



UNIVERSITY OF NAIROBI

FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND INFORMATICS

**A LINGALA AUTOMATIC SPEECH RECOGNITION
SYSTEM FOR RADIO STATIONS IN KINSHASA**

By

MBAYA MOLOLA HONORE

(P60/40490/2021)

SUPERVISED BY

DR. ENG. LAWRENCE MUCHEMI

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE, DEPARTMENT OF COMPUTER SCIENCE, UNIVERSITY OF NAIROBI

July 17, 2023

DECLARATION

This project is my original work and to the best of my knowledge this work has not been submitted for any other award in any University.



HONORE MBAYA MOLOLA

P60/40490/2021

Date: **July 17, 2023**

This project report has been submitted in partial fulfillment of the requirements of the Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University supervisor.



Dr. Eng. Lawrence Muchemi

Department of Computing and Informatics

Date: **July 17, 2023**

EPIGRAPH

Simplicity is the highest goal, achievable when you have overcome all difficulties.

Frederic Chopin

Imagination is more important than knowledge.

Albert Einstein

Neither a lofty degree of intelligence nor imagination nor both together go to the making of a genius. Love, love, love, that is the soul of a genius.

Wolfgang Amadeus Mozart

ACKNOWLEDGEMENTS

To God, the Only Wise, Immortal, Invisible be glory and praise for ever.

We would like to begin by expressing our sincere gratitude to Dr. Eng. Lawrence Muchemi (UoN), for agreeing to oversee the design and writing of this thesis. May he read this as a statement of our deep gratitude for his high-quality supervision.

We are grateful to M. Moturi for his indulgent availability and his willingness to help us follow meticulously the research constraints in accordance with the University Guidelines. May by him all the panel members receive our sincere acknowledgments.

Creating a labeled speech corpus is an enormous undertaking. This project would not have been possible without the assistance of many friends who generously helped us to transcribe the collected data: Fulgence Ntieni, Ruben Tongotani, Stanislas Kambashi, Joseph Baliki, Blaise Ikiniwewa. Moreover, we want to thank M. Desiré Kanyama for providing us with detail information about the Lingala program at Okapi radio.

We would like to appreciate the support from Dr. Philippe Nzoimbengene (UCL) who helped us to understand many linguistic aspects of the standard Lingala. His linguistic expertise was highly beneficial to us.

We acknowledge and thank the great support from Dr. Emmanuel Kalunga (UVQS) and Ebbie Dorcas (UoN) for their valuable insights. We also want to thank Allan Ggita for his helpful feedback on the manuscript.

We give thanks to all the Jesuit members of the Loyola Community for their concern, encouragement and kindness towards us. May through them, the whole Jesuit Province of Eastern Africa accept our profound gratitude.

Finally, we express a special gratitude to all our large network that goes from family to friends and acquaintances, classmates to UoN faculty members for their unforgettable contribution to our formation.

ABSTRACT

Considerable number of multi-lingual ASR systems supporting Lingala have been developed in recent years. However, most of them still perform poorly especially when applied to a specific application domain.

This study attempts to develop a Lingala Automatic Speech Recognition (ASR) System for broadcasting domain in Kinshasa. To this end, a 3 hours Lingala speech corpus was created using publicly available radio audio archives. We ran several experiments on the created corpus to train ASR models using the traditional supervised ASR modeling approach and two of the current state-of-the art pretrained modeling techniques, whisper(Radford et al., 2022) and the Massive Multilingual Speech (MMS)(Pratap et al., 2023) models. The best classical model yielded 55% of WER while the whisper tiny and the MMS finetuned models output 43% WER and 31% WER respectively. The final model achieved 25 % WER after fine-tuning the whisper base checkpoint on a mixed dataset resulting from combining our custom corpus with the Google's fleurs dataset. This final model was integrated as backend engine to a Lingala ASR web transcription prototype platform. Despite the promising results obtained, the ASR model performance needs to be improved by first applying further data quality check and normalization steps, and then adding more data from diverse sources in the target domain. This project has confirmed fine-tuning of existing ASR pretrained models as the best approach to create Lingala ASR system for broadcasting domain.

We make four core contributions. First, the construction of a domain specific Lingala speech dataset that will foster further speech translation research in similar context. Second, the release of a replicable pipeline for the creation of speech corpus from existing audio news and broadcasts from other Radio Stations in Kinshasa. Third, a baseline Lingala ASR model for broadcasting that can serve as a starting point for further research in the same domain. Fourth, a transcription platform prototype to encourage Lingala document preservation.

TABLE OF CONTENTS

DECLARATION	ii
EPIGRAPH.....	iii
ACKNOWLEDGEMENTS.....	iv
ABSTRACT.....	v
LIST OF FIGURES	viii
LIST OF TABLES.....	ix
LIST OF ABBREVIATIONS.....	x
DEFINITION OF TERMS	xi
CHAPTER 1: INTRODUCTION.....	12
1.1 Background.....	12
1.2 Problem statement.....	14
1.3 Objectives	15
1.4 Research questions.....	15
1.5 Scope and assumptions	15
1.6 Significance.....	15
CHAPTER 2: LITERATURE REVIEW	17
2.1 Introduction.....	17
2.2 Some state-of-the art Multilingual ASR models.....	17
2.3 Multilingual speech dataset.....	18
2.4 Existing Lingala speech corpora.....	19
2.5 Research Gaps.....	19
2.6 Conceptual framework.....	20
2.7 Conclusion	22
CHAPTER 3: RESEARCH METHODOLOGY	23
3.1 Introduction.....	23
3.2 Research design	23
3.3 Proof of concept.....	24
3.4 Methodology	25
3.4.1 Data collection	25
3.4.2 Preprocessing and Feature extraction	28
3.4.3 ASR Models.....	28
3.4.4 Evaluation metrics	30
3.4.5 Deployment.....	31
3.5 Experimental Process Model	32
3.6 Conclusion	33

CHAPTER 4: IMPLEMENTATION, RESULTS AND DISCUSSION.....	34
4.1 Introduction.....	34
4.2 Artifact development	34
4.2.1 Feasibility analysis.....	34
4.2.2 Requirements analysis	36
4.2.3 System users.....	36
4.2.4 System design	37
4.3 ASR Experiments.....	37
4.3.1 Tools used for Implementation	37
4.3.2 ASR Models.....	38
4.4 Results.....	40
4.4.1 Cascade ASR	40
4.4.2 Whisper fine-tuning	42
4.4.3 MMS fine-tuning	43
4.5 Interpretation and Discussion	44
4.5.1 Strengths of the model	44
4.5.2 Limitations	47
4.5.3 Ways for improvement	48
4.6 Conclusion	49
CHAPTER 5: CONCLUSION AND RECOMMENDATIONS	50
5.1 Summary of findings.....	50
5.2 Conclusion	51
5.3 Limitations	52
5.4 Recommendations.....	53
REFERENCES	54
APPENDICES	58

LIST OF FIGURES

Figure 1: Conceptual Framework	21
Figure 2: Research Design	24
Figure 3: Speaker gender distribution.....	27
Figure 4: Classic ASR architecture.....	29
Figure 5: E2E Transformer-based ASR architecture	30
Figure 6: Classic ASR Process model	32
Figure 7: System block diagram	37
Figure 8: Corpus vs WER trend for classic ASR.....	42

LIST OF TABLES

Table 1: Corpus size by Gender.....	26
Table 2: Mixed dataset.....	27
Table 3: Whisper pseudo-code.....	32
Table 4: MMs pseudo-code	33
Table 5: Excerpt of the Lingala lexicon.....	39
Table 6: Classic ASR results.....	41
Table 7: Whisper tiny performance	43
Table 8: Whisper base performance	43
Table 9: MMS performance.....	43
Table 10: Reference transcript vs classic ASR predicted transcript.....	45
Table 11: Reference transcript vs Whisper base predicted transcript.....	46

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech recognition
CNNs	Convolutional Neural Networks
CTC	Connectionist Temporal Classification
DNN	Deep Neural Network
DRC	Democratic Republic of Congo
E2E	End-to-end
HMMs	Hidden Markov Models
MFCCs	Mel-frequency cepstral coefficients
MuST-C	Multilingual Speech Translation Corpus
NLP	Natural Language Processing
RNNs	Recurrent Neural networks
SSL	Semi-Supervised Learning
TTS	Text-to-speech
WER	Word Error Rate

DEFINITION OF TERMS

Artificial Intelligence (AI) – Refers to the creation and use of computer systems or machines that are capable of doing activities that traditionally require human intelligence. It entails modeling intelligent activity, including perception, learning, problem-solving, and decision-making.

Acoustic Model – An acoustic model is a part of an Automatic Speech Recognition (ASR) system. It illustrates the link between phonetic units (such as phonemes or sub-phonetic units) and audio features (such as spectral information).

Frequency spectrum – Also known as Frequency domain representation, it describes the individual frequencies that make up the signal and how strong they are.

Hidden Markov Model (HMM) – It is a statistical tool for modeling sequential data, including speech and text. It is based on the idea of a Markov process, where the system is thought to exist in a set of states but is hidden from direct observation. Only a series of observations that were emitted from the hidden states are seen instead.

Language Model (LM) - A language model is a computational model that seeks to predict the likelihood of word or character sequences in a particular language. In order to produce writing that is coherent and meaningful, it captures the statistical dependencies and patterns in a language.

Mel-spectrogram – Commonly used in Speech Recognition and Machine Learning, the mel scale is a perceptual scale that approximates the non-linear frequency response of the human ear.

Phoneme – Refers to the smallest distinguishable unit of sound in a language. It serves as the basic building block of spoken language and symbolizes the distinct sounds that make up words.

Sampling Rate – Number of samples (measurements or data points) taken from an analog signal to transform per second. It is frequently measured in Hertz (Hz) and denotes the sampling frequency.

Spectrogram – Plots the frequency content of an audio signal over time, allowing visual representation of time, frequency and amplitude all on one graph.

Utterance – An utterance refers to a unit of speech or written text produced by a speaker or writer in a given context. It can be a complete sentence, a phrase, or even a single word.

CHAPTER 1: INTRODUCTION

1.1 Background

The field of Natural Language Processing (NLP) exploits two main modalities to make computers able to understand and process human languages: text and speech (Jurafsky & Martin, 2020). From these two modalities derives a wide range of subfields that consider the NLP problem from a different perspective but with the same end goal. Automatic Speech Recognition (ASR), also known as Speech-to-text (STT), is one of the subdisciplines of NLP focused on speech modality with the objective of accurately converting spoken language into written text or executing commands based on spoken instructions. It has attracted significant attention and made great advancements in recent years. Some application examples of ASR technology include Dictation Systems, Voice Assistants, transcription services, call center automation, and many more in various domains.

The history of ASR system is broadly characterized by a progressive recognition capability in terms of vocabulary size. It goes from a purely frequency detector able to detect only one word to large vocabulary continuous Speech Recognition with Hidden Markov Models first, and then using Deep Neural Network based approaches.

This journey has shaped the whole speech recognition process which involves multiple stages. Initially, a microphone or other audio recording devices are used to record the voice signal. The next step is preprocessing the recorded audio, which might include eliminating background noise, adjusting the loudness, and breaking up the speech into smaller chunks.

The process of extracting pertinent acoustic elements from the spoken signal known as feature extraction comes next. These features record details like pitch, intensity, and spectral information that are then utilized to describe the voice signal in a way that machine learning algorithms can understand.

