one or more local languages as their primary means of communication. Sukuma (or Kisukuma as referred to by natives), a Bantu language spoken by an estimated 10 million people in northern Tanzania (Lestrade, 1948; Matondo, 2006), remains absent from publicly available speech datasets and models. This scarcity hinders the development of automated speech and language technologies that could enable inclusive communication and support language preservation.

We address this gap by introducing Sukuma Voices, the first open speech corpus for Sukuma, curated from publicly available Bible readings and validated through native-speaker verification. The dataset comprises 7.47 hours of transcribed audio. Using this corpus, we train and evaluate baseline ASR and TTS systems to establish reference performance for future research. Our contributions are

[1] Code: https://github.com/Sartify/sukuma-voices; Data: https://huggingface.co/datasets/sartifyllc/Sukuma-Voices-ACL;

threefold: (1) We release *Sukuma Voices*, the first Sukuma speech corpus, curated with the participation of native speakers. (2) We provide baseline ASR and TTS performance metrics to guide future research. (3) We discuss key challenges and opportunities for advancing low-resource speech technologies in the African context.

2 Related Work

In the general public and less-documented language spaces, several initiatives have focused on building multilingual speech corpora. Ardila et al. (2019) developed a multilingual speech corpus encompassing 29 languages through crowdsourcing as part of the *Mozilla Common Voice* initiative. This is by far the largest publicly available speech dataset. Despite its linguistic diversity, the dataset included only one African language (Kinyarwanda) as the only low-resource language from the African context.

Communities in these settings have also adopted similar strategies to develop datasets for languages at risk of endangerment or requiring domain-specific resources. Using community-driven approaches, Katumba et al. (2025) curated datasets in Luganda and Kiswahili, crowdsourced through Mozilla Common Voice (versions 12.0 and 15.0, respectively). The authors used native female speakers to validate the audio clips, supporting the development of text-to-speech systems in resource-constrained contexts. In a similar effort, Sikasote and Anastasopoulos (2021) developed BembaSpeech, a corpus comprising over 24 hours of Bemba, a language spoken by more than 30% of the Zambian population. The recorded phrases were drawn from the Bemba literature, illustrating a domain- and context-specific approach to building linguistic resources for African languages.

Kimanuka et al. (2024) introduced Lingala and Congolese corpus from the records of the Read Speech and Radio stations. The data consists of two corpora: a 4-hour labeled corpus and a 741-hour unlabeled corpus of four local languages. Similarly, (Wiafe et al., 2025) curated a multi-country corpus of five languages widely spoken in both Ghana and Togo, which share linguistic similarities. Their approach involved recording approximately 1,000 hours of spoken descriptions of culturally relevant images per language, with 10% (100 hours) transcribed, yielding parallel corpora to support ASR development. Finally, (Gutkin et al., 2020) released an open-source corpus for Yoruba, a language widely spoken in West Africa, representing a user base of more than 22 million people, thus expanding the coverage of speech technology for another primary African language. Ahia et al. (2024) further introduced Voices Unheard (YORÙLECT), a high-quality Yoruba corpus designed for ASR and speech-to-text translation.

Several efforts have explicitly targeted low-resource African languages to expand multilingual speech and language coverage. Godard et al. (2017) developed a speech corpus for Mboshi, a highly spoken Bantu language in the Republic of the Congo (Congo-Brazzaville), but poorly represented in the literature and research. The dataset comprises 5,000 speech utterances aligned with French text, collected through a realistic language documentation process. Biswas et al. (2022) built an ASR system with English as the primary language and evaluated it on four indigenous code-switched South African languages: Sesotho, Setswana, isiXhosa, and isiZulu. Similarly, (van der Westhuizen and Niesler, 2019) introduced a corpus of South African Bantu languages compiled from soap operas, which supports the development of conversational ASR in multiple languages. In a related study (van der Westhuizen et al., 2021), a multilingual ASR model was trained for two severely under-represented low-resource Malian languages: Bambara and Maasina Fulfulde. Sharma et al. (2025) fine-tuned OpenAI's Whisper-Tiny model on 5,520 Swahili audio samples, and they identified key challenges such as phonetic misinterpretations and reported a few performance gains. Finally, (Pratap et al., 2020) proposed a massively multilingual acoustic model capable of recognizing multiple languages within a single framework, demonstrating improved ASR performance for low-resource languages.

