بسم الله الرحمن الرحيم

Sudan University of science and technology

College Of Graduate Studies

College of Computer Science and Information Technology

A Proposed Automatic Speech Recognition model for the Sudanese Dialect

نموذج مقترح للتعرف الآلى على الأصوات فى اللهجة السودانية

A Thesis Submitted in Partial Fulfillment of the Requirements for MSc degree in Computer Science

By:

Ayman Abdelaziz Elhassan Mansour

Supervised by:

Dr. Wafaa Faisal Mukhtar

September 2020

الآية

قال تعالى:

(يُؤْتِي ٱلْحِكْمَةَ مَن يَشَآءُ ۚ وَمَن يُؤْتَ ٱلْحِكْمَةَ فَقَدْ أُوتِيَ خَيْرًا كَثِيرًا ۗ وَمَا يَذَّكَّرُ إِلَّآ أُو۟لُوا۟ ٱلْأَلْبَٰبِ)

البقرة [٢٦٩]

Dedication

I would like to express my gratitude to my family; my father who always inspires me to execute to my utmost potentials, my mother for she is the pillar that is always there to offer support throughout my endeavors, my friends who regularly check on me and exchange advice, my mentors whom I owe all my success till this moment.

Sincerely grateful...

Acknowledgment

I wish to express my deepest appreciation to all those who helped me, in one way or another, to complete this project. First and foremost I thank Allah almighty who provided me with strength, purpose, and success throughout this endeavor. Special thanks to my supervisor Dr. Wafaa Faisal Mukhtar for all her patience, expert guidance, and support during the execution of this research. And also Dr. Nizar Habash and Dr. Ahmed Ali for their consultation and advice. My thanks also go to Dr. Nawar Halabi and Mr. Kamil Ciemniewski for making their resources available for researchers and students.

Declaration

I hereby declare that the work reported in this MSc thesis titled "**A Proposed Automatic Speech Recognition model for the Sudanese Dialect**" submitted for the Sudan University of Science and Technology, is an authentic record of my work carried out under the supervision of Dr. Wafaa Faisal Mukhtar. And never has been submitted elsewhere for other degrees.

Ayman Abdul-Aziz Elhassan Mansour

Dr. Wafaa Faisal Mukhtar
Supervisor

Abstract

Nowadays, speech recognition plays a major role in designing a natural voice interface for communication between human and their modern digital life equipment. It is presenting an easy way to cross the language barrier between monolingual individuals. But the obvious problem with this field is the lack of wide support for several universal languages and their dialects; while most of the daily interaction is done using them.

This research comes to ensure the viability of designing the Automatic speech recognition model for the Sudanese Dialect. The researcher focused on building a dataset by collecting represented resources and perform pre-processing to construct the dataset. The Automatic speech recognition model was built by training the model to recognize each character of the Sudanese Dialect. The model's architecture followed the end-to-end speech recognition approach. Each building block of the model was formed using Convolution Neural Networks rather than Recurrent Neural Networks, the usual choice of the speech-related task, and the training was done using the Connectionist Temporal Classification learning algorithm.

In this research, a Sudanese dialect dataset was built overcoming the lack of annotated data and reached an average label error rate of 73.67%. The proposed model will enable the use of the collected dataset in any Natural Language Processing future research targeting the Sudanese Dialect. The designed model, with its performance, provided some insights about the current recognition task. The model can reach a much better label error rate by deploying any improvement such as a language model. The applications for this research are vastly available from designing archives for the Sudanese content with its text format to develop real-time speech recognizer.

المستخلص

في الوقت الحاضر، يلعب التعرف على الكلام دوراً رئيسياً في تصميم واجهة صوتية طبيعية للتواصل بين كل من الإنسان وجميع معدات الحياة الرقمية الحديثة (الهواتف، الأجهزة القابلة للإرتداء، التليفزيون، السماعات، السيارات، المنازل الذكية، إلخ). والأهم من ذلك أنه يقدم طريقة سهلة لعبور حاجز اللغة بين الأفراد متكلمي اللغة الواحدة. لكن المشكلة الواضحة في هذا المجال هي عدم وجود دعم واسع للعديد من اللغات العالمية ولهجاتها بينما يتم إستخدامها أثناء التفاعل اليومي.

ركز الباحث على بناء قاعدة بيانات من خلال جمع موارد تعكس اللهجة السودانية وإجراء العديد من عمليات المعالجة المسبقة، ثم تصميم نموذج للتعرف التلقائي على الكلام للإستفادة من قاعدة البيانات التي تم بناءها من خلال تدريب النموذج على التعرف على كل حرف من اللهجة السودانية، إتبعت بنية النموذج نهج end-to-end للتعرف على الكلام. كل لبنة بناء من هيكلية النموذج تم تصميمها باستخدام Convolution Neural Networks بدلاً عن Recurrent Neural Networks والتي تمثل الاختيار المعتاد للمهام المتعلقة بالكلام، وقد تم التدريب باستخدام خوارزمية Classification Connectionist Temporal للتعلم.

حقق هذا البحث أهدافه من خلال بناء قاعدة بيانات للهجة السودانية للتغلب على نقص البيانات وحقق النموذج متوسط معدل خطأ (LER) **73.67** ٪. ستمكن نتائج هذا البحث من استخدام قاعدة البيانات المجمعة في أي بحث مستقبلي يستهدف اللهجة السودانية في مجال معالجة اللغات الطبيعية، قدمت النتائج بعض الرؤى حول مهمة التعرف الحالية، في المستقبل القريب يمكن ان يحقق النموذج معدل خطأ أفضل بكثير من خلال تجربة أي تحسين مثل إستخدام نموذج اللغة. وتتوفر لهذا البحث تطبيقات بشكل كبير مثل تصميم أرشيف للمحتوى السوداني بتنسيق نصي، وتطوير أداة للتعرف على الكلام في الوقت الفعلي.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

TERM	MEANING
ASR	Automatic Speech Recognition
MSA	Modern Standard Arabic
CTC	Connectionist Temporal Classification
OOV	Out-Of-Vocabulary
WER	Word Error Rate
E2E	End-To-End
HMM	Hidden Markov Model
CNN	Convolution Neural Networks
RNN	Recurrent Neural Networks
CODA	Conventional Orthography For Dialectal Arabic
LSTM	Long Short-Term Memory
DFT	Discrete Fourier transform
LVCSR	Large Vocabulary Continuous Speech Recognition

TERM	MEANING
RCA	Radio Corporation of America
MIT	Massachusetts Institute of Technology
CMU	Carnegie Mellon University
HTK	Hidden Markov Model Toolkit
GALE	Global Autonomous Language Exploitation
DARPA	Defense Advanced Research Projects Agency
PRLM	Phone Recognition followed by Language Modeling
DBN	Deep Belief Network
BP-DBN	Backpropagation Deep Belief Network
AM-DBN	Associative Memory Deep Belief Network
PER	Phone Error Rate
SGMM	Subspace Gaussian Mixture Model
ICSI	International Computer Science Institute
ALASR	Arabic Loria Automatic Speech Recognition

TERM	MEANING
TARIC	Tunisian Arabic Railway Interaction Corpus
TIMIT	Texas Instruments/Massachusetts Institute of Technology
LDC	The Linguistic Data Consortium
TDT-4	Topic Detection and Tracking
CSV	Comma-Separated Values

List of Equations

CHAPTER Ⅰ

Introduction

1.1 Background

Automatic Speech Recognition (ASR) is a key technology for a variety of industrial and IT applications. ASR is playing a growing role in a variety of applications, such as hands-free operation and control, automatic query answering, telephone communication with information systems, automatic dictation (speech-to-text transcription), government information systems, etc. Speech communication with computers, PCs, and household appliances is envisioned to be the dominant human-machine interface (AbuZeina et al., 2011).

Arabic is a Semitic language, and it is one of the oldest languages in the world. It is the fifth widely used language nowadays. Standard Arabic has 34 basic phonemes1, of which six are vowels, and 28 are consonants. Arabic has fewer vowels than English has. It has three long and three short vowels, while American English has at least 12 vowels(Satori et al., 2007).

Recognition research on Arabic compared to other languages. The first works on Arabic ASR have concentrated on developing recognizers for Modern Standard Arabic (MSA). The most difficult problems in developing highly accurate ASRs for Arabic are the predominance of non-diacriticized text material, the enormous dialectal variety, and the morphological complexity.

Sudanese Dialect is the product of complex historical and social conditions. This complexity is reflected in both the form and content of the Dialect. Language, however, is not divorced from the life of the people using it. It actually reflects their cultural existence. The same complexity is discernible in Sudanese life and hence, in Sudanese character. It is evident that the backbone of the Dialect is Arabic. Even many Arabic words and concepts were adapted to suit the conditions of life in Sudan. However, the fact remains that Arabic or acclimatized Arabic is the dominant element. This very dominance tends, however, to over-shadow the important contributions of other non- Arab(Gasim, 1965).

1.2 Problem Statement

The Sudanese dialect as well as all Arabic language dialects suffer from the lack of annotated resources and tools which are needed for ASR development. Giving the fact that among the Arabic dialects, there is no work accommodated to represent or implement the Sudanese dialect in any system or application.

1.3 Research Questions

This research comes to answer the following questions:

- Do the existed Modern Standard Arabic speech recognition models have the ability to recognize the Sudanese dialect?
- Which speech recognition model's types is better for the giving recognition task?

1.4 Objectives

This research is going to deal with the stated problem by applying speech recognition to the Sudanese dialect. The main objective of this research is to design a speech recognition model for the Sudanese Dialect, which will be achieved by:

- Collecting recordings and textual data that reflect and represent the Sudanese Dialect.
- Building a simple dataset to overcome the lack of resources and make them available to future researches.
- Designing a model to recognize the speech and map it into a textual format, to fulfill the aim of this research.

1.5 Methodology

The methodology of this research to design the ASR model is to find a way to collect data and recordings then pre-processing data for it is a necessary step for preparing them for the next stage.

Second, taking the data from the last step and designing a simple dataset to make the data organized and ready for the module to train and improve using it.

By using the breakthrough technology of Deep Neural Networks to benefit from its flexibility, performance, and efficiency handling the data, so it does not require as much data to achieve the same performance of other ASR technologies,

therefore a model is going to be built using the same method. Besides, apply it in the field of Arabic language Recognition, mainly the Sudanese Dialect.

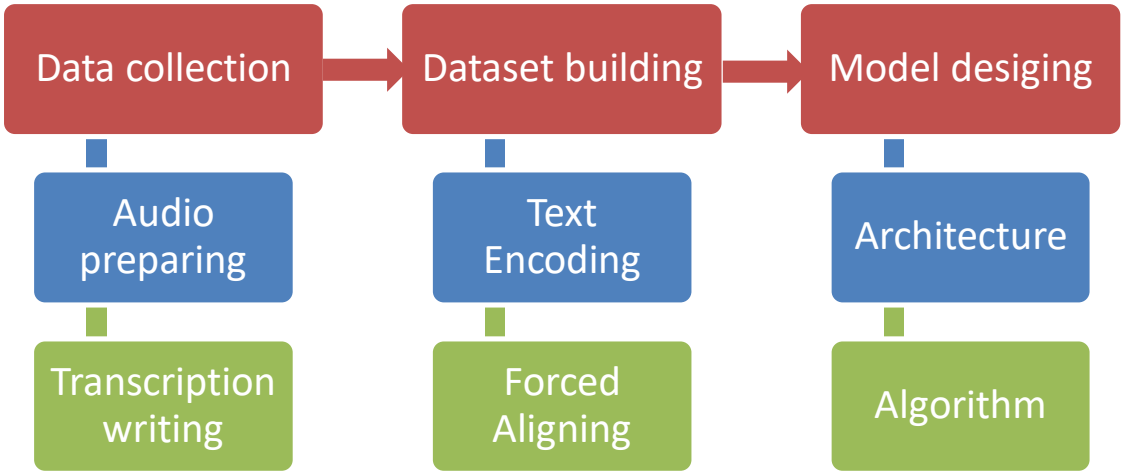

Figure 1-1: Research Methodology

Figure 1-1 shows the chosen methodology to accomplish the objectives of this research each of the objectives is concomitant, the data collection phase is followed on by dataset building and finally model designing.

1.6 Research Scope

This research aims to deal with one of the Artificial Intelligence applications, which is Automatic Speech Recognition, and the use of it in the dialectal Arabic language mainly the Sudanese dialect.

1.7 Thesis Organization

This research organized as follows:

Chapter I contains the research problem statement and objectives. Chapter II discusses the literature review and related work. Chapter III describes the research methodology and the implementation of the techniques used. Chapter IV presents the Experiments and their results. Finally, Chapter V concludes this research and presents Recommendations for future works.

CHAPTER II

Literature Review

2.1 Introduction

This chapter gives some ideas and reviews of the techniques related to Automatic speech recognition showing some of the obsolete methods from the commencement of the field in the 1950s, until the present and the state of the art techniques. Also includes insights about the structure, major components, and functionality of an automatic speech recognizer.

Alongside several linguistic information about the Arabic language and its dialects and their geographical distribution, and some aspects and origins of the Sudanese dialect. Lastly mentions the literature and related work both in Modern Standard Arabic and dialectal speech recognition and gives a summary for all of them. Finally, discuss the proposed model design from its structure and the proposed learning algorithm is going to be used in the empirical part of this research to train and optimize the automatic speech recognition model.

2.2 Automatic Speech Recognition Technologies

In the era of “OK Google...”, “Alexa…” and “Hey Siri...”, the rising of personal digital assistants and their voice-enabled interface is staggering, even since the time of Stanley Kubrick and Arthur C. Clarke masterpiece 2001: A Space Odyssey and their famous computer HAL, a voice interacted machines have foreseen to become a reality.

The begging of Automatic Speech Recognition researches was in the 1950s when various researchers tried to exploit the idea of acoustic-phonetics. In 1952, researches in Bell Laboratories built a system for a single speaker that recognize isolated digit, followed by research by RCA laboratories in 1956 to distinct 10 syllables talker. In 1959 at University College in England, a recognizer was built to recognize four vowels and nine consonants, in the same year at MIT Lincoln Laboratories a system for recognizing 10 vowels in a speaker-independent manner(Rabiner and Juang, 1993).

The evolution and milestones of Automatic speech recognition from 1960 to 2002 have been summarized in figure 2-1, besides some of automatic speech recognition types from (small vocabulary acoustic phonetics based to very large vocabulary semantic multi-model based).

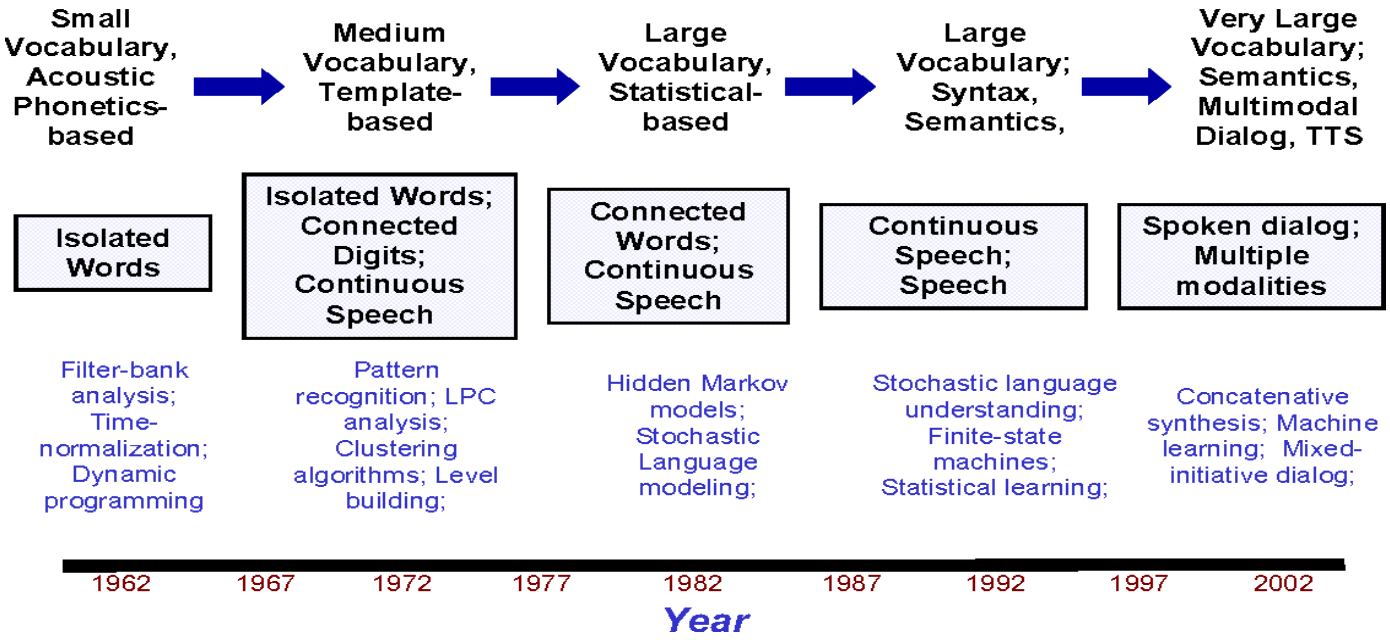

Figure 2-1: Milestones in Speech Recognition from 1960 – 2002 (Francisco et al., 2020)

In 2006 when the concept of neural network has revisited by some researchers and by the arrival of new and powerful hardware (CPUs and GPUs), the algorithms of training neural networks proved that it has the potentials to be more efficient and out-performed the previous methods of speech processing. Since then the machine learning community began to use and build-up on the same ideas and implement them for a wide range of researches, then the era of deep learning began.

