A Dialectal Speech Corpus and Automatic Speech Recognition Models for Four Jordanian Sub-Dialects

Final Project Report Submitted to
The Department of Computer Science
Faculty of Computer and Information Technology
Jordan University of Science and Technology
In Partial Fulfillment of the Requirements for the Degree of
Bachelors of Science in Computer Science

Prepared by:

Majd Fakhri Omari [163099]
Rania Khalil Ali Hushoush [163339]

Supervisor:
Dr. Rasha Obeidat

January 2026

نموذج حقوق الملكية الفكرية لمشاريع التخرج في قسم علوم الحاسوب

يتم قراءة وتوقيع هذا النموذج من قبل الطلاب المسجلين لمشاريع التخرج في قسم علوم الحاسوب

تعود حقوق الملكية الفكرية لمشاريع التخرج ونتائجها (مثل براءات الاختراع أو أي منتج قابل للتسويق) إلى جامعة العلوم والتكنولوجيا الأردنية، وتخضع هذه الحقوق إلى قوانين وأنظمة و تعليمات الجامعة المتعلقة بالملكية الفكرية وبراءات الاختراع.

بناءا على ما سبق أوافق على ما يلي:

1) أن أحفظ كافة حقوق الملكية الفكرية لجامعة العلوم والتكنولوجيا الأردنية في مشروع التخرج.
2) أن ألتزم بوضع اسم جامعة العلوم والتكنولوجيا الأردنية و أسماء جميع الباحثين المشاركين في المشروع على أي نشرة علمية للمشروع كاملا أو لنتائجه. و يشمل ذلك النشر في المجلات و المؤتمرات العلمية عامة أو النشر على المواقع الإلكترونية أو براءات الاختراع أو المسابقات العلمية.
3) أن ألتزم بأسس حقوق التأليف المعتمدة في جامعة العلوم والتكنولوجيا الأردنية.
4) أن أقوم بإعلام الجهة المختصة في الجامعة عن أي اختراع أو اكتشاف قد ينتج عن هذا المشروع و أن ألتزم السرية التامة في ذلك و أن أعمل من خلال الجامعة على الحصول على براءة الاختراع التي قد تنتج عن هذا المشروع.
5) أن تكون جامعة العلوم والتكنولوجيا الأردنية هي المالك لأي براءة اختراع قد تنتج عن هذا المشروع و تشمل هذه الملكية حق الجامعة في إعطاء التراخيص و التسويق و البيع كمؤسسة راعية و داعمة لكافة الأنشطة البحثية. و يكون حق للطالب شمول اسمه على براءة الاختراع كأحد المخترعين، و في حال تم إعطاء تراخيص أو تسويق و بيع لأي من منتجات المشروع يمنح المخترعون بما فيهم الطالب نسبة من الإيرادات حسب تعليمات البحث العلمي في جامعة العلوم والتكنولوجيا الأردنية.

إسم الطالب مجد فخري العمري   التوقيع majd

إسم الطالب رانيا خليل العشوش   التوقيع

إسم الطالب   التوقيع

إسم المشرف د. رشا عبيدات   **التوقيع**

تاريخ ....

Bridging the Dialectal Gap: A Systematic Investigation of Speech Recognition across Four Key Jordanian Localities

Hajar Alhadaris
Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan
hsalhadaris22@cit.just.edu.jo

Rania Khalil Ali Hushoush
Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan
rkhushoush22@cit.just.edu.jo

Majd Fakhri Omari
Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan
mfalomari221@cit.just.edu.jo

Nour Abu Beidar
Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan
nmabubeidar22@cit.just.edu.jo

