Investigating Accurate Deep Learning Solutions for Recognizing Arabic Speech in the Jordanian Dialect

By

Sarah Ayman Almadi

Supervisor

Dr. Gheith Ali Abandah, Prof.

This Thesis was Submitted in Partial Fulfillment of the Requirements for the Master's Degree in Computer and Networks Engineering

Faculty of Graduate Studies

The University of Jordan

May, 2024

COMMITTEE DECISION

This Thesis (Investigating Accurate Deep Learning Solutions for Recognizing Arabic Speech in the Jordanian Dialect) was Successfully Defended and Approved on

Examination Committee	**Signature**
Dr. Gheith Abandah, (Supervisor) Prof. of Computer Engineering	----------
Dr. Andraws Swidan, (Member) Prof. of Computer Engineering	----------
Dr. Musa Alyaman, (Member) Assoc. Prof. of Mechatronics Engineering	----------
Dr. Esam alQaralleh, (Member) Prof. of Computer Engineering Princess Sumaya University for Technology	----------

ACKNOWLEDGEMENT

Many thanks to Prof. Gheith Abandah for his guidance, patience, and encouragement. I am grateful for giving me the chance to learn machine learning with him.

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ASR	Automatic Speech Recognition
CNN	Convolutional Neural Network
CER	Character Error Rate
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
PMC	Pulse-Code Modulation
SASSC	The Standard Arabic Single Speaker Corpus
VAD	Voice Activity Detection
WER	Word Error Rate
WADA	Waveform Amplitude Distribution Analysis
WSJ	Wall Street Journal

Investigating Accurate Deep Learning Solutions for Recognizing Arabic Speech in the Jordanian Dialect

By

Sarah Ayman Almadi

Supervisor

Dr. Gheith Ali Abandah, Prof.

ABSTRACT

Automatic speech recognition of Arabic language and its dialects is a challenging task. This is due to the variety of dialects, lack of labeled data, and the morphological complexity. Several machine learning techniques have been used to enhance speech recognition of Arabic language such as recurrent neural networks ~~(RNN)~~, machine learning models as ~~Hidden~~ hidden Markov ~~Models~~models, and convolutional neural networks.

The recent usage of self-supervised learning for speech recognition is providing great results. In this thesis, we use self-supervised learning to enhance a speech recognition system of the Jordanian dialect. We investigated two state-of-the-art models built by Meta AI company: Wav2vec2.0 wrapper model and Wav2vec2.0-XLSR model. These models were trained and fine-tuned on Jordanian dialect dataset from the Massive Arabic Speech Corpus (MASC).

Due to the lack of enough labeled data for the Jordanian dialect, the Wav2vec2.0 wrapper model that is trained on the available labeled Jordanian dialect was unable to correctly recognize the Jordanian dialect. The Wav2vec2.0-XLSR model is pretrained on 53 languages and is pretrained on the similarity between languages. This model performs better than the first model when fine-tuned on the available Jordanian dialect dataset. We

have fine-tuned this model on two dataset sizes and evaluated the fine-tuned model using the word error rate (WER) metric on a Jordanian dialect test dataset. When fine-tuning on 20-hour and 35-hour datasets, the Wav2vec2.0-XLSR model achieves 25% and 23% WER, respectively. This is a good improvement compared with the model version that is fine-tuned on a dataset of Modern Standard Arabic ~~(MSA)~~, which achieves 29% WER on the test dataset. Additionally, the model fine-tuned on the longer Jordanian dialect dataset subjectively provides good recognition of the Jordanian dialect speech.

Chapter 1
Introduction

Automatic speech recognition (ASR) is the process of converting a direct speech from microphone or an audio file to text (Malik *et al*., 2021). ASR systems are of major interest for research since speech is the most preferred method of communication between humans (Alharbi *et al*., 2021).

Since the development of ASR systems began, it has helped a lot of people to connect and communicate with each other. They are being used widely in computer assistance systems, such as Siri in Apple systems and Google Assistant. ASR systems are also used in developing language translation applications such as Google translation application, social media applications (WhatsApp, Facebook Messenger, *etc.*), gaming and many other systems.

1.1 Automatic Speech Recognition Background

Previously, in early 2000s, feed forward artificial neural networks (ANN) were used in combination with hidden Markov model (HMM) to reproduce human language and speech. Recently, recurrent neural networks and combination of deep learning techniques are used to build ASR systems (Malik *et al*., 2021).

ASR process, as shown in Figure 1, starts with processing the audio signals which includes cleaning the audio signal from unwanted noises, detecting speech and length of the audio which will make the ASR more efficient (Labied *et al*., 2022). The next step is feature extraction that is used to get useful speech components for the linguistic content (Labied *et al*., 2021), and finally neural networks are typically used for recognizing the speech.

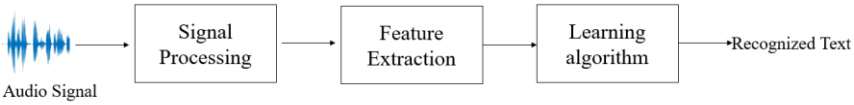

Figure 1. The process of a speech recognition system

1.2 Challenges and Limitations in ASR

Several limitations and challenges face the research while building ASR systems, such as variation of accents, also ASR systems are almost never designed to be used with children, Noises (Basak *et al*., 2023).

There is lot of research aimed for ASR for many languages, but with lack for Arabic language (Alsayadi *et al.,*2022), this lack is observed due to the challenges which are not observed in other languages. According to (Abdou *et al*., 2019) Arabic language challenges can be summarized as the variety of dialects, morphological complexity, the dominance of non-diacritized text, furthermore the lack of training data for dialectical Arabic and the variety of the dialects in the same country as sometimes people in a village speaks different dialects than in city, also the dialect Arabic lacks Arabic writing standards.

As Arabic dialects are one of the major challenges for Arabic ASR systems, there are active studies for different Arabic dialects using machine learning algorithms to enhance its performance, but it is observed that nearly none of these studies is dedicated for Jordanian dialect which was the encouragement to start a basis for Arabic ASR system for Jordanian dialect.

1.3 Thesis Objectives

The main objectives of this research are to discuss recent deep learning techniques for ASR, study recent research for dialect ASR systems and enhance an ASR model for Jordanian dialect.

1.4 Thesis Questions

This thesis aims to answer the following questions:

- What is the most efficient machine learning algorithm to be used to enhance ASR system for Jordanian dialect?
- Is it possible to achieve high accuracy of recognizing Jordanian Arabic on Arabic speech recognition systems using self supervised learning?
- Does self supervised learning overcome the issue of lack of data for Arabic language?

1.5 Thesis Contribution

People across Jordan speak different dialects in villages and their words may differ from governorate to another, this may decrease the efficiency of recognizing the speech of Jordanian people. In this thesis, Meta models (will be explained further in next chapters) are adopted to enhance the speech recognition process for Jordanian dialect. The model is trained from scratch on Jordanian data and fine-tuned on another pretrained model to study and get the best performance of the approaches.

1.6 Thesis Methodology

The methodology steps are the following:

- **Survey of related work**: review the related work that deals with ASR systems for Arabic and other languages.
- **Evaluation of related work**: show the results of the related work and outline the tools that been used.
- **Dataset gathering and preparation**: prepare the dataset namely MASC, extract the Jordanian dialect data and preprocess it to be usable for our model.
- **Applying Meta model to recognize Arabic speech**: prepare the model and apply it to Jordanian data.
- **Documentation**: document our work and show the results.