Finally, at the same time in 2006, Alex Graves proposed Connectionist Temporal Classification CTC, which allows for training an acoustic model without the need for frame-level alignments between the acoustics and the transcripts. At first, CTC used with phoneme output targets, by 2015 CD-phoneme based CTC models achieve state-of-the-art performance for conventional automatic speech recognition. These studies prepare for End-to-End Speech Recognition, which is a system that directly maps a sequence of input acoustic features into a sequence of graphemes or words. A single end-to-end trained sequence-to-sequence model, which directly outputs words or graphemes, could greatly simplify the speech recognition pipeline. Since then End-to-End Speech Recognition has been an active area of study to add

and contribute to simplifying the way in which conventional automatic speech recognition systems work. (Graves and Jaitly, 2014)

2.3 Automatic Speech Recognition System Structure

The conventional ASR system consists of several modules for acoustic modeling, pronunciation lexicon, and language modeling. All of which work to achieve the process of recognizing speech, identify a particular speaker, detect numerous dialects, etc. merely from a speech signal. (Goldenthal, 1994).

2.3.1 Automatic Speech Recognition System Components

A block diagram of the major components of an ASR system is shown in Figure 2-2 Typically, the samples of the continuous speech signal are first processed to form a discrete sequence of observation vectors. This operation is denoted by the Signal Processing block in the figure.

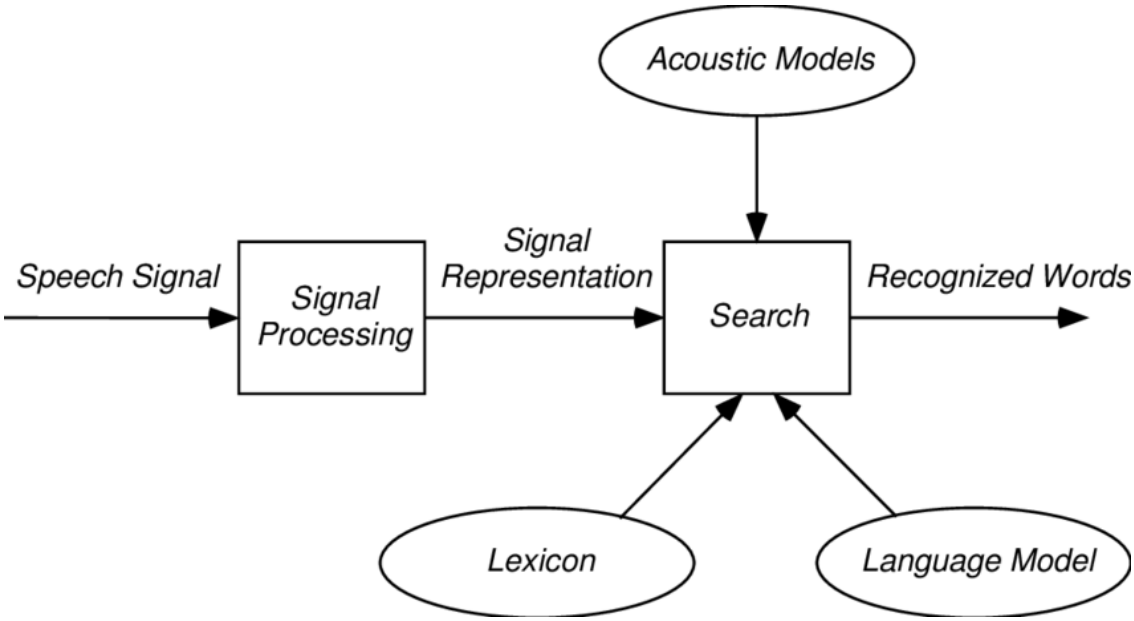

Figure 2-2: major components of an ASR system (Goldenthal, 1994)

The resulting components of the observation vectors are the acoustic attributes that have been chosen to represent the speech signal. DFT discrete Fourier transformation based on spectral coefficients or auditory model parameters.

Each observation vector called a frame of speech and the sequence of T frames comprises the Signal Representation

$$X = \{\vec{x}1, \vec{x}2, \ldots, \vec{x}T\} \qquad \text{Eq. 2-1}$$

- A search is then conducted over the frame sequence, X, to produce hypothesized word sequences.
- Acoustic models are used to score the individual frames or multiple frame sequences, known as segments.
- Language models, which contain information about allowable sequences of speech units in the lexicon (e.g. phones, words, etc.), also incorporated into the scoring process. (Goldenthal, 1994).

2.3.2 Automatic Speech Recognition System Functionality

The representation, models, search, and scoring procedures are key design components of the system. As the number of words in the lexicon becomes large, the task of training individual acoustic models for each word becomes prohibitive. Consequently, an intermediate level of representation is generally used. A common representation involves describing the pronunciation of a word in terms of phonemes. A phoneme is an abstract fundamental unit of a language. By definition, changing a phoneme changes the meaning of a word. For example, if the phoneme /p/ in the word pit changed to a /b/, the word becomes a bit. (Goldenthal, 1994).

The acoustic variability that can occur when realizing the same phoneme is part of what makes the task of identifying a phoneme so challenging, the acoustic models generally trained to recognize some set of phones (the exact set being a design decision). The task of decoding a phone sequence known as phonetic recognition and the resulting output is a sequence of probabilities from which phonetic transcriptions are hypothesized. The phonetic probabilities are of fundamental importance to the ASR task since they are the foundation upon which the word string search is based. (Goldenthal, 1994).

All large vocabulary speech systems utilize phonetic models as a component in the speech recognition system. (Goldenthal, 1994).

2.4 Arabic Language and Arabic Dialects

“ Arabic is an official language for more than 22 countries also the religious instruction in Islam (Holy Quran) ” (Kirchhoff et al., 2002).

“Arabic is a semantic language and one of the oldest languages in the world, it the fifth widely used nowadays” (Satori et al., 2007).

As mentioned in (Habash, 2010) the Arabic language can be classified into three main categories :

- Classical Arabic: the older form and it is the language of the Holy Quran transcript.
- Modern Standard Arabic: contains modern vocabularies than classical form, used in media (newspapers, radio, and TV) courtrooms, etc.
- Dialectal Arabic (colloquial): has a weak connection to the classical form, formed from several ancient dialects and foreign languages (colonial), local, daily langue fairy tales, and traditional songs. Usually spoken but not written.

Arabic dialects vary according to their speakers and other factors, as shown in figure 2-3 which describes Arabic Dialects based on the geographical dimensions of the Arabic world.

The following table 2-1 clarify with additional details Arabic dialects and their regions alongside their speakers and subcategories.

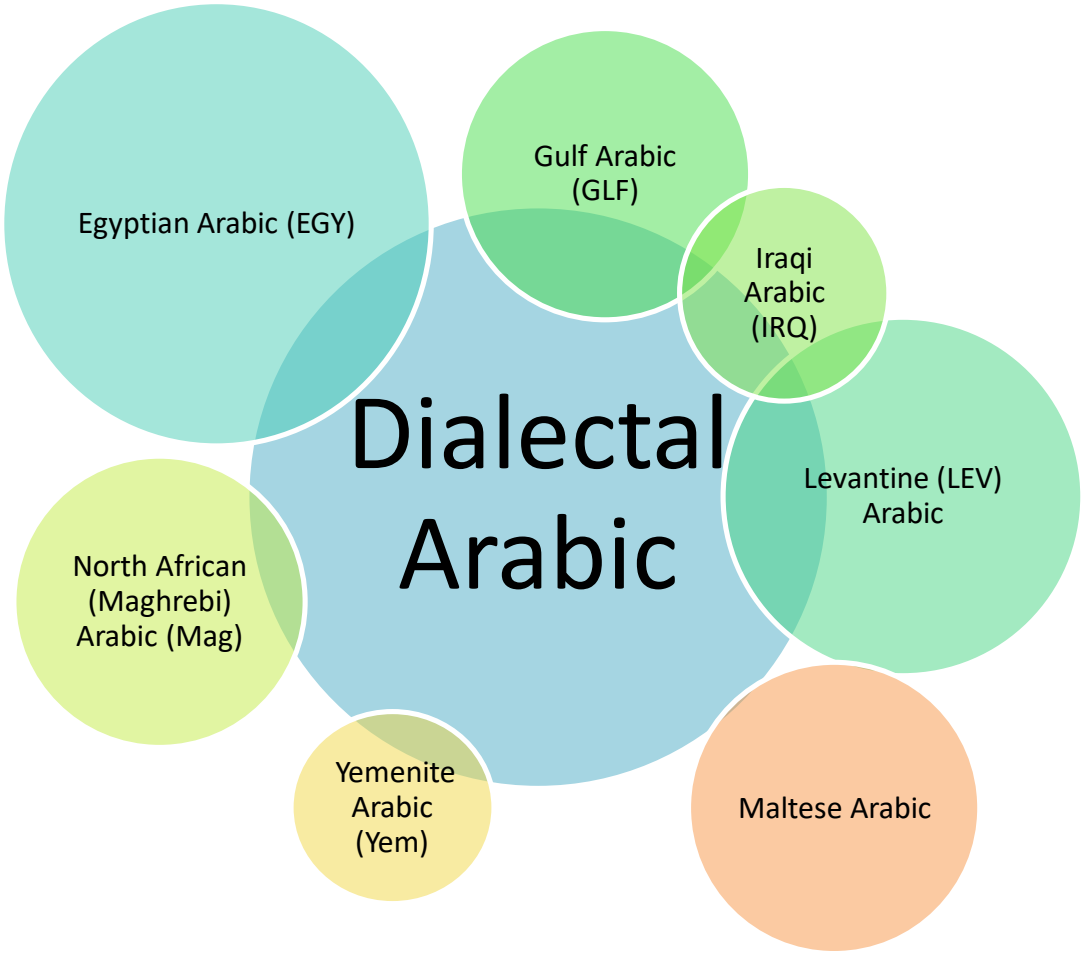

Figure 2-3: Arabic Dialects

Table 2-1: Description of Arabic Dialects

Dialect	Description
Egyptian Arabic (EGY)	Covers the dialects of the Nile valley: Egypt and Sudan.
Levantine (LEV) Arabic	Includes the dialects of Lebanon, Syria, Jordan, Palestine, and Israel.
Gulf Arabic (GLF)	Includes the dialects of Kuwait, the United Arab Emirates, Bahrain, and Qatar. Saudi Arabia is typically included although there is a wide range of sub-dialects within it. Omani Arabic is included some times.
North African (Maghrebi) Arabic (Mag)	Covers the dialects of Morocco, Algeria, Tunisia, and Mauritania. Libyan Arabic is sometimes included.
Iraqi Arabic (IRQ)	Has elements of both Levantine and Gulf.
Yemenite Arabic (Yem)	Often considered its own class.

Dialect	Description
Maltese Arabic	Not always considered an Arabic dialect. The only Arabic variant that considered a separate language and written with the Roman script.

2.5 The Sudanese Dialect

The coming of the Arabs into Sudan was the turning point, which produces ethnic and linguistic changes and effects on the cultural structure of this country, the Arabs movement concentrated in the center, which is very similar to the Arabian environment.

Arabic dialects vary according to the social environment and the degree to which the people are concerned and influenced by elements using tongues other than Arabic. Sudanese dialect can be considered as the product of complex historical and social conditions. Arabic is a desert language for her speakers and their resident in the Arabian desert, however, Sudan is one of the Nile valley countries, and most of its northern land is preserved for cultivation, and the dwellers do farming and shepherding. Arabic language and its vocabulary did not fulfill the Sudanese lifestyle for it is the language of Arab tribes and their environment. Hence, it has to borrow a handful amount of terms and words to adapt itself to this new territory. (Gasim, 1965)

Several myriad tongues have contributed to form this dialect such as Arabic, Nubian, old Egyptian, and others. Nonetheless, Arabic is the general language and considered the mother tongue for the nation, however; several changes prove that every tribe and region has its own version of Arabic. It is; legitimate to take the dialect of Khartoum and its vicinity as a common medium of exchange intelligible to most, if not all, who speak Arabic in Sudan especially in towns. (Gasim, 1965)

For instance, terms from the Nubian tongue were adopted to be used, particularly those relevant to the Nile, agricultural tools, some unfamiliar species of vegetation, and their products all of which were borrowed, and others of Nubian origin were incorporated into the language. (Gasim, 1965)

Table 2-2: Nubian Terms

Word	Arabic Word	meaning	origins
sãgia	ساقيا	The water-wheel	Nubian
utfa	عطفه	frame wheel carrying water-bucket chain	Nubian
akudêk		excavation in riverbank beneath water-wheel	Nubian
toreig	طوريق	horizontal driving spindle	Nubian
gurayr	قرير	[newly formed alluvial soil	Nubian
mosoore	موسور	flood season	Nubian
karu	كارو	land behind the true sãgia, liable to flood	Nubian
ingâya	إنقايا	a special agricultural plot	Nubian
wäsüg	واسوق	a special broad wooden shovel pulled by ropes	Nubian
koreig	كوريك	ordinary shovel	Nubian
khãsa	خسا	knife	Nubian
weika	ويكا	okra	Nubian
mãreig	ماريق	dura	Nubian
hangüg	عنكوك	stiff reeds used for brooms	Nubian
ashmeig	عشميق	palm fiber	Nubian

Beside Nubian, Arabic borrowed a lot from Badawi or Bejawi tongue like Animal names, important terms like in everyday items, and some food-related terms besides obscure words the origin of which is not easily recognizable. (Gasim, 1965)

Table 2-3: Badawi Terms

Word	Arabic Word	Meaning	Origins
marfa'in	مرفعين	wolf	Badawi
ba'ashüm	بعشوم	fox	Badawi
angareib	عنقريب	bed	Badawi
karkab	قرقاب	wooden slippers	Badawi
funduk	فندك	wooden mortar	Badawi
däna		pumpkin container	Badawi
suksuk	سكسك	beads	Badawi
dõf	دووف	boneless meat	Badawi
gangar	قنقر	corn ears	Badawi
unkoleib	عنكوليب	sweet stalk	Badawi
darfûn	درفون	child	Badawi
dabas	دباس	swelling of the skin caused mostly by the heat of the sun	Badawi

Word	Arabic Word	Meaning	Origins
shanab	شنب	moustache	Badawi
shallüfa	شلوفة	thick lip, originally meaning trunk of the elephant	Badawi

Old Egyptian and Coptic language has also been borrowed from, Terms related to the Nile such as damira [high flood], sheima [whirlpool], shulbãya [a species of fish], miraisi and tayyãb [southern and northern winds], are of Egyptian ancestry.

Finally, yet importantly, African languages have also been a subject of borrowing as shown in numerous words containing the sounds "nya" and "cha" derive their existence from this source. Words, the origin of which is still a mystery, such as ga'ünja [frog], tâmbïra [a kind of fish], girinti [hippo], and many others. (Gasim, 1965)

All these ancient remnants cited above and many more passed into the main-stream of the dialect, and Arabic was unable to obliterate them, because their continuance was a social necessity, since they performed special functions no other Arabic words could equally perform. It is, however, a credit to Arabic that it was flexible and elastic enough to be able to incorporate and eventually assimilate them to such a degree that their origins could not be detected without academic research. (Gasim, 1965)

The Arabic language itself has been affected by some changes whether it is structural or semantic to suit the style and the tongue of the speakers. These changes besides the way of pronouncing numerous words to ease the use of the language revamp Arabic from its classical form to be the dialect that we all speak as Sudanese people, these some of the changes:

- **Structural changes:**

Table 2-4: Structural Changes

Replacement					
Method	**Original form**	**Arabic Word**	**Dialectal form**	**Arabic Word**	**Meaning**
Replacing Dh with D	jabadh	جبده	jabad	جبد	pulled
Replacing Dh with D	dhabah	ذبحه	dabah	ضبح	slaughtered

Method	Original form	Arabic Word	Dialectal form	Arabic Word	Meaning
Replacing Th with T	thalätha	ثلاثة	talata	تلاتة	3 (three)
Replacing J with D	jaysh	جيش	daysh	ديش	army
Replacing A with ayn	ja'ar	جعر	ja'ar	جعر	lowed [for bull]
Replacing M sometimes with B	minbar	منبر	banbar	بنبر	seat
Inversion					
Method	**Original form**	**Arabic Word**	**Dialectal form**	**Arabic Word**	**Meaning**
Reversing the position of letters in the body of the word	dajja	ضجة	jadda	جضة	screamed
	nadij	نضج	nijid	نجض	ripened
	jalada	صلدة	dalaja	دلجة	hard ground
	batt	بتة	tabb	تب	absolutely
	zawäj	زواج	jawãz	جواز	marriage
Omission					
Method	**Original form**	**Arabic Word**	**Dialectal form**	**Arabic Word**	**Meaning**
The shortening of Longer words	walad	ولد	wad	ود	boy
	bint	بنت	bit	بت	girl
	imra'a	إمرأ	mara	مرا	woman
	nisf	نصف	nus	نص	half
	marhabãbik	مرحبا بك	habäbak	حبابك	you are welcome
Addition					
Method	**Original form**	**Arabic Word**	**Dialectal form**	**Arabic Word**	**Meaning**
A letter or more may be added to facilitate pronunciation	tawwaha	طوح	tõtah	طوطح	flung
	lawwaha	لوح	lõlah	لولح	waved a thing about
Assimilation					
Method	**Original form**	**Arabic Word**	**Dialectal form**	**Arabic Word**	**Meaning**
Similar or allied letters are assimilated	inta	إنت	itta	إتا	you
	gulta	قلت	gutta	قتا	you said
	Kunta	كنت	kutta	كتا	you were

Amalgamation					
Method	**Original form**	**Arabic Word**	**Dialectal form**	**Arabic Word**	**Meaning**
The processes of assimilation and abbreviation	mã 'alayka shay'	ما عليك شئ	ma'alaysh	معليش	never mind
	bilã shay'	بلا شئ	baläsh	بلاش	free of charge
	hãdhi as-sã'a	هذه الساعة	hassa'	هسا	now
	ilä as-sã'a	إلا الساعة	Lissa'	لسه	up to now

- **Semantic changes:**

Table 2-5: Semantic Changes

Word	Arabic Word	Original Meaning	Dialectal Meaning
dagala	ضقلة	a weak, slender goat	a small watermelon
zaghrada	زغرد	the groaning of camels	trilling shrills by women in weddings
zuwâ'a	زواعة	the driving and dispersing of camels	roaming about [men]
kuwãsa	كواسة	the walk on three feet of a hamstrung camel	identical with roaming
khashaba		pasturing of dry pasture	a kind regime where a person abstains from fats and salts and may take some dried herbs
janfa	جنف	she who goes astray	indicate the left-handed or the woman who puts on her "tob" in the reverse position
tashlïg	إتشلق	splitting a thing lengthwise in general	an eye operation by a local doctor
sagar		the green or black color of a bird mixed with a reddish or yellowish color	rust

The dialect has not been immune to further changes, in the 19th and 20th centuries the country has been subjected to foreign domination; Turkish and European cultures offered new avenues.