Different algorithms and models are used by speech recognition systems to convert the extracted features into text. Currently, research in speech is marked by a technological shift in terms of ASR methods : the switch from conventional methods like Hidden Markov Models (HMMs) based on 3 sub-models architecture to end-to-end (E2E) modeling, which directly

converts the input speech sequence into the output token sequence (characters or even words) using a single network (Li, 2022).

During the training phase, a big dataset of matched speech and related transcriptions is fed to the speech recognition system. Between the audio features and the accompanying textual representations, the algorithm learns statistical patterns and correlations. The machine can properly detect and record speech thanks to this training. By comparing the predicted transcribed text to the original one, the system is evaluated and fine-tuned using evaluation metrics like word error rate (WER) or accuracy. To raise the system's performance, iterative methods of refinement and optimization are used.

Several sources of variability make the ASR task a difficult problem. In fact, an ideal system should be able to deal with differences in accents, pronunciation, and speech patterns, manage loud settings, and effectively transcribing technical jargon or domain-specific words.(Jurafsky & Martin, 2020). These challenges highlight the critical role of quality dataset for training good ASR systems. Given the cost related to getting more data and/or expertise, less data-intensive methods are being used as workaround, depending on the task, to solve the data constraint. Examples include the Semi-supervised learning (SSL)(Zhang et al., 2022), which uses unlabeled data to improve the performance of labeled tasks, Few-shot learning which involves training a model using one example per class (k-shot)(Roger, Farinas, & Piquier, 2022), and Transfer learning, which uses a pre-trained model either as a feature extractor or as a parameter initializer. Another promising research direction consists of relying on large amounts of untranscribed speech-text with only a small amount of paired audio-text to train a large single universal ASR model(Zhang et al., 2023).

However, since up to now these less data-intensive strategies still rely on labelled data, they cannot completely replace the need for transcribed audio data which remain a key requirement for ASR systems. This explains why considerable efforts are being made to find efficient ways and strategies to construct quality datasets. In this work, we follow the same research direction of addressing the data scarcity in low-resource settings. Specifically, we constructed a Lingala speech corpus which allowed us to develop an ASR system useful for many applications, in principle, with radio or television news transcription as the primary use case.

1.2 Problem statement

Over the last ten years, speech technology has advanced significantly and been incorporated into a wide range of business applications including smartphones, home assistants, voice-controlled devices, dictation software, machine transcribers, and live caption generators (Roger et al., 2022).

Despite these advances, current ASR systems are far from being perfect in terms of both accuracy and language coverage. In fact, while using these systems, we must have noticed that they are not always accurate under non-ideal conditions similar to the data on which they were trained. Oftentimes, noisy environments, diverse languages, accents, and background noises impair the performance of ASR applications. We also require various kinds of speech datasets based on the speech recognition model's use case. And in relation to language coverage, many of the 7000 languages spoken over the world are either not yet supported or, if supported, the ASR systems perform poorly on them (Pratap et al., 2023). For example, one of the largest start-of-the-art general purpose ASR models that supports Lingala, one of over 200 native languages spoken in the DRC (Palma, 2022), achieves only 75.6 % of WER as of now (Radford et al., 2022).

Consequences of such poor language coverage include accelerating the disappearing of endangered languages as well as preventing an alternative natural interface for illiterate users. In the context of DRC, a robust ASR system can foster the use of speech technology in many application domains. A typical application need is that of Radio stations in Kinshasa. Not only do most of them lack a robust system to organize and archive their daily data, but nearly none of them is able to properly document and archive news data in local languages. For instance, for one Radio station called Okapi, audio and text of news in French are easy to find. But for news in local languages such as Lingala, one rarely finds a corresponding text to many of the audio available. Yet, the use of a robust ASR could well facilitate the archiving and digitization of documents.

Data scarcity is clearly one of the issues that impact the performance of ASR system. Specifically, the data required by ASR systems needs to be, ideally, in the form of paired speech-transcription which is even difficult to find in the context of poor document digitization as it is still the case for the DRC. In this work, we face this challenge of dataset to improve the performance of ASR system for the Lingala language.

1.3 Objectives

The primary goal of this research is to build a Lingala ASR system for Radio stations in Kinshasa. To this end, we need:

- 1) To create a Lingala speech corpus from publicly available radio news and broadcast audio data.
- 2) To experiment on which ASR modeling approach can yield promising performance results on the created corpus.
- 3) To develop a transcription platform prototype that uses the trained ASR model as a backend engine.

1.4 Research questions

The main driving question of this research is: How can we improve the performance of the ASR system for the standard Lingala spoken in Kinshasa?

Specifically, we investigate the following:

- a) How can we construct a reliable speech corpus from radio archives data?
- b) What are the optimal modeling architectures and training strategies that best fit our case study and dataset?

1.5 Scope and assumptions

This research is restrained to the standard version of Lingala spoken in the DRC. The standard version is the one used by official institutions and media communication. It has the particularity of bridging the gap between the formal Lingala and the daily Lingala spoken by common citizens in Kinshasa.

1.6 Significance

The primary beneficiaries of this project are radio stations since the ASR is trained on the data from this domain. The proposed transcription platform will highly assist Lingala journalists in their daily tasks. But other use case applications like in business are still possible given the relatively diverse topic range that characterizes radio news data. Extensively, with such ASR

system machine translation systems can be used, for instance, to allow Lingala native speakers with poor literacy levels to communicate easily and access various services. Moreover, given that Lingala is an international language, the impact of the proposed system can empower business activities across countries that use Lingala.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter reviews and synthesizes existing literature in relation to ASR system for the Lingala language. Our end goal is to develop a Lingala ASR system able to provide transcription service for various use cases with the transcription of radio news recording being the target application. Because, as language, Lingala has many variants depending on where it is spoken, this work considers only the standard version of Lingala spoken in Kinshasa and often used by Radio Stations. The following theme are reviewed and summarized before establishing the gap and adopting a conceptual framework. We first examine some state-of-the-art multilingual ASR models and then review existing Lingala speech corpora with a special attention on those intended for ASR task.

2.2 Some state-of-the-art Multilingual ASR models

Two broad categories of learning strategies are currently shaping state-of-the-art multilingual ASR systems: self-supervised pre-training and weakly-supervised pre-training. The rationale behind the self-supervised approach is to mitigate the constraint of large human labelled audio data as requirement for building robust ASR system by pre-training the model on a huge amounts of unlabeled audio data (Conneau, Bapna, et al., 2022). Pre-trained on a new dataset built from readings of freely available religious texts, the Massively Multilingual Speech (MMS) model (Pratap et al., 2023) yield the character error rate (CER) of 4.0% and 4.3% respectively on the Lingala dev and test sets of the Fleurs dataset after fine-tuning on the training set of the same corpus. Despite these promising results, the model still needs to be fine-tuned on a domain specific dataset for it to be deployed in a production solution.

Using the same self-supervised pre-training approach, (Kimanuka, Maina, & Buy, 2023) managed to obtain a WER of 21.4% with the CdWav2Vec multilingual model pretrained on the Congolese speech radio corpus and subsequently fine-tuned using the Lingala Read corpus. Again, this model cannot be used as it is for our application use case.

The Google's USM model (Zhang et al., 2023) is pre-trained on 12 million hours of exclusive YouTube audio in 300 different languages, and it has been fine-tuned to perform ASR for up to 100 languages on a labeled dataset of 90 thousand hours, outperforming Whisper

significantly. Results from this experiment confirmed the importance of in-domain data as the most effective way to improve the performance of ASR system for a given domain.

Whisper (Radford et al., 2022) is the current outstanding multilingual ASR pretrained on 680K hours of weakly supervised audio in 99 languages. Based on a sequence-to-sequence architecture Whisper model achieves great performance at a very large-scale. However, one of the limitations of the whisper model is its poor performance on many low-resource languages.

2.3 Multilingual speech dataset

Building speech-to-text corpora has been extensively studied in the literature. To mention some few recent investigations that inspire our corpus creation pipeline, in the area of Speech translation, (Iranzo-Sánchez et al., 2020) built a multilingual Spoken Language Translation (SLT) corpus, which contains paired audio-text examples from and into 6 European languages for a total of 30 possible translation directions. The corpus was developed using publicly accessible videos from debates in the European Parliament. In the same line, (Cattoni, Di Gangi, Bentivogli, Negri, & Turchi, 2021) used English TED Talks as the foundation to provide a sizable and openly accessible Multilingual Speech Translation Corpus (MuST-C). Their corpus creation pipeline includes the following steps: Download data, segmentation, and text-level alignment; Audio to text alignment; filtering; feature extraction. To meet the necessity for a specialized parallel resource required by the current state-of-the-art methods in Speech translation, (Salesky et al., 2021) created a Multilingual TEDx corpus as a way of promoting speech recognition and speech translation research across multiple languages.

In the context of low-resource African languages, (Doubouya, Einstein, & Piech, 2021) were able to publish the West African Radio dataset, a corpus of 17,091 audio clips, each lasting 30 seconds, sampled from archives gathered from six Guinean radio stations. Using this corpus, they test the effectiveness of unsupervised speech representation learning for downstream tasks aimed at West African languages. While the leverage of radio broadcasting archives as a source of speech corpus is similar to our strategies, our target language and the corpus creation process makes a big demarcation from their work.

2.4 Existing Lingala speech corpora

Literature on speech corpora for Congolese languages, especially for Lingala is progressively growing for various downstream tasks. African voices(Ogayo, Neubig, & Black, 2022) is a project that attempted to create Text-to Speech dataset for 12 low-resource African languages, including Lingala. In addition to the fact of relying only on the Bible as data source for Lingala, this dataset is rather suitable for speech synthesis task which is not our focus.

LisTRa (Kabenamualu, Marivate, & Kamper, 2022) is another effort that proposes an English to Lingala Automatic Speech Translation dataset using the Bible as data source. FLEURS (Conneau, Ma, et al., 2022) is an n-way parallel speech dataset in 102 languages, including Lingala, with around 12 hours of voice supervision per language, built on top of the machine translation FLoRes-101 benchmark. This dataset can be used for different speech tasks, including Speech recognition. We rely on it to increase the size of our corpus.

Recently, (Kimanuka et al., 2023)made available to the research community two new datasets: the Congolese Speech Radio Corpus, which contains 741 hours of unlabeled audio in four of the main spoken languages in the Democratic Republic of the Congo, and the Lingala Read Speech Corpus, which consists of 4 hours of tagged audio clips. While the created Lingala Read Speech dataset is reliable, it does not cover as much subjects as needed for the media use case. In order to be able to capture the context and order of words in Lingala, a corpus of semantic and syntactic questions has been manually constructed by (Maniamfu, Kiketa, Muepu, & Kabongo, n.d.). The performance of the trained language model is also promising, but the relatively small size and poor domain coverage of the constructed datasets limits its applicability to more robust systems.

2.5 Research Gaps

From the above, it is clear that the performance of existing Speech recognition system is still poor for most of the low-resource languages like Lingala. Furthermore, despite considerable efforts of supporting as many languages as possible, there will always be a need for domain specific ASR system, since general purpose or multidomain systems are currently far from being a reality in the AI field.

Regarding the dataset constraint, the domains covered by the existing Lingala speech corpora, namely the literature and the religious domains, are still narrow plus the fact that some of the Lingala corpora are not directly appropriate for the ASR task. This situation leaves the ground

open for any contribution with specific data to any application domain not yet covered as is the case for news broadcasting. And from the Linguistic perspective, digital Lingala ingredients toward tools like automated dictionary or spell checker remain insufficient.

