

Amharic Print Documents Transcription to Audio Using Machine Learning Techniques



Sinishaw Getachew Chala

A Thesis Submitted to
The Department of Computer Science and Engineering
School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's
in Computer Science and Engineering

Office of Graduate Studies
Adama Science and Technology University

September 2022
Adama, Ethiopia

Amharic Print Documents Transcription to Audio Using Machine Learning Techniques

Sinshaw Getachew Chala

Advisor: Dr. Tilahun Melake

A Thesis Submitted To
The Department of Computer Science and Engineering
School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's
in Computer Science and Engineering

Office of Graduate Studies
Adama Science and Technology University

September 2022
Adama, Ethiopia

Approval Sheet

I/we, the advisors of the thesis entitled “**Amharic Print Documents Transcription to Audio Using Machine Learning Techniques**” and developed by **Sinshaw Getachew challa**, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

_____	_____	_____
Major Advisor	Signature	Date
_____	_____	_____
Co-advisor	Signature	Date

We, the undersigned, members of the Board of Examiners of the thesis **Sinshaw Getachew chala** have read and evaluated the thesis entitled “**Amharic Print Documents Transcription to Audio Using Machine Learning Techniques**” and examined the candidate during the open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science Engineering

_____	_____	_____
Chairperson	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
<u>Eyob N.</u>		_____
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis are contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

_____	_____	_____
Department Head	Signature	Date
_____	_____	_____
School Dean	Signature	Date
_____	_____	_____
Office of Postgraduate Studies, Dean	Signature	Date

Declaration

I hereby declare that this Master Thesis entitled “**Amharic Print Documents Transcription to Audio Using Machine Learning Techniques**” is my original work. That is, it has not been submitted for the award of any academic degree, diploma, or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation.

Name of the Student

Signature

Date

Recommendation

I/we, the advisor(s) of this thesis, hereby certify that I/we have read the revised version of the thesis entitled “**Amharic Print Documents Transcription to Audio Using Machine Learning Techniques**” prepared under my/our guidance by **Sinshaw Getachew chala** submitted in partial fulfillment of the requirements for the degree of Master’s of Science in Computer Science and Engineering. Therefore, I/we recommend the submission of the revised version of the thesis to the department following the applicable procedures.

Major Advisor

Signature

Date

ACKNOWLEDGMENT

I am humbly grateful to my Lord for guiding me and helping me all the way through. I owe a special debt of gratitude to **Tilahun Melake (Ph.D.)**, whose guidance and critical advice, as my thesis advisor, greatly contributed to the making, letting me undertake this thesis and showing me lots of kindness whenever I visited his office.

My special gratitude goes to my best friend **Lishan E., Eyob G. and Dagem T.** who had been helping me with commenting and providing direction to my work.

Lastly, I would like to express my special thanks, but certainly not least to my family, especially my mother who helped me immeasurably with her encouragement, and kept me going at the most desperate moments, their continuous support and encouragements throughout my studies.

CONTENTS

ACKNOWLEDGMENT	iv
CONTENTS	ix
LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF ACRONYMS AND ABBREVIATIONS	xiv
ABSTRACT	xv
CHAPTER ONE	1
1. INTRODUCTION	1
1.1. Background	1
1.2. Statement of the Problem	3
1.3. Research Questions	4
1.4. Objectives	4
1.4.1. General Objectives	4
1.4.2. Specific Objectives	4
1.5. Significance of the study	4
1.6. Scope and Limitation	5
1.6.1. Scope of the Study	5
1.6.2. Limitations of the Study	6
1.7. Organization of the Thesis	6
CHAPTER TWO	7
2. LITERATURE REVIEW	7
2.1. Overview	7
2.2. History of Amharic language	8
2.3. Origin of the Amharic Writing System	9
2.3.1. Symbols of the Amharic Writing System (Fidälat)	11
2.3.2. Order Formation in the Amharic Writing System	16
2.4. Digitization of Amharic printed documents	17
2.5. Challenges In Building an OCR For Amharic Scripts on The Printed Documents	18
2.5.1. A large Number of Characters in the Scripts	18
2.5.2. printed documents Variations	18
2.5.3. Some Amharic letters are Similar	18

2.6.	Stages in Developing an OCR system	19
2.6.1.	Document Image Analysis and Recognition	20
2.6.2.	Reviews of Binarization Methods	20
2.6.3.	Skew Estimation	24
2.6.4.	Page Layout Analysis	25
2.6.5.	Text Recognition.....	27
2.7.	Deep Neural Networks	29
2.7.1.	ANN	29
2.7.2.	CNN	32
2.7.2.1.	CNN Architecture	32
2.7.2.2.	Convolution Layer.....	32
2.7.3.	RNN	33
2.7.4.	Speech signals and speech processing.....	34
2.7.5.	Bidirectional Recurrent Neural Networks	34
2.8.	Long Short-Term Memory Units (LSTM).....	35
2.9.	Historical Background of TTS	37
2.9.1.	Fundamental of Speech Synthesis.....	39
2.9.2.	Phonetic Analysis.....	39
2.9.3.	Speech Synthesis Techniques	40
2.10.	Related Researches in Local languages	41
2.10.1.	TTS for Afaan Oromo.....	41
2.10.2.	TTS for Tigrigna	41
2.11.	Related work.....	33
CHAPTER THREE		41
3.	METHODOLOGY.....	41
3.1.	Introduction	41
3.2.	General Architecture of the Proposed System.....	41
3.3.	Pre-processing Experimental Setup.....	42
3.3.1.	Image Binarization	42
3.3.2.	Image Skew Estimation	43
3.4.	Page Layout Analysis	43
3.5.	Training OCR Engine Setup	44
3.5.1.	Training Data Preparation	44
3.5.2.	Performance Evaluation	45

3.6.	Research Tools.....	45
3.6.1.	Design Tools.....	45
3.6.2.	Data Preparation Tools (Scanner)	45
3.6.3.	Programming Tools.....	45
3.6.4.	Frameworks	46
3.6.5.	Hardware Tools.....	47
CHAPTER FOUR.....		48
4.	EXPERIMENTAL RESULTS AND DISCUSSION	48
4.1.	Introduction	48
4.2.	Binarization Document Images.....	48
4.3.	Skew Document Image Estimation	51
4.4.	Comparison.....	51
4.5.	Discussion of Results	52
4.6.	Results and Discussion of The Page Layout Analysis	53
4.7.	Results and Discussion of Training the OCR Engine	58
4.8.	Training Data Preparation	58
4.9.	Running the Training Process.....	59
4.10.	Performance Evaluation	61
4.11.	Comparison.....	61
4.12.	Discussion of Results	62
4.13.	Text Analysis.....	62
4.13.1.	Syllables connection	63
4.13.2.	Evaluation and Discussion.....	63
4.13.3.	Speech Materials	63
4.13.4.	Methods.....	64
4.13.5.	Results and discussion.....	64
CHAPTER FIVE.....		66
5.	CONCLUSIONS AND FUTURE WORKS	66
5.1.	Conclusions	66
5.2.	Future Works	67
Reference.....		68

LIST OF TABLES

TABLE	PAGE
Table 2: 1 The core syllables in the Ge'ez writing system (Scelta, 2001).....	13
Table 2: 2 Labialized symbols (Labiovelars) in the Ge'ez writing system (አቤሴሎም, 2012 ዓ.ም)	14
Table 2: 3 Punctuation marks in the Ge'ez writing system (ተግባሩ, 2008 ዓ.ም).....	14
Table 2: 4 T Numerals in Ge'ez writing system (አቤሴሎም, 2012 ዓ.ም)	15
Table 2: 5 Composition and total number of symbols in the Amharic writing system (ተግባሩ, 2008 ዓ.ም)	16
Table 2: 6 High inter-class similarity of the AmharicTable	19
Table 2: 7 Research works on OCR for Ge'ez and Amharic scripts [17th May 1997 – 30th Sep. 2020].....	33
Table 4: 1 Overall evaluation results and the final ranking of the binarization methods..	50
Table 4: 2 Evaluation results of binarization for each test image with respect to the metrics used.....	50
Table 4: 3 Frequency table of the ground truth skew angle distribution	51
Table 4: 4 Evaluation results and comparison of skew estimation methods	52
Table 4: 5 Page layout evaluation results of success rate for each document	58
Table 4: 6 Performance measure evaluation of Amharic synthesizer	64

LIST OF FIGURES

FIGURE	PAGE
Figure 2: 1 language family of the Ge'ez language (አቤሴሎም, 2012 ዓ.ም)	11
Figure 2: 2 Multilayer Perceptron	30
Figure 2: 3 Multilayer Perceptron, Highlighting the Feed Forward And Backpropagation	31
Figure 2: 4 One iteration of Gradient Descent	31
Figure 2: 5 Architecture of a CNN	32
Figure 2: 6 Convolution Operation.....	33
Figure 3: 1 Architecture of the proposed for Amharic Print Documents Transcription to Audio Using Machine Learning Techniques system.....	42
Figure 4: 1 Example of binarization results of Adaptive Thresh Mean Threshold method	51
Figure 4: 2 Page layout description of each document image in the testing set.....	55
Figure 4: 3 Sample of text line detection in Image1.....	56
Figure 4: 4 Sample of word detection	56
Figure 4: 5 Ground truth of Image3 created manually using PyQt	57
Figure 4: 6 Sample of text line images with their corresponding ground truth.....	59
Figure 4: 7 Comparison of OCR performance by the base and new models	62

LIST OF ACRONYMS AND ABBREVIATIONS

AED	Average Error Deviation
ANN	Artificial neural networks
AVX-512	Advanced Vector Extensions 512
BRNN	Bidirectional Recurrent Neural Networks
CER	Character Error Rate
CNN	Convolutional Neural Network
EMIP	Ethiopian Manuscripts Imaging Project
EMML	Ethiopian Manuscript Microfilm Library
G2P	grapheme-to-phoneme conversion
GT	ground-truth
GUI	graphical user interface
IVR	Interactive voice response
LSTM	Long short-term memory
NLP	Natural language processing
NN	Neural Network
OCR	Optical character recognition
RGB	red-green-blue
RNN	Recurrent neural network
RNN	Recurrent Neural Networks
TD-PSOLA	Time Domain Pitch Synchronous Overlap
TTS	Text-to-Speech

ABSTRACT

Printed document recognition of remains a largely unsolved problem in the field of pattern recognition. This thesis examines how modern deep learning techniques can improve printed document recognition in the context of the Amharic language.

Although Amharic was a literary language in Ethiopia, it is underrepresented in the research field of document image recognition and analysis. Therefore, a printed document recognition based on real-world large-scale digitization scenarios is proposed

. Its architecture is made up of functions: pre-processing (binary and discriminant estimation), page layout analysis, model recognition, and post-processing. For each task, a test setup is prepared. In the binarization task, four binarization methods (Otsu's global method, Otsu's local method, Adaptive_Thresh_Mean_method, and Adaptive_Thresh_Gaussian method) were investigated using FM, p-FM, PSNR, and DRD evaluation parameters. Adaptive_Thresh_Mean_method is superior to all other methods in all dimensions.

In the document image bias estimation task, a method based on the Howe transform was investigated by testing and investigating its effect on the dataset. The evaluation parameters CE, AED and TOP80 were used and values equal to 86.00,0.4215and 0.068 were obtained. The performance of the Leptonica open-source C library was investigated in the page layout analysis function, and the results were found to be highly successful in various formats at the region and line level.

Amharic language page layouts. The final test setup was designed to build a recognition model using the Tesseract OCR engine. Since it was difficult to prepare large training data with realistically accurate printed documents, a calibration method based on the Amharic language was proposed and implemented. A total of 286 images of 200 lines of text collected from 15 different pages were processed and a recognition model with a character error rate of 2.632% was constructed.

Overall, the experiments conducted with the prototype approach have yielded encouraging results, so the full development of an OCR system for Amharic printed document recognition is applicable and vital for multi-purpose task specially for education purpose.

Key word OCR system, recognition, page layouts.

CHAPTER ONE

1. INTRODUCTION

1.1. Background

One of the fields of computer science under research is machine learning. Understanding and processing human languages is the focus of NLP, a branch of machine learning. Many systems exist now that can read and pronounce a variety of languages, including English, Spanish, French, Japanese, and other tongues that are widely spoken throughout the globe. We use printed papers in the field of machine learning.

Ethiopia is wealthy in a variety of ways. The nation also contains a large number of scripts written in Geeze and Amharic. In addition to libraries, we may find many manuscripts written in Amharic, Geeze, and Arabic when we visit historic churches and mosques.

The users are the ones who can and can access them because they are printed in papers. However, folks who cannot read for a variety of reasons need technology that can turn these documents into audio files utilizing deep learning. We can make it accessible to everyone, including those who are unable to read due to a variety of conditions, such as a vision issue. Additionally, by using this technique, we can safeguard these manuscripts against damage.

Recognition of printed documents is a difficult task. due to the quantity of classes, the scarcity of annotated data sets, and the high caliber of the documents. challenge for printed historical, scientific, and other materials because to the poor document quality and the numerous complicated page layouts¹. A function may be used to identify printed text in order to make a duplicate of a certain document. Copying is a laborious operation that needs patience and expertise. Consequently, there is a rising need for pattern recognition to streamline the creation of this text.

However, the difficulties posed by their diversity make the procedures complicated. Therefore, automatic document identification has been a crucial area of study in document and pattern recognition from the dawn of time. It has also been a significant issue for many years, but only recently has it come to the public's attention because of the potential benefit of having access to data held in printed papers.

¹ https://cedar.buffalo.edu/papers/articles/Online_Offline_2000.pdf

Document detection uses a variety of pattern recognition methods for characterisation and classification tasks, including template matching, statistical, structural, and syntactic methods (Fischer, A. (2012). However, more recently, machine learning techniques inspired from statistical teaching theory have expanded their focus on pattern representation.

Machine learning, as opposed to straightforward artificial neural networks, is used both for mapping and for learning to represent. Representation education is the name given to this strategy (Goodfellow, I., Bengio, Y., & Courville, A. (2016). Delegations with education frequently outperform those with written documents. Therefore, the requirement for a distinctive hand-drawn behavior approach is diminished by the potential to develop powerful automated behaviors in deep learning. Machine learning methods and Google's recently announced multilingual Tesseract OCR engine have been used. Image placement, object detection, video editing, natural language processing (NLP), and speech recognition are some more applications that have been effective in higher education (Goodfellow, I., Bengio, Y., & Courville, A. (2016).

Generally speaking, machine learning techniques enable a computer to learn on its own by identifying patterns in unstructured data. using RNN and CNN methods. The issue enables the learning system to understand the connections between input data and raw pixels while recognizing printed documents.

The formulated of recognition is described as a series of pattern categorization. The issue is that 2D (two-dimensional) digital printed document pictures that have been optically scanned are being utilized as raw data. This investigation has two (2D) major obstacles: Variability in printed documents and low-quality processing.

The main difficulty in recognizing printed documents is the first one, printed document variety. It changes depending on the period, location, culture, etc. This type of individual variation is known as inter-variability. The printed papers, however, have a unique distinctive character that may differ for a number of reasons, including mechanical (such as the writing instrument or surface used), physical (such as disease or age), and psychological (e.g., excitement, anger, mood, etc.). These variations are known as inherent sources of differences and pose serious obstacles to accurate detection.

The second difficulty includes writing and writing-related media. The writing style is the result of several tools and resources working together. The text may be huge or tiny if the typewriter does not precisely color the letters, or it may be inaccurate if the printing paper is thin or thick.

Recognition issues with printed documents result from such variations in writing media quality. Similar to this, the manner in which writing is produced has an effect of its own.

1.2. Statement of the Problem

This study explores how OCR and TTS approach for document recognition and document conversion to audio operate to digitize and convert documents written in Amharic.

Ethiopia is a very prosperous country in many different ways. When you visit the National Archives and Library of Ethiopia, as well as historic churches and mosques, you may find a large number of printed materials that are written in Amharic and Geez. Users who have the capacity and access to read them are the only ones who can access printed papers. However, individuals who cannot read for a variety of reasons do not have access to it. If we don't convert these documents into audio, we'll never be able to make them accessible to everyone, including those with reading disabilities brought on by conditions like poor vision. By using OCR and TTS, we can do this. Moreover, we can shield the papers from any damage brought on by things like rain, sunshine, etc.

Nobody used a performance-based evaluation method for the pre-processing. The tasks of binarization and skews estimation have so far been documented in Amharic and Geez. Additionally, earlier OCR studies did not take page layout analysis into account. since OCR the text and non-text portions of the scanned pages are frequently insufficient. It is necessary to conduct a thorough analysis of the regional level and page division page layout.

We have to read that a number of European nations have perfected the capacity to convert written papers into audio. However, there is some work to be done in this area in our nation. There are numerous printed papers in Amharic, as is well known, but they are only accessible to individuals who can read. For people who are unable to read for a variety of reasons, it is preferable to have these materials available in audio form to circumvent the issue.

1.3. Research Questions

The study looks into how printed documents are recognized and converted into audio using deep learning methods in the context of Amharic language. The research aims to respond to the following queries.

1. What appropriate pre-processing techniques should we use to improve the quality of image to improve recognition performance?
2. What kind of suitable page layout analysis methods to apply in order to extract the region of interest to proceed with OCR?
3. How to adjust and improve the ability to a capacity of deep neural networks?
4. To what extent does the prototype image recognition perform?

1.4. Objectives

1.4.1. General Objectives

The main objective of the study is to Amharic Print Documents Transcription to Audio Using Machine Learning Techniques by accepting the Amharic text as an input.

1.4.2. Specific Objectives

The primary objectives of this research were to get a complete understanding of all of its many components and achieve the following specific objectives. Additionally, it tried to remain committed to answering the research questions and achieving the study's ultimate objective.

- Study the OCR technology, in general, and its development tools, techniques, procedures, and methodologies in particular giving special
- Perform pre-processing tasks (binarization and skew estimation) on degraded printed documents in order to improve the recognition process.
- Perform page layout analysis and to study the effectiveness of page segmentation and region classification.
- Build a deep neural network model for printed document recognition of Amharic language.
- Evaluate the performance of the built model for Amharic printed document recognition.

1.5. Significance of the study

The study is important in many sectors and individuals. Digital libraries and archives are the most important of them. Additionally, there are several advantages to the research. End users

will get access to the Amharic documents to change by OCR and TTS system for a variety of applications, such as assistive technology for those who are visually impaired, when the character recognition and voice synthesizer have been implemented. The end-user gains access to a variety of potential applications by using such a synthesizer, some of which are as follows:

- I. Is used to facilitate minimized space in replacing huge collections of documents and electronically preserve irreplaceable materials in archives.
- II. Aid to disabled persons
- III. Language education
- IV. Talking books and literature (narration)
- V. Multimedia, man-machine communication
- VI. It helps for Fundamental and applied research in this area

Any of the aforementioned apps can be used, particularly for individuals who are visually impaired and persons with reading problems, as assistive technology after designing and implementing the OSR and TTS system for Amharic language.

1.6. Scope and Limitation

1.6.1. Scope of the Study

This study prototype only takes into account the printed texts that have been turned into audio systems for the Amharic language. Before pre-processing with OCR, the study looked at the pre-processing steps of binarization and skew estimation using page layout analysis. The Adaptive Thresh Mean technique, the Adaptive Thresh Gaussian method, the Otsu's local method, and the Otsu's global approach were all studied utilizing an experimental assessment method on a number of tests. Nine (9) exemplary photographs make up the document.

A method based on the Bicubic interpolation was examined in the document image bias estimation task by conducting an experiment and evaluating the outcomes on a dataset. There are 200 distinct skew angle samples in the dataset. Leptonica, an open-source C library for effective document image analysis tasks, performs best when given printed documents with five (5) pages of complicated graphics. 257 lines of text from 15 distinct sites were prepared as training examples with deep learning applications during the development of the recognition model. The pre-processed data will be turned into audio using the audio segmentation approach after which the speech will be produced by concatenating the syllables. we conduct a systematic study on the

current TTS systems, which can help us to find the problems, and is equally important (if not more) than solving the problems.

1.6.2. Limitations of the Study

The study's main limitation is that it doesn't discuss how to select a deep neural network's architecture for the task of printed document recognition. In other words, it doesn't offer analysis or recommendations for deciding which architecture to employ for the printed text depending on the situation. Additionally, it doesn't take into account language models or other post-processing chores when doing recognition.