3 Datasets and Methods

3.1 Dataset Structure and Overview

Our dataset comprises a comprehensive collection of paired audio-text data specifically curated for Sukuma processing. The dataset contains 2,588 samples (7.47 hours), each consisting of an audio recording paired with its corresponding textual transcription. The data structure follows a standardized format with two primary components. Audio component: High-quality audio recordings stored in standard audio format, with durations ranging

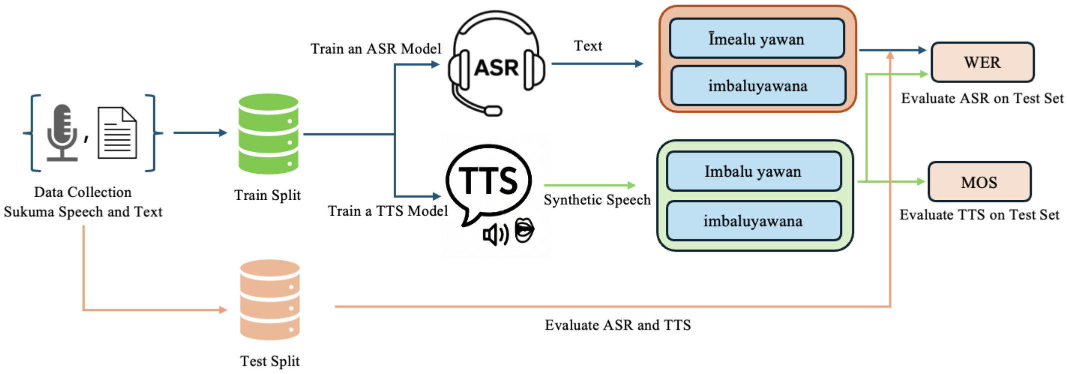

Figure 1: Overall experimental workflow for building and evaluating SUKUMA-ASR and SUKUMA-TTS. The data is split into a train set and a test set. The ASR model is trained on the speech–text pairs and evaluated on the test set using word error rate (WER). The TTS model is trained on the text–speech pairs, and its synthetic speech is evaluated on the test set using Mean Opinion Score (MOS)

Metric	Value
Total Samples	2,588
Total Duration	7.47 hours
Average Duration	10.39 ± 4.20 sec
Duration Range	1.74 - 30.36 sec
Total Words	51,107
Unique Vocabulary	10,007
Avg. Words/Sample	19.7
Speaking Rate	115.5 WPM

Table 1: Sukuma Voices Dataset Statistics

from 1.74 to 30.36 seconds per sample; Text component: Corresponding transcriptions in string format, capturing the spoken content with appropriate linguistic markers and punctuation.

3.2 Data Collection

The Sukuma Voices dataset was systematically constructed from audio recordings and textual transcriptions of the Sukuma New Testament 2000 translation, both sourced from the Bible.com platform [2]. This digital biblical resource provides authoritative Sukuma language content, serving as the foundation for our speech-text paired dataset. The selection of biblical text as source material offers several methodological advantages: (1) standardized orthographic conventions ensuring transcription consistency, (2) diverse linguistic structures encompassing narrative, dialogue, and theological discourse, (3) cultural relevance to Sukuma-speaking communities, and (4) availability of both audio recordings and verified textual transcriptions from the same authoritative source.

4 Experiments and results

We utilize the dataset presented in this paper to explore complementary speech processing tasks. First, we implement a text-to-speech (TTS) system that synthesizes natural-sounding Sukuma speech from Sukuma text. Second, we develop and evaluate an automatic speech recognition system to transcribe Sukuma speech into text. Table 2 presents examples of Sukuma phrases with their English translations for the reader's understanding.

4.1 System Architecture

Figure 1 illustrates the system architecture designed for this study, which employs a dual-stream pipeline to simultaneously develop and evaluate the ASR and TTS systems. First, the dataset is divided into training and test sets. Following this split, the ASR and TTS models are trained in parallel: the ASR model learns to map audio to text, while the TTS model is optimized to synthesize speech from textual input.

