Longa: An automated speech recognition tool for Bantu languages

Nelson Mganga[a], Eliot Jones-Garcia[b*], Andrea Gardeazabal Monsalue[c], and Jawoo Koo[d]

[a]*Zindi, Cape Town, South Africa*
[b]*University of Nottingham, Nottingham, England*
[c]*International Maize and Wheat Improvement Center (CIMMYT), Texcoco, Mexico*
[d]*International Food Policy Research Institute (IFPRI), Washington, DC, USA*

INFO

Submitted	31 December 2023
Keywords	AI, Audio Analytics, Transfer Learning, Bantu
Flagship	Co-LAB
Work Package	Enabling environment

EXECUTIVE SUMMARY

Farm Radio International (FRI) and the CGIAR Initiative for Digital Innovation have collaborated on the development of an end-to-end, automatic speech recognition pipeline for the transcription, translation and analysis of Swahili and Luganda. This task is particularly challenging due to the number of languages used by FRI's clients, and the limited training data available for speech recognition in African languages. The tool is named 'Longa', or 'Lets chat' in Swahili. Longa will be used to answer the surfeit of phone calls currently being received from small-holder farmers asking questions about radio programs which FRI do not presently have the capacity to address. When fully implemented, Longa should allow FRI to design their broadcasts more intricately in line with the needs of farmers, and better deliver insights to those most in need, such as female and youth farmers. Key results from the collaboration include a proof of concept for Longa, building on open-source models and open access corpora, to be shared with the developer community upon completion of the final tool; a series of design principles iteratively and collaboratively developed to reflect the common values and goals of FRI and the CGIAR. These are being used to give shape to, and evaluate, the usability of Longa as development progresses; a 10% improvement upon the state-of-the-art automatic speech recognition in Luganda-Radio-Speech performance and accuracy; some improvement of the performance with audio enhancement processes. Running on real world data, Luganda is a notoriously difficult language due to intonation and low training resources. This is further complicated by quality of phone call audio, and files were not recorded with speech recognition in mind. Presently only around 10% of the existing FRI data is of sufficient quality; and proof that fine-tuning is an effective approach to expanding Longa to new languages. By transferring knowledge of language, grammar, syntax and meaning from Swahili and Kinyarwanda to Luganda, we see great promise in moving forward with other Bantu languages and other challenging languages on the African continent. The next steps of the collaboration will focus on the analysis and interpretation of an aggregation of farmer phone calls, and integration with the existing FRI workflow and software. The CGIAR team will workshop a series of methods to investigate and action

1. Problem Statement

Farm Radio International (FRI) is an international non-profit which hosts agricultural talk-radio shows in 40 countries across sub-Saharan Africa, to make expert advice accessible to and comprehensible by small-scale farmers. They work on the principle that 'Farmers have a lot to say', and 'as nations, organizations, and individuals, we all must commit to listening and taking action together… sharing knowledge and giving voice to farmers.' As such, not only do they broadcast discussions between experts, tailored toward specific local problems, but they also receive and answer questions from farmers about their practices and demands. Unfortunately, a significant portion of these responses goes unanswered, as without the staff to manually translate each phone call, the information remains inaccessible. This presents a vast and rich dataset, speaking directly to the needs and desires of smallholders.

In September of 2023, FRI launched their Nature Based Solutions (NBS) project. This aims to initiate "actions to protect, sustainably manage and restore natural and modified ecosystems in ways that address societal challenges effectively and adaptively, to provide both human well-being and biodiversity benefits," according to the International Union for Conserva-

*Corresponding author. *Email address:* eliot.jones@nottingham.ac.uk

tion of Nature (IUCN). In six countries (Burkina Faso, Côte d'Ivoire, Ethiopia, Ghana, Uganda and Zambia), "On Air Dialogues" with rural communities are being hosted by 20 radio stations to discover local priorities about, ideas for, and experiences with, harnessing nature to adapt to a changing climate. Based on the results, more than 200 entertaining radio documentaries will be produced that showcase local solutions, and high-impact interactive radio programs will support rural Africans in bringing nature-based solutions to their communities. The best ideas that emerge from these programs will be spread Africa-wide through scripts and stories shared with a network of over 3,500 broadcasters in 38 countries.

These radio programs aim to create an environment where women and youth have an increased role in contributing to conversations about the climate — and that decision and policy makers are paying attention to their voices.

2. Proposal

As outlined in Figure 1, it has been our intention to investigate the possibility of automating the process of transcribing, translating and analysing these data and to propose a tool—named *LONGA* (Swahili for '*chat*' or '*let's talk*')—which might collect rich information from a broad pool of farmers, leading to better-suited and longer-lasting interventions. Having completed a proof of concept in 2022, this report details a visit to the FRI offices in Uganda to better understand their challenges, the progress of model development, and the intentions of the collaboration for the next year. Stages 5 to 8 of Figure 1 will be the next focus of this work, devising the way in which data can be best interpreted, insight interfaced, and research findings established.