Thus the dialect has a big number of words suffixed or prefixed with the Turkish term "bãsh" [senior], in addition to all the terms of the military hierarchy, likewise, all words ending with the suffix "khãna" [place].

Table 2-6: Turkish Terms

Word	Arabic Word	Meaning	Origins
bãshkãtib	باشكاتب	senior clerk	Turkish
hakïmbâshi	حكيمباشا	senior medical officer	Turkish
amiralãy	أميرلاي	Brigadier	Turkish
bikbâshï	بيكباشا	colonel	Turkish
yüzbäshi	يوزباشا	captain	Turkish
shafakhãna	شفخانا	dressing point or dispensary	Turkish
ajzakhãna	أزدخانا	pharmacy	Turkish
adabkhäna	أدبخانا	toilet room	Turkish

The borrowings from English, French, and other European languages are vast. Thousands of these words have been assimilated beyond recognition.

Table 2-7: European Terms

Word	Arabic Word	Original form	Meaning	Origins
barnïta	برنيطا	borreta	hat	Italian
kabbût t	كبود	cappotto	overcoat	Italian
battãria	بطاريه	batteria	battery	Italian
baia		balla	bale	Italian
sigãla	سقالا	scala	wooden plank on scaffolding	Italian
askila	أسكلا	scala	quay	Italian
tãwla	طاولة	tavola	table	Italian
kimbiyãla	كمبيالا	cambiale	voucher	Italian
lakùnda	لوكندا	locanda	hotel	Italian
fätüra	فاتورة	fattura	receipt	Italian
awantaa	أونطا	aventa	trick	Italian
girish	قرش	Groschen	Piaster	German
nimra	نمره	number	number	English
fatil	فايل	phial	phial	English
sirwis	سرويس	service	service	English
budra	بودرا	powder	powder	English
maiz	ميز	mess	mess	English

Word	Arabic Word	Original form	Meaning	Origins
warsha	ورشه	workshop	workshop	English
tirilla	تريلا	trailer	trailer	English
dush	دوش	douche	shower	French

All of which discussed above, manifest that the Sudanese dialect is the product of several ingredients contributed over time to make a unique yet crucible language, similar to the Sudanese society and its various characteristics. From myriad tongues, one dialect has formed to encapsulate all the key factors (terms, words, and vocabularies) of the dialect, which until now known and used to form an easy way of communication.

2.6 Related Automatic Speech Recognition Researches

ASR for Arabic researches focused mainly on modern standard Arabic, nevertheless; the performance of most of them was not satisfactory. One of the researches in the John-Hopkins summer workshop an effort of developing novel speech recognition models for Arabic has reported; (Kirchhoff et al., 2002) explored methods of making MSA data usable for training models for colloquial Arabic and develop novel statistical models in particular language models. To exploit the small amount of training data available for dialectal Arabic by investigating three different types of language models designed to exploit the available Egyptian colloquial Arabic data (particle models, morphological stream models, and factored language models).

Language modeling is an essential part of the speech recognition process however it is a difficult problem for languages with rich morphology Arabic has a large number of affixes that can modify a stem to form words. This causes a high out-of-vocabulary (OOV) rate for typical lexicon size and Leeds to a potential increase in WER. hence (Vergyri et al., 2004) Show that the use of morphology-based LMs at different stages in an LVCSR system for Arabic leads to word error rate reductions by 1.8%. Dialectal Arabic is no exception to morphology issues for they share the Arabic rules and some of its grammar. Therefore (Afify et al., 2006) presents a word decomposition algorithm, that uses popular Arabic affixes, for constructing the lexicon in Iraqi Arabic speech recognition. The net effect is about 13% relative improvement in WER.

There are a good amount of freely available tools and systems for speech recognition such as the CMU (Carnegie Mellon University) Sphinx speech recognition system which currently is one of the most robust speech recognizers in English. CMUSphinx enables research groups with modest budgets to quickly begin conducting research and developing applications, and Cambridge Hidden Markov Model Toolkit (HTK) which is a portable toolkit for building and manipulating hidden Markov models. (Satori et al., 2007)Demonstrated the use of the CMUSphinx System to build an ASR system for Arabic language Hello_Arabic_Digit the possible adaptability of this system to Arabic speech. Also (Elshafei et al., 2008) developed an Arabic speech recognition system based on the Carnegie Mellon university Sphinx tools. And also used the Cambridge HTK tools for the process of utilizing at various testing stages. The system trained on 4.3 hours of the 5.4 hours of Arabic broadcast news corpus and tested on the remaining 1.1 hours. The Word Error Rate (WER) came to 9.0%.

Another example is the Global Autonomous Language Exploitation (GALE) project, which is a project funded by Defense Advanced Research Projects Agency (DARPA) and executed by IBM to make foreign language (Arabic and Chinese) speech and text accessible to English monolingual people, particularly in military settings. The first study in the GALE project was by (Soltau et al., 2007) they opted for a flat-start approach with pronunciations generated automatically using the Buckwalter morphological analyzer and studied some aspects related to coverage, HMM topologies, and pronunciation probabilities, these advances were instrumental in lowering the word error rate by 42% relative over one year.

The neural network model allows for more robust generalization and is be able to fight the data sparseness problem (Emami and Mangu, 2007) The NN models improved considerably over the baseline 4-gram model, resulting in reductions of up to 0.8% absolute and 3.8% relative in WER. (Alotaibi, 2008) compared to The ANN and HMM-based recognition system they achieved 99,5% and 98.1% correct digit recognition in the case of multi-speaker mode, and 94,5%, and 94.8% in the case of speaker-independent mode respectively.

Another group of researchers drew attention to language rules, grammar, and linguistic behaviors to gain an advantage in the recognition process (Biadsy et al., 2009a) obtain an improvement in absolute accuracy in phone recognition of 3.77%–7.29% and a significant improvement of 4.1% in absolute accuracy in ASR by

applying linguistically motivated pronunciation rules and the MADA morphological analysis and disambiguation tool. followed by (Biadsy et al., 2009b) study which investigated four Arabic colloquial dialects (Gulf, Iraqi, Levantine, and Egyptian) plus MSA and found that they can be distinguished using a phonotactic approach with good accuracy using Phone Recognition followed by Language Modeling (PRLM) approach.

When a group of researchers revisited the concept of training deep belief networks –later to be known as deep neural networks- the era of deep learning has started, one of them was the study by (Mohamed et al., 2009) which investigated two types of Deep Belief Network for acoustic modeling; the backpropagation DBN (BP-DBN) and the associative memory DBN (AM-DBN) architectures. DBNs consistently outperform other techniques and the best DBN achieves a phone error rate (PER) of 23.0% on the TIMIT core test set.

In the GALE project Phase 3.5 machine translation evaluation (Saon et al., 2010) presented a set of techniques for Arabic broadcast transcription that taken together lead to word error rates below 9%. Techniques like (Subspace Gaussian Mixture Model) SGMM acoustic modeling, neural network acoustic features, variable frame rate decoding, exclusion of conversational training data, and the use of unpruned n-grams language models and neural network language models. In phase 4 of the GALE project (Kingsbury et al., 2011) described improvements made over the past year that led to a word error rate of 8.9% on the 2009 evaluation data and a year-to-year, absolute reduction of 1.6% word error rate on the unsequestered 2008 evaluation data. By using context-dependent modeling in vowelized Arabic acoustic models; the use of neural-network features provided by International Computer Science Institute ICSI; Model M language models; a neural network language model that uses syntactic and morphological features; and improvements to our system combination strategy.

Followed by GALE project phase 5 machine translation evaluation (Mangu et al., 2011a) described improvements made over the past year that led to a word error rate of 7.4% on the 2011 evaluation data and a year-to-year, absolute reduction of 0.9% word error rate on the unsequestered 2009 evaluation data. New techniques that contributed to this improvement include Bayesian Sensing HMM acoustic models, improved neural network acoustic features, MADA vowelized acoustic model,

improved word and syntax neural network language models, and enhanced classing Model M and discriminative language models.

To support voice search, dictation, and voice control for the general Arabic speaking public, including support for multiple Arabic dialects. Google contributed to this domain by (Biadsy et al., 2012) who designed and described the ASR system for five Arabic dialects, with the potential to reach more than 125 million people in Egypt, Jordan, Lebanon, Saudi Arabia, and the United Arab Emirates (UAE). Achieved an average of 24.8% word error rate (WER) for voice search.

Most of the researches dealt with Arabic dialects as one group in the process of recognition although every dialect has its significant identity and circumstances, especially the dialects of North African (Maghrebi) for there been influenced by for instance French language. nevertheless, individual researches conceder studying them like: the feasibility study of Algerian dialect by (Menacer et al., 2017) presented ALASR, a speech recognition system dedicated to MSA with (a WER of 14.02). Moreover, tested on Algerian dialect a new acoustic model by combining two models: one for MSA and one for French. This combination leads to a WER of 65.45. And the work of (Masmoudi et al., 2018) who created a spoken dialogues corpus in the Tunisian dialect in the Tunisian Railway Transport Network domain called TARIC, developed an automatic speech recognition system of the Tunisian dialect. This ASR reaches a word error rate of 22.6% on a held-out test set.

Traditional automatic speech recognition (ASR) systems employ a modular design, with different modules for acoustic modeling, pronunciation lexicon, and language modeling, which trained separately. In contrast, end-to-end (E2E) models trained to convert acoustic features to text transcriptions directly, potentially optimizing all part for the end task; E2E ASR has attracted attention in both academia and industry(Belinkov et al., 2019). The E2E system is based on a single deep neural network that can be trained from scratch to directly transcribe speech into labels (words, phonemes, etc.). (Ahmed et al., 2018) paved the road and presented the first end-to-end recipe for an Arabic speech-to-text transcription system using the lexicon free Recurrent Neural Networks (RNNs).Reported Word Error Rate (WER) of 12.03% for non-overlapped speech.

2.7 Summary of Literature

Table 2-8: Summary of literature

AUTHOR, YEAR	OBJECTIVES	TECHNIQUES	DETAILS OF THE DATA	RESULTS
(Kirchhoff et al., 2002)	Explores methods of making MSA data usable for training models for colloquial Arabic	Language modeling, HMM	Training set (80 conversations 146298 words), the development set (20 conversations 20 32148 words), and the evaluation set (20 conversations), (15584 words). And text corpora (An-Nahar 72 million words, Al-Hayat 160 million words, Al-Ahram 61 million words, AFP 65 million words)all of which newspaper text, (Al-Jazeera 9 million words) from TV shows.	Develop novel statistical models, in particular language models, in addition to investigating three different types of language models designed to better exploit the available Egyptian colloquial Arabic data (particle models, morphological stream models, and factored language models).
(Vergyri et al., 2004)	investigates the use of morphology-based language models at different stages in a speech recognition system for conversational Arabic	Language modeling, HMM	LDC CallHome corpus of Egyptian Colloquial Arabic (ECA). The training set consists of the training, hub5 new, and eval96 subsets and contains 120 conversations (˜180K words) in total	The proposed system demonstrates word error rate reduction by 1.8%
(Afify et al., 2006)	Presents a word decomposition algorithm, that uses popular Arabic affixes, for constructing the lexicon in Iraqi Arabic speech recognition	Language modeling, HMM	The training data consists of about 200 hours of dialectal Iraqi Arabic collected in the context of a speech-to-speech translation project, The training corpus consists of about 2M words.	The net effect of the decomposition algorithm about 13% relative improvement in WER.
(Satori et al., 2007)	Investigates the Arabic language from the speech recognition problem.	Language modeling, HMM	A corpus consists of 300 tokens.	Demonstrated the possible adaptability of CMUSPHINX SYSTEM to Arabic speech recognition.
(Elshafei et al., 2008)	Reports the progress in research towards achieving large vocabulary, speaker-independent, natural Arabic automatic speech recognition system	HMM	The system was trained on 4.3 hours of the 5.4 hours of Arabic broadcast news corpus and tested on the remaining 1.1 hours	The Word Error Rate (WER) of the system came to 9.0%.

AUTHOR, YEAR	OBJECTIVES	TECHNIQUES	DETAILS OF THE DATA	RESULTS
(Soltau et al., 2007)	Presents a set of techniques for the Arabic transcription system for broadcast news.	Language modeling, HMM	LDC data, large-scale discriminative training on 1800 hours of unsupervised data, automatic vowelization using a flat-start approach, use of a large vocabulary with 617K words and 2 million pronunciations .	The system achieved a lowering of the word error rate by 42%.
(Emami and Mangu, 2007)	Studies the use of neural network language models for Arabic broadcast news and broadcast conversations speech recognition	A neural network, language modeling	Transcripts of audio data from a variety of sources released by LDC (7M words) Arabic Gigaword corpus, 5 parts (approx. 400M words) Web downloaded data from CMU (95M words) Web downloaded data from Cambridge University (200M words) Web text, namely newsgroups and weblogs, collected by LDC (28M words)	The NN model outperformed the base n-gram model and achieved a reduction of up to 0.8% absolute and 3.8% relative in WER
(Alotaibi, 2008)	Compare, analyze, and discuss the outcomes from two recognition systems Hidden Markov Model (HMM) and Neural Networks model (NN) for testing automatic Arabic digits recognition.	HMM, Neural network	The database consists of 10 repetitions of every digit produced by each speaker, totaling 1,700 tokens.	The NN model achieved 99.5% and 98.1% correct digit recognition in the case of multi-speaker mode, and the HHM system achieved 94.5% and 94.8% in the case of speaker-independent mode.
(Biadsy et al., 2009a)	Designed an ASR system by the using of linguistically motivated pronunciation rules that improve phone recognition and word recognition results for MSA	HMM	Used the broadcast news TDT4 corpus, divided into 47.61 hours of speech (89 news shows) for training and 5.18 hours (11 shows).	The system achieved absolute accuracy in phone recognition of 3.77%–7.29% and a significant improvement of 4.1% in absolute accuracy in ASR

AUTHOR, YEAR	OBJECTIVES	TECHNIQUES	DETAILS OF THE DATA	RESULTS
(Biadsy et al., 2009b)	Describes a system that automatically identifies the Arabic dialect (Gulf, Iraqi, Levantine, Egyptian, and MSA) of a speaker given a sample of his/her speech	HMM, language modeling	398 speakers from corpora (75.7 hours of speech), holding out 150 speakers for testing (about 28.7 hours of speech.)	The result produced by the system found that the most distinguishable dialect among the five variants they tested is MSA (F-Measure is always above 98.00%). EG (F-Measure of 90.2% with 30s test-utterances), LEV (F-Measure of 79.4%, with 30s test). Iraqi and Gulf (F-Measure of 71.7% and 68.3%, respectively, with 30s test utterances).
(Mohamed et al., 2009)	Investigates two types of Deep Belief Network: the backpropagation DBN (BP-DBN) and the associative memory DBN (AM-DBN) architectures for acoustic modeling.	Deep belief networks	TIMIT corpus	The deep Belief Network approach achieved a phone error rate (PER) of 23.0% on the TIMIT core test set.
(Saon et al., 2010)	Presents a set of techniques for the Arabic transcription system phase 3.5 for broadcast news.	HMM, language modeling, neural networks	85 hours of FBIS and TDT-4 audio with transcripts provided by BBN, 1500 hours of transcribed GALE data provided by the LDC	The new version led to word error rates below 9%.
(Kingsbury et al., 2011)	Presents a set of techniques for Arabic transcription system phase 4 for broadcast news.	Neural networks, language modeling	Use an acoustic training set composed of approximately 1800 hours of transcribed Arabic broadcasts provided by the Linguistic Data Consortium (LDC) for the GALE Phase 4 evaluation and 85 hours of FBIS and TDT-4 data with transcripts provided by BBN.	The new version led to a word error rate of 8.9% on the 2009 evaluation data and a year-to-year, absolute reduction of 1.6% word error rate on the unsequestered 2008 evaluation data.
(Mangu et al., 2011b)	Presents a set of techniques for Arabic transcription system phase 5 for broadcast news.	HMM, language modeling, neural networks	1800h of speech transcribed by LDC for the GALE program. Other notable sources: Arabic Gigaword corpus, 29M words of transcripts harvested from the web (Archive), we used a Phase 3 unvowelized recognizer trained on 1500 hr. of acoustic data to decode an un-seen 300 hr. set. This un-seen training set is provided in Phase 4.	The new version led to a word error rate of 7.4 % on the 2011 evaluation data and a year-to-year, absolute reduction of 0.9% word error rate on the unsequestered 2009 evaluation data.

