Building an Ewe Language Dataset: Towards Enhancing Automatic Speech Recognition Technologies for Low-Resource Languages

Isaac Wiafe[1] Akon Obu Ekpezu[2*] Raynard Dodzi Helegah[1]
Fiifi Baffoe Payin Winful[1] Elikem Doe Atsakpo[1] Charles Nutrokpor[1]
Kafui Kwashie Solaga[1]
[1]Department of Computer Science, University of Ghana
[2]Department of Information Processing Science, University of Oulu, Finland
akon.ekpezu@oulu.fi

Abstract

Automatic Speech Recognition (ASR) systems rely on large-scale, high-quality training datasets. However, low-resource languages, such as Ewe, remain underrepresented in the development of these systems. This study presents the development of a large-scale open-source speech dataset for Ewe, a Niger-Congo language spoken across Ghana, Togo, and Benin. Using supervised crowdsourcing, participants recorded descriptions of preselected culturally relevant images using a customized Android app. We collected 203,336 validated speech samples (1130 hours) from 1937 speakers, along with 107 hours of transcribed audio. To demonstrate the utility of the dataset for ASR, we fine-tuned Whisper base models, which were originally trained on Shona and Yoruba. The evaluation results suggest that both base models adapted well to Ewe and achieved a word error rate of 37%, an orthographic error rate of 45%, and a character error rate of 12%. A qualitative error analysis identified challenges including orthographic inconsistencies, morphological complexity, phonetic confusion, and dialectical variations. Thus, highlighting the need for dialect-sensitive and morphologically aware ASR modeling. The open-source release of this dataset provides a critical resource for advancing ASR research and linguistic preservation efforts for underrepresented African languages. Future work will explore self-supervised learning techniques to further improve performance using the unlabeled Ewe speech corpus.

1 Introduction

Africa's linguistic diversity poses major challenges for automatic speech recognition (ASR) because of the limited availability of high-quality open-source speech and text data for low-resource languages (LRLs). This is further exacerbated by the fact that over 1500 languages are endangered and may be lost by the end of the century (Bromham et al., 2022). Thus, prioritizing endangered languages via ASR development is crucial to preserving linguistic heritage, ensuring inclusivity (Chizzoni and Vietti, 2024), and preventing the loss of valuable cultural and historical knowledge (Jimerson et al., 2018).

Although natural language processing (NLP) has made significant progress particularly, ASR modeling in high-resource languages such as English, Mandarin, and Spanish, only a small fraction of the world's languages are supported by these technologies (Peterson et al., 2021). This is evident in the limited availability of annotated datasets, computational tools, and research funding for LRLs. Interestingly, this is not different in Ghana, which is a multilingual country with over 80 languages, yet all are LRLs. Existing multilingual speech datasets such as the Common Voice project (Ardila et al., 2020), African Speech Dataset (Olatunji et al., 2023), and GlobalPhone (Schultz, 2002) do not include the Ewe language.

Ewe is a Niger-Congo language spoken by approximately eight million people in Ghana and neighboring countries, Togo and Benin and neighboring countries. Yet it lacks the diverse corpora needed to develop ASR models and speech technologies. Although some studies (Antwi-Boasiako and Agyekum, 2022; Dei, 2024) have attempted to document and preserve some Ghanaian languages including Ewe, these initiatives are constrained by cost, lack of expertise, and technological support. These languages are also limited by standardized spelling conventions, dialectal variations, unspecified orthographies, and potential code-switching. For instance, Ewe has different dialectal variants across different regions, primarily in terms of orthography and pronunciation. These variations stem from the fact that Ewe is spoken across multiple countries; Ghana (including different regions), Togo, and Benin, and each of these countries or regions have its own sociolinguistic influences. While its core grammatical structure remains consistent, dialectal differences manifest in phonetics, vocabulary, and spelling conventions (Sam and Agbloe, 2024). Additionally, digitization efforts are limited by the lack of a standardized Ewe digital keyboard that can be installed on computers. This complicates digitization and makes data collection and transcription more challenging. Other previous attempts to digitalize Ewe are limited to context such as religious texts (Resnik et al., 1999), which may introduce domain-specific biases during ASR model training.