1.7 Thesis Outline

This thesis is structured as follows: Chapter 2 reviews the related work for speech recognition for different languages and models. Chapter 3 presents the Meta models used in this thesis in details. Chapter 4 describes the data pretraining and processing along with the fine-tuning. Chapter 5 provides the results of the experiments. Finally, Chapter 6 summarizes the conclusions and future work.

Chapter 2

Literature Review

In this chapter, the previous research related to ASR for the Arabic language and other languages is discussed. Most of the previous research used supervised learning techniques for ASR systems and recently unsupervised learning is being used for ASR systems.

2.1 ASR Systems Previous Studies

Masmoudi *et al*. (2018) presented the features and the linguistic characteristics of the Tunisian dialect in different levels such as phonological, morphological, syntactic, and lexical, also the authors also created TARIC corpus which was labeled manually and used for ASR experiments using Kaldi tool kit and they have achieved 22% WER.

SpecAgment is data augmentation technique that was developed for speech recognition, this augmentation technique is applied to the neural network feature inputs, three distortions were used to create the augmentation policy, first is time wrapping via sparse image warp function in TensorFlow and second distortion is frequency masking which is applied to mask the consecutive Mel frequency channels, lastly time masking is applied to mask the consecutive time steps, listen, attend and spell (LAS) network was used for the purpose of testing the effect of the of the data augmentation technique on the ASR system, upon testing it was shown that SpecAgment greatly enhances the performance of speech recognition by approximately 10% when comparing to other work done on LAS model for the same used dataset (Park *et al*., 2019).

Schneider *et al*. (2019) pretrained a multi-layer convolutional neural network on large amount of unlabeled audio data to compute representations that can be used as input to speech recognition system, in other words they applied unsupervised pre-training to improve supervised speech recognition, the authors used Wall Street Journal (WSJ) dataset for the experiments and achieved 2.34% WER, the model is considered as the first unsupervised pretraining ASR.

Moritz *et al*. (2020) developed a transformer based automatic speech recognition system for online application (streaming), the output should be predicted shortly after speaking, the proposed method worked on achieving this goal by applying time restricted self attention and triggered attention concept to control the latency output of the encoder and the decoder respectively, the experiment was done on LibriSpeech dataset, which is a corpus for English language, the system scored 2.8% WER.

Alsayadi *et al*. (2021) studied end-to-end deep learning for diacritized Arabic automatic speech recognition and presented three approaches, first model is CTC-based automatic speech recognition, second model is CNN-LSTM and lastly an attention-based approach to diacritized Arabic ASR, these approaches were applied to ESPnet and Espresso frameworks on the Standard Arabic Single Speaker Corpus (SASSC) dataset, the experiments showed that CNN-LSTM with an attention framework performance is better and achieved 5% WER

Messaoudi *et al*. (2021) developed an ASR system for Tunisian dialect, using DeepSpeech Mozilla model, for the purpose of the system a dataset “TunSpeech” was created to overcome the lack of Tunisian labeled data, and the best WER achieved 24% was after combining modern standard Arabic (MSA) data with the Tunisian data.

Alsayadi *et al*. (2022) created a CNN-LSTM based model to recognize multi dialectal Arabic speech recognition, the authors applied data augmentation technique to increase the training data, the experiments were done on MGB-3 and SASSC datasets and the model achieved 57% WER.

Showrav (2022) used a multilingual speech model pretrained on 40 Indian languages and fine-tune it on labeled data for Bengali language, which is considered to be low resource language, the dataset used for fine-tuning is Bengali Common Voice Corpus, the model achieved 3.819 Levenshtein Mean Distance.

Amari *et al*. (2022) proposed and compared two convolutional neural networks (CNN) based models for single word Arabic speech recognition, the first model uses CNN for feature extraction and long short term memory (LSTM) used for recognition, and the second model CNNs with deep architecture for feature extraction and recognition, the experiments were done using ASD dataset, after testing it was shown that the model with fully CNN's achieved best performance with 98%

Nasr *et al*. (2023) proposed two ASR models for Jordanian dialect and Yemeni dialect and multi-dialect Arabic, a Bidirectional Long Short-Term memory (Bi-LSTM) with attention model and DeepSpeech2 Mozilla model, the dataset used was collected by the authors and the results obtained from the experiments are in Table 1.

Bartelds *et al*. (2023) studied the effect of data augmentation for low resource languages on ASR performance, the study used four languages West Germanic, West-Frisian, Besemah and Nasal, the used model was XLS-R, they have shown that using data augmentation results lower word error rate results, and the best performance achieved was noticed when the amount of labeled data used for fine-tuning was increased, the best achieved WER was 14%.

Table 1. Summary of the results obtained for (Nasr *et al*., 2023)

Model	Dialect	WER
Bi-LSTM	Yemeni	59%
Bi-LSTM	Jordanian	83%
Bi-LSTM	Multi-dialect	53%
DeepSpeech2	Yemeni	31%
DeepSpeech2	Jordanian	68%
DeepSpeech2	Multi-dialect	30%

In Table 2, we summarize the previous works based on the language that been targeted, the method that been used, the evaluation metric that been used, and the score that been reported.

Table 2. Summary of the Literature Review

Authors(s)	Language	Method	Metric	Score
Masmoudi *et al*. (2018)	Arabic (Tunisian dialect)	Kaldi tool kit	WER	22%
Park *et al*., (2019)	English	SpecAgment	WER	Performance increase 10%
Schneider *et al*. (2019)	WSJ dataset	Multi-layer convolutional neural network	WER	2.34%
Moritz *et al*. (2020)	LibriSpeech dataset	Transformer	WER	2.8%
Alsayadi *et al*. (2021)	SASSC dataset	CNN-LSTM	WER	5%
Messaoudi *et al*. (2021)	Arabic (Tunisian dialect)	DeepSpeech	WER	24%
Alsayadi *et al*. (2022)	MGB-3 and SASSC	CNN-LSTM	WER	57%
Showrav (2022)	Bengali language	CNN	Levenshtein Mean Distance	3.819
Amari *et al*. (2022)	ASD dataset	CNN-LSTM	WER	98% performance increase
Nasr *et al*. (2023)	Jordanian and Yemeni Dialect	Bi-LSTM and DeepSpeech	WER	Table 1
Bartelds *et al*. (2023)	West Germanic, West-Frisian, Besemah and Nasal	Wav2vec2.0-XLS-R	WER	14%

Formatted: Font: Bold, Complex Script Font: Bold

Formatted: Centered, Space After: 8 pt

Formatted: Space After: 8 pt

2.2 Arabic Datasets for Speech Recognition

According to Chadha *et al*. (2022) building a good ASR model mainly depends on the used dataset, the used audios should be originated from different sources, it shouldn't contain a lot of background noise and music, it needs to be balanced for genders and a speaker contribution in the dataset shouldn't exceed 90 minutes, unfortunately Arabic datasets are limited, as many collected corpora are not publicly shared causing lack of Arabic datasets when building ASR for Arabic language and its dialects, in this

subsection the available Arabic datasets are reviewed in Table 2, presenting a brief overview about the dataset and its content.