AUTHOR, YEAR	OBJECTIVES	TECHNIQUES	DETAILS OF THE DATA	RESULTS
(Biadsy et al., 2012)	Supports voice search, dictation, and voice control for the general Arabic speaking public, including support for multiple Arabic dialects.	HMM, language modeling	Used about (EG train 604 245K 223 hours test 29 15K 12.4 hours), (JO train 848 260K 224 hours test 21 15K 10.7 hours), (SA train 745 299K 226 hours test 29 15K 10.6 hours), (AE train 587 235K 193 hours test 29 15K 10.6 hours), (LB train 795 264K 219 hours test 29 15K 13.8 hours).	The designed system reached an average of 24.8% word error rate (WER) for voice search.
(Menacer et al., 2017)	Presents new automatic speech recognition named ALASR (Arabic Loria Automatic Speech Recognition) system	Deep neural networks, HMM, language modeling	the acoustic data, 63 hours extracted from Nemlar1 and NetDC2 corpora The data are split randomly into three parts (Train, Dev, and Test): 83% are used for training (315K words), 9% for tuning (31K words), and the rest (8%, 31K words) for evaluating the performance of the system. Also, the Gigaword Arabic corpus.	A speech recognition system dedicated to MSA with (a WER of 14.02), for MSA and one for French. This combination leads to a WER of 65.45.
(Masmoudi et al., 2018)	Focuses on the design of speech tools and resources required for the development of an Automatic Speech Recognition System for the Tunisian dialect	HMM, language modeling	Used Training data of 8 h and 57 Min which consist of 18027 statements or 3027 words, Dev data of 33 min and 40 consists of 1052 s statements or 612 words, Test data of 43 min and 14 s consist of 2023 statements or 1009 words	Build the first ASR system for the Tunisian dialect with a word error rate of 22.6% on a held-out test set.
(Ahmed et al., 2018)	Presents the first end-to-end recipe for an Arabic speech-to-text transcription system using the lexicon free Recurrent Neural Networks (RNNs).	Recurrent neural networks	using 1200 hours corpus of Aljazeera multi-Genre broadcast programs	On the development, the system reported a Word Error Rate (WER) of 12.03%

2.8 Model Designing

As mentioned earlier in this chapter end-to-end speech recognition has gained wide acceptance by numerous groups of researchers, for it proved to outperform most of the conventional methods of speech recognition. The E2E system is based on a single deep neural network that can be trained from scratch to directly transcribe speech into labels (words, phonemes, etc.) (Ahmed et al., 2018).

The suggested model is going to be following the same concept as Deep Speech (Hannun et al., 2014) research which examined end-to-end (E2E) speech

recognition, Deep Speech architecture is significantly simpler than traditional speech systems yet proven to outperform previously published results, hence by using the same approach for recognizing Sudanese dialect it may result in better training outcomes and simpler design.

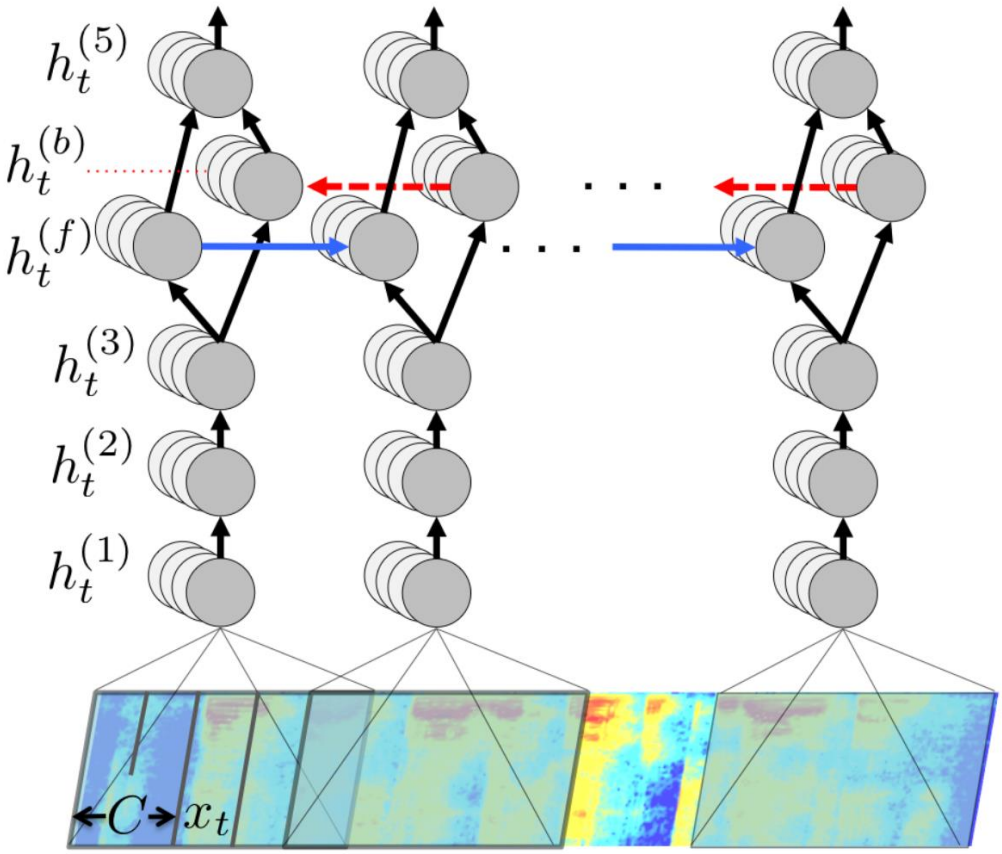

Figure 2-4: Deep Speech Architecture (Hannun et al., 2014)

Figure 2-4 illustrates Deep Speech architecture from audio spectrogram input to its text outputs, the right to left dotted arrows represent RNN.

2.8.1 Model Architecture

While the proposed model is going to use the same design principle as Deep Speech nevertheless, it will differ from the used neural networks type. Deep Speech uses a Recurrent neural network (RNN) which is good for sequence model, for instance, time series and speech recognition tasks., what makes RNN so unique for speech processing, in particular, Long short-term memory (LSTM) type, is their ability to have the capacity for keeping very long contexts in their internal state (memory) especially when applied to very long sequences. Nonetheless, the proposed

model will use Convolutional neural networks (CNN) which are concerned with computer vision for instance image recognition.

Moreover, WaveNet (Oord et al., 2016) showed very promising results and introduced Dilated Convolutions (also called `a trous, or convolution with holes) which is a convolution where the filter is applied over an area larger than its length by skipping input values with a certain step, WaveNet experimented with a generative model (generating raw audio and music) also dedicated some work for speech recognition resulted in better improvements than RNN LSTM model and also the work of (Liptchinsky et al., 2017) proved that Gated ConvNet has the ability to outperform LSTM for WaveNets we have shown that layers of dilated convolutions allow the receptive field to grow longer in a much cheaper way than using LSTM units.

Figure 2-5 shows a single stack of dilated CNN and the concept of dilation of the visualized stack has the dilation of 1, 2, 4, and 8.

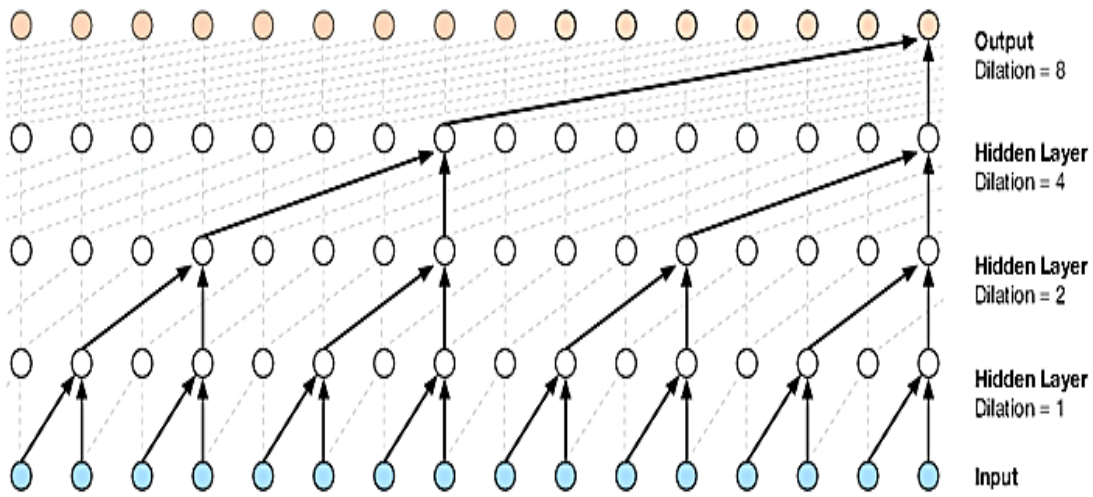

Figure 2-5: Visualization of a Stack of Dilated Causal Convolutional Layers. (Oord et al., 2016)

Figure 2-6 demonstrates in detail the entire architecture of a residual block of the previous stack which uses the same gated activation unit.

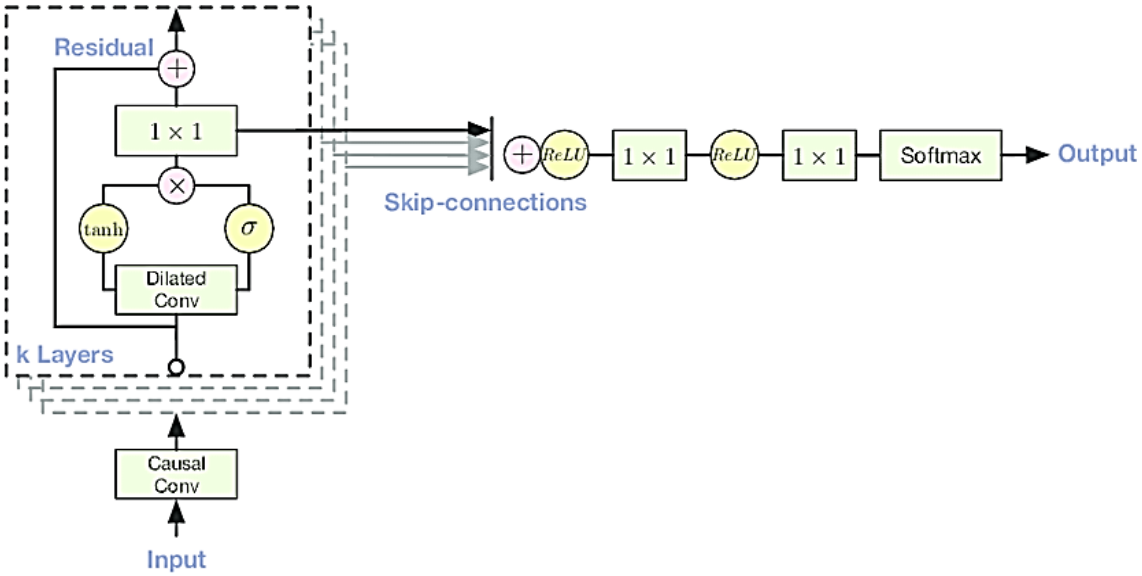

Figure 2-6: Overview of the Residual Block and the Entire Architecture. (Oord et al., 2016)

And the following equation demonstrates the mathematical representation for each block:

$$\boldsymbol{z = tanh(W_{f,g} * x) \odot \sigma (W_{gk} * x),} \qquad \textbf{Eq. 2-2}$$

Where $*$ denotes a convolution operator, $\odot$ denotes an element-wise multiplication operator, $\sigma(\cdot)$ is a sigmoid function, k is the layer index, f and g denote filter and gate, respectively, and W is a learnable convolution filter. And parameterized skip connections are used throughout the network, to speed up convergence and enable the training of much deeper models (Liptchinsky et al., 2017).

Hence this research uses the same type of WaveNets neural networks CNN, due to several reasons one of which RNN tends to allocate a considerable amount of resources doing training and optimization compared to Dilated Convolutions.

2.8.2 Algorithm

The most paramount part of any machine learning project is the learning algorithm, due to its responsibility dealing with all the heavy lifting -classification or regression- and in fact, it is the main gear that drives all training and optimization

phases, since the process of choosing the suitable learning algorithm for each problem is a very challenging task.

Connectionist Temporal Classification (CTC) (Graves et al., 2006) is the proper learning algorithm for end-to-end speech recognition, from the time of its introduction -CTC- it eliminated the need for pre-segmented data and allows the network to be trained directly for sequence labeling. The main idea behind CTC is to transform the network outputs into a conditional probability distribution over label sequences.

Given an input sequence X of length T, CTC assumes the probability of a length T character sequence C is computed as follows:

$$\boldsymbol{P(C|X) = \prod_{t=1}^{T} P(c_t|X),} \qquad \textbf{Eq. 2-3}$$

Where the network output at different times is conditionally independent given the input. Afterward, the total probabilities of anyone label sequence can then be found by summing the probabilities of its different alignments(Graves and Jaitly, 2014).

$$\boldsymbol{CTC(X,W) = \sum_{C_W} P(C|X)} \qquad \textbf{Eq. 2-4}$$

$$\boldsymbol{CTC(X,W) = \sum_{C_W} \prod_{t=1}^{T} P(c_t|X).} \qquad \textbf{Eq. 2-5}$$

CTC(X, W) is the likelihood of the correct final transcription W which requires integrating over the probabilities of all length T character sequence C.

To conclude CTC aim is to maximize the likelihood of giving acoustic features as input with their sequence characters labels, on other words, in a nutshell, CTC is a generic loss function that train sequence systems without any known alignment between the giving inputs (features) and the sequence outputs (labels).

2.9 Summary

This chapter reviewed surveys and journal papers about automatic speech recognition and mentioned most of the techniques related to it, and showed that the field of automatic speech recognition is broadly investigated for both Modern standard Arabic and dialectal Arabic. Also, the design of the proposed model was reviewed from its architecture to the learning algorithm that is going to be used to assess the process of training and optimizing. Moreover, the literature review showed that a lot of work must be done in the field of speech recognition particularly for dialectal speech recognition, therefore this research adopts a methodology based on machine learning to design a speech recognition model for the Sudanese dialect which is going to be explained in the next chapter.

CHAPTER III
Methodology

3.1 Introduction

This chapter is carried out to emphasize and describe the methodology used to fulfill the sheer objective of this research which is developing an automatic speech recognition model for the Sudanese dialect, the review of most used techniques in the field of speech recognition was investigated both their implementation and complications alongside the datasets used respectively with each research mentioned in chapter two. Furthermore, the main objectives of this research as mentioned in chapter one are:

- Collecting recordings and textual data that reflect and represent the Sudanese dialect.
- Building a simple dataset to overcome the lack of resources and make them available to future researches.
- Designing a model to recognize the speech and map it into a textual format, to fulfill the aim of this research.

The methodology was attained by contemplating each of the objectives thoroughly as showed in Figure 1-1, from collecting representable audio and textual data for the Sudanese dialect, building a dataset by performing several stages of preprocessing of the collected data, and designing a model using a deep learning approach by implementing Convolution Neural Networks (CNN) and Connectionist Temporal Classification (CTC) loss function.

3.2 Data Collection

The collected data for this research was a great hurdle; for it consumed a considerable amount of time, in fact even using official channels – national Radio and TV Commission - did not result in any access or acquisition of the requested data. So the majority of the data came from online resources (YouTube videos) extracting audio from them, by using Format Factory version 3.3.4 program to convert the videos to audio files which resulted in nearly 70 audio files, that reflect and represent the characteristics of the Sudanese dialect, mainly the middle of Sudan dialect - Khartoum in particular- and have some northern tendency.

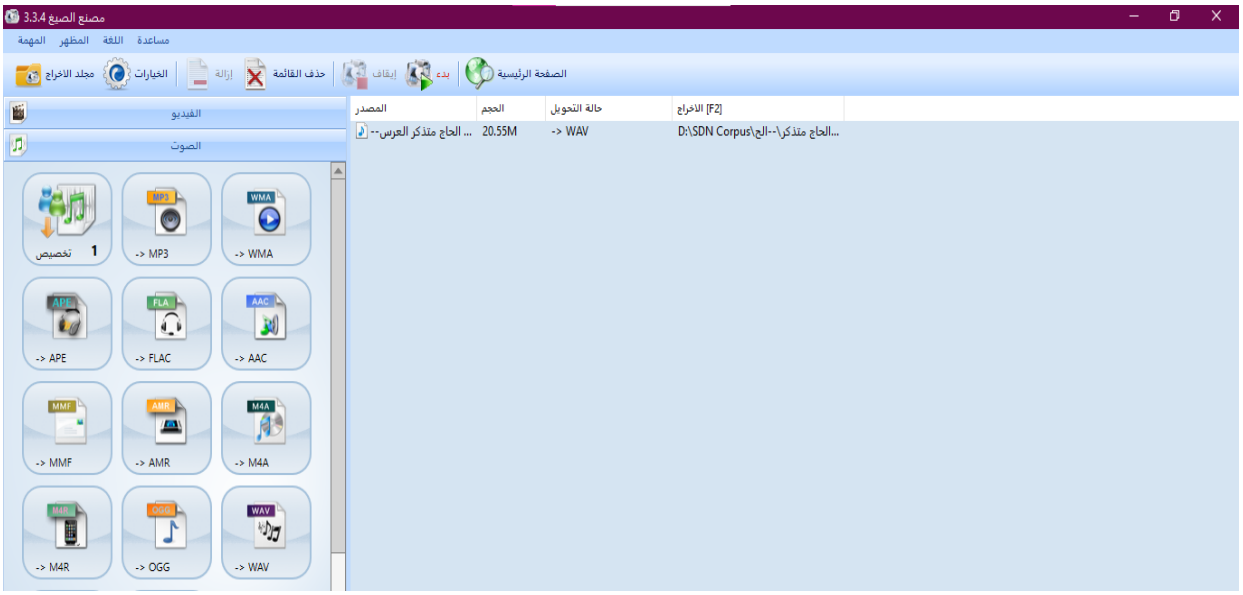

Figure 3-1: Format Factory Program

3.2.1 Data Preparing

This phase is essential for cleaning the audio files and get rid of any included music or other types of unwanted frames, to ensure the quality of the data and make it more efficient and useful in the long run of this research and future researches.