2 Related Work

Current methods of speech data collection including sentence reading and uncontrolled crowdsourcing are expensive, time-consuming, and logistically complex. This makes large-scale dataset development in low-resource environments (LREs) challenging. Although existing speech data collection approaches have been demonstrated to be effective in some jurisdictions (Ragano et al., 2020; Panayotov et al., 2015) they may not be appropriate for collecting Ewe. For instance, the sentence reading approach utilized by studies (Ibrahim et al., 2022; Georgescu et al., 2020; Gutkin et al., 2020) may be ineffective for languages with limited standardized orthographies. Moreover, many indigenous speakers of Ewe may lack the functional literacy required to accurately read sentences written in Ewe. Given the linguistic complexity of Ewe, crafting sentence prompts that capture the full range of natural speech and spontaneous utterances would require considerable effort.

While (Callison-Burch and Dredze, 2010) utilized Amazon's Mechanical Turk for uncontrolled crowdsourcing speech data collection, this method may not be feasible in regions with limited digital literacy and internet access. Also, uncontrolled crowdsourcing (Ardila et al., 2020) often results in inconsistent recording conditions, varying audio quality, and a lack of standardized quality checks, which affects the reliability and usability of the resulting dataset. This necessitates the design of a more structured and contextually appropriate approach for collecting Ewe speech data.

Accordingly, this study seeks to use a scalable and cost-effective approach to collect, curate, and evaluate a large speech dataset for Ewe. Specifically, it aims to collect at least 1000 hours of spontaneous speech data and 100 hours of transcribed text in Ewe language. The dataset will be evaluated by training an ASR model in Ewe. This study is expected to make several key contributions to theory and practice. Firstly, it will address the critical data scarcity challenge by providing an open-source, large-scale, high-quality speech dataset for Ewe. This is expected to significantly expand available linguistic resources for ASR development. Also, by employing a scalable and cost-effective data collection approach, this study offers a replicable framework that can be adapted for other LRLs. This study will contribute to the development of ASR technology by leveraging existing ASR models such as Whisper Small to finetune and evaluate an ASR model for Ewe. Ultimately, this study seeks to advance linguistic preservation efforts, enhance digital inclusivity, and set a foundation for future advancements in speech technology for Niger-Congo languages.

3 Methods and Materials

3.1 Ewe Speech Data Collection Pipeline

Ewe is linguistically complex. Hence, although it has a simple grammatical structure that makes it easy to decompose polysyllabic words into monosyllabic roots, it is characterized by unique phonological, morphological, and syntactic features. It is a tonal language with three main tones (high, mid, and low), that are used to distinguish the meaning of words. Hence, a phonetic structure may have different meanings depending on the tone used (e.g., ”to” means ”mountain” in one tone and ”ear” in another). This makes it a challenge for speech recognition when compared to non-tonal languages such as English. Ewe also has a complex morphophonemic process (vowel harmony and nasalization) which affects the pronunciation of words depending on their syntactic environment. It is characterized by significant dialectal variations, where there are differences in pronunciation, vocabulary, and grammar based on regions (Sam and Agbloe, 2024). These variations make the development of a standardized speech recognition system a challenge. Nonetheless, the development of an Ewe dataset would augment ASR research and provide opportunities to develop technologies for over eight million speakers to support education, healthcare, and government services.

Existing speech data collection approaches are not well-suited for a LRL such as Ewe. Thus, a more structured and contextually appropriate approach for collecting Ewe speech data is necessary. To determine the most efficient approach, a focus group discussion with both functionally and non-functionally literate participants was conducted. The discussion revealed that since this study seeks to ensure a diverse representation of the Ewe language, as well as capture all possible complexities and scenarios of the language, then sentence reading would be impractical. This is because most of the study participants would be unable to read Ewe. Thus, image descriptions were considered the most suitable approach for collecting speech data in Ewe. This approach will facilitate the collection of a diverse range of spoken words in the form of sentences and also address the challenges of performing sentence segmentation of audio data manually (Uliniansyah et al., 2016).

Over 8000 images were initially extracted from online sources including Pinterest and Google images. Out of which a subset of 1000 images cutting across 50 categories (such as Sanitation, Tourism, Weather, Technology, Automobile, Security, TransportatioRobbien, Architecture, Fashion, Food, Trading, Hospitality, Lifestyle, Health/Medicine, Agriculture, through Entertainment, Arts/crafts, Science, Mining, Education, Governance, Leisure, Home/Housing, Religion, Engineering, Accidents, Sports, Culture, Family, and Nature) were selected during the focus group discussion. Selected images were required to be easily describable in at least three different ways between 15 seconds and 30 seconds. In addition, images were screened to ensure they were devoid of any nudity or profanity. Context specificity was another consideration. This was to ensure that the selected images were culturally and linguistically relevant for the native speakers of Ewe. The images were uploaded onto an Android mobile app (UGSpeechData) that was developed to collect the data. The images alongside the URLs were integrated into the app's image database. The app was designed to operate on-device and with/without the Internet. Figure 1 shows the data pipeline from image selection to data finalization.

