Sudan University of Science and Technology

College of Graduate Studies

Automatic Recognition and Identification for Mixed Sudanese Arabic – English Languages Speech

التعرف الآلي وتحديد اللغات في الكلام المختلط العربي – الانجليزي السوداني

A dissertation submitted in fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science

By

Mohammed Osman Eltayeb Elfahal

Supervisor: Prof. Dr. Mohammed Elhafiz Mustafa

Co-supervisor: Prof. Dr. Rashid A. Saeed

August 2019

Student's Declaration

I know the meaning of plagiarism and declare that all of the work in the document, unless properly acknowledged, is my own.

_____

Mohammed Osman Eltayeb Elfahal

Date: _____

Supervisor's Declaration

I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in terms of scope and quality for the award of the degree of Doctor of Philosophy in Computer Science.

Name: _____

Signature: _____

Date: _____

Copyright Transfer

I agree to assign my copyright ownership of this thesis to:

College of Graduate Studies

Sudan University of Science and Technology

_____

Mohammed Osman Eltayeb Elfahal

Date: _____

DEDICATION

To my parents' souls

To my wife

To my kids

To my supervisors

ACKNOWLEDGMENTS

In the name of Allah, the Most Gracious and the Most Merciful

Alhamdulillah, thank you my God for strength given me, take my hands at the time of frustration and Paved my way to complete this work

My thank goes to Prof. Dr. Mohammed Hafiz, who accepted the supervision of this research.

My special gratitude goes to my co-supervisor Prof. Dr. Rashid A. Saeed, who really directed my thoughts to right track and made my effort fruitful, who encouraged me and raised my confidence and self-challenges to the level of distinction. His guidance made this work possible; he is and academic life advisor much more than just a supervisor.

My appreciation goes to those who gave time to participate collecting and recording mixed speech data used to build our corpus for this research. Special thanks to my friends in WhatsApp groups Lamtna (undergraduate classmate, Sudan University of Science & Technology) and Alzaidab high secondary school classmate group, who actively responded to the collection campaign, to my friends in Sudan University of Science and Technology and Fedail Hospital. Special gratitude and appreciation go to my wife's colleagues at Petro Energy exploration and oil production company, who gave their valuable time recording speech samples for this research.

My thank goes to my batch one PhD student's classmate, who kept asking for the status of my research offering unlimited assistance, special thanks to Dr. Wafa Faisal Mukhtar and her family for encouragement and giving time participating in recording audios for training.

Much thanks to my close friends, Dr. Hisham Ibrahim Izzeldin, Dr. Tarig Karar and Dr. Abu Sabah Elfatih for their encouragement to complete the work for this study.

I would like to acknowledge Sudan University of Science and Technology for partially sponsoring my PhD studies, great thanks to colleagues at Computer Center – SUST for their support and maintaining my internet access.

Dr. Mohammed Mustafa Ali, the symbol of hope, am happy being his friend, his discussion, encouragement and assistant made this work possible.

My kids, Abdelrahman, Abdel Nasir and Osman, beside their real contribution recording speech sample, they kept their smile when am back home tired, frustrated and bored.

My wife, Ilham Medani, the light living in the heart of our home, she committed to share me all time at the bright and dark moments along the journey to conclude this work. Her efforts are notable, listening to me, discussing my ideas and leading the family team with her kids recording all speech samples for this works. She reviewed and formatted the whole thesis.

ABSTRACT

Mixed speech is the phenomena of using more than one language in a single sentence, this occurs in communication between bilinguals to express their ideas and thoughts using vocabulary of both languages, even occurs among none bilingual people to describe product originally from second language.

This thesis addresses the problem of mixed speech communication in multilingual communities. This is regional problem faces shortage resources and studies. Sudanese Arabic and English languages are the two languages selected for this research to build a generalized mixed speech and language identification model, the first is common and formal language among the Sudan and the latter is international, language of science and primary lesson in Sudan education systems.

For experimental purposes, mixed speech corpus was built including most frequent daily life Sudanese Arabic and English mixed sentences, collected through social media applications campaign, 75% of this collection is read by 87 bilingual Arabic natives in office environment resulting in 2289 audio files associated with their transcription for training purpose, considering speakers and code-switch types, environment as factors affecting performance of the model at recording time.

Based on the assumption that native language dominance others in mixed speech, proposed solution for generalizing recognition model is centered around Sudanese Arabic language. The solution keeps the original words for each language participates in switching in all components of the model such as mixed phonetic dictionary, mixed languages lexicon, etc., except for Acoustic Model (AM) Arabic language is used instead of its original language based on assumption that native speaker does not suddenly reconfigure his articulation organs to produce sounds as natives do.

Open source CMU SHPINX is adapted for this mixed speech task, proposed model, which is consider effected by native language dominance, outperforms existing single pass and multi pass models achieving overall accuracy of 33.05% in term of Word Error Rate (WER). Mixed speech produce hybrid language not belong to each participating language, interface for further linguistic computation is provided to deal with this new language. The interface contains recognized word, its order in the sentence, recognition confidence and its language identity. Language identification in

the model is simply looked up identity from mixed languages lexicon to avoid effects of unclear language discrimination attributes in such speech.

Achieved results, prove the possibility to generalize the model based on Arabic language, module for phonemes clustering and comparison needed to serve as front-end to detect new language phonemes that are not included in phonemes set in order to add new language to the model.

المستخلص

يعرف الكلام المختلط على انه استخدام اكتر من لغة في جملة واحدة اثناء الحديت، وهي ظاهرة تنتشر وسط المتحدثين باكثر من لغة للتعبيرعن افكارهم بشمل ادق، وقد تحدث وسط المتحدثين بلغة واحدة فقط باستخدام عبارات او اسماء منتجات غير معرفة لديهم في لغتهم الام.

هذا البحث يقدم مقترحات للتعرف علي الكلام واللغات في الحديث المختلط بشكل عام، والحديث المختلط في الحياة العامة السودانية بين اللغة العربية كلغة ام وهي اللغة الرسمية، لغة التعليم الساسي ولغة التواصل بين مكونات المجتمع المختلفة واللغة الانجليزية كلغة ثانوية وهي لغة عالمية الانتشار ولغة العلم والتكنولوجيا الحديثة، وتسمي الاولي لغة الاساس والثانية اللغة المضمنة. من معوقات البحث العلمي في هذا المجال نقص المصادر التي تستخدم في اجراء التجارب وتقييم النتائج، هذه المشكلة اكثر وضوحا وتـأثيرا عندما تكون مشاكل البحث محلية او اقليمية وليست من مشاكل المناطق ذات الانتاج العلمي الغزير. لاغراض هذا البحث، عبر اطلاق حملة عن طريق وسائل التواصل الاجتماعي تم جمع الجمل المختلطة بين العربية والانجليزية الاكثر استخداما في الحياة اليومية السودانية، 75% من هذه الجمل تمت قراءتها بواسطة عدد 87 متحدث منتجة عدد 2289 ملف صوتي فب بيئة المكتب العادية، مع مراعاة العوامل التي تؤثر علي الصوت مثل نوع المتحدثين واعمارهم ولهجاتهم وذلك لبناء نموذج صوتي معتمدا علي العربية السودانية، وذلك لافتراض ان المتحدث لايعيد تشكيل اعضاء انتاج الكلام عند الانتقال من لغة لاخرى وانما يستخدم تشكيل لغته الام.

تم تكييف برنامج SPHINX مفتوح المصدر للتعرف علي الكلام من جامعة كارنيق ميلون الامريكية، وتم الحصول علي اداء بنسبة خطأ قدرها 33.05% . تعتبر اللغة المختلطة لغة هجينة لا تنتمي لاي من اللغات الموجودة في الحيث، وتصبح بالتالب غير معرفة خارج نطاق النموذج، لذلك تم اقتراح واجهة للربط مع النظم الاخرى التي تعتمد علي معرفة هوية اللغة في المعالجة، تحتوي الواجهة علي الكلمة المتعرف عليها، ترتيب الكلمة في الجملة، موثوقية التعرف اضافة علي هوية لغة الكلمة التي تم الحصول عليها بالبحث عنها في قاموس اللغات الذي تم تصميمه بغرض التعرف علي اللغات في بيئة الكلام المختلط حيث ان السمات التي تميز لغة عن اخرى تكون غير واضحة.بناءا علي هذه النتائج يمكن تعميم هذا النموذج باستخدام اللغة العربية،

مع اضافة معالج يصنف ويقارن اصوات اللغة المراد اضافتها مع الاصوات الموجودة في النموذج ثم يتم اضافة غير الموجودة

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artificial Intelligence
AM	Acoustic Model
ANN	Artificial Neural Network
ASM	Acoustic Segment Models
ASR	Automatic Speech Recognition
ATIS	Air Travel Information Service
B LSTM	Bidirectional Long Short-term Memory
CMU	Carnegie Mellon University
DARPA	Defense Advanced Research Projects Agency
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transformation
DNN	Deep Neural Network
DTW	Dynamic Time Wrapping
EM	Expectation Maximization
ERR	Error Rate
FACST	French Arabic Code-switching Triggered
GLDS	Generalized linear Discriminant Sequence
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
ICT	Information and Communication technology
IFP	Interface for Further Processing
IPA	International Phonetic Alphabet
JFA	Joint Factor Analysis

LDC	Linguistic Data Consortium
LID	Language Identification
LPC	Linear Predictive Coding
LRE	Language Recognition Evaluation
LSA	Latent Semantic Analysis
LSTM	Long Short-term Memory
LVCSR	Large Vocabulary Continuous Speech Recognition
MFCC	Mel Frequency Cepstral Coefficients
MGB	Multi-Genre Broadcast
MLLR	Maximum Likelihood Linear Regression
MSA	Modern Standard Arabic
NIST	National Institute for Standard and Technology
NLP	Natural Language Processing
OGI	Oregon Graduate Institute
OGI -ST	Oregon Graduate Institute of Science and Technology
OOV	Out Of Vocabulary
PAM	Prosodic Attribute Model
PDF	Probability Density Function
PLP	Perceptual Linear Predictive
PMC	Model Composition
POS	Part Of Speech
PPR	Parallel Phone Recognition
PRLM	Phone Recognition following by Language Model
PSE	Phone Selection by Elimination

PSID	Phonological Segment Inventory Database
PSR	Perceptually Significant Regions
RNNs	Recurrent Neural Networks
RTF	Real Time Factor
SDC	Shifted-Delta-Cepstral
SEAME	South East Asia Mandarin-English
SMAP	Structural Maximum A Posteriori
SUST	Sudan University of Science and Technology
SVM	Support Vector Machine
TDNNs	Time-delay Neural Networks
TOPT	Target-Oriented Phone Tokenizer
UCLA	University of California, Los Angeles
VQ	Vector Quantization
VSM	Vector Space Model
WER	Word Error Rate
WRR	Word Recognition Rate

CHAPTER I

INTRODUCTION

1.1. Overview

Speech is the natural, easy and common approach of communication between humans. People are expert in their native languages in terms of speech production and understanding, every group of people use specific languages to communicate and deliver messages among them. The language itself is a set of codes, their meaning is predefined for its speakers, particular language codes are not defined to every human speaking other languages. The technology of today bring up the world to your computer and cellphone near your physical location, your language alone is not enough tool to communicate globally And to exploit human production in all aspects of life in the world with 7111 living languages, 3,116 of them are spoken only without written systems (Eberhard, 2019). Existence, availability and ease of use of social media and communication applications eliminates literacy barrier from global communication in term of using speech rather than written messages, it is easy for every one that uses cell phone to establish conversation using speech facilities with other people, even the language used has orthographic system or not (Montag et al., 2015). It is not possible for anyone to know all the languages he needs in his daily life affairs. If someone who knows a large number of languages, he can transfer this knowledge to others through training and education that requires great effort and time, instead one learned machine could be replicated to unlimited numbers of other machines without extra effort.

Linguistic computation studies are researches aim to build computer software that analyze, produce, interpret and understand human speech and act accordingly. Speech production and interpretation is a primary task of human articulation system as speech producer and brain as interpreter, these easy and basic functions of humans manipulating, cost decades of research to make machines partially address speech as human. Automatic Speech Recognition (ASR), or simply talk to computer, is a part of linguistic computation researches that uses computer software to process incoming speech and output the equivalent text as the result, this may be ultimate goal or the

outputs may use for another speech-enabled process where ASR considered a front-end module in this case. Speech production and understanding is a language dependent tasks, in other words, each language has its own alphabets, phonetic inventory, words vocabulary and linguistic rules, for generalization automation of speech processing in multilingual environment, a front-end module for language identification is essential to determine language identity contained in speech and directed to suitable speech processing process for specific language (Jurafsky and Martin, 2009).

Usually, speech communication is done using one language both speaker and the listener understand, which is called monolingual conversation, but bilingual and multilingual speakers tend to use more than one language in a single sentence. This phenomenon is called mixed speech or code-switching which is defined as using more than one language in a single discourse (Mabule, 2015). Mixed speech is a great challenge of speech recognition models because the number and location of participating languages are not known.

The technology of speech processing, makes dealing with machines verbally possible. Human can instruct robots that handle dangerous task at unreachable places, at the time when typing is not possible or the hands are needed to perform other tasks, well trained machines to recognize speech play a major role as intermediate translator between people speaking different languages. Speech as biometrics, could be used to control computerized personal devices, such as cell phone, personal computers and saving box utilizing human voiceprint for secure access. Speech processing technologies aid the visually impaired people by enabling speech interaction with machines, and hearing impaired by changing speech to readable and understandable visual representation for them. Speech interface could be designed for information inquiry and instructions execution; moreover, existence of built-in speech input/output interface in the machines facilitates the human-machine interaction (Matarneh et al., 2017).

In spite of the existence of Arabic language worldwide as a first or second language, its automatic language processing researches are still far from concrete in comparison with achievements for other languages like English due to the lack of resources e.g. Arabic speech corpora and linguistic complexity and dialects variation (M. Osman Eltayeb and Mohammed Elhafiz Mustafa, 2013) (Panayotov et al., 2015).

According to Ethnologue (Eberhard, 2019), Arabic languages comes on the fifth order ranking based on the number of speakers following Chines, Spanish, English and Hindi languages. Arabic is a language with formal version Modern Standard Arabic (MSA), which is only used in education and news broadcast and of four regional dialects with too many slangs for each one at each smaller geographical regions and ethnic groups (Ali et al., 2015), Arabic is Islam religion language for 1.65 billion people uses it in their daily religion activities representing 24.0% of the world population with expected to be one out of for in 2020 and one out of three in 2075 (Kettani, 2010).

Arabic language variations makes resource lack severe for some dialects more than other, since speech processing tools developed for MSA are ill-suited to generalize for dialects recognition due to difference in context meaning, pronunciation and phonemes' articulation (Kwaik et al., 2018).

Sudanese Arabic dialect is a common communication language for the people of Sudan. It is distinct version of Arabic language that affected by existence of 75 spoken regional and ethnic groups local languages, such as *Zaghawa* language spoken by 180.000 people in Darfur and the part of chad (Eberhard et al., 2019 online version: http://www.ethnologue.com.). The effect of local languages in Sudan result in generating more slangs for Arabic, which leads to generate hierarchical structure of languages and dialects.

In spite of existence of large number of local languages in Sudan, English dominant as a second language for educated Sudanese people, because it is a formal principal language beside Arabic, exist in education system, language of learning in universities for long time, it is also language of international communication, knowledge sharing and the language of smart machines, so, appearance of English words within Arabic sentence is normal in daily life, embedding English words in Arabic conversation is common behavior among young educated urban. Mixed speech becomes communities accepted phenomenon. Among the Sudan where local language is used and dominant, mixed speech occurs between local languages and Arabic, the latter considered second language in conversations.

1.2. Motivation

Availability of smart machines, in terms of existence and cost affordability,

gives owners benefits of performing their daily life tasks ranging from greeting someone, transferring massive cash or perform mathematically complicated tasks.

Utilizing these capabilities is limited due to difficulties facing some people dealing with the machines through its specified interface. This happens due to variation in education level, foreign interface's language or disabilities hinder their hands doing their job. Speech, a popular and human mastering tool that used for communication, regardless their education levels, with other people or even with other creatures shared living surrounding environment e.g. dogs, is an appropriate tool to deal with machines also, by raise machines level of smartness to interpret, understand and react in response to human speech. In our communities among the Sudan, where literacy rate is low, hand disabilities resulting from various civil wars is high, existence of speech interface to smart machines is essential to perfectly utilized its capabilities, taking into account existence of local languages in daily use among Sudanese communities.

1.3. Problem Statement

Speech is foremost human to human communication's tool. Every adult without disability is expert and well mastering his native language in terms of speech production and understanding. Advances in technologies get machines and smart devices closer to human, to take advantage of these human capabilities to communicate with machines. Automatic Speech Recognition (ASR) and Language Identification by machine are ongoing research topics for decades have achieved significant performance in terms of robustness, accuracy and computation cost (Sriram et al., 2018).

However, achievements in this domain assume speech is composed from a single language, which its words vocabulary, phonetic inventory and linguistic rules are previously known. A hybrid language that is composed from two or more languages is a reality and communities' acceptable phenomena among bilinguals and multilinguals, even worse, in limited cases, among monolingual borrowing terms from foreign language e.g. ***Mobile***. This phenomenon is raised by dominance of some languages e.g. English or by diverse existence of multiple local languages among ethnic and geographical groups shared daily life activities such as in Sudan. This phenomenon is known as mixed speech or code-switching (Adel et al., 2015). In Sudan, the country with named 75 local languages, national Arabic language and dominance of English (Turrini, 2018), the usage of such hybrid language is apparent in daily life

communication.

The language attributes and cues exploited by machine learning techniques to recognize speech and identifies language in mixed speech is obscured due to dominance of first language speech articulation process over second language (Hassan and Hassan, 2014). This concludes, adapting existing speech recognition methodologies bias towards native language.

Mixed speech usage is a local problem. Most probably mother language participates in the phenomena as the first language, this resulting in the lack of speech resources (corpus) to utilize by researchers for the purpose of experiments trying different approaches and techniques.

1.4. Aims and Objectives

This research is aimed to build a generalized solution interface to be adapted for various speech enabled applications and smart machines instructions based on Sudanese unique pronunciation and speech articulation process, using their native languages either monolingual or mixed with other languages, and establish required resources for researches on linguistic computation for Sudanese monolingual and hybrid languages. The following are objectives to achieve these aims:

- To collect and manipulate mixed speech corpus of Sudanese Arabic – English frequent daily life mixed speech sentences.
- To adapt developed speech processing technique for mixed speech recognition and language identification.
- To model a recognizer to handle the problem of native language effects of other languages in mixed sentences.

1.5. Research Questions

The goal of this research is to build speech recognition and languages identification model that is generalized for any language, dialects, type of speech in monolingual or mixed modes, specifically the following questions are addressed for this thesis:

1. What types and size of existing mixed speech corpora for Sudanese Arabic – English languages?

2. Why are existing speech recognition and language identification models not suitable for processing mixed Sudanese Arabic – English languages?

    a. What are limitations and drawbacks of existing approaches of speech recognition and language identification in mixed speech mode?

    b. How a single front-end speech and languages recognizer for speech enabled application built and generalized based on Arabic language in multilingual and accents variation communities such as HAJ?

    c. How existing monolingual speech recognition and language identification software could be adapted to serve mixed speech processing:

3. What is the impact of dominance of Sudanese Arabic language pronunciation on other languages participates in mixed speech?

    a. Is model will biased toward native language?

    b. Are others language's discriminative attributes exist and clear?

1.6. Thesis Contributions

1. **Investigates and emphasizes the major reasons behind mixed language trends:** the study concluded that the phenomenon is very prevalent among people, who's their native languages are non-English. As the majority of the technical jargon and terminologies are originally from English, which is the lingua franca in science and in research as general, non-English speakers may not be able to express their ideas in their native languages only, but instead they used mixed sentence form from English words with their native language. The phenomenon is apparent if we consider the fact that non-English speaking countries represents the majority of the

world population.

2. **Build a mixed speech corpus for Sudanese Arabic - English languages: for** the purpose of experiments of this research a mixed speech language speech corpus with both Arabic and English has been built. One unique feature in this corpus that makes it different form other mixed languages corpora is that its Arabic language is primarily a Sudanese dialect. According to the best of my knowledge there is no such speech corpus in terms of languages it was contained and structure of dictionary and language model. The corpus will freely be released as contribution in the domain of research in mixed speech processing, Sudanese languages linguistic studies, biometric information extraction (gender, age, etc..) and languages, accents and dialects identification. The corpus has been gathered by launching a campaign using several resources like social media applications, e.g., WhatsApp and Facebook, SMSs and direct interviews. Hundreds of sentences have been collected this way. For the experiments in this study, 201 distinct sentences were chosen to build the corpus form 20 participants. In particular, a more than one hour of speech with 2289 audio files for 87 Sudanese native speakers along with their associated transcription are created. The selection criterion is mainly focused on the diversity of the sentences and speaker gender, age and ethnic group as these factors have a significant role in the phoneme articulation process. Along with this corpus, a distinct list containing 474 Arabic – English words phonetic dictionary is also built and accompanied by a tri-grams statistical mixed language model for the chosen set. A mixed language lexicon utilized for language identity look up is also created.

3. **Develop a set of tools and resources to support mixed languages processing:** these are: a mixed phonetic dictionary generator, which is based on the richness of the Arabic language phonemes set to represents sound of each language. And audio transcription tool was developed. Since the process of audio transcription is a time-consuming process, such a tool could have a significant impact on reducing such text dependent manual transcription. Based on the richness of phonemes representation in Arabic, a generalized phoneme set that representing most world languages phones, with ability to add new phoneme not seen, were created for the

purpose of this research and other related researches, this articulation-based sound representation could help enrich resources for spoken only languages.

4. **Develop speech recognition and language identification techniques of mixed speech:** this is the major contribution of this study. In which, error propagation problem resulting from different processing stages was eliminated to a single point, unclear transition at switching points and effects of native language on other languages are resolved, and language identification accuracy is no longer affected by incoming utterance length.

5. **Develop a generic language interface to outside world:** for this hybrid language, to facilitate its usage in language dependent speech processing environment, an interface to deal with outside world was developed, it contains recognized word, its original language identity tag and recognition confidence for Permissible threshold.

6. **Develop approach for identifying languages presents in the mixed speech:** In developed technique, language identity is looked up and matched against a set of entries exists in languages lexicon for each output word from mixed speech recognizer. This approach eliminates effect of error propagation in the multi-pass approach, articulation process and pronunciation dominance of native languages over other languages and cancel the bad influence of ununiform speech length to language identification accuracy.

1.7. Out Lines of Thesis

Speech processing in general and for mixed speech specially were introduced in chapter 1 of this thesis, the problem of code switching is emphasized, motivation for building model for mixed speech recognition also addressed, aim and objectives are explained along with community and knowledge contribution of this work.

In chapter 2, types of speech used in speech recognition are addressed with the most challenging issues of the process that are not solved yet or should be utilized for more model robustness, approaches that researchers follows for the task of speech recognition. Researchers works in the domain of speech recognition and language identification were reviewed and analyzed.

Proposed solution of mixed speech and language identification are explained in chapter 4, design and collection methodologies of special mixed speech corpus creation will be investigated, steps of building universal acoustic model, mixed language model, phones set, mixed phonetic dictionary and mixed language lexicon were detailed and explained. Setup for testing the performance of mixed speech acoustic model are showed.

In chapter 4 of this thesis, mixed speech corpus analysis and statistics are showed, gender and switching types besides words distribution of each language also analyzed.

Experiments and results are the topic of chapter 5 of this thesis, Results for mixed and monolingual speech experiments are analyzed and discussed, their performance discussion addressed and stated.

Chapter 6 concludes the work done in this thesis and hints future direction to fill the gaps towards generalized speech processing systems.

CHAPTER II

AUTOMATIC SPEECH RECOGNITION AND LANGUAGE IDENTIFICATION

Humans always find a way to communicate even if they speak different languages. In fact, they interact with machines in their daily life with a personal computer, smart phone, car navigator computer, aircraft autopilot, service meter (electrical, Gas, etc.), etc. Transferring human ability of producing and interpreting speech to machine helps much in dealing with it in a natural way instead of using keyboard to write commands or pressing instructions keys, keeping in mind human are experts in their speech by nature in terms of speech production and understanding. Speech processors nowadays exist in many applications that need human-machine interface, such as phone calls processing, security access applications, query-based information like travel information systems or voice-based information retrieval, weather reports or price inquiries.

Automatic Speech Recognition (ASR) is the research domain that tries methodologies to make machines mimic human in producing and react to speech through utilization of information technologies techniques either hardware circuits solutions or software applications ASR (Forsberg, 2003, Juang and Rabiner, 2005). Rresearches started early in 1920 by Bel Lab. Since then, software solutions are dominant exploiting statistical and Artificial Intelligence (AI) algorithms and theories. Although it is slow processing compared to hardware solutions, but it is cheaper in terms of developing requirements, maintenance and its lifetime. Research progress in speech recognition is variant based on the language, because it is a language dependent task (Bhuvanagirir and Kopparapu, 2012), some language researches go further e.g. English and Japanese in terms of resources availability and standard for works evaluation and real applications deployment, and poor for other languages like Arabic in spite of it is the fifth world language (Kwaik et al., 2018, Eberhard et al., 2019).

Language identification (LID) is a front-end module for speech enabled applications in multilingual communities where people speak different languages. For the purpose of direct speech to suitable processor because speech processing task is

language dependent it is the goal of the automatic language identification, to transfer most accurate human ability of identifying languages to the machine (Muthusamy, 1993).

Availability of labeled or transcribed data, which is time consuming task, for language helps researchers much to proceed in developing speech enabled applications for target language (Muthusamy, 1993). Phenomena of mixed speech in bilingual speakers' environment were raised up by social media applications and became acceptable in societies (Rallabandi et al., 2018). The following sections give detailed background of speech recognition process and review the works done and their achievements exploiting different algorithms and techniques.

2.1. Overview

Human speech perception mechanism is the methodology utilized by machine learning, linguistic computation and speech processing researchers for speech processing tasks. Cochlea, fluid filled spiral part in inner ear, respond to eardrum vibration generated from incoming sound wave pressure. Each part along the cochlea respond to specific frequency band of the vibration and generates nerve impulses. Cochlea electrical signal travel through auditory system to the brain, where they are analyzed and interpreted according to their excitation regions. Different experiments with linguistic and acoustic language properties shows human fast and accurate ability of identifying his native languages and other languages with little knowledge. Beside human reasonable evaluation of completely unknown languages identities, ability of human infant of identifying languages raises the assumption of acoustic signals properties conveys much language dependent information, since infant knows nothing about language linguistic rules (Ramus and Mehler, 1999, Elliott and Shera, 2012).

2.2. Types of Speech

The ultimate goal of speech processing system is to process speech at continues mode as produced and interpreted by human. Continuous speech is subjects to too many variabilities factors; researches are done for different type of speech. The following types of speech were used in speech processing (Saini and Kaur, 2013, Benzeghiba, 2007), each type serve specific purposes:

2.1.1 Isolated Words Speech

This type of speech recognition model accepts speech word by word in isolated manner. The ASR system process a single word of speech at a time, no need for determining the start and end of the input. Applications of this type of systems are suitable for auto calling telephone, using voice print as a biometric verification access control and speakers related identification such as identity and gender or for languages and dialects or accents recognition and verification (Ananthi and Dhanalakshmi, 2013).

2.1.2 Connected Words Speech

Speech recognition for connected one, is the system that deals with continuous speech recognition system type where it is parts separated clearly and in a uniform manner, so the methodology for determining the beginning and end of each part was uniformed and cleared. The system of connected word speech, beside all those for isolated words, are used for read speech and broadcast news transcription (Benzeghiba, 2007).

2.1.3 Spontaneous Speech

Speech recognition systems for isolated words or that for connected words separated in a uniform manner are considered trails along the line of building robust speech recognition system for free conversation of spontaneous speech. The most challenging issues in the tasks of mimic the ability of human interpreting speech in such situation were detecting the start and end of words, since they are affected by the context (previous and after words) and mode of speakers along with environment effects (Pascale Fung, 2008)

2.3. Speech Signal Analysis

Each sound produced by a human is shaped by his vocal tract and other parts of body participates in this production process, which is called articulation process. The job of features extraction techniques is to capture this shape in terms of attributes that uniquely describe one sound (phoneme). Naturally, speech signal beside original message (phoneme) contains a lot of information regarding speech (discreate, continuous and read or spontaneous), speaker (gender, mode, age, accent), languages

identities used in conversation and environmental effects such as other speaker or TV speech interference or background noise, capturing device quality and its recording parameters (channel type, Hertz and resolution), speech signal analysis or feature extraction is essential shared step for dimensionality reduction and extract most task related information contained in the signal for all speech processing tasks. Step of speech analysis and features extraction efficiency impact the overall system performance. Extract task most relevant information and dimensionality reduction are the main goals of this step. Digital signal processing is a science responsible for analyzing speech signal to extract its spectral features. The following approaches based on human auditory and perception mechanism were proved to be most effective methodologies (Verdet, 2011).

2.3.1. Perceptual Linear Predictive (PLP)

Uses concepts of critical-band spectral resolution, equal-loudness curve and intensity-loudness power law of psychophysics of human hearing process to derive auditory spectrum (Hermansky, 1990) for further automatic speech processing.