Table 3. Arabic available datasets for speech recognition

Dataset	Brief
Mozilla Common Voice Dataset (Ardila *et al*., 2020)	Multilingual speech corpus designed for speech recognition purposes, it contains 7,335 hours of data from 60 languages (MSA, English, German, *etc*.) and considered to be the largest corpus for speech recognition that is in public
MediaSpeech Dataset (Kolobov *et al*., 2021)	The dataset designed for speech recognition testing and consist of four languages (French, Arabic, Spanish and Turkish) 10 hours for each language
Quran Ayat Speech to text (B-Ubuntu *et al*., 2022)	A public release dataset that contains Quran Ayat audios and the labeled transcription of the Ayat for multiple readers
Massive Arabic Speech Corpus (MASC) (Al-Fetyani *et al*.,2023)	Multi-regional and multi-dialect dataset contains 1000 hours of speech that is intended to advance the research in the development of Arabic technology, it contains speech of multiple Arabic dialect (Egyptian, Saudi, Jordanian, *etc*.)

Chapter 3

Neural Networks in Automatic Speech Recognition Systems

3.1 Introduction

In this chapter, we will discuss neural networks that are usually used in automatic speech recognition systems, a brief overview about self supervised learning, convolutional neural networks (CNNs), a comparison between CNN's and recurrent neural networks (RNN). The model's structure and how it processes the data, preprocessing techniques, and the evaluation metric will be discussed.

3.2 Machine Learning Basic Techniques

Machine learning can be recognized as the operation of solving problems by collecting data, creating models and mathematical algorithms to make predictions and resolve the problem (Burkov, 2019).

Machine learning can be classified into three techniques that are supervised learning, unsupervised learning and reinforcement learning, self supervised learning is type of unsupervised learning. This learning technique was adopted in the model used for this research and it will be discussed in the next subsections, in Table 3 a brief about each type is explained, supervised learning depends mainly on human annotations and labeled data, it faces many obstacles such as generalization errors, adversarial attacks, and spurious correlations, this has pushed the research to find the alternative techniques (Jaiswal *et al*., 2020).

Table 4. Machine learning techniques classification

Machine learning type	Brief
Supervised Learning	Supervised learning is a class of machine learning that is mainly based on human notes and observations (Theobald, 2017), it depends on labeled datasets to map the features between input and output (Burkov, 2019).
Unsupervised Learning	Unsupervised learning is a class of machine learning, which is based on feeding the model unlabeled data. The model will transform the data into values or vectors and the algorithm should be able to find patterns to resolve the problem. (Burkov, 2019).
Reinforcement Learning	In this class of machine learning, the model should be able to predict the output based on its interaction with the environment, it can preserve the state of its environment as a feature, it implements many actions and obtain the feedback as a reward. (Burkov, 2019).

Self supervised learning falls under the category of unsupervised learning, its process for speech recognition is mainly two stages as in Figure 2. The pretraining stage where a pretext task is formed for the deep learning algorithm to learn features and relationships from massive amounts of unlabeled data, pseudo labels are generated for this task that depends on special features of the input data this stage is called the pretraining stage, once this process is accomplished the model can be used to solve the primary task such as detection, classification and recognition. Since SSL algorithms can generate pseudo labels from the unlabeled data, it doesn't need the human annotations (Gui *et al*., 2023), SSL has shown its effectiveness in many tasks such as natural language processing (NLP) that can include translation, language models and sentence classification (Jaiswal *et al*., 2020).

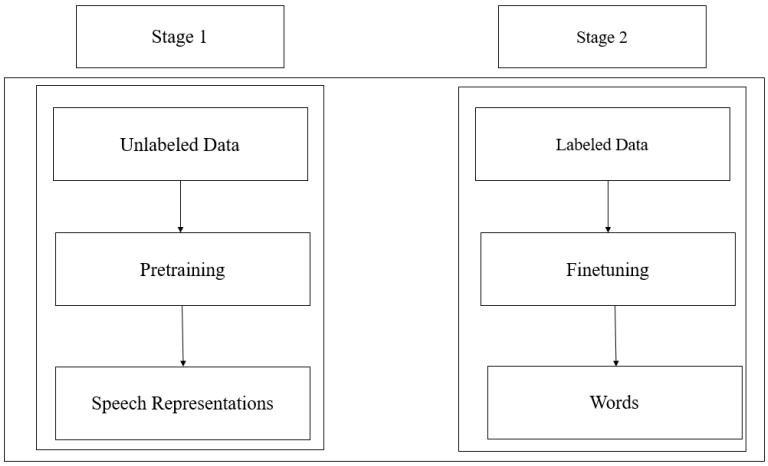

Figure 2. Self-supervised learning process

The progression of the SSL has got the attention of the researchers in many fields, in 2021 approximately 18,900 research was published on google scholar with SSL related subject. (Gui *et al*., 2023).

3.3 Convolutional Neural Networks for Speech Recognition

Convolutional neural networks are type of deep learning models that are designed to process data that has grid pattern, they are based on three layers, that are convolution, pooling, and fully connected layers as in Figure 4, the first two layers are aimed for the feature extraction process and the third layer maps features to the output data (Yamashita *et al*., 2018).

For ASR, the convolutional layer uses kernel sets (learnable filters) to convolve over the input speech signals (x), the operation can be represented in the Equation 1, were (z) represent the output of i-th filter, $w_i$is the wight and b_i is for the bias. When the convolution operation is done, an activation function should be applied to the output elements of the feature map. This is done to provide non-linearity, a common example of activation functions is ReLU and GeLU, this can be illustrated in Equation (2), $a_i$is the output after applying the activation function (Rahman *et al*., 2024).

$$z_i = (x \times w_i) + b_i \quad (1)$$

$$a_i = ReLU(z_i) \quad (2)$$

The pooling layer is used to down sample the output of the convolutional layer, this is done to reduce the computational complexity. Usually, this process is done using max-pooling by choosing the highest value in every local area, and can be illustrated in Equation (3), here p_i is for the output of the pooling layer (Rahman *et al*., 2024).

$$p_i = \max\_pool(a_i) \quad (3)$$

Lastly, after multiple convolutional and pooling layers, the output is inputted to fully-connected layer for the predict generation.

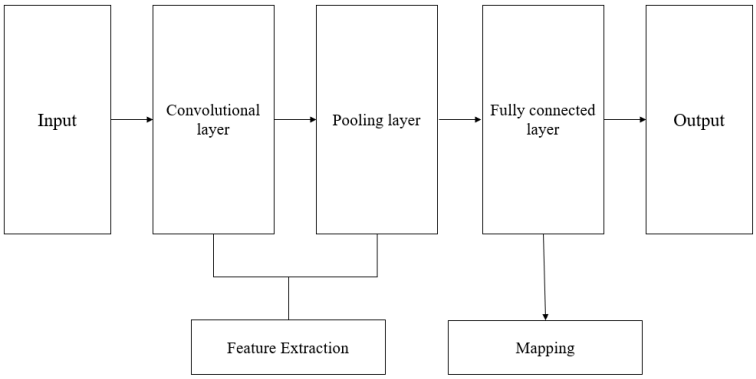

Figure 3. Theoretical model of convolutional neural network

3.4 Meta Models

Existing ASR systems requires enormous amount of labeled data to accomplish a respectable performance, unfortunately this is not available for Jordanian dialect. Based on known datasets, only MASC dataset contains Jordanian dialect data. In this Thesis self supervised learning is adopted for the goal of enhancing speech recognition for Jordanian dialect. SSL process learns the features from unlabeled data, after that fine-tuning the model on small amount of labeled data has been proving its efficiency in NLP and gaining researcher's interest (Baevski *et al.,* 2020).