3.2 Study Participants and Speech Data Collection

Almost 2000 volunteers from diverse Ewe-speaking regions were recruited using convenience sampling and snowballing. Participants signed up and were trained to use the app to record image descriptions following a set of predefined rules. They signed the consent form and provided relevant demographics, including their age, gender, and recording environment. In addition, the app retrieved the device's name and the recording timestamp. Subsequently, all this data was stored in a file linked with the audio files.

Participants were required to describe the selected 1000 images in Ewe. Each image was limited to a single recording by a specific participant, and the app would only allow recording to start when there was little or no background noise. Participants could save, replay, and delete their recordings. However, the app was designed to only permit an audio file to be saved if it was between 15 and 30 seconds; if there was less than a three-second pause during the description; and if there were no excessive speech mannerisms/fillers in the description.

Furthermore, to ensure that the recruited participants could speak Ewe fluently, they were initially assigned 10 images and were required to record descriptions of the 10 images

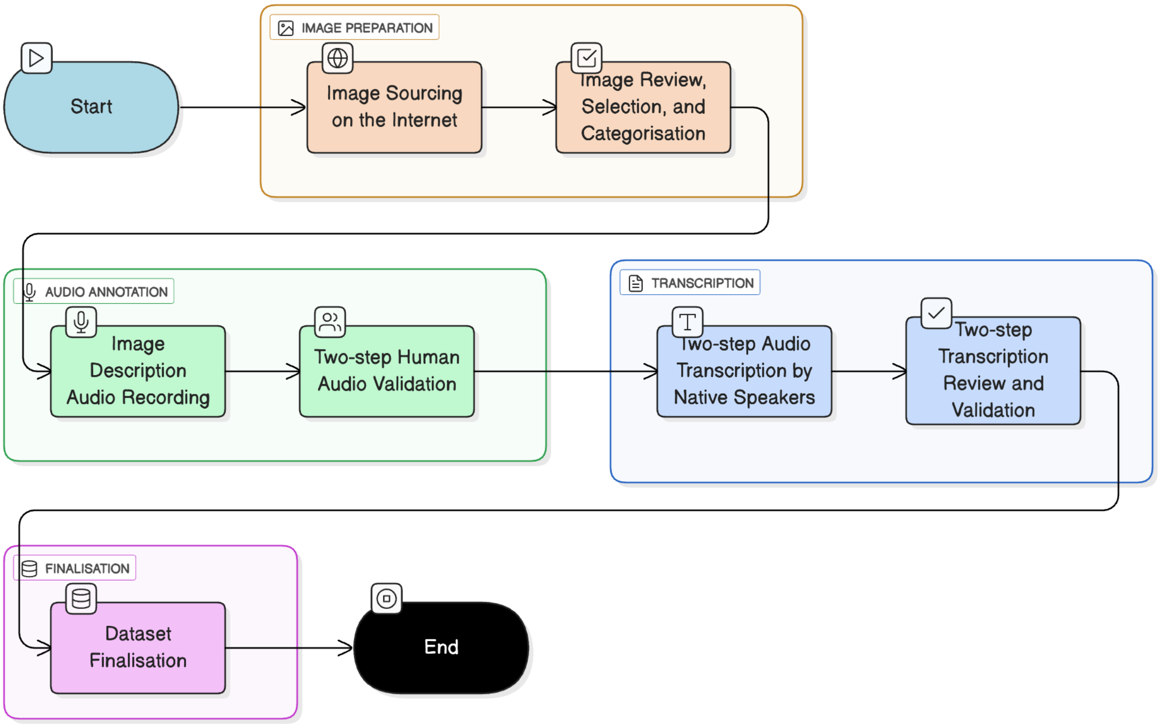
Start
IMAGE PREPARATION
Image Sourcing on the Internet
Image Review, Selection, and Categorisation
AUDIO ANNOTATION
Image Description Audio Recording
Two-step Human Audio Validation
TRANSCRIPTION
Two-step Audio Transcription by Native Speakers
Two-step Transcription Review and Validation
FINALISATION
Dataset Finalisation
End