Audio files have been cleaned from laughter and noises beside music and some muddled frames, the result set has 4 hours of relatively clean speech from various speakers.

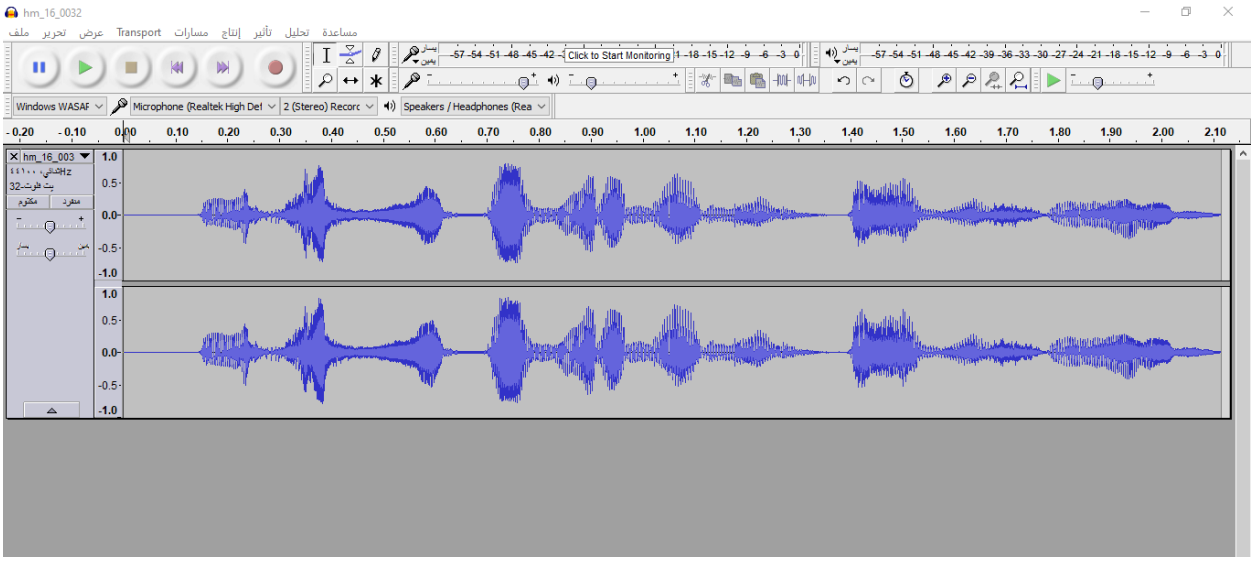

Figure 3-2: Audacity Program

All the required preparation for the audio files has been done using the Audacity program.

3.2.2 Transcriptions Writing

Having the represented text of the audio files is a great advantage but as stated in chapter two it is rare to find annotated resources especially for dialectal Arabic, hence to get the best of the collected audio, manual work has been done by listening to every single audio file repeatedly to write each conversation as it occurs and make sure that every word is written as said by the speakers.

Transcriptions have written without diacritics Arabic alphabet has been used, in a manner that reflects the Sudanese way of speaking, therefore, any correction to the noticeable mistakes was not applied to get rid of any biases and make the data representative.

3.3 Dataset Building

One of the main contributions of this research is building a Sudanese dialect dataset to compensate for the lack of resources, nevertheless, the amount of data is not considered sufficient yet it is a good step to test the viability of such a recognition task, hence to be the first building block for the most representative data and bridge the gap of the lack of annotated data, to enable further investigation and future studies, the process of corpus building features two major phases.

3.3.1 Forced Aligning

To make it possible for the model to train on the data, it is required to align the transcriptions of the audio file and save it as (comma-separated values) file to feed it to the model, each of the CSV file rows contains two values the first column represents audio filename and the second column represents the respected text to each audio file.

All of what mentioned above has been done and the resulted CSV file totaled at 3549 records represents the audio files as long as their transcription all aligned together.

Figure 3-3 shows a preview of the dataset and each of its columns.

	Filename	Text
0	hm_01_0001.wav	تعبان تعبان خالص تعبان
1	hm_01_0002.wav	تعبان تعبان خالص تعبان
2	hm_01_0003.wav	نحولك العناية المكثفة
3	hm_01_0004.wav	الإسم الكريم
4	hm_01_0005.wav	محسوبك النعيسان الفي الشعر حسان
...	...	...
745	wb_03_0242.wav	أووووو أهلا أهلا أهلا يا ود الحسين
746	wb_03_0243.wav	إزيكن
747	wb_03_0244.wav	مبسوط كدي مبسوط كدي مالك يا ود الحسين
748	wb_03_0245.wav	أنا ما جبتا لكم الليلة خبر
749	wb_03_0246.wav	أنا ما مشيت المدرسة بس قالو لي الإمتحانات السن...

3549 rows × 2 columns

Figure 3-3: Aligned Dataset Preview

3.3.2 The Text Encoding (Transliteration)

Transliterations allow for simple non-lossy mapping from Arabic to Roman script and back (Habash, 2010), which allows the model to compare the audio to each character (training process). Several types of research examined this stage and there are two methods for encoding Arabic text first is using the Buckwalter dictionary to transliterate Arabic text to English characters, and CODA (Habash et al., 2012) which stands for (Conventional Orthography for Dialectal Arabic).

The proposed methodology tends to use the Buckwalter dictionary, for it has wide implementation and ease of use but for dialectal Arabic processing CODA is preferred giving the fact it is designed primarily for the purpose of developing computational models of Arabic dialects(Habash et al., 2012), hence it may come in handy for the future study or a follow-up for this research.

Table 3-1 shows some rows of Buckwalter dictionary to transliterate Arabic text to English characters, the entire dictionary is obtainable in Appendix A.

Table 3-1: Buckwalter Transliteration Dictionary

UNICODE			BUCKWALTER	
Decimal	**Hex**	**Glyph**	**ASCII**	**Orthography**
1569	U+0621	ء	**'**	Hamza
1571	U+0623	أ	**>**	Alif + Hamza Above
1572	U+0624	ؤ	**&**	Waw + Hamza Above
1573	U+0625	إ	**<**	Alif + Hamza Below
1574	U+0626	ئ	**}**	Ya + Hamza Above
1575	U+0627	ا	**A**	Alif

In figure 3-4, the dataset is transliterated using the Buckwalter transliteration dictionary been applied to the Sudanese dialect dataset and resulted in Arabic text transliteration using English characters.

	Filename	Text
0	hm_01_0001.wav	تعبان تعبان خالص تعبان
1	hm_01_0002.wav	تعبان تعبان خالص تعبان
2	hm_01_0003.wav	نحولك العناية المكثفة
3	hm_01_0004.wav	الإسم الكريم
4	hm_01_0005.wav	محسوبك النعيسان الفي الشعر حسان
...	...	...
745	wb_03_0242.wav	أووووو أهلا أهلا أهلا يا ود الحسين
746	wb_03_0243.wav	إزيكن
747	wb_03_0244.wav	مبسوط كدي مبسوط كدي مالك يا ود الحسين
748	wb_03_0245.wav	أنا ما جبتا لكم الليلة خبر
749	wb_03_0246.wav	أنا ما مشيت المدرسة بس قالو لي الإمتحانات السن...

	Filename	Text
0	hm_01_0001.wav	tEbaa'n tEbaa'n xAA'lS tEbaa'n
1	hm_01_0002.wav	tEbaa'n tEbaa'n xAA'lS tEbaa'n
2	hm_01_0003.wav	nHuu0'lk lEnaaii0 lmk^f
3	hm_01_0004.wav	aal<i0sm lkrii0'm
4	hm_01_0005.wav	mHsuu0'bk lnEii0saa'n lfii0' l$Er Hsaa'n
...	...	...
745	wb_03_0242.wav	<uu0uu0uu0wuu0' <a'hlaa <a'hlaa <a'hlaa yaa' u...
746	wb_03_0243.wav	<i0zii0'kn
747	wb_03_0244.wav	mbsUU0'T kdii0' mbsUU0'T kdii0' maa'lk yaa' uu...
748	wb_03_0245.wav	<a'naa maa' jbtaa' lkm lii0'l xbr
749	wb_03_0246.wav	<a'naa maa' m$ii0't lmdrs bs qAA'luu0 lii0' l<...

Figure 3-4: Dataset Preview before and after Transliteration

Figure 3-5 summaries the major steps applied to collect the required data needed for this research, as stated the majority of the data came from YouTube videos, the first process was to convert them from videos to audio files by converting them by Format Factory program which extracted the audio signals from the videos,

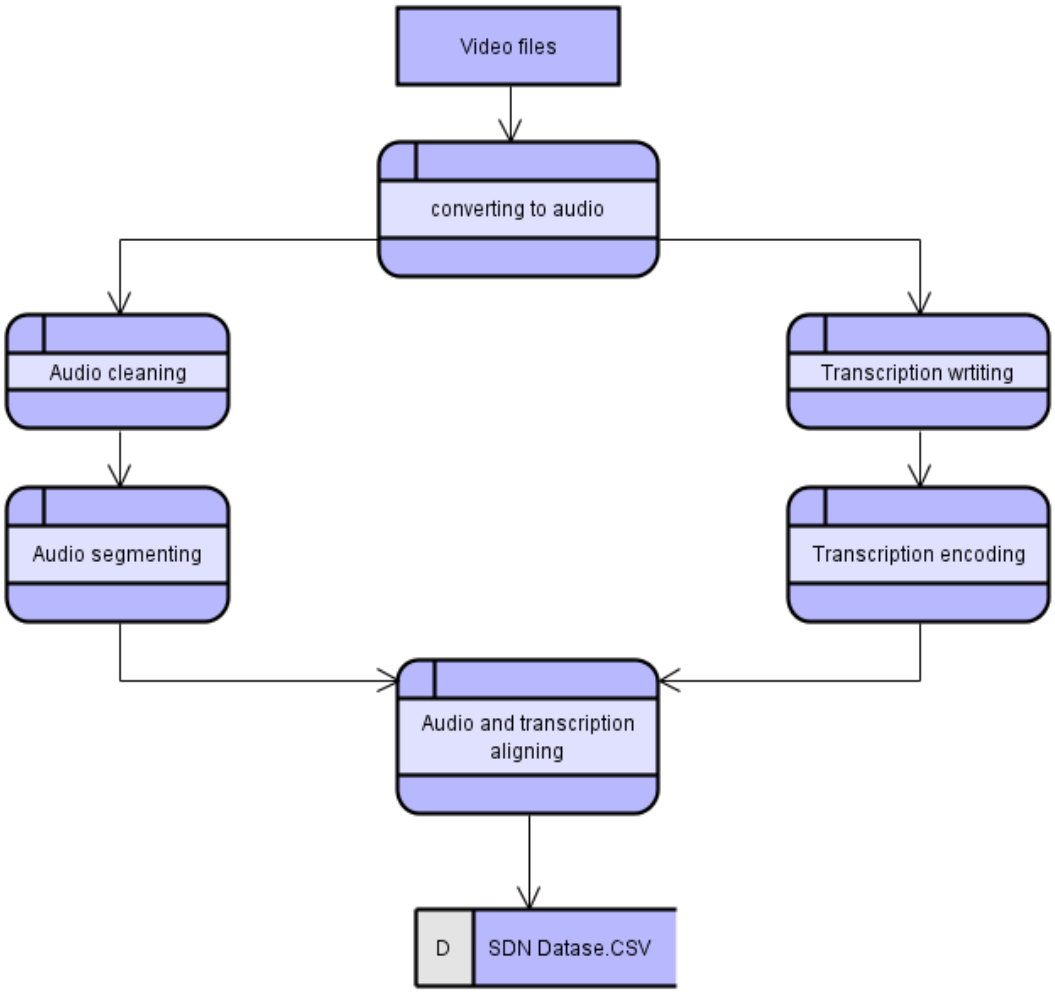

Figure 3-5: Data Collection Diagram

followed by two processes applied to the audio, first the process of cleaning the audio files from unwanted frames and segments (laughter and music), then the process of segmenting the audio files to small segments to ease the building of the dataset both audio preparing processes been done using Audacity program.

From the process of converting the videos to audio files came two processes related to the textual data, first Transcription writing as mention earlier been done by listing to each audio file carefully to fill the speaker's dialogue exactly as been said. Then the process of encoding the transcription using the Buckwalter transliteration dictionary.

The final process is building the dataset by aligning the segmented audio files with the encoded transcription and save them all as comma-separated values file format SDN dialect.CSV, which is going to be fed to the proposed model to perform training and optimization to recognize the Sudanese dialect, by converting the spoken audio to its textual form.

3.4 Operation Environment

The implementation of the proposed methodology executed by adopting several tools and concepts and choosing the proper platform for developing the speech recognizer, all the tools are mentioned below in a brief introduction.

3.4.1 Juypter Notebook

Jupyter is a free, open-source, interactive web tool known as a computational notebook, Jupyter stands for Julia (Ju), Python (Py), and R combined, which researchers can use to combine software code, computational output, explanatory text, and multimedia resources in a single document.

A Jupyter notebook can work either locally or on the cloud. Each document is composed of multiple cells, where each cell contains script language or markdown code, and the output is embedded in the document. Typical outputs include text,

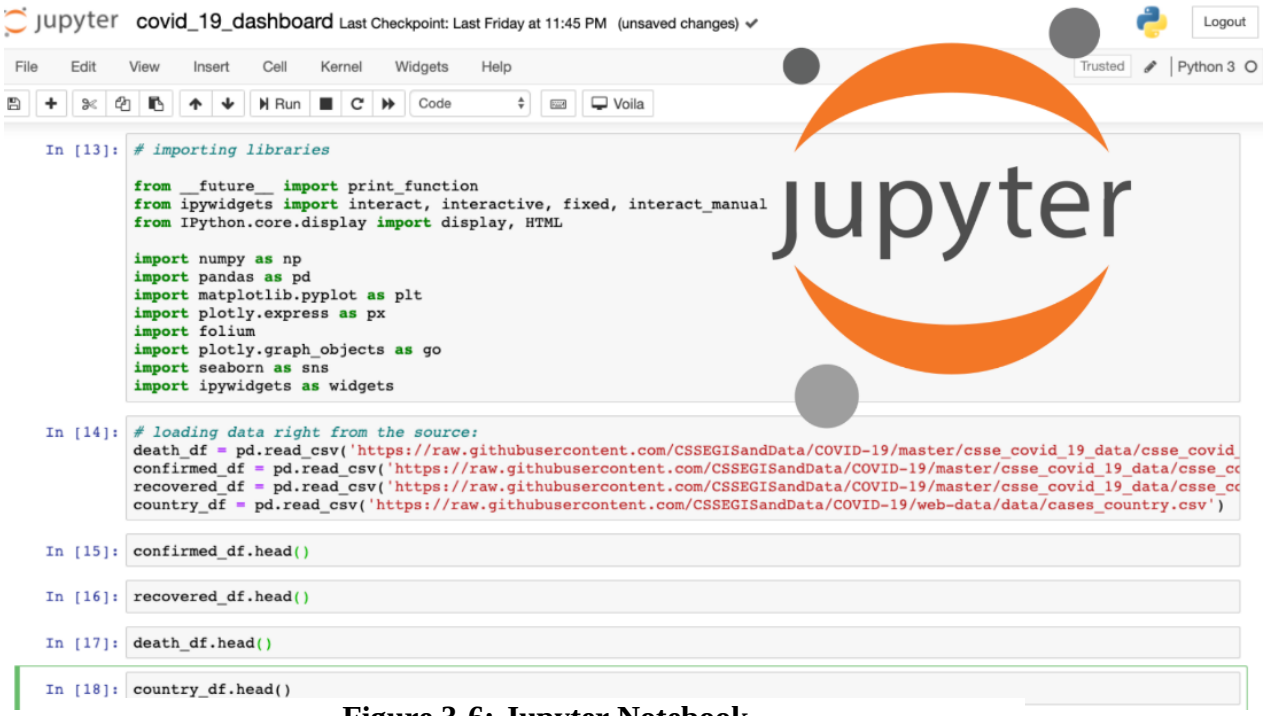

Figure 3-6: Jupyter Notebook

tables, charts, and graphics. Using this technology makes it easier to share and replicate scientific works since the experiments and results are presented in a self-contained manner (Perkel, 2018).

3.4.2 TensorFlow

TensorFlow is a machine learning system that operates at a large scale and in heterogeneous environments, it maps the nodes of a dataflow graph across many machines in a cluster, and within a machine across multiple computational devices,

including multicore CPUs, general-purpose GPUs, and custom-designed ASICs known as Tensor Processing Units (TPUs).

TensorFlow enables developers to experiment with novel optimizations and training algorithms, also supports a variety of applications, with a focus on training and inference on deep neural networks. Several Google services use TensorFlow in production, it released as an open-source project, and become widely used for machine learning research (Abadi et al., 2016).