One of the most famous SSL models in the field of speech recognition is Wav2Vec developed by Meta Company which was first developed in 2019 and enhanced to Wav2Vec2 which has many pretrained models also, Cross-Lingual Speech Representation (XLSR-53) which is pretrained on 53 languages.

3.4.1 Wav2vec 2.0 Architecture

Wav2vec2.0 model in Figure 4 is structured with a feature encoder which consists of convolutional layers, transformer encoder, and the quantization model, the feature encoder, is built with temporal convolution blocks, normalization layer and GELU activation function as in Figure 6, the role of the encoder is to transform the raw input data into latent representations $x \rightarrow z$.

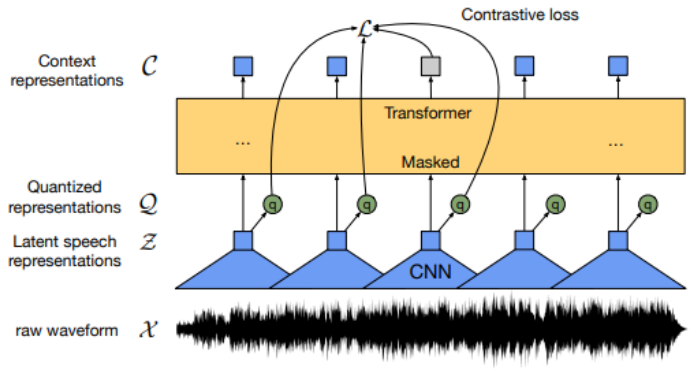

Figure 4. Wav2vec 2.0 model structure (Baevski *et al.*, 2020)

The output of the feature encoder (z_t) is fed to the transformer as in Figure 6 which builds the representations $c_1 \rightarrow c_T$, the role of the transformer is to capture information from the entire sequence, also the output of the feature encoder is fed to the quantization module to discretize the representations to $Z \rightarrow Q$ and represent the targets using the product quantization.

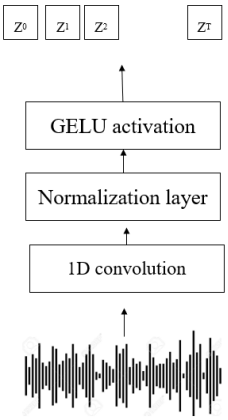

Figure 5. Wav2vec 2.0 feature encoder

The model pretraining process is based on masking an amount of the latent representation output and replacing them with trained feature vector shared between all masked time steps before they are fed to the transformer, the training objective is identifying real time quantized latent representation within a set of distractors for the masked time step, according to that the overall loss equation is:

$$L = L_m + \alpha L_d \tag{4}$$

Where L_m is the contrastive loss, L_d is the diversity loss used to inspire the model to use the code book elements equally often and α is hyperparameter used to control the diversity loss amount to be added to overall loss.

$$L_m = -log \frac{exp(\frac{sim(c_t, q_t)}{k})}{\sum_{\tilde{q}}^{Q_t} exp(\frac{sim(c_t, \tilde{q})}{k})} \tag{5}$$

$$sim(a, b) = a^T b / \|a\| \|b\| \tag{6}$$

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\tilde{p}g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \tilde{p}(g,v) \log \widetilde{\boldsymbol{p}}(\boldsymbol{g},\boldsymbol{v}) \quad (7)$$

$$p_{g,v} = \frac{\exp(l_{g,v}+n_v)/\Gamma}{\sum_{k=1}^{v} \exp(l_{g,k}+n_k)/\Gamma} \quad (8)$$

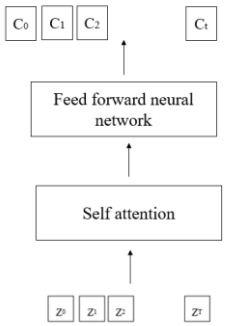

Figure 6. Wav2vec2 transformer encoder

where c_t is the output of context network, q is quantized latent speech representation, V is the number of entries in each codebook, G is the number of codebooks, and H is entropy of the Gumbel-SoftMax distribution maximization the next stage is the fine-tuning stage for the learned representations on labeled data (Baevski *et al*., 2020).

3.4.2 Cross Lingual Representation Learning for Speech Recognition

The cross lingual representation model in Figure 7 idea is to pretrain the model on speech from multiple languages, which will pull similar representations from other languages, the model is built to solve the contrastive task, which is to learn quantized representations shared between languages, similar to the Wav2vec2.0 model, Wav2vec2-XLSR contains a feature encoder that has convolutional layers, a transformer network and a quantization module.

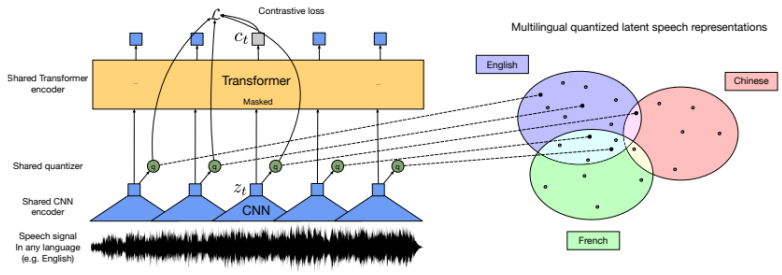

Figure 7. Wav2vec 2.0-XLSR model structure (Conneau *et al.*, 2020)

For the pretraining process, the authors used $p = 0.065$ for the masking as a start index, M=10 time steps and K = 100 distractors and Q_t from other time steps is in Equation 5.

To explain how the similarity of language work, the authors experimented the impact of language similarity on Italian language as a low resource language, low resource means that the amount of unlabeled data is low as like 5 hours. The authors pretrained the model on 5 hours of unlabeled data of Italian language and 50 hours from different language, then the model was fine-tuned on 1 hour of labeled Italian data, the results showed that adding more unlabeled data enhances the performance, adding related

language data such as Spanish improves the performance the largest (Conneau *et al*., 2020).

The similarity can be virtualized in Figure 9, as the authors analyzed the quantized latent speech representations shared across languages. Frequency vector of the tokens was calculated, these vectors are the probability of the distribution across the latent, after that the Jensen-Shannon symmetric similarity is computed between vectors and the resulting clusters defined by colors are the similarities obtained, this shows that pretraining on multiple languages should improve the performance of the model.

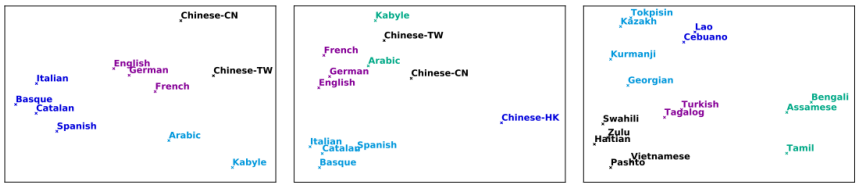

Figure 8. Language similarity virtualization (Conneau *et al*., 2020)

3.5 Voice Activity Detection

Voice activity detection (VAD) is the process of detecting speech and non-speech frames within the audio signal. It is one of the most important preprocessing steps for speech processing systems such as speech and speaker recognition (Wilkinson *et al*.,2021). VAD can be done in supervised and unsupervised models, usually to train the model for supervised VAD speech and non-speech labels are required and features like energy, zero crossing rate are investigated, on the other hand unsupervised VAD doesn't require huge amounts of labeled data and they are faster to train due to simple architecture (Dinkel, 2021).