3. Design Principles

During our collaboration with FRI, the following design principles have been synthesized to give shape to, and evaluate, Longa:

1) Collaborative: the tool will be based on the needs, values and expertise of actors through the FRI value chain.
2) User-centered: the tool will be capable of transcribing, translating, and delivering actionable insight.
3) Longevity: the tool will eventually be handed off to FRI for independent use.
4) Data driven: the tool will be sufficiently robust to be trained on real world data, to satisfy a real world use case
5) Flexibility: the tool will be able to be expanded to include new languages, without requiring significant input of the end user.
6) Advance the field: the tool will utilize state-of-the-art technology, enhancing the FRI workflow whilst contributing to the broader AI community.
7) Gender equality: the tool is intended to expand the number of female and youth farmers in voicing agricultural concerns.

These principles were developed through iterative discussion and collaboration between CGIAR and FRI staff and give shape to our design pathway, as described below.

4. Design Pathway

Following a significant leap in artificial intelligence (AI) research and development, the number of farmer that can be served by agricultural extension is increasing, as is the quality of advice and support by collecting data about their clients and providing site-specific, evidence based information. Natural language processing (NLP) tools, in particular, have enhanced remote-digital communication channels allowing for cross-language interactions with little to no human intervention, overcoming barriers of illiteracy and geographic isolation.

Two issues have, however, lessened the impact of AI and NLP in agricultural extension.

The first of these is the continued lack of access to smart phones or other advanced information communication technologies by farmers. Smart phones are often looked upon with suspicion, as digital spaces may introduce new, unsafe spaces, thus commanding little trust within rural communities. FRI, on the other hand, through their decision to exploit the well-established and highly trusted medium of radio across Africa, provide a lifeline to underserved farmers. FRI collaborate with locally operated radio stations, connecting the acquisition of information with the legitimacy of government and the draw of entertainment.

The second issue is in relation to design. The advance of African NLP has been severely limited due to the low digital presence of these languages, relative to more widely used western languages. Known as low-resource languages, it is difficult to attain sufficient training data from which to model and deliver the benefits of these tools to communities that are most in need.

Recent efforts have attempted to increase the representation of low-resource languages for NLP through initiatives such as HuggingFace and Mozilla's Common Voice project. These

Figure 1 FRI augmented workflow (Source: Authors)

open-source networks of contributors have benefitted African languages not only with models, but user-contributed corpora that can be used to build more accurate and advanced tools. Other efforts have attempted to transfer learning from higher resource languages, utilizing their linguistic properties to reduce the necessary training data to achieve a successful output. These are often difficult to implement in practice, however, owing to the relatively complex and uninterpretable model designs required to achieve higher accuracies, as well as the fact that the datasets on which the models are trained are not representative of everyday interactions.

This study is aimed at one aspect of NLP, adapting automatic speech recognition (ASR) technologies to low resource languages, with a focus on the Bantu group, building a tool that

Figure 2 FRI, CGIAR, and Radio Simba (Source: Authors)

can readily be integrated into workflows for real-world use. This project is therefore meeting a significant gap in the advancement of AI, bringing benefit to populations which would not otherwise have access due to low investment and skills in the private sector, localizing these tools and knowledge to support future development.

4.1. Collaborate

Year two began by developing a more robust, end-to-end ASR pipeline, trained and tested on open-source speech data. This was to be used in an in-person meeting with Farm Radio International, scheduled for late June. Figure 2 shows the first author attending a meeting between FRI and one of their client radio stations.

The visit was to serve two key objectives:

- A chance for FRI and CGIAR staff to collaborate, discuss current progress as well as next steps and deployment options for the ASR model.
- To attend the launch of a new FRI initiative with radio stations in Uganda.

The initiative launched by the Digital Innovation team involved opening up Uliza Interactive—the software used by FRI to manage the radio show recordings—for public use, allowing radio stations to use the application for programs other than those related to FRI. The software was redesigned to allow for a more general-purpose and user-friendly interface that would allow users to set up profiles (known as campaigns) for different shows and manage each independently.

Discussions around the ASR model involved introducing the Digital Innovation team to the concept underlying the model's design and demonstrating the model's capabilities. This was to help understand the motivation behind the current approach as well as to better align expectations for Longa's performance. The discussions also kickstarted conversations on how the

model would be integrated into the FRI's existing workflow as well as the necessary work needed to deploy the model into production.

Table 1 is a summary of the activities along with the output that resulted from each as well as some notes on a few potentially important observations.

4.2. User-centered

The current strategy utilized by FRI aims to inform farmers' decisions through radio. This makes radio stations the primary target of FRI interventions who then help to indirectly impact farming outcomes through the various programs. For this model to prove effective, FRI operates under the assumptions that:

- Radio is the primary source of information within the community, commanding significant respect/trust among farmers.
- Listeners are representative of the entire farmer population and therefore data collected during interviews represents shared views of the farming community. These are disaggregated to different catchment areas, however, with distinct communication and content according to local attributes.
- Information shared through radio programs spreads from listeners to the community thus ultimately impacting the larger population. This is often restricted by social factors, for example, in male headed households, women are not allowed to participate.