3.4.3 Google Colaboratory (Colab)

Google Colaboratory or Colab is a project that has the objective of disseminating machine learning education and research. Colab provides either Python 2 or 3 runtimes pre-configured with the essential machine learning libraries, such as

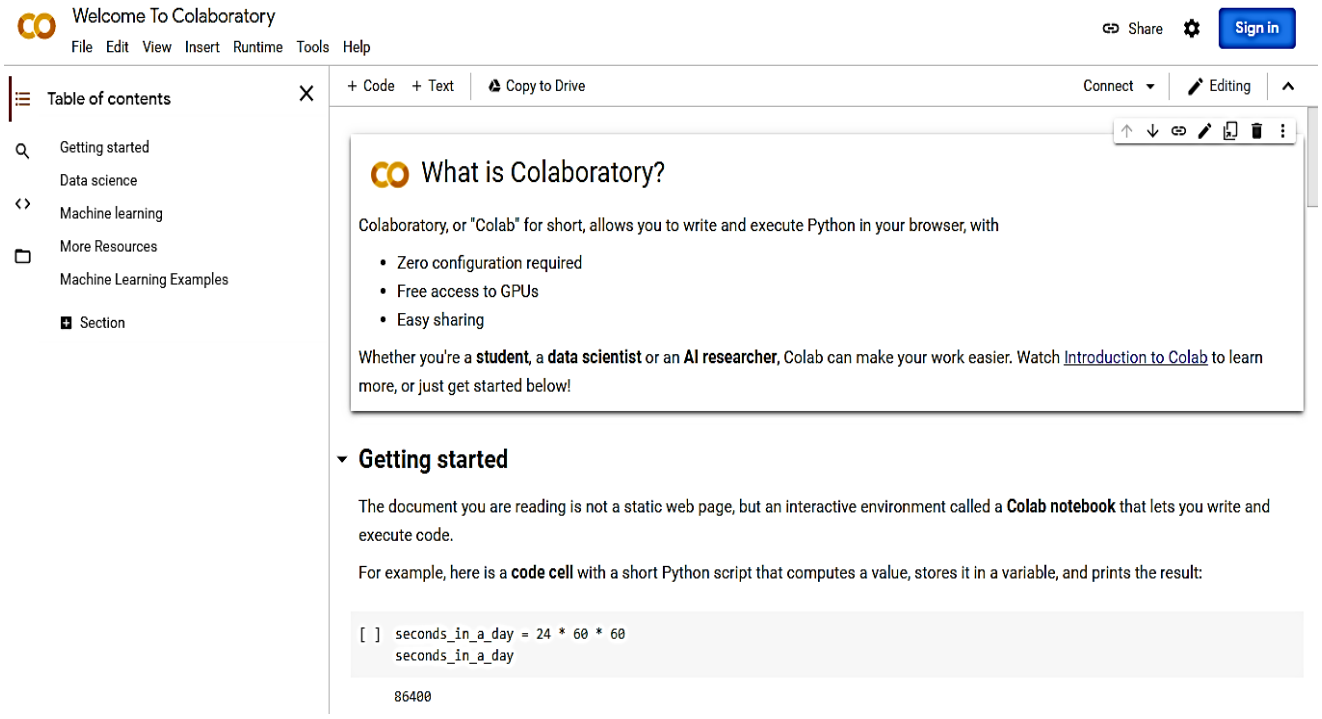

Figure 3-7: Google Colab Environment

TensorFlow, Matplotlib, and Keras. Colab Operates under Ubuntu 17.10 64 bits and it is composed of an Intel Xeon processor (not specified) with two cores @ 2.3 GHz and 13 GB RAM. It is equipped with an NVIDIA Tesla K80 (GK210 chipset), 12 GB RAM, 2496 CUDA cores @ 560 MHz (Carneiro et al., 2018).

The proposed model is going to be computed using Colab which is free to use the Jupyter notebook environment. Also using TensorFlow as a library to design the components of the proposed model structure and track each step in the training and

validation processes using TensorFlow's visualization toolkit (TensorBoard) to get a visual representation of the model performance at each step.

3.5 Sudanese Dialect ASR Model Structure

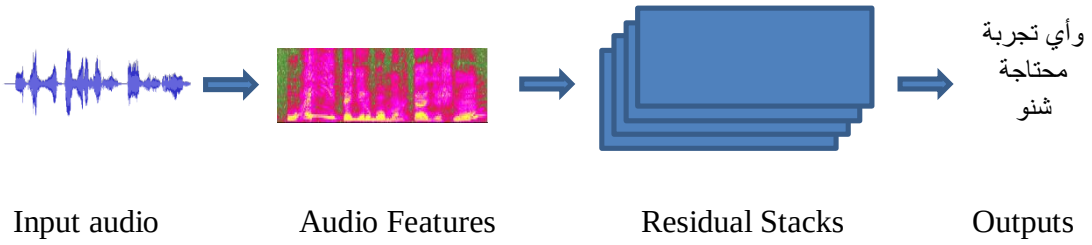

Figure 3-8: Sudanese Dialect ASR Model Overview

Figure 3-8 gives an overview of the proposed ASR model which consists of four main steps, the input step accepts raw audio file then the audio features extraction which performs Log Mel Spectrogram to get the audio features to feed it to the next step, residual stack step conceptualize all the components of the deep learning model, and output step finally produces the transcripts for the inputted audio file.

3.6 Sudanese Dialect ASR Model

The proposed ASR model has explored briefly earlier, this section is demonstrating the main parts needed to give the general idea of this research, here in this section a detailed design is going to be presented to elaborate the previous illustration. The adopted design follows the same concepts from (Ciemniewski, 2019) Blogpost.

3.6.1 ASR Model Components

The main building part of the Sudanese dialect model is the Residual block which primarily all the rest of the pieces are built upon its design (kernel), from Residual Stacks which are simply a group of Residual Blocks stacked together and connected via skipping connections (dilation) to speed up the convergence and allow much deep design.

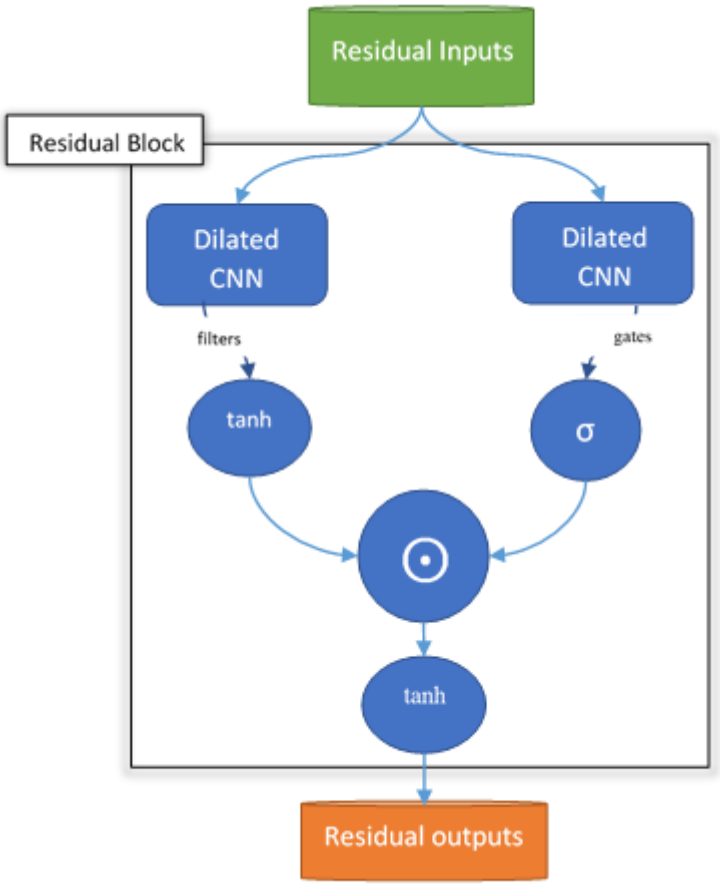

Figure 3-10: Structure of the Residual Block

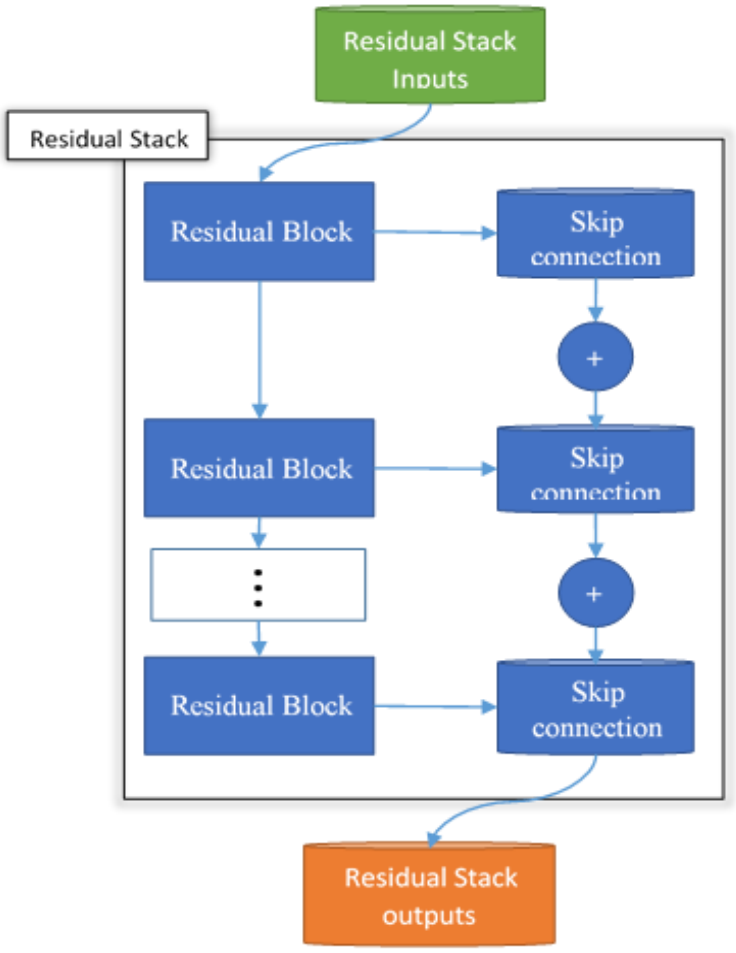

Figure 3-9: Structure of the Residual Stack

From what been illustrated Residual Blocks as in figure 3-9 combined build Residual Stack as in figure 3-10, and Residual stacks together build the core components architecture of the neural network which builds the Sudanese dialect model along with features extractors and some inputs preprocessors (Log Mel Spectrogram) and at last outputs Logits (layer).

Figure 3-11 depicts the simplified general design of the Sudanese dialect model.

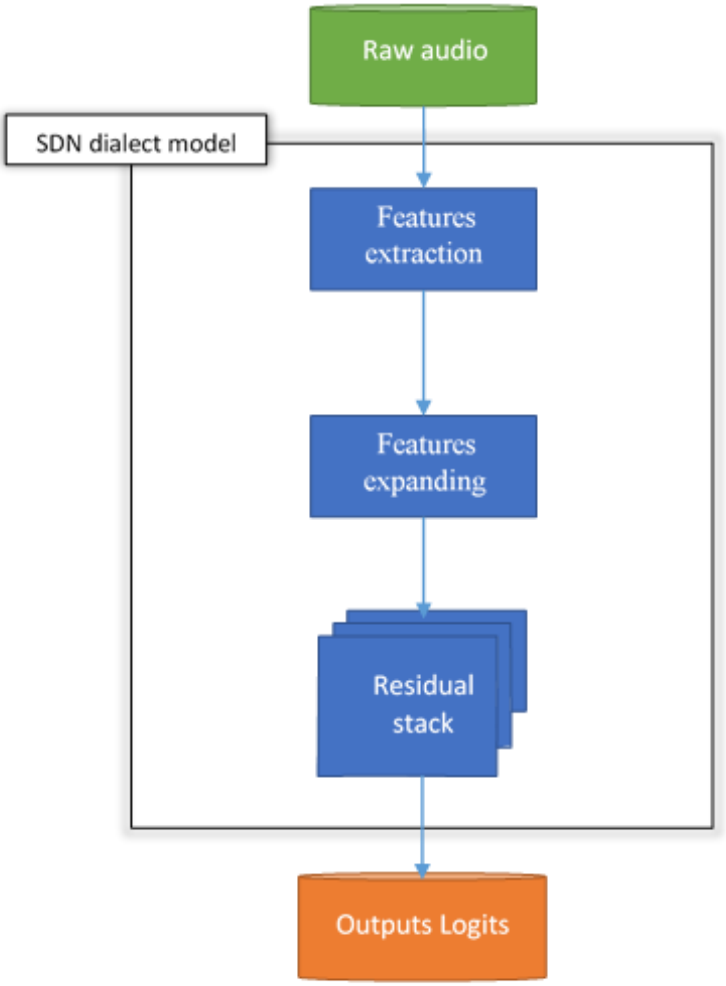

Figure 3-11: Sudanese Dialect ASR Model

3.6.2 ASR Model Setup

While the previous section gave a broad look at the general components of the model, moreover, this section gives the model setup from the number of the kernels – Residual Block – in each stack and the number of the Residual Stack in the suggested model, besides some preprocessing steps applying to each the raw audio files entered to the model.

Each of the Residual Stacks has an equal number of kernels which is 4 Residual Blocks together build a single Residual stack, and each stack in the suggested model obeys a certain dilation sequence which in 1, 3, 9, and 27, the number of the Residual Stacks in the suggested model is going to be investigated in the empirical part of this research in the next chapter.

Each audio file inputted into the model via the input layer been dealt with it in a certain manner from performing several preprocessing stages, from selecting the number of files (batch) fed to the model per epoch which is 18 files in each round of training, and performing unified sampling rate which is 16k to every single audio file.

Last but not least randomly combine selected noise files to make the model more robust in handling more noisy inputs and give reasonable performance for future inspecting of real-life experiments. At last performing random stretch to the audio files to compensate for the difference in length for each input file.

All of what was mentioned above implemented using Tensorflow in the Colab platform together as one unit which is both compatible as Google product. To resolve the problem that the Colab platform loses data at the end of every session, the resulted data stored directly using Google Drive cloud storage platform which has 15GB free to use.

However, the given setup may seem haphazardly chosen but in the next chapter, some configurations are going to be applied with further investigation and discussion.

3.7 Summary

This chapter described the proposed methodology for this research, from the first step of collecting the proper data through its several stages of preprocessing phases. Form audio files preparing and transcribing them, to the designing of the Sudanese dialect dataset by manually aligning the transcriptions and cropped audio files. Then the transcriptions were been transliterated using the Buckwalter transliteration. Also demonstrated some of the essential tools their underlining technology and features needed to apply the methodology of this research, the researcher gave a general overview of the structure for the proposed Sudanese dialect ASR model, moreover illustrated in details the design of the suggested ASR model from its major components and suggested configurations.

CHAPTER IV

Experiments and Results

4.1 Introduction

This chapter is carried out to test the viability of this research by conducting various experiments using the proposed methodology. From presenting the empirical part of this research to reach the prober results by inspecting each step in the training and optimizing processes, moreover, presents the resulted outcomes of this research.

4.2 Empirical Implementation

The process of designing any machine learning model obeys two major phases training and validating, both of these phases are essential for measuring the performance of the designed model, below are each of these phases of developing the speech recognition model. Below are each of these phases of developing the speech recognition model.

4.2.1 Training

As mentioned in the last chapter the adopted methodology of this research is an end-to-end approach for its empirical simplicity and state of the art performance, giving the previous model setup from the last chapter the process of training was performed by using the designed Sudanese dialect dataset for it is already built considerably to be fed to an end-to-end model. Nevertheless, regarding what has been said before the Sudanese dialect dataset does not consist of much data to produce informative results, therefore, an addition MSA corpus (Halabi and Wald, 2016) been used combined to the dialectal Sudanese dataset like similar researches add MSA data to their dialectal corpus (Menacer et al., 2017, Masmoudi et al., 2018) to add some benefits and more linguistic features because the Sudanese dialect inherits some MSA characteristics (Gasim, 1965).

By combing the Sudanese dialect dataset which contains nearly 3549 records, with the MSA corpus which contains 1813 records totaled at 7 hours and 50 minutes, representing the audio files as long as their transcription. To train the model, 5083 records were used from the heterogeneous dataset, and the rest of the data reserved for the process of evaluating the model performance.

The training process executed using the Colab platform as VM to compensate for the lack of the prober hardware needed for the training for the obtained hardware just composed of an Intel Core i7 processor with two cores @ 1.9 GHz available to reach 2.5 GHz and 8 GB RAM. It is equipped with AMD Radeon graphic processor (HD 8500m/8700m) 2 GB RAM, @ 400 MHz, and a lack of GPU acceleration support.

Initially, three setups been used to further examine the proposed design each of them shares a similar number of Residual blocks which is 4 blocks, and a similar dilation sequence which in 1, 3, 9, and 27, but differs in the number of the Residual stack, the first setup contains 6 Residual stacks and trained for 103.1K iterations (steps) using the combined dataset which resulted in an average of 85.24% label error rate (LER) which is not good result as the smaller LER the better, the second setup contains 7 Residual stacks and trained for 117.9K iterations (steps) using the combined dataset which resulted in an average of 70.55% label error rate (LER), and the last setup contains 8 Residual stacks and trained for 10.46K iterations (steps) using the combined dataset which resulted in an average of 78.01% label error rate (LER).

Figure 4-1 shows the training performance graph for the first setup which has 6 Residual stacks, the y-axis represents edit distance value and the x-axis represents the number of iterations (steps).