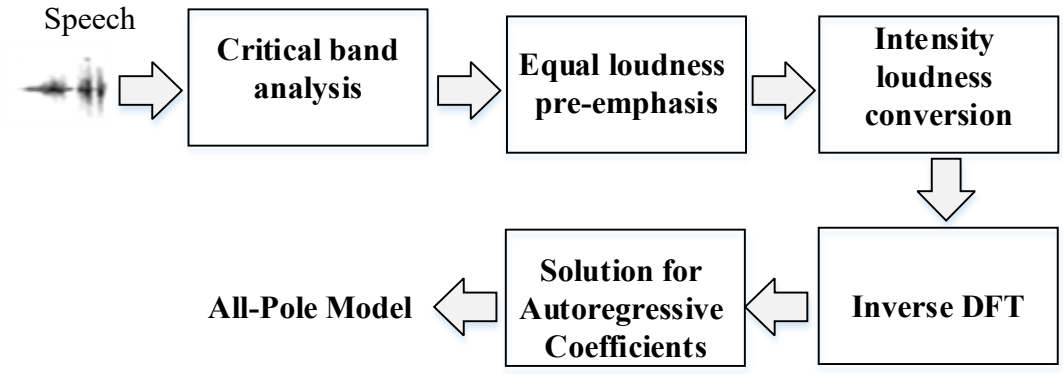


Figure 2.1: PLP Speech Analysis

2.3.2. Mel-Frequency Cepstral Coefficient (MFCC)

Mel Frequency Cepstral Coefficients (MFCC) is a state-of-the-art feature extraction technique that mimic human process of interpreting and understanding speech signal. As human cochlea does, MFCC firstly determined the frequencies component in each frame. Since the cochlea has a problem to differentiate between too

close frequency bands, the Mel Filter Bank is used to sum up adjacent frequency bands, which gives the idea of how much energy exists in specific group. Mel filter banks are linearly spaced along the logarithmic Mel frequencies that human perceived speech and raise the sound accordingly instead of Hertz frequencies. Equations 2.1 and 2.2 change between Mel and Herts frequencies respectively (Haizhou Li, 2013).

$$M(f) = 1125 \ln\left(1 + \frac{f}{700}\right) \tag{2.1}$$

Where f the frequency in Hertz

$$H(m) = 700\ \left(e^{\frac{m}{1125}} - 1\right) \tag{2.2}$$

Where m the frequency in Mel scale

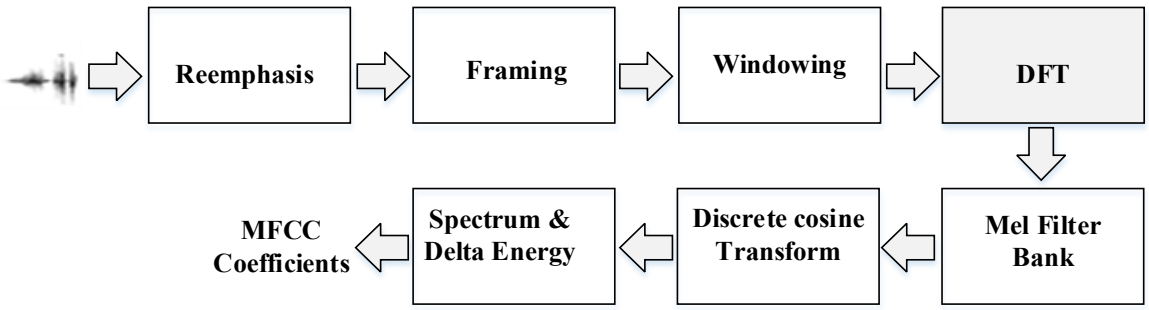


Figure 2.2: MFCC Features extraction technique

As illustrated in Figure 2.2, for MFCC analysis low frequencies boost, then signal framed to stationary parts and each frame edge smoothed using windowing. To get frequencies that are represented in the signal Discrete Fourier Transformation (DFT) technique were applied. Then, applying Equation 2.1 frequencies changed from Hertz to Mel scale, these scales are spaced evenly using tringle filter bank. Logarithmically filter bank returned to time domain using Discrete Cosine Transform (DCT). The delta applied to capture nonstationary information for this representation.

2.3.3. Linear Predictive Coding (LPC)

The idea is to get correlation coefficients that linearly predict current sample from previous samples, with error approaching zero. The method encodes spectral envelope (to extract formants) of good quality speech at low bit rate (Haizhou Li, 2013).

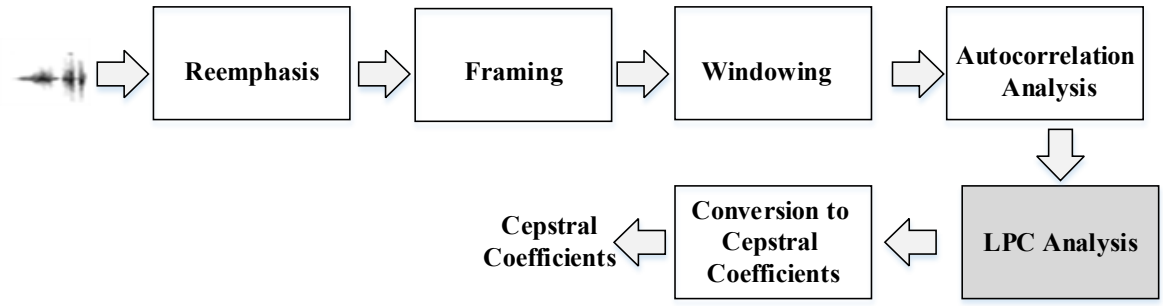


Figure 2.3: LPC speech analysis

2.4. Speech Modeling Techniques

The building block of speech is the smallest unit which is called phoneme. These units together form a syllable, word, sentence and other bigger form of the discourse (Hall Jr, 1961). Human is baseline and most effective speech interpreter and recognizer system, analyze these sequences of phonemes to understand speech. For a machine to understand and react to speech, research in the domain of linguistic computation and natural language processing focused in finding a way to properly recognize parts of speech based on previous knowledge. It is better to consider the smallest and building block unit of speech for recognizing any part of speech either phonemes, words or a whole utterance, the effort is directed towards separating each phoneme form the rest with respect to surrounding context of the utterance. Figure 2.4 shows 5 phonemes in 2-dimensional space built based on feature space X= [x1, x2], the size of ellipses for each phoneme and overlap between them clearly represents the feature of speech variability nature that complicates speech processing task.

Acoustic Model (AM) is a statistical representation of speech features that assign each phoneme to a phonetic cluster based on some similarity measure, it is a relationship between signal and associated sound, each distinct phoneme in the language has a unique statistical representation as illustrated in the figure 2.4 built using

training algorithm with speech training data, ellipses represent phonemes clusters while closed circles represent a new part of speech (phonemes) to be classified according to some similarity of known phonemes representation (clusters). Most world high population languages have monolingual trained acoustic model globally and freely available for research use. This is not true for recent opened research domain of mixed speech and language recognition. For this bilingual mixed speech task, for collection of mixed speech based on Sudanese accent variation and dialects for both Arabic and English languages our own universal acoustic model were built.

Figure 2.4 represents clusters for 5 phonemes. The clusters shapes and overlapped raised a question of how to measure the similarity between data point and clusters. Traditionally used of k-means clustering algorithm is the simple answer which is based on the distance measurement between data point and centroid of each cluster, which is defined as the mean of the cluster, with the assumption that size and shape of clusters are same, the nearest cluster to new data point is picked as its label (phoneme symbol).

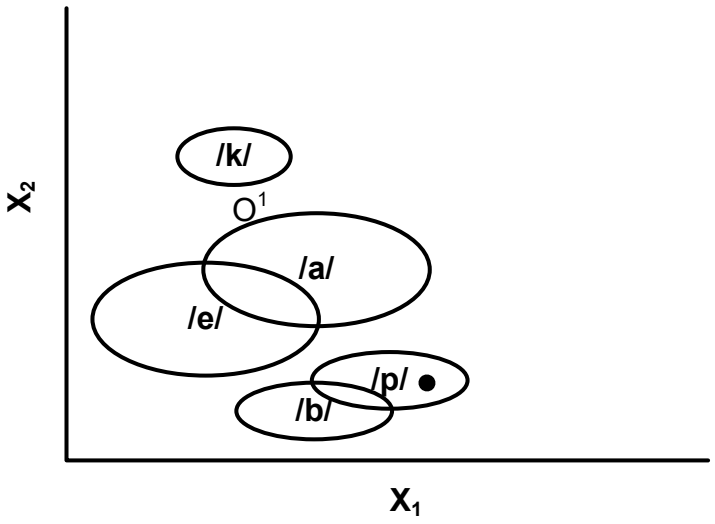


Figure 2.4 Phonemes statistical representation and its search space

Similarity measurement method based only on the mean value does not take into account the variability of speech represented by the ellipse shape. This is clear when measuring similarity for O^I data point which is close enough to centroid of cluster /k/ even it is placed at the edge of cluster /a/, so, this concludes that the distance from the

data point to the center of the cluster is not a suitable attribute for similarity measurements for clusters with different sizes and shapes.

For speech modeling, Gaussian Mixture Model (GMM) introduce soft clustering technique which is a way to add specific data points to certain clusters to some probability ratio. Soft clustering determines the similarity of objects to certain cluster based on the center of the cluster (mean) and how the data sparse from this center (size or shape). A version of normal data distribution is the Gaussian distribution, which is used to form components of mixture of gaussians for each phoneme over its Probability Density Function (PDF) (Law et al., 2004).

2.4.1. Gaussian Mixture Model (GMM) Clustering

Due to variability of speech, which its signal features defined as stochastic random data, phonemes were not deterministically clustered, we have to find a way to add specific phoneme to certain cluster to some probability. One phoneme may span the wide area that interferes with another phoneme's ellipse as illustrated in Figure 2.4.

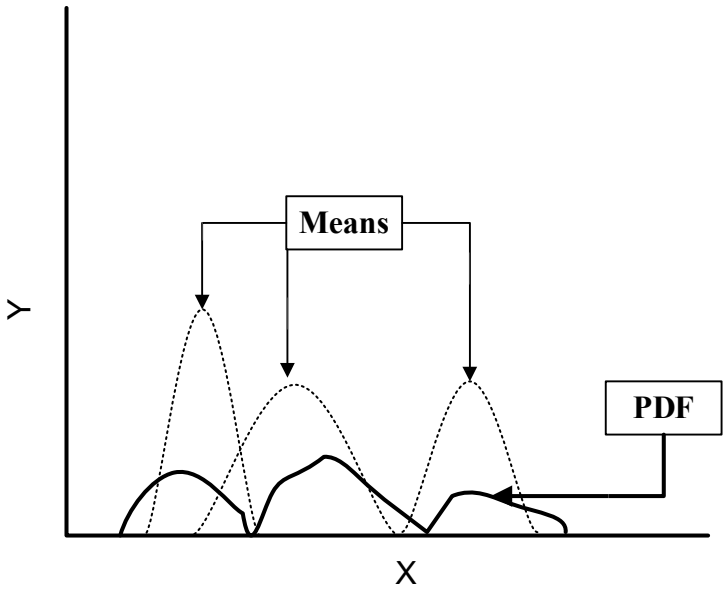


Figure 2.5: Gaussian components and their PDF

In speech processing, Gaussian Mixture Model (GMM) is defined as a parametric model that shaped using features of speech signal (Reynolds, 2015). Figure 2.5 shows the concept of Gaussian Mixtures, three gaussians (components) represented by dotted lines, each has its own curve, mean and variance, which collectively form their PDF, which is represented by solid line in the figure. The figure shows that

components of PDF are interfered and overlapped. Phoneme features vector is considered belong to specific cluster according to some probability ranging from 0.0 to 1.0. Each mixture is the Gaussian (normal) probability distribution represented by Equation 2.3, that takes into accounts the shape of data distribution based on the following parameters:

- Mean: Represents the centroid of the data (μ)
- Variance: Horizontal measure represents how data is sparse out the mean (σ)
- Weight: Vertical measure represent the size of the data that component contains (the height) (w).

For each feature vector its mean determined the centroid of the curve, variance determined how data point far from the center of features vector (mean) and the weight determine the size of the data under the curve, in other words, weight measures how many vectors that composed mixture component.

$$N(X|\mu,\sigma) = \frac{1}{\sqrt{2w\sigma^2}} e^{-\frac{(\mu-\sigma)^2}{2\sigma^2}} \quad (2.3)$$

In N-dimensional space of features, there is a mean vector and covariance matrix composed multiple probability distribution, resulting in components equal to existence phonemes. This representation is called Gaussian Mixture Model. The GMM is a normal probability distribution of the features produce one or more peaks represent mixtures components or phonemes. Equations 2.4 represents parameters of the GMM of sample data with mean and variance.

$$\lambda = (\mu, \Sigma, \mathrm{w}, \mathrm{k}) \quad (2.4)$$

Where:

μ Vector of mixtures means

Σ Matrix of covariance

w Vector of mixtures weights

k Number of mixtures (components)

Subjects to:

$$\sum_{i=1}^{k} w_i = 1,\ 0 \leq w_i \leq 1 \qquad (2.5)$$

The question is, how parameters of Equation 2.4 are obtained to correctly classify a new data point. These are model parameters, for each mixture component, its parameters are extracted at training step from training data using Expectation Maximization (EM) algorithm (Juang and Rabiner, 1991). The EM is an iterative algorithm, with convenient property that the maximum likelihood of the data strictly increases at each subsequent iteration, in other words, it is guaranteed to approach a local maximum. The algorithm works on suitable initial values for model parameters, iterates along the training data and update parameters values at each turn depends on maximum likelihood of feature vector (Do and Batzoglou, 2008).

To measure similarity of the model to new data point, using Equation 2.6, responsibility function value assigned to each mixture and the cluster with maximum probability is chosen as the label of new data point.

$$P(X|C_i) = P(X|\big(N(\mu_i, \sigma_i, w_i)\big) \qquad (2.6)$$

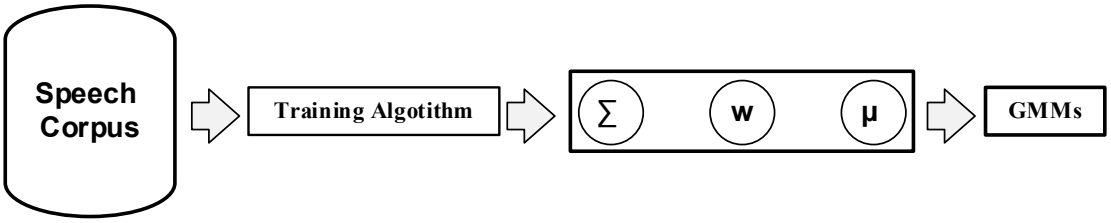


Figure 2.6: GMMs parameters generation

The GMM is a good tool for soft clustering of discrete observations. In speech processing, where observations are in continues and in sequence manner, the GMM leaves us with unresolved issue, which is determination of transitions from cluster to other. Occurrence and cooccurrence of language phonemes should be determined from training samples (Shental et al., 2004).

2.4.2. Hidden Markov Model

Hidden Markov Model (HMM) is transitions probabilities from states to others, are calculated based on the theory of Markov Chain (Juang and Rabiner, 1991, Eddy, 2004), for example, in speech processing the cluster of sound /p/ is much likely cluster of sound /b/, we have to find a way to differentiate between them according to some probability of cooccurrence from current cluster (sound) e.g. /h/ sound (Guttorp and Minin, 2018), this stochastic behavior of speech need uncertainty representation which is provided by HHM (Preeti Saini and Kau, 2013). Figure 2.7 raise phonemes clustering overlap and interference problem, which is occurred due to the speech variabilities result from environment or speaker, the figure shows that the sound /s/ followed by either sound /p/, /b/ or /t/, the three sounds have some degree of similarity. This scenario indicates that hard clustering is not well suited to utilize. To resolve the issue, probability ratio is assigned to each transition link between two sounds. The probability is modeled during training stage and formed structure of HMM for previously created GMM. Transition probability from sound /s/ to sound /t/ represented by P(st) and to sound /b/ by P(sb) and so on, the probabilities out of specific phoneme must summed up to 1.

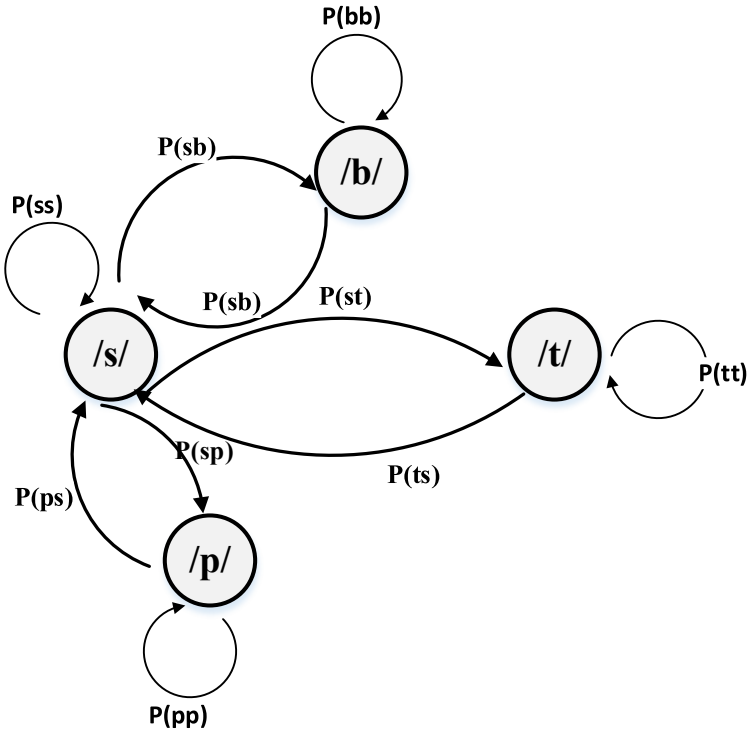


Figure 2.7: HMM concept

Actually, the sound represents the state is hidden, what is given is some kind of observations or features, for speech, are the spectral features extracted from speech signal. Modeling smallest speech unit (phoneme) grantee ease of adding new bigger speech units formed using phonemes sequence, such as syllable, words or sentences. Each node represents phoneme model, its HMM, that may have three or more states. HMMs are linked by transitions probabilities, each branch ends up representing one word from training set vocabulary. All HMMS formed what is called words lattice. The responsibility to select right branch (word) during recognition time is the job of Language Model (LM) (Saini and Kaur, 2013).

2.4.3. GMM - HMM

The GMM is a good tool for soft clustering phonemes of speech, which takes into accounts the environment and speakers' effects, such as background noise and accent variations. The GMM is not considered effects of phoneme location in speech utterance (context), which is resolved by HMM (Ibrahim M. M. El-emary et al., 2011). A GMM-HMM is left-to-right three states Hidden Markov Model, where observation given to state of HMM is the parameters of GMM (μ, σ) and transitions probabilities of states will be calculated using training data set and GMM parameters as illustrated in Figure 2.8, where:

μ, σ are picked from GMM models for the phonemes

$P(i|j)$ is transition probability from state i to state j calculated at training time as follows:

$$P(i|j) = \frac{occurence\ of\ phoneme\ of\ state\ i\ followed\ by\ phoneme\ of\ state\ j}{occurence\ of\ phoneme\ of\ state\ i\ in\ the\ whole\ training\ data} \qquad (2.7)$$

HMM three states represent link to previous state, static part of the sound in the middle and the end state which represents transition station to next phoneme.

The ultimate goal of this training process is to generate acoustic model, which is actually words lattice that gives speech recognition decoder the values of posterior property of the term $P(w|O)$, probability that word is W given features O

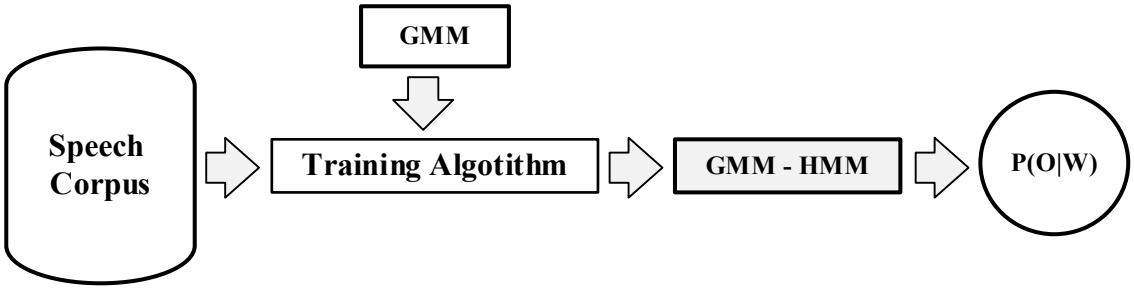


Figure 2.8: GMM – HMM acoustic model training

In this model, GMM is used to model the distribution of the acoustic characteristics of speech, whereas HMM is for model time sequence of speech signals, that statistically represents smooth transition of speech articulation process from phoneme to other. GMM – HMM acoustic model gives a maximum likelihood posterior probability of word regardless of the context of its utterance as illustrated by Equation 2.8, which gives the HMM model (states sequence) from the set of trained models that maximally generate same observation as speech part do.

2.5. Automatic Speech Recognition Processing

The following sections of this chapter discusses the concept behind process of speech recognition and reviewed the works and efforts done by the researchers towards building robust speech recognition application.

2.5.1. Speech Recognition Approaches

During decades, researchers investigates varies techniques and exploits different types of attributes that speech signal conveys. These techniques are grouped in the following three main approaches (Preeti Saini and Kau, 2013).

2.5.1.1. Acoustic Phonetic Approach

This approach developed upon assumption that each language has distinct set of a phonetic units called phonemes. Each phoneme is characterized by its own acoustic attributes that corresponding speech signal conveys. Incoming speech signal were segmented to small stationary parts to overcome change over time property of speech. Then, spectral analysis that change the signal to discrete representation in frequency domain were applied, features extraction techniques suitable for extracting most

relevant task attributes that speech signal contained applied, that considered distinctively describes speech unit. The classification task for this approach is responsible for selecting most appropriate prelabeled reference phoneme that maximally match the features in current speech signal as recognized phonemes. Acoustic phonetic approach is cost effective compared to that approaches exploit language dependent linguistic rules with minor accuracy deficiency. Adding new language to exist model of this type just need a bulk of labeled data of target language without needs to linguistic experts.

2.5.1.2. Pattern Recognition Approach

This approach is based on grouping speech parts according to some similarity measurement techniques. Two main stages for this approach were conducted, firstly, building model or reference templates of phonemes, syllables or words are created using predefined and labeled parts in training stages, and then, new speech pattern were processed in the same way and compared to a set of reference models created previously, to pick most similar one using efficient classification technique. For training phase, labeled speech parts were fed to features extractor and modeling steps to build a reference model for each phoneme in the training data set. To test the performance of the model, unseen test speech pattern follows the same signal processing and features extraction in training step. The resulting features vector compared to each trained model, to pick the best match according to pattern classification technique used. Template comparison method faced by the problem of existence of nondistinctive template for the same sound due to speech variability factors. To resolve the later issue, the concept of soft clustering that add a single data point to each cluster to some probability degree was introduced. The HMM is successful statistical method used for this approach, which is model both speech context as transitions probability and observations as spectral attributes of speech represented as model's states.

2.5.1.3. Artificial Intelligence Approach

Capturing surrounding attributes at speaking time that humans use to interpret speech, were targeted by researchers to build smart model utilizes this surrounding knowledge. Fusion between acoustic phonetic and pattern recognition approaches

results in developing classification rules for speech pattern. This approach faced a problem of experts' unavailability. They are needed for transferring their linguistics knowledge to machine as classification rules to machine in interpretable ways.

2.5.2. Challenges of Automatic Speech Recognition

Automatic speech processing systems, in compare to human base line system, faced by many factors that decrease their accuracy and performance, their stability and reliability subject to many environmental variability and context factors (Benzeghiba et al., 2007, Forsberg, 2003).

2.5.2.1. Speech Context Difference

ASR system, have a current speech signal and trained model to recognize incoming speech, whereas humans in the same situation utilize his background knowledge of speaker and his mode, subject they talking about and most probably have a good knowledge of language used in communication with its syntax and linguistic rules. Statistically, context knowledge could be achieved during speaking, but speaker status is still far to optimally achieved. Capturing and utilizing visual features of speaker at speaking time is utilized and enhance recognition rate (Corona et al., 2017)

2.5.2.2. Absence of Linguistic Rules

Linguistics rules that govern sentence formation are always relates to written medium of communication, written message is a one-way communication while spoken messages occurs in two-way. The speech is most probably occurring informally in environment subject to interference between speakers or environmental sounds, disfluencies and accents variation, subject and speaker's mode changes. The situation gets worse for those languages with no written systems at all (Thanh, 2015).

2.5.2.3. Environment Effects

Humans could make reasonable judgment to understand and interpret a message embedded in speech, even if it is partially heard. Part of message may be lost due to environmental effects, such as another sound presents at speaking time not related to conversation. Speech capturing device or other devices or software that used to change waveform to digital or discrete content may also affect original signal quality. ASR

should determine such overhead and developed techniques to filter and isolate them away from the original message (Seltzer et al., 2013).

2.5.2.4. Speaker Effects

Speech articulation process differ from human to other. This results in distinctive sounds pronunciation for everyone. Speaker's current mode and conversation context may instantly change, accordingly, his speaking style changed, people are speaking differently at different situation and time. Emotional status clearly expressed during discourse, naturally, speaking style changes based on the place, to whom one speaks and present emotional situation. Focal tract difference between male and female is come up with distinctive properties of speech based on speaker gender and age, results in different formant and fundamental frequencies (Floccia et al., 2006, Ververidis and Kotropoulos, 2006). Accent variations and language dialects, variant language spoken by specific geographical or social community, are challenging factors that should be overcome along the line of speech recognition model generalization. Bilingual speakers that uses foreign and mixed speech during single discourse add another complexity dimension of automatic speech processing system.

2.5.2.5. Capture Body Language Messages

Human can communicate silently using only signs language or expressed their speech with help of body movements. Body language most probably clarify the mode of the speaker, if these visual movements captured and efficiently processed, especially for the parts of body directly participates in sound articulation like mouse and lips, will enhanced the degree of speech recognition accuracy (Liew and Wang, 2009).

2.5.3. Speech Recognition System

Building speech recognition system more or less tried to follow human being methodology of producing, interpreting and understanding message embedded in speech. To transfer this ability to machines, researchers tried a lot of methodologies and techniques. Speech processing by machine is multidisciplinary domains includes linguistic computation, statistics, signal processing and machine learning. Training and testing are the main phases of automatic speech recognition as illustrated in Figure 2.9. Training phase is a step were algorithms and techniques used with a sufficient amount

and variants of labeled and transcribed speech data to create a reference models of speech. For unseen speech at testing time, the utterance goes through same process of preprocessing and feature extraction before presents to trained model for recognition (Benzeghiba, 2007, Preeti Saini and Kau, 2013, Matarneh et al., 2017)

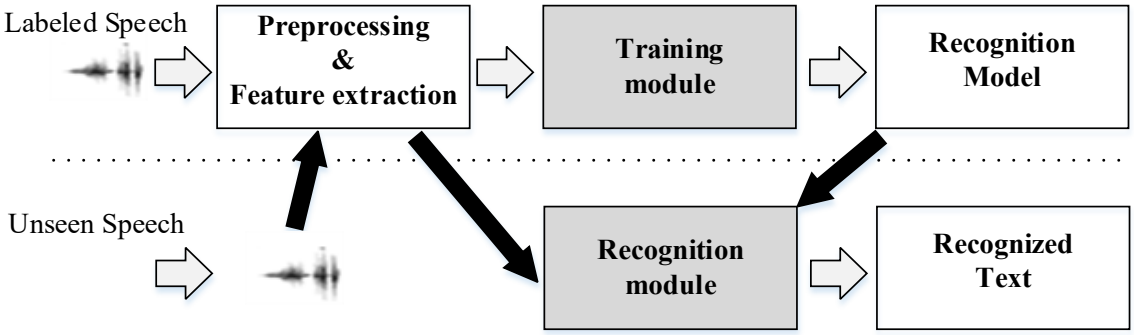


Figure 2.9: Automatic speech recognition

2.5.3.1. Training Phase

In speech processing, training phase, accept task related labeled and transcribed speech data. For the task of speech recognition, speech audio files accompanied with their transcribed files feed to suitable training algorithm, in help with phonetic dictionary to produce a recognition model. At recognition time the model used to recognize unseen speech message as illustrated in Figure 2.10.

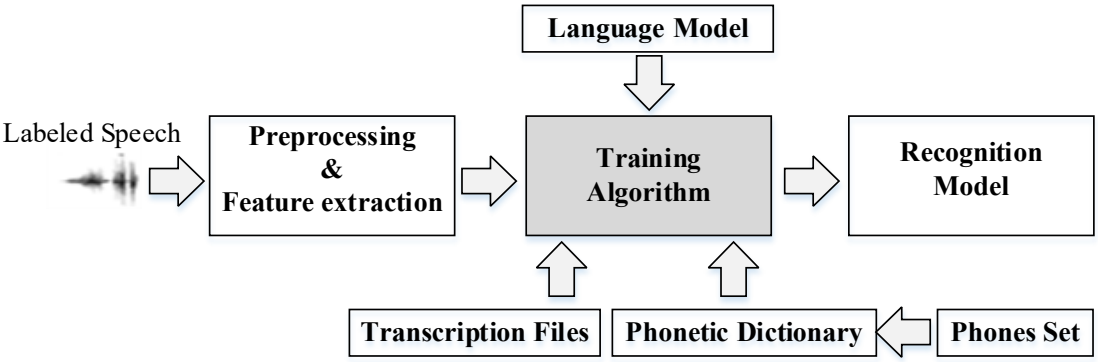


Figure 2.10: Training phase

The following modules and resources, collectively or individual, are generalized requirements for each automatic speech recognition system.

a. Preparing Data

Speech corpus, is a collection of speech audio files recorded with specific parameters for in hand task. These files accompanied by its transcription (text of speech) and phonetic dictionary. These components make up a training corpus for building recognition model. The following are steps of preparing training data.

b. Data Acquisition

Automatic speech processing aims to learn machine how to interpret and understand human speech. This done using bulk of speech audio files, recoded using specific device, preset quality and environment properties. The audios files are accompanied by their transcription files contains text of speech inside the file.

c. Speech Transcription

To generate speech recognition model, speech sample files with its associated text were needed at training time to generate a model. Transcription, a process of write speech utterance content in equivalent text, is a most time-consuming task along the line of developing speech recognition system. Most probably is done manually by native speakers to ensure correctness, because it is essential factor affect the ultimate goal of the model, which is achieve high accuracy.

d. Phones Set Preparation

Phonetic codes, are the symbols representing language sounds (phones), used to write speech in sounds codes rather than alphabets, keeping in mind some living languages have no written system. International Phonetic Alphabet (IPA) is standard collection of phonetic symbols representing most live languages phones, its symbols categorized based on articulation and production mechanism of phones by human. For specific task and languages, developers can build their own special and related phonetic

symbols that represents all phones in hand because some languages phones are not all included (Association and Staff, 1999).

e. Building Phonetic Dictionary

Phonetic dictionary, in a simple case, it is a list of two columns, the first represents the word written in orthographic units (letters) for specific language, and the second column is the same word written in phonetic symbols of speech (phonemes).

f. Language Model

A single word may have different meanings depend on its context, the situation getting worse for the language like Arabic where the meaning of the word differ depends on its context and long vowels or nunation that appears with its letters. Word context is defined by its precedence and following words. Language model is statistical representation of word context based on occurrence and cooccurrence words of current word, the designer chooses order and length of the context by determining the number of words that participates in probability calculation that appear before current word, which is called language model order, e.g. A 2-gram language model, means occurrence probabilities calculated for standalone appearance, and for cooccurrence with each other word.

2.5.3.2. Testing Phase

Speech recognition model built in training phase, should be tested to measure its accuracy and efficiency of recognizing speech sample not encountered during training phase. Incoming input speech sample goes through same process of preprocessing and extracting task related features.

With the help of in hand recognition model, and its assistant components (phonetic dictionary, phones set and language model), speech recognizer decode incoming speech features to their equivalent text. Figure 2.11 illustrates the follows and module of testing speech recognition model phase.

