Voice Interaction in Moore Language Study on Isolated Word Recognition in Audio Samples

Moumini KABORE[1], Rodrique KAFANDO[2], Aminata SABANE[2], Abdoul Kader KABORE[2], Tégawendé F. BISSYANDE[2]

{ka.munkab@gmail.com[1], rodrique.kafando@citadel.bf[2], aminata.sabane@ujkz.bf[3], abdoulkader.kabore@citadel.bf[4], tegawende.bissyande@citadel.bf[5]}

CITADEL-Université Virtuelle du Burkina Faso[1]

Département d'Informatique, UFR/SEA-Université Joseph KI-ZERBO[2]

Abstract. This paper explores the optimization of telephone functionalities through voice interaction in the Moore language, prevalent in Burkina Faso. Data gathered from 492 individuals in Ouagadougou, representing diverse dialects and vocal intensities across age groups, informs the study. Employing K-Nearest Neighbor (KNN), Random Forest (RF), and Recurrent Neural Networks (RNNs), the analysis focuses on 29 Moore language commands, prioritizing practicality and user interaction. The findings suggest promising prospects for RNNs, achieving a 63% accuracy in recognizing isolated words. This success hints at potential advancements in RNNs, incorporating attention mechanisms and end-to-end technology, catering to the voice-controlled mobile device needs of Moore speakers.

Keywords: Isolated Word Recognition, RNN, Mel-Frequency Cepstral Coefficient

1 Introduction

Based on the 2021[1] and 2023[2] reports on digital and social media in Burkina Faso, it is observed that in a population of 22.96 million, there are 25.86 million mobile telephone subscribers, with 84% of the traffic routed through phones equipped with the Android OS. Additionally, 33% of Burkinabè have a Mobile Money account, and 95% of the 2 million social media users are on Facebook. According to the fifth report of the General Population and Housing Census[3], 52.9% of Burkina Faso's total population speaks Moore, a language also spoken in neighboring countries such as Ghana, Côte d'Ivoire, and Mali [1]. So we're seeing a lot of people with Android OS phones

[1]https://rb.gy/cxqq2x
[2]https://rb.gy/5szone
[3]https://rb.gy/bgk2th

JRI 2023, December 18-20, Ouagadougou, Burkina Faso

DOI 10.4108/eai.18-12-2023.2348169

doing a lot of things with those phones.

Given the prevalence of Android-equipped phones and the diverse functionalities performed on them, this study explores the application of machine and deep learning techniques to leverage common telephone features through voice interaction. The research aims to (i) identify a list of isolated words in the form of telephone commands, (ii) collect these words in audio format from phones, (iii) explore techniques for recognizing these isolated words in audio sequences, and (iv) implement isolated word recognition when spoken into a microphone.

According to the report on digital and social media in Burkina Faso in 2021[4] and 2022[5], out of a Population of 22.96 million, there are 25.86 million mobile Telephone subscribers - 84% of the traffic passed on phones equipped with Android OS - 33% of Burkinabè have an Orange Money account and - 95% of Facebook users out of 2 million social media users.

According to the fifth report of the General Population and Housing Census [6], 52.9% of the total population of Burkina Faso speaks Moore. This language (moore) is also spoken in neighboring countries such as Ghana, Côte d'Ivoire and Mali [1].

We witness that a large number of persons own telephones equipped with an Android operating system and complete a certain number of functionalities with these telephones. We therefore wonder how Machine and deep learning techniques could make it possible to exploit the most common features of telephones through the interaction of human voices.

The goals pursued by this research work are (i) identifying the list of isolated words under the form of commands to the telephone, (ii) collecting these words in audio format from the telephone, (iii) identifying a technique for recognizing these isolated words in audio sequences by exploring the multiple existing techniques and (iv) implementing an isolated word recognition, when uttered from a microphone.

2 State of the art

Automatic Speech Recognition, commonly referred to as ASR (Automatic Speech Recognition) is a computer technique that allows the analysis of a word or phrase captured by means of a microphone to transcribe it in the form of a text that can be used by a machine[7]. The field of RAS, which has largely occupied researchers since the 1950s, has benefited from the recent advances in deep learning and big data.