***Abstract*—Many Jordanian Arabic dialects are still inadequately researched in speech recognition, which limits the capabilities of modern systems for Jordanian speech. Most current systems are designed to work with Modern Standard Arabic or general Jordanian dialects, reflecting a limited ability to understand Jordanian speech. This research aims to address this gap by investigating Jordanian dialects. The research focuses on the collection of speech data from four different locations in the state of Jordan: Irbid in the north of the state, Tafilah in the south of the state, Amman as the capital of the state, and Southern Aghwar. The texts were originally written in the local dialect of the state of Jordan. The texts were then recorded in audio format to represent the pronunciation of the spoken words. The collected speech data was then used to train various speech recognition systems. From the experimental results, it is observed that training the models on dialect-specific speech data improves the accuracy of Jordanian dialectal speech recognition. The proposed system demonstrated greater strength in pronunciation and accent variability than a general speech recognition system. Finally, the paper highlights the need to use localized dialect data to improve the system's accuracy.**

I. PROJECT GOALS AND OBJECTIVES

The primary goal of this project is to enhance voice recognition systems' comprehension of Jordanian dialects by utilizing data that accurately represents regional speech patterns.

The specific objectives of this project are as follows:

- Collect speech data from four important parts of Jordan: the Southern Aghwar, Irbid, Tafilah, and Amman. Ensure the data reflects regional pronunciation or style.
- Before making an audio recording, translate written Arabic sentences into a real, daily Jordanian dialect.
- Use this dialectal speech to refine Arabic-focused Wav2Vec 2.0 models and assess their performance.
- Use WER and CER to investigate how different training conditions and hyperparameters impact model quality.
- To determine the comparative advantages of dialect-aware models and a general-purpose ASR system, compare them.
- To Examine mistakes qualitatively to identify the phonological and dialectal difficulties these models face.

II. INTRODUCTION

Arabic Automatic Speech Recognition (ASR) has remained a challenge due to the great diversity of spoken dialects and limited speech resources. Recent large-scale efforts, such as the Casablanca Corpus [1], have expanded dialectal representation by providing multi-dialectal speech in Arabic across several Arabic-speaking countries, including Jordanian Arabic.

In this regard, speech-based research usually depicts Jordanian Arabic at a general dialectal level, either as a single category or as a component of larger Levantine groupings [2]. Although Jordanian-spoken audio is available in the current resources, the granularity of these datasets is typically low, making it challenging to capture variation in Jordanian speech.

Dialectal variation in Jordanian Arabic is strongly influenced by both geographic and social factors. In regions with stable communities and little population movement, local dialects are usually well-preserved and deeply embedded. On the other hand, governorates that attract migrants for work, education, or urban living often bring together speakers from diverse linguistic backgrounds. On the other hand, governorates that attract migrants for work, education, or urban living often bring together speakers from diverse linguistic backgrounds. Instead of a single, stable local dialect, this interaction results in the emergence of mixed or fluctuating speech patterns

Inspired by this reality, this work highlights that it is no longer adequate to treat the Jordanian dialect as a homogenous entity. Instead, it is essential to understand the current dialect in each governorate, particularly in areas where demographic changes directly affect everyday pronunciation. The essential factor here is knowing which dialect people in each governorate use in their daily speech, especially in areas where population shifts affect daily speech patterns. A review of the existing literature suggests that previous studies—both within Arabic contexts and in other languages—generally conceptualize dialects as broad regional categories, often failing to account for intra-regional variation and the influence of population movements on dialect evolution. Furthermore, Arabic ASR research has historically focused on Modern Standard Arabic or a limited set of widely studied dialects, leaving many low-resource varieties, including Jordanian dialects, insufficiently represented.

This paper presents a governorate-level study of Jordanian Arabic for ASR, focusing on speech from Irbid, Tafilah, Amman, and the Southern Aghwar. The proposed dialect-aware data collection pipeline adapts written sentences into local Jordanian dialect before recording them as spoken audio, preserving natural pronunciation and accent. Using the resulting speech–text dataset, several ASR models are evaluated, including Arabic-focused Wav2Vec 2.0 variants and a generic Whisper baseline. Model performance is assessed using Word Error Rate (WER) and Character Error Rate (CER), enabling a comparative analysis of how effectively each model captures dialectal variation across different Jordanian regions