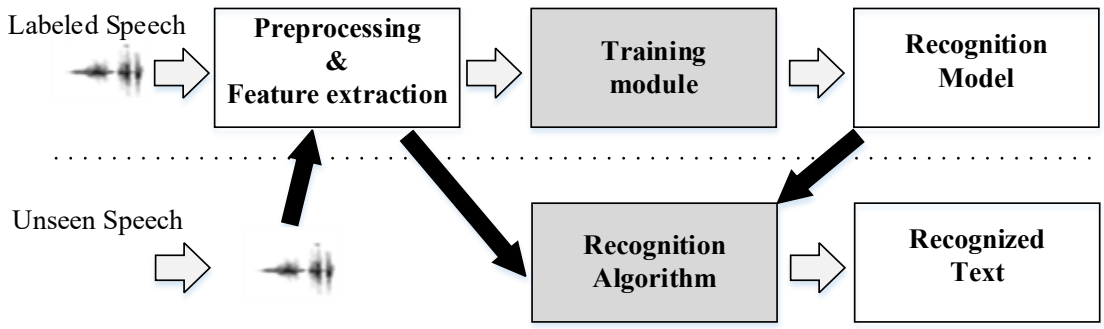


Figure 2.11: Testing phase

2.5.4. ASR Performance Measurements

The performance of automatic Speech recognition system is measured by two factors, accuracy that measure recognition correctness and speed that measure how fast the system process incoming speech (Preeti Saini and Kau, 2013).

Accuracy is generally measured by calculating Word Error Rate (WER) ,which estimates how result differ from the reference in terms of detecting deletion, insertion and substitution of words that occurs in the result compared to its reference (M.A.Anusuya and S.K.Katti, 2009). WER is calculated as:

$$WER = \frac{D + I + S}{N} \tag{2.9}$$

Where:

N is number of words in the references

D the number of missing words from the result that found in the reference

I the number of words in the result not found in the references

S the number of words that wrongly recognized

WER is a metric of how the result differ from the references, and Word Recognition Rate (WRR) that measure the correctness of the result in compare to its reference could be infer from WER as:

$$WRR = 1 - WER \tag{2.10}$$

The second performance measure of speech recognition system is how fast the recognition process completed in specific machine. The metric of Real Time Factor (RTF), is calculated to measure system speed as:

$$RTF = \frac{T}{D} \tag{2.11}$$

Where:

T is the time of processing

D is duration of the input

For the real time performance time needed to complete the task should be equal or less of the duration of the input, ratio should not exceed 1 (Ivanov et al., 2016).

When measuring the performance of automatic speech recognition system, some factors should be kept in mind, especially, when comparing the performance of the systems. Implementation environment in term of machines specifications, and speech variabilities ranging from speaker, existence of background sounds to capturing device specifications that may differ from session to other. These performance measurement issues, tell that metrics should be calculated in a single test environment or the result accompanied by its environment specifications of speech and processing computer.

2.5.5. Related Works

The Radio Rex, a toy that seem it is a voice recognizer, was manufactured in 1920, is considered the first trial along the research journey of linguistic computation (David and Selfridge, 1962). Rex is a dog toy comes when he called out his cage. The idea is based on mechanical movement that occurs due to release of lever on the back of the cage push Rex out in response to sound. The arm released in response to 500 Hz resonance as first formant generated by sound of vowel /e/ in the Rex name. The Voder (Voice Operating Demonstrator), speech synthesis machine, developed for speech synthesis, presented at the World fair in New York city in year 1939. The machine generates speech based on buzz and hissing voice followed by special keys representing bank of filters (Dudley, 1939). The first attempt of exploit statistical theories and linguistic rules of allowable phonemes cooccurrence were appeared in 1959 in England for English language, performance of a phoneme's recognizer of four

vowels and nine consonants were enhanced utilizing information of allowable sequences of phonemes in the English words contains two or more of targeted phonemes (Denes, 1959). A limited speed of the computers during the mid of 20th century, motivate the researchers to build special hardware computers for speech processing (J.Suzuki and K.Nakata, 1961, Sakai and Toshiyuki, 1961, Nagata, 1964). In spite of speed and accurate speech processing by the hardware comparing with a software, using hardware for the task does not goes further. Building special purpose hardware is commercially expensive, need periodical maintenance and big well-suited space. Furthermore, the researches in computer processor speed and size compaction goes further adding power to software processing in general.

Speech recognition systems for isolated words achieve encouraging performance utilizing hardware or software solutions. Naturally, speech is continuous with many factors of variabilities through time scale of speech events. To overcome this problem, a time-normalization method based on detecting start and end of speech were developed. This approach significantly reduces the effects of variabilities for the systems. Comparing two speech segments regardless of their time span, is considered the starting point of connected words recognition (Martin et al., 1964). This motivates utilizing Dynamic Time Wrapping (DTW) algorithm, for comparing two speech segments in different time scale (Vintsyuk, 1968).

The statistical model based on transitions representation power of Hidden Markov Model (HMM) were initially applied during 1980s. The HMM is utilized instead of pattern matching approaches that dominant 1970s researches, exclusively in IBM, Institute for Defense Analysis and Dragon Systems laboratories. It is capturing transitions from phonemes to other represented as states based on weighted (probabilities) links (J. Ferguson, 1980). The SPHINX software for continuous speech processing is developed, in collaboration between DARPA and CMU utilizing HMM (Lee, 1990).

DARAP with collaboration of IBM released a real speech-to-text systems. They aim to make it easy for humans detecting, extracting, summarizing and translating important information embedded in text instead of listening to speech over telephone or unrestricted conversations, especially for foreign languages. The Effective Affordable Reusable Speech-to-Text (EARS) system is developed with more sophisticated features

or dealing with spontaneous broadcast and foreign conversation. The system has abilities to detect sentences boundaries in ununiform speech manner. It is handling fillers and non-speech sounds during discourses and disfluencies of foreigners' (Y. Liu, 2005, H. Soltau, 2005).

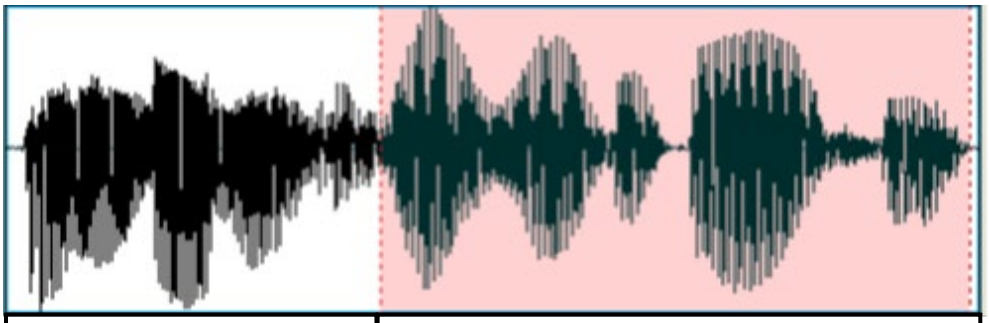
A Deep Neural Network (DNN), a type of Artificial Neural Networks (ANN), that uses more than one hidden layer between its inputs and outputs layers. This ANN configuration exploits error backpropagating derivative trained from nonlinear data, were applied for task of automatic speech recognition as a generative model in place of its traditional usage as discriminative one for acoustic modeling. The technology is proved to outperform the GMM when large amount of nonlinear training data was in-hand, and a hardware that suitable for multi-hidden layers processing is available. Determining the optimum number of hidden layers, network structure, computation cost and overfitting reduction techniques are still research challenge (Hinton et al., 2012). For the realized tasks, where interference by reverberation, background noise or other talker occurs, a DNN were used when the microphone is far from target speaker and in a circular geometry. Experiments done in clean English utterance corrupted by adding reverberation and different types of noise. Used methodology, shows significant of improvement in multichannel automatic speech recognition (Sainath et al., 2017). Based on studies that shows ability of Neural Network to model sequences features and capturing temporal context, a deep learning topologies such as feed-forward Deep Neural Networks (DNNs), Time-delay Neural Networks (TDNNs), Long Short-term Memory (LSTM) networks and Bidirectional LSTMs (BLSTMs) were utilized with different versions of Multi-Genre Broadcast (MGB) Arabic broadcast corpora that contains conversations, interviews and reports speech for Modern Standard Arabic (MSA) and other Arabic dialects for different types of broadcasting ranging from news reports to movies and comedy programs. Kneser-Ney smoothed a 4-gram language model were created. The setup achieves overall performance of 24.25% in WER measurement (Najafian et al., 2017).

The performance of speech recognition systems was acceptable in type of uniform speech in specific environment. Robustness of such systems under speech variability factors is still challenging task. Mixed speech, which is worldwide phenomena of bilingual and multilingual speakers, has a great researchers' interest in

the regions of the world where mixed speech sentences are occurred in daily life communication, following section define the phenomena and reviewed the work done.

Using more than one language in the single discourse called mixed speech or code-switching. Bilingualism and Multilingualism encourages people to use words and phrases from different languages in their speech to easily express their thoughts and describes things perfectly. The habit of such daily life mixed conversation is globally increased and acceptable, especially in Asia and Africa. Code-switching classified of two types based on the position of switching point, Inter-Sentential type, it is switch occurs at sentence boundaries, either at the beginning or end of the sentence, e.g. **عندنا *presentation, remote* تحت الشاشة.** Inter-Sentential type, were switching occurs in the middle of the sentence with no pause or indication of the switch, e.g. ***party*** **حضرت الشركة.** (Myers-Scotton, 2017).

Figure 2.12 illustrates female speaker's signal pronounced mixed sentence " **عندنا Presentation**". The figure shows the signal of speech, words pronounced, language of each part and translation of complete sentence in Arabic respectively. From my being a member of the Sudanese society, I ensured this sentence is always pronounced like this mixed way, because it is used among educated people and relates to new technology usage it is terminology in Arabic, even exist, is not known. Such short of terminologies motivate speakers to deliver their idea in others languages.



Signal		
Words	**عندنا**	**Presentation**
Languages	**Arabic**	**English**
Translation	**عندنا عرض تقديمي**	

Figure 2.12: Arabic-English mixed speech signal

Traditional speech recognition system assumes single language in incoming speech utterance. In mixed speech processing, another dimension of complexity was added. The number and locations of languages are not previously known. Researches in the task of mixed speech is recent, two approaches were dominant, multi-pass approach illustrated in Figure 2.13, in which, the point of switching was detected first followed by determining the identity of the language in the segment. Then utterance processed by particular language dependent speech recognizer (Lyu and Lyu, 2008).

This approach is subject to performance degradation upon error propagation. The methodology is determining switching point, identify the language of the segment and the language dependent speech recognizer applied. Each step may propagate its error to subsequent steps. Explicit language identification is affected by native language pronunciation on other languages participate in speech. Insufficient requirements to optimally determines the identity of the language in the segment may encounter due to short segment length leads to obscure of language discriminates attributes.

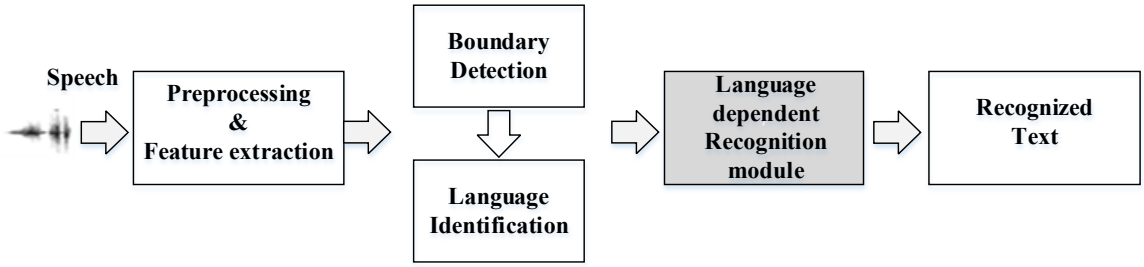


Figure 2.13: Multi-pass mixed speech recognition

The second approach, is a one-pass recognition methodology, where determining languages in the utterance is not explicitly done. Recognition process is launched directly through multiple recognition models of each participant language or by using a unified single model for all languages followed by decision methodology to select appropriate result for each approach (Imseng et al., 2011, Yeh and Lee, 2015).

Mixed speech is occurred in informal conversations. Is classified as under-resourced process, because it is recorded resources is rare, most of the available collections are task specific collections.

For investigating effects of Cantonese to English in Hong Kong, where two languages may use together in single discourse, mixed speech corpus between them were built. It contains 17 hours of code-switched speech, recorded by a youngsters gender balanced graduated 80 speakers, whose Cantonese is native language and fluent in English (Chan et al., 2005). Switches are manually determined and transcription of each word is done in its original language.

Motivated by language identification research in code-switching, South East Asia Mandarin-English (SEAME) was initialized. In year 2015 the corpus made globally available through Linguistic Data Consortium (LDC), that contains 129 hours of transcribed spontaneous speech, to serve all types of automatic speech processing in mixed speech mode. The speech is conversation and interviews with young university staff members and students of balanced gender (Lee et al., 2017).

Annotated with sentence boundaries and switches points, 5 hours for 3614 sentences of mixed speech from Turkish-German languages conversations were created. Linguistic experts are responsible for collecting and annotated processes. 28 university students were participating in recording speech. Transcription is done in original languages (Çetinoğlu, 2017).

Read speech, that simulates spontaneous, of Algerian Arabic and French mixed speech corpus was collected. A university graduated bilingual adults 20 speakers (10 males and 10 females), who lived part of their life in Algeria and other in France were selected based on their usage of code-switching in their daily life to answer questions by linguist for recording in sound proof or calm room of total 7.5 hours of speech. The speech was then segmented to average 6 seconds length segment base on speaker and language change. Each French segment transcribed in French script, while Algerian Arabic writing uses modified Buckwalter Arabic transcription. Even the conversation led by linguistic towards code-switching occurrence, the French language is dominant and being the first language and Algerian Arabic is second language of this mixed speech corpus (Amazouz et al., 2018).

Code switching between Japanese and English is more frequent in Japanese daily life, speech recognition for such speech is needed, but resources collection is expensive and time-consuming task. Utilizing text-to-speech synthesis tool for Japanese and English languages, a 280 thousand mixed utterance were created. The text was

crawled from existing bilingual databases e.g. travelling information database (Nakayama et al., 2018).

A first Arabic code-switching spontaneous speech corpus were recorded in quiet closed room from informal interviews about technical topic with bilingual Egyptian Arabic-English languages speakers. A young teaching assistant of both genders whose native language is Arabic and fluent in English were participates in collection. Transcription was done manually by annotator fluent of both languages, non-speech sounds were transcribed by special fillers tags. 4.3 hours of mixed speech were recorded and transcribed for 12 speakers evenly distributed for both genders corresponding to 1234 sentences; with combination of 62.1 % Arabic words and the rest are English words. For analysis purpose part of the corpus is annotated for Part-Of-Speech (POS) tags, the tagging analysis conclude English nouns are mostly participates in switching, and for Arabic one third switch to English follows and article (Hamed et al., 2018).

It is clear that most available collection either for regional languages or dialects, proved our claim of problem statement regarding the regionality nature of the task. For Arabic language available data set are for Algerian and Egyptian Arabic dialects mixed with French and English respectively. These Arabic corpora, even it is for spontaneous speech, were recorded in quiet and closed room for specific dialects and not publicly available so far.

A multi-pass and one-pass approach that trained on monolingual speech data were tested and compared for Chinese dialects (Taiwanese- Mandarin) code switching. Share of common formal written language of Chinese dialects encourage a single-pass approach utilizing a universal acoustic model and lexicon for Chinese character-based recognition utilizing distinction of some Taiwanese phones. compared to complicated multi-pass and stage effected results, the study concludes promising performance achievement for simpler single-pass approach (Lyu et al., 2006).

A monolingual and parallelized automatic speech recognition models for isolated words for five European languages were tested. In the parallelized setup, the language identity is not previously known. Language identity and utterance recognition chosen through maximum likelihood calculation. The study concludes that significant difference between two types of models. It is advised universal phoneme set for lexicon generation (Imseng et al., 2010).

For mixed speech of Hindi as first language and English as a second language, a one-pass recognition system was built. Ready CMU English acoustic model was exploit to test mixed utterance read from sheet written in Hindi script for 225 mixed sentences. Phonetic dictionary was built using only English phones set. The Hindi phones not exist in English were produce by English phones combination. This approach gain performance of 53.73% in the metric of WER (Bhuvanagiri and Kopparapu, 2010).

Code-switching in conversational speech is more challenging task than uniform speech. Based on phones merging strategies for Mandarin-English in conversational mode, a performance of 36.6% were achieved in term of WER measurement. A 36 hours of speech data from SEAME corpus for Mandarin-English speech was used, recording environment is a close-talk microphone in a quiet room. A bilingual dictionary for English and Mandarin were created from merged two CMU dictionaries of these languages. Some approximation were done to overcome pronunciation difference (Vu et al., 2012).

A deep learning neural network were exploit for Frisian and Dutch languages mixed speech, models of phone set for each language and merged phone set of two languages were tested. The merged phone set of Frisian and Dutch languages mixed speech gives better performance of 59.5% WER (Yılmaz et al., 2016).

Recognition of under resourced South African isiZulu language mixed with the formal English language of the country speech were performed. Monolingual and mixed speech for both languages from broadcast South African soap operas were collected and transcribed. The corpus contains 75% of monolingual English language and the rest is mixed between two languages. Pool phones set were created, language dependent and independent setups were investigated. The performance of all setup is above 80% WER. The study concludes language dependent setup is performed well due to dominance of phonetic characteristic of language being spoken in time. The researchers noted that, confusion occurs between phonetically similar speech parts (van der Westhuizen and Niesler, 2016)

Algerian Arabic mixed with French data corpus were utilized to test detection of code switching. The technologies of automatic language identification and automatic speech recognition were exploited. The length speech segment needed to complete the task is considered. Firstly, Maghrebian broadcast for entertainment shows speech

collection contains speech for Algeria, Morocco and Tunisia speakers were used to measure code-switching frequency for target speakers. For the rest goals, FACST (Amazouz et al., 2018), speech corpus were used for testing two languages dependent models trained over hundred hours of speech for each language. For ach language, Algerian Arabic and French, speech recognition system was used to calculate recognition confidence score to determine the identity of the language in each segment (Amazouz et al., 2019).

Reviewing previous researchers' efforts for mixed speech processing, concludes that effort is still primitive and focused regionally. For Arabic language and its dialects, Algerian and Egyptian start focusing the task since 2018 by developing and testing mixed speech collections with French and English respectively. Availability of such resources will assist in researches in this type of speech.

2.6. Automatic Language Identification

In multilingual communities, either it is physical or virtual, people speak different languages, dialects and even worse mixed speech in their daily life communication. In such communities, determine the language identity conveys in speech utterance is essential frontend step before any further speech-enabled task, such as routing phone call to human operator fluent in identified language or to language dependent speech instant translator computer application. The ultimate goal of the automatic language identification, is to transfer most accurate human ability of identifying languages to the machine (Muthusamy, 1993, Marc A. Zissman, 2001, Haizhou Li, 2013). Language variations and dialects (Torres-Carrasquillo et al., 2002b), mixed languages speech (Wu et al., 2006), spoken only live languages with no orthographic system and linguistic rules (Simons and Fennig, 2017) generates more challenges of language identification task.

2.6.1. Language Identification Relative Information

Transferring human ability of language perception to the machine need investigation of language distinctive properties that human uses to determine or evaluate language identity conveyed in speech utterance (Ramus and Mehler, 1999, Navratil, 2001). The following four main broad categories were used throughout decades of domain researches:

Phonetic Inventory: Phoneme is smallest spoken unit that human articulation system produces. Phonetic inventory differs from language to other in terms of size of the set, consonants-vowels count and unique-shared phonemes. These inventory properties conclude that, even phonemes are shared among languages each language has a unique set of phonemes (Balleda et al., 2000).

Phonemes Co-occurrence: Each language has set of constraints govern the co-occurrence of phonemes called phonotactics. Even inventory set shared among languages the way phonemes structured and ordered differ. For example, د ز are shared phonemes between Arabic and Persian but constrained by their ordered in Arabic language (Zissman, 1996, Penagarikano et al., 2010, Penagarikano et al., 2011).

Expressive Properties: To express meaning, punctuation and sentence structure, speakers change their articulation system configuration to show what called prosody features of stress, duration, intonation and syllable (Ng et al., 2010, Martinez et al., 2012)

Linguistics Properties: Each language has phonological rules that govern word formation and syntactic rules followed to build a sentence (Adda-Decker et al., 2003, Soufifar et al., 2011).

2.6.2. General Form of Automatic Language Identification

Model creation and model testing are two main phases comprised the process of automatic language identification. Language dependent model is created in training phaser using speech samples from each language in the set, as illustrated in Figure 2.14, where n represents languages number in the set (Zissman, 1996) (Muthusamy, 1993).

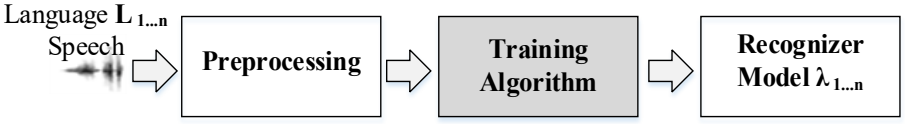


Figure 2.14: Language Identification training phase

In the testing phase, the task goes through same process of preprocessing and features extraction along with model created in training stage as illustrated in Figure 2.15 The selection of language identity decision is taken according to conditional probability as:

$$\hat{l} = arg_l \max p(l_{1..n}|\boldsymbol{\lambda}_{1..n}) \quad (2.12)$$

where: $\hat{l}$ target language, l_n *l*anguage set, $\boldsymbol{\lambda}_n$ *l*anguage models

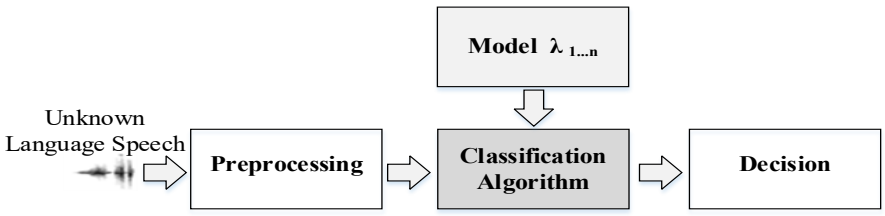


Figure 2.15: Language Identification testing phase

2.6.3. Automatic Language Identification Speech Corpora

Human speech environment is highly variable. Speech signal, even for same word in specific language, subject to effects of speaker (mode, age, gender and accent), surrounding environment (background noise) and recording equipment configuration and status. These variabilities make comparison of models developed in different environment not applicable and may give misleading results. Comprehensive development effort of common speech resources were held at Oregon Graduate Institute of Science and Technology (OGI -ST) in 1993 (Muthusamy, 1993), two speech corpus where developed, first one contains high quality speech for four languages (American English, Japanese, Mandarin Chinese and Tamil), chosen based on variations of native speakers in United States. The speech automatically segmented using neural network-based segmentation algorithm to vowels fricatives, stops, closures (silence or background noise), pre-vocalic sonorant, inter-vocalic sonorant and post-vocalic sonorant. In the second corpus, more realistic telephone speech collected for ten languages (English, Farsi (Persian), French, German, Korean, Japanese, Mandarin Chinese, Spanish, Tamil and Vietnamese) selected based on linguistics properties and availability of native speakers in United States (Muthusamy et al., 1992), the corpus then automatically segmented to previously mentioned seven broad phonetic transcription. These two speech corpora were globally available and extensively used

for development, comparison and evaluation of language identification models (NIST, 2017a, NIST, 2017b).

2.6.4. Automatic Language Identification Approaches

Different approaches and methods applied to the domain of automatic languages identification. These approaches extract and analyse speech attributes that conveys language discriminating information humans use to determine language identify. Approaches could be grouped into three broad categories as shown in the below sections.

2.6.4.1. Phonotactics Based Approaches

Phonotactics, rules that govern speech formation were used as a backend classifier in the automatic language identification model to specify identity of language. Phonemes and words set, phones co-occurrence, syllable structure and lexical information are examples of such rules. For decades, the approach gives high accurate performance, with shortcomings of needs a large amount of labeled speech data for each language in the set, linguistics experts to define rules and relatively long processing time to test rules against incoming speech utterance. The speech tokenizer, a front-end module to break down speech utterance into smallest units either it is frames, phonemes, syllables or words, is essential frontend part for language identification. Phonotactics approach as illustrated in Figure 2.16 and Figure 2.17, which shows models creation phase and model testing phase respectively. The performance of language identification as a front-end module affect the overall system accuracy (Matejka et al., September 2005).

Figure 2.16: Phonotactics approaches training phase

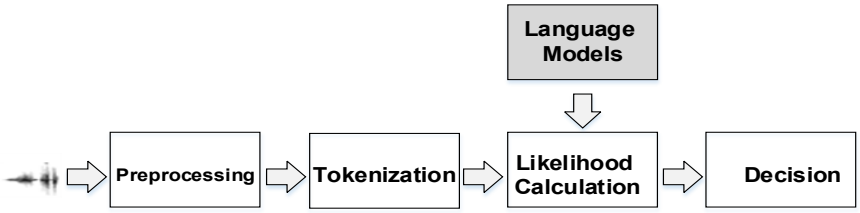


Figure 2.17: Phonotactics approaches testing phase

In the following sections, the backend Phonotactics classifier categorized according to the level or type of decoder output.

a. Sequence of Language Key Sounds

Inspired by the fact that each spoken language has a set of distinct sounds (key sounds), early attempt of languages classification for purpose of monitoring communication channels based on language key sounds sequence classification. Both, automatic and manually approach of identifying reference sounds were investigated. Automatic approach gave 64% classification accuracy for seven language, whereas human key sounds preparation approach degrades overall system automation with higher accuracy of 80% for five languages (Leonard and Doddington, August 1974, Leonard, March 1980). The success of Large Vocabulary Continuous Speech Recognition systems (LVCSR) encourage researcher to use more Phonotactics constraints at the different tokens level (Mendoza et al., 1996, Schultz et al., 1995, Schultz et al., 1996). Different algorithms with individual or combined source of information were examined. These studies conclude that using higher linguistics information improve model accuracy and raises the effects of approach drawbacks. Clustering mechanism, based on significant language sounds (key sounds) and sounds co-occurrence used for five Indian languages with VQ to avoid supervised training which is most challenging process in spoken languages identification. This method achieves promising result on utterance length between 100 and 150ms (Balleda et al., 2000).

b. Phone Based Phonotactics

Phones tokenizer followed by n-gram language model, (model that statistically computes co-occurrence probability of tokenizer output sequences), were compared in configuration of Gaussian mixture model (GMM) acoustic based classifier with no labeled data. A single-language phones tokenizer followed by n-gram language dependent model (PRLM), parallel PRLM; which uses multiple single-language phone recognizers, each trained in a different language in the set; and language- dependent parallel phones tokenizer along with its n-gram model(PPR) (Zissman, 1996). Different experiments, when applicable, were held for both 10- and 45-seconds utterance length. The comparison concludes that parallel PRLM obtains high performance with drawbacks of slow processing and needs for labeled data for each language in the set.

A Hybrid Neural networks and Viterbi algorithm phonemes tokenizer employing temporal pattern is used. The study emphasis dependency between ERR of tokenizer and final output, its concludes that less well-trained tokenizer is better than multiple with poor training (Matejka et al., September 2005). 5-gram language model following broad phonemes tokenizer achieves performance of 93.7% for phone set of 80 member for 6 seconds utterance length (Adda-Decker et al., 2003). Comparison of human perception and machine identification is conducted in the same environment, shows that for the short utterance, 1.5 – 2 seconds length, the performance of human and machine both are below the theoretical assumption.

Benefits from vector geometrics that measures similarity as a distance between two vectors, unified phone tokenizer output fed to language n-gram model. The language dependent n-gram model victoried token sequence based on the bag_of_sounds concept. This approach is evaluated with National Institute for Standard and Technology Language Recognition Evaluation (NIST LRE) 1996 and proven successful classification with EER of 14.9% (Li and Ma, 2005). To eliminate need for large amount of labeled data and linguistic experts for phonotactics approach, a general computationally efficient GMM tokenizer based on acoustic characteristics of speech signal followed by language model have been created. The computationally efficient tokenization step is easily expanded to new languages. In a subset of 12 languages from CALLFRIEND corpus (Linguistic Data Consortium, 1996), this model produces error rate of 17% (Torres-Carrasquillo et al., 2002a). Significant improvement achieved of this low-cost approach by incorporating speech signal temporal information (shifted-delta-cepstral SDC) (Torres-Carrasquillo et al., 2002b). This language identification

technique applied to dialect identification for dialects in Call Friend and Miami corpora. Accuracy of 13% and 30% ERR achieved of dialects in both corpora respectively (Torres-Carrasquillo et al., 2004).

Phone Selection by Elimination (PSE), where mutual information used to select best phones set from set of languages and those phones not selected either removed or substituted followed by language model gives 7.58% EER (Kumar et al., 2010) while target-oriented phone tokenizer (TOPT), where a phones' subset that best discriminates between target languages selected from whole recognizer's inventory. This approach gives 9.26% for 30 seconds length (Tong et al., 2009).

Phone co-occurrence at the frame level using cross-decoder that considered time aligned information along with frequency of occurrence model slightly improve performance of language identification of the Phonotactics approach (Penagarikano et al., 2010). Motivated by this result, with assumption of co-occurrence is language specific, approaches of phone n-gram co-occurrences and co-occurrences of phone n-gram improve baseline Phonotactics approach by 16% (Penagarikano et al., 2011). A JFA a front-end to i-vector for 3-gram counts language model with SVM backend shows slight improvement over baseline Phonotactics model which indicates higher order of n-gram models most probably gives further improvement with less computation cost (Soufifar et al., 2011).

c. Syllable based Phonotactics

Inspiring by the motivated result of preliminary experiment for eight languages, manually broad transcription (stop, fricative, vowel, silence) fed to Hidden Markov Model (HMM) to model sequential and co-occurrence properties of speech patterns (House and Neuburg, 1977), syllables segments for five languages representing two languages families achieved 80% classification accuracy (Li and Edwards, 1980), with real male read speech. The study shows that syllable is perfectly differentiate between two languages families. Automatic segmentation of speech signal based on fundamental frequency (F0) and temporal trajectory of short-term-energy output broadly categorized to Vowels, diphthongs, glides, schwa, stops, nasal, fricatives, and flaps. This frame by frame segment fed to language dependent trigram model of 12-CallFriend languages corpus. The trigram model had 24% ERR for 30 seconds utterance length. The study

concludes that prosodic information is significant in classifying some languages, such as Mandarin Chinese (Adami and Hermansky, 2003). With assumption that, even shared phones and co-occurrence spread over languages, sound duration is different based on language, context and speaker. Automatic normalized duration vector of UV (Unvoiced, Voiced) segments front-end for n-gram language model achieves 19.7% ERR on NIST LRE 2005 (Yin et al., 2009). With Prosodic Attribute Model (PAM), attempt is held to model language-specific co-occurrence of compact prosodic attributes.