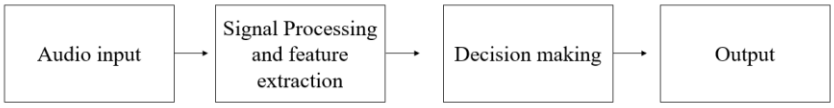

Figure 9. VAD algorithm block diagram

VAD algorithms work on processing a digitized audio signal, extracting the features of this signal, and making a decision based on the thresholds as in the block in Figure 9. There are many features that can differ for a VAD model such as Fourier coefficients, zero-crossing rates and periodicity, and these can be described based on statistical or heuristics models (Kola *et al*., 2011).

3.6 Waveform Amplitude Distribution Analysis

Waveform amplitude distribution analysis (WADA-SNR) was introduced by (Kim and Stern 2008), an algorithm to estimate signal to noise ratio (SNR) which assumes that the amplitude distribution of clean speech is categorized by Gamma distribution with fixed shape parameter value between 0.4 and 0.5 and the noise of the background is gaussian. WADA-SNR is the most recommended SNR measurement technique according to NIST 2016.

3.7 Evaluation Metrics

In this section, the most common evaluation metrics for ASR systems will be presented. These metrics measures the performance of ASR model by evaluating the difference between two sentences.

3.7.1 Word Error Rate

One of the basic methods to evaluate the performance of system ASR is WER, it is computed by aligning the recognized speech by the system with a reference transcription according to the equation:

$$\text{WER} = \frac{\text{S+D+I}}{\text{N}} \times 100\% \quad (10)$$

Where S refers to sum of substitutions, I for the insertions, D for the deletations and N denotes the total number of words in refferance to the transcirbted sample (Ali et al., 2018)

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Check spelling and grammar

Formatted: Font: Not Italic, Complex Script Font: Not Italic

3.7.2 Character Error Rate

Similar to WER, character error rate CER is the incorrect characters ratio used to evaluate ASR as in Equation 11, where S is the substitutions number, D is for the deletations, N denotes the total number of characters in refferance to the transcirbted sample, and I for the insertions.

Formatted: Font: 12 pt, Complex Script Font: 12 pt, Check spelling and grammar

$$CER = \frac{S+D+I}{N} \times 100\% \quad (11)$$

3.7.3 Cosine Similarity

Cosine similarity determines how close two sentences in vector space, as defined in Equation 12, the vectors of attributes are A and B, $A_i$and B_i are the components of the vector attributes, Euclidean norm of vector A is ||A||, and lastly ||B|| is Euclidean norm of vector ||B||.

$$\cos(A,B) = \frac{AB}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} AiBi}{\sqrt{\sum_{i=1}^{n}(Ai)^2}\sqrt{\sum_{i=1}^{n}(Bi)^2}} \quad (12)$$

3.8 The Speech Recognition Process

For this thesis, the block in Figure 10 will be the process adopted for the experiments in this thesis.

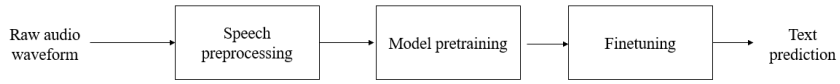

Figure 10. Block diagram for the process that will be followed in this research.

Starting with speech preprocessing which will include labeled data cleaning and noise reduction techniques as it will provide a great support for the model training. After that the model pretraining on the unlabeled data, finally fine-tuning on labeled data. To evaluate the ASR model, WER will be used to measure the performance.

Chapter 4

Data Preparation and Model Building

This chapter discusses the preprocessing steps for the audio signals and how it was prepared for each model. It also describes the experiments done on Wav2vec2.0 model and Wav2vec2.0-XLSR.

4.1 Data Preprocessing

Speech signal preprocessing is the most important step for ASR process, it is the first step to be done for speech recognition system. Preprocessing includes data cleaning, normalizing, and noise removal, the goal of this is to make the ASR more efficient. In the next subsections the preprocessing steps will be explained.

4.1.1 Preprocessing Steps for Audio Pretraining

According to Long (2021), a preprocessing step for pretraining, is that each audio should contain one person speaking and the pulse-code modulation (PMC) is to also be changed to 16 bits, silence is removed using VAD with aggressive Level 2 as was decided after experimenting, this step is only done in the pretraining process.

4.1.2 Data Cleaning

Starting with the labeled data, the non-alphabetical characters were removed, such as (dots, question marks, commas, *etc.*), the variations of hamza on Alef (أ ! إ) were normalized to *alef* (ا), and *teh marbuta* (ة) to *heh* (ه), as these characters usually doesn't affect the text understanding and will induce noise.

4.1.3 Audio Resampling

Resampling is a process where the sample rate is changed for the audio signal. Wav2Vec2 model originally was trained on 16 kHz audios, based on that all the audio files were resampled at sampling rate 16 kHz.

4.1.4 Data Chunking

To get the best audio data for ASR model, audios must be broken down to smaller chunks with durations (1-15) seconds. Although the audios can be broken down to small chunks, this process may include breaking the words at the boundary of the new audio causing noise during the process of training, VAD should be a critical task for this step to assure that the audios are clipped on silences and words are not broken at boundaries.

For this purpose, Google developed VAD known as WebRTC-VAD was used as it is fast and easily available according to (Ko *et al*., 2018). The used frame duration is 30ms and padding duration is 300ms, the aggressive level which is an integer number between 0-3, was chosen after testing on a sample of the audio data and found that level 2 is more suitable for the data.

4.1.5 Noise Reduction

The chunks are filtered to remove chunks with high background noise using WADA-SNR, the evaluation approach assumes that the amplitude distribution of clean speech is categorized by Gamma distribution with fixed shape parameter, after experimenting on sample of data, the audio files with SNR value less than 20 dB were excluded.

4.2 Wav2vec2.0 Model Training and Fine-tuning

Long (2021) was able to pretrain Wav2vec2 model on large amount of unlabeled data from the Vietnamese language and fine-tune it on labeled data, the author achieved good performance by pretraining the model on 100 hours of unlabeled data and trained the model on one hour of labeled data. For this experiment, this procedure will be followed for pretraining and training the model for Jordanian dialect, as there is lack of unlabeled data, the data will be used for pretraining and training the model.

1. Model training

After preparing the audios, the training process started for the Jordanian audios, same data was used for the pretraining and fine-tuning with split 20.5 hours for training, 14 hours validate, with total 35 hours of data, hardware specifications are in Table 4.

Table 5. Hardware specifications for the experiments

Aspect	Specification
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz, 2 cores, 56 MB cache
GPU	Nvidia P100 @ 1.32GHz @12GB
RAM	52GB
Python Library	3.10

2. Model fine-tuning

This full experiment training and fine-tuning took around 10 days to be completed, as there was a lack in pretraining audio data, this phase was expensive and

considered to have low prediction, the model was unable to predict any audios in the text and the output of the prediction was only random incorrect letters.

4.3 Fine-tuning Wav2vec2-XLSR on Jordanian Dialect

A public released Wave2vec2 model: Wav2Vec 2.0 Large-XLSR-53 which has 24 transformer blocks, model dimension 1024, inner dimension 4096 and 16 attention head. The model is already pretrained on 56k hours of data, the experiment started by preprocessing the labeled data as was done on the first experiment to get the best output of the ASR system.

Formatted: Normal, Indent: Before: 0"

Following the same training and validation set split percentage in common voice for modern standard Arabic (MSA) data set, according to that the training and validation are 13 and 21 respectively for 35 labeled data, 8 and 11 for the 20 hours labeled data the model and 1 hour for testing on both experiments.