Figure 1: Data Collection Pipeline Using Image Prompts

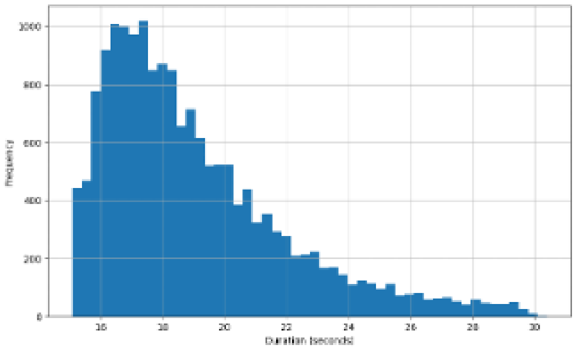

Figure 2: Distribution of audio clip duration in the dataset

to check for adherence to the rules of recording, language fluency, and audio quality. Restrictions to continue recording audio descriptions of the remaining 990 images were removed by the authors if at least 8 out of the 10 descriptions were validated and accepted. An audio file was valid and accepted if: the image description was in Ewe, there was no conflicting background sound in the recording, the audio was naturally audible, the description matched the displayed image and did not contain excessive use of English words, or filler words. Participants with less than 8 accepted recorded audio files were blocked and compensated but could no longer participate in the study. Out of the 2000 participants who were initially recruited, 1905 including 1076 males and 816 females passed the pre-selection phase. Their ages ranged between 18 and 74 years old with a majority between 18 and 45. See Table 1 for a summary of the participant's demographics and the number of audio files. A total of 203,391 audio samples, equivalent to 1,198 hours were recorded. Although participants were required to provide audio descriptions of 1000 images, they were at liberty to stop the recordings at any point. The audio durations range from 15 to 30 seconds, with most clips concentrated between 15s and 20s, and a gradual decline in the number of longer clips from 21s to 30s. Figure 2 shows the distribution of audio duration in seconds.

3.3 Audio Validation and Transcription

Following the collection of speech samples, thirty participants who recorded were reassigned to validate the audio based on the predefined rules. They were further trained on the stringent validation rules specified earlier (see Section 2.1). Out of the 203,391

Gender	No. of recorders	No. of audio files recorded
Male	1076	121,116
Female	816	81,684
Other	13	591
Total	**1905**	**203,391**

Age range	No. of recorders	No. of audio files recorded
18–25	751	56,361
26–35	606	84,544
36–45	287	32,676
46–55	149	17,175
56–65	71	9,796
66–75	30	2,613
Unspecified	11	226

Table 1: Distribution of participants' demographics across audio recordings.

(1198 hours) collected speech samples, 203,336 (1130 hours) speech samples passed the quality checks and formed the Ewe speech dataset. Furthermore, twenty linguists were trained to transcribe the validated speech samples using a structured workflow. We sought to transcribe at least 100 hours out of the 1130 hours of validated speech samples. A maximum of 240 audio files were randomly assigned to a transcriber every 48 hours. Each file was transcribed by two linguists and in situations where there are conflicts in the transcription, the audio will be passed on to a third linguist for conflict resolution. To facilitate transcription, a custom Ewe keyboard was developed to incorporate diacritics and special characters essential to the language. The keyboard utilizes the standard QWERTY keyboard layout and incorporates all special characters to support the Ewe orthography (i.e., including diacritics and tonal marks). The Ewe alphabet consists of 30 characters including the 26 letters of the English alphabet, excluding c, j, and q which were replaced by ɔ, ɣ, and ƒ respectively. In addition to the standard alphabet, the Ewe keyboard includes these special characters: ɖ, ŋ, ɛ, ɔ, ɣ, ʋ.

4 The Ewe Speech Dataset and Automatic Speech Recognition Experiment

The generated dataset consists of 203,336 (1,130 hours) validated audio speech samples, along with 19,152 (106.4 hours) of transcribed audio containing 31,756 unique words. Each audio file is between 15 and 30 seconds long.

Audio speech samples were received from participants in two regions of Ghana: Greater Accra and the Volta Region. Within the Volta Region, data was collected from eight towns, namely Anloga, Keta, Peki, Akatsi, Ho, Juapong, Kpando, and Sogakope. The recordings were done in different environments, but the majority were done outdoors. Specifically, 118,193 recordings were done outdoors, 74,169 were done indoors, 2,465 were done in offices, 66 were done in studios, 144 in buses, and 6,755 in unspecified environments. The dataset is open-source and available at GitHub and Science Databank (Wiafe et al., 2025). See Table 2 for a summary of the Ewe dataset. Next, using the transcribed audio recordings, we test the suitability of the generated speech corpus for automatic speech recognition and also conduct a qualitative error analysis of the predicted transcriptions.