Since single language dialects most probably share phonetic inventory and syllable structure and written script, syllable tokens fed to n-gram model along with Latent Semantic Analysis (LSA) to capture more phototactic constraints. For three Chinese dialects (Mandarin, Cantonese, Shanghai), a 99.23% classification accuracy were achieved (Lim et al., 2005).

2.6.4.2. Acoustic Based Approaches

In spite of employing higher linguistics information for automatic languages recognition achieved most identification accuracy; it is computation complexity and linguistics experts' dependency force researchers to find a language dependent information conveyed into speech waveform that human with linguistics knowledge or not uses to determine language identity (Combrinck and Botha, 1997, Cimarusti and Ives, 1982). Based on infant ability to discriminates between language with no previous linguistics knowledge; French native speakers discriminates well between two different unknown languages having different rhythms using rhythm prosodic property (Ramus and Mehler, 1999).

The major drawback of linguistic based approach is the difficulties of adding new languages to the system without needs of linguistic experts and more training data. Acoustic approach is seeking solution of this problem. Rare and detectable languages features are significant regional discriminant languages properties. Features of occurrence of nasalized vowels, labial-velar stops and of retroflex consonants were examined against The University of California, Los Angeles (UCLA) Phonological Segment Inventory Database (PSID) of 451 languages representing all languages families with at least one language for each family(Ian Maddieson and Precoda, 1984).

These three features together eliminates set members to only three languages which represents 0.7% of languages from the corpus (Hombert and Maddieson, 1999).

Figure 2.18: Acoustic Model approach training phase

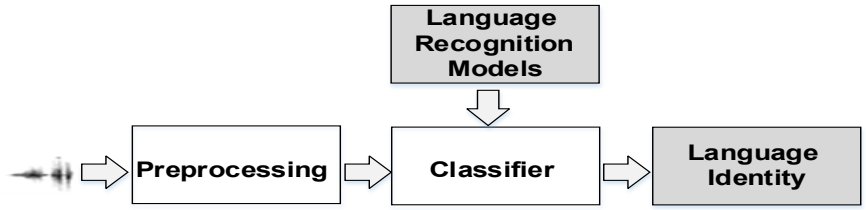


Figure 2.19: Acoustic Model approach testing phase

2.6.4.3. Spectral Information Based

This approach is only based on spectral attributes contained in speech signal for determining the identity of language in speech utterance. Acoustic features, including Autocorrelation, cepstral, filter and formant frequencies, etc. were extracted using LPC (Cimarusti and Ives, 1982, Ives, may 1986) and PLP (Muthusamy et al., 1993, Chu-Carroll and Carpenter, 1999) features extraction techniques. Based on the fact that human can makes reasonable judgement for unknow language using only acoustic features; a novel approach to determine a language specific information significant for language recognition (Perceptually Significant Regions (PSR) in speech utterance) were introduced (Braun and Levkowitz, 1998). Recurrent neural network trained with PSR achieves 9% performance enhancement over other training approach.

With SVM classifier, i-vector technique outperform direct JFA state-of-the-art model (Martınez et al., 2011), acoustic model tested on NIST LRE 2009 shows promising results (Dehak et al., 2011). Double reduction of SDC acoustic speech feature with deep neural network bottleneck feature (feature from layer with few hidden nodes) followed by i-vector representation shows significant improvement with low computation cost for LRE2009 tests of 30, 10 and 3 seconds, specifically for short duration, which achieves 9.71% EER (Song et al., 2013). Inspired by the success of

artificial neural network in acoustic modeling, feed forward neural network with deep learning (DNN) with PLP outperforms i-vector approach and achieves 70% improvement on 3s utterance length, specifically when large amount of training data is available. (since i-vector saturated DNN stay learning) (Lopez-Moreno et al., 2014). Deep learning for speaker and language recognition good performance encourage researchers to use as a front-end (Richardson et al., 2015) or as i-vector extractor for building generalized language identification model with attention to within phone transitions (Song et al., 2015).

2.6.4.4. Speech Token based Identification

Inspired by different languages have different phone sets, phonemes segmentation followed by maximum likelihood scoring for model of HMM and neural network were exploit for this, method achieving comparable (Lamel and Gauvain, 1993, Lamel and Gauvain, 1994, Muthusamy et al., 1992, Muthusamy et al., 1993).

Inspired by phoneme superset used in work (Muthusamy et al., 1993),a mono-phoneme set of each language in the set which convey language discriminant information were used instead (Yan and Barnard, 1995). The study shows that using mon-phoneme superset reduces feature space with insignificant performance loss, vowels articulation distinction were also utilized (Pellegrino and André-Obrecht, 2000).

Vector space modeling (VSM); the dominant technique in information retrieval (IR) research, were used for language recognition employing unsupervised approach. VSM discriminately measures the similarity between test (query) vector and target language vector, based on distance between them (Li et al., 2007, Tong et al., 2009, Siniscalchi et al., 2009).

2.6.4.5. Prosodic Information Based

Prosody is a study of tune and rhythm and how they contribute in speech meaning. It is characterized by vocal pitch (fundamental frequency), loudness (acoustic intensity) and rhythm (phoneme and syllable duration), it is playing significant roles of human process of identifying spoken languages (Ramus and Mehler, 1999). Studies in (Farinas and Pellegrino, 2001, Rouas et al., 2003, Rouas et al., 2005) assumes languages could be grouped into rhythmic classes. (Ng et al., 2010) utilize difference of sound

duration between languages to determine their identities, whereas (Mohanty and Swain, 2010) study based on jitter (variability of F0) and shimmer (amplitude of vibration) as new information source. For the prosodic GMM features (rhythm, stress and intonation) is evaluated for i-vector reduced feature space. The result shows fusion i-vector prosodic model with new techniques gives comparable performance to acoustic Phonotactics model (Martinez et al., 2012).

2.6.5. Evaluation and Comparison Studies

Comparison and evaluation of automatic languages recognition research output is impractical in different environment due to speech variability. In 1996 NIST begins publish common evaluation environment including speech corpus and test plan (NIST, 2017a). Since then evaluation and competition held every two years, for year 2017 (LRE17), eighth Languages recognition evaluation plan is for language detection for 5 languages clusters (Arabic, Chinese, English, Slavic and Iberian) with 14 languages (NIST, 2017b).

Automatic recognition of language in speech utterance for languages from same families or that shares many sounds are confusable and add another complexity dimension of the task (M. Osman Eltayeb and Mohammed Elhafiz Mustafa, 2013). For NIST LRE 2009, tasks includes language identification, target language detection and discriminate between confusable language pairs comparison (Torres-Carrasquillo et al., 2010b, Torres-Carrasquillo et al., 2010a).

The task of Albayzin2012 Language Recognition Evaluation (LRE), effort made by the Spanish/Portuguese community for benchmarking language recognition technology, is to output likelihood scores for the YouTube extracted audio for each target languages (English, Portuguese, Basque, Catalan, Galician and Spanish) along with score for out-of-set languages (French, German, Greek and Italian) that have no training data. State-of-the-art total variability (i-vector) model mostly used for participants submissions (Rodriguez-Fuentes et al., 2013).

Pear in mind issues of short utterance recognition and linguistics processing content requirements for language identification task, comparison studies were held for 3 seconds and shorter utilizing different techniques for shared corpus e.g. NIST

LRE2009 (Gonzalez-Dominguez et al., 2015, Lozano-Diez et al., 2015, Zazo et al., 2016, Geng et al., 2016).

2.6.6. Challenges of Automatic Language Identification

In spite of great achievements obtained in this demanding front-end module in multilingual communities' communication as detailed in previous section; it is performance is far from human based line system. Apparently, some domain performance challenges arise from speaker and speech environment and others from linguistics structure of languages.

2.6.6.1. Identifying Unseen Languages

In spite of good performance achieved of phonotactics based language recognition approach, needs for linguistic experts and an huge amount of labeled training data slowdown its progress due to model limitation in term of adding new languages to the model and classifying unseen language of world with a lot of language and dialects, some of them are spoken only.

2.6.6.2. Recognition Time

Fast and accurate automation of language recognition are ultimate goals of the domain, but it is still far away comparable to human performance, the tradeoff between recognition accuracy and recognition time is hot research issue. Some real time speech processing system need recognition time comparable to human performance such as instant translation bearing in mind the task is just pre-process of translation main task, while others systems concentrates on system accuracy such as speech biometrics used for access control.

2.6.6.3. Dialects and Accents Variations

Most challenging issue in the domain of speech processing in general and language recognition specifically is language variations (dialects) and speakers' accents difference. For accents variation techniques of total variability that separates language attributes form channel attributes reduce its effects in performance. Unseen language

dialects greatly affect the performance of the system. To some extent, this challenge similar to problem of adding new languages to the system.

2.6.6.4. Multilingual and Mixed Speech Utterance

Bilingualism, heavily affected community daily life communication. They represents half population of the world (Ramus and Mehler, 1999). This fact adds a practical complexity dimension of speech processing enabled system. Bilinguals tends use more than one language in a single sentence to express their thought. In this type od speech, languages used are not previously known. For mixed speech, the study shows that performance of speech recognition system in this environment greatly enhanced by perfect language identification using less than 3 seconds utterance length before speech recognition process start (Ma et al., 2002). A practical use of spoken language recognition for mixed speech mode (Wu et al., 2006), is applied as a front end for multilingual speech recognition system of English and Mandarin languages.

For decades, researchers investigate varies approaches and techniques for ultimate goals to fast and accurately identifying language in speech utterance based on human ability of such task. Variabilities of speech and effects of its environment complicates the process of extracting most relevant language information. Develop efficient and best fit training algorithms with minimum needs for linguistics experts still challenging task. Language dialects, mixed languages speech and languages with no linguistics rules add other dimensions of task complexity. Exists of standard multilingual corpora, greatly enhanced model performance by offer single evaluation and comparison environment of varies techniques.

CHAPTER III

MIXED SPEECH AND LANGUAGE IDENTIFICATION MODEL

3.1. Introduction

Speech recognition system is in practical use nowadays and is commercially available e.g. medical record transcription. In multilingual communities, such as Sudan, a hybrid language composed of mixed local Arabic dialect and English are in daily use. Bear in mind, many local languages are in mixed use or interchangeably with Arabic in daily life and work domains. This led to need of generalized speech processing solution in such communication environment without extra model training for adding new language to the system.

Mixed speech formed type of hybrid language that is difficult to classify as any of languages participating in utterance production, since it is linguistically different of each. The researches done till now is not much as detailed in sections 2.5.6 and 2.6.6.4 for mixed speech recognition and language identification respectively. This is due to regionality nature of the problem and it is not globally addressed by researchers worldwide. Mixed communication occurs in informal communication and rarely written, so resources availability is the problem facing progress of research. This is needed to train, test, evaluation and studies performance comparison. Since most speech processing tasks are language dependent, resource available for one language may not be sufficient for other languages, in particular in mixed speech mode when native pronunciation is dominant.

The biggest problem hindering development and prevents the dissemination of solutions in linguistic computation, and in particular mixed speech processing, is the absence of a single solution or model composed from many disciplines share their experience and knowledge to promote its efficiency. This is due to speech variability nature as detailed in chapter 2 of this thesis.

This research proposes a universal solution for speech recognition in monolingual, multilingual and hybrid (mixed) mode based on the richness of Arabic language phonetics representation that could be exploited and extended to represents phonemes and sounds for other languages, specifically for languages with no orthographic systems. Two languages are participating in the proposed model. In the proposed model, Sudanese Arabic and English both pronounced by Sudanese Arabic native speakers.

This research is intended to establish computerized based speech recognizer solution for hybrid language created by a bilingual Sudanese Arabic. At this stage, the research concentrates on language version that is mixed between Sudanese Arabic and English languages. These two languages are nationally used throughout the Sudan and acceptable in the societies where it is used. They are heavily used in communication between social media friends and followers, universities classmates and in work domains where technical phrases are shared and frequent.

Bear in mind, the fact that mixed speech may be more complicated, when a single sentence contains more than two languages or dialects. Proposed model has a flexible capability of adding new dialects and languages in the future, specifically for Sudanese native without model retraining and reasonable effort. Mixed language is hybrid and not defined elsewhere, developed solution establishes definition layer to outside world through module called Interface for Further Processing (IFP).

3.2. General Form of the Proposed Model

Figure 3.1 shows proposed model of Sudanese Arabic – English mixed speech recognition system, the diagram is an abstract view using standard flowchart notation to show processes and their links. Each process will be detailed later in this chapter. Proposed approach adapts traditional model of two processing phases, training and testing, for the sake of providing generalized solution, a layer that provides subsequent processes the necessary information to complete their jobs is designed, for example, language identity is essential piece of information for translation between languages.

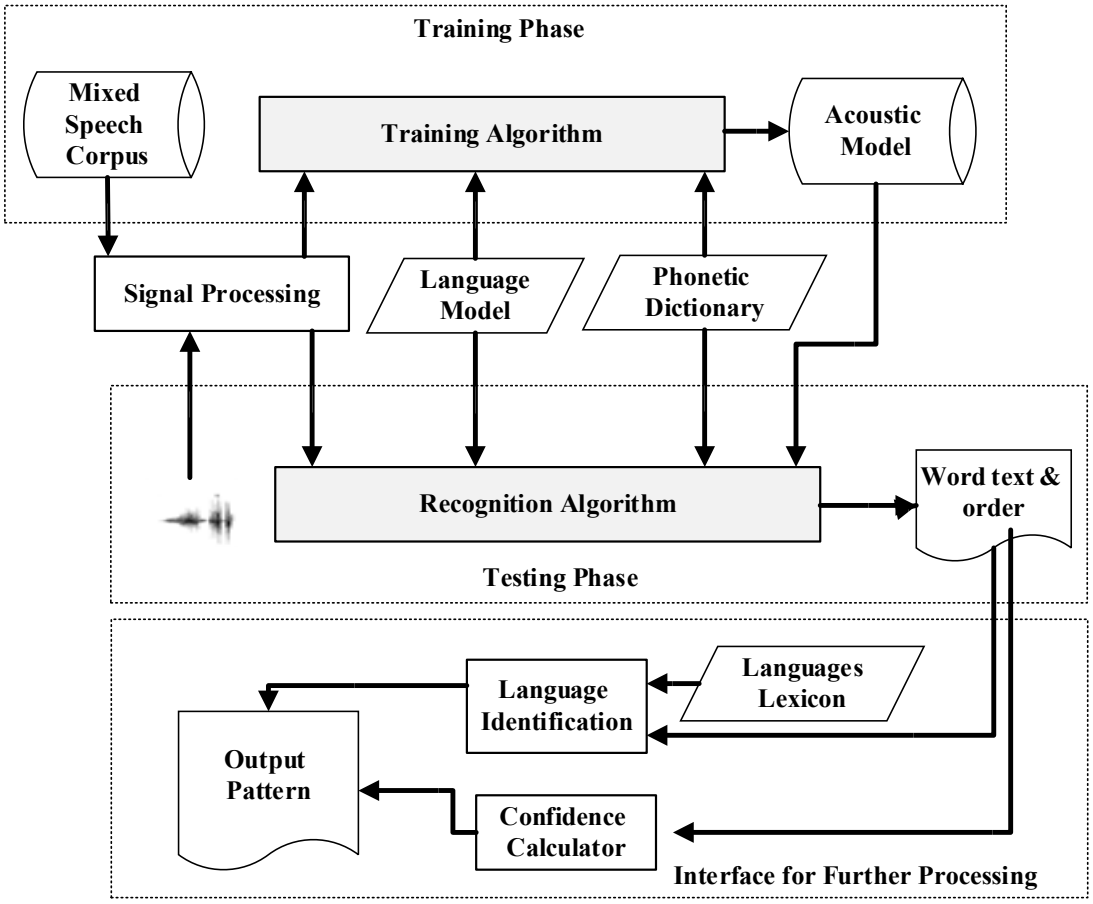


Figure 3.1: Proposed model of mixed speech and language identification

In the following sections, each process and resource utilization and creation will be described in details.

3.3. Sudanese Arabic – English Mixed Speech Corpus

Creation of Sudanese Arabic – English mixed speech corpus is motivated by the lack of resources that support research and studies evaluation in the domain, to the best of my knowledge, only two collections are published so far, French - Algerian Arabic corpus (Amazouz et al., 2018) which have a French as the main language and Algerian Arabic as second language and Egyptian Arabic - English (Hamed et al., 2018) for technical terminologies. Both collections are not publicly available for research or commercial use, with attention that Arabic content in such corpus does not serve our work for Sudanese Arabic due to Sudanese accent difference from both Algerian and Egyptian e.g. غ and ق are interchangeably pronounced in Sudanese pronunciation, ماعارف and ماعرفش both means "I do not know", the first in Sudanese and the latter in

Egyptian, the worst comes when two words used for the same meaning e.g. **شديد** and **جامد** both means strong in Sudanese and Egyptian respectively, whereas the latter is means "solid" in Sudanese Arabic dialect. Differences between Sudanese and Algerian Arabic are very apparent e.g. word **بكري** is a proper noun in Sudanese whereas it means earlier in Algerian dialect. Daily life Sudanese Arabic content with English of native Sudanese speakers gives this corpus its uniqueness compared to other similar corpora.

Figure 3.2 shows the main processes followed to build Sudanese Arabic – English mixed speech corpus, which start by collecting most frequent mixed sentences in daily life and end up with complete dataset composed of speech recording, transcription files, phonetic dictionary and language model.

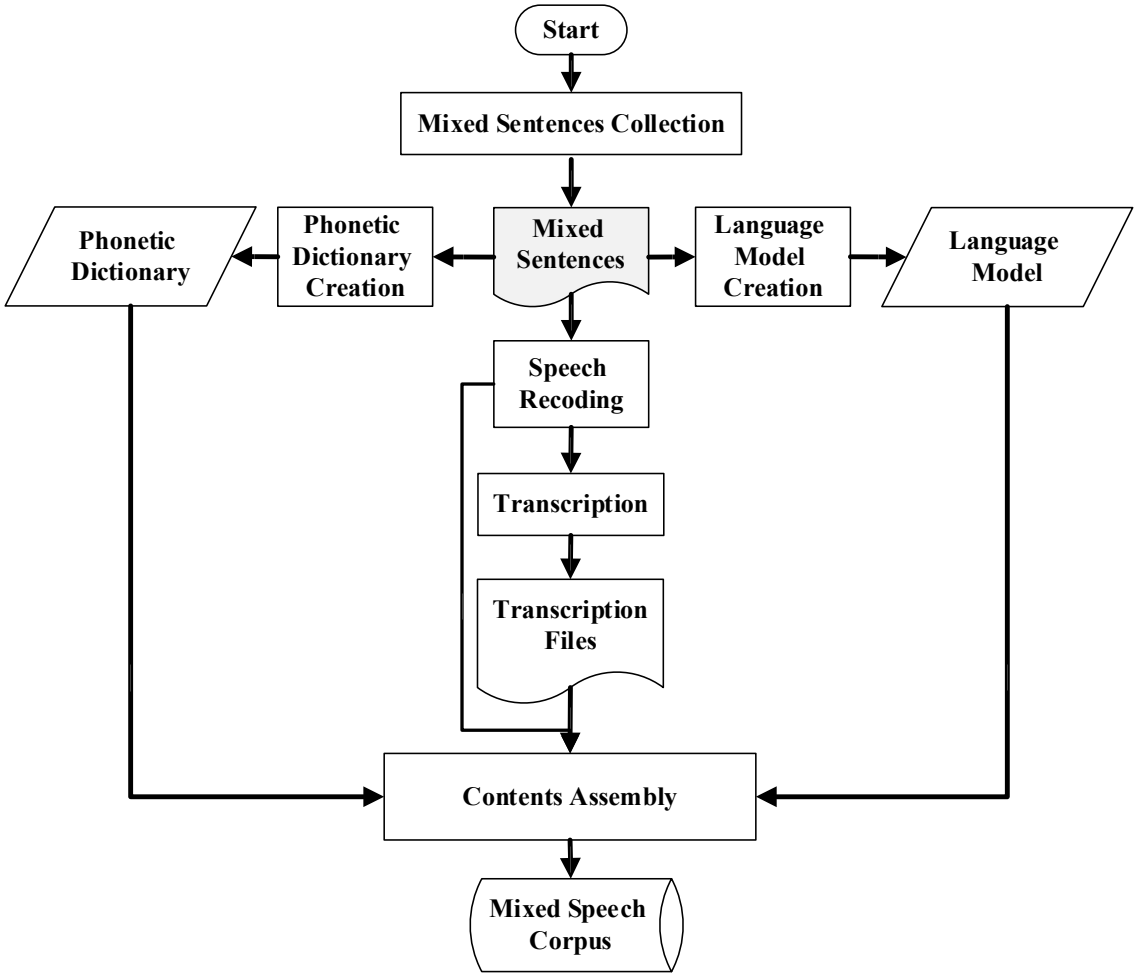


Figure 3.2: Mixed speech corpus generation

The corpus is composed of two languages, Sudanese Arabic, the native language of Sudanese people and English as international language. The existence of a lot of ethnic groups in Sudan encouraging creating such corpus, most probably each group has its own language or dialect, some of them are considered under-resourced languages because they are spoken only without a written system or alphabets, mixed in daily life conversation with national and formal Arabic language. Sudanese Arabic – English mixed speech communication is daily life and societies acceptable phenomenon. It is common between universities and international students, between family members living in different countries and in technical domains such as Medical and Information and Communication technology (ICT) work areas, language of education in many universities and international schools and it is a primary education mandatory class, even occurred in mixed speech communication of people who do not speaks English languages, especially for new technical terminologies such as mobile, 4G and 3G.

The following sections describe mixed speech corpus of Sudanese Arabic and English languages in details, mixed sentences collection method and strategy will be explained, speech recording specification and transcription techniques will be discussed. Used phonemes set along with language lexicon and phonetic dictionary setup are investigated.

3.3.1. Collection of Daily Life Mixed Sentences

In todays' world, people keep in touch with their social media application in daily basis for social and business conversations, most applications offer push to talk verbal communication in reasonable cost and effective in terms of delivering exact message, freehand usage and eliminate time required for writing. These applications open world to everyone have access to this global virtual community, they ease communication between people either they know each other's', speaking same languages and from same cultural background or not. In such loose boarder communities, occurrence of mixed speech is potential and people tend to use more than one language in single conversation to express their ideas accurate and clear.

For the purpose of collecting the most frequently used mixed sentences in Sudanese daily life, a campaign was launched through social media applications, in particular WhatsApp, Facebook and Short Message (SMS), direct interview and personnel community observation. Collected sentences were filtered to 201 distinct

sentences. The campaign terminated when the ratio of sentences repetition from participants is high. Then, each sentence is uniquely identified by distinct numerical identifier, along with its type of switching. Switching type is based on position of embedded language in the sentence. Symbol is used S for start inter-sentential, E for end inter-sentential and M for intra-sentential). As illustrated in table 4.1, sentence should be read from right to left, with attention that English phrases should be read left to right. The direction complication is clear in sentence with identifier 00003 in table 4.1, this complication is not apparent in speech because the sentence pronounced in order.

Table 3.1: mixed speech sentences formatting

Identifier	Sentence	Switch Type
00001	ماتاخد الموضوع personal	E
00003	سافر Let us say	S
…	.....	
00099	اعملها double لو سمحت	M

The collection contains mixed sentences from different domains, such as social communication, technical terminologies, proper names, etc... Well representation of participants in terms of gender, age, ethnic group and work domain is considered. The collection is containing all types of code switching, either Inter-sentential or Intra-sentential is considered.

3.3.2. Speech Recording

Speech recording is a time consuming and challenging task in the corpus creation process. Office environment was chosen as comparable environment to reality where the model will be tested. It gives average variabilities effects between street noise and quiet room that may occur in real daily life communication regarding surrounding noise. Air condition noise, background noise, telephone ring, etc. create additional noise that may exist in recording.

The recording process illustrated in Figure 3.3 end up of with audio file for each sentence for each speaker.

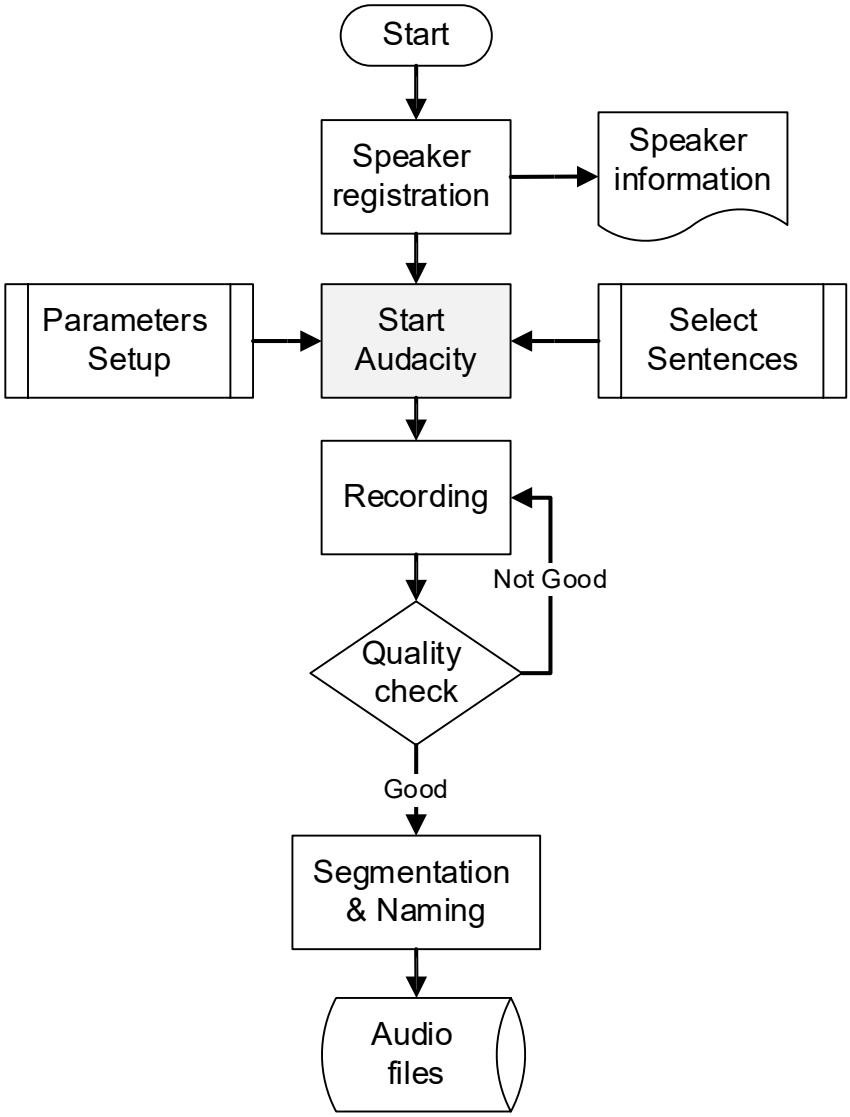


Figure 3.3: Speech recording process

Audacity version 2.1.0 software[1] for audio processing is used for speech recording, with parameters described in Table 3.2. Each speaker is asked to read ordered part from mixed sentences list. Each recording is given a full name representing information of sentence identifier and speaker identity for link with his profile that contains name, gender, age and languages speaking. Speakers selection considers distribution of age, gender, education, ethnic group and education system and level.

[1] Available online: https://www.audacityteam.org

Table 3.2: Recording parameters

Parameter	Value
Environment	Office
Software	Audacity 2.1.0
Microphone	Headphone
Sample rate	16000 Hz
Resolution	16-bit
Channel	Mono
Format	WAV

3.3.3. Speech Transcription

Transcription is the process of taking an audio file and write its content in alphabet system of chosen language, in other words is the process of mapping speech content to text. Audio transcription is essential process that tells training algorithm what audio file contains, it is boring and time-consuming task needs a lot of manual efforts to transcribe sufficient number of audios for each recording in the collection for training purpose.

A most famous and accurate open source software for speech processing is called SPHINX developed at Carnegie Melon University (CMU) was chosen to build AM for mixed speech task of this research (Matarneh et al., 2017), SPHINX training algorithm accept transcription format of speech content delimited at each end by silence tag accompanied by associated audio file name as follows:

<S> text of speech content </S) (audio filename)

To eliminates time-consuming manual efforts needed for transcription and formatting in required training format, the process was automated by developing our own audio transcription and formatting software tool called **SUD_audio_transcription** using C programing language. The pseudo code of this software tool is provided in Figure 3.5, which implement flowchart sequence illustrated in Figure 3.4 for

transcription and formatting. This tool will be publicly released for reading speech transcription as contribution of SUST in the domain of speech processing.

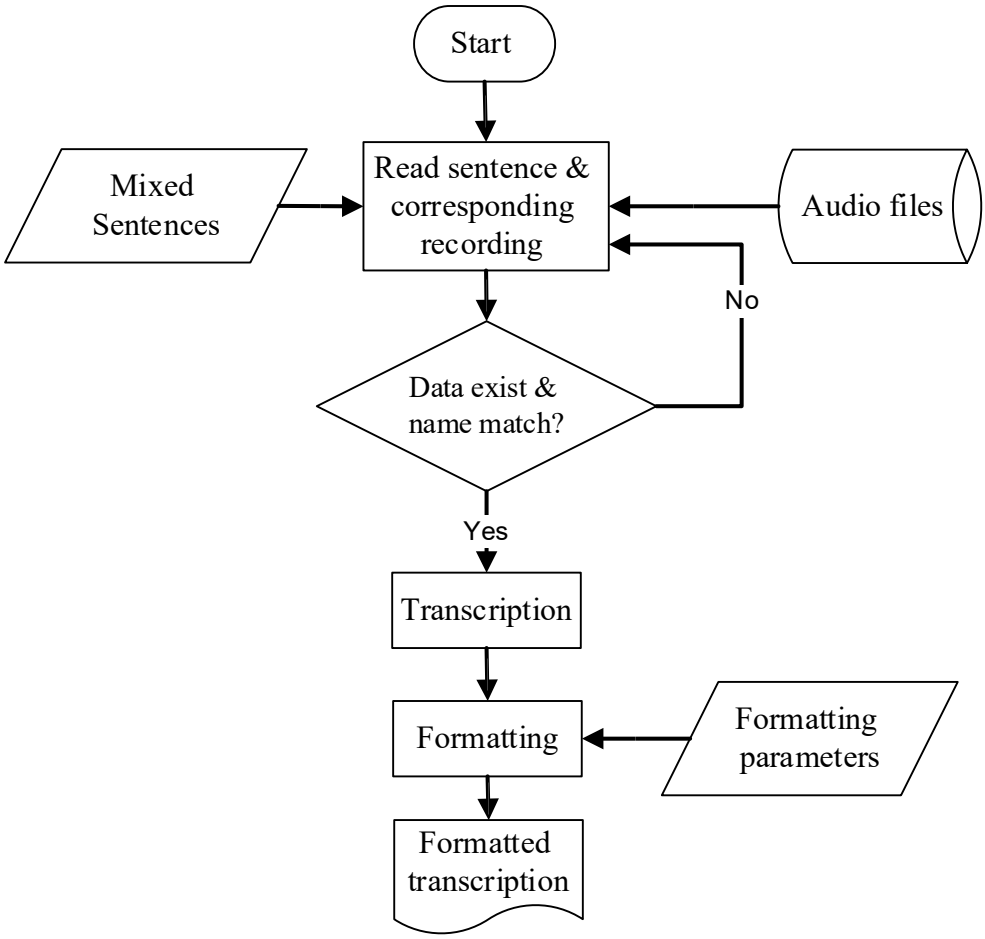


Figure 3.4: Audio transcription and training strings formatting

Table 3.3 shows the part of transcription process output according to require format of the purpose of this research, each speech file content written in a separate line contains from left to right in order start tag, content of the file, end tag and name of corresponding speech file.