Finally, since ASR systems are useful only when deployed and integrated into a practical user application, the need of ASR based tools such Transcription platform is acute and can be a way of boosting language preservation through document conservation.

2.6 Conceptual framework

The interaction of the keys elements for designing, developing and evaluating ASR systems is as follows.

The Input Speech recorded using microphones or other audio capture hardware is passed to a Preprocessing component involving a variety of preparation steps such as noise removal, speech segmentation, for further analysis. The Feature extraction component converts the preprocessed speech signal into a usable representation like Mel-frequency cepstral coefficients (MFCCs), spectral properties, or pitch information important for analysis (Jurafsky & Martin, 2020).

The Acoustic modeling unit uses Statistical methods like hidden Markov models (HMMs) or more recent Deep-learning based methods to simulate how the retrieved acoustic features relate to the relevant phonetic or language units. To enable the system to produce more precise and contextually appropriate transcriptions, the language modeling component is often added to capture the statistical characteristics of real language. Recurrent neural networks (RNNs), transformers, n-gram models, and other methods that record word or phrase probabilities can all be used as the foundation for language models (Latif et al., 2023).

The Decoding process that follows after consists of aligning the acoustic features with the associated language units, such as phonemes, words, or sub-word units. For this alignment, the most likely transcription of the input speech is determined using a variety of methods, including dynamic temporal warping (DTW) and beam search. The Post-processing step entails editing the output transcription to enhance its coherence and readability. Applying language-specific norms, grammar and spell checks, or post-editing by human annotators are a few examples of this.

The evaluation step assesses and contrasts the output transcription with the original transcription by using metrics like word error rate (WER) or accuracy. This step aids in assessing the effectiveness of the system and locating potential areas of improvement.

Once evaluated, the speech recognition system must be integrated into the programs or services that will use it as the last stage. Voice assistants, transcription services, contact center automation, voice-controlled devices, and other tools fall under this category. Considerations for the integration include the scalability, user interfaces and system requirements.

Continuous learning and modification can help speech recognition systems perform better over time. To improve accuracy and personalization, this may entail updating language models, retraining acoustic models with fresh information, or implementing user-specific adaption strategies.

Figure 2-1 below summarizes the workflow of a typical ASR system

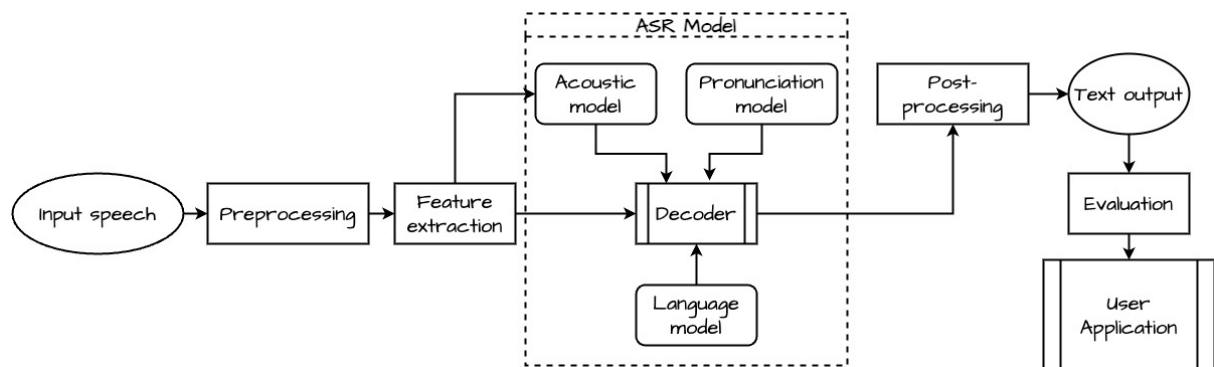


Figure 2-1: *Conceptual Process Model*

2.7 Conclusion

In this chapter we reviewed existing literature that is closely related to our research objectives. We found that there is still room for improvement with regards to the current performance of existing ASR pre-trained models on low-resource languages such as Lingala. More domain specific speech corpora need to be created in order to build performant specialized ASR systems. In the case of DRC, practical ASR tools are also needed to boost language preservation through document conservation.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter is devoted to the methodological prerequisites that mark out the path. The data collection process, the data preparation and analysis are exposed here. The chapter is split into two major parts. The first part presents the research design which provide a mind-map and rationale for carrying out the research. The second part dives deep into the specific speech recognition methods including techniques used for data collection, data preparation, Model training and evaluation.

3.2 Research design

The purpose of this study is to develop and evaluate a Lingala Speech Recognition System for Radio Stations in Kinshasa. The overall study is a mixture of qualitative and quantitative research methodology with prototyping as proof of Concept. We started by reviewing current multilingual ASR systems and existing Lingala speech datasets to better identify the gaps. Two strategies are used for data collection. First, we created a custom dataset suitable for the broadcast domain, and then we mixed it with existing dataset to increase the corpus size. The ASR modelling step was largely experimental. We explored conventional models like Hidden Markov Models (HMMs) as well as deep learning-based models, such as Transformers using transfer learning strategy. The standard Word Error Rate (WER) metric was used to evaluate the speech recognition system's performance. This metric measures the accuracy and quality of the transcriptions produced by the system. Additionally, qualitative evaluations are conducted by comparing visually the predicted transcription with the reference transcription annotated by human.

The diagram below illustrates and summarizes the research road-map adopted.

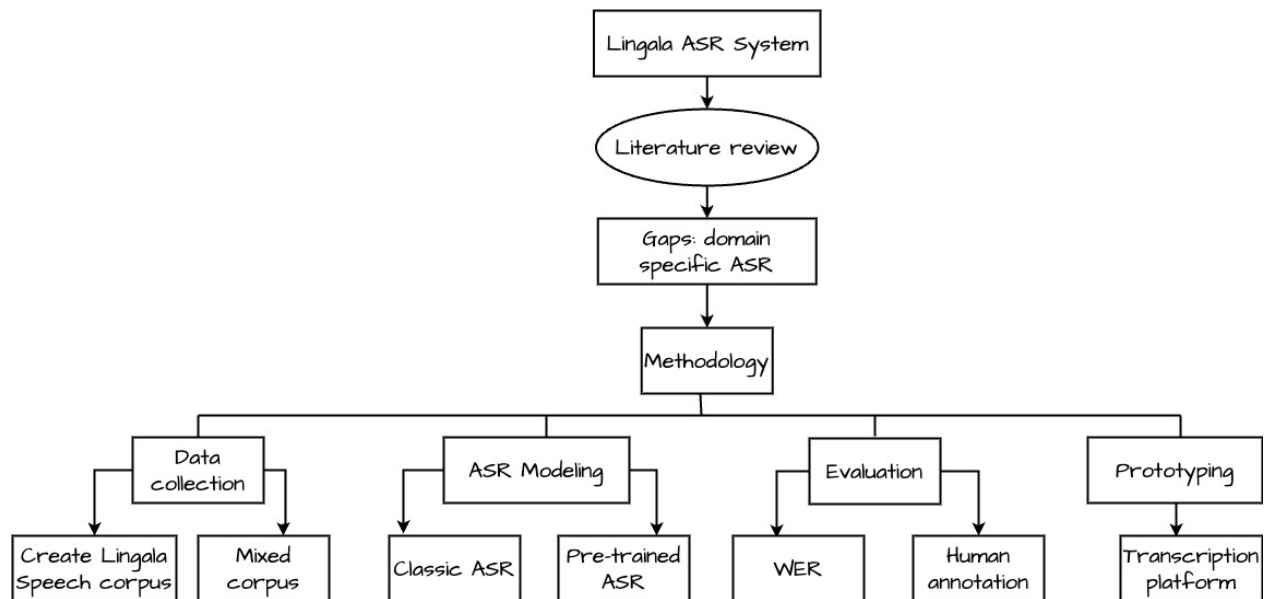


Figure 3-2: *Research Design*

3.3 Proof of concept

The creation and testing of a prototype that can automatically transcribe Lingala radio broadcast recordings served as proof of concept for this project. The chosen sample of speech data used to calibrate and assess the performance of the ASR models was transcribed by Lingala native speakers. The outcomes of training and testing several ASR models using the transcribed data sets were then documented and examined. The Cross Industry Standard Process (CRISP) model, which has six sections, was employed in this study to build the ASR system.

3.4 Methodology

3.4.1 Data collection

Data availability remains a big challenge to build robust Speech recognition systems (Sambasivan et al., 2021). That is why several data curation strategies have been developed in the research community. Inspired by the works presented in the literature section, our corpus creation process is as follows.

a) Data sources

The main source of our data was the Lingala non-transcribed audio archives provided by the Okapi Radio, one of the famous and most listened-to radio stations in the DRC under the auspices of the United Nations Organization. Each audio file is about 15 minutes long on average and contains news on a variety of topics. The audio files are also characterized by background noise and music at the beginning and toward the end of the recording. Multiple adult speakers, both male et female, intervene in general in each audio.

Ideally, the dataset required for the ASR task needs to be in the form of Audio-transcripts pairs. To guarantee the performance and the robustness of the ASR system, the dataset should be diverse in terms of audio recording environment as well as in terms of speaker demographic aspects. Given the domain coverage limitations of current Lingala speech corpus available, we opted for a manual data transcription process by relying on Lingala native speakers to create a labelled Lingala corpus. The corpus creation and the transcription processes are detailed in the following sections.

b) Lingala Corpus creation

A key step in the creation of the needed corpus was the transcription process. We followed the process proposed by (Awino et al., 2022) by adapting it to the specificity of the Lingala and the situation of the audio data available. We used a single transcription process in which audio files are given to many Lingala transcribers for transcription. Given the size of the corpus and the small number of transcribers, this strategy was practical. We used an intelligent transcription approach that involves some degree of editing to enhance readability and clarity. In fact, for some use cases like media, interviews, commercial or legal papers, and other forms

of content requiring a high level of accuracy and professionalism, this type of transcription is frequently utilized.

The overall corpus creation pipeline consists of the following steps.

Step 1: Scraping (downloading) the audio files from the Radio website. The period from 21/11/2022 to 01/05/2023 was selected.

Step 2: Data preparation. This comprises noise removal and converting to wav format. we removed the beginning (about 1 minutes) and the ending audio sections, which contain generic introductory (outlines) and closing remarks of a given broadcast.

Step 3: Segment audio into manageable short chunks. Since the average length of audio received was 13 seconds, we had to clip the files into small chunks based on 250 milliseconds silence interval. The range from 5-35 seconds was adopted as a good compromise between technical and intra-linguistical considerations.

Step 4: Renaming the files using a consistent pattern

Step 5: Creating a paired csv file to each audio file for transcription

Step 6: Transcription

The following transcription guidelines were explained and forwarded to 10 Lingala native speakers to whom we assigned 30 minutes of audio to transcribe:

- 1) Transcribe exactly only intelligible speech from a given audio
- 2) Comply with standard (official) Lingala writing rules.
- 3) Ignore words that would be cut off at the beginning or end of the audio.
- 4) Use punctuation wherever possible.
- 5) Ignoring character accents in Lingala
- 6) In the event of code switching and/or code mixing, i.e., switching to another audible language (very often French), also transcribe said language exactly.

Up to now, we managed to obtain 3 hours of transcribed audio corpus. Table 3-1 below shows the gender distributions of the Lingala speech corpus.