Same preprocessing steps used in the first experiment to fine-tune the data on Wav2vec2 model was used to fine-tune the data for this experiment with the same data split for training and validation and same hardware specifications in Table 4. the fine-tuning process took approximately three days to be completed, and the model performance in terms of training loss and validation loss is in Figure 12 and 13.

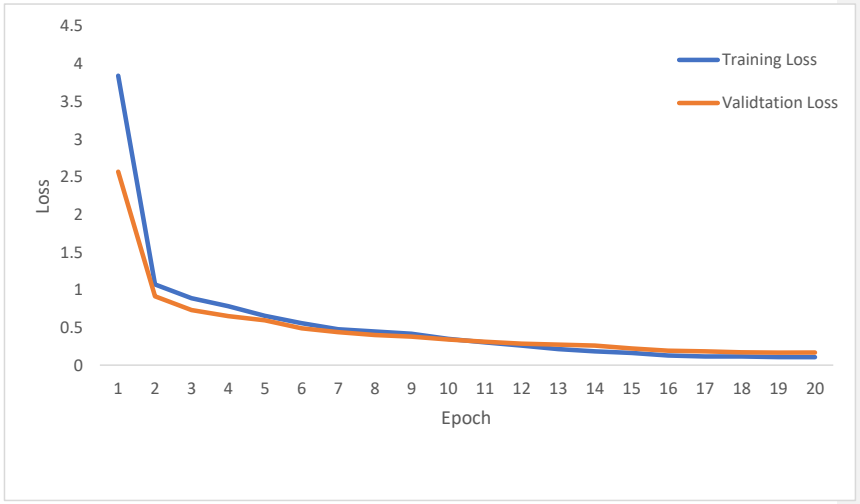

Figure 11. Model Loss on training and validation sets for 20 hours labeled data

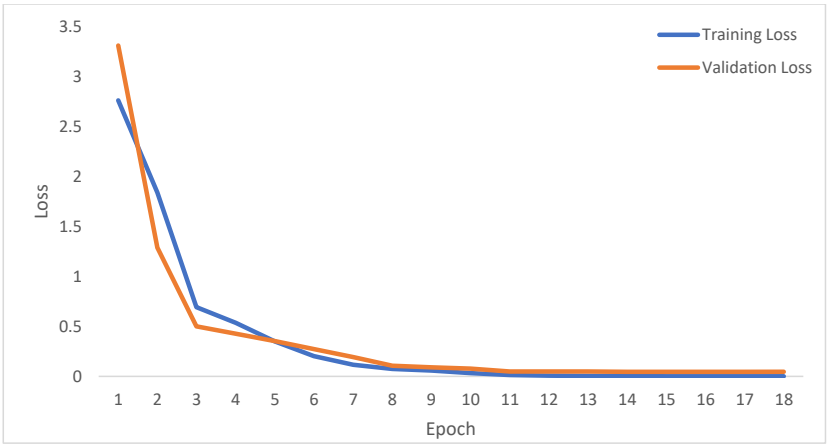

Figure 12. Model Loss on training and validation sets for 35 hours labeled data

In summary, we have presented three experiments done on Jordanian data, the first experiment was to fully train Wav2vec2.0 wrapper model and the model was unable to predict anything due to lack of training data based on the recommendation.

Formatted: Justified, Line spacing: Double

The second and third experiments were done on Wave2vec2.0-XLSR model on two different dataset sizes. Resulting good prediction in both datasets with higher results for the larger dataset.

Chapter 5
Experiments and Results

In this chapter, the model parameters and results obtained from the experiments conducted in Chapter 4 will be shown. a comparison between the results, the metric output, and an example of the predicted text.

5.1 Wav2vec2.0 Monolingual Model Training and Fine-tuning

The performance of this model was not efficient, as this model was unable to predict any words, the approach was to pretrain and fine-tune the model on the train and validate dataset using the wrapper version of Wav2vec2.0 and using the same preprocessing done by Long (2021) experiment. To evaluate the model on the test dataset, but due to the lack of pretrain data and as the model was only trained on Jordanian dialect, the pretraining was time consuming, model performance was not good, none of the predicted test was readable which is expected.

5.2 Wav2vec2.0-XLSR Fine-tuning

As stated, this model is pretrained on multiple languages and looks for the similarities between languages. This model fit perfectly for Jordanian dialect as it provided great results. For this experiment, the approach was to fine-tune the model on different amounts of labeled data to show the effect of increasing the labeled data for the model, the model was trained on 20 and 35 hours respectively following the parameters in Table 5. The results were compared with Sinai model published on hugging face as it is the highest performance for Arabic language with 18% WER on MSA common voice dataset.

Table 6. Fine-tuned Wav2vec2.0-XLSR Model Configurations

Parameter	Configuration
Learning rate	10^{-5}
Training batch size	64
Eval batch size	8
Epoch	20
Optimizer	Adam

5.2.1 Fine-tuning Wav2vec2.0-XLSR on 20 Hours of Data

The model performance was considered good for this experiment as the predict text is readable and somewhat accurate, with small errors as shown in Table 6, the results can be assumed as flawless in comparison to the Sinai model and the model WER was 25% which is considered good for Arabic speech recognition model.

Table 7. The fine-tuned model on 20 hours labeled data model performance.

Reference	Predicted Text
هلا انا عم بحكي عن اســـرتي قديه كانت متعاونه وبضلني احكي لولا امي وابوي ما وصلت لمحل ما انا واصله هلا	هلا انا عم بحكي عن اســـرتي ت متعاونه وبضـــلني احكي لولا امي وابوي ما وصـلت لمحل ما انا واصـله هلا
مرحبا انا اسمي ريم محيسن والده بتول بنتي معها شلل دماغي وانا كنت اشـــتغل بالتربيه وتقاعدت عاســـاس اتفرغلها واشتغل معها وتشتغل معي	مرحبا انا اسمري ريم محاسن والده بتول بنتي معها لل دمـاغي وانـا كنـت اشـــتغل بـالتربيـه وتقـاعدت عس اتفرغلها واشتغل معها وتشتغل معي
طيب عندي فكره شـــو رايكم نعمله في ملعب القدم الي جنب البيت	طيب عندي فكره شـــو رايكم نعمله في ملعب القدم الي جب البيت

5.2.2 Fine-tuning Wav2vec2.0-XLSR on 35 Hours of Data

As it is seen that the model performance for Arabic language was enhanced after fine-tuning on 20 hours, and to see the effect of increasing the amounts labeled data on the model, this time the model was fine-tuned on 35 Hours and the prediction was more accurate as expected and could be shown in Table 7 and the WER for the test dataset was 23%, according to the experiment results, it is confirmed that as long as the model is being fed with labeled data, the performance will get better.

Table 8. Fine-tuned on 35 hours labeled data model performance

Reference	Predicted Text
هلا انا عم بحكي عن اســـرتي قديه كانت متعاونه وبضلني احكي لولا امي وابوي ما وصلت لمحل ما انا واصله هلا	هلا انا عم بحكي عن اسرتي اد كانت متعاونه وبضلني احكي لولا امي وابوي ما وصلت لمحل ما انا واصله هلا
مرحبا انا اسمي ريم محيسن والده بتول بنتي معها شلل دماغي وانا كنت اشـــتغل بالتربيه وتقاعدت عاســـاس اتفرغلها واشتغل معها وتشتغل معي	مرحبا انا اسمي ريم محاسن والده بتول بنتي معها شلل دماغي وانا كنت اشـــتغل بالتربيه وتقاعدت عســـس اتفرغلها واشتغل معها وتشتغل معي
طيب عندي فكره شـــو رايكم نعمله في ملعب القدم الي جنب البيت	طيب عندي فكره شـــو رايكم نعمله في ملعب القدم الي جنب البيت

5.3 Discussion and Result Comparison

For this thesis, Wav2vec2.0 wrapper model was pretrained and fine-tuned on Jordanian dialect, as there was lack of unlabeled data, it is seen that the model was unable to predict anything from the test dataset. The results for the successful experiments is in Figure 13.