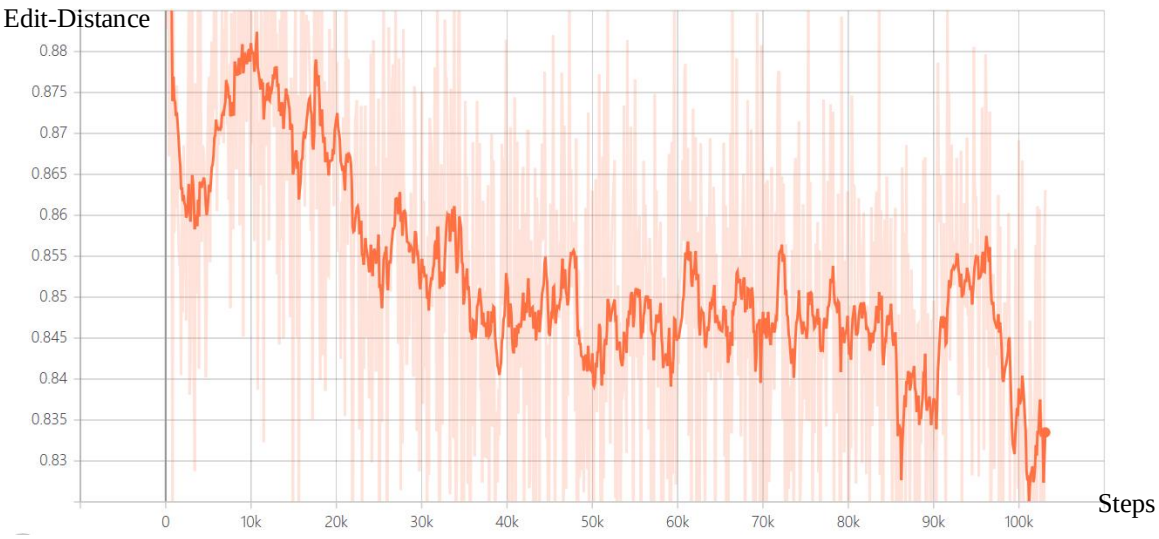

Figure 4-1: First Setup 6 Residual Stacks Training Graph

Figure 4-2 shows the training performance graph for the second setup which has 7 Residual stacks.

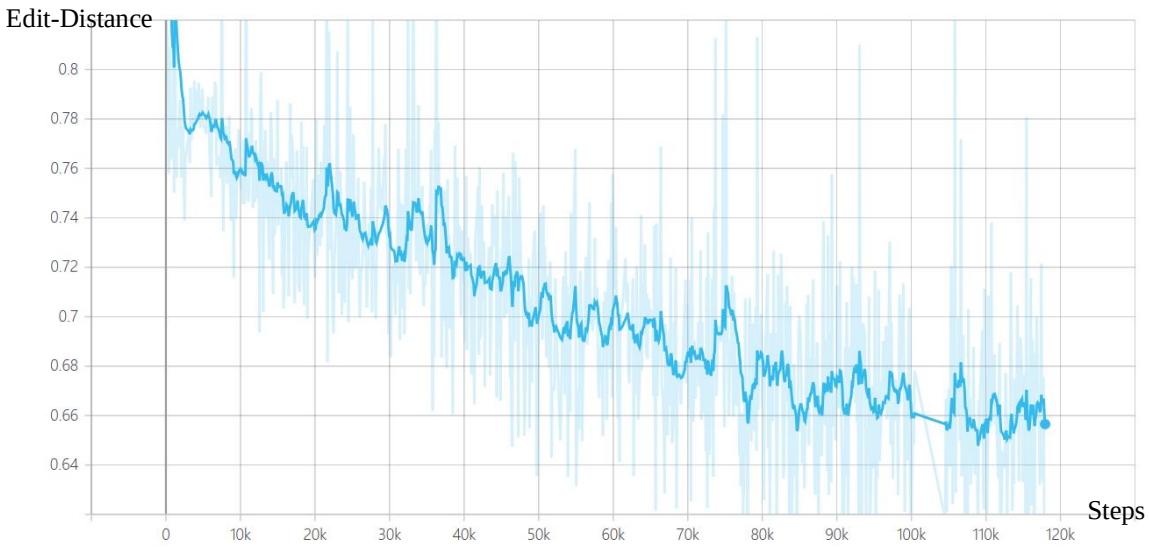

Figure 4-2: Second Setup 7 Residual Stacks Training Graph

Figure 4-3 shows the training performance graph for the last setup which has 8 Residual stacks.

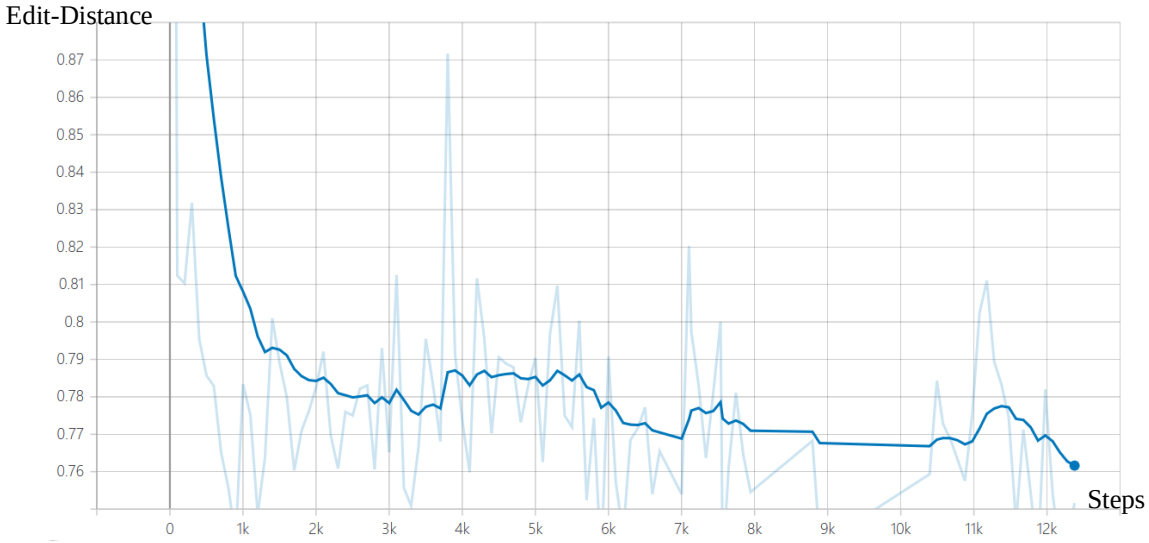

Figure 4-3: Last setup 8 Residual Stacks Training Graph

Relatively the best setup is the last one with the 8 Residual stacks because it reached better LER than other setups in smaller iterations (steps), but it has some complications for it uses more computational resources also more storage and memory to complete each Iteration compared to other setups.

The second setup with 7 Residual stacks used as the accepted design for this research, because it is promising and moderate in its performance as shown in figures (4-1, 4-2, and 4-3), compared to the first setup (6 Residual stacks). And has the ability to get a better result, for instance, at 10K iteration the 6 Residual stacks setup produced an 87.5% label error rate (LER), in contrast, the 7 Residual stacks setup produced a 76% label error rate (LER), also its better than the last setup (8 Residual stacks) in using the computational resources, at 10K iteration the 7 Residual stacks setup produced 76% label error rate (LER), but the 8 Residual stacks setup barely produced 77% label error rate (LER) compared to the resources that have utilized.

Regarding the three initial setups, unfortunately, the unintentional problem has occurred during the training process. During each execution, the reserved data for the validation phase leaked to the training phase which led to false results regarding LER. But hopefully, by using the second setup, which contains 7 Residual stacks, an alternative experiment started to eliminate the false results produced earlier, to assure a clear view about the feasibility of this research and produce reliable results.

The alternative experiment lasted nearly 4 weeks and reached 302.1K iterations (steps), by using the combined dataset which resulted in an average of 66.14% label error rate (LER), as figure 4-4 shows the alternative experiment training graph.

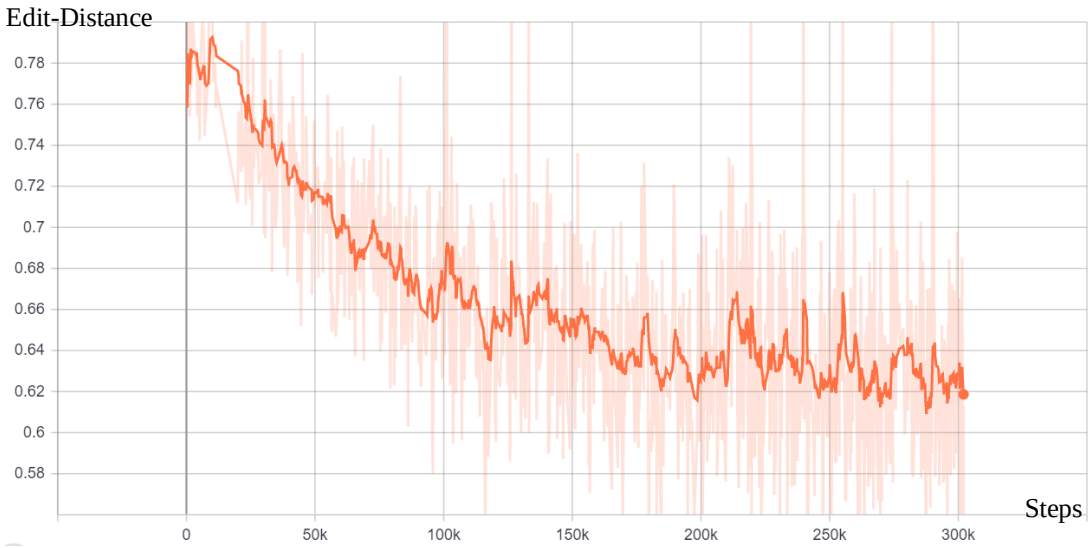

Figure 4-4: Alternative Experiment (7 Residual Stacks) Training Graph

Table 4-1: Model Setups Comparison

Model setup	Dataset Size	Iteration	Training (avg.)
The first model with 7 Residual stacks	7h and 50m	103.1K steps	85.24% LER
The second model with 7 Residual stacks	7h and 50m	117.9K steps	**70.55%** LER
The third model with 8 Residual stacks	7h and 50m	10.46K	78.01% LER

Table 4-1 show all the model setups and compare between them from their performance regarding the used stack configuration and the resulted outcomes in label error rate

4.2.2 Validation

While training results seem important as good initial proof of concept and model assessment measures but the most paramount results are those related to the validation phase giving the fact that the data used for this phase are not available to the model during the training phase, and considered as a challenge to the model resembling a real-world test.

Giving the problem with the initial experiment regarding the leakage of the validation data to be processed along with the training data, while the stated problem introduced great insights about the proper setup configurations for the model during training, despite that it caused the produced results to be falsified, the results of each setup of the initial experiment been completely ignored especially those related to the validation phase.

Therefore, the only available validation results to be submitted as acceptable outcomes for this research are those related to the alternative experiment which as discussed before contains 7 Residual stacks, the validation phase resulted in an average of 73.67% label error rate (LER). While the sheer results may seem underwhelming compared to the training result, but we should keep in mind that the data fed into the model during validation never seen before in the training phase, therefore, the performance varied tremendously during the validation phase.

Figure 4-5 shows the alternative experiment validation graph the y-axis represents the edit distance value and the x-axis represents the number of iterations (steps).

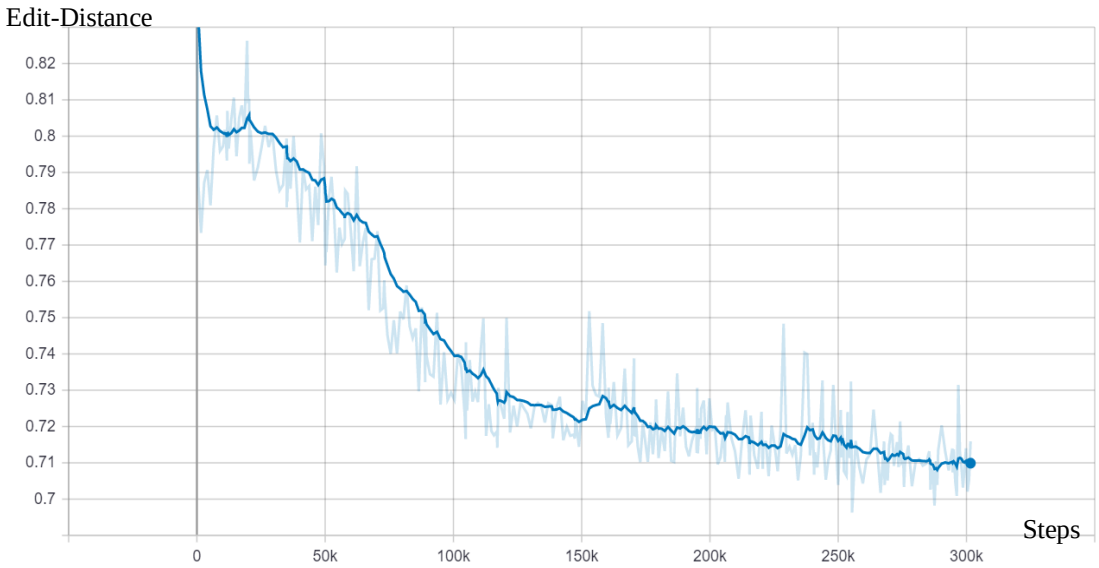

Figure 4-5: Alternative Experiment (7 Residual Stacks) Validation Graph

4.3 Results and Discussion

Giving the size of the used data to train the proposed end-to-end model (7hours and 50 minutes) which is not sufficient compared to an end-to-end approach, and the limited access to the Colab platform (12 hours for each session), in addition

to its hardware limitation (one processor, 13 GB RAM, and single GPU), also Google Drive storage available as free (15 GB).

The total training time was nearly one-month executing the Colab session at least one time per day but normally executed more than once, only for the acceptable results (the alternative experiment) not counting the initial experiment with its different configuration.

The produced results as shown earlier in the last section are illustrated together in the flowing table 4-2.

Table 4-2: Results of the Suggested ASR Model

Model	Language model	Dataset size	Iteration	Training (min.)	Training (avg.)	Validation (min.)	Validation (avg.)	Model Accuracy (avg.)
CNN/CTC	None	7h and 50m	302.1K steps	50.38% LER	66.14% LER	69.63% LER	73.67% LER	26.33% LER

The results may seem not satisfactory, but as compared to (Ahmed et al., 2018) in some aspects, especially the 8-hours experiment from the same research yielded much better results than those accomplished by this research. Not forgotten that (Ahmed et al., 2018) research has much performance advantage to be utilized, nonetheless, this research proved that it is feasible to design a model giving the moderate dataset and slightly reasonable available hardware.

In summary, the results show that the approach followed in this research is quite effective, but giving the shortage of more data, and proper available hardware, the model performance would have been better and LER been smaller.

4.4 Summary

This chapter demonstrated the empirical part of this research both the training and validation phases have been deliberately investigated, and finally presented the results accomplished by implementing the suggested setup and configuration to fulfill the paramount goal of this research which is designing an ASR model for Sudanese dialect.

CHAPTER V

Conclusion and Recommendations

5.1 Conclusion

The Sudanese Dialect lacks broad representation in the field of speech recognition for Arabic dialects. This research addressed the possibility of designing an ASR model for the Sudanese Dialect. For the Sudanese Dialect, no data set was available. The researcher started with the collection of related data and then went through the process of designing the represented dataset. A suitable learning algorithm was chosen, and the structure of the model was built by training. And evaluating the produced results from the machine learning model.

From the beginning of this research, there have been a lot of hurdles to overcome, especially the two major steps (building a dataset and training the model). Manual transcription and aligning of the collected data were achieved, and training the model took time because of the shortage of computation power.

The researcher designed the first Sudanese dialect dataset model, which will be very helpful for relevant future researches, and developing Sudanese dialect recognizer to convert spoken speech to corresponding text form. The machine learning model is based on dilated CNN architecture and using CTC as a learning algorithm, and training it on the heterogeneous dataset. The data set consists of MSA corpus (1813 records) and the Sudanese dialect dataset (3549 records) at nearly over four weeks which reach 302.1K steps (iterations). The results were evaluated to measure the performance of the proposed model. The free version of Google Colab platform was used for training and validation.

The results showed that the model has potentials in converting spoken Sudanese Dialect to text format, the model reached an average LER of 66.14% and minimum LER of 50.38% at the training stage, and reached an average LER of 73.67% and minimum LER of 69.63% at the validation stage. The used dataset consisted of merely 7 hours and 50 minutes.

5.2 Recommendations

This research is considered to break new ground for the Sudanese Dialect speech recognition. Yet future work is required to build upon this research results by further data collecting for the Sudanese Dialect and improving the model

performance, to make a contribution for the field of Arabic dialectal speech recognition, and to enrich available resources particularly for the Sudanese Dialect, two main points are recommended to continue the pursuance of this research mission.

5.2.1 Data Collection

The researcher recommends collecting data from many sources and designing a more diverse dataset to reflect the Sudanese Dialect's unique characteristics to avoid any misrepresentation and biases.

5.2.2 Model Improvement

To further improve the performance of the designed model, the following is recommended:

- Implementing a language model inside the proposed design.
- Use some APIs to tweak hyper-parameters such as weight & biases API.
- Transfer learning instead of building new models from scratch.

References

ABADI, M., AGARWAL, A., BARHAM, P., BREVDO, E., CHEN, Z., CITRO, C., CORRADO, G. S., DAVIS, A., DEAN, J. & DEVIN, M. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

ABUZEINA, D., AL-KHATIB, W., ELSHAFEI, M. & AL-MUHTASEB, H. 2011. Cross-word Arabic pronunciation variation modeling for speech recognition. *International Journal of Speech Technology,* 14**,** 227-236.