Table 3-1: *Corpus size by Gender*

Corpus version	Total size	Female size	Male size	Mixed size
1	3 hours	160.8 minutes	21.3 minutes	4.6 minutes

Figure 3-3 illustrates the gender distribution of the final dataset

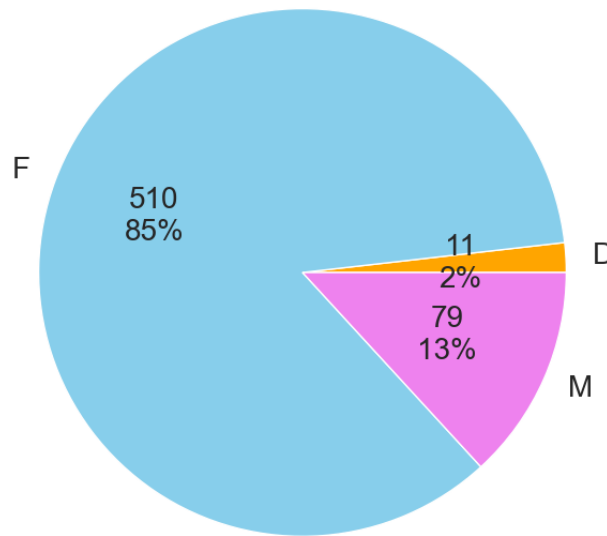


Figure 3-3: *Speaker gender distribution*

c) Supplementary dataset

To increase the dataset size, we explored mixing the created corpus with the Google’s Fleurs dataset which is a speech recognition evaluation dataset covering 102 languages, including Lingala (Conneau, Ma, et al., 2022). It comes from the FLoRes-101 dataset, a corpus of 3001 machine translations of sentences from English to 101 additional languages. The sentence transcriptions are narrated by native speakers recorded in the original languages. Each source-target language combination in the training sets has 10 hours or more of supervised audio transcription data. As shown in the Whisper research, such an increase in training data can have a big impact on performance later on.

Thus, we increased the dataset to 15 hours of transcribed audio in total. Table 3-2 below show how we split the combined dataset.

Table 3-2: *Mixed dataset*

Dataset	Training	Validation
Custom	2 hours	1 hour
Fleurs	10 hours	2 hours
Total	12 hours	4 hours

3.4.2 Preprocessing and Feature extraction

Regardless of the modelling techniques used, the preprocessing step for audio data often include the following tasks among others: removing silence from some audio, discarding audio without any speech, filtering audio based on their length. The input speech is also often normalized to 16kHz sampling rate and Mono channel to meet the requirement of many ASR models. Once prepared, the features such as *Log-mel* Spectrogram or the Mel-frequency cepstral coefficients (MFCCs), are then extracted and passed to the ASR model.

3.4.3 ASR Models

This study uses a broad Supervised learning paradigm for Training ASR models. Specifically, we used the traditional statistical ASR approach and current state-of-the art pretraining techniques.

1) Classical ASR

A typical ASR system has three main components, the acoustic model, the pronunciation model and the language model. The role of the acoustic model is to convert acoustic features from speech waveform into phonemes. Hidden Markov Models are one of the paradigms used to learn the mapping from the acoustic features to phonemes with the probability distribution being a mixture of Gaussian. In the so-called hybrid models, a DNN is used to replace the Gaussian mixture models for the assessment of the acoustic likelihood. The DNN-hidden Markov model is a popular and effective hybrid model (DNN-HMM) example which combines deterministic models (like DNNs) with probabilistic ones (like HMMs). Hybrid models are used primarily because they are optimized for production (Li, 2022), despite the fact that they perform poorly in terms of accuracy.

The pronunciation model which provides rules for mapping words to their corresponding phonemes. As for the language model component, it assigns probability to words occurring together in a large text corpus. It also helps to disambiguate between similar acoustics.

The architecture of a classical ASR is illustrated in Figure 3-4 below

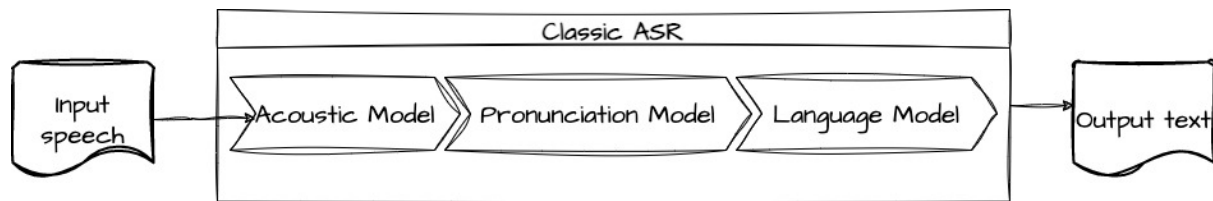


Figure 3-4: *Classic ASR architecture*

2) Pre-trained modeling for ASR

Pretraining refers to some learning strategies preconized for dealing with limited amount of labelled data in Supervised Machine Learning settings (Raschka, n.d.). Some of these strategies used in this work are:

Transfer learning. It is a deep learning technique that consists of pre-training a model on a large general labelled dataset. The pretrained model is then used to train (fine-tune) a final model on a small domain-specific labelled dataset.

Self-supervised learning. Also known as unsupervised pretraining, this technique is similar to transfer learning but with the particularity that the labels are automatically extracted from unlabeled data.

Weakly supervised learning. This technique leverages an external label source to generate labels for an unlabeled dataset.

Regarding the architecture, there are two broad categories of ASR pre-trained models:

The Connectionist Temporal Classification (CTC) Models which use only the encoder part of the transformer architecture with a linear classification head on top. MMS (Pratap et al., 2023) is one of such models that we attempt to fine-tune in this study. The second category is that of Sequence-to-sequence Models. These are encoder-decoder models with a cross-attention mechanism in between. Whisper (Radford et al., 2022) is an example of such pretrained model we finetuned. Both these categories fall under the End-to-end ASR paradigm.

The End-to-end (E2E) modeling uses a single network to directly translate an input speech sequence into an output token sequence. With this approach, all the modeling component in the traditional ASR systems are no longer required. Among the several advantages of these models are the simplification of the ASR pipeline and good accuracy performance in most benchmarks (Li, 2022). In this study we used one of the popular end-to-end techniques called

Attention-based Encoder Decoder implemented as Transformer to ensure the model captures long-term dependencies.

An encoder-decoder speech recognizer receives a sequence of acoustic feature vectors as input. Then these inputs go through a compression stage before the encoder decoder stage. The output can be letters or words.

Figure 3-5 below show the architecture of the E2E model used in this study.

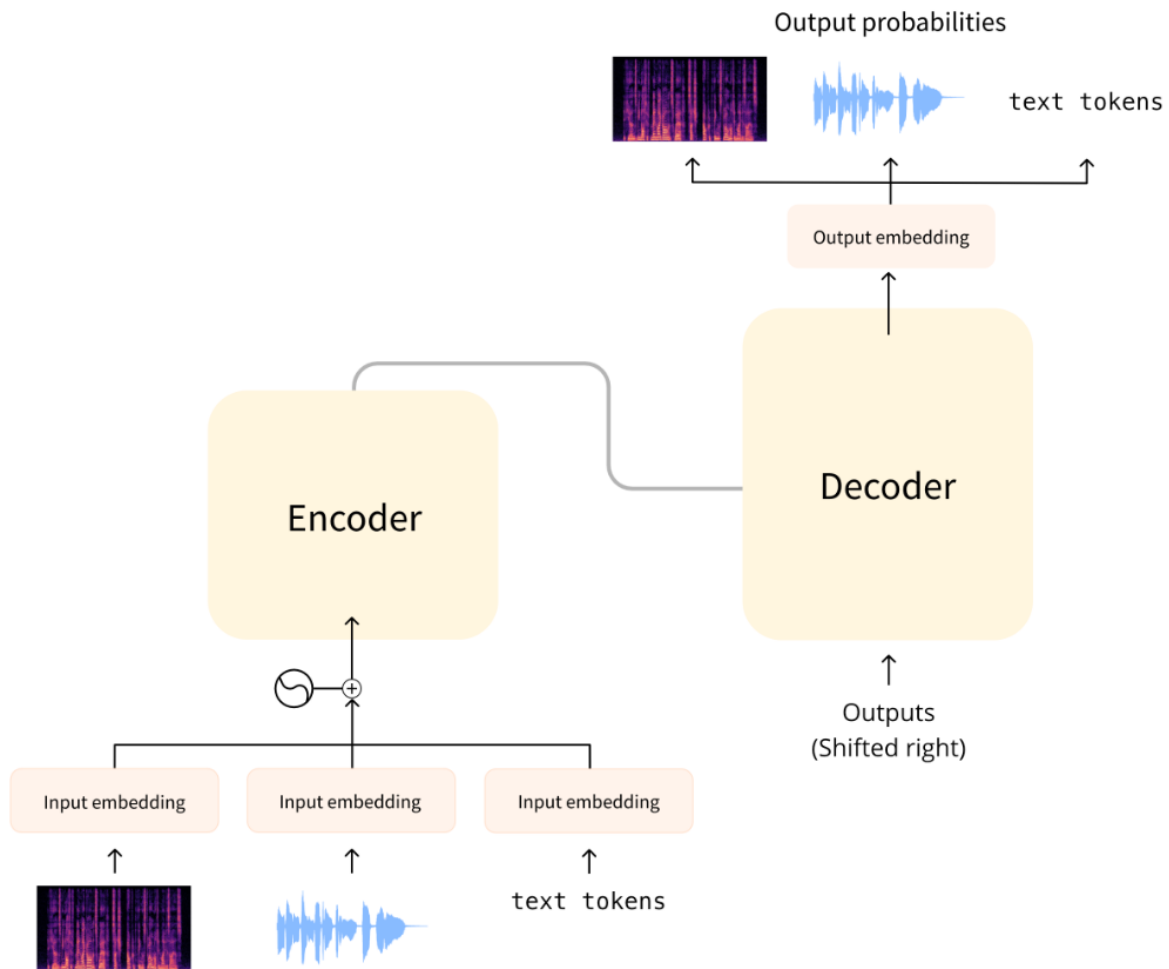


Figure 3-5: *E2E Transformer-based ASR architecture*

3.4.4 Evaluation metrics

We performed the evaluation of the trained models for both modeling techniques using qualitative and quantitative approaches. Qualitatively, we relied on feedback from Lingala native speakers to get an indication of users' confidence in the trained models. Regardless of accuracy performance, the most important question should be to assess the usability of the

transcript for the task at hand. A good ASR system should be able to recognize important words for the analysis of the task we are interested in (Ritchie et al., 2022).

The quantitative evaluation was based on the Word Error Rate (WER) which is the standard metric used to assess the performance of ASR model on a different test dataset. It compares the accuracy of the transcript produced by the system to the original transcripts annotated by humans. Specifically, the word error rate (WER) reveals the number of words that the system transcribed incorrectly. The errors are classified into one of these three categories:

Substitution (S): number of words wrongly transcribed in the prediction.

Insertion (I): number of extra words added in the prediction.

Deletion (D): number of words removed in the prediction.

Hence the formula

$$WER = \frac{S + I + D}{N}$$

Where N = total number of words in the reference sequence.

For a correct WER computation all the transcripts need to be normalized in terms of capitalization, punctuation, and numbers. Lower WER means better system performance.