Developing a deep neural network to address the issue of printed recognition. In other words, it doesn't offer analysis or recommendations for deciding which architecture to employ for the printed text depending on the situation. Additionally, it doesn't take into account language models or other post-processing chores when doing recognition.

1.7. Organization of the Thesis

The backdrop of the study is introduced in Chapter 1 of this thesis report, which is followed by the explanation of the statement of the problem, objective, importance of the investigation, the scope of the study, limitations of the study, and Organization of the Thesis. In-depth examinations of the literature and related works are used in Chapter 2 to examine several academic publications on document image analysis, OCR research, transcription of Amharic scripts, and theoretical foundations of deep neural networks. The method used to accomplish the aim and the specified experimental settings are extensively described in Chapter 3. Results and a discussion of the experimental findings are presented in Chapter 4. Chapter 5 concludes the discussion and outlines potential future extension projects.

CHAPTER TWO

2. LITERATURE REVIEW

2.1. Overview

When we attempt to digitalize a large number of documents particularly printed papers, what comes first in our mind is character recognition. It is very significant for our issue. Some research shows that now days, new technology demand is incising highly for document analysis and identification purpose.

In order to solve these problems, to convert text files in to digital documents numerous solutions have been introduced and a number of research initiatives have been carried out.in the field of OCR. Despite the fact that, these solutions have been processed printed documents, in terms of the low quality of handwritten papers are unsatisfactory.

When you come to Amharic language the focus area of this research is to change the Amharic printed documents in to digitalization. Both the printed and handwritten papers in Amharic on script are of poor quality. These tools could not create Amharic scripts as a consequence of handwriting variations and low quality. The goal of this study is to create an OCR and TTS system for printed documents.

On the other hand, there has been tremendous advancement in the area of language technology in the field of speech science and technology, particularly in the integration of speech and voice recognition.

Speech has long been used as a communication tool and as a medium of communication. Organs ranging from the diaphragm and lungs to the vocal cords and nose are all part of the human speech production system (Henock, L. 2003).

A machine that communicates with people is the simplest way to define speech integration. Modern algorithms and technologies are used to produce fake sounds from recordings of natural sounds while maintaining the initial speaker's acoustic characteristics in text-to-speech technology (Alessandro, N. D. 2009).

As was said in Chapter 1, the major goal of this work is to develop and use a generic discourse model to make printed Amharic documents sound more authentic. With the aim of providing textual information to machine. through voice, speech synthesis is the process of translating text

from speech. this technology has the capacity to turn random text into audio (Dutoit, Thierry. 2008).

Additionally, human speech is based on artificial intelligence and is frequently utilized to aid disabled persons (T. Anberbir and T. Takara, 200) Discover and debate (Lemmetty, S. 1999), TTS Synthesis is a procedure that makes it easier for people to connect with machines for a range of applications, including screen readers, IVR systems, and hands-free information access. software for people with visual impairments.

Text analysis, in which the input text is converted into an audio representation, is the main category for speech integration text (i.e., standard), phonetic (i.e., G2P), and phonetic analysis (compression and intonation). Speech waves are developed and produced using a very complex process since the audio output is based on information from sound and hearing.

The ultimate goal of text-to-speech technology is to create programs that users and audiences can't tell whether from automated or human sources (Yamagishi, J., Kobayashi, T., Renals, S., King, S., Zen, H., Toda, T., & Tokuda, K. 2007), which can change the spelling. Separating it from human speech In TTS, the speech processor is frequently used. Depending on the design of the system, the modifier produces audible, organic noises. A number of linked articles will also cover the fundamental processes and techniques of OCR and TTS.

2.2. History of Amharic language

Amharic is an Afro-Asiatic language that belongs to the Semitic family and it is one of the Ethiopians Semitic languages. Next to Oromifa², Amharic is the primary tongue used in major Ethiopian cities and towns.

For a long time, Amharic was an official language spoken in Ethiopia. It has long served as the court's official language. The language of the palace since the end of the 12th century was Amharic, which is utilized for trade and communication. It is one of Ethiopia's official legal languages at the moment, along with Oromifa, Somali, Afar, and Tigrinya.

Amharic is derived from Geez which is the root language of the South African Semitic family of Afro-Asiatic languages. It is also a language used by the Tewahedo Ethiopian Orthodox Church. The letters used to write Geez are slightly different than those used to write Amharic. There are

² https://en.wikipedia.org/wiki/Languages_of_Ethiopia

34 fundamental letters in the Amharic alphabet, and each one has seven different forms depending on the vowel's gender.

Amharic is the language spoken by 21.6 percent of Ethiopia's population. More modern sources Nearly 32 million people in Ethiopia speak their first language as of 2019, while another 25 million people speak as their second language.³

Amharic is written left-to-right using a system that grew out of the Ge'ez script⁴. In Ethiopian Semitic languages, the writing alphabet is known as fidäl (ፈደል). Script, alphabet, “letter”, or “character” are all denoted (Abay Teshage, R. Sethuraman 2015) by the word Fidäl From the first four symbols, the writing system is also known as Abu Ghraib, Hence the word Abugida of the day⁵.

2.3. Origin of the Amharic Writing System

Although, there is considerable debate around this claim, many historians feel that the Ge'ez script is descended from the Epigraphic South Arabian script, which has Proto-Sinaitic roots. Other researchers also hold that the Ge'ez script is related to Egyptian hieroglyphics. The Ethiopian Orthodox Tewahedo Church has a belief that Enos, the grandson of Adam the First Man, had a divine revelation of the script. Ge'ez, however, is written from left to right in contrast to other Semitic scripts. (ተግባሩ, 2008 ዓ.ም).

The Ge'ez script has evolved through remarkable developments. For instance, vocalization of the Ge'ez occurred in the 4 centuries, and it is believed many other changes on the script (such as, change in writing direction, order formation, alphabet order change from አ-በ-ገ-ደ to ሀ-ለ-ሐ-መ, and new symbols for numerals) were also occurred during this time. Many scholars believed the symbols for numerals (፩ ፪ ፫ . . .) are derived from the Greek (አቤሴሎም, 2012 ዓ.ም). According to the language family of contemporary linguistics study, Ge'ez is grouped in the parent Afro-Asiatic category. Many notable scholars in linguistics, such as Baye Yimam, Dessie Qeleb, Yoseph Greenberg, and etc. agree with this classification. Based on this classification, the following figure shows Ge'ez and some other Ethiopian languages which are grouped in the parent Afro-Asiatic category (አቤሴሎም, 2012 ዓ.ም).

³ https://en.wikipedia.org/wiki/Languages_of_Ethiopia

⁴ <https://library.bu.edu/amharic/Home>

⁵ <https://en.wikipedia.org/w/index.php?title=Amharic&oldid=1077727613>

According to Donald Levine, a language expert, the Afro-Asiatic family began in the eastern Sahara or southwest Ethiopia. Around the fourth or fifth millennia BC, Proto-Semitic, Proto-Cushitic, and Proto-Omotic languages began to be spoken in Ethiopia highland settlers. Soon after, speakers of Proto-Semitic, Proto-Cushite, and Proto-Omotic languages crossed the Sinai Peninsula and arrived in Asia Minor.

The migration of people from southern Arabia to Ethiopia brought Semitic languages. Semitic speakers once lived in the empire about 500 BC. Semitic languages have been spoken in Ethiopia since 2000 B.C. this shows that they are there. Levine claims that around the turn of the century⁶, the majority of people living in Greater Ethiopia were Afrom-Mediterranean Afro-Asians who spoke Semitic, Cushitic, and Omotic.

Amharic is the outcome of the linguistic process of coding with the Cushitic sub division and the Semitic supervisor to facilitate communication between speakers of other languages, claims Lionel M. Bender. (H. Yazid and H. Arof, 2008).

The number of Cushitic impacts from Proto-Ethio-Semitic languages on Amharic is comparable to that of Gutturals, analogous Cushitic influences in other southern Ethiopian languages, and a significant number of Cushitic languages (such as A Romanian influence that started in the 16th century) (Gurage, Tigrinya, etc.).

According to one theory, the Semitic-speaking people of Yemen first migrated in the Axum area of northern Ethiopia⁷, intermarried with the Agwu and other Cushites, and then slowly went south to their modern homeland of Amhara (Abay Teshage, R. Sethuraman 2015). One hypothesis is that the Semitic-speaking people of Yemen initially arrived in the northern Ethiopian province of Axum, intermarried with the Agwu and other Cushites, and then eventually moved south to their current homeland of Amhara (Abay Teshage, R. Sethuraman 2015).

6

https://books.google.com.et/books?id=TtmFQejWaaYC&pg=PA12&source=gbs_toc_r&cad=4#v=onepage&q&f=false

7

https://books.google.com.et/books?id=TtmFQejWaaYC&pg=PA12&source=gbs_toc_r&cad=4#v=onepage&q&f=false

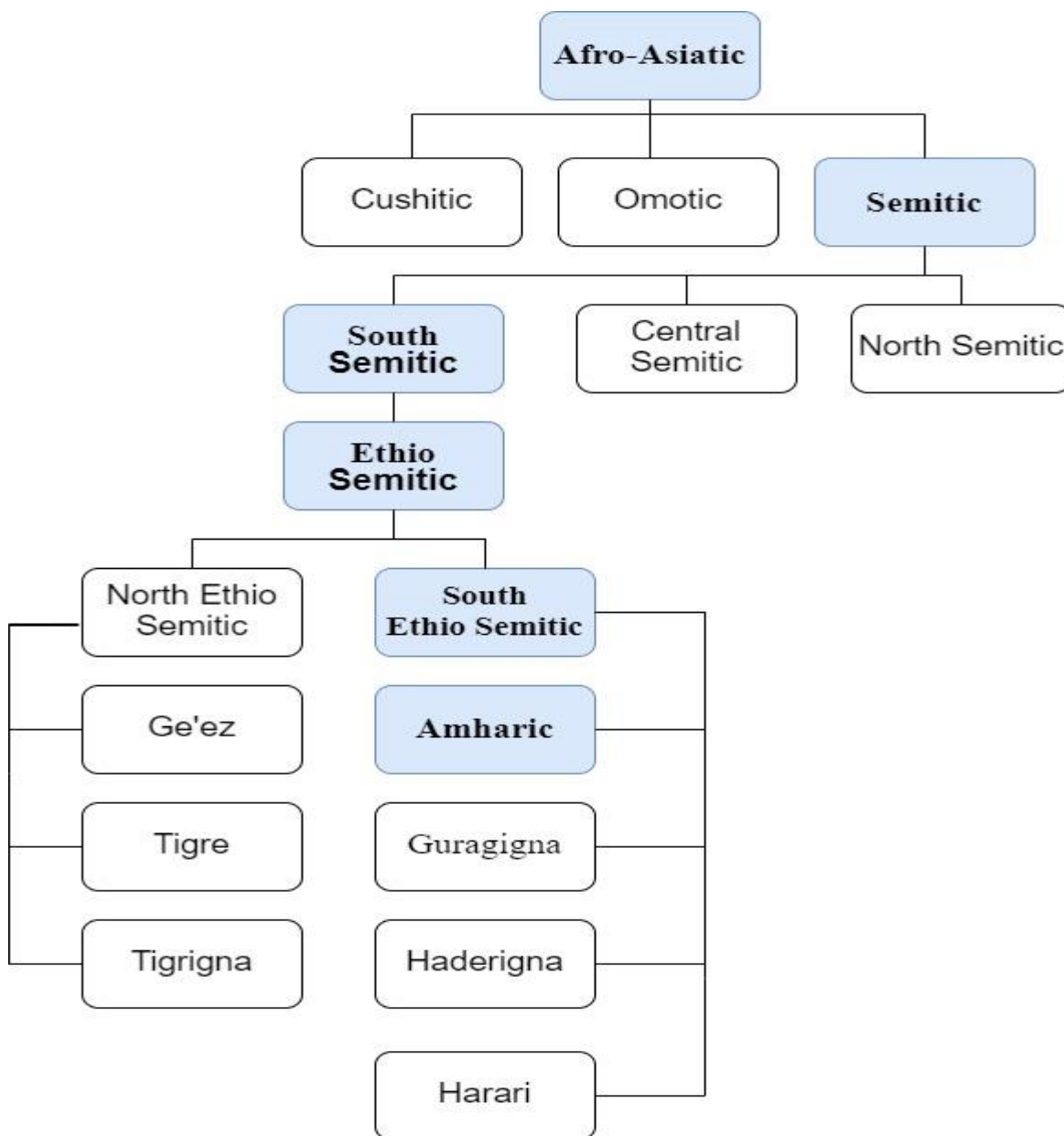


Figure 2: 1 language family of the Ge'ez language (አቢሰላም, 2012 ዓ.ም)

2.3.1. Symbols of the Amharic Writing System (Fidälat)

Abugida is the name of the Amharic alphabet, which is also known as the Graph alphabet. (D. Bradley and G. Roth, ,2011) The Geez alphabet is a modified version (R. Meyer, 2006) Some consonant typefaces (such as the /, /s, /ts, and /h/ series) have many character sets (the last four have different letters). This is due to the fact that these letters formerly stood for various sounds, but due to a shift in sound, they were combined.

The first column in the file, which is the consonant form ä, is the citation form for each series. The Amharic script is a part of Unicode, and all of the main operating systems' fonts support glyphs. This is the reason why the Amharic spelling system frequently uses alphabetical syllables rather than alphabets. The 34 important or fundamental symbols come in 7 distinct forms, as indicated in Table 2.1 below, bringing the total capital to 182. (34x7)⁸.

⁸ <https://en.wikipedia.org/wiki/Amharic>

Table 2: 1 The core syllables in the Ge'ez writing system (Scelta, 2001)

		ə	o	^w ä/ ue	^w i/u	^w a/ ua	^w ē/ uē	^w ə					
<i>a</i>	<i>ē</i>	[i], ø		[^w ə]	<i>i</i>	<i>ua</i>	<i>uē</i>	[^w i/ ū]					
<i>h</i>	/h/	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ					
<i>l</i>	/l/	ለ	ሉ	ሊ	ላ	ሌ	ሎ	ሎ			ሏ		
<i>h</i>	/h/	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ			ሐ		
<i>m</i>	/m/	መ	ሙ	ሚ	ማ	ሜ	ም	ሞ			ሚ		
<i>s</i>	/s/	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ			ሢ		
<i>r</i>	/r/	ረ	ሩ	ሪ	ራ	ራ	ር	ሮ			ረ		
<i>s</i>	/s/	ሰ	ሱ	ሲ	ሳ	ሴ	ሰ	ሶ			ሰ		
<i>š</i>	/ʃ/	ሸ	ሹ	ሺ	ሻ	ሼ	ሽ	ሾ			ሸ		
<i>q</i>	/k'/	ቀ	ቁ	ቂ	ቃ	ቄ	ቅ	ቆ	ቈ	ቀ	ቁ	ቂ	ቃ
<i>b</i>	/b/	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ			ቢ		
<i>v</i>	/β/	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ			ቪ		
<i>t</i>	/t/	ተ	ቱ	ቲ	ታ	ቲ	ት	ቶ			ተ		
<i>č</i>	/tʃ/	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቸ			ቸ		
<i>h</i>	/h/	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ	ኸ	ኀ	ኁ	ኂ	ኃ
<i>n</i>	/n/	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ			ና		
<i>ñ</i>	/ɲ/	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ			ኸ		
<i>'</i>	/ʔ/	አ	ኡ	ኢ	ኣ	ኤ	አ	ኦ			አ		
<i>k</i>	/k/	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ	ኸ	ከ	ኩ	ኪ	ካ
<i>x</i>	/h/	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ	ኸ	ኹ	ኺ	ኻ	ኼ
<i>w</i>	/w/	ወ	ወ	ወ	ወ	ወ	ወ	ወ					
<i>'</i>	/ʔ/	ዐ	ዐ	ዐ	ዐ	ዐ	ዐ	ዐ					
<i>z</i>	/z/	ዘ	ዘ	ዘ	ዘ	ዘ	ዘ	ዘ			ዘ		
<i>ž</i>	/ʒ/	ዝ	ዝ	ዝ	ዝ	ዝ	ዝ	ዝ			ዝ		
<i>y</i>	/j/	የ	የ	የ	የ	የ	የ	የ					
<i>d</i>	/d/	ደ	ደ	ደ	ደ	ደ	ደ	ደ			ደ		
<i>ǧ</i>	/dʒ/	ጀ	ጀ	ጀ	ጀ	ጀ	ጀ	ጀ			ጀ		
<i>g</i>	/g/	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ	ገ
<i>t</i>	/t/	ጠ	ጠ	ጠ	ጠ	ጠ	ጠ	ጠ			ጠ		
<i>č</i>	/tʃ/	ጠ	ጠ	ጠ	ጠ	ጠ	ጠ	ጠ			ጠ		
<i>p</i>	/p/	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ			ጸ		
<i>ʃ</i>	/ts'/	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ	ጸ			ጸ		
<i>ʃ'</i>	/ts'/	ፀ	ፀ	ፀ	ፀ	ፀ	ፀ	ፀ					
<i>f</i>	/f/	ፈ	ፈ	ፈ	ፈ	ፈ	ፈ	ፈ			ፈ		
<i>p</i>	/p/	ፐ	ፐ	ፐ	ፐ	ፐ	ፐ	ፐ			ፐ		

Symbols that have been labialized are also used in the Amharic alphabet. These are symbolized by the four fundamental forms short-wow, ከ።/kwä/, ኧ።/gwä/, /qwä/, and ከ።/hwä/. The laboratory symbols only contain five instructions, as opposed to the center symbols' seven commands as previously mentioned.

Table 2: 2 Labialized symbols (Labiovelars) in the Ge'ez writing system (አቤሴሎም, 2012 ዓ.ም)

	ä	i	ə	a	e
kw	ከ። /kwä	ከ። /kwi	ከ። /kwə	ከ። /kwa	ከ። /kwe
gw	ኧ። /gwä	ኧ። /gwi	ኧ። /gwə	ኧ። /gwa	ኧ። /gwe
qw	ቈ። /qwä	ቈ። /qwi	ቈ። /qwə	ቈ። /qwa	ቈ። /qwe
hw	ከ። /hwä	ከ። /hwi	ከ። /hwə	ከ። /hwa	ከ። /hwe

The Amharic alphabet is also composed of punctuation and numbers. Punctuation marks are basic words, 's a sentence-divider or full stop (።), and other symbols similar to the English comma (፣), semicolon (፤), and as a question mark (?), Exclamation mark (!), Quotation marks ("") and parentheses ().

Table 2: 3 Punctuation marks in the Ge'ez writing system (ተግባሩ, 2008 ዓ.ም)

Punctuation marks	Amharic	English
፡	ሁለት ነጥብ	Ethiopic Wordspace
።	አራት ነጥብ	Ethiopic Fullstop
፣	ነጠላ ሰረዝ	Ethiopic Comma
፤	ድርብ ሰረዝ	Ethiopic Semicolon

For 10 multipliers (10, 20, 30, 40, 50, 60, 70, 80, and 90), for 100 and 10,000, the numbering scheme uses 1 to 9 digits. Numbers, decimals, and zeros cannot be negative. In the numerical system of Amharic. Table 2.4 below shows the numbers.

Table 2: 4 T Numerals in Ge'ez writing system (አቤሴሎም, 2012 ዓ.ም)

Numerals	In Ge'ez words	In Amharic words (English transcription)
፩ - 1	አሐዱ	አንድ (ande)
፪ - 2	ክልኤቱ	ሁለት (hulät)
፫ - 3	ሠለስቱ	ሶስት (sost)
፬ - 4	አርባዕቱ	አራት (arat)
፭ - 5	ሓምስቱ	አምስት (amïst)
፮ - 6	ስድስቱ	ስድስት (sidist)
፯ - 7	ሰብዓቱ	ሰባት (säbat)
፰ - 8	ሰመንቱ	ስምንት (simint)
፱ - 9	ተስሃቱ	ዘጠኝ (zät'änn)
፲ - 10	ዓሠርቱ	አስር (asir)
፳ - 20	ዕሥራ	ሃያ (haya)
፴ - 30	ሠላሳ	ሰላሳ (sälasa)
፵ - 40	አርባዕ	አርባ (arba)
፶ - 50	ሓምሳ	ኃምሳ (hamsa)
፷ - 60	ስድሳ	ስልሳ (silsa)
፸ - 70	ሰብዓ	ሰባ (säba)
፹ - 80	ሰማንያ	ሰማንያ (sämanyä)
፺ - 90	ተስዓ	ዘጠና (zät'äna)
፻ - 100	ምእት	መቶ (mäto)
፪፻ - 10, 000	እልፍ	አስር ሺህ (asir ših)

The Amharic script generally consists of 286 characters for syllables, labialized, punctuation, and numbers. Table 2.4 following provides an overview of the composition and total number of symbols (Fidālat) in the Amharic writing system.