To measure performance, we evaluate the ASR system using Word Error Rate (WER) and the TTS system using the Mean Opinion Score (MOS). Beyond these standard independent metrics, we implement an integrated validation step in which the synthetic output of the TTS model is fed back into the ASR engine. This "cross-model" approach allows us to assess the ASR system's robustness on out-of-distribution data and verify if the synthetic

[2] https://www.bible.com/versions/1512-snt00-186-snt00-sukuma-new-testament-200

Language	Sentence
Sukuma	Umunhu ngwunuyo agabhalelaga chiza abhanhu bhakwe, kunguyo ya kikalile kakwe akagubhatogwa na gubhambilija abho bhali mumakoye.
English	This person raises his people well, because of his good behavior, of loving people and helping his colleagues who are in trouble, in their lives.
Sukuma	Uweyi agabhalangaga na bhanhu bhakwe inzila ja gwigulambigija uguitumama imilimo yabho, kugiki nabho bhabhapandikile jiliwa abhanhu bhabho.
English	He teaches his people to be diligent in work, so that they can provide for their daily needs.

Table 2: Parallel Sentences in Sukuma and English

audio maintains sufficient phonetic fidelity to be accurately recognized by the speech recognition model.

4.2 Text-to-Speech (TTS) Training and Synthesis

In this experiment, we train a TTS model that processes the Sukuma text into the Sukuma speech. To evaluate the generated audio, we conducted a subjective quality study with 2 native Sukuma speakers. The participants were asked to rate each synthesized utterance using a 5-point Likert Scale, following the Mean Opinion Score (MOS) protocol, where one indicates "bad" and 5 indicates "excellent". Our study achieved a MOS of 3.9 ± 0.15, close to human recordings (4.6 ± 0.1).

Experimental Setup. We used Orpheus 3B v0.1 (Labs, 2024), a neural codec language model designed for speech generation, as our base TTS system. The model uses a hierarchical neural audio codec (SNAC) to tokenize and reconstruct audio at a 24kHz sampling rate. We fine-tuned the model using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) instead of full fine-tuning to reduce computational costs while maintaining generation quality. The LoRA configuration used a rank of 64, alpha value of 64, and zero dropout, targeting all projection layers (q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj).

Before training, all audio data from the Sukuma dataset was preprocessed and encoded into discrete acoustic tokens using the SNAC codec. Audio samples were resampled to 24kHz to match the codec's expected input format. The SNAC encoder produces a three-layer hierarchical representation: the first layer captures broad acoustic patterns, while subsequent layers encode fine-grained temporal details. These representations were interleaved into a sequence of seven tokens per frame, with consecutive duplicate frames removed to improve efficiency. Each training example was formatted as a conversational turn, with special tokens delineating human input (text prompts with speaker identifiers) and AI responses (speech tokens), enabling the model to learn the text-to-speech mapping.

Training was carried out for four epochs using a batch size of 1 with accumulation of gradients in four steps (effective batch size of 4). We used the AdamW 8-bit optimizer with a learning rate of 2e-4, cosine learning rate scheduling, and a warmup ratio of 0.1. We set the weight decay to 0.01 for regularization. We trained the model using mixed precision (bfloat16 on Ampere GPUs, float16 otherwise) to optimize memory usage and training speed. The maximum sequence length was set to 16,384 tokens to accommodate longer utterances.

For inference, we generated synthetic speech for all test set examples by conditioning the model on the corresponding text transcriptions prefixed with a speaker identifier (`<sukuma>`). Generation used nucleus sampling with temperature 0.6, top-p 0.95, and repetition penalty 1.1 to balance quality and diversity. The maximum generation length was set to 4,096 new tokens. The generated acoustic tokens were decoded back to audio waveforms using the SNAC decoder, producing 24kHz audio files, which were then used to evaluate the performance of the ASR model on TTS-generated speech versus natural recordings.

4.3 Automatic Speech Recognition (ASR)

In this experiment, we trained a lightweight speech recognition model on the Sukuma Voices dataset to evaluate the model's performance in two key metrics: the quality and utility of both the original and synthetic speech data. We divide the dataset into an 80% train set and a 20% test set. We report the word error rate (WER) as our primary evaluation metric, with lower values indicating better performance, as shown in Figure 2 and Figure 3.