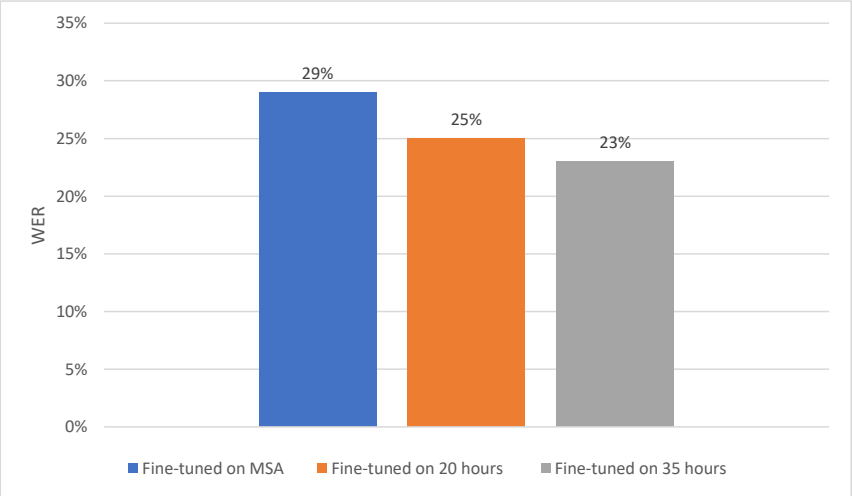

Figure 13. Fine-tuned model performance in terms of the WER on the test set

Looking into the pretrained XLSR model as it is pretrained on multiple languages and explores the similarities in the language, it was a great fit for Jordanian data as it lacks for the prepared datasets for speech recognition purposes. Comparing the results with Sinai model , it is seen that the model fine-tuned on 35 hours of labeled data provides the best performance.

Chapter 6

Conclusions and Future Work

In recent years, self-supervised learning models have proven to be powerful models that can handle natural language processing (NLP) tasks very well, it is good method to work with unlabeled datasets, in this thesis, Meta models which are the most effective way for ASR systems with SSL are CNN based models which starts with raw audio input, generates representations and solves a contrastive task, learned representations are used to enhance ASR.

Pretraining the cross-lingual model Wav2vec2.0 and fine-tuning the model resulted a bad performance due to lack of the pretraining data, Wav2vec2.0-XLSR model that is built to learn mutual speech unit across several languages was fine-tuned on Jordanian dialect was a good fit for Jordanian dialect case as it solve the problem of both lacking in labeled and unlabeled data, The obtained results were compared with Wav2vec2.0-XLSR model that was fine-tuned on MSA, the fine-tuned model on 35 hours of labeled data with 23% WER.

For future work it is suggested to use data augmentation in the preprocessing to increase the audio data as using augmentation enhances the WER, using other new models for self-supervised learning.

REFERENCES

Abdou, S., and Moussa, A. (2019), Arabic Speech Recognition: Challenges and State of the art. **Computational linguistics, speech and image processing for Arabic language**, 1-27.

Al Fetyani, M., Al-Barham, M., Abandah, G., Alsharkawi, A., and Dawas, M. (2023), MASC: Massive Arabic Speech Corpus. In **2022 IEEE Spoken Language Technology Workshop (SLT),** (pp. 1006-1013). IEEE.

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R.,and Almojil, M. (2021), Automatic Speech Recognition: Systematic Literature Review. **IEEE Access**, *9*, 131858-131876.

Ali, A., and Renals, S. (2018), Word Error Rate Estimation For Speech Recognition: E-WER. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics** (Volume 2: Short Papers) (pp. 20-24).

Alsayadi, H., Abdelhamid, A., Hegazy, I., and Fayed, Z. (2021), Arabic Speech Recognition Using End-to-End Deep Learning. **IET Signal Processing**, 15(8), 521-534.

Alsayadi, H., Abdelhamid, A., Hegazy, I., Alotaibi, B., and Fayed, Z. (2022), Deep Investigation Of The Recent Advances In Dialectal Arabic Speech Recognition. **IEEE Access**, 10, 57063-57079.

Alsayadi, H., Al-Hagree, S., Alqasemi, F., and Abdelhamid, A. (2022), Dialectal Arabic Speech Recognition Using CNN-LSTM Based on End-to-End Deep Learning. In 2022 2nd **International Conference on Emerging Smart Technologies and Applications (eSmarTA)** (pp. 1-8). IEEE.

Amari, R., Noubigh, Z., Zrigui, S., Berchech, D., Nicolas, H., and Zrigui, M. (2022), Deep Convolutional Neural Network For Arabic Speech Recognition. In **International Conference on Computational Collective Intelligence** (pp. 120-134). Cham: Springer International Publishing.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., and Weber, G. (2019), Common voice: A Massively-Multilingual Speech Corpus. **arXiv preprint arXiv***:1912.06670*.

B-Ubuntu, Quran Ayat Speech to Text, openSLR, Czech Republic, 2022.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020), Wav2vec 2.0: A Framework For Self-Supervised Learning of Speech Representations. **Advances in neural information processing systems**, 33, 12449-12460.

Bartelds, M., San, N., McDonnell, B., Jurafsky, D., and Wieling, M. (2023), Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. **arXiv preprint arXiv**:2305.10951.

Basak, S., Agrawal, H., Jena, S., Gite, S., Bachute, M., Pradhan, B., and Assiri, M. (2023), Challenges and Limitations in Speech Recognition Technology: A Critical

Review of Speech Signal Processing Algorithms, Tools and Systems. **CMES-Computer Modeling in Engineering & Sciences**, 135(2).

Burkov, A. (2019), The Hundred-Page Machine Learning Book (Vol. 1, p. 32). Quebec City, QC, Canada: Andriy Burkov.

Chadha, H. S., Gupta, A., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R., and Raghavan, V. (2022), Vakyansh: ASR Toolkit for Low Resource Indic Languages. **arXiv preprint arXiv**:2203.16512.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020), Unsupervised Cross-Lingual Representation Learning for Speech Recognition. **arXiv preprint arXiv**:2006.13979.

Dinkel, H., Wang, S., Xu, X., Wu, M., and Yu, K. (2021), Voice Activity Detection in The Wild: A Data-Driven Approach Using Teacher-Student Training. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, *29*, 1542-1555.

Gui, J., Chen, T., Zhang, J., Cao, Q., Sun, Z., Luo, H., and Tao, D. (2023), A Survey on Self-Supervised Learning: Algorithms, Applications, and Future Trends. **arXiv preprint arXiv***:2301.05712*.

Jaiswal, A., Babu, A., Zadeh, M., Banerjee, D., and Makedon, F. (2020), A Survey on Contrastive Self-Supervised Learning. *Technologies*, *9*(1), 2.

Javed, T., Doddapaneni, S., Raman, A., Bhogale, K.,Ramesh, G., Kunchukuttan, A., and Khapra, M. (2022), Towards Building ASR Systems for the Next Billion Users. In **Proceedings of the AAAI Conference on Artificial Intelligence** (Vol. 36, No. 10, pp. 10813-10821).