[4]https://rb.gy/fr183h
[5]https://rb.gy/5szone
[6]https://rb.gy/bgk2th
[7]https://rb.gy/o1qf8g

Nowadays, ASR has many uses and we can mention among others: language learning; voice commands, automatic subtitling, meeting comprehension and synthesis, TV remote control, etc.

Structurally, ASR attempts to solve three families of tasks [2]: (i) Voice Activity Detection (VAD) for the identification of speech in a speech sequence, (ii) Segmentation for identifying the beginning of a word in a voice sequence, (iii) for associating the meaning of each word pronounced in a sentence within a voice sequence. This latest family encapsulates Two sub-families: (a) speech synthesis for the generation of speech from another one, and (b) Recognition of single words for the association of meaning with a word spoken in a vocal sequence. Within the frame of this work, we will focus on this last subfamily of the family of Classification.

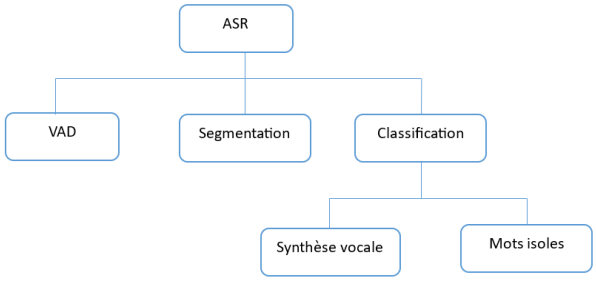

Fig. 1. Family of ASR Techniques

Recent advances in the family of Isolated word recognition tasks have encompassed a wide range of machine learning and deep learning algorithms. Within the frame of our research, we chose two classifiers of Machine Learning, KNN and RF, and a Neural Network-Based Deep Learning Classifier, RNN.

The choice of these classifiers can be accounted for by:

- the work of Muhammad Atif Imtiaz and Gulistan Raja [3] entitled Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW KNN using MFCC, KNN and DTW characteristics with a rate of precise prediction of isolated words around 98.4%. In view of the main advantages of RF such as (i) the reduction of risk of over adjustment given that the average number of trees decorrelated reduces the overall variance and prediction error. (ii) the flexibility which guarantees accuracy when part of data is missing with feature bagging and (iii) easy determination of the importance of the function, we chose to explore the RF in addition to the KNN. It should be noted that the choice of MFCC[8] features aligns with our Balancing Collection approach which balances[9] out classes of words to be predicted.

- DIPANWITA PAUL's work entitled AUTOMATED SPEECH RECOGNITION OF ISOLATED

[8]Mel-Frequency Cepstral Coefficients

[9]MFCC is very sensitive to class imbalance for prediction tasks

WORDS USING NEURAL NETWORKS presenting a methodology for neural networks automatic recognition of isolated words regardless of the speakers. The researcher used an eristic characteristic vector consisting of a combination of the first three formant frequencies of the vocal tract and the mean rate of passage to zero (ZCR) of the audio signal. The level of accuracy indicates that the set of characteristics performs better than anywhere in contemporary works in the existing literature.

3 Methodological approach

3.1 Data collection

The data used in our work come from traders and students of the city of Ouagadougou. We mobilized three (03) students over a period of seven (07) days to interview 492 interviewees with a questionnaire numbered with KoboToolbox[10] technology among which, 27% speak only the Moore language and 73% speak another language in addition to the Moore language. It should be noted that the majority of respondents to this study are between the ages of 18 and 38 (96%). A total of 29 commands in the Moore language were previously established on the basis of frequencies of use of the functionalities of the telephones that have been collected of which we will analyze some waveform and spectrum-shaped graphs.

Table 1: List of Services in Moore Language

Services	**Command in Moore**
Dial a number	Boole
Send Orange Money	Toole Orange Mone
Orange Money deposit	Digle Orange Mone
Send Moov money	Toole Moov Mone
Moov Money deposit	Digle Moov Mone
Launch Youtube	Pake Yutube
Launch whatsApp	Pake WhatSapp
Send WhatsApp vocal	Toole vocal
Launch Facebook	Pake FaceBook

It should be noted that the limitation of this list of commands [4] is simply conventional, it may be extended to the limit of the actual needs for voice commands. We have benefited from a well-balanced database as shown in Figure 4 with average-quality data analyzed with wave graphs and spectrogram of a sample of audios.