Table 2: 5 Composition and total number of symbols in the Amharic writing system (ተግባሩ, 2008 ዓ.ም)

No.	Type of symbols	No. of symbols
1	Core symbols (34 x 7)	238
2	Labialized symbols (4 x 5)	20
3	Punctuation marks	8
4	Numerals	20
Total		286

2.3.2. Order Formation in the Amharic Writing System

The Amharic script comprises seven commands, as stated in Section 2.4. The other six commands are designed to express additional vowel sounds by adding further strokes and alterations to the fundamental forms, which are represented by Basic Forms (Initial Sequence), which is a representation of vowels. To put it another way, after the first sequence, all six of the remaining orders are specified by adding little links to the right or left, above or below, shortening or lengthening one of the major strokes, and returning to the first sequence for everything else. There are extremely few exceptions to commandments 6 and 7, which are followed by the second, third, fourth, and fifth commandments.

The second sequence is built by adding a horizontal stroke to the center of the basic shape to the right (e.g., ሀ from ሀ; ለ from ለ; and መ from መ). Similarly, the third order is created by adding a horizontal stroke at the base of the basic character of the basic character (e.g., ሊ from ለ; ቢ from ቢ; and ኪ from ከ). The fourth order is based on adding a one-legged character to the foot by adding a diagonal stroke or extending the right leg of a two- or three-legged character (e.g., ቸ from ቸ; ዳ from ዳ; and ጣ from ጣ). sequence is constructed from the ground by adding a ring to the lower part of the right leg (e.g., ኤ from ኤ; ኘ from ኘ; and ሐ from ሐ) (Worku, A. 1997).

All in all, the 2nd, 3rd, 4th, and 5th commands above are based on high standard patterns, others, 6 and 7, are very informal (Bender, Cooper, and Ferguson, 1972). The sixth sequence is constructed by adding a stroke, loop, or other shape on each side of the basic shape. Take, syllables ህ from ሀ; ል from ለ; ሞ from ሙ; ስ from ሰ; ር from ረ; and ገ from ገ. Similarly, the seventh sequence is formed from the basic shape by lengthening the left leg or adding one turn at the top or right. For instance, syllables ሆ from ሀ; ሎ from ለ; ቆ from ቀ; ሶ from ሰ; ሞ from ሙ; and ም from ወ.

It is essential to have a thorough grasp of the Geez OCR system's format as well as the methods available for using it when designing the system. According to recent study (Birhanu et al. 2019). (Fidālat) the Amharic OCR system was developed using multi-task learning factored CNN.

2.4. Digitization of Amharic printed documents

Scholars have had access to a significant number of manuscripts during the past forty years because to numerous microfilm and digitization projects. The Ethiopian Document Microfilm Library (EMML), which has 9, 9,238 manuscripts, is crucial when it comes to microfilms.

The Collection has expanded over the past 40 years as a center for the study, recording, digitization, and indexing of Ethiopian manuscripts. It is kept at the Hill Museum and Handwriting Library at St. John's University College in Ville, Minnesota. with others). He recently digitized a number of significant collections (e.g. the Gunda Gunde Monastery Library).

The British Archives Disaster Program (EAP10) and Arcadia Fund are funding additional digitization initiatives. Thousands of photographs and historical collections of significance are gathered by the University of Toronto's Mazgaba Seelat project. (EMIP), which was established in 2005, has digitized minor collections at academic libraries, vendors, and private owners largely in North America but also in the UK, Israel, and Kenya.

More than 2,000 excellent digital photographs of Ethiopian manuscripts have been added to the Ethio-Spare 11 Christian Ethiopian Cultural Heritage, which is funded by the University of Hamburg's Research Council (Zervas, P., Potamitis, I., Fakotakis, N., & Kokkinakis, G. 2001).

2.5. Challenges In Building an OCR For Amharic Scripts on The Printed Documents

It is difficult to perform OCR from document pictures printed in Amharic script. The following challenges must be resolved in order to create a functional OCR system for Amharic scripts.

2.5.1. A large Number of Characters in the Scripts

The first obstacle is caused by the number of characters in the class, and because is the total amount of symbols in the written letters. a total of 286 characters, including cells, numerals, and punctuation, omitting tone symbols.

The existence of so many symbols in the writing system makes it extremely difficult to create an effective and clever OCR system, we must consider the mathematics and logical criteria for such a vast number of signals.

2.5.2. printed documents Variations

The size of printed papers or printed script in a variety of sizes is one of the key challenges to understanding the Amharic alphabet.

There are variances in the same words published at various dates even in publications with a same author (Yoshiko Fujisaki, 1998). Amharic script varies in spelling, size, and style due to variations in print. Characters of the Amharic script are distinctive because of these fonts, styles, and sizes. Making the representation compatible with handwritten variants is challenging since it requires balancing size diversity by using standardized procedures and providing useful features.

Nearly always; Character Identification Problem: Because various printing machines provide different characters from the script that understands the shape of the same characters, The differences and distortions of print script are the only things that are being considered.

2.5.3. Some Amharic letters are Similar

The Amharic alphabet has a lot of identical characters, making it often difficult to distinguish between them. For each character in the right category or class, it's crucial to create strong discriminating behaviors. (Abay Teshage, R. Sethuraman 2015).

Table 2: 6 High inter-class similarity of the AmharicTable

Ge'ez <i>1st Order</i>	Ka'eb <i>2nd Order</i>	Salis <i>3rd Order</i>	Rab'e <i>4th Order</i>	Hamis <i>5th Order</i>	Sadis <i>6th Order</i>	Sab'e <i>7th Order</i>
ለ and ሰ	ለ and ሰ	ለ and ሰ	ለ and ሰ	ለ and ሰ	ለ and ሰ	-
ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ
ዐ and ፀ	ዐ and ፀ	ዐ and ደ	ዐ and ደ	ዐ and ደ	ዐ and ፀ	ዐ and ደ
ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ	ደ and ደ
ኀ and ኀ	ኀ and ኀ	ኀ and ኀ	ኀ and ኀ	ኀ and ኀ	ኀ and ኀ	ኀ and ኀ
ፈ and ረ	ፈ and ረ	ፈ and ረ	ፈ and ረ	ፈ and ረ	ፈ and ረ	ፈ and ረ

2.6. Stages in Developing an OCR system

Use a flat-panel, handheld scanner, digital camera, or smartphone with a wavelength sensor to first digitize the document. To create digital document pictures, the two-dimensional symbols that were collected are sampled and adjusted in size. Human perception is capable of analyzing, integrating, reading, and comprehending every piece of information present in digital document images, including texts, photos, and so on (Fitsum Demissie 2011, Lasko, Laskov, 2006). However, using an intelligent (OCR) technology is necessary to be able to decipher such information in document image.

Numerous scholars have suggested various OCR system operating procedures. (Lasko, Laskov, 2006) provide interconnected functional modules a system for handwritten text recognition. Pre-processing, which refers to the image representation input, is the initial phase. The second involves separating text lines from image in the document that were extracted when looking for text lines. The word division process, which includes separating words from text lines, comes next. The next phase is to create character and behavioral recognition algorithms for the papers. The language post-processing is the last step and is used to correctly assess how language limitations are used and to identify text. (Lasko, Laskov, 2006) The two primary phases of preparation and recognition are described in linear sequence by the traditional OCR systems. Sound removal, binary, bias, page analysis, and classification are all included in the preliminary stage, while behavior and classification are included in the level of cognition.

The aforementioned procedures have been used in the past to design a number of algorithms for the OSR system. However, in recent years, innovative systems based on machine learning approaches have combined many operating levels to reflect changes in the OCR system. Short-term memory (LSTM) networks, a form of repeating neural network (RNN), have seen the most advancement in behavioral identification because to recent developments in OCR technology

(W. A. Mustafa, H. Yazid, and S. Yaacob, 2011). For instance, these concepts are the only ones that can be used to create Google's recently announced Tesseract OCR engine. For historical documents, the OCR technology still presents a scientific challenge. Thorough analysis of the literature reveals that there is no set method for creating a document recognition system.

Therefore, in this study, allowing for a number of information processing is the recommended method for creating a printed recognition system. Following a thorough review of the literature, the printed recognition system architecture challenge is addressed using the following modular method.

The foundation of the suggested system is extensive digitization. Pre-processing (binary and bias estimates), page layout analysis (page division and area classification), model identification, and post-processing are some of the built-in functions that make up its architecture. These efforts are the subject of the parts that follow in this chapter, with an emphasis on the methodologies and procedures employed in the research.

2.6.1. Document Image Analysis and Recognition

When it comes to printed recognition for historical documents, the goal of creating machine reading systems that can read, interpret, and assimilate printed content is still very far off. (Shafii, M. 2014).

Strong OCR engines and document image analysis struggle to accurately recognize the text in the original printed documents due to the complex page structures and artifacts present. Researchers have previously employed binarization, noise reduction, and document image enhancement techniques to solve the deteriorating problem during pre-processing. However, the preference for using a single approach to solve these problems throughout the binarization process has changed recently. Different document image binarization methods have been developed in the contemporary period to enhance the quality and increase the resistance of printed documents. The next section goes into great depth on binarization.

2.6.2. Reviews of Binarization Methods

Some of the most common binary techniques applied in this study are described in this section. Binary processes frequently have an impact on poor image quality caused by r, illumination, camera artifacts, and image distortion

The process used to transform color photos into grayscale is frequently thought to have no impact on recognition performance in image recognition. We show that this assumption is incorrect by comparing four distinct types of picture descriptors with thirteen different grayscale algorithms: Even when utilizing some color-to-grayscale methods,

Robust to changes in illumination characteristics. These techniques are evaluated on face, object, and image datasets with a contemporary descriptor-based image recognition framework utilizing a small number of training dataset. We pinpoint two effective approaches for image recognition as well as a simple technique that performs well in general for face and object detection. (R. C. Gonzalez and R. E. Woods, 2008, M. A. Bakhshali, 2011).

a) Gradient Based Thresholding

By creating a home page, Haniza and Arof unveiled a novel binary technique that utilizes a gradient-based roof (H. Yazid and H. Arof, 2009). Three crucial phases are involved in this technique: first, create the inverted picture $T(I, j)$; Second, determine the value of k between -255 and 255; and third, use binary to distinguish between the object and backdrop.

$$result(i, j) = \begin{cases} 0, (object) & \text{if } I(i, j) < T(i, j) + k_0 \\ 255, (background) & \text{if } I(i, j) > T(i, j) + k_0 \end{cases} \quad (2.1)$$

Where k_0 is the minimal sum of the absolute strength differences and (I, j) is the strength of the original picture. Based on ME, the suggested binary produced outcomes that were better and more effective than those from other techniques.

b) Niblick Method

The Niblick method's primary objective is to establish the beginning value in accordance with the local variation and the local average. Each pixel's size has a fixed value (Bataineh et al 2011, Ntogas & Ventzas, 2008).

$$T(x, y) = m(x, y) + k\delta(x, y) \quad (2.2)$$

Where, the standard difference $\delta(x, y)$ and the area average $m(x, y)$ are determined using a window size of 80×80 (Bataineh et al., 2011, Ntogas & Ventzas, 2008). and the standard k value is -0.2. This method will not work properly if the image suffers from inconsistent lights.

c) Otsu Method

The method automatically determines a beginning value based on the discrepancy between the class and the global variance. In order to offer the final solution, Otsu assumes that the non-uniform picture comprises two places that are dark and light (N. Otsu, 1997, Bataineh et al., 2011, Ntogas & Ventzas, 2008). Last but not least, the precise Otsu limit

$$T = \frac{\delta^2 B}{\delta^2 G} \quad (2.3)$$

d) Nick Method

By lowering the thresholding value, Khurshid et al. suggested brand-new approaches to enhance the Niblack technique (Bataineh et al., 2011, Ntogas & Ventzas, 2008). Following is the equation used to determine the threshold:

$$T(x, y) = m(x, y) + k\delta(x, y) \quad (2.4)$$

The k factor value is similar with Niblack and the windowing size is defined as 15 x 15, while I and N represent the intensity pixel and mean of grey scale image. N represents the image size.

e) Bernsen Method

The estimated value of the area limit for each pixel serves as the foundation for Bersonon's method. Only when the difference between the least and highest gray values is higher than the initial k is this value categorized as a boundary value. If not, it is presumed that the window area only contains front- or back-part pixels. The equation is as follows for the default window size (w);

$$T(x, y) = \frac{Z_{max} + Z_{min}}{2} \quad (2.6)$$

f) Local Adaptive Thresholding

A fundamental and straightforward mechanism for separating the front face from erratic illumination is called local adaptation limit. It is necessary to figure out the threshold for each pixel in the image. The background value will be used if the pixel value is less than the default; otherwise, the face value will be used. The area limit (T) is 0.05 and the default area window (w) is 15 by 15 (R. C. Gonzalez and R. E. Woods, 2008).

$$T = \frac{\max + \min}{2} \quad (2.7)$$

g) F-Measure (FM)

F-measure (FM) is the harmonic mean of Precision (P) and Recall (R), which in turn are defined by the number of True Positives (TP), False Positives (FP), and False Negatives (FN).

$$FM = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

$$\text{where } \text{Recall} = \frac{TP}{TP + FN} \quad \text{Precision} = \frac{TP}{TP + FP} \quad (2.8)$$

h) pseudo-F-Measure (ps-FM)

Pseudo-F-Measure (ps-FM), as described by Ntirogiannis et al. (2013), employs pseudo-Recall ps-R and pseudo-Precision ps-P. (following the same formula as F-Measure shown in Equation 2.8). With regard to the contour of the GT characters, distance weights are used in the pseudo-Recall/Precision metrics. The weights of the GT foreground are normalized according to the local stroke width in the case of pseudo-Recall. These weights typically have a delimiter of [0, 1]. When using pseudo-Precision, the weights are restricted to a region that stretches to the GT backdrop while taking into consideration the closest GT component's stroke width. The weights are normally bounded between (1, 2) inside this region, although they are equal to one beyond it.

i) Peak Signal-to-Noise Ratio (PSNR)

A measure of how closely a picture resembles another is called PSNR. The resemblance between the two photos increases with increasing PSNR values. Keep in mind that C is equal to the difference between foreground and background.

$$\text{PSNR} = 10 \log \left(\frac{C^2}{\text{MSE}} \right)$$
$$\text{where MSE} = \frac{\sum_{x=1}^M \sum_{y=1}^N (I(x,y) - I'(x,y))^2}{MN} \quad (2.9)$$

2.6.3. Skew Estimation

One of the most crucial jobs in the document image analysis stage of an OCR system is the identification and correction of document skew, which is known as "skew estimation" (Papandreou et al., 2013). Document image skew correction is necessary as a pre-processing step before OCR since it is certain that skew will be introduced when a document is scanned or recorded using a digital camera. The prevailing orientation of the text lines in a picture of a document is said to deviate from the horizontal axis, and this is known as the skew angle. An observer using a human eye may be able to see skew angles larger than 0.10. The performance of subsequent processing, such as page layout analysis and text recognition, may be significantly hampered by the presence of skew (Papandreou et al., 2013, Boudraa et al., 2020).

Skew identification and correction continue to be difficult and complicated problems, particularly for texts that contain pictures, charts, figures, or different font sizes. According to a thorough analysis of the literature, there are many different skew detection approaches that may be used and that essentially fall into several groups based on the fundamental methodology they use. According to (Fitsum Demissie., 2011), there are four categories: interline cross correlation approaches, nearest-neighbor clustering methods, Hough transform methods, and projection profile-based methods. includes more categories, including those based on morphological transformation, background image analysis, statistical mixture models, principal component analysis, radon transformation, and fourier transformation, as well as others.

The most prominent techniques for problem-solving that are also used in this study are discussed in the paragraphs that follow. hand is discussed.

a) Projection profile-based method (PP)

This method is one of the most popular skew estimation methods and it was initially proposed by (Postl.,1986). In this method, histograms of the number of black pixels along horizontal parallel sample lines through the document for a range of angles are calculated. For a non-skewed document, horizontal projection profile will have peaks whose width are equal to the characters' height with maximum peak heights at the text lines and valleys whose width are equal to the between-the-line spacing. Therefore, for each angle a measure of the variation in the bin heights (such as variance) along the projection profile is tracked and the angle with the most variation gives the skew angle.

b): Hough transform method (HT)

Duda and Hart in 1972 GC introduced the Hough transform as a simple linear transform to detect a straight line. In this method, each point in Cartesian space (x, y) is mapped to a sinusoidal curve in ρ - θ Hough space using the following transform function:

$$P = x\cos(\theta) + y\sin(\theta) \quad (2.10)$$

When multiple points are on the same line, their transformation will intersect at the same point on the transform plane. Therefore, an accumulator is defined to track number of intersections that sinusoidal curves have at various ρ and θ values. As the number of intersections increases at a particular value of θ so does the possibility of having a line in the original image corresponding to that θ value. Finally, peaks at each ρ value give the angles at which straight lines can be fit through the original pixels. The skew angle of the documents is found by averaging the θ s with highest accumulator peaks.

2.6.4. Page Layout Analysis

The document's picture is composed of a variety of elements, including text, graphics, tables, backdrops, and plain text. Automatic text recognition requires a distinct analysis of various portions of the document picture. Page layout analysis is the term used to describe this procedure (Shafii, M., 2014).

To format the document page partitioning procedure in the machine reading system, binary image representation is required. In actuality, the majority of page fragmentation techniques are created for binary, biased document pictures. As a result, the effectiveness of the binary and skew correction procedure is crucial to the process of page layout analysis.

Page division and area allocation, or grouping pixels into regions or objects, are the primary objectives of page layout analysis, according to (Shafii, M., 2014). There are several approaches to page layout analysis. Typically, they are split up into three sections based on the actions that follow. techniques that are top-down, bottom-up, and hybrid (Shafii, M., 2014).

The top-down approaches start with the document's general structure before breaking it down into smaller categories. Project profile approaches, histogram analysis, X-Y piece algorithms, spatial modifications, and other top-down methods are examples. Although these techniques are frequently quick, their effectiveness depends on the user's prior document type knowledge.

Contrarily, up-and-down approaches are started and repeated in short document portions, such as link analysis, region development method, run length smoothing, Voronoi diagram, do castrum method, etc (connected pixels or sections). To form larger, more comparable territories, they combined. Although they are more adaptable, they are susceptible to storage problems.

Additionally, a bottom-up and top-down approach can be combined. The Gabor filtering method, wavelength, fractional analysis, and other techniques are included in the hybrid approach. The sort of work needed will determine which technique is best as each method has benefits and drawbacks (Khurshid, K., 2009).

Following the aforementioned tactics, many algorithms have often been employed in the past to assess page layout. Modern free and open-source software layout analysis tools with an emphasis on document complexity and extensive digitalization have been created in recent years. The tools Pure-Alethia, Larex, Acrope, and OCRFeeder are only a few of those listed.

Python is utilized in this thesis for both landscape analysis and landscaping (using the GUI PQT5 framework). PQT5 framework is renowned for its power and script flexibility throughout the document analysis process. It may also be used with a variety of operating systems, including Linux, Windows, and Macs. Its primary duties include page partition creation and viewing, as well as OCR field truth. PAGE and XML content is the native storage format. As a result, it may be used as a reader for partitioning and OCR products made by independent software that accepts page formatting, including page analysis and ground truth aspects. PQT5 framework, which was recently published, offers PAGE functions for page splitting and regional level performance assessment layouts, which are useful to pinpoint particular algorithmic issues when comparing results and help developers improve.

2.6.5. Text Recognition

Supervised or unsupervised categorization is the main objective of creating pattern recognition techniques, such as printed recognition systems (Jain et al., 2000). The specification of pattern classes, pattern representation, feature extraction and selection, classifier design and learning, selection of training and test samples, and performance evaluation are just a few of the challenges that need careful consideration (Jain et al., 2000., Fischer, 2012). For example, a multiclass pattern classification problem or a succession of pattern classification problem can be used to build a handwriting recognition issue.