To ensure effectiveness of the intervention, FRI oversees the design of each radio program, incorporating research techniques. This allows the Monitoring, Evaluation, Research, and Learning (MERL) team to assess each project's impact and in turn inform policy. Moreover, to ensure effective targeting, the radio shows are aired to the public but only in a language predominant in a specific community. This is done to encourage participation from local communities so as to collect data from within the desired population.

Table 1 Uganda visit activities

Date	Activity	Output	Comment
21 June 2023	• Introduction to FRI and overview of the organization's operation. • Introduction to the digital innovation team and the Uliza app.	PowerPoint presentation from FRI	The various departments operate independently with each team focusing on its objective e.g., the DI team working on the technology and MERL specializing in impact evaluation
22 June 2023	• Introduction to Machine Learning and the ASR model. • Discussion on the workflow and how best to integrate.	Presentation from Nelson and discussion with the digital innovation team	DI team seem to appreciate the research and are keen on seeing Longa integrated with Uliza
23 June 2023	• Field work at Radio Simba. • Unveiling Uliza for general purpose use.	Pitch meeting with key personnel from the radio station	The radio station was keen to take up Longa for use with shows other than farm radio programs
26 June 2023	• ASR model demo on Luganda. • Human evaluation of model's performance.	Transcriptions of several recordings from FRI and evaluation by native speakers	Model appears to have learnt the acoustic features but fails to incorporate grammatical/syntactical information/knowledge
27 June 2023	• Training at Radio Simba for the general purpose Uliza Interactive. • More model testing and evaluation.	Uliza interactive handover to Radio Simba	Radio stations have full access and control of the app with freedom to organize based on their needs. Further model testing revealed Luganda to be a complex language that even natives find difficult to formalize.

Uliza Application

Uliza is the software that FRI uses to manage the various projects running with radio stations. The software comprises Uliza Interactive—the user interface through which the radio programs and interview responses/recordings are managed—as well as Uliza Log which comprises the database and is used to store and manage data related to the various radio programs.

The design of the software would allow for easy integration of the model into FRI's current workflow. Since Uliza interactive already supports manual transcription of recordings, incorporating the model would only require the software to call the model, run inference on the speech data collected and output the transcriptions for the user.

Uliza Interactive for General Use

Other than radio programs directed at farmers, FRI has opened Uliza Interactive to partner radio stations allowing them to make full use of the platform for their own personal needs. This involves a change in the UI with more management options, such as adding users and assigning roles, creating customizable campaigns for different shows. This initiative came as a result of previous requests by radio stations to have a more general-purpose platform that would better integrate with their workflow.

The launch of the initiative was included as one of the objectives of the visit where the FRI team pitched the technology to Radio Simba who appeared keen to implement it. Along with the improved interface, radio stations would receive up to $2,000 worth of airtime to help cover the costs of calling back listeners while carrying out radio interviews.

ASR Model

During the visit, the concepts underlying the model's functioning were discussed. This was a chance to gage the transferability of the tool, as well as expanding on some of the choices made in the design, giving and receiving feedback as to the utility of the output.

It was revealed that it would be most useful to have Longa integrated within the NBS program where it would allow for field-testing and evaluation. The program is currently ongoing in Uganda with a focus on shows presented in Luganda. This demanded a change in strategy and focus of model development.

Luganda ASR

Once Luganda was decided on as the language of focus, the model was finetuned to open-source Mozilla data. This included around 100 hours of annotated speech data with about 50 hours of it used to train the model and the remaining used to evaluate performance on unseen data. The model was trained for 28 epochs achieving a validation (unseen data) WER of around 34%.

Following evaluation with a portion of the open-source data, the model was used to transcribe sample data collected by the FRI team which included around 97 audio recordings. Having not been manually transcribed, the model's performance on the recordings was evaluated with help from some of the Uganda FRI staff. This not only gave insight to the model, but also the complexity of the language itself.

Luganda Linguistics

Once the model was run on the recordings, native Luganda speakers aided in evaluating the performance by listening to recordings and comparing the transcribed text with the utterances in the audio files.

The exercise revealed that the model managed to accurately identify the acoustic patterns in the audio but failed to incorporate the grammatical rules of the language. This was most evident in recordings with muffled speech where the model would output tokens that sounded most similar but were grammatically incorrect as pointed out by native speakers.

Apart from technical difficulties, however, having native speakers help interpret the results revealed that Luganda is governed by relatively complex rules that are even more complicated for the written language. Moreover, it appears the language is predominantly spoken with many not being able to write it. This indicates that some aspects of the spoken language may not have necessarily developed in written form, especially in digital media.

Along with the complex structure of Luganda, the language is often used interchangeably with Lusoga, another language most spoken by natives. Mixing the two languages is such common practice that natives would not identify the switch unless explicitly instructed to. This is the case with the FRI data collected where some interviews were carried out in Lusoga and it was only through interpreting the model results that this was pointed out.

4.3. Longevity

To approach the morphological complexity Luganda demanded a review of the model pipeline. While uneven, this space of AI is advancing at a fast pace, in well-documented, open-source spaces – and we were able to draw on state-of-the art

models to inform our work.