[10] A free and open source tool developed by Harvard Humanitarian Initiative & Brigham And Women's Hospital to cater for data collection needs with mobile devices in crisis and natural disaster environments.

Table 2: List of Numbers in the Moore Language

Values	Command in Moore
Zero (0)	Zaalem
One (1)	Yemele
Two (2)	Yiibu
Three (3)	Taabo
Four (4)	Naasse
Five (5)	Nu
Six (6)	Yoobe
Seven (7)	Yopue
Eight (8)	Nii
Nine (9)	Wouai

Table 3: List of values in the Moore Language

Values	Command in Moore
Five thousand (5 000)	Tusri
Ten thousand (10 000)	Tussa yiibu
Fifteen thousand (15 000)	Tussa taabo
Twenty thousand (20 000)	Tussa naasse
Twenty-five thousand (25 000)	Tussa nu
Thirty thousand (30 000)	Tussa yoobe
Thirty-five thousand (35 000)	Tussa yopue
Forty thousand (40 000)	Tussa nii
Forty-five thousand (45 000)	Tussa wouai
Fifty thousand (50 000)	Tuss piiga

3.2 Methods

For the prediction of isolated words spoken in the audio sequences, we used the Mel-Frequency Cepstral Coefficient (MFCC) characteristics for two types of machine learning algorithms (KN and RF) as well as spectrograms for a deep learning algorithm using neural networks (RNNs). We have chosen to optimize the parameters of our KNN models and RF using the Grid Search technique.

1. *KNN [5]:* The K-nearest neighbor method is one of the monitored / supervised learning algorithms ,which can be used for a regression and classification problem. However, it is still mainly used for classification problems. This algorithm trains on 'labeled' data and attempts to predict the class associated with variables by calculating the distance between the explanatory test variable and the training points. In other words, this method records the training points and then classifies them depending on their similarity to the recorded data.

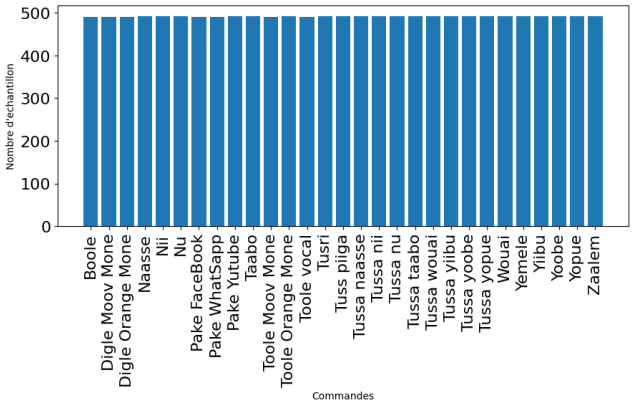

Fig. 2. Distribution of the number of audios by command class

2. *RF [6]:* random forest formed by random forests is a learning technique based on decision trees. It includes several trees, and as such it is part of the methods of set learning, i.e. Methods that use the wisdom of crowds[11] .These methods were first proposed by Ho in 1953 and formally proposed in 2001 by Leo Breiman[12] and Adele Cutler. This algorithm combines the concepts of bagging (parallel set methods) for the data selection phase, and for random subspaces. The decisional tree forests learns on multiple deciduous trees carried on slightly different subsets of data.

3. Recurent Neural Network [7]: is an artificial neural network with recurring connections . A network of recurrent neurons is made up of interconnected units (neurons) interacting in a nonlinear way, and for which there is at least one cycle in the structure. The units are connected by arcs (synapses) that have a weight. The output 'of a neuron is a nonlinear combination of its inputs.

4 Results and discussion

Our research shows that deep learning models using neural networks are better suited to the recognition of isolated audio words. We note the overall precisions (Accuracy) of *9%*, *21%* and *63%* respectively for *K-Nearest Neighbour*, *Random Forest* and *Recurent Neural Network*. In order to refine the likely errors that Our models could produce, we explored the confounding matrices of our models.