Experimental Setup. We used Whisper Large V3 (Radford et al., 2023) as our base ASR model, using the Unsloth framework for efficient training. The model was configured for full fine-tuning rather than parameter-efficient methods, allowing complete adaptation to the Sukuma language. We set the language parameter to Sukuma (sk) and configured the model for transcription tasks. To ensure optimal performance for this low-resource language, we disabled token suppression and allowed the model to freely generate transcriptions without forced decoder IDs.

In the training process, we used a batch size of 2 with gradient accumulation in four steps, resulting in an effective batch size of 8. We trained for four epochs using the AdamW 8-bit optimizer with a learning rate of 1e-4, linear learning rate scheduling, and a warmup period of 5 steps. Weight decay was set to 0.01 for regularization. All audio inputs were resampled to 16kHz to match Whisper's expected input format. We trained the model with 32-bit floating point precision to maintain numerical stability. We evaluated the model's performance every 20 training steps using two test sets: one containing original recordings and another containing synthetic speech generated from text. Performance was tracked using Weights & Biases, with WER computed using the standard evaluation metric from the HuggingFace evaluate library.

5 Results and Analysis

Table 3 presents comprehensive statistics that compare the performance of the models trained on the original versus synthetic speech data over 81 training steps.

Architectural Comparison and Model Selection. We evaluated both Whisper and Wav2Vec2-large-XLSR-53 architectures to identify the optimal approach for low-resource Sukuma ASR. While Wav2Vec2 demonstrated the characteristic CTC training dynamics (Baevski et al., 2020a), it consistently exhibited prolonged convergence periods and failed to outperform the baseline, ultimately achieving inferior WER despite extensive hyperparameter tuning. In contrast, Whisper's sequence-to-sequence framework with its pretrained multilingual representation proved significantly more effective for rapid adaptation to Sukuma's tonal and diacritic-rich phonology. Consequently, Whisper demonstrated superior performance, achieving substantial learning progress and WER reductions of 82.94% and 78.93% on original and synthetic data, respectively. The final model evaluated on original human-recorded data achieves 25.19 WER, while synthetic data reaches 32.60 WER- an absolute difference of 7.41 points (29.4% relative). Given these empirical results and the resource constraints of our low-data regime, we focus our analysis exclusively on the Whisper architecture, which provided both superior performance and more stable training dynamics for Sukuma speech recognition.

Metric	Original	Synthetic
Min WER	22.01	29.97
Max WER	170.65	179.45
Mean WER	41.53	51.15
Median WER	30.93	41.52
Std Dev	28.27	28.35
CV (%)	68.07	55.43
Final WER (Step 81)	25.19	32.60
WER Reduction	82.94%	78.93%

Table 3: ASR performance comparison between original and synthetic speech data across 81 training steps.

Convergence and Stability. Analysis of the final five training steps (77-81) reveals that both evaluations have reached stable convergence, with the original model showing a mean WER of 25.27 and a standard deviation (σ=0.44) and the synthetic model at 33.03 (σ=0.51). The low standard deviations indicate minimal performance fluctuation in the late training stages.

Consistency Across Training. Interestingly, synthetic-trained data exhibit greater evaluation stability, as evidenced by a lower coefficient of variation (55.43% vs 68.07%). This suggests that although the synthetic data achieves a higher absolute WER, it exhibits more predictable learning dynamics throughout training.