Table 3.3: Formatted audio transcription examples

<s> مشيت ليو already </s> (000003_spk105)
<s> I think لغي الموضوع </s> (000004_spk013)
……..
<s> عندنا presentation </s> (000005_spk105)
<s> انا في meeting </s> (000006_spk004)

```
{
    1. Open mixed sentences file for read
    2. Open transcription file for write
    3. Loop
       {
          a. Read sentence
          b. Extract sentence identifier
          c. Check the existence of corresponding audio file
                  If found then
                  {
                     1. Format the sentence
                     2. Write formatted text to transcription file
                  }
       } // end loop
    4. Close files
} // end program
```

Figure 3.5: Pseudo code for audio transcription and formatting

Automatic transcription process depends on audio files naming convention, where each recorded audio file name is composed of speaker unique identifier with it is corresponding sentence identifier separated by underscore symbol e.g. 000014_spk029 which is a pointer to file contains speech of sentence number 14 for speaker 29.

The tool loop throughout the file system directory where recordings files are stored, takes file name, extracts associated speaker and sentence identifiers, sentence identifier is used to search mixed sentences text collection list for associated sentence. Text and recoding file name with help of formatting specification for CMU training algorithm entered formatting tool to output training string for each recoding that match required format and specification by training algorithm as illustrated in table 4.3.

Table 3.4: Audio transcription and training format

<**S**> وين ال cheque </S> (filename_00001)
<**S**> قون double kick <**/S**> (filename_00002)
<**S**> ..... <**/S**> (…….)
<**S**>اعملها double سمحت لو<**/S**> (filename_00099)

3.3.4. Mixed Phonetic Dictionary Generation

Each language has specific set of sounds, may be represented by a single equivalent written symbol or by combination of more than one symbol. For languages with written systems, most phonemes are represented by equivalent alphabets (graphemes), for those languages with no orthographic system, IPA standard system are used to represent its sounds and approximates sounds not exist in standard.

Phonetic dictionary is defined as mapping sound to written symbol, in its simple format is a list of two columns, first column represents the word written in orthographic units (letters) for specific language, and the latter is same word written in equivalent chosen written symbols. Phonetic dictionary is essential component in the process of modeling speech, were each phoneme is represented in AM according to its signal features with consideration of speech variability factors such as speaker, environment effects and speech context. This model is then used at recognition time to label output recognized symbols (phonemes) as understandable text by mapping back phoneme to letters utilizing this phonetic dictionary.

For the task of this mixed speech processing of two different languages with different sets of sounds and orthographic symbols, special comprehensive phonetic symbols were proposed based on Sudanese Arabic language pronunciation. That could easily be extended and generalized for new other languages. The following subsection describes the process of building our own universal mixed phonetic dictionary.

3.3.4.1. Hybrid Language Phonetic Symbols

In spite of the fact that Arabic is official language of the Sudan, many factors affect its pronunciation, which result in deviation of some sounds from the original MSA, this result in special regional Sudanese Arabic dialects. These factors include local languages effects, tribes' interferences within the Sudan and neighboring countries and languages of different education systems. These collective parameters generate a distinct version of Arabic language we named it for the purpose of this study Sudanese Arabic, which is originally based on Khartoum Arabic dialect. Of course, English sounds are also affected by native language.

The International Phonetic Alphabets (IPA) (Ladefoged and Halle, 1988, Association and Staff, 1999), was introduced in year 1820, and updated regularly to represent the sounds of the world languages. It is categorized based on the standard articulation procedures of speech production process. For Arabic language, MSA which contains 28 consonant sounds, 3 long vowels and 3 short vowels is considered in IPA without emphatic sounds **(ص، ض، ط، ظ)** and **ع**, it should be noted that MSA is only officially used now days.

Existing representation of Arabic phonemes are not sufficient in multilingual and mixed speech environment, because they are targeting transcription of Arabic language only. It is neglecting nunation and represents Arabic sounds using English graphemes, substitute some English grapheme for Arabic sounds not present in IPA, e.g. (Sawalha et al., 2014) represents Arabic phone **ق** by English sound Q, which is needed to represents itself in multilingual communities.

For the aims to produce a universal solution for mixed speech processing based on Sudanese Arabic language, taking into account the failure of the IPA to represent Arabic sounds and the requirements for the IFP that every orthographic symbol should phonetically be represented for further speech processing tasks, such as speech to text synthesis, a new phonetic comprehensive symbols set was proposed. It includes symbols for every Arabic orthographic and diacritic symbol. Roman characters are used to represent these sounds avoiding substitute distinct Arabic sounds with English grapheme. For example, sound **ق** is represented by **Q-** in our set reserving English phone **Q** as it is. Avoiding using a Unicode symbol is considered.

Appendix B (Hybrid language phonetic symbols) shows the full list that contains 47 symbols categorized in 8 groups. The groups include 3 consonant categories

which are normal consonants sounds, the sound of "ع" which is distinct for Arabic language and four emphatic sounds "ص", "ض", "ط", "ظ" that IPA system does not include. Two groups contain long and short vowels, group for nunation and other special Arabic symbols, plus ***shaddah*** group that represents sound stress and emphasis, table 3.5 shows part of the phone set symbols and their categories.

Table 3.5: Part of phone set symbols and categories

Orthographic Symbol	Phonetic Symbol	Category
ت	T	Consonants
ج	J	
ح	H-	
ا	A:	Long Vowels
و	U:	
ٌ	UN	Nunation
ٍ	IN	

3.3.4.2. Diacritics and Word List Preparation

Towards the creation of mixed phonetic dictionary, each sentence in mixed collection is divided into its words, English words were transliterated into Arabic (written in Arabic script), all words now rewritten in Arabic script based on Sudanese pronunciation either for Arabic and English. The words list was refined and distinctly filtered by removing repetition based on its sounds not alphabets, complete set then alphabetically ordered and manually diacriticked using short vowels and nunation symbols described in section 3.3.4.1 to exactly represents its composed sounds.

Figure 3.6 illustrates the process of produce distinct diacritic words list of collection as pronounced by Sudanese Arabic native speakers. The process target maintains speaker accent regardless of the original pronunciation of the phoneme. A single word may have multiple pronunciation and multiple dictionary entries based on its speakers' actual

articulation process and context; this is handled by numbering each instant of words e.g. Arabic word قلم it will be قَلَمْ or قَلْمْ which mean pen and prune respectively.

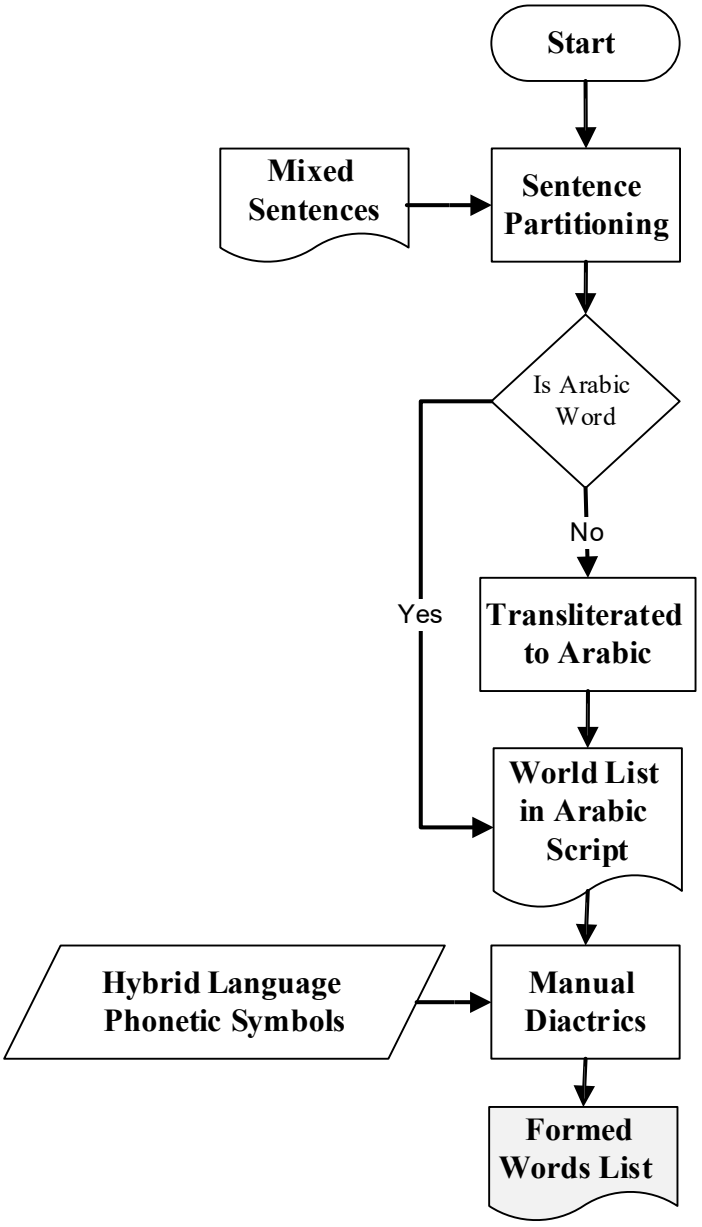


Figure 3.6: Formed words list generation

Table 3.6 shows Arabized and diacriticked version of the mixed sentence" جيب ***file*** من الدولاب", which means "get a file from cupboard".

Table 3.6: Arabized and diacriticked sentence example

#	Original word	Language	Arabized & Diacriticked word
1	جيب	Arabic	جِيبْ
2	file	English	فَايلْ
3	من	Arabic	مِنْ
4	الدولاب	Arabic	الدُولَابْ

3.3.4.3. Build Mixed Phonetic Dictionary

Phonetic dictionary is a two columns list that maps letters (alphabets) used for words orthographic of specific language to corresponding phonemes used to represent its sounds as illustrated in Table 3.7. A software tool was developed to assist building a hybrid phonetic dictionary that contains words of both Arabic and English for the purpose of this research. Arabic phonetic set described in section 3.3.4.1 along with diacritic words list created in section 3.3.4.2 were utilized by the tool to generate entry for each word in the list, this hybrid dictionary is multi-lines document contains word without diacritics in the first column along with associated phonetic symbol of diacriticked version of the same word in the row. Figure 3.7 illustrates input parameters and sequence of process to generate dictionary. Figure 3.8 shows pseudo code for developed tool[2] to generate mixed phonetic dictionary using PERL scripting language.

Table 3.7: Part of phonetic dictionary

Word in letters	Word in phonetic
مشيت	M AE SH Y T
ليهو	L IY H W
already	E UW R AA D IY
already (1)	E L UW R AA D IY

[2] Adapted from freely available source code: https://www.youtube.com/watch?v=3XmgaQltpGw

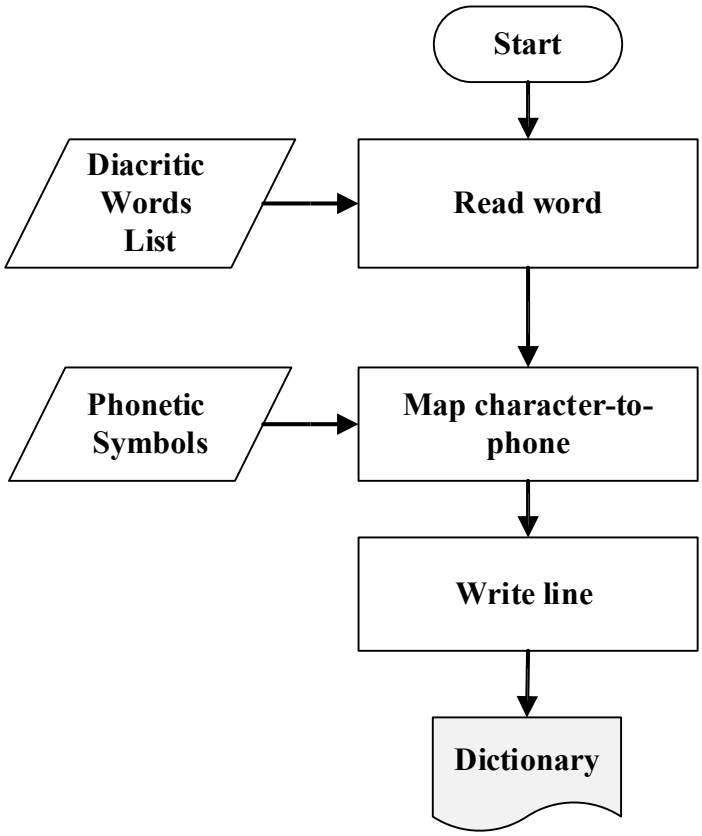


Figure 3.7: Phonetic dictionary generation

Table 3.7 shows part of generated multilingual phonetic dictionary of mixed Arabic and English languages of the sentence "***مشيت ليهو already***", which means “I have already gone to him”, that is composed of three words, two Arabic words and one in English language. The word is placed in the dictionary as it is in original language. Note that the single alphabet may represented by one or more symbols. English word **already** hast two entries in the dictionary depending on it is pronunciation difference between Sudanese speakers and those whose mother tongue is English.

{

1. Open diacritic word list file for read
2. Open phonemes set file for read
3. Open dictionary file for write
4. **Loop**
    a. Read word *w* from word list
        Loop
        {
            1. Read character w_i form the word
            2. Get associated phonetic symbol from phonemes file

        }// **end loop**
5. Write *w* and w_n to dictionary file

} // **end loop**

}

Figure 3.8: Pseudo code used for dictionary generation

To this end, each distinct word in the mixed sentence list have at least one entry in the dictionary accompanied by its phonetic symbols. Table 4.5 shows that our universal dictionary maintains alphabets of original word. The dictionary has the ability to repeat a single word as many times according to phonetic symbols based on pronunciation difference, this clear for word ***already*** which has two different entries for its different pronunciation.

3.3.5. Language Model

Language model is statistical representation used to select most probable next word that may follow current word in the sentence when more than one option for the next word is possible. For this research 3-gram language model is created using **SRILM**[3] open source tool (Stolcke et al., 2011), exploiting mixed sentences collection

[3] SRILM – The SRI Language Modeling Toolkit: http://www.speech.sri.com/projects/srilm

described in section 3.3.1 as example to calculate these probabilities. SRILM tool requires specific sentence format as illustrated in Table 3.8, each sentence occupies separate line and delimited by SIL tag at each end.

Table 3.8: Text format for language model generation tool

<S> cheque وين ال </S>
<S> double kick قون </S>
<S> ..... </S>
<S> لو سمحت double اعملها </S>

For the purpose of this research and contribution offers to other researchers to build their language model from the existing text collection and examples, a software tool in Figure 3.9 was developed using C programing language and used produce formatted list as illustrated in Table 3.8.

```
{
    1. Open mixed sentences file for read
    2. Open LM formatted text file for write
    3. Loop
       {
          a. Read sentence from mixed sentences file
          b. Delimit sentence by SIL tag at each end
          c. Write sentence to formatted text file

          }// end loop

    4. Close files

    }
```

Figure 3.9: Pseudo code used to prepare examples for language modeling

Language model maintains statistical probability for each word based on its context, previous set of words in the utterance, used to differentiate between words having similar sounds utilizing probability of current word following previous one.

Probability distribution represents occurrence probability of the word in the context calculated from whole provided to SRILM tool.

The order of language model (n-gram) is defined by how many previous words considered in calculating cooccurrence probability. The n-gram model is suitable for this research task since its focus in predicting the current word in respect to $n-1$ preceding words as illustrated in Equation 3.1.

$$P(w_i) = P(w_i|w_{i-n}, \dots, w_{i-1}) \quad (3.1)$$

Subjects to $\sum w_i = 1$

For the whole sentence comparison, the probability product for words in the sentence is taken as illustrated in Equation 3.2.

$$P(w_1, \dots \dots, w_k) = \prod_{i=1}^{k} P(w_i|(w_{i-1}, \dots, w_{i-n}) \quad (3.2)$$

$P(w_i)$ calculated at training stage using as much as possible sentences to represent a priori probability of each word in the language, that guides and constraints search among different words possibilities during recognition. For Arabic – English code switching, mixed language model trained with SRILM using mixed sentences collection described in section 3.3.1 applying Equations 3.1 and 3.2.

Figure 3.10 shows sequences of processes and resources needed to generate mixed language model for Sudanese daily life collection, the process ends up with database contains occurrence and cooccurrence probabilities of each word in the collection, preceding by one word and two words in the set, table 3.9 illustrates model for the word الفصل.

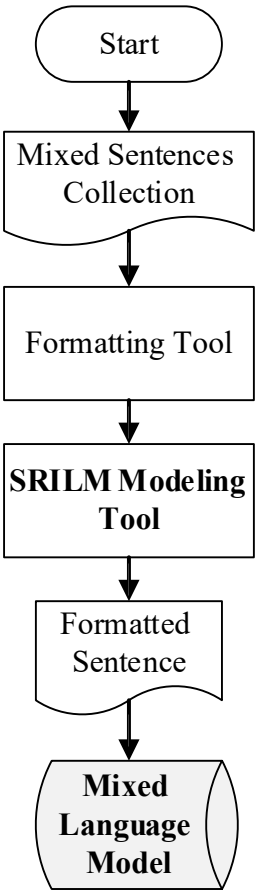


Figure 3.10: Training mixed language model

Table 3.9: LM model for word **الفصل** in the collection

Words	Probability	Model	Comment
الفصل	0.0017	1-gram	Over all occurrence
mode الفصل	0.1234	2-gram	Preceded by word mode
رفعت mode الفصل	0.1283	3-gram	Preceded by words mode and **رفعت**

3.3.6. Mixed Languages Lexicon

In addition to mixed speech recognition model proposed in this work, language identification is also offered. Most speech processing tasks are language dependent, whereas mixed speech phenomenon produce a hybrid language does not belong to any participating languages a mixed languages lexicon is built to facilitate language identification process when number of languages and switching locations are not previously known in mixed speech conversation. It is two columns document, contains words in its original language in the first column associated with its identity in adjacent

column, KHAR code is used for Sudanese Arabic and SE is used for Sudanese English, selection of identities codes takes into accounts future expansion of lexicon to contain a dialects and variants of others languages.

This lexicon is built to support proposed method of language identification in mixed speech mode, Table 3.10 shows part of this lexicon which is manually created for each word in first the column of mixed phonetic dictionary described in section 3.3.4

Table 3.10: Languages Lexicon Part

Word	Identity code
desktop	SE
meeting	SE
……	SE
حفظتو	KHAR
……	KHAR

3.4. Training Phase and Acoustic Model Generation

Automatic speech processing training phase is the collection of processes that use speech samples to produce models and resources essential for recognizing speech samples at testing time unseen at training time. For the purpose of this research model for phonemes, smallest speech unit, is considered. In the following subsections signal processing, GMM-HMM and words lattice generation are described in details.

2.1.1. Signal Processing and Features Extraction

Speech signal in general, in addition to speech attributes, conveys much information related to languages, speaker and environment including background noise and speaker emotion. The signal processing and speech parametrization is a methodology to extract part of information contained in speech signal are much assist in automatic speech processing task in hand. The process eliminates much processing cost

because it reduces efforts and computation for task related information and discard others.

For feature extraction for the purpose of this research MFFC were used as described in section 2.3.2. The process ends up with a vector of 12 coefficients (attributes) for each frame to represent the power spectral envelope of its speech signal. To capture the variability attributes over time delta and delta-delta spectral for each coefficient is taken and added to features vector.

2.1.2. Universal Mixed Acoustic Model Creation

The building block of speech is smallest unit which is called phoneme, these units together form a syllable, word, sentence and other bigger form of the discourse. Human, are baseline and most effective speech interpreter and recognizer system, analyze these sequences of phonemes to understand and interpret speech. For the machines to understand and react to speech, research in the domain of linguistic computation and natural language processing focused in finding a way to properly recognize part of speech based on previous knowledge. It is better to consider smallest and building block unit of speech for recognizing any part of speech either phonemes, words or a whole utterance, the effort is directed towards separates each phoneme form the rest with respect to surrounding context of the utterance. Training phase ends with two structures, cluster for each phoneme in training dataset and words lattice, they are all together called acoustic mode.

For the purpose of this research, AM for each phoneme in training set is modeled and considered as building block of the speech recognition process as illustrated in Figure 3.11. This model is softly clustering each phoneme GMM and calculate transition probability from each GMM to other forming phonemes lattice as described in section 2.4 and its subsequent sections.

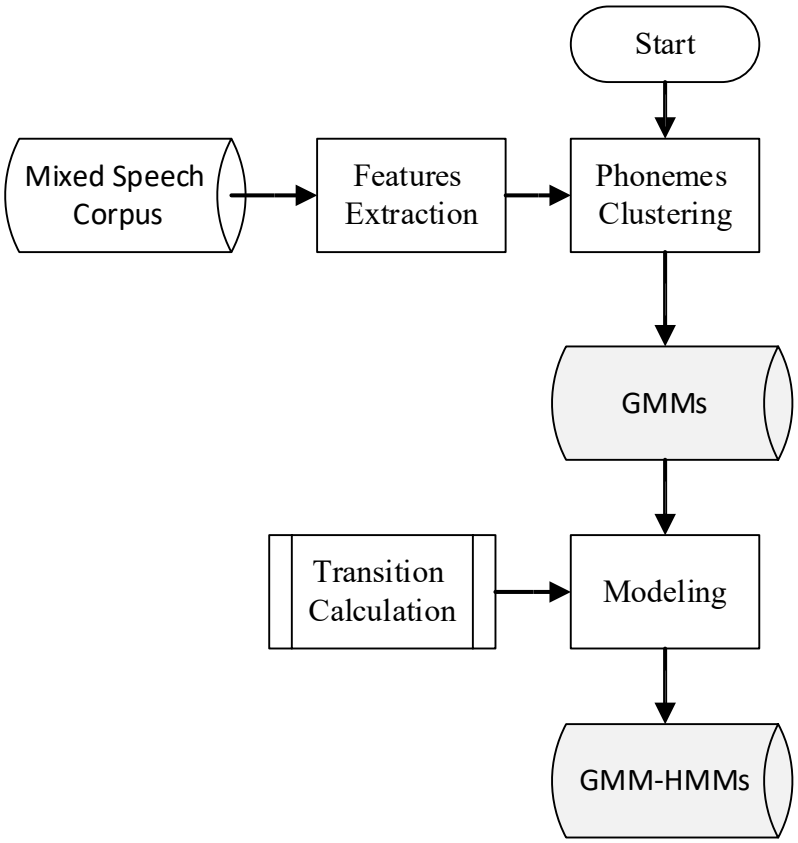


Figure 3.11: GMM-HMMs generation Processes

The second structure produced at this stage is phonemes lattice illustrated in Figure 3.12.

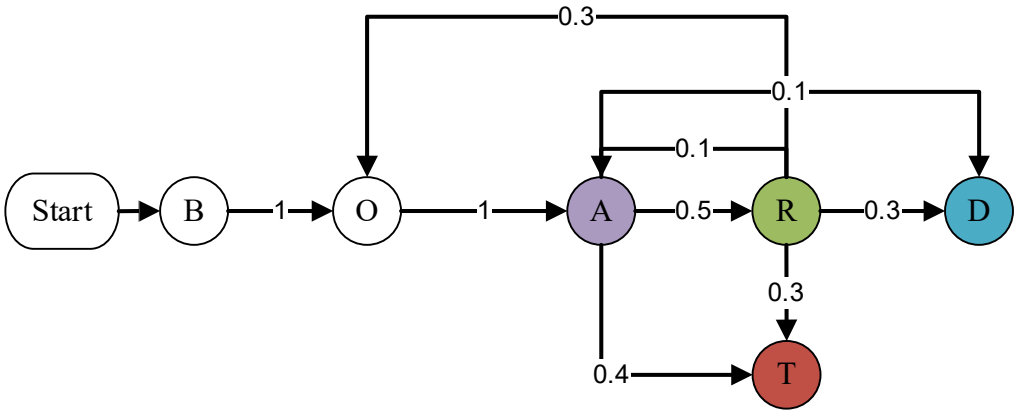


Figure 3.12: Phonemes lattice example

Lattice creation is based on transitions probability from node to others, outgoing links from each node summed up to 1. GMM – HMM is actually composed of GMM for each phoneme and its links from this phoneme to all other phonemes, ends with words: Boa, Boar, Board, Boat and Rat in the lattice are illustrated by filled circle.

3.5. Mixed Speech Recognizer

Mixed speech recognizer is the software tool that follows the same process of speech analysis and feature extraction techniques used in training phase. CMU SPHINX speech recognizer is adapted to recognize test speech and output text corresponding to incoming speech utterance. Mixed language model, mixed phonetic dictionary and acoustic model described above are all utilized as illustrated in Figure 3.13 which describes process sequence and resources required.

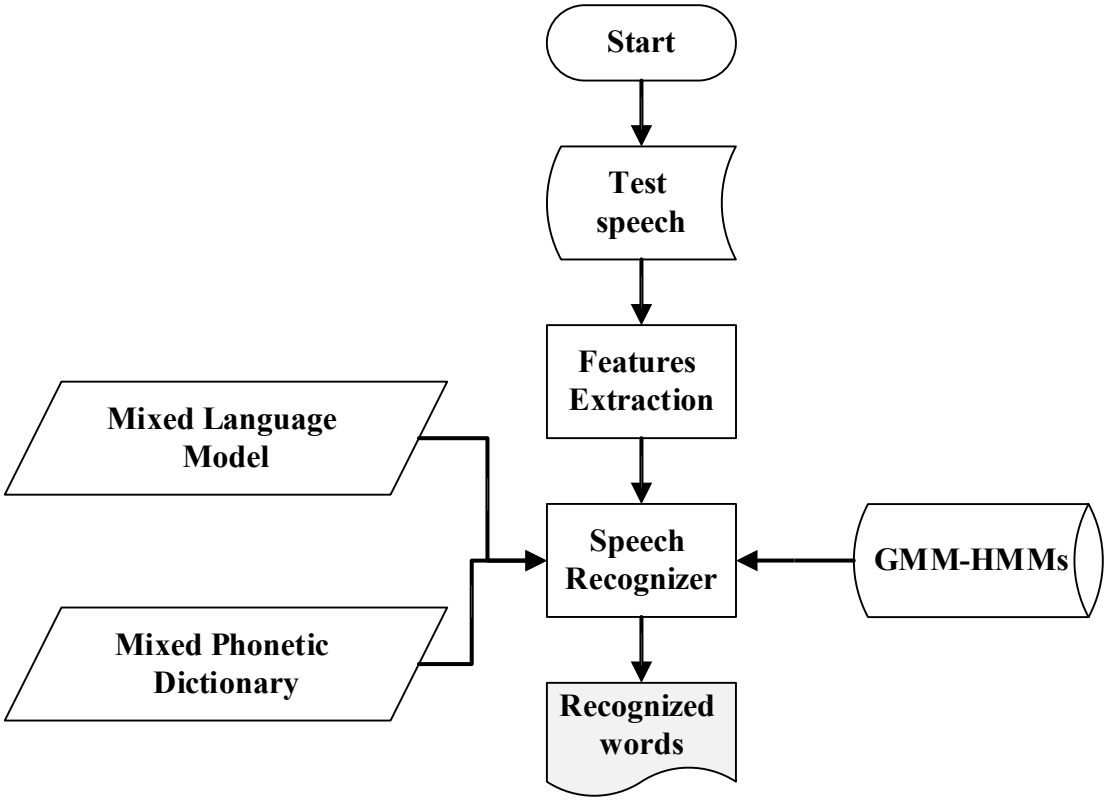


Figure 3.13: Mixed speech recognition process

CMU SPHINX recognizer process receives test utterance, for consistency, same features extraction technique used at training phase were applied. Phonemes

composing words are determined and clustered assisted by GMMs, then according to phonemes lattice single or many words may be formed applying Equation 3.3.

$$P(w_i) = \underset{w}{argmax}(P(O|w_i) \quad (3.3)$$

Where:

w_i word to be recognized

O feature vectors

$P(w_i)$ gives the maximum probability for the feature vectors O

GMM – HMM recognition process does not take into account the word context or language phonotactic regulation that govern sentence structure. This property is provided by language model that was previously built upon text collection to determine words cooccurrence probability, so the final recognition equation is modified as follows:

$$P(w_i) = \underset{w}{argmax}(\underbrace{P(O|w_i)}_{Acoustic\ Model} \quad \underbrace{P(lm_{wi})}_{Language\ Model}) \quad (3.4)$$

Where:

$P(lm_{wi})$ It is the prior probability given by language model for each word in its context, the recognizer will follow the path that have maximum prior probability of language model in words lattice.