Chanie et al. (2023) developed a multilingual ASR model for Kinyarwanda, Swahili, and Luganda. The study made use of Mozilla's Common Voice initiative to access transcribed audio data for each language, combining the datasets and jointly training the model to better capture the patterns within the languages and achieve a multilingual speech recognition model that performs well across the languages.

Mukiibi et al. (2022) built an end-to-end speech recognition system used to monitor radio conversations for keywords that can later be utilized to carry out further analysis. The researchers also collected over 150 hours of radio speech data that was used extensively in this study to evaluate model performance.

Recordings from Mozilla's Common Voice initiative proved to be less representative of radio/phone conversations owing to relatively more background noise and, in many cases, having less clearer utterances which significantly impacted model performance.

Longa was then finetuned and evaluated on two open-source datasets that include a Luganda radio speech corpus collated by Mukiibi et al. (2022) as well as a Swahili broadcast news corpus from Gauthier et al. (2016). The Luganda radio speech corpus comprised around 155 hours of speech data, only 20 of which was transcribed and ultimately used to finetune and test the models, while the Swahili dataset comprised 10 hours of transcribed speech data. Each dataset was partitioned into a train, validation, and test set which were used to finetune and evaluate model performance.

This approach has allowed us to satisfy a series of concerns raised by FRI that:

- Longa would at some point become unavailable, whether removed from the internet or protected by a paywall.
- FRI would not be able to build and improve upon Longa over time
- FRI would be dependent on CGIAR for maintenance and hosting.

Utilizing open-source software and corpora ensures that the input data and models will not disappear, rather they will grow and become more reliable through the support of user communities and, being well documented, FRI engineers may adapt and build on the model as and when necessary.

That said, having manual transcriptions of the remaining data will allow for training on a much larger dataset and potentially have a significant impact on model performance.

This includes the ability to deal with code switching as well as to better capture the natural language. Moreover, the radio recordings allow for the model to pick up on issues such as varied sound quality and intonations that are less represented in data that is sourced in a manner as done by the Common Voice project but are more prevalent in phone conversations carried out on air.

4.4. Data-driven

Data Preprocessing

To prepare for real-world data, the model had to be first trained the open-source corpora.

Files containing the location, duration, and transcript for each of the audio files were prepared from the tsv data frame of each dataset. The model then used these to locate the audio

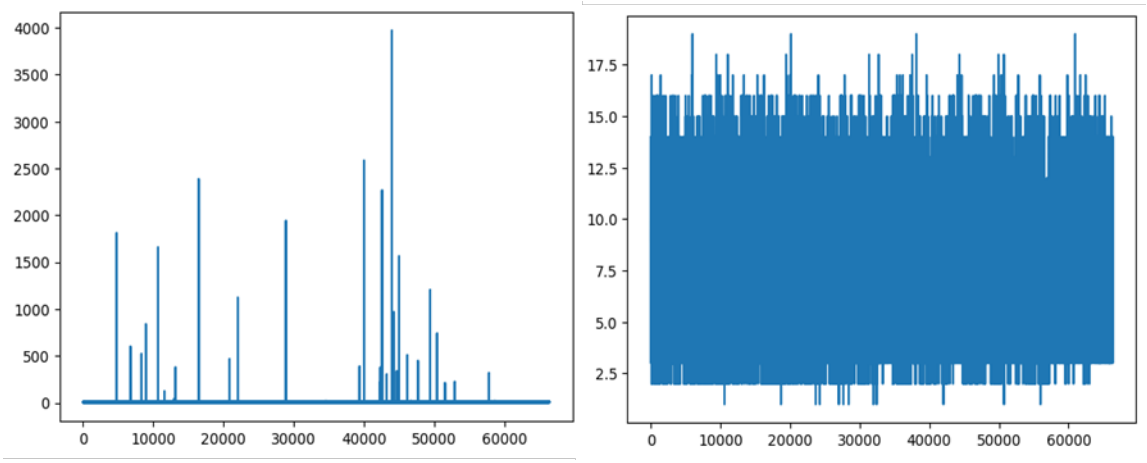

Figure 3. Outlier samples (left) discarded to have more uniform sequence lengths (right) (Source: Authors)

files which would be read and passed into the model during the training and validation process.

Owing to the operation of the model, longer sequences require larger resources such as GPU, and as such, samples containing more than 20 words in the transcript were discarded from the train set (Figure 3). This was to ensure the model trained without running into out of memory (OOM) errors when met with long text sequences. Following the cleaning, each dataset comprised audio files roughly 20s long with no more than 20 words in each.

Model Architecture

Wang et al. (2019) and Prabhavalkar et al. (2023) review various ASR architectures, outlining the pros and cons from the more traditional hidden Markov models to more recent developments in end-to-end models. The articles investigate the various architecture designs and identify the end-to-end model as being more advantageous compared to the traditional architecture. The articles do further comparisons on the performance of three end-to-end mechanisms (CTC, Attention, RNN-Transducer) and identifies RNN-Transducer and attention-based models as having higher performance compared to CTC models, albeit at the cost of model complexity.

KoSpeech—an open-source toolkit for end-to-end Korean speech recognition—was initially used to study the various morpheme-based model architectures. Their approach implements several designs including attention-based models as well as models based on the CTC loss (kospeech, n.d.)