Performance Gap Analysis. Synthetic speech consistently produces higher WER across all training stages, with a mean absolute difference of 9.62 WER points (RMSE: 9.84). The relative performance gap averages 28.34%, ranging from a minimum of 3.70% (Step 55) to a maximum of 40.75% (Step 36). Statistical testing confirms that this dif-

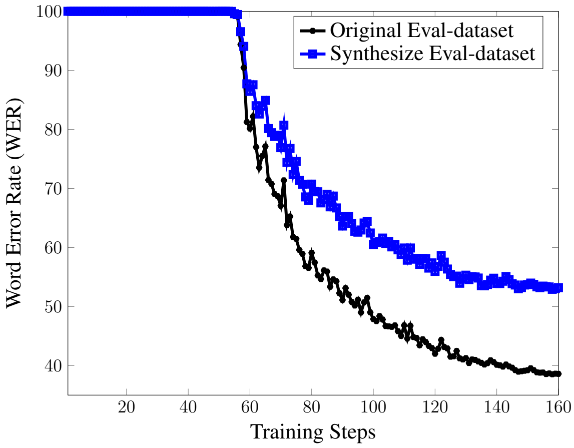

Figure 2: Word Error Rate (WER) during wav2vec2-large-xlsr-53 fine-tuning on the Sukuma corpus. The model was evaluated on two held-out test sets at regular intervals: the original human-recorded evaluation set (black line) and the TTS-synthesized in-distribution evaluation set (blue line). Both curves exhibit the characteristic CTC training pattern: an initial 100% WER plateau followed by rapid improvement. The synthetic data consistently achieves lower WER throughout training (≈5–15% absolute reduction), suggesting better acoustic alignment with the training distribution despite domain mismatch.

ference is significant (paired t-test: $t = -41.41$, $p < 0.001$; Wilcoxon test: $p < 0.001$).

Stage-wise Performance. We partition the training process into three stages to examine the learning dynamics. In the early stage (Steps 1-27), the performance gap is 9.97 WER points. This gap increases slightly to 10.77 in the midstage (steps 28-54) before narrowing to 8.11 in the late stage (steps 55-81), suggesting that the relative disadvantage of synthetic data diminishes as training progresses.

Correlation Analysis. Despite the consistent performance gap, both training curves exhibit a remarkably strong correlation (Pearson's $r = 0.997$, $p < 0.001$; Spearman's $\rho = 0.993$, $p < 0.001$), indicating that the synthetic and original datasets produce highly similar learning trajectories. This strong correlation suggests that the synthetic speech effectively captures the essential acoustic–phonetic characteristics of the original data, albeit with some degradation in absolute performance.

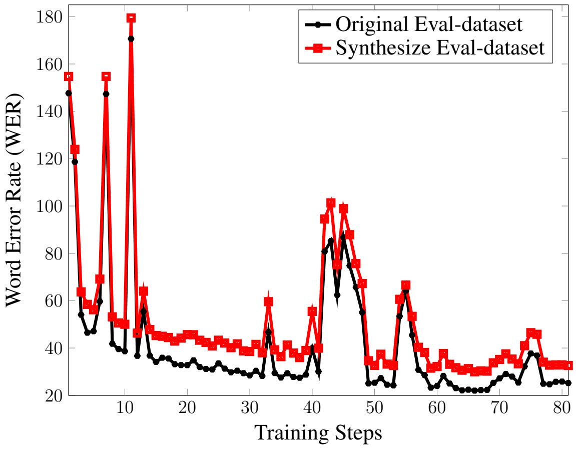

Figure 3: Word Error Rate (WER) across training steps for the SUKUMA-ASR model evaluated on two datasets: the original human-recorded evaluation set (black line) and the TTS-synthesized evaluation set (red line). The comparison highlights performance differences between original human-recorded and synthetic test data during model training.

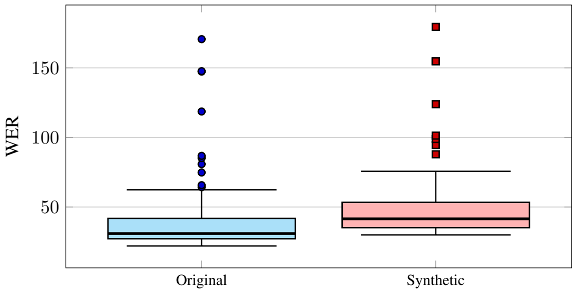

Figure 4: Distribution of Word Error Rate (WER) for the SUKUMA-ASR model evaluated on two test sets: the Original human-recorded evaluation set (blue) and the Synthetic TTS-generated evaluation set (red). The boxplots summarize the median, interquartile range, whiskers, and outliers, highlighting that the synthetic test data generally yields higher WER and greater variability compared to the evaluation set of the original recordings.