3.4.5 Deployment

After approval of the best model based of the evaluation metric, we started the process of integrating the ASR model into a web platform designed for document transcription. This implies defining first, preprocessing and standardization step to be followed once an audio file is uploaded by the user before performing the inference.

3.5 Experimental Process Model

In this section we present the schematic architecture implanted for the classic ASR experiments, and the pseudocode for finetuning the end-to-end ASR models.

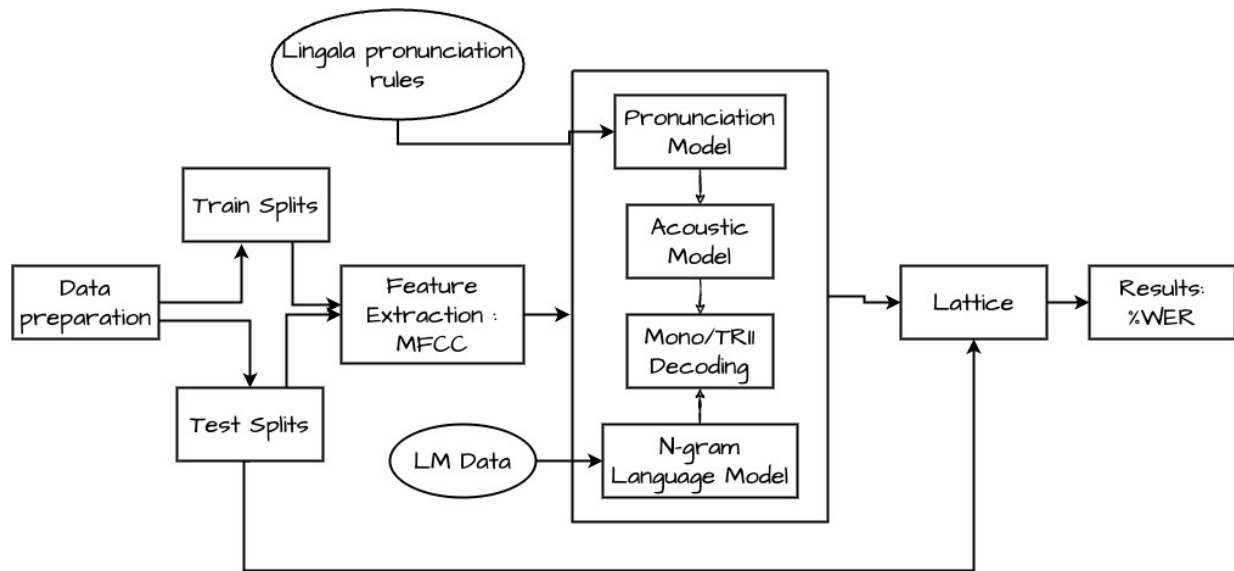


Figure 3-6: The developed Lingala ASR model

Table 3-3: Whisper pseudo-code

Step	Task
1	Loading the dataset
2	Preparation: <ol style="list-style-type: none"> 1. Feature Extractor <ol style="list-style-type: none"> 1) Pad/Truncate audio inputs to 30 seconds length 2) Convert audio input to log-Mel spectrogram 2. Tokenizer: Performs seq2seq mapping
3	Combine Feature Extractor and Tokenizer
4	Data Pre-processing <ol style="list-style-type: none"> 1. Resample the audio to 16kHz 2. Compute log-Mel spectrogram 3. Encode the transcription to label IDs
5	Training and Evaluation <ol style="list-style-type: none"> 1. Define and initialize the data collator 2. Define the evaluation metrics (WER) 3. Load a pre-trained Checkpoint 4. Define the Training configuration 5. Train the Model

Table 3-4: *MMs pseudo-code*

<i>Step</i>	<i>Task</i>
1	Loading the dataset
2	Preparation: Tokenizer: <ol style="list-style-type: none"> 1. Normalize transcripts 2. Define a new vocabulary 3. Add a CTC's "Blank token" to the vocabulary Define a Feature Extraction pipeline
3	Combine Feature Extractor and Tokenizer
4	Data Pre-processing <ol style="list-style-type: none"> 1. Resample the audio to 16kHz 2. Encode the transcription to label IDs
5	Training and Evaluation <ol style="list-style-type: none"> 1. Define and initialize the data collator 2. Define the evaluation metrics (WER) 3. Load a pre-trained Checkpoint 4. Define the Training configuration 5. Train the Model

3.6 Conclusion

This chapter presented the methodology adopted for achieving the objectives of this research. The overall research design is more experimental oriented with a mixture of both qualitative and quantitative approach for the assessment of the ASR system. Given the lack of dataset in the target domain, we aim to create a custom speech corpus using a rigorous transcription guideline. This dataset can then be combined with the fleur's dataset for finetuning. As for modelling, we opted for classic supervised ASR as well as cutting-edge pretraining techniques. The WER is used as evaluation metrics.

CHAPTER 4: IMPLEMENTATION, RESULTS AND DISCUSSION

4.1 Introduction

The third chapter dealt with setting up the methodological framework of the study. In this new chapter, we first describe the development phases of the final artifact, focusing more on the feasibility and requirement analyses. Then we engage in the experimental part of the adopted modeling techniques and report the results. We finally interpret and discuss the obtained results by considering various practical aspects of the trained models.

4.2 Artifact development

The final product of this project is a system integrating an ASR model with the lowest WER percentage. The system can convert a Lingala speech utterance into text. Below is the summary of the outcomes from the various development phases.

4.2.1 Feasibility analysis

We conducted a feasibility study to assess the viability and practicality of implementing a Lingala ASR system for media use case. We typically considered various aspects including technical feasibility, economic feasibility, operational feasibility as well as legal and ethical considerations.

1) Technical Feasibility:

From a technological point of view, tools required for developing a speech recognition system are easily accessible. There are well-established algorithms, models, and techniques, such as probabilistic and deep learning-based approaches, that can be used. In the context of media application use case, the system is compatible with current hardware and software infrastructure, including operating systems, CPUs, and microphones.

However, in addition to the long learning curve for mastering some of these ASR tools, there can be technical difficulties when addressing accent and speech pattern variances or effectively

identifying speech in noisy environments. To solve these issues, research and development initiatives are needed.

2) Operational feasibility:

For a positive impact of a speech recognition system on existing operations and workflows, the system should seamlessly integrate into the existing processes without causing disruptions. Users and stakeholders must be willing to adopt the system and go through the required training in order to utilize it efficiently. The system's scalability and adaptability must be taken into account to support future expansion and shifting needs. Different domains and accents should all be supported by the system.

To assess such impact, a prototype of a web transcription platform was developed with Lingala Journalists as primary target end users.

3) Economic Feasibility:

The costs considered for the development and implementation of the Lingala ASR system include software development, hardware infrastructure, training data acquisition.

Overall, it takes a lot of human resources to compile a speech corpus from the accessible sources. Within our means we were able to create 3 hours of audio-transcription data as a minimum fixed size for the development of the Lingala ASR proof of concept with the help of some volunteers. Computation resources for training the models was another constraint that implied a cost. We relied on the affordable hardware accelerator plans proposed by Google Cloud Platform to deal with this constraint.

Despite these constraints, the benefits of such system are worth the cost. These include increasing the productivity of journalists, promote the preservation of Lingala documents, and improve customer satisfaction.

With the end goal of improving Lingala ASR system using data from radio station, the amount of time required to develop a proof-of-concept system in a timely manner was judged feasible.

4) Legal Feasibility:

There were not special legal and ethical constraints attached to the data used, since they are publicly accessible by nature. However, we made sure to implement user anonymization measures to protect personal information.

4.2.2 Requirements analysis

a. Functional requirements

The system delivers the following functionalities:

1. Uploading audio file in multiple formats
2. Transcribe an audio file
3. Editing transcription
4. Downloading the produced transcription
5. Provide APIs or integration capabilities to enable seamless integration with other applications, platforms, or devices.

b. Non-functional requirements

Regarding the performance, the system is expected to produce promising results in terms of WER, such that with more data it can approach standard production ready ASR engines. In relation to reliability, the system is capable to provide transcription service to the end users within a reasonable amount of time.

4.2.3 System users

Native speakers of Lingala, radio news journalists, data scientists, linguists, and system administrators are some of the stakeholders involved in the development of the system. Lingala native speakers played a critical role of transcribing the audio data used for training the ASR model. Although Lingala Journalists are the primary target users, the system can serve as a general transcription platform, given the diversity of topics covered.

4.2.4 System design

The Web platform and the Lingala ASR engine are the two major components of the system. The web application mainly consists of a user interfacing and the API which interacts with the ASR model to provide transcription service to the user. Additional features offered by the web platform include translating a text from Lingala to Swahili. The ASR model acts as the backend engine that performs the speech-to-text task.

The overall system design is shown in Figure 4-7

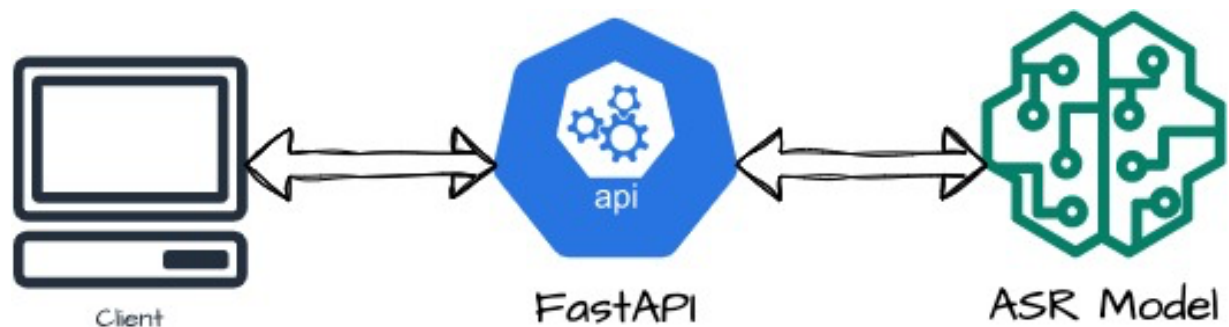


Figure 4-7: System block diagram

4.3 ASR Experiments

4.3.1 Tools used for Implementation

Python was the main programming language used for the implementation of the project. We wrote several python scripts to automate the preparation steps before the transcription process. To run baseline ASR experiments using the classic approach, we used Elpis (Foley et al., 2018), a user-friendly ASR toolkit from University of Queensland developed in python. It simplifies the access to training ASR models using various engines, namely *Kaldi*, *ESPnet*, *Hugging Face Transformers* library. Elpis is wrapped in a docker container to facilitate its execution across multiple platforms. A key advantage of Elpis is that it allows training ASR system with very little amount of data (1 hour).

To meet the transcription format currently supported by the Elpis Kaldi model, we converted the transcriptions files into *eaf* format using the *ELAN* software.

For the end-to-end ASR finetuning, we mainly used the *Transformers* library by Hugging Face which supports the two leading deep learning frameworks: *Pytorch* and *TensorFlow*. The

official Whisper and MMS checkpoints implementation used for the finetuning are also available on Hugging face.

The web application API was implemented with *FastAPI*, a modern high-performance python-based web framework specifically design for creating production ready APIs. The user interface of the web transcription platform was developed with *Streamlit*, an open-source Python library designed for rapid prototyping of web apps for machine learning and data science.