III. Related Work

A. The Multi-Dialectal Dilemma: A Global Technical Frontier in ASR

Arabic language is among one of the largest linguistic foundations of humanity and serves as one of the major communication mediums for approximately 400 million people and a vital bridge between civilizations. As one of the six official languages of the United Nations [3]. The complexity of Arabic arises not only from its huge vocabulary, but also from its highly productive ”root and pattern” morpho-syntactic system, which provides a difficult challenge for modern Natural Language Processing (NLP) and Speech Recognition [4]. Beyond formal structure, Arabic's complexity is further rooted in its unique state of diglossia. As noted in linguistic research, there is a substantial discrepancy between Modern Standard Arabic (MSA), the language of formal education and the many regional dialects spoken in spontaneous, daily contacts [5]. These dialects evolved from a common source into different systems with significant pronunciation and vocabulary differences, sometimes leaving typical AI models unable to bridge the gap between formal training and real-world speech

This search for the “authentic voice” of speakers is no isolated Arabic challenge; it is a global technical frontier. To appreciate the scale of this challenge, we need to see how other key languages deal with similar fragmentation. In English, for example, dialectal changes in syntax have been shown to be not only geographically driven but are ”moving targets” that shift over space and time [6]. In the Far East, the Chinese language landscape exhibits an even more intense struggle with ”phonological overlap,” where dialects share a writing system but possess entirely different phonetic realizations, providing significant difficulties for audio models [7].

The problem further extends to the ”rhythmic melody” of speech. For Korean dialects, the true identity of the speaker in terms of his region usually lies not in the words but in the intonation pattern, and complicated deep learning models are needed to detect such subtle musical cues [8]. Similarly, information theory has been used as a tool to quantify the distance between European and Brazilian Portuguese, showing that closely related forms may be indistinguishable for standard AI systems [9].

By examining Arabic through this global lens, it becomes clear that the ”dialect problem” is a universal linguistic barrier. The immense scale and complexity of the Arabic language [10] making it the ultimate testing ground for ASR technology. These international efforts highlight the imperative for a system to go beyond “standardized” models and embrace the rich, chaotic and beautiful reality of human dialects.

B. Evolution of Arabic ASR: From Traditional to Neural Approaches

Over the past decade, Arabic ASR has evolved significantly, primarily driven by the need to address the scarcity of labeled data and the linguistic complexity of Arabic. Existing work spans traditional acoustic modeling, data augmentation methods, and the recent shift toward self-supervised learning.

One of the earlier contributions to Arabic ASR was presented in [11], where a hybrid GMM-HMM acoustic model enhanced with DNN fine-tuning was developed for MSA. This work highlighted critical challenges, including rich morphology and pronunciation ambiguity due to the absence of diacritics. To address the limited size of Arabic corpora, data augmentation emerged as a practical solution. In [12], techniques such as noise injection, speed perturbation, and pitch shifting were applied to the SASSC corpus, resulting in a 4.55% reduction in WER. More recently, the ArabRecognizer system [13] introduced architectures inspired by DeepSpeech2, utilizing stacked GRU layers with CTC decoding to improve accuracy and computational efficiency specifically for Arabic phonetics.

C. Self-Supervised Learning and Multidialectal Advancements

The rise of **self-supervised learning (SSL)** has enabled substantial progress for low-resource Arabic ASR. Research on the **Wav2Vec 2.0** framework [14] demonstrated that multilingual models such as **XLSR-53** significantly outperform English-only models when fine-tuned with Arabic data. This advancement reached a new scale with the introduction of **XLS-R** [15], a model pre-trained on 436,000 hours of speech, providing a robust universal foundation for low-resource dialects.

The focus has subsequently shifted toward capturing the rich diversity of regional dialects. A landmark effort in this area is the **MGB-5** challenge [16], which utilized 2,000 hours of broadcast media to prove that unified models trained on multiple dialects achieve better generalization than those limited to a single variety. Following this, Shon et al. [17] established an influential baseline for the identification of dialects using CNN-based encoders. Further pushing these boundaries, the **Casablanca project** [1] recently introduced a massive community-driven data set that covers eight dialects, including Jordanian. Their findings confirmed that while single-dialect models fail on unseen varieties, **unified multidialectal training**—such as **Whisper-Mixed**—is essential for robust performance.