RNN-T Model

The RNN-Transducer (Figure 4) is an advancement on the CTC overcoming its two main limitations where a CTC architecture assumes independence of elements (words, characters, etc) in the output sequence and can only map an input sequence to output sequences shorter than itself.

Moreover, the RNN-T model has language modelling capabilities, owing to its joint network, and supports streaming speech recognition, a feature not supported by attention-based models. The morpheme-based model is thus based on the RNN-T model trained using the RNN-T loss.

The model is composed of 3 components/subnetworks:

- Transcription Network (acoustic encoder/model) which receives the acoustic input sequence and maps it to a feature sequence, mapping onto a high-level representation.
- Prediction Network (language model) which is an RNN network that models the interdependencies within the out-

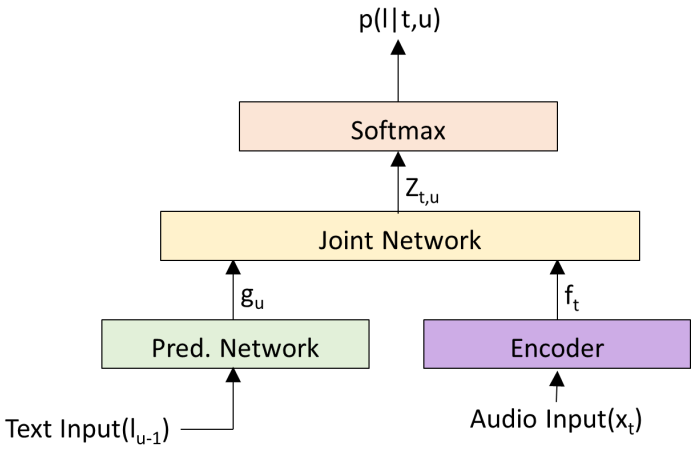

Figure 4 Recurrent Neural Network Transducer (RNN-T) Structure (Source: Authors)

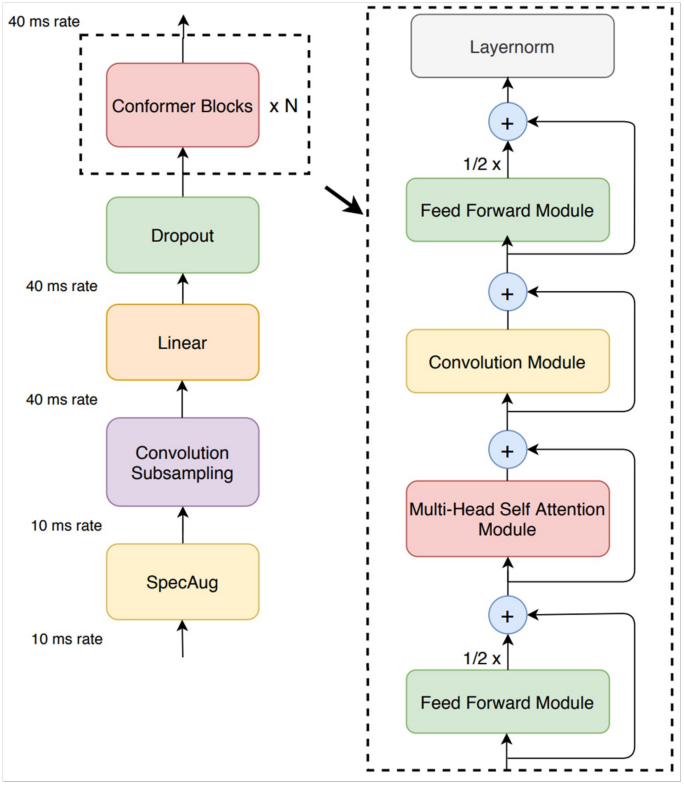

Figure 5 Conformer encoder architecture (Source: Authors)

put label sequence by autoregressively incorporating previously emitted symbols into the model.

- Joint Network that aligns the input and output sequences mixing both acoustic and language model outputs via a monotonic alignment process (Albesano, 2022).

Initial designs of the model utilized the LSTM unit as the encoder, however, since RNNs are only effective at capturing short-term dependencies, the use of transformers for encoders is preferred since it captures global context, allowing for improved performance over RNN models.

Conformer Model

Combining a convolutional neural network (CNN), effective at capturing local context, with a transformer network to replace the RNN encoder improves modelling capabilities and hence performance (Gulati, et al., 2020). Moreover, owing to how the model effectively captures both local and global contexts, a relatively simpler conformer architecture can be implemented to achieve results that are comparable to, or potentially better than, a transformer model.

For example, (Koo, 2021) designed a conformer-based Korean speech recognition system (Figure 5) and compared its performance to a transformer variant observing improved performance with the conformer model. The model utilized in this project will therefore be based on a conformer architecture similar to the one implemented in this study including the parameter selection employed in the model design.

4.5. Flexibility

The experiments conducted in this study entailed making use of openly available pretrained models from Nvidia's GPU Cloud, specifically the Kinyarwanda pretrained model , as well as the multilingual model from Chanie et al. (2023). This was done to not only cut down on the training time but to also highlight the ease with which it is to incorporate new languages with LONGA as well as to establish a reference point to evaluate the tool's performance, accuracy and efficiency.