Figure 3.14 shows a software code written in C language to adapt open source SPHINX speech recognizer, it receives mixed acoustic model, mixed language dictionary and mixed phonetic dictionary as input parameters, finally recognized words written to test_result.txt file.

```
#include <pocketsphinx.h>
#include <math.h>
#include <stdio.h>
#include <string.h>
#include <locale.h>
#include <dirent.h>
#define MODELDIR
int main(int argc, char *argv[])
{
   ps_decoder_t *ps;    cmd_ln_t *config;    FILE *fh;    char const *hyp, *uttid;    int16 buf[512];
   int rv;    int32 score;   int32 prob;
          char lang = NULL; char c[1000];  FILE *lang_class; char myword[100];      char lang_flag[4];
          FILE *out_file = fopen("E:\\PHD_SOFTWARE\\arabic_model4\\test_result.txt", "a"); // to write
recognized words
   config = cmd_ln_init(NULL, ps_args(), TRUE,
                        "-hmm", MODELDIR("E:\\PHD_SOFTWARE\\arabic_model4\\model4_hmm"), // path
of trained acoustic modell
                        "-lm", MODELDIR("E:\\PHD_SOFTWARE\\arabic_model4\\etc\\arabic_model4.lm"), //
path of mixed language model
                        "-dict", MODELDIR("E:\\PHD_SOFTWARE\\arabic_model4\\etc\\arabic_model4.dic"),
//path of mixed phonetic dictionary
                          NULL);
   if (config == NULL) {
          fprintf(stderr, "Failed to create config object, see log for details\n");
          return -1;
   }
   ps = ps_init(config);
   if (ps == NULL) {
          fprintf(stderr, "Failed to create recognizer, see log for details\n");
          return -1;
   }
   // file of test speech utterance
                      fh =
fopen("E:\\PHD_SOFTWARE\\arabic_model4\\test_audio\\test_spek_007_all_u.wav", "rb");
   if (fh == NULL) {
          fprintf(stderr, "Unable to open input file \n");
          return -1;
   }
   rv = ps_start_utt(ps);
   while (!feof(fh)) {
          size_t nsamp;
          nsamp = fread(buf, 2, 512, fh); // read 512 sample - one frame 2 ms
          rv = ps_process_raw(ps, buf, nsamp, FALSE, FALSE);
   }
   rv = ps_end_utt(ps);
   hyp = ps_get_hyp(ps, &score); //  recognized words sequence
  fprintf(out_file, "%\n", hyp); \\ write result to test_result.txt
}
```

Figure 3.14 C language code to adapt SPHINX speech recognizer

3.6. Automatic Language Identification

Automatic speech processing is language dependent task, in other words, most ASR processes depends on specific linguistic rules and attributes of the target language. The obvious example is bilingual dictionary used for translation between two languages, where the user should specify both languages that participates in translation process e.g. Google Translator application. The language identity of output of traditional speech recognition system is known in monolingual or multilingual mode either is previously known or identified by language identification front module, this favor is not existing in mixed speech conversation, when more than one language may exist in a single sentence. In this research mixed speech is defined as a hybrid language does not belong to any participating languages, the number and locations of switching is not known, this requires language identification module for every word separately. This research proposes simple language identification process avoids problems that occurs in the works in the literature. Boundary detection approach and explicit language identification of each words, both suffers from the effects of native speakers that makes cues used for language identification obscure, no sufficient utterance length to perform language identification and error propagated from step to other.

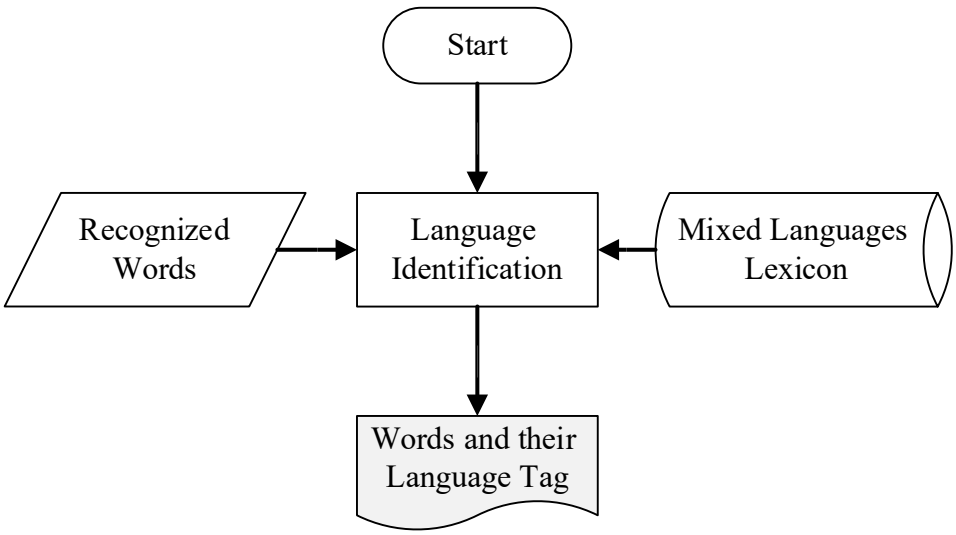


Figure 3.15: Automatic mixed language identification

```
{
1. Receive recognized word from speech recognizer w
2. Open language lexicon for read
3. Search word in the language lexicon
                If found then
                        {
                        Read corresponding language identity Li
                        }
4. Output w and Li
5. Close file
}
```

Figure 3.16: Pseudo code for language identification

Figure 3.16 shows C language code used to determine language identity of each recognized word. This approach is simple and avoids most problems of language identification in mixed speech mode, only keep the dependency of accuracy of mixed speech recognizer.

3.7. Interface for Further Processing (IFP)

To facilitate integration between this hybrid language composed of two languages or more and other linguistic computation dependent processors, such as translator, part of speech tagger, speech to text synthesis, etc., language to language interface were proposed. As illustrated in Figure 3.17, each recognized word accompanied by its language identity, its order in the utterance and recognition confidence. Language identity exploit by language dependent processors to trigger appropriate module, order is used to assembly sentence over network or sentence reformation in other language when the order of the words in the sentence may be different and confidence is provided for use by speech enabled application that accept recognition certainty to some threshold for building query.

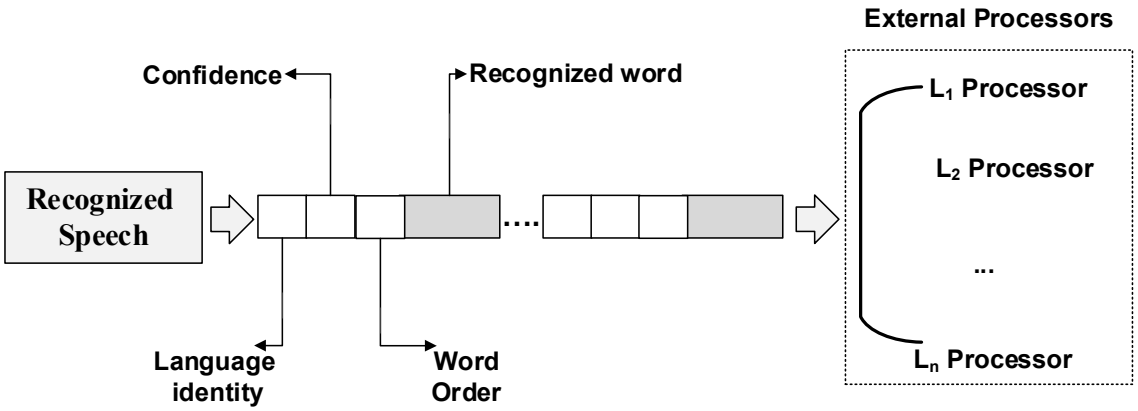


Figure 3.17: Mixed speech interface to further processing

Confidence is calculated by adapting SPHINX API dedicated to calculate recognition confidence as additional service no originally included as showed in Figure 3.18. Its calculation is based on the number of potential hypothesis available for the exact word.

```
ps_seg_t *iter = ps_seg_iter(ps, NULL);
while (iter != NULL)
{
                int32  pprob;
                float confd;
                ps_seg_frames(iter, &sf, &ef); // get words between start &sf and end
&ef
                pprob = ps_seg_prob(iter, NULL, NULL, NULL); // get confidence
probabilty using SPHIX API
                // using ps_seg_prob function in lograthmeic form
                confd = logmath_exp(ps_get_logmath(ps), pprob); //  change to natural
logarithm
                int y = confd * 100;  // give confidence as readable percentage
}
```

Figure 3.18: C code for calculating recognition confidence

CHAPTER IV

GENEREATION AND ANALYSIS OF MIXED SPEECH CORPUS

Motivated by the lack of resources necessary for experiments and creation of speech enabled application, a plan was setup to build mixed speech corpus for the purpose of this research and it will be freely available for researchers in linguistic computation domain. Some unique points give this corpus its strength over similar mixed speech corpora contains Arabic language in the literature, this corpus builds upon daily life mixed speech occurrence, recorded in office environment, Arabic is the first language and both contents pronounced in Sudanese local accent. In the following sections this corpus is

4.1. Mixed Sentences Collection

Based on the fact that Sudanese bilinguals are educated people who most probably use social medial applications in daily basis, a campaign through WhatsApp social media application, which is most usable chat application in Sudan, were launched. WhatsApp used for announcement, setup direct interview appointment and media for mixed sentences collection. Table 4.1 displays statistics for collection methodologies.

Table 4.1: Sentences collection statistics

Collection method	Sentences count	Percentage
WhatsApp Application	137	68.16%
Picked form communities	39	19.40%
Direct Interview	18	8.96%
Short Text Message	7	3.48%

(SMS)		

Campaign announcement includes restriction for own real-life usage, the response is good. A total of 201 distinct daily life mixed sentences were selected out of large amount of collection from different participant that may include same sentence, which is expected in a single community, the process is terminated when repetition from participants got high rate.

The collection was prepared for recording and split into training and test sets as Table 4.2 explains. The ratio is 8:1 between training and test sets, which is under most likely 80:20 traditional method, respectively. This ratio is justified as split happened for raw mixed sentences data before recording, more training sample gives model creation algorithm wide range of speech variabilities representation, accuracy is intended for real usage instead of testing time, test data set ensured to represent mixed speech types (inter-sentential, intra-sentential), recording environment and speakers related variability factors such as gender, age and accent.

Table 4.2: Mixed sentences statistic for training and test sets

Set	**Sentences count**	**Percentage**
Training set	176	87.56%
Test set	25	12.43%

The distribution of words for training set were analyzed and displayed in Table 4.3 excluding most frequent Arabic triggers words in the set that appear in Table 4.4.

Analysis shows mixing is occurred at average 1.5:1 ratio, for Arabic and English languages respectively, in mixed sentence formation. High distinction ratio of words is achieved through following effective filtering and elimination strategies at collection time. High occurrence of trigger words is due to speakers trying smoothing his transition and switching between two languages **updated نظامه ما**

Table 4.3: Training set words distribution

Language	**Words count**	**Percentage**	**Distinct Words**	**Distinction Ratio**
Sudanese Arabic	267	59.20%	227	0.85
Sudanese English	202	40.80%	191	0.95

Statistically, participation of each Arabic words in training set is 0.433%, table 4.4 shows non distinct words and their occurrence ratio.

Table 4.4: Arabic trigger words occurrences in training set

Word	**Occurrence**	**Occurrence Percentage**
في	10	4.33%
ما	7	3.03%
من	5	2.16%
فيها	4	1.73%

Figure 4.1 shows 75% of bilingual Sudanese speakers tend to opening and concluding his mixed sentences by embedded language, approximately 60% of them opening their idea by English, and only quarter of common daily life mixed sentences have and embedded words between two Arabic phrases e.g. **دخل من turn غلط.**

Inter-sentential tends of speakers reduces the number of switches in the sentence, therefor, the effect of transition and pronunciation interference between languages is also minimized.

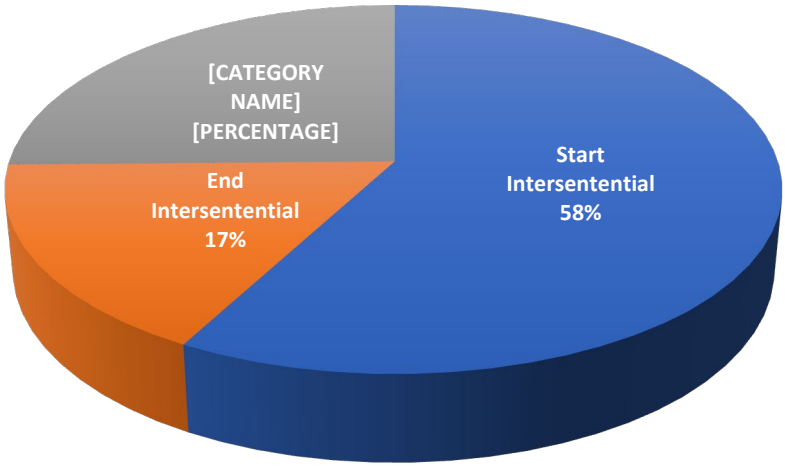


Figure 4.1: Speakers recognition representation

The test part of the corpus is randomly selected, Table 4.5 shows the total number of words from each language and their occurrence distinction ratio for test collection.

Table 4.5: Testing words distribution

Language	Words count	Percentage	Distinct Words	Distinction Ratio
Sudanese Arabic	43	62.32%	35	0.81
Sudanese English	26	37.68%	26	1

Figure 4.2 shows the distribution of collection switching types, which indicates behavior of speakers expressing their thoughts.

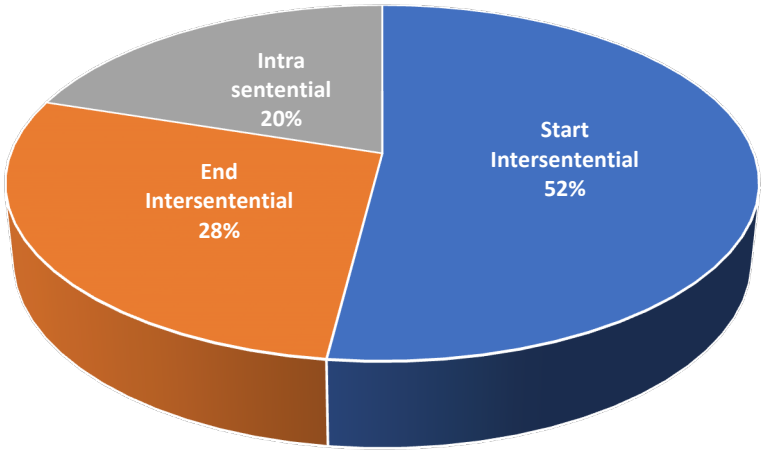


Figure 4.2: Test set switch types distribution

Table 4.6: Comparison of words distribution in the collection

Language	**Total words**		**Distinct Words Ration**	
	Training	Test	Training	Test
Sudanese Arabic	59.20%	62.32%	0.85	0.81
Sudanese English	40.80%	37.68%	0.95	1

The reality of speech corpus contents that is used for training and testing automatic speech recognition model, has a fatal effect on the performance and robustness of the system in the real usage. Comparison of the parts of the collections in terms of words distribution illustrated in Table 4.6 and in terms of switching types showed in Table 4.7 indicates the community behavior of bilingual when expressing their ideas in more than one language. Keeping in mind the part of testing set is randomly selected from the whole collection.

Table 4.7: Comparison of switching types in the collection

Language	Training Set	Testing Set
Start Inter-Sentential	58%	52%
End Inter-Sentential	17%	28%
Intra-Sentential	25%	20%

4.2.Training Set Statistics

Bilingual Sudanese adult speakers of age range between 20-60 years participated in recording training collection in their place (own office environment), training set is prepared to include as much as possible speech variation that ensured modeled most cases regarding human communication. Table 4.8 shows the speakers participation statistics for total speech length of 1.07 hours.

Table 4.8: Training set speakers and recording analysis

Remark	Count	Percentage
Male speakers	62	71.26%
Female speakers	25	28.74%
Male audio files	1646	71.91%
Female audio Files	643	28.09%

For best modeling of speech, the following attributes regarding variations of speakers, environment and type of code switching were maintained:

1. Used most frequent daily life mixed speech sentences.
2. Represents all types of code-switching.
3. Both genders are participating with 3:1 ratio for male and female respectively.

4. Cover age range of 20 – 60 years old.
5. Speakers represents a wide range of ethnic groups and accent variations.
6. Recording at speaker location ensured environment variation in the set.
7. Each sentence read by at least 10 speakers.
8. Each speaker read at least 10 different sentences with guarantee representation of all types of code switching.

CHAPTER V
RESULTS AND DISCUSSIONS

Speech enabled applications is ultimate goal of linguistic computation for decades, keeping in mind wide range of speech variabilities factors, performance robustness was achieved for some types of speech, specifically for isolated parts of speech or text and speaker dependent speech. Human – machine interaction through speech were successfully applied in real life application, such as utilizing voice as a biometric measure for access control, querying and retrieve information from weather or travel database, software that deaf and blinds dealing with computers and machines and cars or robot's navigator programs and much other software.

Active research issues in the domain of linguistic computation such as language dialects, accent variation, speaker environment effects and mixed speech were addressed and studies in this work using our own mixed speech corpus and applying appropriate algorithms towards building reliable and generalized model has a reasonable robustness against these variabilities. In the following sections, proposes mixed speech and language identification models were tested, for model generalization, monolingual experiments for Arabic and English languages were done. To test model capability of enroll new languages without model retrain small mixed speech collection of Donglwai and Arabic were used for this experiment.

5.1. Speech Recognition Experiments

Mixed speech corpus and model were designed to recognize speech in mixed or monolingual manner. Sudanese Arabic and Sudanese English mixed speech test were done, monolingual test for Arabic and English were also performed. For generalization purpose, a test that measures mixed speech between Arabic and local language were performed.

5.1.1. Mixed Speech Recognition Experiments

A test collection described in chapter III and analyzed in chapter IV of this thesis were used to test performance of mixed speech model of Sudanese Arabic and Sudanese English languages based on Sudanese pronunciation. Table 5.1 shows a test of 6 speakers with of 3:1 gender ratio representation of male and female respectively,

which matches the training ratio, overall 33.05% WER were achieved for mixed speech model

Table 5.1: Mixed speech test statistic

Speaker	Gender	Substitution %	Insertion %	Deletion %	WER
Speaker1	M	26.09	2.90	**2.9**	**31.88**
Speaker2	M	**8.70**	2.90	4.35	**15.94**
Speaker3	M	20.29	5.80	20.29	**46.38**
Speaker4	F	20.29	11.59	15.94	**47.83**
Speaker5	M	10.14	5.80	5.80	**21.47**
Speaker6	F	15.94	**0**	18.84	**34.78**

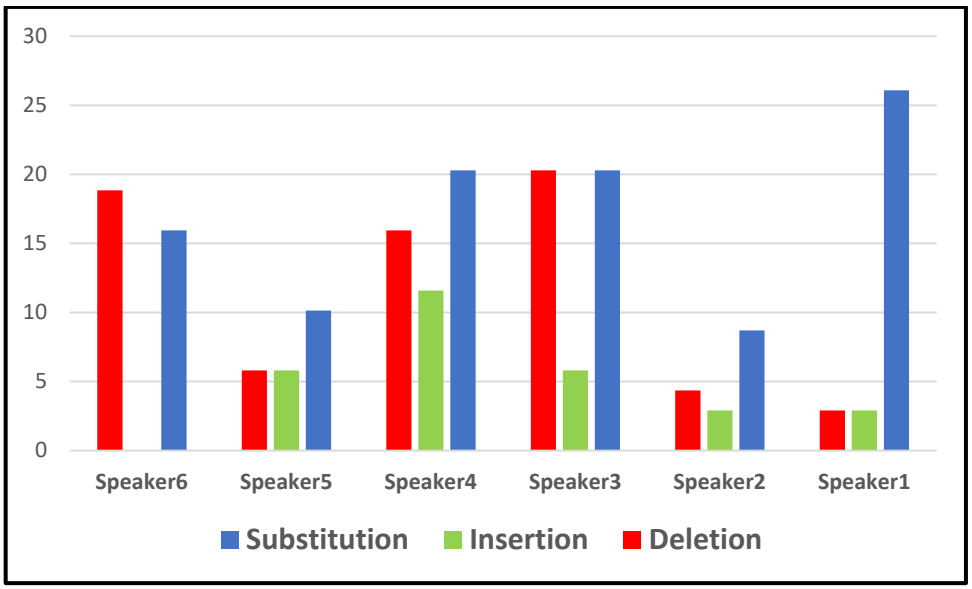


Figure 5.1: Speakers wrong recognition distribution

Figure 5.1 shows wrong recognition percentage distributed among error factors for each speaker. Insertion of the words is due to environment effects in recording process, such as background noise, recording equipment and utility setting or other talk

interference from TV, Telephone or other person talked in the office. The most factor of substitution is misrecognition to Out of Vocabulary (OOV), which is result from a big margin of accent variation between Sudanese speakers. Deletion is occurred when speech signal is under sound activation level of sound, or the signal add to previous one due to wrong segmentation.

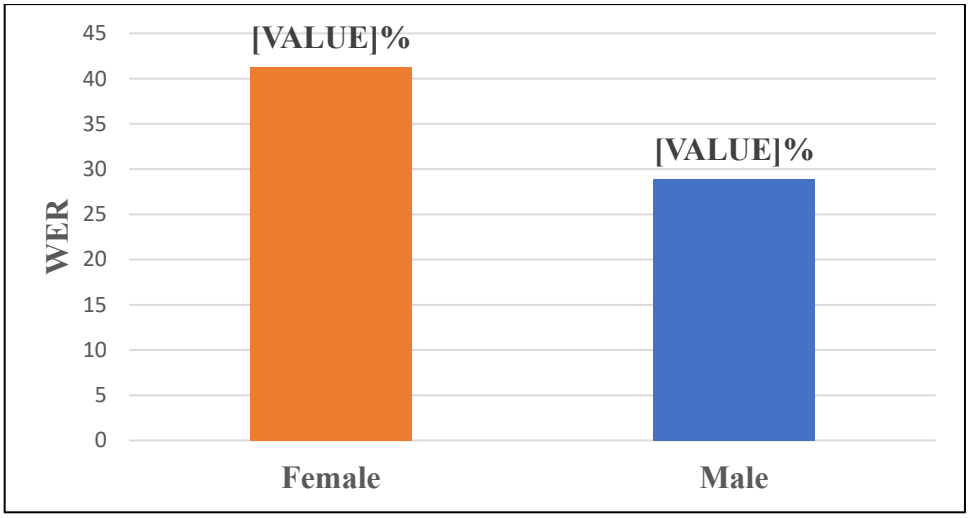


Figure 5.2: Gender Average WER Percentage

Female WER error is higher due to high rate of substitution, because they try to stick to right pronunciation for both languages. Table 5.2 shows example of substitution for Arabic to Arabic, English to Arabic, Arabic to English and English to English words.

Table 5.2: Substitution Examples

Word	Substitution
جديد	لقيت
Result	رسلت
واقف	Wi Fi
Facebook	System

Figure 5.3 shows performance for selected six speakers against mixed speech test corpus along with representation of wrong recognition on WER metric.

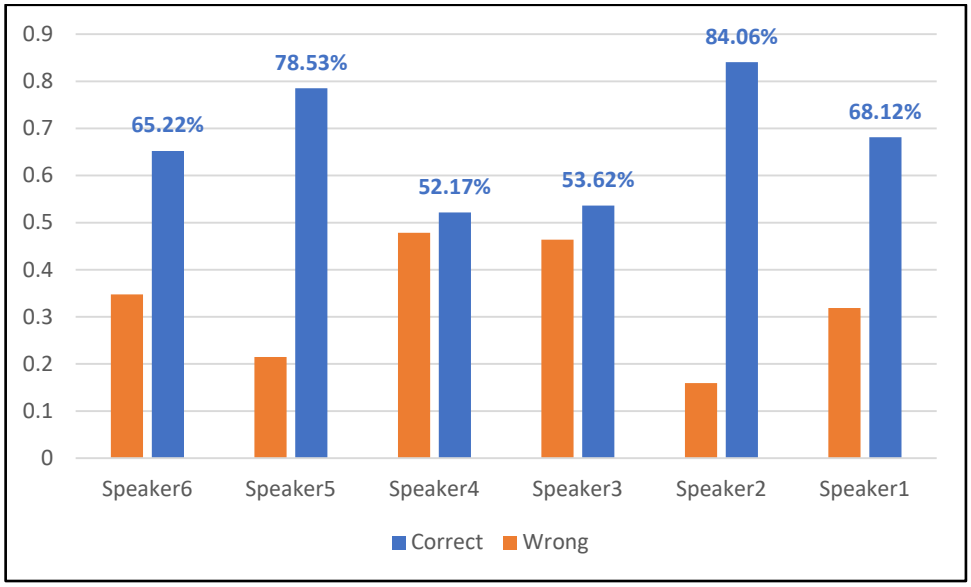


Figure 5.3: Speakers Performance Representation

Table 5.3: Previous model recognition and environment comparison

Year	**Languages**	**Type of speech**	**Performance**
2006	Mandarin - Taiwanese	TV program	Promising
2010	Hindi - English	Read sentences	53.73% WER
2012	Mandarin-English	Conversational in quiet room	36.60% WER
2016	Frisian - Dutch	Radio broadcast	59.5% WER
2016	IsiZulu – English	Broadcast Operas	80% WER
2019	Algerian Arabic -French	Radio broadcast	Confidence Score
2019	Sudanese Arabic - English	Read Speech	33.05 % WER

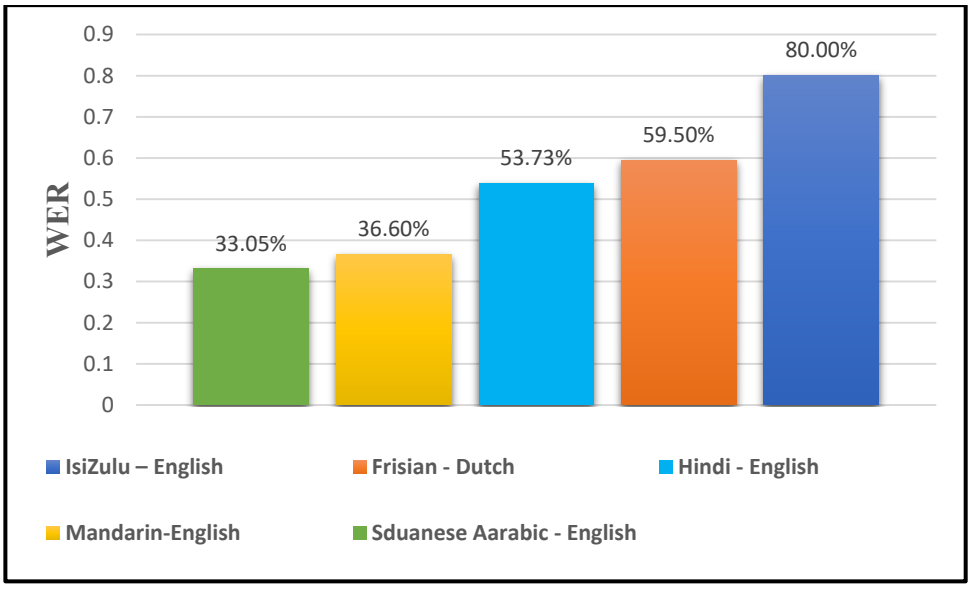


Figure 5.4: WER for five mixed speech recognizer

Table 5.3 and figure 5.4 show our proposed universal mixed speech recognizers based on pronunciation and phonetic representation of Sudanese Arabic outperform methods of multi-pass recognizers, monolingual language model, a single and universal phone set of languages participates in code switching, this achieved by taking effects of native language, which is Arabic in this case, into account avoiding clustering and classification at each stage depends on language property, instead this model completely based on native pronunciation for both languages participating in switching.

To measure effect of one language over other in mixed speech communication, result statistic was calculated as illustrated in Table 5.4 and Figure 5.5 which compares overall performance, sentences starting with Arabic word and sentences start with English word. Overall and language started with Arabic word performance is consistent, whereas, deviation of performance occurred in sentences started with English, we conclude this is due to Arabic native speakers tend to speak English correctly and rearranged articulation organs to produce sounds as English native, whereas English smoothly pronounced after Arabic word without reconfiguration of articulation organs.

Table 5.4: WER % Based on sentence initialization language

Speaker	Overall	Start with Arabic	Start with English
Speaker1	31.88	23.68	44.44
Speaker2	15.94	15.79	19.44
Speaker3	46.38	36.84	63.89
Speaker4	47.83	42.12	63.89
Speaker5	21.47	34.21	25
Speaker6	34.78	23.68	55.56

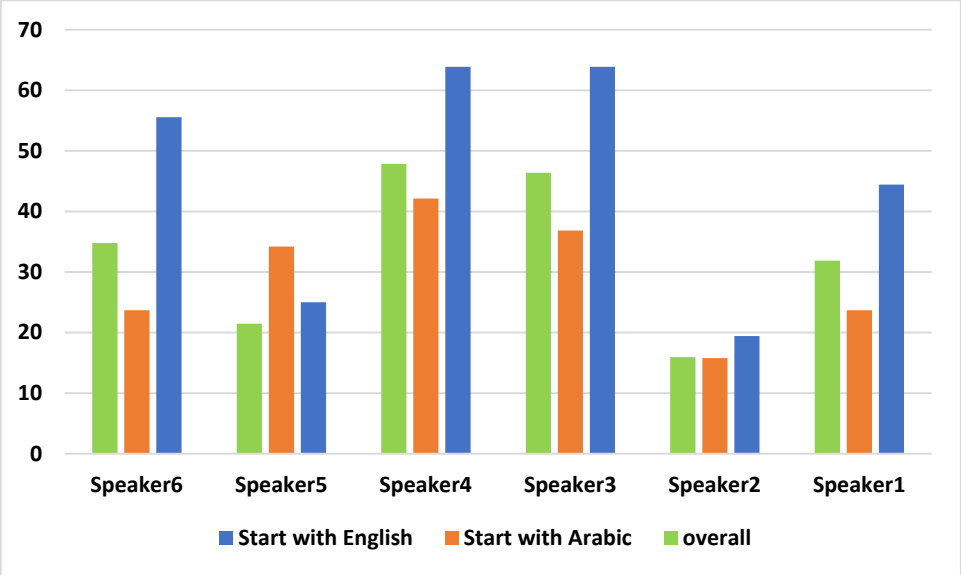


Figure 5.5: Comparison of WER based on Sentence Initialization Language

5.1.2. Monolingual Speech Recognition Experiments

Small test corpus for monolingual recognition of Arabic and English languages were created to test generalization capability of mixed speech model to recognize monolingual speech utterance for languages participates in model creation. Table 5.5 and Figure 5.6 illustrate performance in WER metric for both languages.

High WER rate of monolingual recognition is due to lack of monolingual sample in the language model, test sample is small. Arabic language has a better performance because it is a mother language for speakers in its sample is 18% more the English sample.

Table 5.5: Monolingual Test of Mixed Speech Model

Language	WER
Sudanese Arabic	55.25%
Sudanese English	71.0%

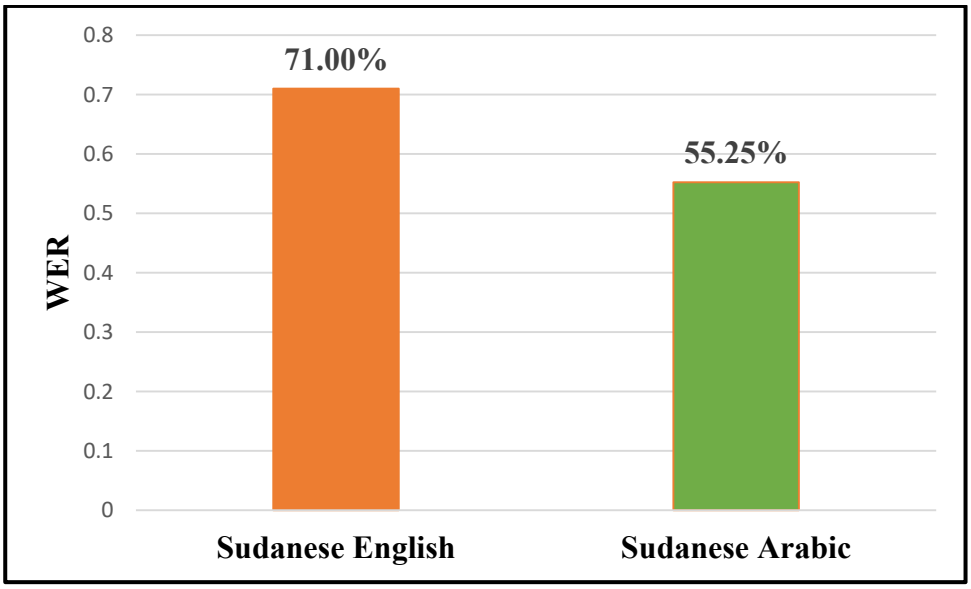


Figure 5.6 Monolingual test of mixed speech model

Other reason that increase substitution rates people tends to pronounce English as its native speakers when speaks English, Table 5.6 shows confusion matrix or substitution between languages, high substitution of English to English supports the assumption of mimicking native speakers. The results indicate articulation process based on speaker more than language.