Kim, C., and Stern, R. (2008), Robust Signal-to-Noise Ratio Estimation Based on Waveform Amplitude Distribution Analysis. In **Interspeech** (pp. 2598-2601).

Ko, J. H., Fromm, J., Philipose, M., Tashev, I., and Zarar, S. (2018), Limiting Numerical Precision of Neural Networks to Achieve Real-Time Voice Activity Detection. In 2018 **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)** (pp. 2236-2240). IEEE.

Kola, J., Espy-Wilson, C., and Pruthi, T. (2011), Voice Activity Detection. **Merit Bien**, 1-6.

Kolobov, R., Okhapkina, O., Omelchishina, O., Platunov, A., Bedyakin, R., Moshkin, V., and Mikhaylovskiy, N. (2021), Mediaspeech: Multilanguage ASR Benchmark and Dataset. arXiv preprint arXiv:2103.16193.

Labied, M., and Belangour, A. (2021), Automatic Speech Recognition Features Extraction Techniques: A Multi-Criteria Comparison. **International Journal of Advanced Computer Science and Applications**, 12(8).

Labied, M., Belangour, A., Banane, M., and Erraissi, A. (2022), An Overview of Automatic Speech Recognition Preprocessing Techniques. In **2022 International Conference on Decision Aid Sciences and Applications** (DASA) (pp. 804-809). IEEE.

Long, M. (2021), Self-Supervised Speech Recognition with Limited Amount of Labeled Data. **Github** https://github.com/mailong25/self-supervised-speech-recognition

Malik, M., Malik, M., Mehmood, K., and Makhdoom, I. (2021), Automatic Speech Recognition: A Survey. **Multimedia Tools and Applications**, *80*, 9411-9457.

Masmoudi, A., Bougares, F., Ellouze, M., Estève, Y., and Belguith, L. (2018), Automatic Speech Recognition System for Tunisian Dialect. **Language Resources and Evaluation**, 52, 249-267.

Messaoudi, A., Haddad, H., Fourati, C., Hmida, M., Mabrouk, A., and Graiet, M. (2021), Tunisian Dialectal End-to-End Speech Recognition Based on Deep Speech. **Procedia Computer Science**, 189, 183-190.

Malik, M., Malik, M., Mehmood, K., and Makhdoom, I. (2021), Automatic Speech Recognition: A Survey. **Multimedia Tools and Applications**, 80, 9411-9457

Moritz, N., Hori, T., and Le, J. (2020), Streaming Automatic Speech Recognition With the Transformer Model. In **ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)** (pp. 6074-6078). IEEE.

Nasr, S., Duwairi, R., and Quwaider, M. (2023), End-to-End Speech Recognition for Arabic Dialects. **Arabian Journal for Science and Engineering**, *48*(8), 10617-10633.

Oliver, T. (2017). Machine Learning for Absolute Beginners. **A Plain English Introduction**.

Park, D., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E., and Le, Q. (2019), Specaugment: A Simple Data Augmentation Method for Automatic Speech Recognition. **arXiv preprint arXiv**:1904.08779.

Rahman, A., Kabir, M. M., Mridha, M. F., Alatiyyah, M., Alhasson, H. F., and Alharbi, S. (2024). Arabic Speech Recognition: Advancement and Challenges. **IEEE Access.**

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019), Wav2vec: Unsupervised Pre-Training for Speech Recognition. **arXiv preprint arXiv**:1904.05862.

Showrav, T. (2022), An Automatic Speech Recognition System for Bengali Language Based on Wav2vec2 and Transfer Learning. **arXiv preprint arXiv***:2209.08119*.

Wilkinson, N., and Niesler, T. (2021), A Hybrid CNN-BiLSTM Voice Activity Detector. In ICASSP 2021-2021 **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)** (pp. 6803-6807). IEEE.

Wiseman, J., and Bondarenko, I. (2016), Python Interface to the Webrtc Voice Activity Detector. **Python Interface** to the WebRTC voice activity detector.

Yamashita, R., Nishio, M., Do, R., and Togashi, K. (2018), Convolutional Neural Networks: An Overview and Application in Radiology. **Insights into imaging**, *9*, 611-629.

تقصي حلول التعلم العميق للتعرف بدقة على الكلام العربي باللهجة الأردنية

إعداد

ساره ايمن مسعد الماضي

المشرف

الأستاذ الدكتور غيث علي عبندة

ملخص

يعتبر التعرف التلقائي الآلي على الكلام للغة العربية ولهجاتها مشكلة شائعة. يعود ذلك إلى تنوع اللهجات، ونقص البيانات المصنفكتوبة وتعقيد المورفولوجيا. لقد تم استخدام العديد من تقنيات التعلم الآلي سابقا لتحسين التعرف على الكلام للغة العربية، مثل الشبكات العصبونية المعصبية المتكررة، ونماذج التعلم الآلي مثل نموذج هيدننماذج ماركوف المخفية، والشبكات العصبونية المعصبية التلافيفية المتكرارية.

استخدام التعلم بالإشراف الذاتي للشبكات العصبونية المذاتي المشرف للتعرف على الكلام يؤدي حاليًا إلى نتائج رائعة. في هذه الرسالة، نستخدم النوع من التعلم المذاتي المشرف لتحسين نظام التعرف على الكلام للهجة الأردنية. قمنا بدراسة نموذجين حديثين تم تطويرهما من قبل شركة Meta AI: نموذج الغلاف Wav2vec2.0 ونموذج Wav2vec2.0-XLSR حيث تم تدريب هذه النماذج وضبطها باستخدام مجموعة بيانات للهجة الأردنية من مشروع الكلام العربي الضخم (MASC).

نظرًا لنقص البيانات الموسومة المصنفة بما يكفي للهجة الأردنية، لم يتمكن نموذج الغلاف Wav2vec2.0 الذي تم تدريبه على اللهجة الأردنية المصنفة للموسومة المتاحة من التعرف بشكل صحيح على اللهجة الأردنية. لذلك تم تدريب نموذج Wav2vec2.0-XLSR المدرب مسبقًا على 53 لغة وعلى التشابه بين اللغات. هذا النموذج أداءه أفضل من النموذج الأول عند ضبطه تدريبه تربيًا إضافيًا على مجموعة بيانات اللهجة الأردنية المتاحة. قمنا بضبط هذا النموذج على حجمين مختلفين من مجموعات البيانات وقمنا بتقييم النموذج المضبوط باستخدام معيار معدل الخطأ في الكلمات (WER) على مجموعة بيانات اختبار فحص للهجة الأردنية، حيث تم تدريبه عند ضبط النموذج Wav2vec2.0 XLSR على مجموعتيات البيانات التي مدتها 20 ساعة و35 ساعة، يحقق وقد حقق النموذج WER معدل خطأ بنسبة 25% و23% على التوالي. هذا تحسين جيد مقارنةً بالنموذج الذي يتم ضبطه إعادة تدريبه على مجموعة بيانات اللغة العربية الفصحى الحديثة فقط (MSA)، والذي يحقق معدل خطأ نسبة WER تبلغ 29% على مجموعة البيانات التجريبيةالفحص. بالإضافة إلى ذلك، يوفر النموذج الذي تم ضبطه على مجموعة بيانات اللهجة الأردنية الأطول تعرفًا جيدًا للكلام باللهجة الأردنية.