- Nvidia NeMo: A conversational AI toolkit built to help researchers from industry and academia to reuse prior work (code and pretrained models)—was used to build, experiment, as well as finetune the ASR models. The toolkit comes equipped with a variety of pretrained models and is built on top of pytorch and pytorch lightning, making it relatively easier to experiment and keep record of the results.
- Google Colab: The experiments were carried out through Google Colaboratory, making use of the pro+ version which gives access to up to 40GB of GPU memory and 80GB of RAM on the A100 Nvidia GPU runtime. Owing to the data size and model configurations, the experiments took only around 25GB of VRAM and about 16GB of RAM to run. However, this was at the trade-off of running for longer periods of up to 8 hours for each experiment.
- Nvidia Riva: A GPU-accelerated software development kit(SDK) for building Speech AI applications—will be used to integrate the ASR model(s) with FRI's current workflow. This includes designing a workflow that integrates NeMo as well as Uliza to allow for users to more seamlessly train, finetune, and deploy models of different languages of interest. The latest Riva toolkit is free and open to researchers allowing processing of up to 1,000 hours of audio per day for ASR projects.

By incorporating these back end tools, it is hoped that Longa should be accessible to users beyond the AI community, allowing for use without requiring advanced AI or programming expertise.

4.6. Advance the field

Using Nvidia's NeMo toolkit, the model parameters, such the number of layers in each network as well as the number of epochs to train the model, are specified using config(.yaml) files which can easily be set using readymade scripts from the toolkit's repository. Moreover, for many the ASR models, and especially the Conformer models, there are predefined configurations along with their requirements and use cases in the repo that can be downloaded and modified accordingly to suit experiment needs. Each of the pretrained models used in our experiments was based on the Large Conformer-Transducer

Table 2 Model configuration (Source: Authors)

Component	Size
Encoder Hidden Units	512
Encoder Layers	17
Encoder Attention Heads	8
Decoder Units	640
Decoder Layers	1
Batch Size	16
Epochs	100
Tokenizer Vocab Size	1024
Tokenizer Maximum Sentence Length	4192

Table 3 Test set WER and CER

	Swahili		Luganda	
	WER	CER	WER	CER
rw	44.51	27.92	54.36	26.96
rw (frozen encoder)	45.42	28.6	52.56	22.77
rw-sw-lu	39.86	19.97	51.58	21.66
rw-sw-lu (frozen encoder)	40.29	21.45	52.10	23.13

configuration and utilized a SentencePiece Byte-Pair Encoding (BPE) tokenizer from HuggingFace (Table 2).

Model Training and Evaluation

Each of the models were finetuned on both the Luganda and Swahili datasets for a total of 100 epochs. The experiments were first conducted without altering the model's configurations and thereafter the models' encoder layers were frozen for the first 50 epochs. This was done to allow the decoder to first pick up on patterns previously learned by the models when first trained prior to our experiments and after have the encoder and decoder train simultaneously for the remainder of the experiments.

The experiments made use of two pretrained conformer models, i.e. a Kinyarwanda mode(rw) from Nvidia and a multilingual Kinyarwanda, Swahili, and Luganda model(rw-sw-lu) from the Makerere study on radio speech. With each model finetuned both with and without freezing the encoder layer, there were a total of 8 experiments with 8 corresponding model checkpoints. The models' performance was evaluated based on both word and character error rates (Table 3).

The models appeared to perform better overall on the Swahili dataset with the multilingual model showing the best performance across the board. This, along with evidence from previous studies (Chanie et al., 2023), indicates that Swahili might be a relatively easier language to model than Luganda. It is important to note however that the difference in performance of the models is perhaps a result of the data collection and cleaning process since each language's corpus was acquired from different sources.

Observing the WER scores on the validation set (Figure 6), the multilingual model appears to start off at a much lower error rate compared to the Kinyarwanda model. This may be a direct result of the model learning the patterns not only within the language but also across several languages. Moreover, freezing the encoder layer seems to get the Kinyarwanda mod-