6 Conclusion

In this paper, we introduced the Sukuma Speech Dataset, developed to address the scarcity of low-resource languages in natural language processing (NLP) research. We demonstrated the dataset's utility through two key tasks: automatic speech recognition (ASR) and text-to-speech (TTS). Our baseline experiments achieved competitive word error rates (WER) for ASR and promising mean opinion scores (MOS) for TTS, establishing reference benchmarks for future research.

Beyond these baseline results, the dataset opens several avenues for exploration. First, it enables the study of cross-lingual speech processing between Swahili, a widely spoken lingua franca, and Sukuma, the widely spoken ethnic language in Tanzania. Second, the dataset offers practical implications for developing inclusive technologies, including educational tools, accessibility applications, and digital resources for cultural preservation.

We hope that this resource will facilitate the development and evaluation of speech technologies that have traditionally been limited to high-resource languages. In addition, the dataset is expected to catalyze the continued research on low-resource speech technologies in Africa and beyond. Future work will focus on expanding the size and diversity of the dataset, incorporating additional languages and domains, and exploring broader downstream applications. By making this resource publicly available, we aim to support the broader NLP and speech communities in advancing inclusive and equitable language technology.

7 Limitations and Future Work

Several challenges persist, particularly for low-resource languages with diacritics, such as Sukuma. We present key challenges learned during the study and provide directions for future research.

7.1 Data Collection Challenges

Obtaining native speakers was a significant challenge, given that Sukuma is a poorly documented yet widely spoken language. The language has two written forms, one that employs diacritics and another that does not, but it lacks standardized usage across written materials. As a result, creating a high-quality transcribed dataset is expensive and time-consuming. The dataset presented in this paper was primarily collected from biblical texts, which may not accurately represent the nuances of language in everyday usage. We collected nearly 7.47 hours of corpus, which may not generalize well to everyday language use.

7.2 Evaluation Metrics Challenges

Another limitation was the significant differences in the representation of the diacritic and non-diacritic versions of the text in the corpus. The existence of both in the corpus presents a considerable challenge, as it can increase vocabulary and character error rates at the token level when directly evaluating our models using standard metrics, such as WER. Consequently, our study only focused on the non-diacritic version of the corpus.

7.3 Quality of Generated Speech

Our study is also limited by the lack of diversity in the training dataset; using only 7.47 hours of corpus from a single source (Bible) limits the generalizability of our results. Through experiments, we achieved a 28.34% increase in WER in the synthetic speech generated by the TTS compared to the original recordings, suggesting that the transcription system did not accurately capture the acoustic fidelity of the original Sukuma speech.

7.4 Future Directions

Our study exposes several limitations and provides a way for future exploration including: (1) curation of a more diverse Sukuma speech corpus to at least 100 hours beyond religious data; (2) explore alternative hybrid evaluation metrics that account for diacritic variations in the language; (3) extend the human evaluation part through community led-initiatives; and (4) explore computationally efficent adaptation methods suitable for low-resource development and development of speech techlogies.

8 Ethical Considerations

In cases where human subjects were required to complete the annotation of the data, consent was obtained from the participants. No direct use of humans as research subjects was conducted. The participants were informed about the limitations of the technology they were evaluating and the possible impacts that could result from the misuse of the tools were clearly communicated.

Additionally, we acknowledge that any model trained on large-scale speech data, such as the Bible corpus, may perpetuate social or other kinds of bias, which we are aware of as a limitation. Although our work does not directly address bias issues, the primary objective of this research is to serve academic purposes and develop speech technologies that benefit humanity.

Acknowledgments

The authors express their gratitude to Sartify Company Limited [3] and Pawa-AI [4] for providing the data access necessary to develop our models. We also extend our sincere thanks to all the volunteers who generously dedicated their time to the evaluation process, as their contributions were vital to the completion of this work.

[3] https://www.sartify.com/

[4] https://www.pawa-ai.com/

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, and 1 others. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A Etori, Abraham Owodunni, and Moshood Yekini. 2024. Performant asr models for medical entities in accented speech. *arXiv preprint arXiv:2406.12387*.