4.3.2 ASR Models

As described in the previous chapter, we used both the traditional Supervised and the modern pretraining ASR modeling approaches to run experiments on the collected dataset.

a. Classic ASR

This approach consists of training three sub-models sequentially: the acoustic model, the pronunciation dictionary and the language model. We used the Kaldi recipe (Povey et al., 2011) as wrapped in the Elpis toolkit. The data preprocessing involves the following steps:

- 1) Clean audio and transcription (remove junks and change non-lexical form)
- 2) Segment audio at utterance level
- 3) Convert text transcription to ELAN format (.eaf)
- 4) Parallel data of matching files names
- 5) Create a letter to sound file that contains rules for mapping orthography to phonemes.
- 6) Generate a pronunciation lexicon that maps every word in the cleaned corpus to its pronunciation.

The pronunciation lexicon was generated based on the letter to sound rules that we provided to the system. Given the great preference for spoken word over written word in Lingala generally, it is challenging to verify the idea of standard, recognized spelling: In Kinshasa, Lingala is written using the thirteen common consonants (b, p, m, v, f, d, t, z, s, n, l, g, k) and five short oral vowels (i [i], e [e], a [a], o [o], u [u]). Pre-nasalized consonants are then added to this, which are produced when nasal sounds (n or m) are combined with bilabial (b, p), alveolar (d, t, z, s), and velar (g, k) phonemes. The semi-vowels w and y are also part of the Lingála phonological system. Other sounds such as [r] /r/, [ʃ] /sh/, [ʒ] /j/, [tʃ] /tsh/, [dʒ] /dj/, /kp/, /gb/, and /ngb/, and h (muet) are either derived from nearby Bantu languages or borrowed from European languages.

It should be mentioned here that because of frequent code-switching and cod-mixing in the standard Lingala, we included other French phonemes to allow the recognition of French word in a given speech.

The following Table 4-5 shows an excerpt of the dictionary generated by the system and manually refined for the training stage.

Table 4-5: Excerpt of the Lingala lexicon

Word	Pronunciation
akopesaka	a k o p e s a k a
babanzi	b a b a [ⁿz/ⁿʒ] i
bakangami	b a k a [ⁿg] a m i
kotsha	k o [tʃ] a
mbula	[ᵐb] u l a
tozwami	t o z [w] a m i

After feature extraction the Kaldi recipe script provides a number of training methods using the Gaussian Mixture Model/Hidden Markov Model framework. Gaussians are picked at random to estimate each phoneme at the start of the training. To locate the phoneme in the audio during training, the transcription is positioned in relation to the audio. The Gaussians are then adjusted to better match the actual phonemes. Specifically, two training processes happen under the hood. *Monophone* training where each phone is considered independently to its context and the *Triphone* training where context surrounding a phoneme is taken into account.

The language model was generated using the text of all utterances from the custom corpus and the additional Lingala text corpus of news outline scraped from the Okapi web site.

b. End-to-end Fine-tuning with Whisper

To perform transfer learning on the created corpus, we used Whisper, a supervised ASR model, pretrained on a large transcribed multilingual audio corpus. Whisper architecture consists of an encoder-decoder models implemented as Transformer. The Encoder component takes an input audio (the Log-mele spectrogram feature sequence) and produces hidden state representation.

The decoder component acts as language model. It takes transcript tokens as input and predicts the next tokens based on both the previous tokens and the hidden states from the encoder.

c. End-to-end Fine-tuning with MMS

We finetuned the recent MMS pretrained model using the Hugging face *Transformers* library. MMS is a CTC based ASR model pretrained using the self-supervised learning strategy. It makes use of adapters, a technique for optimizing previously trained models while maintaining their original model parameters across several languages. For each target language, a select few adaptor weights are fine-tuned to recognize its particular phonetic and grammatical characteristics.

4.4 Results

We ran several experiments on the created corpus using both cascade and end-to-end approach to find the most efficient model. We combined both manual and automated validation to assess the performance of the ASR models. WER metrics was used to do an automated evaluation, and the results were recorded. WER computes in percentage the number of incorrect words hypothesized by the ASR over the total number of words. The following sections report the performance of the different experiments.

4.4.1 Cascade ASR

We ran incremental experiments to train Kaldi models by assessing the WER on different corpus size and model settings.

The initial model was trained on just 5 minutes of male speakers. It produced 97.06 % of WER. With 27 minutes of female speakers, we got 50.1 % of WER in the second experiment.

The third experiment was based on 50 minutes of transcribed audio from female speakers. It surprisingly gave 60.09% WER, 10 percents worse than the results of the second experiment. This was probably due to some transcription inconsistencies noticed in the corpus used.

In the fourth experiment, we scaled the corpus to 95 minutes including both genders. The resulting WER was of 74.39%. The fifth experiment used 2.5 hours of data from both genders. The trained model yielded 64.6 % of WER.

In the final experiment we used the entire corpus of 3 hours. This time we used 3-gram language model and additional text corpus. We got a WER of 55.6%.

Table 4- below summarizes the results of these experiments.

Table 4-6: *Classic ASR results*

Experiment #	Corpus size	Gender ratio (F/M/D in min)	Language model value	Additional text corpus	WER%
1	5 min	0/5/0	2		97.06
2	27 min	27/0	2		50.1
3	50 min	50/0	2		60.09
4	1.5 h	83/11/1			74.39
5	2.5h	133/15/2	2		64.6
6	3h	160/20/4	3	Yes	55.6

Figure 4-8 illustrates the same results with a graph of WER as function of corpus size

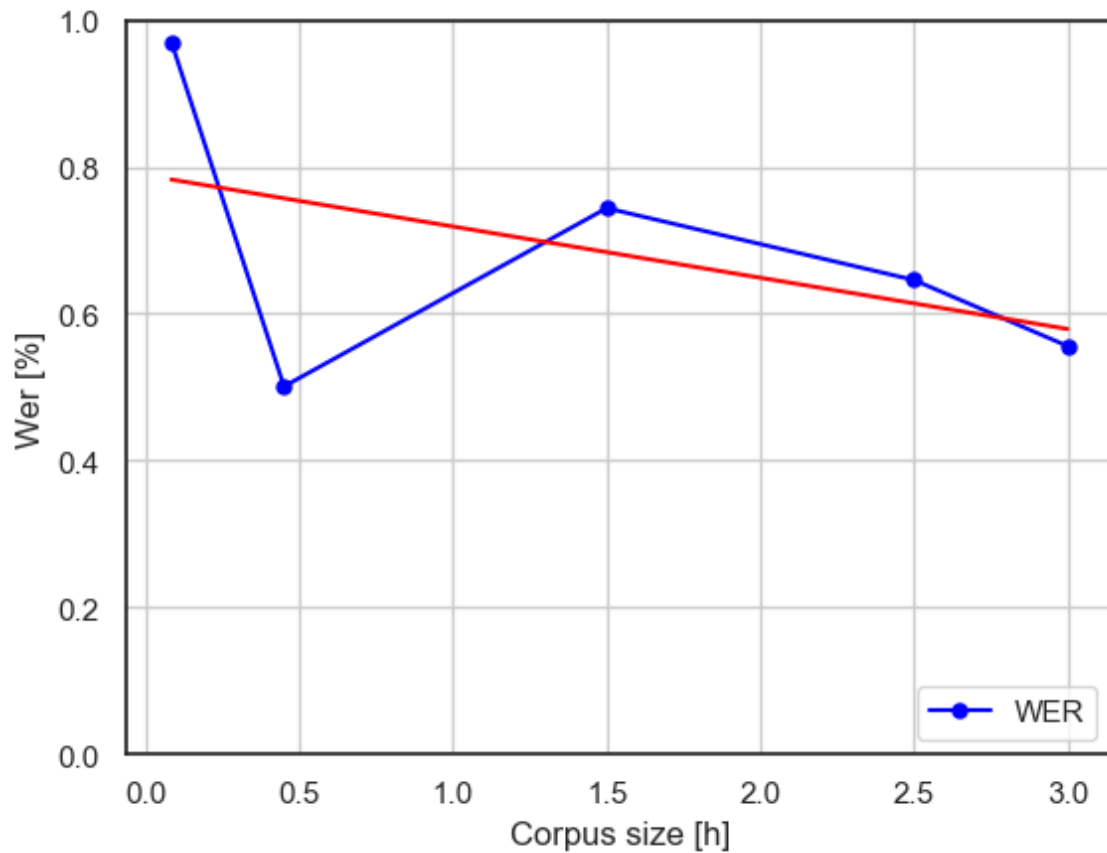


Figure 4-8: *Corpus vs WER trend for classic ASR*

4.4.2 Whisper fine-tuning

We leveraged the official pre-trained Whisper checkpoints for finetuning. We chose the tiny and the base model size configurations to run our experiments respectively on the created corpus and the mixed dataset resulting from combining our custom data with the Lingala fleurs dataset. 43.52% and 25.05 % of WER are the results we got respectively for the tiny and base checkpoints.

The tables below summarize the results of the two checkpoints

Table 4-7: *Whisper tiny performance on custom dataset*

Step	Training Loss	Validation Loss	WER
500	0.1767	0.9122	45.79
1000	0.0191	1.0786	45.38
1500	0.0059	1.1891	47.66
2000	0.0019	1.2661	43.52

Table 4-8: *Whisper base performance on mixed dataset*

Step	Training Loss	Validation Loss	WER
1000	0.0081	0.6218	29.8710
2000	0.0016	0.6865	25.1188
3000	0.0009	0.7152	24.9151
4000	0.0007	0.7265	25.0509

4.4.3 MMS fine-tuning

Due to runtime constraints, we fine-tuned the adapter layers of the mms-1b-all checkpoint on the mixed dataset for very few numbers of steps.

Table 4- below summarizes the results of the two checkpoints

Table 4-9: *MMS performance*

Step	Training Loss	Validation Loss	WER
200	0.4886	0.9371	31.77

4.5 Interpretation and Discussion

In this work, we used a custom created speech corpus, the Google's fleurs datasets and the WER evaluation metric to assess performance of Lingala ASR models trained using the traditional and the modern approaches. We will examine and interpret the results in this discussion section, highlighting the strengths and weaknesses of the speech recognition models as well as possible areas for improvement.

The results of our experiments demonstrate that the trained models achieved promising commendable performance in terms of WER metric. The best performing model trained on 3 hours of custom dataset using the traditional approach achieved 55.6 % of WER.

As stressed in the original whisper paper, we noted a remarkable performance improvement when finetuning the two whisper checkpoints. Compared to the original tiny and base whisper checkpoints which yielded 105% and 96% of WER respectively, we got 43.52% and 25% WER respectively by finetuning these checkpoints respectively on custom corpus and on the mixed dataset.

Both the training loss and the validation WER looks acceptable for the MMS model. With just 200 steps, the model performs quite well.

4.5.1 Strengths of the model

Output transcription from the best classic ASR model show good results in terms of spelling. As illustrated below, we noted that in general, words correctly predicted by the classic ASR suffer less from misspelling issues. This good performance is largely influenced by the quality of the pronunciation dictionary.

Table 4- shows inference examples on some samples of untranscribed audio.