An exhaustive review of the literature suggests that pattern recognition techniques may be classified historically in a number of ways, including the Template Matching Approach, the Statistical Approach, the Structural Approach, and the Syntactic Approach (Jain et al., 2000., Fischer., 2012).

The simplest and oldest method of pattern recognition is template matching. In template matching, the pattern to be identified has a prototype or template (usually a 2D form). The stored template is compared to the pattern to be identified while accounting for any permitted posture (translation and rotation) and scale variations. The available training set may be used to maximize the similarity measure, which is frequently a correlation. Frequently, the training set is used to learn the template itself. Although template matching requires a lot of processing, it is now more practical due to the availability of faster processors. Although useful in some application domains, rigorous template matching has a number of drawbacks. For instance, if the patterns are deformed, it would not work to the imaging process, viewpoint change, or large intra-class variations among the patterns (Jain et al., 2000).

In the statistical method, each pattern is seen as a point in a d -dimensional space and is described in terms of d attributes or measurements. The position of the contour, the direction of the stroke, and the center of mass are common features for photographs of handwritten writing. In a d -dimensional feature space, the objective is to select those features that enable pattern vectors from various categories to occupy compact and disconnected regions. How well patterns from distinct classes can be distinguished determines how effective the representation space (feature set) is. The goal is to create decision boundaries in the feature space that distinguish patterns from various classes given a set of training patterns from each class. The decision boundaries are established using the statistical decision theoretic technique. the probability distributions of the

patterns belonging to each class, which must either be specified or learned (Jain *et al.*, 2000; Fischer, 2012).

Strings, trees, and graphs are used in the structural approach to symbolically express an object's patterns. The most versatile models are graphs, which are made up of nodes connected by edges. While the edges record the binary connections between the subparts, the nodes describe the components of an item. It is possible to label nodes and edges with extra data like symbols and feature vectors. A natural graph model for handwritten writing might use nodes to represent the beginning and ending points of individual strokes and connect them with edges that were labeled with the length and curve of the stroke. Numerous techniques have been developed for graph-based object representation due to its tremendous representational capability (Fischer, 2012).

Deep learning strategies and approaches based on statistical learning theory, however, have lately begun to get more interest in the field of pattern representation. Deep learning, in contrast to basic artificial neural networks, is used to both learn the representation and map it from the input to the output. Representation learning is the name given to this strategy. Learned representations frequently produce significantly higher performance than hand-designed representations can (Goodfellow *et. al.*, 2016). As a result, deep learning's strong automated feature extraction capability lessens the requirement for a separate handmade feature extraction method. For instance, Google's recently announced multilingual Tesseract3 OCR system is likewise entirely constructed using deep learning methods.

Performance evaluation must be taken into account while developing a system for pattern recognition. Levenshtein⁴ distance method is frequently used for performance evaluation, particularly in the case of handwritten text recognition. The edit operation of the Levenshtein distance algorithm, often known as the string edit distance algorithm, uses dynamic programming. The smallest distance needed to convert one string into another is known as the Levenshtein distance. Using the edit actions known as insertion (because to mistakes of spurious symbols), substitution (due to errors of misspelled letters), and deletion, one may calculate the edit distance between two strings (due to errors of lost or missing text). Character Error Rate (CER), which is based on the Levenshtein distance technique, is frequently used as an assessment metric in handwritten text recognition (Drobac & Lindén, 2020).

The CER is calculated using the fewest processes necessary to produce the output from the reference text (ground truth). The difference between the two texts is represented by a greater number. The Levenshtein Distance is used to calculate the recognition accuracy as follows:

$$\text{CER} = (i + s + d)/n \quad (2.11)$$

where i stands for insertion errors, s for substitution errors, and d for deletion errors. The reference text will be converted into the OCR output using the total number of characters (n) and the bare minimum amount of character insertions I substitutions (s), and deletions (d).

2.7.Deep Neural Networks

2.7.1. ANN

Popular computer methods for pattern categorization issues include artificial neural networks. They are composed of fundamental processing elements, weight, and direct connections, which allow the output of certain elements to serve as the input for other elements. Nervous System Since these models resemble biological neurons, they are referred to as "artificial" neurons. Two things about the mind are comparable. The network learns by doing training on representative samples, therefore that is the first commonality. The strength of the connections, sometimes referred to as synaptic weights and utilized to store their information, is another (Goodfellow, I., Bengio, Y., & Courville, A. 2016).

The studies of artificial neurons that McCulloch and Pitts presented in 1943 led to the development of ANN (C. Bento, 2021). A central information processing unit with a minimal quantity of local memory is a neural network. Typically, links or communication channels that transport numerical data are used to connect units in a variety of ways. The majority of ANNs feature training rules for altering communication weights. In other words, ANN picks up new skills from training data. If thoroughly trained, ANN can show that it can summarize more than just training data, i.e., predict precise outcomes for brand-new examples that weren't utilized for training.

The perpendicular, which is comparable to a traditional image and responds to visual patterns as a stimulus, is where the algorithm for altering contact weights is focused (C. Bento, 2021). Following the perceptron model is only one example of the numerous neural network topologies that are provided. To mention the basics Each is explained in the sections that follow.

I) Multilayer Perceptron

Multi-layered Perspex has been created to overcome this restriction. A nonlinear neural network maps the input and output. One or more hidden neurons are placed together in a multi-layered perceptron, which combines input and output layers. The neurons in the multi-layered perceptron

can employ any random activation function, in contrast to the progesterone neurons, which are required to have activated functions that impose constraints, such as ReLU or sigmoids.

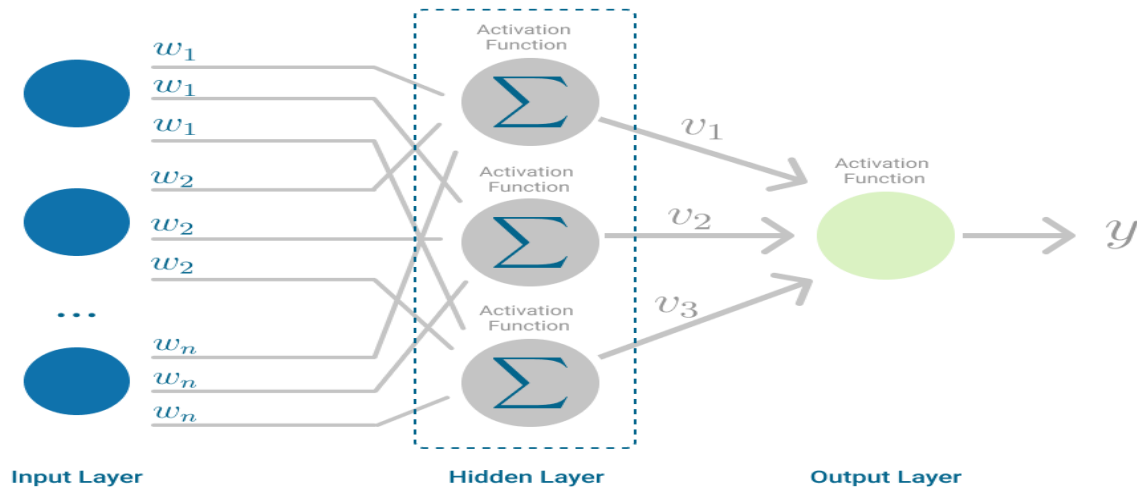


Figure 2: 2 Multilayer Perceptron

Because the components are combined with the original weight and produce a fire-like effect, Multilayer Perceptron falls within the topic of food algorithms. The fact that each linear joint extends to the next level makes a difference, though. Each cover serves as a continuation of the calculation and draws information from the underlying data representation. This progresses from overlapping outcomes to hidden layers. However, he has more. The algorithm will not be able to learn the weights that will lower the cost function if it merely adds up the weights on each nerve, distributes the result to the outcome layer, and ends there. There would be no meaningful teaching if the algorithm was simply calculated once.

II) Backpropagation

A transdisciplinary learning method called backpropagation enables you to continually change the network's weight in order to save money. Back-up distribution must adhere to rigorous guidelines in order to function successfully. It is necessary to change a beginning function, such as ReLU, as well as a function that mixes inputs and weights in neurons. Given that the MultiLayer Perceptron frequently uses gradient descent as a facilitator, these functions ought to have a variety of consequences.

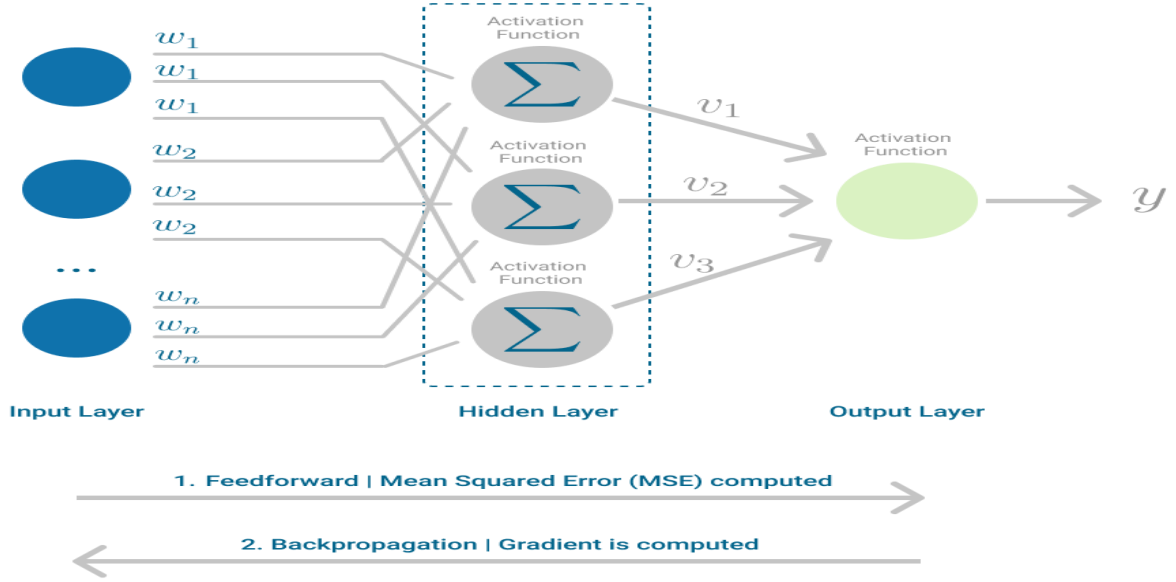


Figure 2: 3 Multilayer Perceptron, Highlighting the Feed Forward And Backpropagation

Every time a repeat is made, the total weights are applied to all layers, and the average square error on all input and output pairs is determined. The gradient value to be spread backwards is then changed with the weights of the first hidden layer. The neural network receives the weights in this manner!

$$\underbrace{\Delta_w(t)}_{\text{Gradient Current Iteration}} = \underbrace{-\varepsilon}_{\text{Bias}} \underbrace{\frac{dE}{dw(t)}}_{\text{Error Weight vector}} + \underbrace{\alpha \Delta_w(t-1)}_{\text{Learning Rate Gradient Previous Iteration}}$$

Figure 2: 4 One iteration of Gradient Descent

The procedure is repeated until each input-output pair is merged, which ensures that the newly computed composition does not deviate from the prior frequency by more than a specific amount of coupling (Abay Teshage, R. Sethuraman, 2015). The Amharic OCR Neural Network was originally shown in an experiment that was documented in (Berhanu, A. 1999). A.D. Graduate MLP attempted to employ OCR in 2016 (Siranesh, G. 2016). for deep neural network transcripts of historical Geez (three hidden layers). The project is regarded as the first known effort to apply machine learning methods to Geez alphabet OCR.

2.7.2. CNN

A network of neural networks called CNN is dedicated to processing data as an image grid. A description of the visual data is provided by the digital picture. Each pixel is made up of a group of pixels that are grouped in pixels to specify the color and brightness of the pixel.

The human brain analyzes a lot of information when we look at an image. Each nerve covers the complete range of vision and functions in its own receptor area. It is also linked to other neurons. Every nerve in CNN acts exclusively in the receiving field, which is the constrained field of view referred to as the receiving field in the biological vision system. The covers are placed such that the simplest designs (lines, curves, etc.) may be recognized first, followed by the most complicated designs (faces, objects, etc.). CNN viewing can be permitted on computers.

2.7.2.1. CNN Architecture

A CNN typically has three layers: a convolutional layer, a pooling layer, and a fully connected layer.

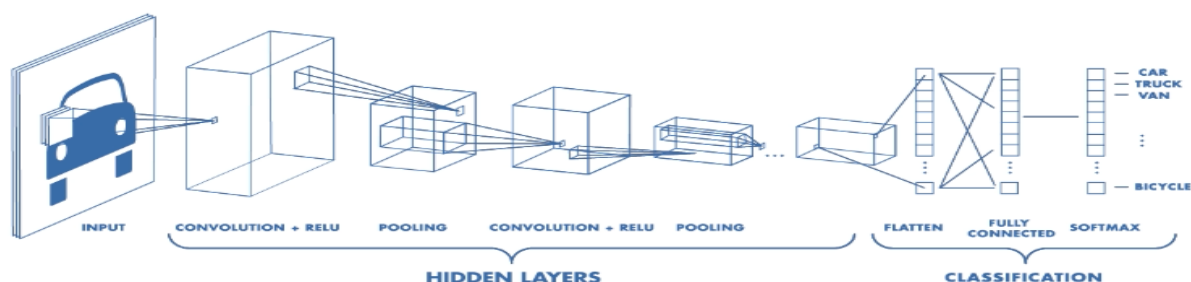


Figure 2: 5 Architecture of a CNN

2.7.2.2. Convolution Layer

The fundamental component of CNN is the convolution layer. The majority of the network's workload is taken up by computation. This layer manages the creation of points between two negotiations, one of which involves a kernel a collection of parameters that may be learnt during negotiation and the other, which involves a smaller portion of the negotiation field. Though somewhat smaller than the picture, the kernel is far deeper. The kernel height and width will be reduced in space if the picture has three RGB channels, but the depth will cover all three channels. The nucleus slides over the height and breadth of the picture as it advances, displaying the receptor region's image. As a result, a two-dimensional activation map picture is produced,

with each place in the image responding to a kernel response. The walker is the size of the main slider.

Quantity Enter the grain size in $W \times W \times D$ and DW using F . The following formula may be used to calculate the size of the output if we know the sizes of PS and P .

$$W_{out} = \frac{W - F + 2p}{s} + 1 \quad (2.12)$$

Formula for Convolution Layer This will yield an output volume of size $W_{out} \times W_{out} \times D_{out}$.

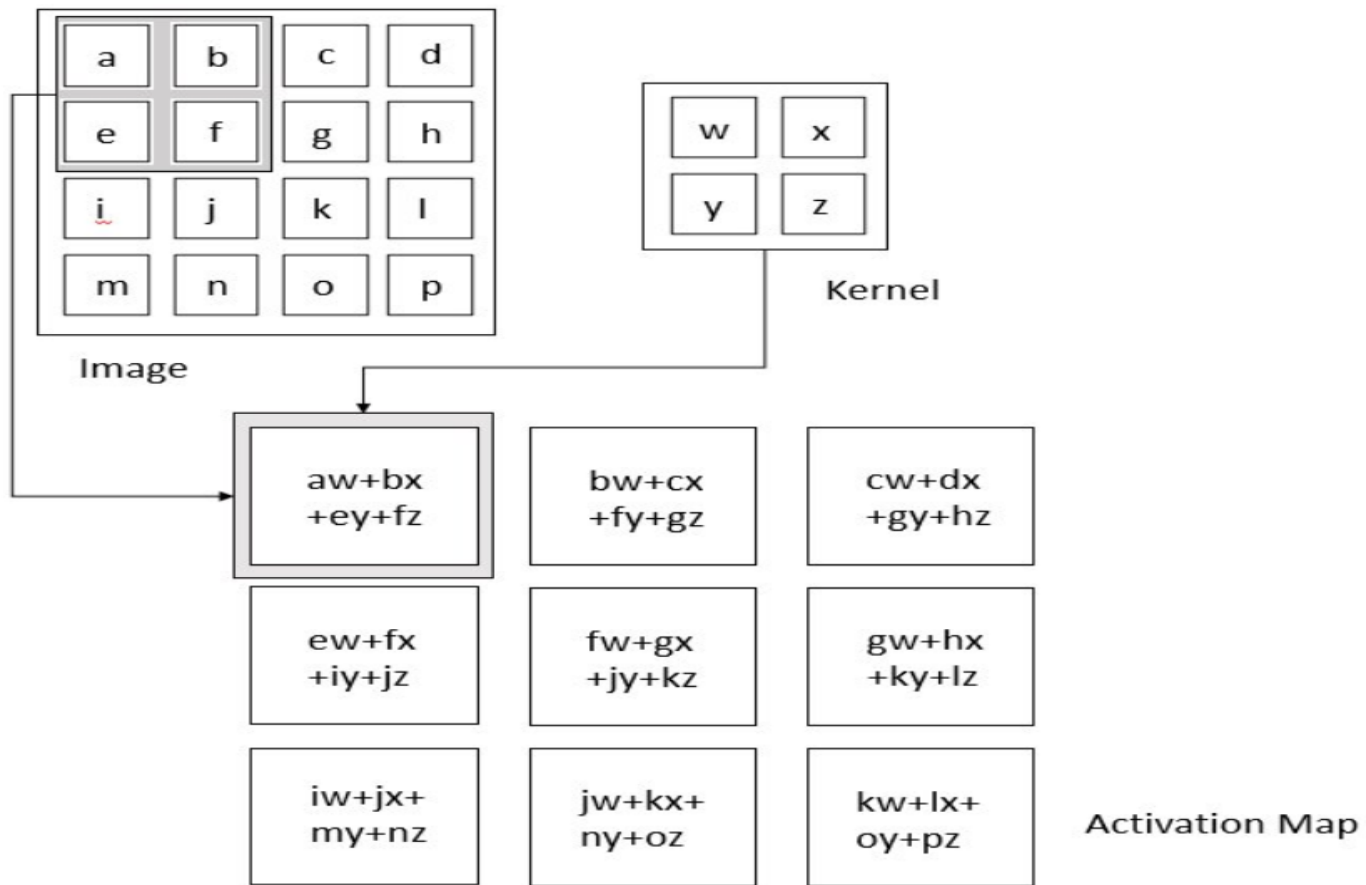


Figure 2: 6 Convolution Operation

2.7.3. RNN

Recurrent neural networks are ones that have an idea of internal state that changes over time and is attained through the network's cycles or loops. RNN is, in other words, a type of neural network that uses sequential data for modeling (such as voice, handwriting, and other types of

recognition). Hopfield networks (Hopfield, J. 1982). are an ancestor of recurrent neural networks, albeit recurrence in these networks was utilized to establish a stable state rather than process time sequences. An MLP with a recurrent layer is the simplest version of an RNN. However, the recurrent layer also feeds data to itself (Bluche, T. 2015). in addition to the inputs from the earlier levels.

2.7.4. Speech signals and speech processing

Technologies come and go. But every time a new technology arrives, it attempts to change the way we work and interact with computers. And each new technology tries to transform the existing one into a simple form. Speech is the most common mode of communication among human beings. Speech offers a new way of interfacing with a computer that lends itself very well to solving the problems of field-based computing. It's natural and intuitive.