Table 5.6: Monolingual substitution confusion matrix

language	Arabic	English
Sudanese Arabic	50	50
Sudanese English	33.3	66.7

5.1.3. Sudanese Arabic - Dongolawi languages mixed speech recognition

Solution generalization is one of the design goals of this work, a simple experiment for mixed speech between the Sudanese Arabic and Donglwai (Andaandi) language, which is local language in northern Sudan spoken by 70,000 people, Arabic is used as one scripting language of its writing system (Eberhard et al., 2019), were performed using mixed Arabic – English model without model retrain. Only dictionary entries, language lexicon entries and text collection for language model to maintain speech context is used to add this language.

Table 5.7 shows Arabic and Donglawi mixed sentences, Donglawi words and translation to Arabic of whole sentence, column contains translation to Arabic language shows the importance of word order as element of IFP for sentence reformation in Arabic. Nevertheless, these examples are very view, recognition test shows promising result and capability of model generalization.

Table 5.7: Arabic transliterated mixed Arabic – Donglawi sentences

Mixed sentences	Donglawi words	Arabic translation
جِيب مين سَاعِقى	من ساعقى	جيب الساعة دي
سُوكَارْكى اتَارى من الدكان	سوكاركى اتارى	جيب سكر من الدكان
غش تَنى دَامُون	تنى دامون	غش مافي تاني
إسِنتَاد الشغل متين	إسنتاد	ماشي الشغل متين

5.2. Language Identification Results

As a part of IFP of this digital language to outside world, Language identity of each recognized word is looked up from universal mixed language lexicon that contains all words in the dictionary associated with its language identity as described in Chapter 3 of this thesis.

This open set language identification task is performed in simple and cost-effective manner in spite of suffering from error propagated from speech recognition module. Language identification in this way for mixed speech processing is better than using separate language identifier model, which is based in language dependent properties that are not clear in this situation due to effect of one language to other as illustrated in table 5.6 where confusion rate is high. Adding new language to this identifier is simple as adding entries of its word to mixed language lexicon.

CHAPTER VI
CONCLUSIONS AND FUTURE WORKS

6.1. Conclusions

For the purpose of this research Sudanese Arabic - English mixed speech corpus was built and will be freely released as contribution to fill the gap of essential resources needed to develop, test and evaluate speech recognition systems for Arabic language in general and Sudanese community languages contained Sudanese Arabic, Sudanese English, local languages and mixed speech daily life usage.

This research concludes that speech recognition model could be generalized based on speakers rather than languages, this research utilizes Sudanese Arabic language, enrolled its native speakers to recognize its speech in monolingual or mixed with other languages for the same natives without retrain, this is done based on assumption of dominance of native language over others, this is proved by outperforming previous model performance in the domain achieving overall accuracy of 33.05% in WER metric. New language can be easily added to the model without the need to retrain mixed speech model.

Mixed speech communication is not belonging to each participating language, is considered hybrid and not defined to other language dependent processors such as translator, in which language identity is essential input to complete its job, a novel approach of determined language identity in the environment where the languages cues used to know which language is spoken not clear, proposed and applied, this approach eliminates error propagation problem of language identification in mixed speech mode, only exists when recognizer substitute word to other language.

6.2. Future Works

This work is considered a beginning of regional natural language processing and linguistic computation in the area of the world full of local languages and accent variations. Mixed speech corpus should be extended to contain speech from all Sudanese languages and dialects in monolingual and mixed manner, representing all speech variabilities that may encounter various speech processing systems. Experiments for none Sudanese native is encouraged to evaluate model behavior in this situation.

To ease adding new language to the model, software tool for new language phonemes clustering and cross checked with existing phones set is needed to perfectly enrolled language's sounds. Develop Arabic language diacritics and transliteration other language to Arabic software tools to eliminate error propagation, boring and time-consuming process to facilitate extend mixed phonetic dictionary, noise normalization and cancelation module are essential to use this recognizer anywhere at any time, and resolving OOV problem.

REFERENCES

ADAMI, A. G. & HERMANSKY, H. Segmentation of speech for speaker and language recognition. INTERSPEECH, 2003.

ADDA-DECKER, M., ANTOINE, F., VASILESCU, I., LAMEL, L., VAISSIERE, J., GEOFFROIS, E. & LIENARD, J.-S. Phonetic knowledge, phonotactics and perceptual validation for automatic language identification. In ICPhS, 2003. Citeseer.

ADEL, H., VU, N. T., KIRCHHOFF, K., TELAAR, D. & SCHULTZ, T. 2015. Syntactic and semantic features for code-switching factored language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 23**,** 431-440.

ALI, A., DEHAK, N., CARDINAL, P., KHURANA, S., YELLA, S. H., GLASS, J., BELL, P. & RENALS, S. 2015. Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.

AMAZOUZ, D., ADDA-DECKER, M. & LAMEL, L. The French-Algerian Code-Switching Triggered audio corpus (FACST). LREC 2018 11th edition of the Language Resources and Evaluation Conference, 2018.

AMAZOUZ, D., ADDA-DECKER, M. & LAMEL, L. Addressing code-switching in French/Algerian Arabic speech. Interspeech 2017, 2019. 62-66.

ANANTHI, S. & DHANALAKSHMI, P. 2013. Speech recognition system and isolated word recognition based on Hidden markov model (HMM) for Hearing Impaired. *International Journal of Computer Applications,* 73**,** 30-34.

ASSOCIATION, I. P. & STAFF, I. P. A. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, Cambridge University Press.

BALLEDA, J., MURTHY, H. A. & NAGARAJAN, T. Language identification from short segments of speech. INTERSPEECH, 2000. 1033-1036.

BENZEGHIBA, M., DE MORI, R., DEROO, O., DUPONT, S., ERBES, T., JOUVET, D., FISSORE, L., LAFACE, P., MERTINS, A. & RIS, C. 2007. Automatic speech recognition and speech variability: A review. *Speech communication,* 49**,** 763-786.

BENZEGHIBA, M., DE MORI, R., DEROO, O., DUPONT, S., ERBES, T., JOUVET, D., FISSORE, L., LAFACE, P., MERTINS, A., RIS, C. AND ROSE, R., 2007.

Automatic speech recognition and speech variability: A review. *Speech Communication,* 10**,** 763-786.

BHUVANAGIRI, K. & KOPPARAPU, S. 2010. An approach to mixed language automatic speech recognition. *Oriental COCOSDA, Kathmandu, Nepal*.

BHUVANAGIRIR, K. & KOPPARAPU, S. K. 2012. Mixed language speech recognition without explicit identification of language. *American Journal of Signal Processing,* 2**,** 92-97.

BRAUN, J. & LEVKOWITZ, H. Automatic language identification with perceptually guided training and recurrent neural networks. International Conference of Spoken Language Process, 1998 Sydney, Australia. 289–292.

ÇETINOĞLU, Ö. A code-switching corpus of Turkish-German conversations. Proceedings of the 11th Linguistic Annotation Workshop, 2017. 34-40.

CHAN, J. Y., CHING, P. & LEE, T. Development of a Cantonese-English code-mixing speech corpus. Ninth European Conference on Speech Communication and Technology, 2005.

CHU-CARROLL, J. & CARPENTER, B. 1999. Vector-based natural language call routing. *Computational linguistics,* 25**,** 361-388.

CIMARUSTI, D. & IVES, R. Development of an automatic identification system of spoken languages: Phase I. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82., 1982. IEEE, 1661-1663.

COMBRINCK, H. & BOTHA, E. Automatic Language Identification: Performance vs. Complexity. In Proceedings of the Sixth Annual South Africa Workshop on Pattern Recognition, 1997. Citeseer.

CORONA, R., THOMASON, J. & MOONEY, R. Improving Black-box Speech Recognition using Semantic Parsing. Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2017. 122-127.

DAVID & SELFRIDGE. eyes and ears of comnputers. IRE, 1962. 1093-1101.

DEHAK, N., TORRES-CARRASQUILLO, P. A., REYNOLDS, D. & DEHAK, R. Language recognition via i-vectors and dimensionality reduction. Twelfth Annual Conference of the International Speech Communication Association, 2011.

DENES, P. 1959. The design and operation of the mechanical speech recognizer at University College London. *Journal of the British Institution of Radio Engineers,* 19**,** 219-229.

DO, C. B. & BATZOGLOU, S. 2008. What is the expectation maximization algorithm? *Nature biotechnology,* 26**,** 897.

DUDLEY, H. 1939. *the Voder, the fisrt voice synthesiser Machine* [Online]. Bel Lab. Available: http://www.whatisthevoder.com/ [Accessed 31/5/2019 2019].

EBERHARD, DAVID M., GARY F. SIMONS & FENNIG, C. D. 2019. *Ethnologue: Languages of the World: Arabic, Sudanese Spoken* [Online]. Ehnologue. Available: https://www.ethnologue.com/language/apd [Accessed 28/5/2019 2019].

EBERHARD, DAVID M., GARY F. SIMONS & FENNIG, C. D. 2019 online version: http://www.ethnologue.com. Ethnologue: Languages of the World. *Twenty-second edition,* 22.

EBERHARD, D. M., GARY F. SIMONS, AND CHARLES D. FENNIG 2019. Ethnologue: Languages of the World, Dallas, Texas: SIL International. *Online version: http://www. ethnologue. com,* 22.

EDDY, S. R. 2004. What is a hidden Markov model? *Nature biotechnology,* 22**,** 1315.

ELLIOTT, S. J. & SHERA, C. A. 2012. The cochlea as a smart structure. *Smart Materials and Structures,* 21**,** 064001.

FARINAS, J. & PELLEGRINO, F. Automatic rhythm modeling for language identification. Seventh European Conference on Speech Communication and Technology, 2001 Aalborg, Denmark.

FLOCCIA, C., GOSLIN, J., GIRARD, F. & KONOPCZYNSKI, G. 2006. Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance,* 32**,** 1276.

FORSBERG, M. 2003. Why is speech recognition difficult. *Chalmers University of Technology*.

GENG, W., WANG, W., ZHAO, Y., CAI, X. & XU, B. End-to-End Language Identification Using Attention-Based Recurrent Neural Networks. INTERSPEECH, 2016. 2944-2948.

GONZALEZ-DOMINGUEZ, J., LOPEZ-MORENO, I., MORENO, P. J. & GONZALEZ-RODRIGUEZ, J. 2015. Frame-by-frame language identification in short utterances using deep neural networks. *Neural Networks,* 64**,** 49-58.

GUTTORP, P. & MININ, V. N. 2018. *Stochastic modeling of scientific data*, Chapman and Hall/CRC.

H. SOLTAU, E. A. The IBM 2004 conversational telephone system for rich transcription. ICASSP, 2005. 205-208.

HAIZHOU LI, B. M., KONG AIK LEE 2013. Spoken Language Recognition: from Fundamentals to Practice. *in proc. IEEE Spoken Language Recognition,* 101**,** 1136 - 1159.

HALL JR, R. A. 1961. *Sound and Spelling in English*, Chilton Books, Educational Division.

HAMED, I., ELMAHDY, M. & ABDENNADHER, S. Collection and analysis of code-switch egyptian arabic-english speech corpus. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018), 2018.

HASSAN, E.-K. M. I. & HASSAN, S.-A. W. A. 2014. Pronunciation Problems: A Case Study of English Language Students at Sudan University of Science and Technology *English Language and Literature Studies,* 4**,** 31 -44.

HERMANSKY, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *the Journal of the Acoustical Society of America,* 87**,** 1738-1752.

HINTON, G., DENG, L., YU, D., DAHL, G., MOHAMED, A.-R., JAITLY, N., SENIOR, A., VANHOUCKE, V., NGUYEN, P. & KINGSBURY, B. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine,* 29.

HOMBERT, J.-M. & MADDIESON, I. The Use of'Rare'Segments for Language Identification. Sixth European Conference on Speech Communication and Technology, 1999.

HOUSE, A. S. & NEUBURG, E. P. 1977. Toward automatic identification of the language of an utterance. I. Preliminary methodological considerations. *The Journal of the Acoustical Society of America,* 62**,** 708-713.

IAN MADDIESON & PRECODA, K. 1984. *UCLA Phonological Segment Inventory Database (UPSID)* [Online]. University of California, Los Angeles (UCLA) Available: http://phonetics.linguistics.ucla.edu/sales/software.htm#upsid [Accessed 11/10/2017 2017].

IBRAHIM M. M. EL-EMARY, MOHAMED FEZARI & ATTOUI, A. H. 2011. Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice

command system: A way to improve the control of remotely operated robot arm TR45 *Scientific Research and Essays,* 6**,** 341-35.

IMSENG, D., BOURLARD, H., DOSS, M. M.-. & DINES, J. Language dependent universal phoneme posterior estimation for mixed language speech recognition. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011. IEEE, 5012-5015.

IMSENG, D., BOURLARD, H. & DOSS, M. M. Towards mixed language speech recognition systems. Eleventh Annual Conference of the International Speech Communication Association, 2010.

IVANOV, A. V., LANGE, P. L., SUENDERMANN-OEFT, D., RAMANARAYANAN, V., QIAN, Y., YU, Z. & TAO, J. 2016. Speed vs. accuracy: Designing an optimal asr system for spontaneous non-native speech in a real-time application. *Proc. of the IWSDS, Saariselk, Finland*.

IVES, R. A minimal rule AI expert system for real-time classification of natural spoken languages. 2nd Annual Artifiial Intelligence and Advanced Computer Technology Confernce, may 1986 Long Beach, CA.

J. FERGUSON, E. 1980. Hidden Markov models for speech. *Institute for Defense Analysis*.

J.SUZUKI & K.NAKATA 1961. Recognition of Japanese Vowels Preliminary to the Recognition of Speech. *J.Radio Res. Lab,* 37 (8)**,** 193-212.

JUANG, B.-H. & RABINER, L. R. 2005. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara,* 1**,** 67.

JUANG, B. H. & RABINER, L. R. 1991. Hidden Markov models for speech recognition. *Technometrics,* 33**,** 251-272.

JURAFSKY, D. & MARTIN, J. H. 2009. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Prentice Hall, Pearson Education International.

KETTANI, H. 2010 world muslim population. proceedings of the 8th Hawaii Internafional Conference on Arts and Humanifies, 2010. 12-16.

KUMAR, C. S., LI, H., TONG, R., MATĚJKA, P., BURGET, L. & ČERNOCKý, J. Tuning phone decoders for language identification. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010. IEEE, 5010-5013.

KWAIK, K. A., SAAD, M., CHATZIKYRIAKIDIS, S. & DOBNIK, S. 2018. Shami: A corpus of levantine arabic dialects.

LADEFOGED, P. & HALLE, M. 1988. Some major features of the International Phonetic Alphabet. *Language,* 64**,** 577-582.

LAMEL, L. F. & GAUVAIN, J.-L. Cross-lingual experiments with phone recognition. Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on, 1993. IEEE, 507-510.

LAMEL, L. F. & GAUVAIN, J.-L. Language identification using phone-based acoustic likelihoods. Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on, 1994. IEEE, I/293-I/296 vol. 1.

LAW, M. H., TOPCHY, A. & JAIN, A. K. Clustering with soft and group constraints. Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), 2004. Springer, 662-670.

LEE, G., HO, T.-N., CHNG, E.-S. & LI, H. A review of the Mandarin-English code-switching corpus: SEAME. 2017 International Conference on Asian Language Processing (IALP), 2017. IEEE, 210-213.

LEE, K. F. An overview of the SPHINX speech recognition system. ICASSP, 1990. 600-610.

LEONARD, R. G. March 1980. Language Recognition Test and Evaluation. Air Force Rome Air Development Center.

LEONARD, R. G. & DODDINGTON, G. R. August 1974. Automatic Language Identification. Air Force Rome Air Development Center.

LI, H. & MA, B. A phonotactic language model for spoken language identification. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, 2005. Association for Computational Linguistics, 515-522.

LI, H., MA, B. & LEE, C.-H. 2007. A vector space modeling approach to spoken language identification. *IEEE Transactions on Audio, Speech, and Language Processing,* 15**,** 271-284.

LI, K. & EDWARDS, T. Statistical models for automatic language identification. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'80., 1980. IEEE, 884-887.

LIEW, A. W.-C. & WANG, S. 2009. *Visual speech recognition: lip segmentation and mapping*, Medical Information Science Reference Hershey, PA.

LIM, B. P., LI, H. & MA, B. Using local & global phonotactic features in Chinese dialect identification. Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on, 2005. IEEE, I/577-I/580 Vol. 1.

LINGUISTIC DATA CONSORTIUM. 1996. *CALLFRIEND Speech Corpus* [Online]. Available: http://www.ldc.upenn.edu/ ldc/ about/callfriend.html [Accessed 22/10/2017 2017].

LOPEZ-MORENO, I., GONZALEZ-DOMINGUEZ, J., PLCHOT, O., MARTINEZ, D., GONZALEZ-RODRIGUEZ, J. & MORENO, P. Automatic language identification using deep neural networks. Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014. IEEE, 5337-5341.

LOZANO-DIEZ, A., ZAZO CANDIL, R., GONZáLEZ DOMíNGUEZ, J., TOLEDANO, D. T. & GONZALEZ-RODRIGUEZ, J. An end-to-end approach to language identification in short utterances using convolutional neural networks. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2015. International Speech and Communication Association.

LYU, D.-C. & LYU, R.-Y. Language identification on code-switching utterances using multiple cues. Ninth Annual Conference of the International Speech Communication Association, 2008.

LYU, D.-C., LYU, R.-Y., CHIANG, Y.-C. & HSU, C.-N. Speech recognition on code-switching among the Chinese dialects. 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, 2006. IEEE, I-I.

M. OSMAN ELTAYEB & MOHAMMED ELHAFIZ MUSTAFA. Acoustic-support vector machines approach to detect spoken Arabic language. Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on, 2013. IEEE, 525-529.

M.A.ANUSUYA & S.K.KATTI 2009. Speech Recognition by Machine: A Review. *International Journal of Computer Science and Information Security (IJCSIS),* 6**,** 181-205.

MA, B., GUAN, C., LI, H. & LEE, C.-H. Multilingual speech recognition with language identification. INTERSPEECH, 2002.

MABULE, D. 2015. What is this? Is it code switching, code mixing or language alternating? *Journal of Educational and Social Research,* 5**,** 339.

MARC A. ZISSMAN, K. M. B. 2001. Automatic Language Identification. *speech comunication,* 35**,** 115-124.

MARTIN, T. B., NELSON, A. & ZADELL, H. 1964. SPEECH RECOGNITION BY FEATURE-ABSTRACTION TECHNIQUES. RAYTHEON CO WALTHAM MASS.

MARTINEZ, D., BURGET, L., FERRER, L. & SCHEFFER, N. iVector-based prosodic system for language identification. Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 2012 Japan. IEEE, 4861-4864.

MARTıNEZ, D., PLCHOT, O., BURGET, L., GLEMBEK, O. & MATEJKA, P. 2011. Language recognition in ivectors space. *Proceedings of Interspeech, Firenze, Italy***,** 861-864.

MATARNEH, R., MAKSYMOVA, S., LYASHENKO, V. & BELOVA, N. 2017. Speech recognition systems: A comparative review. *IOSR Journal of Computer Engineering,* 19**,** 71 - 79.

MATEJKA, P., SCHWARZ, P., CERNOCKý, J. & CHYTIL, P. Phonotactic language identification using high quality phoneme recognition. Interspeech, September 2005. 2237-2240.

MENDOZA, S., GILLICK, L., ITO, Y., LOWE, S. & NEWMAN, M. Automatic language identification using large vocabulary continuous speech recognition. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, 1996. IEEE, 785-788.

MOHANTY, S. & SWAIN, B. K. 2010. Language identification using support vector machine. *Proceedings of OCOCOSDA-2010, Nepal*.

MONTAG, C., BŁASZKIEWICZ, K., SARIYSKA, R., LACHMANN, B., ANDONE, I., TRENDAFILOV, B., EIBES, M. & MARKOWETZ, A. 2015. Smartphone usage in the 21st century: who is active on WhatsApp? *BMC research notes,* 8**,** 331.

MUTHUSAMY, Y. K. 1993. A segmental approach to automatic language identification, PhD Thesis. *Oregon Graduate Institue of Science and Technology*.

MUTHUSAMY, Y. K., BERKLING, K. M., ARAI, T., COLE, R. A. & BARNARD, E. A comparison of approaches to automatic language identification using telephone speech. EUROSPEECH, 1993. 1307-1310.

MUTHUSAMY, Y. K., COLE, R. A. & OSHIKA, B. T. The Ogi multi-language telephone speech corpus. ICSLP, 1992. 895-898.

MYERS-SCOTTON, C. 2017. Code-switching. *In:* COULMAS, F. (ed.) *The handbook of sociolinguistics.* 3 ed.: Blackwell Publishing Ltd.

NAGATA, K. K., YASUO; CHIBA, SEIBI 1964. Spoken Digit Recognizer for the Japanese Language. *NEC Res. Develop.,* 12**,** 336,338, 340, 342.

NAJAFIAN, M., HSU, W.-N., ALI, A. & GLASS, J. Automatic speech recognition of Arabic multi-genre broadcast media. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2017. IEEE, 353-359.

NAKAYAMA, S., KANO, T., DO, Q. T., SAKTI, S. & NAKAMURA, S. Japanese-english code-switching speech data construction. Proc. of Oriental COCOSDA, 2018.

NAVRATIL, J. 2001. Spoken language recognition-a step toward multilinguality in speech processing. *IEEE Transactions on Speech and Audio Processing,* 9**,** 678-685.

NG, R. W., LEUNG, C.-C., LEE, T., MA, B. & LI, H. Prosodic attribute model for spoken language identification. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010. IEEE, 5022-5025.

NIST. 2017a. *Language Recognition Evaluation Plans* [Online]. National Institute for Standard and Technology, Department of Commerce,. Available: https://www.nist.gov/itl/iad/mig/language-recognition [Accessed 9/10/2017 2017].

NIST. 2017b. *NIST 2017 Language Recognition Evaluation* [Online]. National Institute for Standard and Technology, Department of Commerce. Available: https://www.nist.gov/itl/iad/mig/nist-2017-language-recognition-evaluation [Accessed 9/10/2017 2017].

PANAYOTOV, V., CHEN, G., POVEY, D. & KHUDANPUR, S. Librispeech: an ASR corpus based on public domain audio books. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015. IEEE, 5206-5210.

PASCALE FUNG, T. S. 2008. Multilingual Spoken Language Processing. *IEEE Signal processing magazine*.

PELLEGRINO, F. & ANDRE-OBRECHT, R. 2000. Automatic language identification: an alternative approach to phonetic modelling. *Signal Processing,* 80**,** 1231-1244.

PENAGARIKANO, M., VARONA, A., RODRíGUEZ-FUENTES, L. J. & BORDEL, G. Using cross-decoder phone coocurrences in phonotactic language recognition. Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, 2010. IEEE, 5034-5037.

PENAGARIKANO, M., VARONA, A., RODRíGUEZ-FUENTES, L. J. & BORDEL, G. 2011. Improved modeling of cross-decoder phone co-occurrences in SVM-based phonotactic language recognition. *IEEE Transactions on Audio, Speech, and Language Processing,* 19**,** 2348-2363.

PREETI SAINI & KAU, P. 2013. Automatic Speech Recognition: A review *International Journal of Engine ering Trends and Technology,* 14**,** 132 -136.

RALLABANDI, S., SITARAM, S. & BLACK, A. W. Automatic Detection of Code-switching Style from Acoustics. Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching, 2018. 76-81.

RAMUS, F. & MEHLER, J. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *The Journal of the Acoustical Society of America,* 105**,** 512-521.

REYNOLDS, D. 2015. Gaussian mixture models. *Encyclopedia of biometrics***,** 827-832.

RICHARDSON, F., REYNOLDS, D. & DEHAK, N. 2015. A unified deep neural network for speaker and language recognition. *arXiv preprint arXiv:1504.00923*.

RODRIGUEZ-FUENTES, L. J., BRUMMER, N., PENAGARIKANO, M., VARONA, A., BORDEL, G. & DIEZ, M. The albayzin 2012 language recognition evaluation. INTERSPEECH, 2013. 1497-1501.

ROUAS, J.-L., FARINAS, J. & PELLEGRINO, F. Automatic modelling of rhythm and intonation for language identification. 15th International Congress of Phonetic Sciences (15th ICPhS), 2003. 567-570.

ROUAS, J.-L., FARINAS, J., PELLEGRINO, F. & ANDRE-OBRECHT, R. 2005. Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication,* 47**,** 436-456.

SAINATH, T. N., WEISS, R. J., WILSON, K. W., LI, B., NARAYANAN, A., VARIANI, E., BACCHIANI, M., SHAFRAN, I., SENIOR, A. & CHIN, K. 2017. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 25**,** 965-979.

SAINI, P. & KAUR, P. 2013. Automatic speech recognition: A review. *International Journal of Engineering Trends and Technology,* 4**,** 1-5.

SAKAI & TOSHIYUKI 1961. The Phonetic Typewriter : Its Fundamentals and Mechanism. *Studia phonologica,* 1**,** 140-152.

SAWALHA, M., BRIERLEY, C. & ATWELL, E. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning (version 2.0). Proceedings of LRE-Rel 2: 2nd Workshop on Language Resource and Evaluation for Religious Texts, LREC 2014 post-conference workshop 31st May 2014, Reykjavik, Iceland, 2014. The University of Leeds, 42-47.

SCHULTZ, T., ROGINA, I. & WAIBEL, A. Experiments with LVCSR based language identification. proc. ICASSP, 1995.

SCHULTZ, T., ROGINA, I. & WAIBEL, A. LVCSR-based language identification. Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, 1996. IEEE, 781-784.

SELTZER, M. L., YU, D. & WANG, Y. An investigation of deep neural networks for noise robust speech recognition. 2013 IEEE international conference on acoustics, speech and signal processing, 2013. IEEE, 7398-7402.

SHENTAL, N., BAR-HILLEL, A., HERTZ, T. & WEINSHALL, D. Computing Gaussian mixture models with EM using equivalence constraints. Advances in neural information processing systems, 2004. 465-472.

SIMONS, G. & FENNIG, C. D. 2017. *Ethnologue: Languages of the World, Dallas, Texas: SIL International* [Online]. Available: https://www.ethnologue.com/ethnoblog/gary-simons/welcome-20th-edition [Accessed 22/10/2017 2017].

SINISCALCHI, S. M., REED, J., SVENDSEN, T. & LEE, C.-H. Exploring universal attribute characterization of spoken languages for spoken language recognition. Tenth Annual Conference of the International Speech Communication Association, 2009.

SONG, Y., HONG, X., JIANG, B., CUI, R., MCLOUGHLIN, I. & DAI, L.-R. Deep bottleneck network based i-vector representation for language identification. Sixteenth Annual Conference of the International Speech Communication Association, 2015.

SONG, Y., JIANG, B., BAO, Y., WEI, S. & DAI, L.-R. 2013. I-vector representation based on bottleneck features for language identification. *Electronics Letters,* 49**,** 1569-1570.

SOUFIFAR, M., KOCKMANN, M., BURGET, L., PLCHOT, O., GLEMBEK, O. & SVENDSEN, T. iVector approach to phonotactic language recognition. Twelfth Annual Conference of the International Speech Communication Association, 2011.

SRIRAM, A., JUN, H., GAUR, Y. & SATHEESH, S. Robust speech recognition using generative adversarial networks. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018. IEEE, 5639-5643.

STOLCKE, A., ZHENG, J., WANG, W. & ABRASH, V. SRILM at sixteen: Update and outlook. Proceedings of IEEE automatic speech recognition and understanding workshop, 2011.

THANH, N. C. 2015. The Differences between Spoken and Written Grammar in English, in Comparison with Vietnamese. *Gist Education and Learning Research Journal***,** 138-153.

TONG, R., MA, B., LI, H. & CHNG, E. S. 2009. A target-oriented phonotactic front-end for spoken language recognition. *IEEE transactions on audio, speech, and language processing,* 17**,** 1335-1347.

TORRES-CARRASQUILLO, P. A., GLEASON, T. P. & REYNOLDS, D. A. Dialect identification using Gaussian mixture models. ODYSSEY04-The Speaker and Language Recognition Workshop, 2004.

TORRES-CARRASQUILLO, P. A., REYNOLDS, D. A. & DELLER, J. R. Language identification using Gaussian mixture model tokenization. Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, 2002a. IEEE, I-757-I-760.

TORRES-CARRASQUILLO, P. A., SINGER, E., GLEASON, T., MCCREE, A., REYNOLDS, D. A., RICHARDSON, F. & STURIM, D. The MITLL NIST LRE 2009 language recognition system. Acoustics Speech and Signal

Processing (ICASSP), 2010 IEEE International Conference on, 2010a. IEEE, 4994-4997.

TORRES-CARRASQUILLO, P. A., SINGER, E., GLEASON, T., MCCREE, A., REYNOLDS, D. A., RICHARDSON, F. & STURIM, D. The MITLL NIST LRE 2009 language recognition system. 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 2010b. IEEE, 4994-4997.

TORRES-CARRASQUILLO, P. A., SINGER, E., KOHLER, M. A., GREENE, R. J., REYNOLDS, D. A. & DELLER JR, J. R. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features. Interspeech, 2002b.

TURRINI, S. 2018. Multidimensional Change in Sudan (1989-2011): Reshaping Livelihoods, Conflicts and Identities. *African Studies Quarterly,* 17**,** 119-120.

VAN DER WESTHUIZEN, E. & NIESLER, T. 2016. Automatic speech recognition of English-isiZulu code-switched speech from South African soap operas. *Procedia Computer Science,* 81**,** 121-127.

VERDET, F. 2011. *Exploring variabilities through factor analysis in automatic acoustic language recognition.* PhD Thesis, University of Fribourg.