In addition to dialectal diversity, multilingual nuances and post-processing have become key research frontiers. Chowdhury et al. [18] proposed Conformer-based models to support dialectal code-switching between Arabic, English, and French. Simultaneously, advancements like FastConformer [19] have pushed MSA and Classical Arabic (CA) accuracy to state-of-the-art levels. To further refine these outputs, the *CoVoGER benchmark* [20] has introduced generative error correction (GER) using Large Language Models (LLMs) to contextually fix recognition errors.

The work most directly aligned with the present research is the self-supervised framework by Safieh et al. [21], This work leveraged Wav2Vec 2.0 and iterative **noisy-student training** on a 113-hour Jordanian corpus encompassing **urban, rural, and Bedouin** dialects. By employing an iterative five-generation training process and advanced data augmentation techniques such as pitch shifting and time stretching, they achieved a 5% absolute reduction in WER, lowering it from 56.8% to 51.5%.

D. Research Objective and Gap Analysis

While the aforementioned studies have laid a robust foundation for Arabic and Jordanian ASR, a critical gap remains in the granularity of dialectal evaluation. As summarized in **Table I**, existing research predominantly treats the Jordanian dialect as a monolithic entity or divides it into broad socio-linguistic categories—Urban, Rural, and Bedouin—without accounting for specific geographical variants. There is a lack of systematic, empirical evidence on how ASR performance varies across specific regional hubs in Jordan.

This research addresses this limitation by conducting a systematic investigation across four key Jordanian localities. By isolating these regions, we aim to identify specific phonological friction points and lexical variations that contribute to recognition errors. Our goal is to move beyond generalized Jordanian models toward a geographically aware framework that acknowledges the rich linguistic diversity within the Kingdom, ultimately improving the recognition of users' "authentic voice" across all surveyed localities.

IV. APPROACH AND METHODOLOGY

A. Methodology

1) Data Collection and Corpus Description: The textual prompts used were sourced from the Tatoeba corpus. Tatoeba is a free collection of example sentences with translations geared towards foreign language learners. Arabic Sentences were selected for the collection of dialectal speech data. The selected texts from the database cover a wide range of daily-life and conversational domains, including topics such as family relationships, work, transportation, education, travel, emotions, and social interaction. Speech data were collected from native speakers across four regions of Jordan: Irbid, Tafilah, Amman, and Southern Aghwar. For each region, nearly 5,550 Samples will be recorded, resulting in four parallel dialectal versions of the same textual content. Speakers are instructed to read the sentences using their natural local dialect. The dataset includes both male and female speakers across different age groups to ensure speaker diversity and improve the models' generalization capability.

2) Data Preprocessing: The audio files were converted to a unified format (WAV, mono channel, 16 kHz sampling rate) to ensure compatibility with every selected ASR model. To improve audio quality, background noise was reduced using a noise sample taken from the start of each recording and leading and trailing silences were removed. Prior to normalization, the textual data were manually adapted from Modern Standard Arabic (MSA) into Jordanian dialectal forms in order to better reflect natural spoken language. For each sentence, a dialectal version was created on a separate line while preserving the original semantic content, resulting in parallel text entries aligned by unique sentence IDs.

ID	**MSA**	**Irbid Dialect**
14721	ما تكلفة المقاعد الخاصة بالشرفة؟	قديش بتكلف الكراسي تبعت البرندة؟

Fig. 1: Example of MSA to Jordanian Dialect Text Adaptation

Additionally, textual data were normalized to reduce spelling inconsistencies by removing diacritics, punctuation, and numbers and by unifying common Arabic letter variants . This results in cleaner and standardized data for training and evaluation.