2.7.5. Bidirectional Recurrent Neural Networks

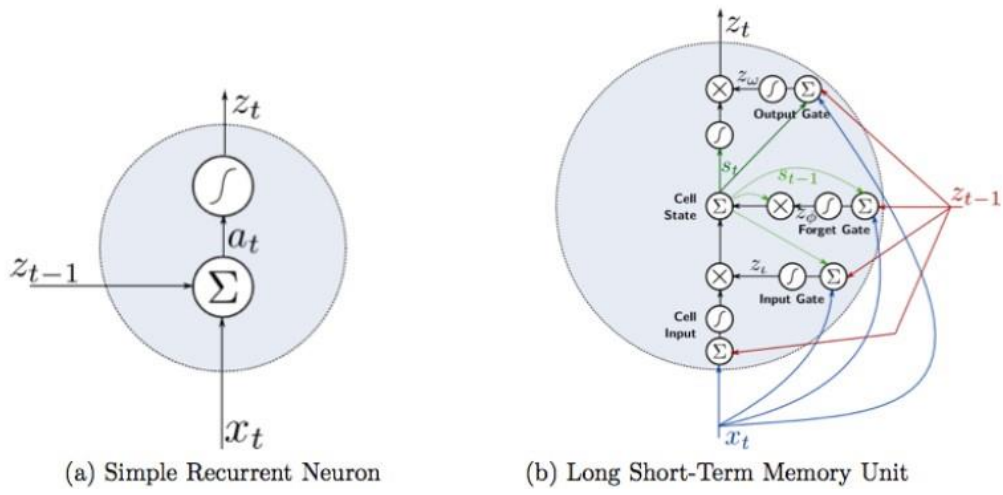
In its purest form, an RNN consists just of a causal structure. To put it another way, the state of a standard RNN simply stores information about the past, $x(1)$, $x(t1)$, and the current input, $x(t)$. To create a forecast of $y(t)$, which depends on the complete input sequence, is something we wish to do in many situations. In speech recognition, for instance, it could be necessary to consider both the present and the past in order to accurately identify the sound as a phoneme. Additionally, this is necessary for a variety of other sequential-to-sequential learning tasks, including handwriting recognition. The stream was processed in both directions using Bidirectional Recurrent Neural Networks (BRNN).

These networks consist of two redundant layers: a forward layer that accepts inputs from the previous time step and a backward layer connected to the subsequent time step. The identical input and output layers are linked to both layers. The one-dimensional nature of standard RNN designs, including BRNNs, makes them appropriate for a single space-time dimension. introduces multidimensional recurrent neural networks (MDRNNs), which handle a four-dimensional input picture in recurrent layers, making RNNs useful in a variety of fields, including vision, video processing, medical imaging, and many more. A thorough review of the literature revealed no examples of OCR research projects using Geez and Amharic scripts and basic RNN, BRNN, or MDRNN designs.

2.8.Long Short-Term Memory Units (LSTM)

Traditional RNNs and the different architectures that have been covered thus far are based on the notion of handling continuous data for modeling. But in practice (i.e. during training), RNN suffers from the difficulties of trickling gradient (i.e. slowly going to zero) or bursting gradient (i.e. gradient to infinity) (i.e. gradient to infinity). The network is unable to learn long-lasting dependencies because of the slowness problem of spillover. Hochreiter and Schmidhuber proposed short-term memory units to address the issue.

LSTM lessons come in a variety of architectural styles. A control cell (the LSTM unit's memory) and three gates (the input gate, output gate, and forget gate) are the basic components of an LSTM design. The cell is in charge of tracking the connections between the input stream's components. An output gate regulates how much is utilized to calculate the value of the cell, while input gates and forget gates regulate how quickly fresh values enter and leave the cell, respectively. activation of the LSTM class's output. Some variations of the LSTM unit lack one or more of these ports or could include additional ports. For instance, GRUs lack an exit door.(Bluche, T., 2015) contrasted the following images of an LSTM cell with a simple recurrent neuron. The top layer triggered in the preceding time step activates all gates and receives cell input from the bottom layer. The behavior of the LSTM class as stated in Bluche, T., 2015) is represented by the following equations.



The Input Gate controls whether the input of the cell is integrated in the cell state

$$a_i^t = \sum_{i=1}^I w_{il} x_i^t + \sum_{h=1}^H w_{hl} z_h^{t-1} + \sum_{c=1}^C w_{cl} s_c^{t-1} \quad (13)$$

The Forget Gate controls whether the previous state is integrated in the cell state, or if it is forgotten.

$$a_{\phi}^t = \sum_{i=1}^I w_{i\phi} x_i^t + \sum_{h=1}^H w_{h\phi} z_h^{t-1} + \sum_{c=1}^C w_{c\phi} s_c^{t-1}$$

The Cell state is the sum of the previous state, scaled by the forget gate, and of the cell input, scaled by the input gate

$$a_c^t = \sum_{i=1}^I w_{ic} x_i^t + \sum_{h=1}^H w_{hc} z_h^{t-1}$$

$$s_c^t = z_{\phi}^t s_c^{t-1} + z_i^t g(a_c^t)$$

The Output Gate controls whether the LSTM unit emits the activation $h(s_c^t)$.

$$a_{\omega}^t = \sum_{i=1}^I w_{i\omega} x_i^t + \sum_{h=1}^H w_{h\omega} z_h^{t-1} + \sum_{c=1}^C w_{c\omega} S_c^t$$

$$z_{\omega}^t = f(a_{\omega}^t)$$

The Cell output is computed by applying the activation function h to the cell state, scaled by the output gate.

$$Z_c^t = Z_w^t h(S_c^t) \quad (2.17)$$

In terms of handwriting recognition, recent research show that LSTM classes with recurrent layers, BRNN, and MDRNN are state-of-the-art and produce great results. The recently launched open-source Tesseract OCR engine uses LSTM networks and offers Unicode support for more than 100 languages, including Amharic.

LSTM Network to Geezer Alphabet OCR System. The network, however, was trained using a dataset made up of artificial text lines in the languages of Tigrinya, Geez, and Amharic. A thorough review of the literature turned up no instances of OCR studies using LSTM on old Geez manuscripts.

2.9. Historical Background of TTS

Even though voice synthesis has a long history, (Anne and Klaus Wedekind, 1991) there have been new developments in this discipline with an emphasis on integrating human speech from a vast corpus to create realistic sounds. For millennia, people have fantasized of creating speech-enhanced artificial intelligence.

More than 200 years ago (Mieko Takada, 2008), both in terms of mechanical and electrical advancement, the first artificial intelligence tests were conducted. A.D. Russian scholar Christian Cratzenstein created an artificial insemination system in 1779 and identified the physiological distinctions between the five long vowels (/ a /, / e /, / I /, and / u) at the time of mechanical development.

Them. Wolfgang von Kempellen unveiled the "Acoustic-Mechanical Speech Machine" in Vienna a few years later, in 1791. This device could create single sounds as well as a variety of noises. The renowned Von Kemplen edition was released by Charles Whiston in the middle of

the 19th century. speaking device, it could create vowels and highly consonant sounds and was a little difficult.

Although there are no accompanying consonants or phrases, the Stewart machine was capable of making single interchangeable vowel sounds in two lower forms. Whiston served as an inspiration for Alexander Graham Bell and his father, who created a comparable typewriter in the late 1800s. He shouts at the dog while holding it between his knees and making sounds that the dog would normally only hear if it were deaf and stupid to speech.

Complete electrical integration was achieved in 1922, the first two forms had noise and at least two resonant circuits, and the third format was introduced to the system in 1923. Researchers found the first gadget, known as the Voder Operating System, which Homer Dudley exhibited at the 1939 New York World Series. mechanism for integrating speech. Scientific study throughout the world is gaining speed and producing spectacular results on a global scale as the integration of speech science and technology into the discourse is separated into main stages of mechanical and electrical development. With a wide range of new applications and services in the industries of telecommunications, entertainment, language instruction, and more since the mid-1990s, there have been major advancements in the field of TTS integration.

The focus of research has changed away from generating more lifelike and comprehensible sounds. These advancements are made possible by information-based techniques and big volume components (i.e., words and phrases), but they are not the only ones. These components now apply to all TTS technologies and systems. The automated creation of speech waves, or speech integration, has been under research for many years (Ralf Kompe, 1997). Sens. with high intelligence have been created thanks to recent improvements in speech synthesis, but the promise of natural and clear clarity remains a major issue. But today's goods are of a caliber that makes them ideal for a range of uses, including multimedia and telecommunications.

The advancement of speech science and technology, particularly in the incorporation of voice for a range of practical applications, has benefited greatly from all prior studies. In order to integrate and extract human-like sounds from the composer, modern speech integration technologies use sophisticated and complicated processes and algorithms. The trade-offs between improved voice quality and decreased memory space, algorithmic complexity, and computing speed and cost are crucial problems for today's speech processors.

2.9.1. Fundamental of Speech Synthesis

The integration of voice and speech recognition has received a lot of attention in the subject of speech science and technology, which is a subfield of language technology. Speech has traditionally been seen as the most natural form of human communication and is the primary method of communication between individuals.

The voice cords, nose, and diaphragm are only a few of the physical elements that make up the human auditory system (Yoshiko Fujisaki, 1998). Speech synthesis It's easy to explain a machine that speaks to people. Modern algorithms and techniques that enable the generation of sounds from real noises while preserving the sound characteristics of the original speaker are the foundation of TTS technology. (Yoshiko Fujisaki., 1998). Speech The technique of transforming text from speech to speech, with the capacity to turn arbitrary text into audio, is known as synthesis. Its purpose is to provide listeners with text-based information via voice (Joyce McDonough., 2002). Additionally, manufactured from raw materials, artificial speech is frequently utilized to assist those with impairments (Jun Xu., 2008, Mieko Takada., 2009).

TTS integration is used for screen readers, interactive voice response (IVR) systems, hands-on and hands-free information access, and human-machine engagement. software for blind people. A speech that blends perceptual and/or multi-modal discourse speech is accompanied by an animated face in Speech Processes, a computer-generated speech created by persons with impairments (Joyce McDonough, 2002). knowledge of signal processing, image integration, physiology, prosthetics, computer languages, phonetics, and experimental psychology. The development of natural language and digital signal processing at the start of the 20th century made artificial intelligence widely applicable.

2.9.2. Phonetic Analysis

After the typical word strings have been extracted through text analysis, the second role of NLP is to determine the proper pronunciation for various contexts in the text. Additionally, as mentioned in Nadew(Ann K. Syrdal, Gregor Mohler, Kurt Dusterho, Alistair Conkie and Alan W. Black,1999), this phase includes preparing information on each word's pronunciation and dialect. The TTS system's most challenging issue may be with a few words known as homography. (Juergen Schroeter, "Text to-Speech (TTS) Synthesis,2008) Consequently, graphite-to-phoneme conversion is another name for phonetic analysis.

The phonetic analysis module's functionality can be categorized in a variety of ways, usually approximately into dictionary-based and rule-based methodologies (Juergen Schroeter., 2005). Most voice knowledge is kept in a lexicon as part of dictionary-based solutions. Inputs are often restricted to solids in order to preserve their rationality, and the pronunciation of the shapes is described in terms of rigid, derivative, and compound morphophic rules that specify how phonetic versions of their morphological components are put together. Some phonetic verbs are frequently used to describe the more subtle aspects of the United Nations after acquiring the initial phonetic transcription of each word. (Juergen Schroeter., 2008).

The majority of dictionary vocabulary abilities are instead transmitted to letter-to-voice (or graph-to-phonetic) rules through rule-based transcription systems (Juergen Schroeter., 2008). In this instance, only the words that are pronounced in a certain manner are retained in a unique lexicon until they become regulations. A somewhat modified dictionary may logically contain more than the number of words in a series since numerous exceptional situations are found in the words that are uttered most frequently. For instance, in English, 2,000 words are sufficient to cover 70% of the text's words (Juergen Schroeter., 2008).

Amharic is virtually phonetic, which implies that most words may be translated into phonetic form using basic grammar because of the language's strong relationship to Amharic phonology and spelling. The homograph effect on Amharic does not affect pronunciation, although there are several methods to communicate the same thing Completely.

2.9.3. Speech Synthesis Techniques

Speech integration approaches may be categorized into two groups, according to Donovan (Abay Teshage, R. Sethuraman., 2015) systematic approach and signal model presentation. Articulatory synthesis is a system model technique that aims to directly simulate human voice chords. On the other hand, the signal model method, which consists of two types: integration and regulation-based formatting, tends to concentrate primarily on the voice signal of the model.

As was already indicated, the manual intervention rule in system design and data-based integration may be used to categorize voice integration. In the first, a composer is managed by a set of rules, whereas in the second, the parameters of the synthesizer are derived from actual speech data. The formal integration is governed even if the communication system is data-based. There is now a novel method for informational synthesis that combines the two in order to make each advantage more superior to the other. (Jun Xu, 2008)

2.10. Related Researches in Local languages

As far as the researcher is aware, some outstanding work has been done so far on speech synthesis for local languages. The following is a discussion of some of the concatenative synthesis-based works.

2.10.1. TTS for Afaan Oromo

For Afan Oromo, Morka (Abay Teshage, R. Sethuraman 2015) created a TTS prototype. Concatenative voice synthesis was the method employed, and the diaphragm was used as the fundamental communication unit to incorporate example texts written in Afan Oromo. The success percentage for native languages in identifying the pronunciation of the phonetic unit replicated by the research was 43.33%. To make the diffusers' transition points easier, it is advised to apply spectrum smoothing techniques.

2.10.2. TTS for Tigrigna

A Tigrinya TTS system prototype to be created. For integrating time-domain Pic-Syntronos Supervision and Addition as well as Tigrigna sample texts, Deepon acts as the fundamental integration unit (TD-PSOLA). On the findings test technique and evaluation for both Tigrinya and both, average remarks were approved.

2.11. Related work

Table 2: 7 Research works on OCR for Ge'ez and Amharic scripts [17th May 1997 – 30th Sep. 2020]

Author name	Year	Location	Thesis Title	Writing mode, digitizer, file [Resolution]	Recognition technique [Level]	Approach followed	Pros	Cons	Application area [Dataset]
Worku Alemu	1997	Addis Ababa University	The Application of OCR Techniques to the Amharic Script	Printed, flatbed scanner, BMP (Binary image) [300 dpi]	Structural [Character]	Adopted a stage-by-stage segmentation algorithm; topological feature based on the structural shape of thinned image characters; binary tree-structured classification	The old recorded of OCR for Amharic script	single font & point type (Washera font style and 12 point type).	Amharic, printed text [Private]
Ermias Abebe,	1998	Addis Ababa University	Recognition of Formatted Amharic text using optical character recognition	Printed, flatbed scanner, BMP (Binary image)	Structural [Character]	Approach followed was similar to & an extension of Worku's work but gave emphasis to Italic and removing underline	Adopted thinning and underline detection & removal. It aimed	Limited handcrafted features were represented.	Amharic, printed text [Private]

				[300 dpi]			formatted Amharic text		
Dereje Teferi	1999	Addis Ababa University	Optical character Recognition of Typewritten Amharic Text	Typewritten & Typed, flatbed scanner, BMP (Binary image) [300 dpi]	Structural [Character]	It was an extension of Worku's work (the first study) but with some adjustments to make it suite to type-written text	Adopted noise detection and removal; had considered image enhancement and character building	Limited handcrafted features were represented (identified from contour analysis).	Amharic, type-written and printed text [Private]
Million Meshesha	2000	Addis Ababa University	A Generalized Approach to Optical character Recognition of Amharic texts	Printed, flatbed scanner, BMP (Binary image) [300 dpi]	Structural [Character]	Adopted and modified the stage-by-stage segmentation of Worku's work, proposed hybrid-thinning; used it for the skeletonization process, topological features based on Worku's procedure were extracted. Then, a feature based binary	The early recorded attempt of Omni-font recognition to Amharic script. It had considered a wide variety of font type & size.	Limited handcrafted features were represented (only primary & secondary global descriptors identified by contour analysis were considered).	Amharic, printed text [Private]

						tree classification was used for the recognition.			
Million Meshesha, and Jawahar	2007	Addis Ababa University	Optical character recognition of Amharic documents	Printed, flatbed scanner, (Gray-scale) [300 dpi]	Structural [Character]	The images were binarized and noise removed (Gaussian filtering), and skew corrected (projection profile). Segmentation followed top-down approach (projection profiles), characters scaled to 20x20 pixels and decomposed to their constituents (connected component analysis). Proposed a two-stage feature extraction scheme; PCA	It is a conference paper presented on the prestigious ICDAR/2005 conference. The study reported the challenges towards the recognition of Omni-font Amharic scripts and possible solutions. The techniques (PCA & LDA)	Objective evaluation methods were not considered for binarization, skew correction and page segmentation tasks. No language model or post-processing in the recognition process. Actual real-life documents were not sampled when the training set prepared.	Amharic, Synthesiz e printed text [Private]

						followed by LDA for optimal discriminant feature extraction. A DAG classifier with SVMs was applied and trained pair-wise using the two-stage feature extraction.	are powerful for feature dimensionality reduction and data representation learning.		
Abay Teshager	2010	Addis Ababa University	Amharic Character Recognition System for Printed Real-Life Documents	Printed, flatbed scanner, JPG (Grayscale) [300 dpi]	Neural Network [Character]	First applied Wiener adaptive filtering method for noise removal then binarization (Otsu's method) in the pre-processing stage. Adopted the stage-by-stage segmentation suggested by Pal and Chaudhuri (1995), normalized	Binarization task (Otsu's global thresholding method) was applied. In the thesis, the body of knowledge of the area was organized & explained in a manner that beginners	Feature extraction task was not considered in the system; the thinning would likely made loss of information about the character to be recognized. Due to the thinning, multiple	Amharic, Printed real-life documents [Private]

						characters into 20x20 pixels by using linear interpolation technique (set by experimentation). Hit- and-miss morphological analysis for thinning was employed. ANN with a network 400-0-1, input-hidden-output nodes was designed &	could see the big picture about OCR.	characters would likely ended up with a similar binary representation. Using a single node on the output layer for a multi-class pattern classification problem was also not a good decision.	
Fitsum Demissie	2011		Developing Optical Character Recognition for Ethiopic Scripts	Printed, flatbed scanner, BMP (Binary image) [300 dpi]	Statistical [Character]	The document was binarized, deskewed and enhanced. Then segmentation applied, and then a total of 1500 samples (100 different samples from selected 15 characters, scaled to 30x30) were split into 1200 for training &	fonts (Nyala, Power Geez & Agafari) & d/t kernel types (linear, polynomial and GRBF) were investigated. NPE, PCA and MDS	Objective evaluation methods were not considered for binarization and segmentation tasks. No language model or post-processing in the	Amharic, Printed character [Private]

						300 for testing. A dimension reduction tools (PCA, NPE and MDS) were investigated and fed to the SVM for training.	techniques were also investigated for feature dimensionality reduction.	recognition process.	
Betselot Yewulu	2018	Parul Institute of Engineering and Technology, Vadodara, India	Amharic Handwritten Character Recognition Using Combined Features and Support Vector Machine	Handwritten, flatbed scanner, (Grayscale) [300 dpi]	Statistical [Character]	In the pre-processing stage: smoothing & noise removal (median filtering), cropping and normalization were applied. A combination of Histogram of oriented gradients (HOG), Local Binary Pattern (LBP) and geometrical features were used and then LDA (for dimension reduction) and DCA to integrate two set of	Large combinations of handcrafted statistical features (3839 features) were considered (along with geometrical features). The proposed model was trained on extra Chars74K benchmark numeric data	The collected data were not sampled from the real-world distribution (just drawn from handwritten characters written in isolation on white papers). Document image analysis and language model were not considered in the recognition	Amharic, handwritten character [Private]

						features applied and investigated, then multiclass SVM with linear kernel was probed for classification.	set, and the model was validated using 10-fold cross-validation.	process.	
B.Belay, T. Habtegebrial, M. Liwicki , G. Belay, D. Stricker	2019	University of Kaiserslautern, Kaiserslautern, Germany, Lulea University of Technology, Lulea, Sweden, Bahir Dar Institute of	Factored Convolutional Neural Network for Amharic Character Image Recognition	Printed	Deep Learning [Character]	Proposed a model of deep CNN (multiple convolutional, ReLU, Maxpooling, fully connected layers & Softmax output) were used. The dataset was curated using OCRopus (Open source OCR system developed by Breuil et al.). Trained over a dataset of 80,000 Amharic character images, normalized into 32x32 pixels. Trained with SGD.	The study considered CNN (auto-derived feature extractor) for feature extraction task (the early recorded attempt to Amharic OCR). Early stopping and dropout regularization techniques were used.	Document image analysis tasks were not considered in the recognition process. No language model or post-processing task in the recognition process. The dataset was limited to synthetic character images of Power Geez font only.	Amharic, Synthetic character [Private]

		Technol ogy, Bahir Dar, Ethiopia							
--	--	--	--	--	--	--	--	--	--

CHAPTER THREE

3. METHODOLOGY

3.1.Introduction

As mentioned in Chapter 1, the purpose of this study is to Digitalization of Amharic print documents by using deep learning for the Amharic document in Audio. In this chapter, we will look at how and in what ways we achieved of goal through Experimental research. This includes method selection of the model and design with a prototyping approach employed to build the proposed printed document recognition and audio conversion system since it is very helpful to improve the system through experiment.