VERVERIDIS, D. & KOTROPOULOS, C. 2006. Emotional speech recognition: Resources, features, and methods. *Speech communication,* 48**,** 1162-1181.

VINTSYUK, T. K. 1968. Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis,* 4**,** 52-57.

VU, N. T., LYU, D.-C., WEINER, J., TELAAR, D., SCHLIPPE, T., BLAICHER, F., CHNG, E.-S., SCHULTZ, T. & LI, H. A first speech recognition system for Mandarin-English code-switch conversational speech. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012. IEEE, 4889-4892.

WU, C.-H., CHIU, Y.-H., SHIA, C.-J. & LIN, C.-Y. 2006. Automatic segmentation and identification of mixed-language speech using delta-BIC and LSA-based GMMs. *IEEE Transactions on audio, speech, and language processing,* 14**,** 266-276.

Y. LIU, E. A. Structural metadata research in the EARS program. ICASSP, 2005.

YAN, Y. & BARNARD, E. An approach to automatic language identification based on language-dependent phone recognition. Acoustics, Speech, and Signal

Processing, 1995. ICASSP-95., 1995 International Conference on, 1995. IEEE, 3511-3514.

YEH, C.-F. & LEE, L.-S. 2015. An improved framework for recognizing highly imbalanced bilingual code-switched lectures with cross-language acoustic modeling and frame-level language identification. *IEEE Transactions on Audio, Speech, and Language Processing,* 23**,** 1144-1159.

YILMAZ, E., VAN DEN HEUVEL, H. & VAN LEEUWEN, D. 2016. Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. *Procedia Computer Science,* 81**,** 159-166.

YIN, B., AMBIKAIRAJAH, E. & CHEN, F. Voiced/unvoiced pattern-based duration modeling for language identification. Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on, 2009. IEEE, 4341-4344.

ZAZO, R., LOZANO-DIEZ, A., GONZALEZ-DOMINGUEZ, J., TOLEDANO, D. T. & GONZALEZ-RODRIGUEZ, J. 2016. Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks. *PloS one,* 11**,** e0146917.

ZISSMAN, M. A. 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on speech and audio processing,* 4**,** 31.

Appendix A: Mixed Sentences Database

No.	Mixed Sentence	Words List	Diacritics Word	Language tag
1	ال file في الدولاب	ال	اَل	A
		file	فَايل	E
		في	فْي	A
		الدولاب	الدُولاَب	A
2	حفظتو في ال desktop	حفظتو	حَفَظْتُو	A
		desktop	دِيسكِتوب	E
3	مشيت ليو already	مشيت	مَشيت	A
		ليو	ليو	A
		already	أُرَّادي	E
4	I think نغي الموضوع	I	أي	E
		think	سِنك	E
		لغي	لْغِي	A
		الموضوع	المَوْضُوع	A
5	عندنا presentation	عندنا	عِنْدَنا	A
		presentation	بِرِذِنْتيشَن	E
6	انا في meeting	انا	أَنَا	A
		في	فِي	A
		meeting	مِيتِنق	E
7	بديت training	بديت	بَدِيت	A
		training	تِرينْيق	E
8	املاها full	املاها	أَمْلَاها	A
		full	فُلْ	E
9	اديك box	اديك	أَدِيْك	A
		box	بُكْسْ	E

10	انت زول cute	إنت	إنْتَ	A
		زول	زوُل	A
		cute	كِيُوتْ	E
11	نجمع share	نجمع	نَجْمَع	A
		share	شيْر	E
12	تمام يا man	تمام	تَمَام	A
		يا	يَا	A
		man	مَّان	E
13	الكلام ده serious	الكلام	الْكَلَام	A
		ده	دَه	A
		serious	صِيرْيَص	E
14	قطعة single	قطعة	قِطْعَة	A
		single	سِنْقِل	E
15	اعمل لي discount	اعمل	أعْمِل	A
		لي	لَي	A
		discount	دِسْكاوٌنْت	E
16	شيل copy	شيل	شِيِل	A
		copy	كُوبِي	E
17	هو floater	هو	هُو	A
		floater	فْلُوتَر	E
18	معقول impossible	معقول	مَعَقُول	A
		impossible	إمْبُسِبُّل	E
19	ماشة ال lab	ماشة	مَاشَّة	A
		ال	ألْ	A
		lab	لاب	E
20	عجيب tension	tension	تنْشَن	E
		عجيب	عَجِيب	A
21	خالص prestige	prestige	بِرِسْتِيج	E
		خالص	خَالِص	A

22	ياخ wicked	wicked	وِكِدْ	E
		ياخ	يَاخْ	A
23	البرنامج ماظابط code	code	كُودْ	E
		البرنامج	البَرْنَامِج	A
		ما	مَا	A
		ظابط	ظَابِط	A
24	دخلت في stress	دخلت	دَخَلْتَ	A
		في	فِي	A
		stress	إسْتْرس	E
25	انا obsessed جدا بالمسلسل	انا	أنَا	A
		obsessed	أبْست	E
		جدا	جِدّاً	A
		بالمسلسل	بِالْمُسَلْسَل	A
26	نحن ما interested فيها	نحن	نِحْنَ	A
		ما	مَا	A
		interested	إنْتَرِسْت	E
		فيها	فِيْهَا	A
27	محتاجة motivationعشان اذاكر	محتاجة	مُحْتَاجَة	A
		motivation	مُتِفِيشَن	E
		عشان	عَشَانْ	A
		اذاكر	أذَاكِّر	E
28	انتي copy paste من امك	انتي	إنْتِي	A
		copy	كُبِي	E
		paste	بيسْت	E
		من	مَنْ	A
		امك	أمِك	A

29	ما عندي اي idea	ما	مَا	A
		عندي	عِنْدِي	A
		اي	أيّ	A
		idea	أيدِيَا	E
30	عايزين delivery	عايزين	عَايْزِيْن	A
		delivery	دِلِفَرِي	E
31	نعمل sharing	نعمل	نَعْمَل	A
		sharing	شيررِنْق	E
32	ما تعمل لينا headache	ما	مَا	A
		تعمل	تَعْمَل	A
		لينا	لِينَا	A
		headache	هِيدِك	E
33	وين cheque المدارس	وين	وِيَنْ	A
		cheque	شيك	E
		المدارس	الْمَدَارّس	A
34	قون double kick	قون	قُون	A
		double	دَبُلْ	E
		kick	كَيكْ	E
35	ظاهر Offside	offside	أوفْسَايت	E
		ظاهر	ظَاهِر	A
36	تمام Third back	third	سيرْ	E
		back	بَاك	E
		تمام	تَمَام	A
37	قاعد corner	قاعد	قَاعِد	A
		corner	كُورنَر	E
38	قوي Defense	defense	دَفينْس	E
		قوي	قَوِي	A

39	لعب through pass حلو	لعب	لِعِبْ	A
		through	سْرُّو	E
		Pass	بَاص	E
		حلو	حِلْو	A
40	الكورة foul	الكورة	الْكُورَة	A
		foul	فَاوُل	E
41	lineman4 رافع الراية	lineman	لايِنْمَان	E
		رافع	رَافِعْ	A
		الراية	الْرَايَة	A
42	خلاص finish	خلاص	خَلَاس	A
		finish	فِنِشْ	E
43	block تاني ناصية	block	بْلُكْ	E
		تاني	تَانِي	A
		ناصية	نَاصِيَة	A
44	باني grid beam	باني	بَانِي	A
		grade	قِرِيْد	E
		beam	بِيْم	E
45	بقيت nervous	بقيت	بِقِيت	A
		Nervous	نيرْفَص	E
46	حضرت party الشركة	حضرت	حَضَرْتَ	A
		party	بَارْتِي	E
		الشركة	الْشَرِكَة	A
47	الاجتماع cancelled	الاجتماع	الإجْتِمَاع	A
		cancelled	كانْصِلض	E
48	اشتريت T-shirt	إشتريت	إشْتَرِيْت	A
		T-shirt	تِي شيرْت	E
49	خيت skirt جميل	خيت	خَيَّتَ	A
		skirt	إسْكيرْت	E
		جميل	جَمِيّل	A

50	jacket السفر	jacket	جَكِت	E
		السفر	ألْسَفَر	A
51	رسمت view	رسمت	رَسمْتَ	A
		view	فِيُو	E
52	remote الشاشة تحت	remote	رِمُوتْ	E
		الشاشة	ألْشَاشَة	A
		تحت	تِحِت	A
53	شلت screenshot	شلت	شِلْتَ	A
		screenshot	إسْكِرِينشُط	E
54	كتبت hand over	كتبت	كَتَبْتَ	A
		hand	هَانْد	E
		over	أفَرْ	E
55	Film السهرة كان شنو	film	فِلِم	E
		السهرة	ألْسَهْرَة	A
		كان	كَان	A
		شنو	شِنُو	A
56	mall الواحة	mall	مُوْل	E
		الواحة	ألْوَاحَة	A
57	بيت loadbearing	بيت	بيت	A
		loadbearing	لُودْبيررنْق	E
58	تسجيل stereo	تسجيل	تَسْجِيلْ	A
		stereo	إسْتِريو	E
59	تصوير studio	تصوير	تَصْوِير	A
		studio	إسْتِديو	E
60	ركبت sound system	ركبت	رَكَّبْتَ	A
		sound	سَّاوون	E
		system	سيسْتِم	E

61	وصلت Cable الجهاز	وصلت	وَصَّلْتَ	A
		cable	كيبِّلْ	E
		الجهاز	الْجِهاز	A
62	الثلاجة عاوزة test	التلاجة	التَلاجَّه	A
		عاوزة	عَاوْزَة	A
		test	تيسْت	E
63	بوكس double cabin	بوكس	بُوكْس	A
		double	دَبِّل	E
		cabin	قَبِينَه	E
64	عربية four wheel drive	عربية	عَرَبِيّة	A
		four	فور	E
		wheel	وِيلْ	E
		drive	درَايِف	E
65	الطيارة captain	captain	كَبْتِن	E
		الطيارة	الْطَيّارة	A
66	البيت فصل freezer	freezer	فِرِيزَر	E
		البيت	الْبيْت	A
		فصل	فَصَلْ	A
67	الشاي جاهز cake	cake	كيْك	E
		الشاي	الْشَّاي	A
		جاهز	جَاهِز	A
68	راجل gentle	راجل	رَاجِل	A
		gentle	جَنْتِل	E
69	انيقة لكن المكياج over	انيقة	أنِيقَة	A
		لكن	لَكِنْ	A
		المكياج	الْمِكياج	A
		over	أووفَر	E

70	عملت make up	عملت	عَمَلْتَ	A
		make	ميِيْك	E
		up	أبْ	E
71	خمسة busy line	line	لاِين	E
		خمسة	خَمْسَة	A
		busy	بِذِي	E
72	رفعت mode الفصل	رفعت	رَفَعْتَ	A
		mode	مُوود	E
		الفصل	الْفَصِل	A
73	عندو villa رهيبة	عندو	عِنْدُو	A
		villa	فِيلْلا	E
		رهيبة	رَهِيبَة	A
74	board الجامعة ملان	board	بُورْد	E
		الجامعة	الْجَامْعَة	A
		ملان	مَلَان	A
75	حصل misunderstanding	حصل	حَصَلْ	A
		misunderstanding	مِسْانْدَرإِسْتَانْدِنق	E
76	حدد deadline	حدد	حَدِد	A
		deadline	ديدْلاِين	E
77	ب side الغرفة	ب	ب	A
		side	صَايِد	A
		الغرفة	الْغُرْفَة	A
78	تلفون من غير camera	تلفون	تَلَفُون	A
		من	مِنْ	A
		غير	غيْر	A
		camera	كَمِيرَا	E
79	internet الشركة قاطع	internet	إِنْتَرنِت	E
		الشركة	الْشَرِكَة	A
		قاطع	قَاطِع	A

80	Wi-Fi القاعة منتهي	Wi-Fi	وَاي فَاي	E
		القاعة	الْقَاعَة	A
		منتهي	مُنْتَهِي	A
81	ادينا example	ادينا	أدّينَا	A
		example	إِقْزَامْبُل	E
82	جد useful	جد	جَدْ	A
		useful	يُوذْفُل	E
83	في data مفيدة	في	فِى	A
		data	دَاتَا	E
		مفيدة	مُفِيدَة	A
84	عايز كم جملة database	عايز	عَايِز	A
		كم	كَمْ	A
		جملة	جُمْلَة	A
		database	دَاتَابيزْ	E
85	بتاع show	بتاع	بِتَاعْ	A
		show	شوْ	E
86	نشيل takeaway	نشيل	نَشِيل	A
		takeaway	تيكْ أوَيي	E
87	عندو case معاهو	عندو	عِنْدُ	A
		case	كيسْ	E
		معاهو	مَعَاهُو	A
88	سجلت minutes لاجتماع القسم	سجلت	سَجَلْتَ	A
		minutes	مِنِتْس	E
		لاجتماع	لِإجْتِمَاع	A
		القسم	الْقِسِم	A
89	المحاضرات محتاجة memory اكبر	المحاضرات	الْمُحَاضَرَات	A
		محتاجة	مُحْتَاجَة	A
		memory	ميمُورِي	E
		اكبر	اكْبَرْ	A

90	اخترت track صح	اخترت	إخْتِرتَ	A
		track	تَرَاكْ	E
		صح	صَحْ	A
91	حضرت lecture مفيد	حضرت	حَضَرتَ	A
		lecture	ليكْشَر	E
		مفيد	مُفِيد	A
92	رسلت ليك SMS	رسلت	رَسَّلْتَ	A
		ليك	ليكْ	A
		SMS	إسيميسْ	E
93	كتبت mail المشرف	كتبت	كَتَبْتَ	A
		mail	ميلل	E
		المشرف	الْمُشْرِف	A
94	نزلت post في صفحتي	نزلت	نَزَّلْتَ	A
		post	بُوسْت	E
		في	فِي	A
		صفحتي	صَفْحَتِي	A
95	طلعتها من google	طلعتها	طَلَعْتَهَا	A
		من	مِنْ	A
		google	قُوقِل	E
96	رسلت ليك WhatsApp	رسلت	رَسَلْتَ	A
		ليك	ليك	A
		WhatsApp	وَاتْسَآب	E
97	Twitter مستخدم اكتر	Twitter	تُوِتَر	E
		مستخدم	مُسْتَخْدَم	A
		أكتر	أكْتَر	A
98	ياسلام new look	ياسلام	ياسَلام	A
		new	نِيو	E
		look	لُكْ	E

99	عايزين chips	عايزين	عَايْزِين	A
		chips	جِّبْسْ	E
100	اشتريت iPhone	اشتريت	إشْتَرِيتْ	A
		iPhone	آيفُون	E
101	حصل displacement	حصل	حَصَلْ	A
		displacement	دِسْبِلِيس	E
102	القصة ignored	ignored	إِقْنُورد	E
		القصة	الْقِصَة	A
103	المسألة to manage	to	تُو	E
		manage	مَنِجْ	E
		المسألة	المَسْأَلَة	A
104	عربيتي وقف Starter	starter	إسْتَارتَر	E
		عربيتي	عَرَبِتِي	A
		وقف	وَقَفْ	A
105	غيرت ليها زيت gearbox	غيرت	غَيَّرتَ	A
		ليها	لِيهَا	A
		زيت	زيْت	A
		gearbox	جَرَبُوكْس	E
106	حجزت في bus الاحد	حجزت	حَجَزْتَ	A
		في	فِي	A
		bus	بَصْ	E
		الاحد	الْاحَد	A
107	نظامه ما updated	نظامه	نِظَامُه	A
		ما	مَا	A
		updated	أبْديتِد	E
108	عندك antivirus	عندك	عِنْدَك	A
		antivirus	أنْتِيفايْرَس	E

109	أمشي step لقدام	أمشي	أمْشِي	A
		step	اسْتيب	E
		لقدام	ليقِدَام	A
110	ما عندو excuse	ما	مَا	A
		عندو	عِنْدُو	A
		excuse	إكْسِكْيُوز	E
111	ضربنا chicken	ضربنا	ضَرَبْنَا	A
		chicken	جِكِن	E
112	الاسبوع كله standby	الإسبوع	الإسْبُوع	A
		كله	كُلُو	A
		standby	اسْتَانْدباي	E
113	ناخد rest شوية	ناخذ	ناخُد	A
		rest	ريسْت	E
		شويه	شوَيّه	A
114	عندك tissues	عندك	عِنْدَك	A
		tissue	تِشُوذ	E
115	بقت issue	بقت	بِقَت	A
		issue	إيشو	E
116	انت sure إنو جاء	إنت	إنْتَ	A
		sure	شُور	E
		إنو	إنُو	A
		جاء	جَا	A
117	مافي مشكلة so far	مافي	مافِي	A
		مشكلة	مُشْكِلَة	A
		so	سُو	E
		far	فار	E
118	Still واقف	still	إسْتِل	E
		واقف	وَاقِفْ	A

119	later on نتناقش	نتناقش	نَتْنَاقَش	A
		later	لِيْتَر	E
		on	أنْ	E
120	see you يلا	يلا	يَلْا	A
		see	سِي	E
		you	يُو	E
121	anyway كلمتو	anyway	ايْنِوي	E
		كلمتو	كَلَمْتُو	A
122	Let us say سافر	let	لِيْتَ	E
		us	أصْ	E
		say	سيْى	E
		سافر	سَافَر	A
123	longtime ما شفناكم	ما	مَا	A
		شفناكم	شُفْنَاكُم	A
		longtime	لُنْتَتَايِم	E
124	around بكون	بكون	بِكُونْ	A
		around	أرَاوُنْد	E
125	message لقيت منك	لقيت	لِقِيتْ	A
		منك	مِنَاكْ	A
		message	مَسِيْج	E
126	mature enough أفتكر إنو	أفتكر	أفْتَكِر	A
		إنو	إنُّو	A
		mature	مَاتْيُور	E
		enough	إِنَفْ	E
127	decision أخت خلاص	أخت	أخَتَّ	A
		decision	دِسِشَّنْ	E
		خلاص	خَلَاس	A
128	relax خليك	خليك	خَلِيْك	A
		relax	رِيلَاكْس	E

129	الموضوع بسيط cool down	الموضوع	المَوْضُوع	A
		بسيط	بَسِيط	A
		cool	كُوُلْ	E
		down	دَاوُنْ	E
130	ارجع لي urgently	ارجع	أرْجَعْ	A
		لي	لَي	A
		urgently	إيْرجَنْتْلِي	E
131	خليك clear	خليك	خَلِيكْ	A
		clear	كِلِيرْ	E
132	منتظر response	منتظر	مُنْتَظِر	A
		response	رِسْبُونْص	E
133	رسل location للمحل	رسل	رَسِل	A
		location	لُوكيشَن	E
		للمحل	لْلمَحَل	A
134	خلاص fine	خلاص	خَلَاس	A
		fine	فايْن	E
135	حصل فيها damage	حصل	حَصَلْ	A
		فيها	فِيهَا	A
		damage	دَمِيج	E
136	حاجة funny	حاجة	حَاجَه	A
		funny	فَنِي	E
137	القصة بقت complicated	القصة	الْقِصَة	A
		بقت	بِقَتْ	A
		complicated	كُمبِلِكيتِد	E
138	Unbelievable	unbelievable	أنْبِلِيفَابْل	E
139	اشتريت lab coat	إشتريت	إشْتَرِيْت	A
		lab	لَابْ	E
		coat	كُوتْ	E

140	اديهو call	اديهو	أدِيهُو	A
		call	كوُل	E
141	كباية double لو سمحت	كباية	كُبَايَة	A
		double	دَبُّل	E
		لو	لَوْ	A
		سمحت	سَمَحْتَ	A
142	رفعت mode الفصل	رفعت	رَفَعْتَ	A
		mode	مُوودْ	E
		الفصل	الْفَصِل	A
143	اعمل like	اعمل	أعْمِل	A
		like	لايِك	E
144	البرنامج فيو zooming	البرنامج	الْبَرْنَامِجْ	A
		فيو	فِيْو	A
		zooming	زُومِينْق	E
145	شغل mice المسجد	شغل	شَغِّلْ	A
		mice	مَايِك	E
		المسجد	الْمَسْجِد	A
146	الماسورة فيها leak	الماسورة	الْمَاسُورة	A
		فيها	فِيهَا	A
		leak	لِيك	E
147	مشغول Line one	line	لاين	E
		one	وَن	E
		مشغول	مَشْغُول	A
148	انتهينا stop	انتهينا	إنْتَهينَا	A
		stop	أسْطُبْ	E
149	مخصوص parking	parking	بَارْكِنْق	E
		مخصوص	مَخْصُوص	A

150	مشيت في runway طويل	مشيت	مَشْيْت	A
		في	فِي	A
		runway	رَنْويْي	E
		طويل	طَوِييل	A
151	اكتشفت نوع mobile رهيب	اكتشفت	إكْتَشَفْتَ	A
		نوع	نُووع	A
		mobile	مُوبَايِل	E
		رهيب	رَهِيْب	A
152	دخل من turn غلط	دخل	دَخَلْ	A
		من	مِنْ	A
		turn	تيْرن	E
		غلط	غَلَط	A
153	case study مكتملة	case	كيْس	E
		study	إصطَدِي	E
		مكتملة	مُكْتَمْلَة	A
154	خزنت في base العمارة	خزنت	خَزَّنْتَ	A
		في	فِي	A
		base	بييز	E
		العمارة	الْعَمَارَة	A
155	لقيت miss call	لقيت	لِقِيْت	A
		miss	مِسْ	E
		call	كُولْ	E
156	Microphone المسجد معطل	microphone	مَكْرَفُون	E
		المسجد	الْمَسْجِد	A
		معطل	مَّعَطّلْ	A
157	في Co-patient صبور	في	أنَا	A
		Co-patient	كُوبيشَنْت	A
		صبور	صَبُورْ	A

158	حولت patient من الطوارئ	حولت	حَوَلْتَ	A
		patient	بيشَنْت	E
		من	مِنْ	A
		الطوارئ	الطَوَارِئ	A
159	فيها security عالي	فيها	فِيْها	A
		security	سِكْيُرِتي	E
		عالي	عَالِي	A
160	دخلت pin code جديد	دخلت	دَخَّلْتَ	A
		pin	بِنْ	E
		code	كُود	E
		جديد	جَدِيد	A
161	خبرتو zero	خبرتو	خِبْرَتُو	A
		zero	زِيْرُو	E
162	enlarged القصة	enlarged	إنْلَارْجْد	E
		القصة	القِصَة	A
163	كلها requirements مهمة	كلها	كُلَهَا	A
		requirements	رِكْوَيَرْمِينْتس	E
		مهمة	مُهِمَة	A
164	درست Linux	درست	دَرَسْتَ	A
		Linux	لِنِكْس	E
165	وصلو مع server سعتو اكبر	وصلو	وَصِّلْو	A
		مع	مَعَ	A
		server	سِيْرفَر	E
		سعتو	سِعَتُو	A
		اكبر	اكْبَر	A
166	حجزنا في bus متاخر	حجزنا	حَجَزْنَا	A
		في	فِي	A
		bus	بَصْ	E
		متأخر	مُتْأَخِر	A

167	ركبت socket تاني	ركبت	رَكَبْتَ	A
		socket	صُوكِتْ	E
		تاني	تَانِي	A
168	نزلت windows	نزلت	نَزَّلْتَ	A
		windows	وِينْدُوزْ	E
169	دفعت cash	دفعت	دَفَعْتَ	A
		cash	كَاشْ	E
170	وصلت filter موية	وصلت	وَصَّلْتَ	A
		filter	فِلْتَر	E
		موية	مُويَة	A
171	calendar السنة	calendar	كَلِنْدَر	E
		السنة	السَنَة	A
172	حالتو stable	حالتو	حَالْتُو	A
		stable	اسْتِيبُل	E
173	اليوم nice	اليوم	اليُومْ	A
		nice	نايِص	E
174	الموضوع personal	الموضوع	المَوضُوعْ	A
		personal	بيرْسُنَل	E
175	جهز refreshment للإجتماع	جهز	جَهِّز	A
		refreshment	رفْرِيشْمِينْت	E
		للإجتماع	للإجْتِمَاع	A
176	شربت عصير fresh	شربت	شِرِبْتَ	A
		عصير	عَصِير	A
		Fresh	فرِيْش	E
177	مسخن league	League	لِيق	E
		مسخن	مُسَخِن	A
178	جبت keyboard جديد	جبت	جِبْتَ	A
		keyboard	كِيبُورْد	E
		جديد	جَدِيد	A

179	story مدهش	Story	استُورِي	E
		مدهش	مُدهِش	A
180	takeoff الطيارة مريح	Takeoff	تيْكُوف	E
		الطيارة	الطَيَارَة	A
		مريح	مُريح	A
181	net سريع شديد	Net	نيْت	E
		سريع	سَرِيع	A
		شديد	شَدِيد	A
182	scratch جيب معاك	جيب	جِيب	A
		معاك	مَعَاك	A
		Scratch	اسْكِرَاتْش	E
183	result مرضية تماما	Result	رِيزَالْت	E
		مرضية	مُرْضِية	A
		تماما	تَمَامْاً	A
184	completion سمنار	completion	كُومبِلِشَن	E
		سمنار	سِمِنَار	A
185	compile اعمل للبرنامج	اعمل	اعمَل	A
		compile	كومبَايل	E
		للبرنامج	لِلبَرنِامِج	A
186	run النظام	Run	رَن	E
		النظام	انِظَام	A
187	service كويس شديد	service	سيرفِس	E
		كويس	كُوِيس	A
		شديد	شَدِيد	A
188	Facebook شفتها في	شفتها	شُفْتَها	A
		في	في	A
		Facebook	فيْسبُوك	E

189	ارفعها على YouTube	ارفعها	ارْفَعهَا	A
		على	عَلَى	A
		YouTube	يُوتِيوب	E
190	كنا في beach بري	كنا	كُنا	A
		في	في	A
		Beach	بِيتْش	E
		بري	بُرْي	A
191	switch المخزن واقف	Switch	سِوِيتْش	E
		المخزن	المَخزَن	A
		واقف	وَاقِف	A
192	تعال نعمل selfie	تعال	تَعَال	A
		نعمل	نَعمَل	A
		Selfie	سيْلْفِي	E
193	شفتها فيِ Snapchat	شفتها	شُفْتَهَا	A
		في	في	A
		Snapchat	اسنَابشَات	E
194	شوف في port فاتح	شوف	شَوف	A
		في	في	A
		Port	بُورْت	E
		فاتح	فَاتِح	A
195	شغال android	شغال	شَغَال	A
		Android	اندِرُويد	E
196	كبيرة benefits	benefits	بينِيفِتْس	E
		كبيرة	كَبِيرة	A
197	كسرها صعب password	password	بَاسوْيرْد	E
		كسرها	كَسْرَها	A
		صعب	صَعَب	A
198	ياخ thanks	Thanks	سَانْكس	E
		ياخ	يَاخ	A

199	شديد thank you	Thank	سَانكِيو	E
		You	يُو	E
		شديد	شَدِيد	A
200	نعمل order	نعمل	نَعْمَل	A
		Order	أُورْدَر	E
201	في complain حاصل	في	في	A
		complain	كُومبلين	E
		حاصل	حَاصِل	A

Appendix B: Hybrid language phonetic symbols

No.	Orthographic Symbol	Phonetic Symbol	Category
1	أ	E	
2	ب	B	
3	ت	T	
4	ث	TH	
5	ج	J	
6	ح	H-	
7	خ	KH	Consonants
8	د	D	
9	ذ	Z	
10	ر	R	
11	ز	ZA	
12	س	S	
13	ش	SH	
14	ص	S-	
15	ض	D-	Emphatic Consonants
16	ط	T-	No IPA equivalents
17	ظ	Z-	
18	ع	A-	Consonants Distinct sound
19	غ	GH	
20	ف	F	
21	ق	Q-	
22	ك	K	Consonants
23	ل	L	
24	م	M	
25	ن	N	
26	ه	H	

27	ا	A:	
28	و	U:	Long Vowels
29	ي	I:	
30	َ	A	
31	ُ	U	Short Vowels
32	ِ	I	
33	ً	AN	
34	ٌ	UN	Nunation
35	ٍ	IN	
36	ّ	+	
37	َّ	+A	
38	ِّ	+I	Shaddah
39	ُّ	+U	
40	ْ	SK	
41	إ	E:	
42	ء	AH:	
43	آ	AA:	
44	ئ	IY:	Special Symbols
45	ؤ	O:	
46	ة	TA:	
47	ى	YA:	

Appendix C: Part of Mixed Phonetic Dictionary

Word	Phonetic Symbol
file	F A A: I: L
الدولاب	A: L D U U: L A: A B
حفظتو	H- A F A Z- T U U:
desktop	D I I: S K I T U: B
مشيت	M A SH I: T
ليو	L I: U:
already	A: U R S: A A: D I:
I	E I:
think	S I N K
لغي	L SK GH I I:
الموضوع	A: L M A U: SK D- U U: A-
عندنا	A- I N SK D A N A:
presentation	B I R I Z I N SK T I: SH A N
انا	E N A A:
في	F I I:
meeting	M I I: T I N Q
بديت	B A D I I: T
training	T I R I: N SK I: Q
املاها	E A M SK L A A: H A:
full	F UN L SK
اديك	E A D I I: SK K
box	B UN K SK S SK
إنت	E: N SK T A
زول	ZA U: U L
cute	K I I: UN U: T SK
نجمع	N A J SK M A A-
share	SH I: SK R
تمام	T A M A A: M
يا	I: A A:
man	M S: A A: N
الكلام	A: L SK K A L A A: M

Appendix D: Part of Mixed Languages Lexicon

Word	Language Tag
الموضوع	A
عندنا	A
Presentation	E
انا	A
سوكاركي	D
في	A
meeting	E
بديت	A
training	E
املاها	A
full	E
إسنتاد	D
اديك	A
box	E
إنت	A
زول	A
cute	E
نجمع	A
share	E
تمام	A
يا	A
man	E
الكلام	A
ده	A
serious	E
قطعة	A
single	E
اعمل	A
لي	A
discount	E

List of Articles Resulting from this Research

Published Works

i- Mohammed O. Elfahal, M. E. Mustafa " Automatic Languages recognition: Approaches and Challenges", " *Journal of Sudan Arabic Academy,* Vol. 13th 2013.

ii- M. O. Eltayeb, M. Musa, " Acoustic-support vector machines approach to detect spoken Arabic language", *presented at the Computing, Electrical and Electronics Engineering, ICCEEE13, IEEE*, 2013

iii- Mohammed O. Elfahal, M. Musa, R. A. Saeed, "AUTOMATIC SPOKEN LANGUAGE RECOGNITION FOR MULTILINGUAL SPEECH RESOURCES", *"Journal of Theoretical and Applied Information Technology*", vol. 96, p. 15, 2018

Prepared for Publication

i- Model Generlization for Mixed speech recognition – Journal article

ii- Language identification in mixed speech mode – Conference Paper