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A Smith, and Yulia Tsvetkov. 2024. Voices unheard: Nlp resources and models for yor\ub\'a regional dialects. *arXiv preprint arXiv:2406.19564*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020a. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Edward Bayes, Israel Abebe Azime, Jesujoba O Alabi, Jonas Kgomo, Tyna Eloundou, Elizabeth Proehl, Kai Chen, Imaan Khadir, Naome A Etori, Shamsuddeen Hassan Muhammad, and 1 others. 2024. Uhura: A benchmark for evaluating scientific question answering and truthfulness in low-resource african languages. *arXiv preprint arXiv:2412.00948*.

Astik Biswas, Emre Yılmaz, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2022. Code-switched automatic speech recognition in five south african languages. *Computer Speech & Language*, 71:101262.

Naome Etori and Maria Gini. 2024. Rideke: Leveraging low-resource twitter user-generated content for sentiment and emotion detection on code-switched rhs dataset. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 234–249.

Naome A Etori, Arturs Kanepajs, Kevin Lu, and Randu Karisa. 2025. Lag-mmlu: Benchmarking frontier llm understanding in latvian and giriama. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 109–120.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélene Maynard, Markus Müller, and 1 others. 2017. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara Rivera, and Kólá Túbòsún. 2020. Developing an open-source corpus of yoruba speech. In *Interspeech*, pages 404–408.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Andrew Katumba, Sulaiman Kagumire, Joyce Nakatumba-Nabende, John Quinn, and Sudi Murindanyi. 2025. A curated crowdsourced dataset of luganda and swahili speech for text-to-speech synthesis. *Data in Brief*, page 111915.

Ussen Kimanuka, Ciira wa Maina, and Osman Büyük. 2024. Speech recognition datasets for low-resource congolese languages. *Data in Brief*, 52:109796.

Canopy Labs. 2024. Orpheus: A neural codec language model for high-quality text-to-speech.

GP Lestrade. 1948. The classification of the bantu languages. *African Studies*, 7(4):175–184.

Masangu Matondo. 2006. Tonal transfer in kisukuma. In *Selected Proceedings of the 35th Annual Conference on African Linguistics*, pages 125–135.

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. *arXiv preprint arXiv:2406.11727*.

Kezia Oketch, John P Lalor, and Ahmed Abbasi. 2025. Benchmarking sociolinguistic diversity in swahili nlp: A taxonomy-guided approach. *arXiv preprint arXiv:2508.14051*.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure FP Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and 1 others. 2023. Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively multilingual asr: 50 languages, 1 model, 1 billion parameters. *arXiv preprint arXiv:2007.03001*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Mardhiyah Sanni, Tassallah Abdullahi, Devendra Deepak Kayande, Emmanuel Ayodele, Naome A Etori, Michael Samwel Mollel, Moshood O Yekini, Chibuzor Okocha, Lukman Enegi Ismaila, Folafunmi Omofoye, and 1 others. 2025. Afrispeech-dialog: a benchmark dataset for spontaneous english conversations in healthcare and beyond. 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 8399–8417.

Avinash Kumar Sharma, Manas Pandya, and Arpit Shukla. 2025. Fine-tuning whisper tiny for swahili asr: Challenges and recommendations for low-resource speech recognition. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 74–81.

Claytone Sikasote and Antonios Anastasopoulos. 2021. Bembaspeech: A speech recognition corpus for the bemba language. *arXiv preprint arXiv:2102.04889*.

Ewald van der Westhuizen and Thomas R Niesler. 2019. Synthesised bigrams using word embeddings for code-switched asr of four south african language pairs. *Computer Speech & Language*, 54:151–175.

Ewald van der Westhuizen, Trideba Padhi, and Thomas Niesler. 2021. Multilingual training set selection for asr in under-resourced malian languages. In *International Conference on Speech and Computer*, pages 749–760. Springer.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, and 1 others. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023.

Isaac Wiafe, Jamal-Deen Abdulai, Akon Obu Ekpezu, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, and Kafui Kwashia Solaga. 2025. Advancing automatic speech recognition for low-resource ghanaian languages: Audio datasets for akan, ewe, dagbani, dagaare, and ikposo. *Data in Brief*, page 111880.