graphicx

4.1 Data Preparation, Fine-tuning and Evaluation

The initial dataset used for modeling comprised 106.4 hours of transcribed audio, encompassing 19,152 audio files. To ensure data quality and to eliminate potentially invalid entries, audio files that exceeded 30 seconds in duration and transcriptions containing fewer than 10 characters were excluded. Following this refinement, the final dataset consisted

Gender	Total no of audio files	Equivalent in hours	Outdoors	Indoors	Other	Office	Car	Studio	Bus
Total no. of audio files by environment									
Male	121 116	673.98	68 300	46 162	3 438	1 850	931	412	23
Female	81 684	453.80	49 851	27 513	3 317	615	17	250	121
Other	536	2.98	42	494	0	0	0	0	0
Totals	**203 336**	**1 130.76**	**118 193**	**74 169**	**6 755**	**2 465**	**948**	**662**	**144**

Summary of Transcribed Files

Gender	Total no of audio files	Equivalent in hours	Outdoor	Indoor	Other	Office	Car	Studio	Bus
Male	10 870	60.39	4 372	5 729	398	210	140	21	0
Female	8 282	46.01	4 929	3 107	114	80	0	52	0
Totals	**19 152**	**106.4**	**9 301**	**8 836**	**512**	**290**	**140**	**73**	**0**

Table 2: Summary of the dataset (validated and transcribed)

of 19,149 audio files with a sampling rate of 16kHz. The dataset was partitioned into training sets (13,382 files, 70%), test sets (3,847 files, 20%), development sets (1,535 files, 8%), and validation sets (385 files, 2%). In terms of speaker distribution, the training set included 163 unique speakers, while the test, development, and validation sets contained 137, 130, and 109 unique speakers, respectively.

We selected the Whisper Yoruba and Shona base models (Radford et al., 2022) as base models due to the linguistic similarities between Yoruba and Ewe. Both languages share a similar writing system, are tonal with three tone levels, and exhibit some lexical overlap. For example, ”mouth” is enu in Yoruba and enu/nu in Ewe, and ”father” is baba in Yoruba and papa in Ewe. We fine-tuned both base models on the prepared dataset using Google Colab with NVIDIA A100 GPUs.

4.2 Training setup

The model was fine-tuned with the following hyperparameters: a batch size per-device of 16, gradient accumulation steps of 1, and a learning rate of 1e-5. We used the AdamW optimizer and applied a constant_with_warmup learning rate scheduler with 50 warm-up steps. Mixed precision training (fp16) and gradient checkpoint were enabled to reduce memory usage. The training process consisted of 2400 steps, and we evaluated the model's performance every 400 steps using the Word Error Rate (WER), arthographic error rate (OER), and character error rate (CER) as the primary metrics. These are widely used metrics for evaluating ASR performance (Fatehi et al., 2025; Mensah et al., 2025). Table 3 summarizes the results of the training loss, validation loss, OER, CER, and WER achieved at each evaluation checkpoint for the Shona and Yoruba base model. It was observed that both models exhibited similar performance trends across metrics. While training loss consistently decreased throughout, validation loss began to plateau after approximately 1600 steps. The lowest error rates across all metrics were recorded at 2000 training steps, with the Yoruba base model achieving an OER of 44.98%, WER of 37.12%, and CER of 12.43%, and the Shona base model achieving an OER of 45.11%, WER of 37.17%, and CER of 12.50%. Beyond this point, error rates showed slight increases, suggesting possible overfitting. Consequently, the 2000-step checkpoint was selected as the best-performing model for Ewe ASR. These results suggest that both base models adapt well to Ewe, with the Yoruba base model slightly outperforming the Shona model on all error metrics. Table 4 shows sample transcriptions predicted by the final model and the corresponding original text using the validation set. Irrespective of the relatively high error rates, the model was observed to make intelligible transcriptions.