3.2. General Architecture of the Proposed System

One of the most common practices in developing a text recognition system is to introduce sequential and modular data processing. This allows the system to process the intermediate output and transforms it into making useful representations.

The proposed system is built on a variety of digital conditions. It features a pre-processing algorithm (skew estimation and binary) that takes into account the various elements of the text recognition process. It also performs page layout analysis and region classification. (trained model).

Th system accepts scanned images as input. All of the functions that are involved in the recognition process are equally important. To ensure that the system is capable of handling the various tasks related to the print document recognition problem, a goal assessment has to be made for each module.

After the initial conversion of a printed document to digital, the system is designed to integrate the digital image output value and speech integration into its architecture. In order to achieve this, the system is built on a combination of techniques and design issues that are based on the Amharic phonology analysis.

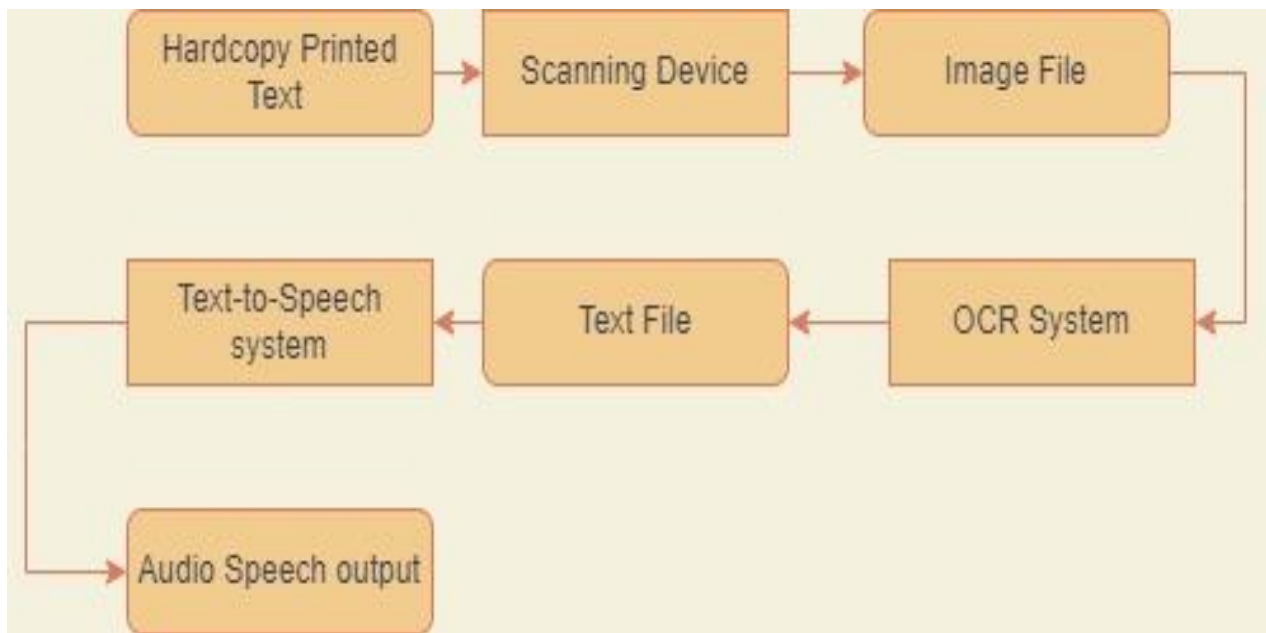


Figure 3: 1 Architecture of the proposed for Amharic Print Documents Transcription to Audio Using Machine Learning Techniques system

3.3. Pre-processing Experimental Setup

One of the main goals of an OCR system is to improve the quality of the scanned images. This process can be carried out through the development of a binarization algorithm. In this setup, the goal is to find the optimal algorithm for the binarization task and also develop to analyze the skew estimation technic.

3.3.1. Image Binarization

A significant portion of the pre-process involves binarization. The majority of printed materials are typically created using black ink on a white background. On the other hand, some printed materials use diverse backgrounds and text colors. In this experimental setting, four binarization approaches are suggested based on their positive reviews in the document image analysis literature. Three of these are the conventional techniques referred to as the Otsu global method, Otsu local method, and the Adaptive Thresh Gaussian method. The adaptive threshold mean approach is the last option. It is necessary to try many methods and evaluate the outcomes to determine which one is the best method. Any printed paper serves as the testing dataset for the binarization of damaged documents with ground truth. There are nine (9) photos in all, each of which shows indicative degradations of the current issue. F-Measure (FM), pseudo-F-measure (ps-FM), Peak Signal-to-Noise Ratio (PSNR), and Distance Reciprocal Distortion (DRD), which are commonly used performance assessment measures, are employed for the evaluation.

All the binarization methods and the evaluation metrics are discussed in detail in Section 2.6.2

3.3.2. Image Skew Estimation

One of the most significant jobs in document image analysis is identifying and fixing deformed documents. There are several document page segmentation methods that can handle pictures without a zero. As a result, the skew correction must be used. The document's page was processed before being divided. as the process of picture identification is indirectly impacted. Therefore, skew identification and rectification were accomplished using a skew estimation technique known as the Bicubic interpolation approach. In the fields of computer vision and image analysis, bicubic interpolation is a well-known method.

The effectiveness of the Bicubic interpolation technique is examined. Analyze the results of the datasets. The Bicubic interpolation method performance is evaluated using the following three parameters: the top 80% average error deviation (TOP80), the percentage of accurate estimates (CE), and the average error deviation (AED), all with a threshold of 0.10. This threshold was selected because a tilt angle larger than it would be noticeable to a human observer.

3.4. Page Layout Analysis

The major goal of this experimental setup is to assess the realistic document layout, the photographs, and the objective analytics system performance. Framework for the Python PQT5 GUI The quality of the documents that are deformed and digitally imaged is continually being degraded while building a thorough OCR system to compile the majority of the documents. This is because separating the scanned pages into text and non-text parts frequently isn't enough. The following points should be included in a thorough study of the page layout.

- The precise interpretation differentiation between the region categories (picture, title, text, and paragraph).
- A reading instruction that reads the entire page of text.
- Accurate text, word, and character line detection.

In order to verify the effectiveness of Leptonica, an open-source C library for effective documentation or printed image analysis, an experimental setup was created. A collection of test pages made up of images with various page layouts are used to assess performance. Laptonia processed each page of the document, and the finished products were saved as an XML partition output file and a black-and-white document picture. Field truth areas and subdivision results

regions were contrasted for assessment. Variations fall under the category of review error (such as merge, partition, error, partial error, incorrect placement, erroneous detection, and general error).

3.5. Training OCR Engine Setup

The development of open-source OCR engines, particularly Google's Tesseract OCR engine, made it feasible to train models to detect printed documents with the maximum level of accuracy. An experimental Tesseract OCR engine design to shown the next chapter. Tesseract was selected as the OCR engine largely due to its compatibility for the Amharic alphabet and open-source nature, which is popular with large-scale digitization. The most recent Tesseract enabled the usage of network-based models that used LSTM and AVX-512. The difficulties now encountered can be interpreted as opportunities for facilitation and issue optimization, which is anticipated to improve the Tesseract recognition model's accuracy across a set of training samples for printed documents. Preparing training data, conducting the training procedure, and evaluating the performance of a currently trained recognition model are the three primary steps in the training of the Tesseract OCR engine.

3.5.1. Training Data Preparation

One of the main issues in the field of recognition is the development of ground truth training samples from printed texts for OCR engines. The method for creating training samples from real texts is described in this section. The Tesseract OCR engine is mostly trained using ground truth, and it is also necessary to assess how well the recently taught recognition model is performing. It is made with the utmost care to guarantee that error-free training samples are collected. Because the training samples must be taken from the original documents, creating the ground truth is challenging. It may, however, be effectively accomplished by employing tools with the least amount of human engagement.

To produce text line image of sample data with the accompanying ground truth for training LSTM-based Tesseract, Python (PQT5 framework) is used as a ground-truthing tool. Python GUI was chosen largely because of its effectiveness and widespread use with the extensive digitalization of printed materials.

3.5.2. Performance Evaluation

The newly trained recognition model is assessed using the CER performance assessment metric. In Section 2.6.5, the evaluation metric is covered in more depth.

3.6. Research Tools

The following different tools were used to prepare the proposed model.

3.6.1. Design Tools

There are different diagrams in this research that need to be properly specified, and the diagram is the center of this research. To create the diagrams for this study, we utilized design working software. draw.io It is a specialized application for making graphs and charts. The software enables you to design unique layouts. They come in a variety of forms that enable us to create distinctive graphs and charts. A great-looking diagram or chart may be easily created using the drag-and-drop capability.

3.6.2. Data Preparation Tools (Scanner)

In order to enter the raw data into a digital device, it must first be scanned to a standard level. We need the caliber of the information gathered during scanning in order to accomplish this. For our investigation, an HP HD Pro scanner with a resolution of 600–1200 DPI was employed.

3.6.3. Programming Tools

Python3 is a high-level, universally applicable, object-oriented, interpretable, and interactive programming language. Between 1985 and 1990, Guido Van Ross also created it. Python source code is also accessible under the GPL, much like Perl. Python is not named for the snake; rather, it is a reference to the television program Monty Python's Flying Circus. Python is one of the most frequently used languages on the data analyses⁹. List of them: -

- The simple to learn Python has a concise syntax, simple structure, and minimal keywords. This makes it possible for the user to learn the language rapidly.
- Python code is easier to read since it is more logically stated and visually appealing.
- Python's source code is comparatively simple to maintain.

⁹ https://www.tutorialspoint.com/python/python_overview.htm

3.6.4. Frameworks

a): Tesseract

The Apache 2.0 license governs Tesseract¹⁰, an open-source text recognition (OCR) engine. It may be used to extract printed text from photos directly or (for programmers) through an API. Many different languages are supported. Tesseract doesn't come with a built-in GUI, however there are plenty on other websites. Tesseract offers containers for a variety of programming languages and frameworks here.

A new neural network subsystem outfitted as a text line detector is part of Tesseract 4.00. It was initially an Osyropus Python-based LSTM implementation that was converted to C++ for Tesseract. Although Tesseract's neural network architecture predates TensorFlow, it is compatible with it since TensorFlow also has a network description language called Variable Graph Specification Language.

We commonly employ a CNN to identify a picture that has a single character. Since the random-length text is made up of a series of characters, RNN and LSTM are a common variety is used to solve the problems overcoming these challenges.

c); Audio play framework

The Access to the fundamental sound-playing equipment offered by Windows systems is supplied through the win sound module. It includes a number of constants and processes.

win audio. Beep (frequency, duration) Speaker on the computer beep. The frequency parameter, which sets the sound's frequency in hertz, must fall between 37 and 32,767. The duration option defines how long the sound should play for in milliseconds. Runtime Error is thrown if the system is unable to beep the speaker. win audio. PlaySound (sound, flags) Use the Platform API to call the underlying PlaySound function. The sound argument can be one of the following: a filename, a system sound alias, audio data in the form of an object with bytes, or None. The value of flags, which might be a bitwise ORed combination of the constants indicated below, determines how it should be interpreted. Any waveform sound that is presently playing is halted if the sound parameter is None. The Runtime Error is produced if the system detects a problem.

¹⁰ <https://github.com/tesseract-ocr/tessdoc>

3.6.5. Hardware Tools

The experiments were performed on a ThinkPad T430 with hardware specifications Intel: Core™ i5-3240 CPU @ 3.40 GHz, 8GB Ram, and x64-based processor. The operating system was Windows 10 Pro.

CHAPTER FOUR

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Introduction

The assessment results for choosing the optimal algorithm or technique for binarization and skew estimation are shown in this section. Examining the outcomes of the experiments and the experimental dataset allows us to choose the best algorithm or approach based on objective metrics. On the basis of the test findings, comparisons and discussions are made next.

4.2. Binarization Document Images

The Nine (9) distorted images and the ground truth that goes with them, which was taken from print documents, are included in the experimental dataset. Images representing an issue resembling the one at the ground-truthing were chosen to be included in the experimental dataset. The evaluation findings for each test image (T1, T2, and T9) technique parameters are shown in Table 4.2. Based on the four various techniques described in Section 2.6.2, the assessment In Table 4.1, the last phase is depicted. First, the total of all measurements for each experimental image is assigned in order to compute it. The final step is to get the overall result given in Table 4.1 by adding up all of the total values for all of the experimental photos.

The precise results for each method for each parameter are shown in Table 4.2. The cumulative value for each experimental image dimension is rated before calculating the "Total score," as indicated in Table 4.1. The final "Total result" displayed in Table 4.1 is the sum of the sum values for each experimental image. Let $R^i(j, m)$ represent the method i rank in relation to the j^{th} image when using the m^{th} measure is used. Then, for each binarization method i , the Total Score S^i is given by the following Equation. Following that, the following Equation gives the Total Score S^i or each binarization technique i .

$$S^i = \sum_{j=1}^K \sum_{m=1}^L R^i(j, m) \quad (4.1)$$

where L is the total number of evaluation metrics (i.e., L = 4), and K is the total number of images utilized in the assessment (i.e., K = 9).

The 'Total Score' for the Adaptive Thresh Gaussian technique is determined as follows:

$$\text{Score (Adaptive_Thresh_Gaussian)} \Rightarrow \mathbf{FM} = 1 + 2 + 2 + 2 + 2 + 3 + 3 + 2 + 3 = 20$$

$$\mathbf{ps-FM} = 1 + 2 + 1 + 1 + 2 + 3 + 2 + 2 + 3 = 17$$

$$\mathbf{PSNR} = 1 + 2 + 2 + 1 + 2 + 3 + 2 + 3 + 3 = 18$$

$$\mathbf{DRD} = 1 + 2 + 2 + 1 + 2 + 3 + 2 + 2 + 2 = 17$$

$$\text{Therefore, Total Score} = 20 + 17 + 18 + 17 = \mathbf{72}$$

The 'Total Score' for the Otsu's local technique is determined as follows:

$$\text{Score (Otsu_Local)} \Rightarrow \mathbf{FM} = 3 + 3 + 1 + 3 + 1 + 1 + 1 + 3 + 2 = 18$$

$$\mathbf{ps-FM} = 3 + 3 + 3 + 3 + 1 + 2 + 3 + 1 + 2 = 21$$

$$\mathbf{PSNR} = 3 + 3 + 3 + 3 + 1 + 1 + 3 + 2 + 2 = 21$$

$$\mathbf{DRD} = 3 + 3 + 3 + 3 + 1 + 1 + 3 + 3 + 3 = 23$$

$$\text{Therefore, Total Score} = 18 + 21 + 21 + 23 = \mathbf{83}$$

The 'Total Score' for the Adaptive_Thresh_Mean_ technique is determined as follows:

$$\text{Score (Adaptive_Thresh_Mean_)} \Rightarrow \mathbf{FM} = 2 + 1 + 3 + 1 + 3 + 2 + 2 + 1 + 1 = 16$$

$$\mathbf{ps-FM} = 2 + 1 + 2 + 2 + 3 + 1 + 1 + 3 + 1 = 16$$

$$\mathbf{PSNR} = 2 + 1 + 1 + 2 + 3 + 2 + 1 + 1 + 1 = 14$$

$$\mathbf{DRD} = 2 + 1 + 1 + 2 + 3 + 2 + 1 + 1 + 1 = 15$$

$$\text{Therefore, Total Score} = 16 + 16 + 14 + 15 = 61$$

The 'Total Score' for the Otsu's global technique is determined as follows:

$$\text{Score (Otsu_Global)} \Rightarrow \mathbf{FM} = 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 = 36$$

$$\mathbf{ps-FM} = 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 = 36$$

$$\mathbf{PSNR} = 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 = 36$$

$$\mathbf{DRD} = 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 = 36$$

$$\text{Therefore, Total Score} = 36 + 36 + 36 + 36 = \mathbf{144}$$

Table 4: 1 Overall evaluation results and the final ranking of the binarization methods

Method	FM	ps-FM	PSNR	DRD	Total Score	Overall Rank
Otsu_Global	8.5677	18.772	9.659	26.295	144	4th
Adaptive_Thresh_Mean_	75.112	80.688	16.134	6.997	61	1st
Otsu_Local	70.1963	76.830	14.93	9.926	83	3rd
Adaptive_Thresh_Gaussian	73.439	79.0182	15.234	7.005	72	2nd

Note: The best results are shown in bold.

The detailed performance for each binarization methods is also given in Table 4.2.

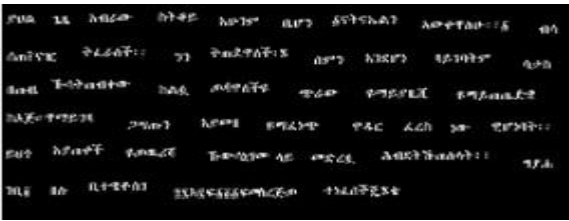
Table 4: 2 Evaluation results of binarization for each test image with respect to the metrics used

Metrics	Method	T1	T2	T3	T4	T5	T6	T7	T8	T9
FM	Otsu_Global	15.906	13.784	9.7289	10.078	8.3993	4.8511	2.9034	3.8951	7.5637
	Otsu_Local	42.905	68.797	49.739	65.01	65.314	93.883	81.853	72.221	92.045
	Adaptive_Thresh_Gaussian_	54.643	74.361	42.253	76.137	78.109	92.315	83.319	80.389	94.482
	Adaptive_Thresh_Mean	57.456	70.739	48.489	75.121	82.982	81.51	78.986	77.201	88.47
ps-FM	Otsu_Global	14.73	11.47	9.3045	109.253	8.7898	6.235	0	2.075	7.0947
	Otsu_Local	45.498	69.622	49.844	65.789	85.092	95.998	88.296	99.237	92.097
	Adaptive_Thresh_Gaussian_	54.07	76.881	51.552	79.87	78.02	94.667	92.501	99.092	99.543
	Adaptive_Thresh_Mean	56.907	72.742	52.339	78.165	83.934	86.82	92.988	98.011	89.258
PSNR	Otsu_Global	5.6985	7.9015	8.0913	7.0432	10.35	11.23	15.99	13.35	7.283
	Otsu_Local	7.9005	11.026	10.711	11.82	17.405	21.54	18.717	15.63	19.68
	Adaptive_Thresh_Gaussian_	9.9125	12.857	12.96	13.456	14.23	21.07	20.567	18.591	21.563
	Adaptive_Thresh_Mean	9.9867	11.91	12.012	13.592	16.902	18.55	19.17	17.901	17.09
DRD	Otsu_Global	47.83	32.036	51.004	31.125	21.74	22.62	8.204	22.0976	24.255
	Otsu_Local	28.027	11.82	22.396	13.229	3.3472	1.7155	4.243	4.5643	2.6551
	Adaptive_Thresh_Gaussian_	19.52	7.9432	14.018	6.933	6.9014	1.7891	2.3933	3.4831	0.8903
	Adaptive_Thresh_Mean	15.939	9.6765	16.036	5.9345	4.9912	4.071	2.4925	3.9049	1.1233

Note: The best results are shown in bold.

The performance is best overall when using the **adaptive threshold mean**. The Adaptive Thresh Mean Method's Binarization Results are shown in Figure. 4.1.

የህልፈው ስትቀይ አሁንም ቢሆን 'ናትናኤልን አውቀዋለሁ።' ብላ
 ለመናገር ትፈራለች። ግን ትወደዋለች። ለምን እንደሆነ ባይገባትም ሳታስ
 በው፣ ሳትጠብቀው ከልቧ ወዳዋለች፤ ጭራው የማይያዝ፤ የማይጠመድ፤
 ከእጅ የማይገባ ጋማውን አያመስ የሚፈነጭ የዱር ፈረስ ነው የሆነባት።
 ይህን አያውቀች የመጨረሻ ውሳኔዋ ላይ መድረሷ እብደት መሰላት። 'ገደል
 ግቢ' ብሎ ቢተዋትስ? "አፍ...ፍ" በረጅሙ ተነፈሰች።



(a)

Figure 4: 1 Example of binarization results of Adaptive Thresh Mean Threshold method

4.3. Skew Document Image Estimation

The Bicubic Interpolation technique is used to assess the quality of distorted estimations using a well-known set of experimental datasets from the printed Document Image Bias Estimation Contest. 200 photos altogether, 20 of which are unique, make up the experimental dataset. For each distinct picture, 10 rotating samples are produced. These rotation angles were chosen at random from the interval (+ 15 °, -15 °). For each sample, the ground truth value of the established angle is manually specified. The distribution of ground truth angles for the pictures in the experimental dataset is shown in the following frequency table.