Table 4-10: *Reference transcript vs classic ASR predicted transcript*

Reference transcript	Predicted transcript
Bakobengaka ba kuluna. Kolandana na misala mya sembo bosambisami boye bookosalema na mbika. Na yango bato banso bakweya makama ya bato mabe baye basengeli koya mpo ya kobafunda.	bakobengaka ba kuluna kolala na misala sembo bosangisami boye bookosalema bandisa yango bato basimba po ya makama ya bato mabe baye basengeli koya mpoya ya kobasunga
Babomaki bato mwa zomi na isato ya sanza ya yambo na tongo. Baye bazalaki kosala bolukaluki bakundoli ete ezali batomboki ya CODECO kouta na mboka Kafende bayingelaki na Nyamamba. Babomi bato, batumbi ndako na boutiques zomi na motoba, mpe bakunzi hotel moko. Mbano ya bolukaluki esalemi na auditorat supérieur elakisi ete sima ya mbeba, bato tuku mibale na minei bakufi mpe bakunzi bibembe na bango na libulu misato. Moko ezali na bato zomi na misato ya mibale na ba	babomaki bato mazwami misato ya sanza ya yambo na tongo baye bazalaki kosala bolukaluki bakumbi eliki ezali batomboki codeco kokoka na mboka asengi bayingelaki nyama na babomi bato batumbi na Kouamouth zomi na motoba mpe makanisi mpo te moko mbano ya bolukaluki soni mino ditu master elakisi ete sima ya mbeba bato tuku mibale na minei bako site bakumbi bibembe na bango na mibu misatu moko ezali na o na misato míme na ngo

The whisper finetuned model shows robustness to noise, handling of both speaker genders and reasonable inference time. Such robustness is ensured by the large corpus on which whisper was pretrained.

Regarding the bias-variance tradeoff, we noted that in our context using the tiny and base checkpoints, prevented the model from overfitting given a relatively small number of parameters of the pretrained checkpoints.

Below are some inference examples on the same audio samples.

Table 4-11: *Reference transcript vs Whisper base predicted transcript*

Reference transcript	Predicted transcript
Bakobengaka ba kuluna. Kolandana na misala mya sembo bosambisami boye bookosalema na mbika. Na yango bato banso bakweya makama ya bato mabe baye basengeli koya mpo ya kobafunda.	bakobengaka bakuluna kolanda na misami ya nsambo basambisa mibu ye boko sallema na mbisa na yango bato banso bakwea makambo ya bato mabebeye basengeli koya po ya kobafunda
Babomaki bato mwa zomi na isato ya sanza ya yambo na tongo. Baye bazalaki kosala bolukaluki bakundoli ete ezali batomboki ya CODECO kouta na mboka Kafende bayingelaki na Nyamamba. Babomi bato, batumbi ndako na boutiques zomi na motoba, mpe bakunzi hotel moko. Mbanu ya bolukaluki esalemi na auditorat supérieur elakisi ete sima ya mbeba, bato tuku mibale na minei bakufi mpe bakunzi bibembe na bango na libulu misato. Moko ezali na bato zomi na misato ya mibale na ba	babomaki bato mwa zomi na istato ya sanza ya mbona ntongo bayeba azalaki kosala boluka luki bakundoli ete ezali batomboki akodeko kotala na mboka kafe nde bayingelaki na nyama mba babomi bato batombinda kona botique zomi na motoba pe bapunzi hotel moko bano ya boluka luki esalemi naudite horrat superiare elakisi ete nsima ya mbeba bato tuku mibale na minei bakufi pe bakundi bibembe bango na libulu misato moko ezalela bato zomi na misato ya mibale na batunnels

Despite the fact of being finetuned for a very small number of steps, the MMS model shows strong robustness in the output transcription as illustrate in table below

Table 4-11: Reference transcript vs MMS predicted transcript

Reference transcript	Predicted transcript
Bakobengaka ba kuluna. Kolandana na misala mya sembo bosambisami boye bookosalema na mbika. Na yango bato banso bakweya makama ya bato mabe baye basengeli koya mpo ya kobafunda.	bakobengaka bakuluna kolanda na misala mya sembo bosambisa miboye bo kosalema na mbisa na yango bato banso bakwea makama ya bato mabe baye basengeli koya mpo ya ko bafunda
Babomaki bato mwa zomi na isato ya sanza ya yambo na tongo. Baye bazalaki kosala bolukaluki bakundoli ete ezali batomboki ya CODECO kouta na mboka Kafende bayingelaki na Nyamamba. Babomi bato, batumbi ndako na boutiques zomi na motoba, mpe bakunzi hotel moko. Mbanu ya bolukaluki esalemi na auditorat supérieur elakisi ete sima ya mbeba, bato tuku mibale na minei bakufi mpe bakunzi bibembe na bango na libulu misato. Moko ezali na bato zomi na misato ya mibale na ba	babomaki bato mwa zomi na isato ya sanza ya yambo na tongo baye bazalaki kosala bolukaluki bakundoli ete ezali batomboki ya kodeko koutaa na mboka cafe nde bayingelaki na nyamamba babomi bato batumbi ndako na butique zomi na motoba mpe bapunzi hotel moko mbanu ya bolukaluki esalemi na auditorat superieur elakisi ete sima ya mbeba bato tuku mibale na minei bakufi mpe bakundi bibembe na bango na libulu misato moko ezali na bato zomi na misato ya mibale naba

4.5.2 Limitations

Despite these strengths, our investigation showed that, given the current WER percentage, the trained models had several shortcomings and thus still far from competing with human baseline performance. Their inability to effectively transcribe speech with uncommon vocabulary was one of their weaknesses. When the models came across fewer common phrases or technical terminology that were not in the training set, they were more likely to mistake. This implies that increasing the vocabulary and including domain-specific data might considerably enhance the models' performance in specialized applications like media communication.

Another limitation was the lower performance while processing speech with fast speech patterns or high speaking rates, particularly with the Kaldi model. Occasionally, the models had trouble effectively in capturing fast-moving speech parts, which raised the word error rates.

Further research on optimizing the models for rapid speech patterns could help mitigate this limitation.

The classic model also struggled to recognize words in audio where both female and male speakers were intervening. This behavior confirms the poor performance reported by (Kimanuka et al., 2023) on multiple speakers. For some noisy audio, the model completely fails to recognize words.

Additionally, we noticed that speech recognition algorithms require better contextual awareness. The models performed well when transcribing single sentences or brief phrases, but they occasionally had trouble grasping complicated sentence patterns and information that depends on context. Enhancing the model's comprehension of contextual cues and dependencies by using large n-gram language model might improve the precision and overall performance of Kaldi model.

Compared to the classic model, the whisper model easily hallucinates on some examples. Part of the reason is the relatively small dataset used for finetuning. On big shortcoming of the two pretrained ASR models we finetuned is their relatively big size. This makes not convenient for finetuning and deployment in the context of limited computational resources.

4.5.3 Ways for improvement

To improve the performance of the trained models, several potential avenues can be considered. First, adding more training data from diverse sources might help improve the vocabulary coverage and enhance the model's generalization capabilities. Augmenting the training data with domain-specific or rare vocabulary could also address the limitations observed in handling specialized terminology.

Second, exploring methods for integrating contextual data with language modeling might improve the models' comprehension of context. The models' capacity to comprehend and predict context-dependent speech may be improved by using language models that take into account the surrounding text.

Based on the performance trend obtained up to now, both modeling can benefit a lot from utilizing more data. Specifically, the classic model can improve its recognition capabilities with a more refined pronunciation lexicon.

4.6 Conclusion

This chapter allowed us to present the implementation phases of the artifact and to discuss the results of the ASR experiments conducted. The feasibility analysis showed that it is possible to build a transcription platform prototype despite the mentioned challenges. Considering practical aspect such as inference speed, robustness to noise, and handling of long audio segment which are crucial for the deployment of ASR model, we opted for the whisper base model that yield 25% WER after fine-tuning on the mixed dataset.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.1 Summary of findings

Objective 1: To create a Lingala speech corpus for broadcast domain.

To reduce the current gap in the availability of domain specific Lingala data for speech tasks this study created 3 hours of Lingala speech corpus¹ using audio archives from Okapi Radio. A key insight revealed by the created corpus is the confirmation that even the so-called standard version of Lingala version spoken in DRC is not that pure. The usage of the language is characterized by many code-mixing and code-switching showing the dynamic of Lingala in terms of simplification for communication. The bias toward female speakers that characterizes the gender distribution of the corpus, is a clear indication of the female gender predominance in the journalism job in Kinshasa.

The following figure summarizes our data creation pipeline.

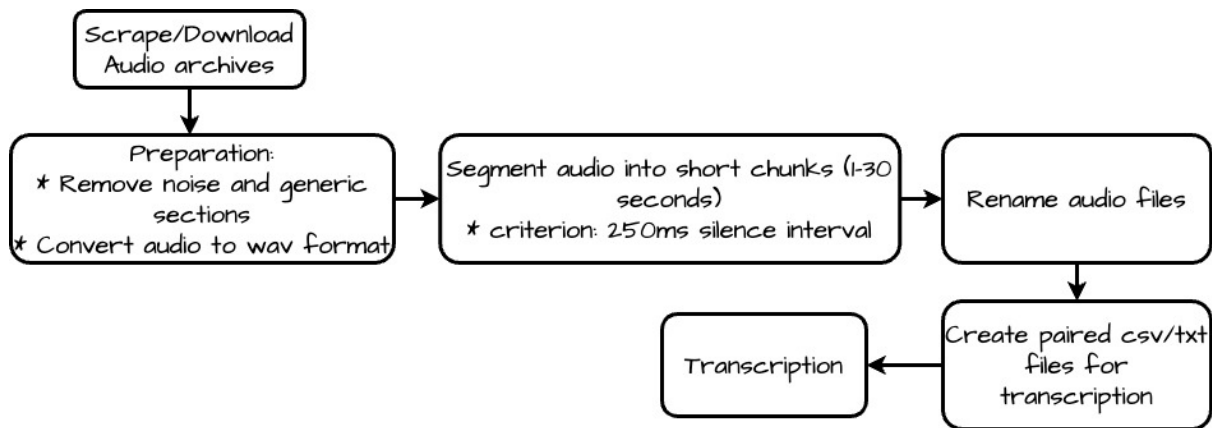


Figure 5-9: *Lingala corpus creation pipeline*

Objective 2: To experiment on which ASR modeling approach yield promising performance results on the created corpus.

Several ASR models were implemented in this research using different modeling approaches and learning strategies. First, baselines experiments were gradually run on the created corpus using the classic Supervised ASR approach. 55.4 % WER was the best result obtained after

¹ The dataset can be found at: <https://huggingface.co/datasets/BrainTheos/ojpl/resolve/main/data/ojpl.tar.gz>

training on the entire 3hours dataset. From the WER vs corpus size curve, a great performance can easily be projected with more data.

Secondly, Transfer learning of some state-of-the art multilingual ASR pre-trained models was explored. The custom dataset was mixed with the Lingala subset of the Google's fleurs dataset to increase the size for fine-tuning. Specifically, two pre-trained ASR model were fine-tuned²: Whisper and MMS. The whisper experiments were based on the tiny and base checkpoints. 43% WER and 25% WER were respectively the result of the two checkpoints. The MMS experiment was based on the Lingala adapter of the mms-1B checkpoint. We obtained 31%WER. Despite the low WER of the MMS, the whisper base model was finally selected for deployment considering its light weight, robustness to noise and ability to handle long input audio.

Objective 3: To develop a transcription platform that uses the trained ASR model as a backend engine.

The research by (Arakawa, Yakura, & Goto, 2022) served as inspiration for the creation of an online transcription platform prototype. This prototype can take lingala audio as input and produce the related text transcription. Additionally, it provides the option to edit, rate, save, and download the predicted transcription.