[11] The idea that a high number of beginners/ amateurs may better respond to a question than a single expert
[12] Renowned statistician at the University of California

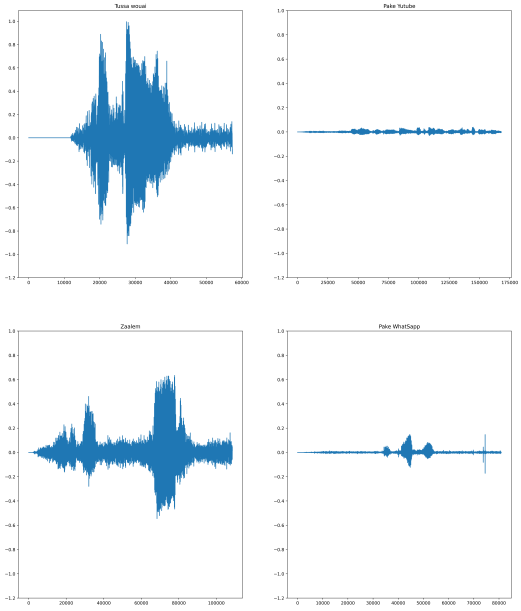

Fig. 3. Audio Waveform Commands: Tussa wouai (45,000), Pake Youtube (Launch Yutube), Zaleem (0) and Pake WhatSapp moving clockwise from top to left

Noticeably, this representation allows us to better understand the disparity in the prediction of isolated words which will be very useful to us in the future tasks of clearing our data set. To account for the relatively poor performance of our models, we make the following assumptions:

- the quality of the audios collected: a sampling of the 14268 audio data reveals sequences in which the voices of the data collector and the interviewee stand out clearly;
- the architecture of the models used.

In order to improve the prediction quality of these models, we plan:

- a clearance of the dataset of audios by isolating empty audios and those with a mixture of the voices of data collectors and interviewees, for KNN and RF model retraining;
- A careful exploration of RNN models with a view to properly isolating the portion of the isolated word in the audio sequence;

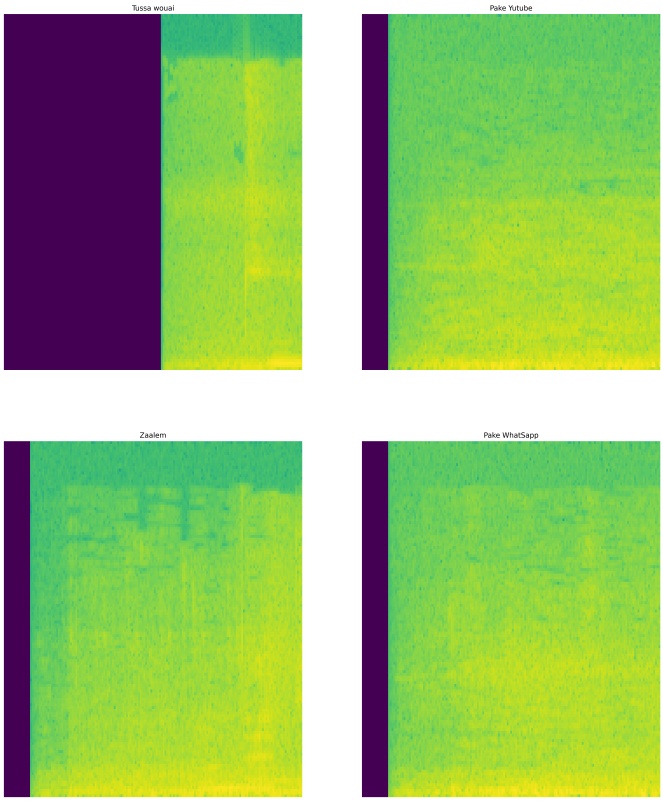

Fig. 4. Spectogram shape of audio commands: Tussa woui (45,000), Pake Youtube (Lancer Yutube), Zaleem (0) and Pake WhatSapp moving clockwise from top to left

- an exploration of the latest end-to-end techniques for speech synthesis for the detection of isolated words, this will not require a clearance of the dataset.