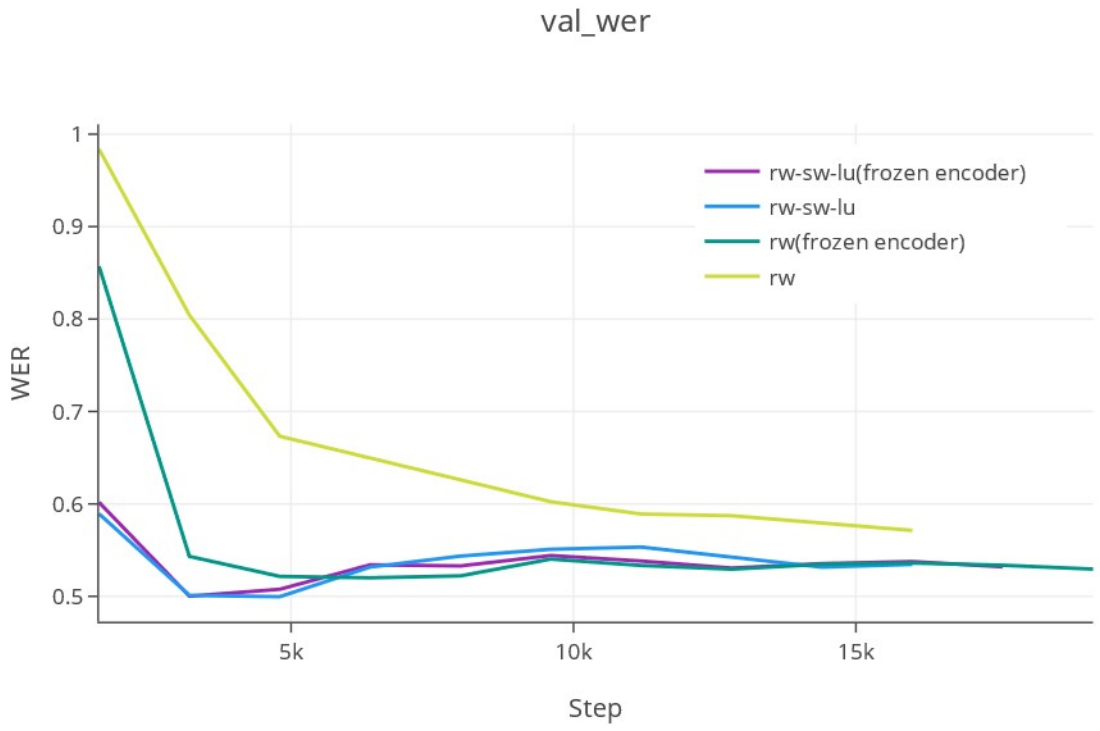

Figure 6 Validation WER

Table 4 Model performance compared to baseline scores (Source: Authors)

	Swahili		Luganda				
Model	Common Voice	ALFFA	Common Voice	Makerere1 (8.35 training hours)	Makerere2 (82.7 training hours)	Makerere3 (3x Makerere2)	FRI
rw	-	95.4	-	97.9	-	-	99.3
rw-sw-lu	17.22+	36.8	21.95+	63.6	-	-	86.4
Makerere	-	-	33+	-	65.1+	47+	
rw (finetune)	-	44.5*	-	54.3*	-	-	93.9
rw-sw-lu (finetune)	-	39.9*	-	51.5*	-	-	95.3
rw (enc-freeze + finetune)	-	45.4*	-	52.6*	-	-	94.2
rw-sw-lu (enc-freeze + finetune)	-	40.3*	-	52.1*	-	-	97.2

el to converge faster and ultimately achieve scores similar to the multilingual model.

Along with evaluating model performance across experiments, it was important to understand how the models compared to baseline results from previous research. To this end, WER scores from Mukiibi et al. (2022), Chanie et al., (2023), as well as Nvidia's experiments with the Kinyarwanda conformer model were used as baselines on which the performance of our models was evaluated.

Furthermore, model checkpoints from both Nvidia NGC and Chanie et al. (2023) were run on the Luganda radio speech corpus as well as the Swahili news broadcast data to give a better reference point on the performance of these ASR systems when applied to radio speech. This was done since the models were originally trained and evaluated on data from Mozilla's Common Voice initiative which is significantly different from radio speech (Mukiibi et al., 2022).

The results, along with scores from our own models have been included in Table 4. It is evident that the models perform better on the Common Voice data compared to radio speech. This is perhaps due to the fact that radio speech faces significantly higher levels of background noise as well as variation in recording equipment which tends to affect the ASR model's ability to recognized and distinguish utterances.

Our models have managed to achieve scores that are significantly lower, compared to those obtained in the original study on Luganda radio speech, using only a small fraction of the data from the research.

4.7. Gender Equality

The final principle, to bring the benefit of AI to farmers, particularly those under served by traditional innovations, will be the basis for the next phase of work. This will involve scaling the tool whilst resolving any remaining technical challenges and exploring the usability of the model with relevant stakeholders. Following the visit, several activities key to pushing the project forward were identified. These can be categorized into technical activities—those needed to ensure Longa is ready for deployment—and scaling activities, which are related to FRI's use of Longa to effectively impact farmers.

Technical Challenges

We are yet to develop a means of integrating Longa into the FRI workflow, both technically and strategically. We need to explore means of making insights actionable by FRI, considering which representations are most helpful to the MERL team and the radio station clients, and creating an interoperable frame-

Table 5 Comparison of Luganda audio recordings (Source: Authors)

Data Source	Maximum Duration (min)	Minimum Duration (min)	Maximum Characters	Minimum Characters	Maximum Words	Minimum Words
Makerere	15.96	3.00	292	16	46	3
FRI	59.74	14.58	922	9	161	1

Figure 7 Results of the CGIAR Scaling Scan on Longa (Source: Authors)

work for exchanging information between the two tools. We intend to propose and conduct a series of workshops in the summer of 2024 with various FRI stakeholders, to ascertain the appropriate design for the final interface.