5.2 Conclusion

In this work we proposed a prototype of a web transcription platform for Radio stations in Kinshasa. The end goal of the artifact is to boost the Language preservation through document conservation. In addition to the suggested corpus creation pipeline, we showed that the resulting ASR model can highly assist Lingala journalists as well as many illiterate people for various applications. With such corpus, we hope to foster the applications of speech technology in the DRC, and thus alleviate the language barrier in such a diverse linguistic country.

Specifically, the initial research questions have been answered as follows:

² The finetuned models are available at <https://huggingface.co/BrainTheos/whisper-base-ln> and <https://huggingface.co/BrainTheos/wav2vec2-large-mms-1b-lingala-colab>

Research question 1: How can we construct a reliable Lingala speech corpus for a broadcast domain?

Leveraging existing Lingala audio archives as data source, this study has set up a corpus creation pipeline that includes technical assets for crawling and preparing Lingala audio for transcription. Based on both technical and linguistical considerations, transcription guidelines and rules were clearly defined and explained to Lingala native speakers.

Research question 2: What are the optimal modeling architecture and training strategies that best fit our use case and dataset?

Two modeling approaches shape the ASR research community. The first is based on traditional statistical method like GMM-HMMs to train ASR system in a Supervised way. The second relies on advanced deep-learning architecture using transfer learning strategies. This research confirmed the fine-tuning of existing large pre-trained models as the best approach to the modeling of a Lingala ASR. Traditional methods, however, still offer additional resources that can aid in a deeper understanding of the linguistic features.

5.3 Limitations

Dataset redundancy and non-normalized corpus

One problem with the created Lingala corpus is the redundancy due rebroadcasting. Despite the efforts to reduce these repetitions, it was challenging for us to track all the duplicate clips, especially after they have been transcribed. This might have affected in some way the generalization performance of the ASR model. In the same line, the Lingala subset of the fleur's dataset was highly unnormalized. This explains the relatively small improvement achieved even after combining this dataset with the custom corpus.

Relatively poor performance of the deployed model

As the key component of the proposed artifact, we expect that the accuracy of the ASR model that recognizes users' speech input should be high. The performance of the deployed model needs to be improved in order to reach a WER that is below 15% which is considered as the human baseline WER.

Robustness of the proposed artifact

It is important to assess the effectiveness of the transcription platform, particularly as it is being used to transcribe a wider variety of speech. For instance, given the nature of broadcasting, it may be more challenging to use the suggested artifact when the speech to be transcribed involves impromptu dialogue between multiple speakers with a lot of overlaps.

5.4 Recommendations

Going forward, several avenues for future work are possible. Among which:

Include audio samples from other Radio stations. Although we made an effort to develop a domain-specific speech corpus, several Radio stations in Kinshasa were left out. Obtaining samples from these radio stations would not only help expand the corpus but also give a more accurate sense of how the standard Lingala is used.

Fine-tuning a domain closer pre-trained model. Another interesting avenue is to fine-tune a ASR model like CdWav2vec (Kimanuka et al., 2023) which was pre-trained on a large unlabeled dataset majority from the broadcast domain.

REFERENCES

- Arakawa, R., Yakura, H., & Goto, M. (2022). BeParrot: Efficient Interface for Transcribing Unclear Speech via Respeaking. *27th International Conference on Intelligent User Interfaces*, 832–840. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3490099.3511164>
- Awino, E., Wanzare, L., Muchemi, L., Wanjawa, B., Ombui, E., Indede, F., ... Okal, B. (2022, October 29). *Phonemic Representation and Transcription for Speech to Text Applications for Under-resourced Indigenous African Languages: The Case of Kiswahili*. arXiv. <https://doi.org/10.48550/arXiv.2210.16537>
- Cattoni, R., Di Gangi, M. A., Bentivogli, L., Negri, M., & Turchi, M. (2021). MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, *66*, 101155. <https://doi.org/10.1016/j.csl.2020.101155>
- Conneau, A., Bapna, A., Zhang, Y., Ma, M., von Platen, P., Lozhkov, A., ... Johnson, M. (2022, April 13). *XTREME-S: Evaluating Cross-lingual Speech Representations*. arXiv. <https://doi.org/10.48550/arXiv.2203.10752>
- Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., ... Bapna, A. (2022, May 25). FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. Retrieved April 18, 2023, from ArXiv.org website: <https://arxiv.org/abs/2205.12446v1>
- Doumbouya, M., Einstein, L., & Piech, C. (2021). Using Radio Archives for Low-Resource Speech Recognition: Towards an Intelligent Virtual Assistant for Illiterate Users. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(17), 14757–14765. <https://doi.org/10.1609/aaai.v35i17.17733>
- Foley, B., Arnold, J. T., Coto-Solano, R., Durantin, G., Ellison, T. M., van Esch, D., ... others. (2018). Building Speech Recognition Systems for Language Documentation: The

- CoEDL Endangered Language Pipeline and Inference System (ELPIS). *SLTU*, 205–209.
- Iranzo-Sánchez, J., Silvestre-Cerdà, J. A., Jorge, J., Roselló, N., Giménez, A., Sanchis, A., ... Juan, A. (2020). Europarl-ST: A Multilingual Corpus for Speech Translation of Parliamentary Debates. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8229–8233. <https://doi.org/10.1109/ICASSP40776.2020.9054626>
- Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Third Edition draft). Stanford.
- Kabenamualu, S. K., Marivate, V., & Kamper, H. (2022). LiSTra Automatic Speech Translation: English to Lingala Case Study. *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, 63–67. Marseille, France: European Language Resources Association. Retrieved from <https://aclanthology.org/2022.dclrl-1.8>
- Kimanuka, U., Maina, C. wa, & Buy, O. B. (2023). *Speech Recognition Datasets for Low-resource Congolese Languages*. Retrieved from <http://repository.dkut.ac.ke:8080/xmlui/handle/123456789/7946>
- Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M., & Qadir, J. (2023, March 21). *Transformers in Speech Processing: A Survey*. arXiv. <https://doi.org/10.48550/arXiv.2303.11607>
- Li, J. (2022). Recent Advances in End-to-End Automatic Speech Recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1). <https://doi.org/10.1561/116.00000050>

- Maniamfu, P., Kiketa, V., Muepu, D., & Kabongo, J. (n.d.). *Word embedding approach in a vector space based on Word2vec and Fasttext for Lingala language representation*. 8.
- Ogayo, P., Neubig, G., & Black, A. W. (2022, July 1). Building African Voices. <https://doi.org/10.48550/arXiv.2207.00688>
- Palma, H. L. (2022). *Native Languages of the Democratic Republic of Congo* (pp. 99–116). Brill. https://doi.org/10.1163/9789004516724_007
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others. (2011). The Kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, (CONF). IEEE Signal Processing Society.
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ... Auli, M. (2023, May 22). *Scaling Speech Technology to 1,000+ Languages*. arXiv. <https://doi.org/10.48550/arXiv.2305.13516>
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December 6). *Robust Speech Recognition via Large-Scale Weak Supervision*. arXiv. <https://doi.org/10.48550/arXiv.2212.04356>
- Raschka, S. (n.d.). *Machine Learning Q and AI (Early Access)*.
- Ritchie, S., Cheng, Y.-C., Chen, M., Mathews, R., van Esch, D., Li, B., & Sim, K. C. (2022, October 4). *Large vocabulary speech recognition for languages of Africa: Multilingual modeling and self-supervised learning*. arXiv. <https://doi.org/10.48550/arXiv.2208.03067>
- Roger, V., Farinas, J., & Pinquier, J. (2022). Deep neural networks for automatic speech processing: A survey from large corpora to limited data. *EURASIP Journal on Audio, Speech, and Music Processing*, 2022(1), 19. <https://doi.org/10.1186/s13636-022-00251-w>

- Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., ... Post, M. (2021, June 14). *The Multilingual TEDx Corpus for Speech Recognition and Translation*. arXiv. <https://doi.org/10.48550/arXiv.2102.01757>
- Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. New York, NY, USA: Association for Computing Machinery. <https://doi.org/10.1145/3411764.3445518>
- Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., ... Wu, Y. (2023, March 2). *Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages*. arXiv. <https://doi.org/10.48550/arXiv.2303.01037>
- Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., ... Wu, Y. (2022). BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6), 1519–1532. <https://doi.org/10.1109/JSTSP.2022.3182537>


APPENDICES

1. Graphical user interface of the transcription platform

Lingala ASR System

Input Audio

Choose a file

 Drag and drop file here
Limit 200MB per file • MP3, MP4, WAV Browse files

 ojpl_25012023_s_2871.wav 1.0MB ✕

Listen to audio

▶ 0:00 / 0:21  🔊 ⋮

Generated Transcripts

Correct if necessary :

ebandisa makimo mosala misato e ya na lombomba chimisela ya moko misami mpoya maponami yango esalemaki na kalaka mokambi ya misala ya maponami oyo ekolo oboso ya ministro oyo oyo akomaki mosala mibale na engumba mokonzi ya etuka ya okatanga zma lukonde asengaki na baélubumbashi basoko kende komikomisa pamba te vango ezali

Click here to download

2. Some excerpts of the predicted transcript

Reference transcript	Predicted transcript
Basengi na basusu kosangana na maponomi ekosalema na sanza ya zomi na mibale ya mobu eye.	basengi na basusu kosangana na ma ponomi e kosalema na sanza ya zomi na mibale ya mobu eye
Na bobangi bopusani ya ba M23 baye baa kobunda uto poso yoko na ba FARDC.	na bobangi bopusani ya ba m20 wana bayebaka kobunda eutoposo yoko na ba farads e
Lisusu bakobandisa pe bobeti ntango po ya milulu ya bokati mangwele na poso ekoya.	lisusu bakobandisa pe bobeti ntango po ya milulu ya bokati mango ele na poso ekoya
Na etuka Tanganyika bowelani bwa ndelo ya mabele kati ya territorne ya Kalemi na eye ya Nyunzo	na etuka tanganika bowellani bandelo ya mabele kat ya territorne ya kalemi na e ya nyonso
Kolongola to kobongola meko eyei ezali mokumbi ya moyengeli ya etuka mpo ye nde azwaki yango.	kolongola tokobongola meko eyei esali mokumbi emoyengele ya etuka koyende ya zwaki yango
Engumba Goma ntina ya likama lina ezali kozanga kotosa mibeko ya kotambola na mayi.	engumba goma ntina ya likamalina ezali kozanga kotosa mibeko ya kotambola na mai
Mpe yambo eyano epesama nasingi ba mbotama ba quartier totangi komibongisa bo elongo bani.	p yambo eyano epesama na singi bambotama ba quartier totangi komibongisa bo elongo bani
Mpo ya mabota mana bosimba ete lisanga lioko litiami mpo ya kolandela makama.	mpo ya mabota mana bosimba ete lisanga lioko litiami mpo ya kolandela makama
Zomi na sambo baye bakangaki bango na secteur ya wamba kuna na territoire ya bagata batalisaki bango epaye ya moyangeli etuka mokolo mwa mosala mibale.	zomi nasambo baye bakangaki bango na secteur a wamba kona na territoire ya bagata batalisaki bango epaye ya moyangeli etuka mokolo ma mosala mibale
Tokundola te uto sanza ya zomi moleki batangaki bakoni nkoto yoko na baweyi zomi na mibale	tokundola te uto sanza ya zomi moleki batangaki bakoni nkoto yoko na bawe izomi na mibale