AFIFY, M., SARIKAYA, R., KUO, H.-K. J., BESACIER, L. & GAO, Y. On the use of morphological analysis for dialectal Arabic speech recognition. INTERSPEECH, 2006.

AHMED, A., HIFNY, Y., SHAALAN, K. & TORAL, S. 2018. End-to-End Lexicon Free Arabic Speech Recognition Using Recurrent Neural Networks. *Computational Linguistics, Speech And Image Processing For Arabic Language,* 4**,** 231.

ALOTAIBI, Y. A. 2008. Comparative study of ANN and HMM to Arabic digits recognition systems. *Journal of King Abdulaziz University: Engineering Sciences,* 19**,** 43-59.

BELINKOV, Y., ALI, A. & GLASS, J. 2019. Analyzing Phonetic and Graphemic Representations in End-to-End Automatic Speech Recognition. *arXiv preprint arXiv:1907.04224*.

BIADSY, F., HABASH, N. & HIRSCHBERG, J. Improving the Arabic pronunciation dictionary for phone and word recognition with linguistically-based pronunciation rules. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2009a. Association for Computational Linguistics, 397-405.

BIADSY, F., HIRSCHBERG, J. & HABASH, N. Spoken Arabic dialect identification using phonotactic modeling. Proceedings of the eacl 2009 workshop on computational approaches to semitic languages, 2009b. Association for Computational Linguistics, 53-61.

BIADSY, F., MORENO, P. J. & JANSCHE, M. Google's cross-dialect Arabic voice search. Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012. IEEE, 4441-4444.

CARNEIRO, T., DA NóBREGA, R. V. M., NEPOMUCENO, T., BIAN, G.-B., DE ALBUQUERQUE, V. H. C. & REBOUCAS FILHO, P. P. 2018. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. *IEEE Access,* 6**,** 61677-61685.

CIEMNIEWSKI, K. 2019. *Speech Recognition from scratch using Dilated Convolutions and CTC in TensorFlow* [Online]. Available: https://www.endpoint.com/blog/2019/01/08/speech-recognition-with-tensorflow.

ELSHAFEI, M., AL-MUHTASEB, H. & AL-GHAMDI, M. Speaker-independent natural Arabic speech recognition system. The International Conference on Intelligent Systems, 2008.

EMAMI, A. & MANGU, L. Empirical study of neural network language models for Arabic speech recognition. Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on, 2007. IEEE, 147-152.

FRANCISCO, I., CERóN, C., GRACIELA, A., BADILLO, G. & ASPLUND, L. 2020. A Keyword Based Interactive Speech Recognition System for Embedded Applications Master's Thesis.

GASIM, A. A.-S. 1965. Some aspects of Sudanese colloquial Arabic. *Sudan Notes and Records,* 46**,** 40-49.

GOLDENTHAL, W. D. 1994. *Statistical trajectory models for phonetic recognition.* Massachusetts Institute of Technology.

GRAVES, A., FERNáNDEZ, S., GOMEZ, F. & SCHMIDHUBER, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. Proceedings of the 23rd international conference on Machine learning, 2006. 369-376.

GRAVES, A. & JAITLY, N. Towards end-to-end speech recognition with recurrent neural networks. International conference on machine learning, 2014. 1764-1772.

HABASH, N., DIAB, M. T. & RAMBOW, O. Conventional Orthography for Dialectal Arabic. LREC, 2012. 711-718.

HABASH, N. Y. 2010. *Arabic Natural Language Processing*, Morgan & Claypool Publishers.

HALABI, N. & WALD, M. 2016. Phonetic inventory for an Arabic speech corpus.

HANNUN, A., CASE, C., CASPER, J., CATANZARO, B., DIAMOS, G., ELSEN, E., PRENGER, R., SATHEESH, S., SENGUPTA, S. & COATES, A. 2014. Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

KINGSBURY, B., SOLTAU, H., SAON, G., CHU, S., KUO, H.-K., MANGU, L., RAVURI, S., MORGAN, N. & JANIN, A. The IBM 2009 GALE Arabic speech transcription system. Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011. IEEE, 4672-4675.

KIRCHHOFF, K., BILMES, J., HENDERSON, J., SCHWARTZ, R., NOAMANY, M., SCHONE, P., JI, G., DAS, S., EGAN, M. & HE, F. Novel speech recognition models for Arabic. Johns-Hopkins University summer research workshop, 2002.

LIPTCHINSKY, V., SYNNAEVE, G. & COLLOBERT, R. 2017. based speech recognition with gated ConvNets. *arXiv preprint arXiv:1712.09444*.

MANGU, L., KUO, H.-K., CHU, S., KINGSBURY, B., SAON, G., SOLTAU, H. & BIADSY, F. The IBM 2011 GALE Arabic speech transcription system. 2011

IEEE Workshop on Automatic Speech Recognition & Understanding, 2011a. IEEE, 272-277.

MANGU, L., KUO, H.-K., CHU, S., KINGSBURY, B., SAON, G., SOLTAU, H. & BIADSY, F. The IBM 2011 GALE Arabic speech transcription system. Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on, 2011b. IEEE, 272-277.

MASMOUDI, A., BOUGARES, F., ELLOUZE, M., ESTEVE, Y. & BELGUITH, L. 2018. Automatic speech recognition system for Tunisian dialect. *Language Resources and Evaluation,* 52**,** 249-267.

MENACER, M. A., MELLA, O., FOHR, D., JOUVET, D., LANGLOIS, D. & SMAILI, K. 2017. Development of the Arabic Loria Automatic Speech Recognition system (ALASR) and its evaluation for Algerian dialect. *Procedia Computer Science,* 117**,** 81-88.

MOHAMED, A.-R., DAHL, G. & HINTON, G. Deep belief networks for phone recognition. Nips workshop on deep learning for speech recognition and related applications, 2009. Vancouver, Canada, 39.

OORD, A. V. D., DIELEMAN, S., ZEN, H., SIMONYAN, K., VINYALS, O., GRAVES, A., KALCHBRENNER, N., SENIOR, A. & KAVUKCUOGLU, K. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

PERKEL, J. M. 2018. Why Jupyter is data scientists' computational notebook of choice. *Nature,* 563**,** 145-147.

RABINER, L. R. & JUANG, B.-H. 1993. *Fundamentals of speech recognition*, PTR Prentice Hall Englewood Cliffs.

SAON, G., SOLTAU, H., CHAUDHARI, U., CHU, S., KINGSBURY, B., KUO, H.-K., MANGU, L. & POVEY, D. The IBM 2008 GALE Arabic speech transcription system. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010. IEEE, 4378-4381.

SATORI, H., HARTI, M. & CHENFOUR, N. 2007. Introduction to Arabic speech recognition using CMUSphinx system. *arXiv preprint arXiv:0704.2083*.

SOLTAU, H., SAON, G., KINGSBURY, B., KUO, J., MANGU, L., POVEY, D. & ZWEIG, G. The IBM 2006 Gale arabic ASR system. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, 2007. IEEE, IV-349-IV-352.

VERGYRI, D., KIRCHHOFF, K., DUH, K. & STOLCKE, A. 2004. Morphology-based language modeling for Arabic speech recognition. SRI International Menlo Park United States.

Appendix A

Publications

1. Ayman Mansour and Wafaa F. Mukhtar. Automatic Dialectal Arabic Speech Recognition: A Survey. The 7th Sudan Conference on Computer Science and Information Technology (SCCSIT7). International University of Africa 2020

Appendix B

The Buckwalter Transliteration Dictionary

UNICODE			BUCKWALTER	
Decimal	**Hex**	**Glyph**	**ASCII**	**Orthography**
1569	U+0621	ء	**'**	Hamza
1571	U+0623	أ	**>**	Alif + Hamza Above
1572	U+0624	ؤ	**&**	Waw + Hamza Above
1573	U+0625	إ	**<**	Alif + Hamza Below
1574	U+0626	ئ	**}**	Ya + Hamza Above
1575	U+0627	ا	**A**	Alif
1576	U+0628	ب	**b**	Ba
1577	U+0629	ة	**p**	TaMarbuta
1578	U+062A	ت	**t**	Ta

1579	U+062B	ث	**v**	Tha
1580	U+062C	ج	**j**	Jeem
1581	U+062D	ح	**H**	HHa
1582	U+062E	خ	**x**	Kha
1583	U+062F	د	**d**	Dal
1584	U+0630	ذ	*****	Thal
1585	U+0631	ر	**r**	Ra
1586	U+0632	ز	**z**	Zain
1587	U+0633	س	**s**	Seen
1588	U+0634	ش	**$**	Sheen
1589	U+0635	ص	**S**	Sad
1590	U+0636	ض	**D**	DDad

1591	U+0637	ط	**T**	TTa
1592	U+0638	ظ	**Z**	DTha
1593	U+0639	ع	**E**	Ain
1594	U+063A	غ	**g**	Ghain
1600	U+0640	ـ	_	Tatweel
1601	U+0641	ف	**f**	Fa
1602	U+0642	ق	**q**	Qaf
1603	U+0643	ك	**k**	Kaf
1604	U+0644	ل	**l**	Lam
1605	U+0645	م	**m**	Meem
1606	U+0646	ن	**n**	Noon
1607	U+0647	ه	**h**	Ha

1608	U+0648	و	**w**	Waw
1609	U+0649	ى	**Y**	Alif Maksura
1610	U+064A	ي	**y**	Ya

Appendix C

Internal view of the designed model

An inner view of the designed model during the training with its preferred 7 residual stacks setup that has been used in this research all came from the TensorBoard model's graph viewer, the first picture gives a bird view of the entire setup with the 7 residual stacks. The second picture illustrates the same design as discussed in chapter three as each stack contains a group of four residual blocks. Finally, the last picture gives a closer look to show each component of the residual block.

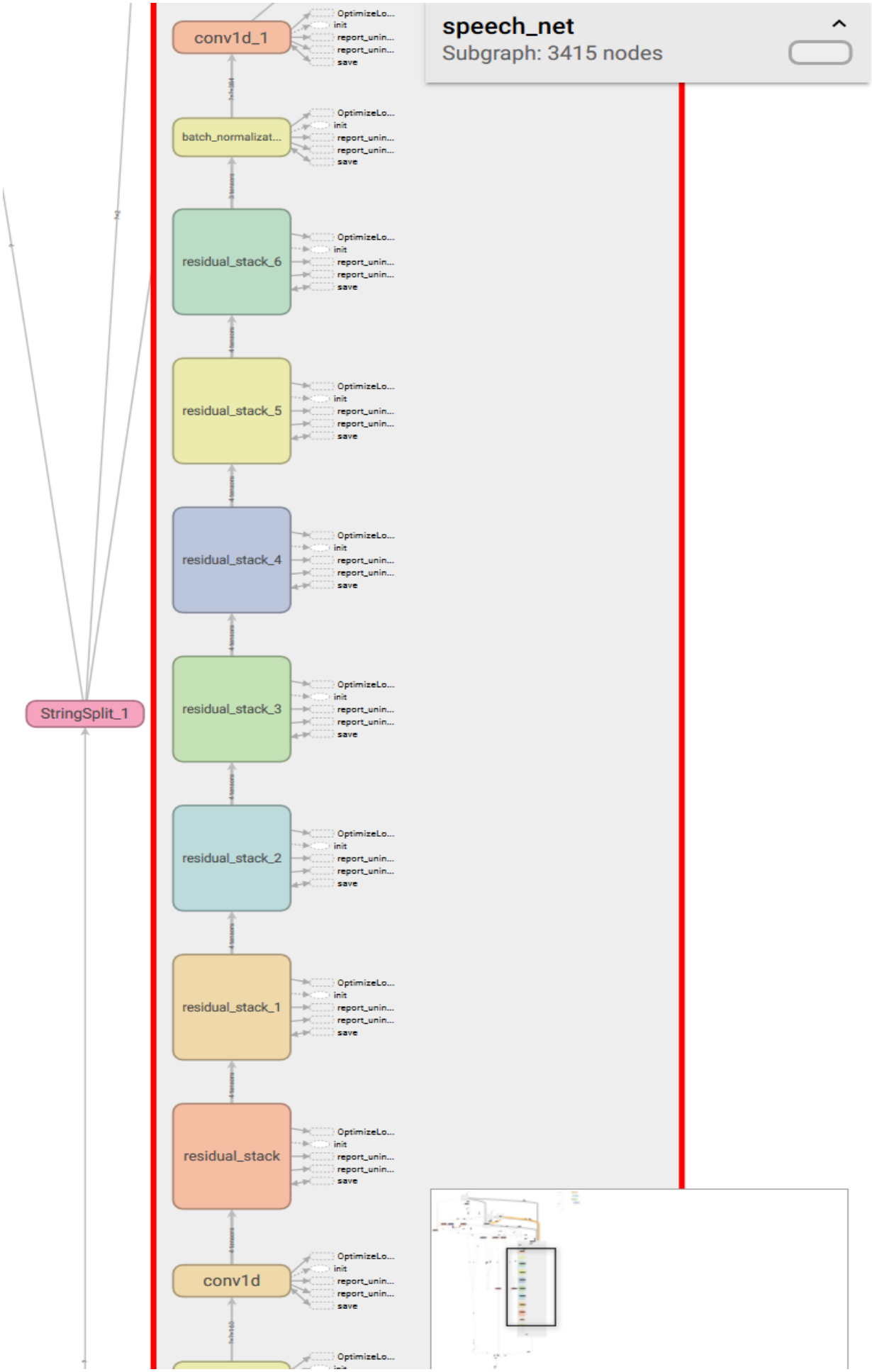

Figure C- 1: The proposed model inner view with 7 Residual Stacks

Figure C- 2: The Residual Stacks inner view with 4 Residual Blocks

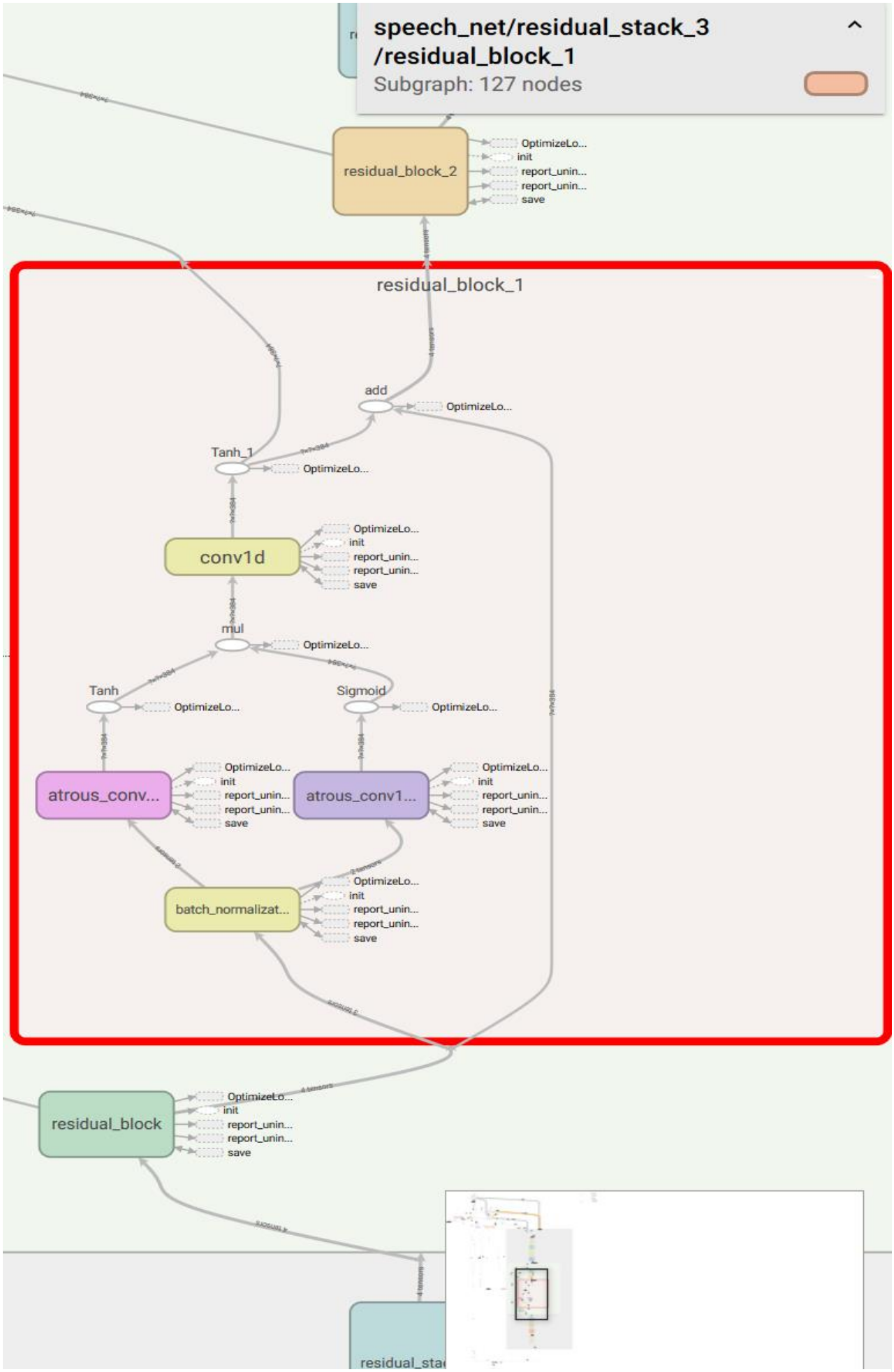

Figure C- 3: The Residual Block inner view