4.3 Qualitative Error Analysis on the Predicted Transcriptions

Although it may be argued that the WER of 37% is high, (Chizzoni and Vietti, 2024) posit

Step	Training Loss	Validation Loss	OER (Shona/Yoruba)	WER (Shona/Yoruba)	CER (Shona/Yoruba)
400	0.50	0.58	52.37/51.88	44.57/44.15	15.05/14.89
800	0.48	0.52	48.49/48.65	40.52/40.66	13.69/13.75
1200	0.38	0.49	47.10/46.80	38.72/38.46	13.22/13.03
1600	0.36	0.48	46.08/45.92	37.86/37.71	12.83/12.70
2000	0.31	0.48	45.11/44.98	37.17/37.12	12.50/12.43
2400	0.31	0.47	45.56/45.43	37.58/37.48	12.86/12.97

Table 3: Model performance for the Shona and Yoruba base model

that the CER is a better evaluation metric in instances where the base model was not trained on the LRL data in question. Regardless, a qualitative error analysis was conducted to understand factors contributing to the relatively high error rates (see Table 3). Although the model generally produced intelligible transcriptions, several recurring challenges were identified across orthographic, linguistic, and acoustic dimensions.

1) **Orthographic inconsistencies**

   a) **Non-standard spelling of English loanwords.** Because most loanwords lack fixed Ewe spellings, transcribers wrote them phonetically. *Example:* "machine" appeared as masini or mashini.

   b) **Dialectal vs. formal spellings.** Mixing of Southern-Ewe forms with the formal standard produced mismatches. *Example:* model: yi nye (Southern) vs. reference: si nye (formal).

2) **Morphological challenges** — the model sometimes mis-segmented Ewe's agglutinative morphemes.

   a) **Reference:** ...enye nugometsi kpakple agbalē gbadza aɖe...
   Prediction: ...enye nugo me tsi kpakple agbalē gbadza aɖe...
   Error: nugometsi → nugo me tsi

   b) **Reference:** Devia ɖewo tsi atsitre…
   Prediction: Devia ɖewo tsatsitre…
   Error: tsi atsitre → tsatsitre

   c) **Reference:** exɔtudzikpɔla aɖe le wo gbɔ
   Prediction: exɔ tu dzikpɔla aɖe le wo gbɔ
   Error: exɔtudzikpɔla → exɔ tu dzikpɔla

3) **Phonetic confusions** — substitutions between phonetically similar consonants, especially affricates vs. stops.

   a) /dz/ ↔ /z/
   Reference: Dzo bi teƒe sia
   Prediction: Zo bi teƒe sia
   Error: Dzo → Zo

   b) /dz/ ↔ /d/
   Reference: Nufiala dzidzim be ya fia nu
   Prediction: Nufiala didim be ya fia nu
   Error: dzidzim → didim

4) **Dialectal pronunciation variation** Ewe exhibits major dialectal differences. The model often defaulted to Southern-Ewe pronunciations, causing mismatches when the reference used another variety. *Example:*
   Reference: Yevuwo wonye (standard Ewe)
   Prediction: Yewuwo wonyo (Ewe-Dome dialect)
   Error: wonye → wonyo

5) **Mistranscription with insertion/substitution** Rare but notable cases where acoustically ambiguous segments led to entirely different words: *Example:*
   Reference: Buno aɖɛ le suku
   Prediction: Nubuno aɖɛ le suku
   Error: Buno → Nubuno

Original Text	Predicted Text
Ŋutsu et̄ɔwo le mashinidɔwɔƒe le dɔ wɔm, wo dometɔ eve tɔ ɖe gakpo gā lɔbɔ aɖe yi le dzi ŋu le ŋku lem ɖe eŋu. Ɖeka tɔ ɖe adzɔge le wo kpɔm. Wodo awu amadede si nye orange eye woɖɔ dɔwokukuwo hā. *Three men are at work in a machine shop, two of them are standing on a big, long steel stick that is above and staring at it. One stood at a distance watching them. They wore orange outfits and helmets.*	Ŋutsu et̃ɔwo le machinidɔwɔƒe le dɔ wɔm. Wo dometɔ eve tɔ ɖe gakpo gã lɔbɔ aɖe yi le dzi ŋu le ŋku lém ɖe eŋu. Ɖeka tɔ ɖe adzɔge le wo kpɔm. Wodo awu amadede yi nye ɔɖɔɛndzi eye woɖɔ dɔwɔ kukuwo hã. *Three men are at work in a machine shop, two of them are standing on a big, long steel stick that is above and staring at it. One stood at a distance watching them. They wore orange outfits and helmets.* WER=39%, CER=11%, Cosine Similarity=95%
Woɖo kpl̄ɔ at̄ɔ ɖe xɔ me. Amewo nɔ kpl̄ɔ ŋu hamehame. Ame bubu aɖewo nɔ wo ŋgɔ eye wonɔ nu ƒom na wo. Ame siwo le kplɔ ŋu la ɖo to hele amea ƒe nu ƒom sem. *They set up five tables in a room. There were all kinds of people around the table. There were others in front of them and talking to them. The people at the table were quiet and listening the talk.*	Woɖo kplɔ at̄ɔ ɖe xɔ me. Amewo nɔ kplɔ ŋu hamehame. Ame bubu aɖe nɔ wo ŋgɔ eye wonɔ nu ƒom na wo. Ame siwo le kplɔ ŋu la ɖo to hele amea ƒe nu ƒom sem. *They set up five tables in a room. There were all kinds of people around the table. There were others in front of them and talking to them. The people at the table were quiet and listening the talk.* WER=18%, CER=4.5%, Cosine Similarity=98%