In practice, it should be noted that the best model resulting from our drive will be converted into *Tensorflow Lite format*, the ultimate format for devices such as phones equipped with the Android/iOS operating system. An Android/iOS application will be developed to 'embark' the model which will be responsible for predicting the text of the command spoken by the user through the

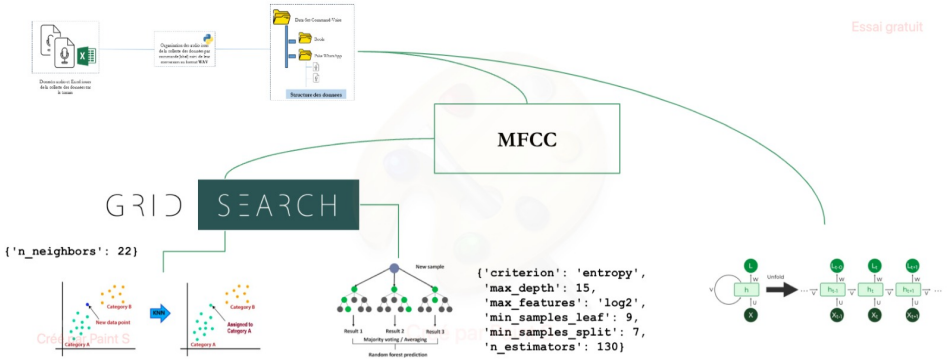

Fig. 5. Model Training data preparation Pipeline

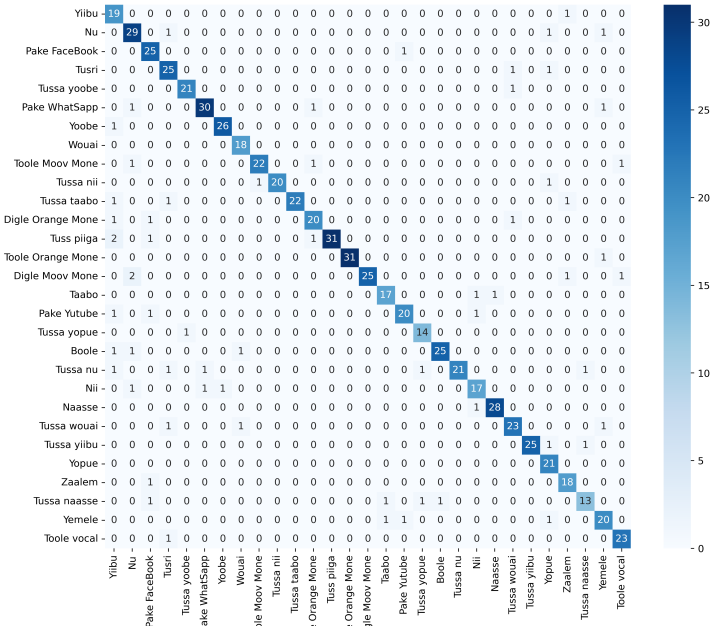

Fig. 6. RNN Confusion Matrix

phone's microphone, which will be used by an internal program to 'route' on the service desired by the user.