Table 4: 3 Frequency table of the ground truth skew angle distribution

Bins	Midpoint	Absolute Frequency	Relative Frequency	Cumulative Relative Frequency	Density
[-15,-10)	-12.5	36	0.17	0.17	0.046
[-10,-5)	-7.5	32	0.18	0.43	0.023
[-5,0)	-2.5	30	0.14	0.54	0.02
[0,5)	2.5	37	0.189	0.756	0.047
[5,10)	7.5	33	0.145	0.74	0.022
[10,15]	12.5	32	0.15	1	0.032

4.4. Comparison

After closely examining the binary assessment findings shown in Table 4.1, it is clear that the Adaptive Thresh Mean approach produced better outcomes. All other approaches can't compare to the adaptive Threshold method in any way. Similar to this, the Adaptive Thresh Gaussian approach, which came in second, is likewise the best method when all metrics are included. The Otsu global technique, on the other hand, performs the poorest across all of the measures. The Otsu approach was primarily used in earlier studies on OCR systems' attempts to incorporate

binary methods into printed documents (Fitehalew, A. 2019), (Siranesh, G. 2016). However, none of them really put the binarization procedure through a meaningful review.

On the same dataset listed in Table 4.4, comparisons are done between the skew estimation results and those of different skew estimation techniques. Using the Canny filter, bicubic interpolation was used to determine the objects' shapes. According to Table 4.4, the suggested technique performs much better than all other methods in the TOP80 criterion when compared to the predictive method in the AED criteria.

Table 4: 4 Evaluation results and comparison of skew estimation methods

Method	AED (°)	TOP80 (°)	CE (%)
Projection profile	0.35	0.254	-
SHT	7.89	5.845	-
Bicubic interpolation	0.4215	0.068	86

While skew estimation techniques based on projection profiles and bicubic interpolation were attempted to be incorporated in earlier research studies of OCR systems for Amharic script, such as (Fitehalew, A. 2019, Siranesh, G. 2016), none of them had implemented an objective evaluation method that takes into account the performance of the skew estimation task.

4.5. Discussion of Results

Experiments have yielded encouraging results, which confirm the validity of the binarization and skew estimation methods. The results show that the global access system is not good enough for low-quality and degrading printed document images. In other words, local or global thresholding methods have demonstrated good records on low-quality and damaged document images. Although The Adaptive Thresh Mean Method is superior to all other measurements used, further research is needed to determine the exact values (i.e., window size and weight). The size of the test set also played an important role in assessing the strength of the strategy. For example, the Otsu local method, in the second place, outperforms all other methods in all parameters in a single test case, T5. But the performance is not the same as in other tests.

The binarization and skew estimation approaches have been found to be effective through experiments, which have produced positive results. The findings demonstrate that the global access system is inadequate for pictures of printed documents that are of poor quality and that degrade. In other words, local or global thresholding techniques have proven effective on

distorted and poor-quality document pictures. Even though the Adaptive Thresh Mean Method outperforms all other metrics, further investigation is required to pinpoint the precise values (i.e., window size and weight). The size of the test set was crucial in determining the effectiveness of the technique. For instance, in a single test scenario, T5, the Otsu local technique comes in second and performs better than all other methods in every parameter. However, the performance varies from those of other tests.

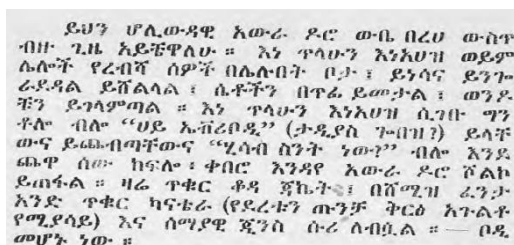
On the other hand, the suggested Bicubic Interpolation technique has a high accuracy of 86%, according to the test results of the skew estimation method. The strategy proved successful in achieving performance below the recognized TOP80 0.10 level.

This demonstrates the method's accurate operation in the specified operating mode, as well as the method's robustness and ability to deal with a wide range of problems. Under the AED standard, however, the technique is ineffective below 0.10. However, there are certain optimization techniques that might make this better but they need further study.

4.6. Results and Discussion of The Page Layout Analysis

The Leptonica, an open-source C library for image analysis, has had its effectiveness examined. Tesseract may be used with the leptonica Library for OCR. the official GitHub¹¹ repository for Leptonica. In this experiment, we'll look at how well PyQt5 performs across platforms when used with the experimental Leptonica page and region segmentation set. Five (5) images of actual printed documents with various complicated designs and physical forms make up the test set. Each document image in the test set is described by its page layout in the picture that follows.

(a)



Description: -

Document Image1 has a resolution of 600 dpi and a size of 800x870 pixels. The document is clear.

¹¹ <https://github.com/DanBloomberg/leptonica>

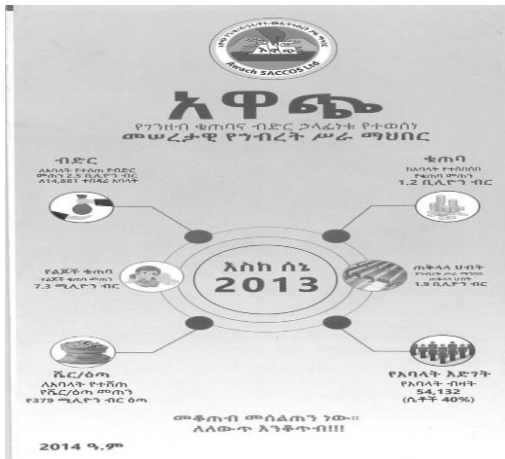
(b)

የህል ጊዜ አብራው ስትቆይ አሁንም ቢሆን 'ናትናኤልን አውቀዋለሁ'። ብላ
ለመናገር ተፈራለች። ግን ትወደዋለች። ለምን እንደሆነ ባይገባትም ሳታስ
በው! ሳትጠብቀው ከልቧ ወደዋለች! ጭራው የማይያዝ፤ የማይጠመድ፤
ከእጅ የማይገባ ጋማውን እያመሰ የሚፈነጭ የዱር ፈረስ ነው የሆነባት።
ይህን አያውቅ የመጨረሻ ውሳኔዋ ላይ መድረሷ እብደት መሰላት። 'ገደል
ግቢ' ብሎ ቢተዋተቅ? 'እ-ና...ና' በረጅሙ ተነፈሰች።

Description:

Document Image2 has a size of 252x200 pixels and a 96 dpi resolution. One column of text makes up the text area.

(c)



Description:

Document Image3 has 620x881 pixel dimensions and a 96 dpi resolution. It is heavily ornamented with artwork, and the text area and marginal notes are surrounded with intertwined ornamentation. It was made of printed materials and decayed inequitably.

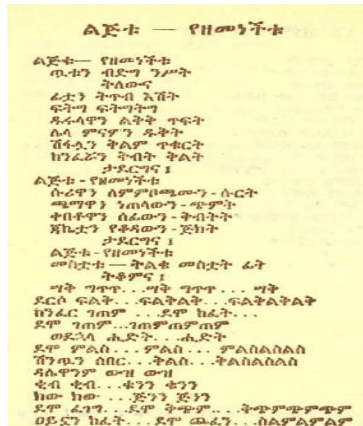
(d)

የዜግነት ክብር በ ኢትዮጵያችን ፀንቶ
ታየ ህዝባዊነት ዳር እስከዳር በርቶ
ለሰላም ለፍትህ ለህዝቦች ነፃነት
በእኩልነት በፍቅር ቆመናል ባንድነት
መሰረተ ፅኑ ሰብዕናን ያልሻርን
ህዝቦች ነን ለስራ በስራ የኖርን
ድንቅ የባህል መድረክ ያኩራ ቅርስ ባለቤት
የተፈጥሮ ፀጋ የጀግና ህዝብ እናት
እንጠብቅሻለን አለብን አደራ
ኢትዮጵያችን ኑሪ እኛም ባንቺ እንኩራ።

Description: -

Document Image4 has a size of 1194x952 pixels and a resolution of 72dpi. It comprises of two pages at a time. Each page features a text area with a single column.

(e)



Description:

Document Image5 has 296x170 pixel dimensions and a 96 dpi resolution. It has a single-column text area with various text line lengths. At the top, there is an interlaced border ornamentation, and there is obvious wear from fading ink and show-through.

Figure 4: 2 Page layout description of each document image in the testing set

Each document image in the test set has been pre-processed before the page layout analysis. Prior to the page layout analysis work, the Adaptive Thresh Mean binarization technique and the Bicubic Interpolation, Bias Detection, and Correction method were carried out. Both approaches were picked based on how well they performed in earlier procedures, as mentioned in section 4.1. Each document image in the test set had a tilt angle that was, respectively, -0.2, -0.65, 2.32, 12.65, and 6.7 degrees.

There are differences between the various sorts of regions (picture, text, paragraph, and title), as well as the order in which the various text sections on the page should be read. On Image1, Image2, and Image4, Leptonica did rather well in terms of area categorization. The region's poorest categorization the entire page is categorized as an image area and a short-length separator, despite the fact that it contains a text region made up of columns. Because of the numerous photos and decorations that surround the text area and connect to one another, the text space is not clearly segregated. But the subject is recognized as such. Similar to Image 5, much of the text section there is incorrectly identified as a header rather than a paragraph. The interwoven design that surrounds a portion of the text space is one of the potential causes of this.

Text box detection in the text area is a crucial factor to take into account while performing the process of page layout analysis. When it comes to complicated layouts and damaged documents, tracing text lines continues to be a difficult operation. Leptonica successfully spotted the text box in this test, though. The text box performance of Leptonica in the pictures of the studied papers is demonstrated in the following image.

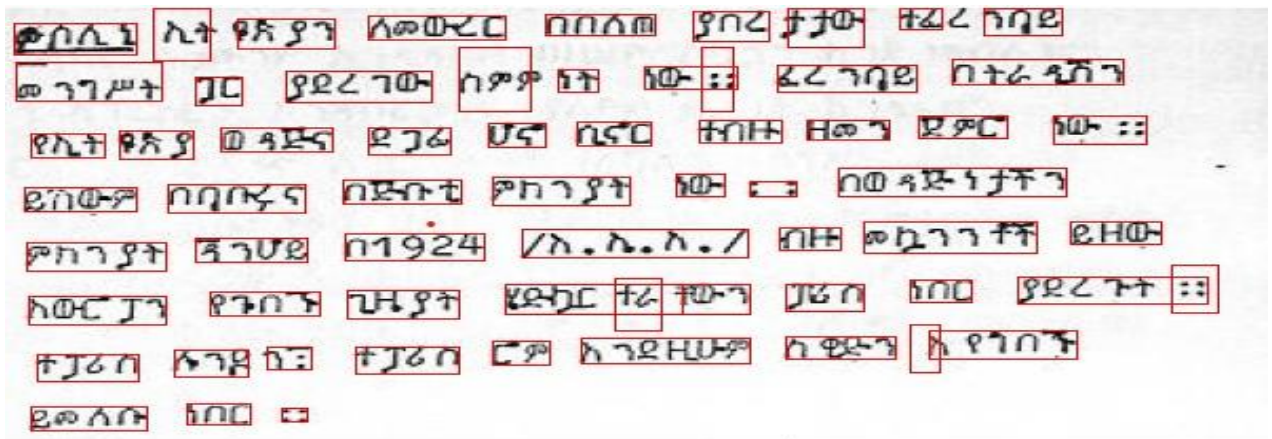


Figure 4: 3 Sample of text line detection in Image1

Word and glyph recognition was also looked at in addition to text detection. However, it was discovered after carefully examining earlier word and glyph identification studies that Leptonica could not achieve an acceptable success rate on all of the document pictures examined.



Figure 4: 4 Sample of word detection

The assessment approach was employed in the examination of the page layout in addition to the qualitative description of the test findings that were previously indicated. Leptonica was used to process each page of the document in order to employ an objective assessment approach. The results were then recorded in xml parameters and compared to the ground truth created manually using a cross-platform xml framework. The assessment utilized the actual field xml file, segmentation results xml, and a black-and-white document picture as input. The distribution result areas were contrasted with the ground truth regions. Disparities were noted as assessment mistakes (such as fusion, segmentation, error, partial error, misclassification, and false discovery relative to success rate). Weights and default settings are taken into consideration. Regions, text,

and groups were the assessment levels taken into account. The following evaluation criteria' success rates were lastly noted.

፱፪ ዐጉል ግስ ግን ተሳቢ የሌለው፤ ሥራው ከዚያው ካድራጊው የማይወጣ ባድራጊው ላይ ብቻ የሚፈጸም ነው። ለምሳሌ ሠነዩ ቀደመ፤ ምዕዘ፤ ጼነው፤ ልህቀ ጥሕረ ነቀው፤ ሞተ ሐይው፤ ቆመ ሐረ፤ ሮጸ በረ፤ ደክመ ነበረ፤ ኖመ ንጎረ፤ ነደ ቄረ፤ እንደዚህ ያለ ይህን የመሰለ ሹሉ ዐጉል ይባላል፤ ፍፍፍንቶ እንደ ማለት ነው። ተገብሮ እንዳይባል ሌላ ገባሪ የለውም፤ ገቢር እንዳይባል ሌላ ተሳቢ የለውም፤ አድራጊውም ተደራጊውም ርሱ ራሱ ስለ ሆነ ዐጓጉል ይባላል። ፈፍፍንታም ሰው ከወንድነት ይልቅ ወደ ሴትነት እንዲያደላ፤ ይህም ከገቢር ይልቅ ወደ ተገብሮ ያደላል፤ ቅጥሩም ከተገብሮ ነው።												
፱፪	ዐጉል	ግስ	ግን	ተሳቢ	የሌለው፤	ሥራው	ከዚያው	ካድራጊው	የማይወጣ	ባድራጊው	ላይ	ብቻ
የሚፈጸም	ነው።	ለምሳሌ	ሠነዩ	ቀደመ፤	ምዕዘ፤	ጼነው፤	ልህቀ	ጥሕረ	ነቀው፤	ሞተ	ሐይው፤	ቆመ ሐረ፤
ሮጸ	በረ፤	ደክመ	ነበረ፤	ኖመ	ንጎረ፤	ነደ	ቄረ፤	እንደዚህ	ያለ	ይህን	የመሰለ	ሹሉ ዐጉል ይባላል፤
ፍፍፍንቶ	እንደ	ማለት	ነው።	ተገብሮ	እንዳይባል	ሌላ	ገባሪ	የለውም፤	ገቢር	እንዳይባል	ሌላ	ተሳቢ
የለውም፤	አድራጊውም	ተደራጊውም	ርሱ	ራሱ	ስለ	ሆነ	ዐጓጉል	ይባላል።	ፈፍፍንታም	ሰው	ከወንድነት	
ይልቅ	ወደ	ሴትነት	እንዲያደላ፤	ይህም	ከገቢር	ይልቅ	ወደ	ተገብሮ	ያደላል፤	ቅጥሩም	ከተገብሮ	ነው።

Figure 4: 5 Ground truth of Image3 created manually using PyQt

Merges: A segmentation result region connects two or more ground truth regions.

Partial Misses: portions of the ground truth that the segmentation result does not completely cover.

Misclassifications: Segmentation of areas of results of different categories from the appropriate regions of ground truth.

False detections: Segmentation result regions with no corresponding ground truth region.

Table 4: 5 Page layout evaluation results of success rate for each document

Evaluation Metrics	Image1		Image2		Image3		Image4		Image5	
	Region level	Text line level	Region level	Text line level	Region level	Text line level	Region level	Text line level	Region level	Text line level
Merge	75.89%	98.44%	100.00%	25.33%	30.32%	25.37%	92.19%	98.22%	54.42%	37.19%
Miss	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	99.17%
Partial Miss	99.97%	96.89%	100.00%	100.00%	99.79%	99.28%	98.74%	99.92%	100.00%	99.89%
Misclassification	96.82%		67.34%		69.97%		99.67%		45.31%	
False Detection	100.00%	39.94%	100.00%	99.90%	100.00%	99.60%	100.00%	99.94%	100.00%	100.00%
Overall success rate:										
Arithmetic mean	63.06%	66.27%	89.57%	59.82%	78.70%	59.06%	88.19%	99.79%	68.73%	68.64%
Harmonic mean	59.51%	55.63%	72.06%	40.29%	58.11%	37.43%	85.45%	99.79%	57.77%	53.62%

The aforementioned finding is supported by a comprehensive examination of the evaluation objective outcomes in Table 4.6. Leptonica, for instance, successfully classified the regions and text in Image 4 with a high success rate, as shown in Table 4.6.

4.7.Results and Discussion of Training the OCR Engine

The Tesseract OCR engine was trained using the training data, and the results are provided below along with a brief outline of the experimental process. To assess the effectiveness of the training, the outcomes of the new recognition model are compared to those of the basic or pre-trained model.

4.8.Training Data Preparation

Deep learning techniques are difficult to adopt because they need a lot of training data. There are, however, several ways to get beyond this obstacle. Fine-Tuning Is One of the Many Training Options. The Adaptive Thresh Mean lets you start with an existing base model and train on certain extra data rather than starting from scratch. The fine-tuning approach is growing in popularity these days since it is challenging to prepare huge training datasets using actual documents. A similar strategy will be suggested and used in this study from the standpoint of printed texts to address the issue.

Real data at the text line level is necessary for the LSTM-based Tesseract training procedure. As a result, real images are used to produce the training data. The Python GUI framework was used to create 286 pictures of text lines that were gathered from 15 distinct pages. A text file contains text files that are UTF-8 encoded and serve as the matching ground truth.



Ground truth

የዜግነት ክብር በ ኢትዮጵያችን ፀንቶ
 ታየ ህዝባዊነት ዳር እስከዳር በርቶ
 ለሰላም ለፍትህ ለህዝቦች ነፃነት
 በእኩልነት በፍቅር ቆመናል ባንድነት
 መሰረተ ፅኑ ሰብዕናን ያልሻርን
 ህዝቦች ነን ለሰራ በሰራ የኖርን
 ድንቅ የባህል መድረክ ያኩራ ቅርስ ባለቤት
 የተፈጥሮ ፀጋ የጀግና ህዝብ እናት
 እንጠብቅሻለን አለብን አደራ
 ኢትዮጵያችን ኑሪ እኛም ባንቺ እንኩራ።

Figure 4: 6 Sample of text line images with their corresponding ground truth

4.9. Running the Training Process

Ground truth at the text line level is necessary for the Tesseract training process. A set of command-line tools including Windows CMD are used to train it. After installing all the prerequisite libraries, training is created. OCR-D for Windows 10 is used to carry out the operation. The procedure is composed of the following steps: -

1. Download the link to dee build your own dataset.

<https://sourceforge.net/projects/vietocr/files/jTessBoxEditor/>

2. Download the visualstudio.exe from the link after running the exe file it installed the C++ compiler

<https://visualstudio.microsoft.com/vs/features/cplusplus/>

3. Download tesseract-ocr-w64-setup-v5.2.0.20220712.exe (64 bit) from <https://github.com/UB-Mannheim/tesseract/wiki>.
4. Provide training images and their ground truth text labels to OCR-D: Provide images with .tiff format and their labels put like this train_amh_exp0.tif extension. We need to place all these files in the data/ground-truth folder inside OCR-D. To do this follow the six steps: - To do this follow the six steps: -
 - 4.1. Extract the jTessBoxEditor-2.3.1 on Desktop then open the file name like this train
 - 4.2. Open CMD on the folder containing the image then run this command: -
tesseract train_amh_exp0.tif train_amh_exp0 batch.nochop makebox
 - 4.3. change the image extension to.TIFF: -
tesseract train_amh_exp0.tif train_amh_exp0 box.train
 - 4.4. Extract the charset from the box files (Output for this command is unicharset file)
unicharset_extractor train_amh_exp0.box
 - 4.5. Create a font_properties file based on our needs.
echo "Abyssinica SIL.exp0 0 0 0 0 0" > font_properties
 - 4.6. Training the data.
amh_train.unicharset train_amh_exp0.tr
 - 4.7. Rename four files (shapetable,inttemp,pffmtable,normproto)
into([langname].shapetable,[langname].inttemp,[langname].pffmtable,[langname].norm
roto)Syntax: rename filename1 filename2
Eg:
rename shapetable amh_train.shapetable
rename inttemp s amh_train.inttemp
rename pffmtable amh_train.pffmtable
rename normproto amh_train.normproto
 - 4.8. Create .traineddata file
combine_tessdata amh_train.