Table 4: Sample of ground truth vs. predicted transcriptions

5 Discussion

The performance of the fine-tuned Whisper Yoruba model on the Ewe dataset, achieving a word WER of 37% and CER of 12% is consistent with expectations for low-resource environments (LREs). Previous studies (Fatehi et al., 2025), have shown that automatic speech recognition (ASR) is dependent on the volume and quality of training data. The Common Voice project (Ardila et al., 2020) demonstrated that while community-driven data collection efforts help address issues of limited labeled speech date, achieving low error rates remains challenging without significant resources for transcription standardization and quality control. In high-resource environments (HREs), models achieve significantly lower error rates of less than 10% because they are trained on tens of thousands of hours of annotated speech (Baevski et al., 2020). However, in LREs such as the Ewe language, even with 106 hours of transcribed data and diligent data collection efforts, the model's performance may have been constrained by the relatively small labeled data, dialectal variation, and orthographic inconsistencies. Besacier et al. (2014) argue that irrespective of advanced modeling techniques, there is an elevated risk of error rates, particularly for tonal and morphologically rich languages that lack large, domain-specific corpora. Dialectal variation and phonetic diversity, as observed in this study for Ewe, introduce substantial complexity. Dialectal shifts across regions (Ghana, Togo, Benin) result in pronunciation and vocabulary differences that pose challenges for ASR systems trained on limited samples. Orthographic inconsistency further exacerbates model error rates, as shown by (Kim et al., 2025) who highlighted the difficulties of building reliable models for languages with non-standardized or emerging writing systems. More specifically, the qualitative error analysis showed that orthographic inconsistencies, particularly with English loanwords and dialectal variations, introduced ambiguities that affected transcription accuracy. It was observed that the morphological complexity of the Ewe language, especially its agglutinative nature, led to frequent segmentation and merging errors. Additionally, phonetic confusion between similar sounds (e.g., /dz/ vs. /z/) and dialectal variations in pronunciation may have compounded the ASR model's challenges. Despite leveraging transfer learning from a linguistically related language (Yoruba), the results show that adaptation alone cannot fully resolve the dialectical diversity or phonetic complexity intrinsic to Ewe.

6 Conclusion

This study introduced a large-scale, validated speech corpus for the Ewe language, comprising 1,130 hours of audio recordings and 106 hours of transcriptions. By employing an innovative image-based prompting method and controlled crowdsourced data collection strategy, this study provides a linguistic resource for advancing ASR development in LRE. Finetuning experiments with the Whisper Yoruba model demonstrated the dataset's utility while also highlighting persistent challenges posed by dialectal variation, orthographic inconsistency, and morphological complexity. Findings from this study affirm that transfer learning from related languages offers practical advantages but cannot fully substitute for in-domain, dialectally representative datasets. This study also suggests the need for morphologically aware and dialect-sensitive modeling approaches to improve ASR accuracy for languages such as Ewe. Future work should prioritize leveraging this study's unlabeled speech corpus through self-supervised learning techniques and explore domain-adapted language modeling to enhance transcription reliability for critical applications such as healthcare, education, and public service delivery. By addressing these linguistic and technological gaps, this research lays the foundation for more inclusive speech technologies that preserve and promote the use of indigenous African languages. The data splits and trained model is publicly available on GitHub and Huggingface.

Limitations

While self-supervised learning approaches, such as wav2vec 2.0 (Baevski et al., 2020), have shown promise in reducing the dependence on labeled data by leveraging large amounts of unlabeled audio, their application is not without challenges. Although the dataset collected in this study comprises 900 hours of unlabeled Ewe speech, the computational constraints limited the feasibility of training with wav2vec. Access to large-scale computing resources remains a significant bottleneck in LRE research. This study argues that, in LREs, model performance is fundamentally constrained by linguistic complexity and computational resources rather than modeling innovations alone. Addressing these challenges is essential to advancing equitable access to speech technologies for underrepresented languages. Future research may build on these findings by prioritizing the development of scalable methodologies and resources that enable the advancement of ASR technologies for LRLs, such as Ewe and other Ghanaian languages.