A continuing major challenge might be the material infrastructure, that is, the equipment currently used to record and process the audio files. The quality of audio recording relative to background noise and the strength of the speaker signal may require further investment, alongside further research in audio preprocessing.

On average, as indicated in Table 5, FRI phone calls involve more words/utterances than the training set, with the average duration indicating more noise/silence, potentially impacting model inference. We have shown from transcription of a limited set of FRI calls, around 236 minutes, that there is an increase in performance. This will, however, need to be supplemented with a larger set of labelled audio.

Scaling

Scaling refers to maximizing the reach of a technology or practice, and thus the potential impact, in relation to the relevant economic, social and environmental systems (Woltering, 2019). It is an essential process to ensure down stream analysis of farmers' opinions to better understand principles underlying agricultural practices.

Figure 7 demonstrates a tool developed by CGIAR to evaluate the scalability of an innovation. We have utilized this method for planning the next year of work, considering the strengths and weaknesses, contributions and improvements to be made.

We consider Longa to be relatively easy to adopt, providing the usability of the tool is maximized relative to the technical expertise required. The tool is not disruptive but augments the existing workflow of the organization. The next stage of work will focus on the user interface, integrating the model with Uliza.

Throughout our interactions, FRI have been extremely receptive to the innovation, as have several other organizations which use audio in their extension processes. It has become the basis for sourcing further funding and a headline success for all stakeholders involved. It is yet unclear of the extent to which this degree of interest is shared throughout the different

FRI departments, and what value they can produce. We hope to include a diversity of staff members in workshopping final interface design in order to ensure a holistic approach to adoption.

There is a solid business case for Longa throughout the value chain, rendering the uptake and analysis of the data more efficient for FRI staff and their partner radio stations, whilst maximizing the value extracted from the data they collect. Farmers will then receive better services, leading to more sustainable farming practices.

The existing value chain is a key strength of FRI. During the visit, we experienced their rapport with their radio station partners, and the trust which this commands with farmer-users. Local networks are a significant determinant of scaling success. Where FRI has focused on the delivery of services to farmers, achieving great success with radio where issues with smart phone apps remain, it is hoped that Longa will contribute to greater internal synergy, ensuring that the value of user feedback is appropriately shared.

The AI in agriculture sector is currently a significant focus of funding and resources. It is a cutting edge and rapidly advancing innovation, with yet unseen impact to be achieved. As such, it is both a significant draw for donors, and there is a large community of developers who share expertise, models and data.

Communicating the value of the model and how it works internally can be challenging. While we expect Longa to work in the background for most users, we are striving to make our communication as engaging as possible, and to reduce the level of technical expertise necessary to make the most of Longa's capabilities. We suggest demonstrating the tool's capabilities by means of a user-friendly interface with access to each of the components in the ASR pipeline. This will allow users to try out the model building functionalities as well as how to deploy and incorporate the built models into an existing workflow by means of software applications.

There is little in the way of direction for adopting NLP in agricultural extension. Early in this project, evidence and learning were highlighted as key aspects to be addressed. By exploring the onboarding process, mapping out the process of research, design, and development of an audio analytics tool for an Agricultural NGO, we hope to share our experience such that future organizations in this field may do the same.

The governance structure for collecting and sharing research data is not particularly friendly for a project of this kind. It was noted that Tanzania has very strict regulations for posting data beyond the country's borders, using foreign server networks or drone piloting. We do, however, have the backing of large institutional group in CGIAR, with strong leadership and diverse technical expertise which should aid us in negotiating these next steps.

Bibliography

Albesano, D. (2022, October 5). Prediction Network Architecture in RNN-T for ASR. Retrieved from What's Next Blog: https://whatsnext.nuance.com/innovation-research/automatic-speech-recognition-on-prediction-network-architecture/

Chanie, Y., Elamin, M., Ewuzie, P., & Rutunda, S. (2023). Multilingual Automatic Speech Recognition for Kinyarwanda, Swahili, and Luganda: Advancing ASR is Select East African Languages.

F.P. Dossou, B., & Emezue, C. C. (2021). OkwuGbé: End-to-End Speech Recognition for Fon and Igbo.

Gauthier, E., Besacier, L., Voisin, S., Melese, M., & Elingui, U. P. (2016). Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., . . . Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition.

Jones-Garcia, E. (2022). Speech Recognition, Machine Translation, and Corpus Analysis for Identifying Farmer Demands and Targeting Digital Extension.

Koo, M.-W. (2021). A Korean speech recognition based on conformer. The Journal of The Acoustical Society of Korea. Retrieved from

This publication has been prepared as an output of **CGIAR Research Initiative on Digital Innovation**, which researches pathways to accelerate the transformation towards sustainable and inclusive agrifood systems by generating research-based evidence and innovative digital solutions. This publication has not been independently peer-reviewed. Any opinions expressed here belong to the author(s) and are not necessarily representative of or endorsed by CGIAR. In line with principles defined in CGIAR's Open and FAIR Data Assets Policy, this publication is available under a CC BY 4.0 license. © The copyright of this publication is held by IFPRI, in which the Initiative lead resides. We thank all funders who supported this research through their contributions to CGIAR Trust Fund.