Move. traineddata file to tesseract programs tessdata directory

5. Create A new folder and open pycharm in the folder then install the Necessary files and start to coding.

4.10. Performance Evaluation

The freshly trained recognition model's character error rate, or the "amh train.traineddata," is 2.632%. (2.6.5)

4.11. Comparison

The outcomes of OCR were compared with and without training in order to assess the impact of training. Here, utilizing the basic model, or "amh.traineddata," suggests there has been no training. We must copy both recognition models to the tessdata directory in order to use them. Or, to put it another way, the "amh.traineddata" and "amh train.traineddata" files must be moved to the database of the Tesseract instance that will be used to do the OCR. The following text line pictures are randomly chosen from the assessment list and examined to compare the OCR outcomes of both algorithms.

Document Image



የዜግነት ክብር በ ኢትዮጵያችን ፀንቶ
ታየ ህዝባዊነት ዳር እስከዳር በርቶ
ለሰላም ለፍትህ ለህዝቦች ነፃነት
በእኩልነት በፍቅር ቆመናል ባንድነት
መሰረተ ፅኑ ሰብዕናን ያልሻርን
ህዝቦች ነን ለስራ በስራ የኖርን
ድንቅ የባህል መድረክ ያኩራ ቅርስ ባለቤት
የተፈጥሮ ፀጋ የጀግና ህዝብ እናት
እንጠብቅሻለን አለብን አደራ
ኢትዮጵያችን ኑሪ እኛም ባንቺ እንኩራ።

Recognition using the base model



የዜግነት	ክብር	በ	ኢትዮጵያችን	ፀንቶ
ታየ	ህዝባዊነት	ዳር	እስከዳር	በርቶ
ለሰላም	ለፍትህ	ለህዝቦች	ነፃነት	
በእኩልነት	በፍቅር	ቆመናል	ባንድነት	
መሰረተ	ፅኑ	ሰብዕናን	ያልሻርን	
ህዝቦች	ነን	ለስራ	በስራ	የኖርን
ድንቅ	የባህል	መድረክ	ያኩራ	ቅርስ
የተፈጥሮ	ፀጋ	የጀግና	ህዝብ	እናት
እንጠብቅሻለን	አለብን	አደራ		
ኢትዮጵያችን	ኑሪ	እኛም	ባንቺ	እንኩራ።

Recognition using the new model

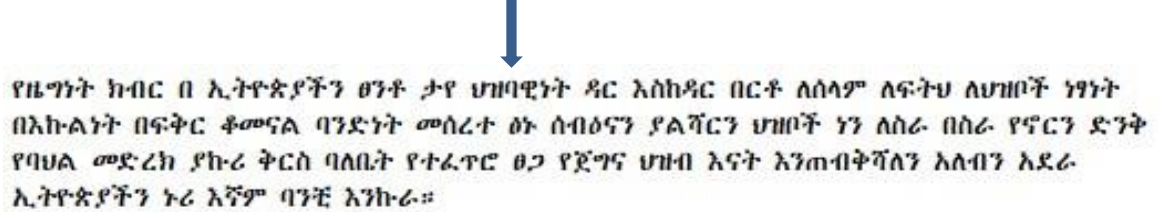


Figure 4: 7 Comparison of OCR performance by the base and new models

4.12. Discussion of Results

The approach demonstrates how continuous training with relatively little training data from printed texts may minimize the Tesseract character error with the amh train.traineddata's base model. The change in mistake rate is significant when viewed qualitatively. However, it requires a big training sample to increase accuracy.

4.13. Text Analysis

The language is mined for information using a text analysis subsystem. Includes prosodic details from the source text. Using the word order rule, the software inputs and repeats the word order. It is simple to switch between transcription using letter-to-sound mapping. Others could classify Amharic as a phonetic language. a straightforward connection between phonology and spelling.

Prosody modeling at its best for any language. For the advancement of high-quality speech processing, it is the most crucial problem. In the Amharic language, the length of the segments is a crucial and significant factor in the management of prosody. Amhara rhythm is mostly characterized as long and short, which seems to be distinct from the English language where accent is primarily determined by stress (vowel). With certain phrasal qualities and in accordance with the way the consonants are formed. To imitate word duration in TTS Amharic language is crucial for this reason. Rule-based compact prosodic generation is used in well-known TTS. The method consists of three steps:

- Consonant duration modeling for PYDUB.
- Managing the sixth order's duration.
- Assignment of the international intonation contour.

Prosody in the difference between sounds and the relative length of speech parts our office It merely covers the fundamentals of prosody. similar to how long the words are and how the phrase ends. PYDUB is envisioned as a long-term residence as indicated by its name. The length

of words is represented as the length of phrases. beneficial for enhancing prosodic quality. Prosodic They are identified in an easy manner utilizing terms. punctuation in text. In order to incorporate the overlapping F0 PYDUB pitch model There are two prosodic modules: phrase level and level.

4.13.1. Syllables connection

We use input-based communication rules for component relations. They consist of syllables. Cepstrum measures are used to represent it. The connections They are used by flexing or smoothing. Cepstral amplitude, F0, and coefficients inside the bounds. There are typically two different syllabic joining patterns. It is the first design where the level goes that has a fluffy and serial connection. Within scope, there is visual integration. Perimeter When the limit is tone-deaf, this pattern appears. An additional pattern of interpolation is an interpolation, which takes place when one or both of the restrictions are established. Syllables are joined to form phonemes. Phonemes are glottal stops or borderline stops if he is. After that, glottal stop pauses of 40 ms length are placed between each pre-plosive closure.

4.13.2. Evaluation and Discussion

We conduct two different sorts of listening tests in our system to assess naturalness and intellect. It was time for the first listening exam. The second listening test was used to evaluate the naturalness of synthesized phrases while also assessing the cleverness of the words combined with the system. We employed listening tests to assess how well our prosodic law employees performed.

4.13.3. Speech Materials

There are 200 words in the first hearing exam that were chosen at random from the Amharic alphabet. The terms chosen are ones that are regularly used and accepted in daily activities. 80 words or 40% of the 200 words that were chosen are gimmicky terms. This proves given how frequently gyms are used. Consequently, two categories are merged by utilizing these terms. Speech information ready: Rule-based synth sounds and analysis/synthesis of Amharic-language TTS voices. The primary exam also includes Speech sounds. the reason for the similarity. We'll utilize five for the second listening exam. sentences using words from the first, sixth, or both orders. From the input text data, the phrases are chosen. Three different forms of speech data will be prepared: original sentences, sentences that have been analyzed, and sentences that have been blended into our system using prosodic principles. We generally prepare 15 speech sounds.

4.13.4. Methods

Both Four persons from Ethiopia who speak the language fluently underwent both listening exams (2 female and 2 male). All of them listen. born and raised in the capital city of Ethiopia; aged 20 to 35. A brief introduction was provided before the hearing comprehension exam for both listening tests, and listening test programs were created for each. Listeners were given the opportunity to hear the first test by hearing each voice once every four seconds. In the space provided on the answer sheet, write the Amharic letters that correspond to the word you heard. The 15 sentences are played simultaneously and randomly for each participant in the second listening test. Additionally, each subject hears 15 sentences and assigns listening evaluation scores. a test procedure that offers the following quality measurement: (very good), (4 good), (3 fair), (2 weak), and (1 awful). by taking into account the natural feature of the system. All listeners We took the listening exam fifteen times. Using the latest 10 tests' outcomes as training, and the first five as tests.

Table 4: 6 Performance measure evaluation of Amharic synthesizer

Performance Measure						
	Correctly Pronounced		Partially Pronounced		Not Correctly Pronounced	
	In number	In %	In number	In %	In number	In %
SWs	37	95	6	12	7	14
NSWs	21	73	3	10	6	17.3
Average %	85.5		11		15.6	

4.13.5. Results and discussion

All listeners' replies are compiled, and we compute Discover the average numbers and the results that follow. The average right quantity for original and analysis-synthesis sounds in the first hearing test was the votes were 98% on a rule-based basis, making them 100%. We discovered the formula. The language is quite brilliant. The main and analysis/summary sentences scored an average of 5.0 and 4.7 on the second hearing test End of opinion (MOS) for complex sentences, respectively. The outcome demonstrates how successful and attractive the prosodic control mechanism employed in our system is. However, relying just on the Duration modeling may not be enough. replicate the real sound faithfully. Correct intonation and appropriate word use rules must also be modeled. then studying an example of an intonation curve for modeling words to

increase syllable and level prosody in communication. It's crucial to use high-quality audio components to mix excellent speech.

CHAPTER FIVE

5. CONCLUSIONS AND FUTURE WORKS

5.1. Conclusions

The studies carried out have shown good findings, confirming the applicability of the suggested research. A variety of algorithms and methodologies for printed document recognition, page layout, and document conversion to audio systems have been tested in this study.

The prototype approach's studies have shown positive results, therefore creating a comprehensive printed documents to audio system for Amharic script using an integrated Tesseract OCR engine, Leptonica Library, PyQt5 library, and OpenCV may be the best course of action. In other words, Tesseract, which is wholly written in the C++ programming language, may be used in conjunction with the OpenCV libraries (of adaptive Threshold, Bicubic Interpolation technique) and Leptonics library. Tesseract is totally written in the C++ programming language. Optimization is the biggest shortcoming.

The duality challenge is not where the Adaptive Thresh Gaussian approach is primarily optimized. This method's disadvantage is that it requires expensive processing, making it unsuitable for real-time applications. Similar to this, the Bicubic Interpolation method detection and correction technique is improved. Due to its ease of use and often superior outcomes, bicubic interpolation is a widely used technique in image interpolation. However, owing of the horizontal and vertical interlacing, distortions like blocking, blurring, and noise are readily present around the margins. Consequently, a further optimization technique with a substantial training sample is required. Furthermore, the foundation for all tasks' objective performance evaluation techniques is ground truth. The assessment's veracity, though, is crucial.

Along with the optimization issue already discussed, which requires additional research, the extension work associated with the suggested written recognition to Audio Converter system also need more focus.

5.2.Future Works

There is still much to learn about the issue of printed recognition for printed documents. Future extension projects to move the existing study on the issue of printed document proof for printed Amharic documents forward are outlined in the following manner.

A script is a way to represent a language. Typically, the detected characters are meaningful words rather than random sequences. Similar to this, word combinations frequently transmit meaning and create sentences that are syntactically, grammatically, and semantically consistent. In other words, lexical sequences must be created using the end transcripts. Future study should think about including statistical language models or natural language processing (NLP) into the recognition process.

Another thing to keep in mind is that the process of extracting a distinct manual feature is sped up by strong automatic feature deep learning. However, it is necessary to look at the results of combining some tasks with automatically derivable elements and printed papers.

The most recent OCR engines can read documents with a fair amount of accuracy, although they could struggle with damaged papers. Therefore, it seems sensible to look at a different alternative approach to the issue of printed document recognition. So, finding keywords is another potential extension work. The aim of keyword spotting identification is to locate particular word groups within an image. The ability of neural networks to discriminate between raw images and keywords while ignoring the rest makes their discriminative capabilities relevant for keyword searches.

Reference

- Fischer, A. (2012). Handwriting Recognition in Historical Documents. PhD thesis, Universitat Bern.
- Abay Teshage, R. Sethuraman, “Artificial Neural Network Approach to the Development of OCR for Real Life Amharic Documents”, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, January 2015
- Abay Teshage, R. Sethuraman, “Artificial Neural Network Approach to the Development of OCR for Real Life Amharic Documents”, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, January 2015
- Alessandro, N. D. (2009). “An Introduction to Text-to-Speech Synthesis”, Laboratoire de Théorie des Circuits et Traitement de Signal Faculté Polytechnique de Mons. Language and Science.
- Ann K. Syrdal, Gregor Mohler, Kurt Dusterho, Alistair Conkie and Alan W. Black, “Three Methods of Intonation Modeling”, AT&T Labs – Research, Florham Park, NJ, USA, 1999
- Anne and Klaus Wedekind, “Amharic stress (beat) rules of linguists, poets and singers: Which beat rules beat which?”, the 11th International Conference of Ethiopian Studies Addis Ababa, 1991.
- Bataineh, B., Abdullah, S. N. H. S., Omar, K., & Faidzul, M. (2011). Adaptive Thresholding Methods for Documents Image Binarization. In J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, C. Ben-Youssef Brants, & E. R. Hancock (Eds.), Pattern Recognition (Vol. 6718, pp. 230–239). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21587-2_25
- Berhanu, A. (1999). Amharic Character Recognition Using Artificial Neural Networks. MSc. thesis, Addis Ababa University.
- Birhanu et al. (2019a). Factored Convolutional Neural Network for Amharic Character Image Recognition. 2019 IEEE International Conference on Image Processing (ICIP). IEEE.

doi:10.1109/ICIP.2019.8804407

Bluche, T. (2015). Deep Neural Networks for Large Vocabulary Handwritten Text Recognition. PhD thesis, Université Paris-Sud.

C. Bento, “Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis,” Medium, Sep. 30, 2021.

D. Bradley and G. Roth, “Adaptive Thresholding Using the Integral Image,” J. Graph. GPU, Game Tools, vol. 12, no. 2, pp. 13–21, 2011.

Dutoit, Thierry. (2008). “A Short Introduction to Text-to- Speech Synthesis”, TTS research team, TCTS Lab. Academic research, 1-16.

E. Kavallieratou & E.Stamatatos, “Improving the Quality of Degraded Document Images”, Proceedings of the Second International Conference on Document Image Analysis for Libraries in University of the Aegean, IEEE, 2006

Fitehalew, A. (2019). Ancient Geez Script Recognition Using Deep Convolutional Neural Network. MSc. thesis, Near East University.

Fitsum Demissie, "Developing Optical Character Recognition for Ethiopic Scripts", Master Thesis Computer Engineering Nr: E4031D, 2011

G.Vamvakas, B.Georgios, et al, “A complete optical character recognition methodology for historical documents", Document Analysis Systems, 2008

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press Cambridge, Massachusetts, London, England

H. Yazid and H. Arof, “Gradient based adaptive thresholding,” J. Vis. Commun. Image Represent., vol. 24, no. 7, pp. 926–936, Oct. 2013

Henock, L. (2003). “Concatenative TTS for Amharic Language”, Addis Ababa University, faculty of

- Informatics', School of Information Science, Addis Ababa, Ethiopia.
- Hopfield, J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences*, 79(8), pp. 2554-8.
doi:10.1073/pnas.79.8.2554
- Jain, A., Duin, R., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *Transactions on pattern analysis and machine intelligence*, 22(1).
- Joyce McDonough, "The Prosody of Interrogative and Focus Constructions in Navajo", Department of Linguistics and Center for the Language Sciences, University of Rochester, Amsterdam, 2002.
- Juergen Schroeter, "Text to-Speech (TTS) Synthesis", AT&T Laboratories, 2005
- Jun Xu, "The Prosody of Interrogatives at Transition-Relevance Places in Mandarin Chinese Conversation", Master of Research Thesis, University of Nottingham, March 2008
- K. Khurshid, I. Siddiqi, C. Faure, and N. Vincent, "Comparison of Niblack inspired Binarization methods for ancient documents," *Proc. SPIE-IS&T Electron. Imaging*, vol. 7247, pp. 1–9, 2009.
- Kasie, E. B (2012) Concatenative Speech Synthesis for Amharic Using Unit Selection Method, Addis Ababa Ethiopia.
- Khurshid, K. (2009). Analysis and Retrieval of Historical Document Images. PhD thesis, Université Paris Descartes.
- Lasko, Laskov. "Classification and Recognition of Neume Note Notation in Historical Documents," *Proceedings of the International Conference of Computer Systems*, Vol. 4. 2006
- Lemmetty, S. (1999). "Review of Speech Synthesis Technology", Master's Thesis, Helsinki University of Technology.
- M. A. Bakhshali, "Segmentation and enhancement of brain MR images using fuzzy clustering based on

- information theory,” *Soft Comput.*, vol. 20, pp. 1–8, 2016.
- McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* volume, 5, 115–133. doi:10.1007/BF02478259
- Mieko Takada and Takako Ayusawa, “The Intonation of Interrogative Utterances in the Japanese Dialogs — Analysis of the TUFS Language Module”,
- N. Otsu, “A Threshold Selection Method from Gray-Level Histograms,” in *IEEE Transactions On Systems, Man, And Cybernetics*, 1979, vol. 20, no. 1, pp. 62–66.
- Nobuyuki Otsu, A Threshold Selection Method from Gray-Level Histograms, *Systems, Man and Cybernetics IEEE Transactions on* , vol.9, no.1, pp.62-66, Jan. 1979
- Ntogas, N., & Ventzas, D. (2008). A binarization algorithm for historical manuscripts. 41–51.
- Plamondon, R., & Srihari, S. N. (2000). On-line and Off-line Handwriting Recognition: A Comprehensive Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22,
- Prachi, Solanki, and Malay Bhatt, "Printed Gujarati script OCR using hopfield neural network", *International Journal of Computer Applications*, 2013
- R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Third. Upper Saddle River, NJ, USA: Prentice Hall, 2008.
- R. Meyer, (2006) “Amharic as lingua franca in Ethiopia,” *Lissan J. Afr. Lang.* Addis Ababa University
- Ralf Kompe, “Prosody Speech Understanding Systems”, *Subseries of Lecture Notes in Computer Science*. Springer-Verlag, Berlin, 1997.
- Sadeghipour, Z., Babaie-Zadeh, M., & Jutten, C. (2009). An adaptive thresholding approach for image denoising using redundant representations. 1–6.
- Sertse Abebe, “Bilingual Script Identification for Optical Character Recognition of Amharic and

- English printed document”, (Mater’s thesis), School of Information Science Addis Ababa University, June 2011
- Shafii, M. (2014). Optical Character Recognition of Printed Persian/Arabic Documents. PhD thesis, University of Windsor.
- Siranesh, G. (2016). Ancient Ethiopic Manuscript Recognition Using Deep Learning Artificial Neural Networks. MSc. thesis, Addis Ababa University.
- Siranesh, G. (2016). Ancient Ethiopic Manuscript Recognition Using Deep Learning Artificial Neural Networks. MSc. thesis, Addis Ababa University.
- T. Anberbir and T. Takara, “Development of an Amharic text-to-speech system using cepstral method,” in Proceedings of the First Workshop on Language Technologies for African Languages - AfLaT ’09, Athens, Greece, 2009, p. 46. doi: 10.3115/1564508.1564517.
- V. Sokratis et al.,” A Hybrid Binarization Technique for Document Images”, at Dept. of Information and Communication Systems Engineering in University of the Aegean, Learning Structure and Schemas from Documents, SCI 375, pp. 165–179, 2011
- W. A. Mustafa, H. Yazid, and S. Yaacob, “Illumination Normalization of Non-Uniform Images Based on Double Mean Filtering,” in IEEE International Conference on Control Systems, Computing and Engineering, 2014, pp. 366–371.
- W. A. Mustafa, H. Yazid, M. Jaafar, M. Zainal, A. S. Abdul-, and N. Mazlan, “A Review of Image Quality Assessment (IQA): SNR, GCF, AD, NAE, PSNR, ME,” J. Adv. Res. Comput. Appl., vol. 7, no. 1, pp. 1–7, 2017.
- Wellner, “Adaptive thresholding for the digital desk,” Tech. Rep. Eur., pp. 93–110, 1993.
- Worku, A. (1997). The Application of OCR Techniques to the Amharic Script. MSc. thesis, Addis Ababa University.

Yamagishi, J., Kobayashi, T., Renals, S., King, S., Zen, H., Toda, T., & Tokuda, K. (2007). “Improved Average-Voice-based Speech Synthesis Using Gender-Mixed Modeling and a Parameter Generation Algorithm Considering GV.” Challenge.

Yoshiko Fujisaki, “Sonority and Its Role for Syllabification”, Department of Humanities, Natural Language, 1995.

Zervas, P., Potamitis, I., Fakotakis, N., & Kokkinakis, G. (2001). “A Greek TTS Based on Non Uniform Unit Concatenation and the Utilization of Festival Architecture”, Wire Communications Lab, Department of Electrical & Computer Engineering University of Patras, 26500, Rion, Patras, Greece. Challenges.