Ethics Statement

Ethical approval for this study was obtained from the Ethics Committee for Basic and Applied Sciences, University of Ghana. All participants, including recorders, validators, and transcribers, were informed of the study's goal and that the collected data would be used for research purposes only. Also, the consent form specified that participation was voluntary and participants could withdraw at any point. All participants were compensated for their respective contributions.

Acknowledgement

This research was funded by Google LLC (Google). We also acknowledge the technical support and iterative feedback provided by the Google Research Ghana team during data collection.

References

Kwame Badu Antwi-Boasiako and Kofi Agyekum. 2022. Globalization, colonization, and linguicide: How ghana is losing its local languages through radio and television broadcast. *International Journal of Humanities and Social Science*, 12(8):142–151.

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.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 12449–12460.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56(1):85–100.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with amazon's mechanical turk. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Ilaria Chizzoni and Alessandro Vietti. 2024. Towards an ASR system for documenting endangered languages: A preliminary study on sardinian. Technical report, Free University of Bozen-Bolzano.

De Graft Johnson Dei. 2024. Sustainability and development of Ewe communities in ghana through indigenous knowledge management practices. *Collection and Curation*, 43(4):111–123.

Kavan Fatehi, Mercedes Torres Torres, and Ayse Kucukyilmaz. 2025. An overview of high-resource automatic speech recognition methods and their empirical evaluation in low-resource environments. *Speech Communication*, 167:103151.

Alexandru-Lucian Georgescu, Horia Cucu, Andi Buzo, and Corneliu Burileanu. 2020. RSC: A Romanian read speech corpus for automatic speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6606–6612, Marseille, France. European Language Resources Association.

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

Umar Adam Ibrahim, Moussa Mahamat Boukar, and Muhammed Aliyu Suleiman. 2022. Development of Hausa dataset: A baseline for speech recognition. *Data in Brief*, 40:107820.

Robbie Jimerson, Kruthika Simha, Raymond Ptucha, and Emily Prud'hommeaux. 2018. Improving ASR output for endangered language documentation. In *Proceedings of the 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 187–191.

Minu Kim, Kangwook Jang, and Hoirin Kim. 2025. Improving cross-lingual phonetic representation of low-resource languages through language similarity analysis. arXiv preprint arXiv:2502.01234.

Mark Atta Mensah, Isaac Wiafe, Akon Ekpezu, Kwame Appati, Jamal-Deen Abdulai, Akosua Nyarkoa Wiafe-Akenten, Frank Ernest Yeboah, and Gifty Odame. 2025. Benchmarking akan asr models across domain-specific datasets: A comparative evaluation of performance, scalability, and adaptability. In *Accepted - Future Technologies Conference.*

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

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public-domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, pages 5206–5210.

Kay Peterson, Audrey Tong, and Yan Yu. 2021. OpenASR20: An open challenge for automatic speech recognition of conversational telephone speech in low-resource languages. In *Proceedings of Interspeech 2021*, pages 4324–4328.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*, pages 28492–28518.

Alessandro Ragano, Emmanouil Benetos, and Andrew Hines. 2020. Development of a speech quality database under uncontrolled conditions. In *Proceedings of Interspeech 2020*, pages 4616–4620.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Jemima Sam and Cynthia Ablah Agbloe. 2024. Lexical variations in the Ewe language spoken in ho in the volta region of ghana. *Journal of Education*, (7):69–87.

Tanja Schultz. 2002. Globalphone: A multilingual speech and text database developed at karlsruhe university. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 345–348.

Mohammad Teduh Uliniansyah, Gunarso, Elvira Nurfadhilah, Lyla Ruslana Aini, Juliati Junde, Fara Ayuningtyas, and Agung Santosa. 2016. A tool to solve sentence segmentation problems on preparing a speech database for an indonesian text-to-speech system. In *Procedia Computer Science*, volume 81, pages 188–193.

Isaac Wiafe, Jamal-Deen Abdulai, Akon Obu Ekpezu, Raynard Dodzi Helegah, Elikem Doe Atsakpo, Charles Nutrokpor, Fiifi Baffoe Payin Winful, and Kafui Kwashie Solaga. 2025. Advancing automatic speech recognition for low-resource ghanaian languages: Audio datasets for akan, ewe, dagbani, dagaare, and ikposo. *Data in Brief*, 61:111880.