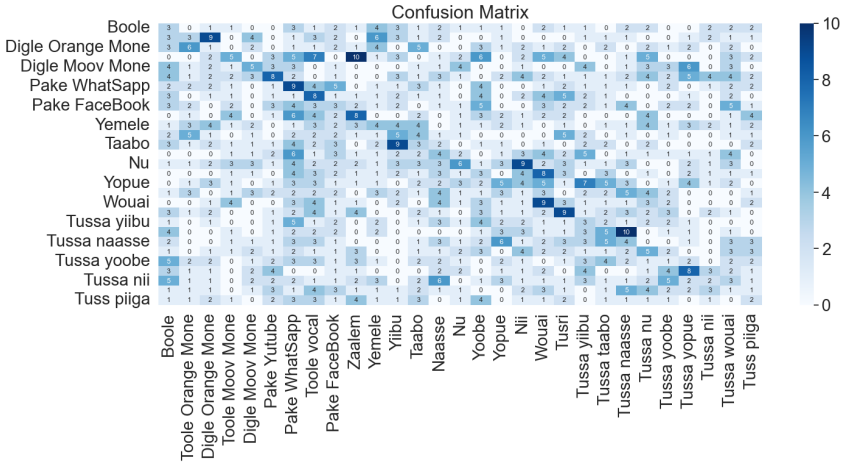

Fig. 7. KNN Confusion Matrix

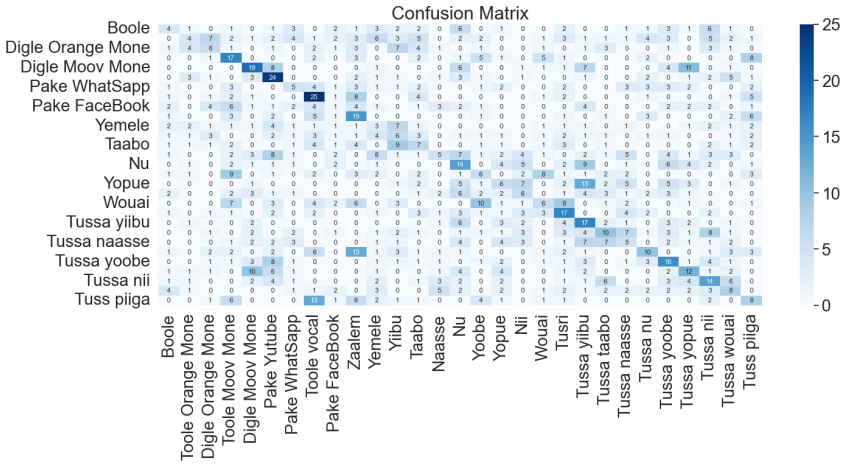

Fig. 8. RF Confusion Matrix

5 Conclusion

This study on the recognition of isolated words in the Moore language has revealed significant advances made in the field of applying artificial intelligence techniques for speech recognition. In using machine learning and deep learning methods, including Recurrent Neural Network (RNN), we have been able to achieve an accuracy of 63% in isolated word recognition, which is a promising result for voice interactions in the Moore language on mobile devices.

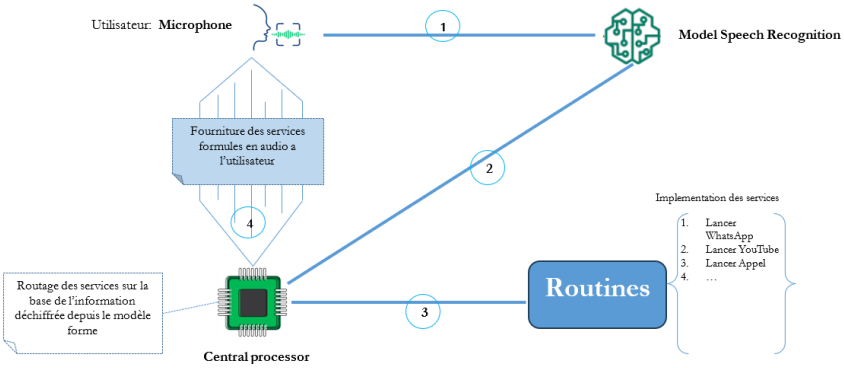

Fig. 9. Global view of the System with the model

The findings highlight the importance of a continuous adaptation of AI models in order to better recognize the linguistic and dialectal nuances of Moore language. The way is now open for in-depth research on RNN models enhanced by attention mechanisms and end-to-end techniques, aiming at optimizing speech synthesis for underrepresented languages. These advances represent a crucial step towards a more inclusive mobile technology, opening up new perspectives for the integration of diverse languages into Voice recognition systems. This research thus contributes to the broader goal of increasing accessibility and efficiency of communication technologies for communities speaking under-represented languages

Acknowledgments

We would like to thank the Virtual University of Burkina (UV-BF) and the Interdisciplinary Center of Excellence in Artificial Intelligence for Development (CITADEL) for their support and assistance during our internship.

References

[1] Leclerc J. Burkina Faso Aménagement linguistique dans le monde; 2015. Available from: https://www.axl.cefan.ulaval.ca/afrique/burkina.htm.

[2] XIAO X. Fundamentals of Speech Recognition; 2023. Available from: https://slpcourse.github.io/materials/lecture$_notes/XiaoGuest_lecture_ASR.pdf$.

[3] Imtiaz MA, Raja G. Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW KNN. IEEE Xplore. 2016:1-7.

[4] Association Motivilliers Nassere. Lexique français-moorè; 2013. Available from: http://www.montivilliersnassere.fr/attachments/francais_moore.pdf.

[5] Zhang Z. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine. 2016;4(11).

[6] Breiman L. Random forests. Machine learning. 2001;45:5-32.

[7] L H. "Explained: Neural networks. MIT News Office. 2022:1-10.