Before Normalization	**After Normalization**
انتبه لأنه جو الجبال بيتغير بسرعة	انتبه لانه جو الجبال بيتغير بسرعه
ماشي، رح أوخذ بس ثنين. شكراً	ماشي رح اوخذ بس ثنين شكرا

Fig. 2: Examples of Text Normalization

TABLE I: Summary of Major Related Work on Arabic ASR and Dialectal Processing

Year	Focus	Dataset / Domain	Model / Method	Main Contribution
2017	MSA ASR	Broadcast News	GMM-HMM + DNN	Improved acoustic modeling for MSA morphology.
2018	Dialect Identification	MGB-3 (5 Dialects)	CNN + Siamese	Established baseline for regional dialect classification.
2019	Multidialectal ASR	MGB-5 (2,000h)	Hybrid DNN-HMM	Proved that multidialectal training improves cross-dialectal robustness.
2021	SSL for Low-Resource	Mozilla Common Voice	Wav2Vec 2.0 (XLSR- 53)	Proved effectiveness of multilingual SSL for Arabic.
2021	Universal Repr	28 Languages	XLS-R (Meta AI)	Massive scaling of cross-lingual represen tations for low-resource ASR
2021	Data Augmentation	SASSC Corpus	Noise/Speed/Pitch	Achieved 4.55% WER reduction via training diversity.
2021	Multilingual/Code switch	QASR, MGB-3/5, CS	Conformer + BPE	Unified model handling dialectal and cross lingual switching
2022	Jordanian Dialect ASR	113h Jordan Corpus	Wav2Vec 2.0 + SSL	Effective use of pseudo labeling for Jordanian Arabic.
2024	Arabic-Specific Arch.	Varied MSA Corpora	DeepSpeech2 + GRU	Optimized inference and accuracy for Arabic phonetics.
2024	Multidialectal Bench.	Casablanca Dataset	Whisper-Mixed	Benchmarked 8 dialects; showed mixed model superiority in generalization
2025	Generative Correction	CoVoGER Benchmark	LLM-based GER	Used LLMs to contextually fix ASR errors across 15 languages.
2025	Unified MSA/Classical	MASC, FLEURS	FastConformer	Achieved state-of-the-art 1.34% WER for Classical Arabic

Each audio file was matched with its corresponding dialect text using unique sentence IDs. Samples with missing or corrupted audio were removed. Random samples were manually checked to confirm the accuracy of the audio-text alignment.

3) Selected ASR Models: Multiple variants of the Wav2Vec 2.0 architecture were used in this study to benchmark performance on Jordanian Dialect Speech. Wav2Vec 2.0 was created by Meta AI as a self-supervised framework to extract speech characteristics from unlabelled audio. [22]

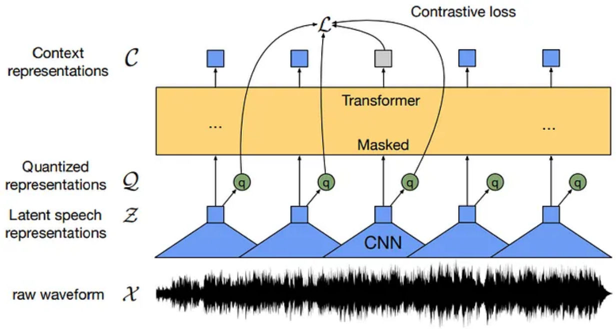


Fig. 3: Architecture of the Wav2Vec 2.0 self-supervised speech representation model [22]

Evaluating different pretraining and fine tuning configurations provides insight into how various model variants affect speech recognition accuracy across Jordanian dialects.

Two Wav2Vec 2.0 models focused on Arabic were chosen: jonatasgrosman / wav2vec2-large-xlsr-53-arabic and mohammed / wav2vec2-large-xlsr-arabic. Both models were pretrained and fine-tuned specifically for Arabic speech, using an XLSR-53 multimodal pre-training structure, which is built to deal with language variances and resource-poor vernaculars. As a baseline, the Whisper model from OpenAI was also included. Whisper is designed to be a generic ASR model, developed with a large multi-lingual and multi-genre audio dataset.

It can perform multiple tasks, such as multilingual speech recognition, translating recognized speech into text, and providing identification of the target identity of the spoken language. Including Whisper will facilitate the direct comparison of the two Arabic-targeted ASR models with a generic ASR system that is built to work with multiple languages.

The combination of evaluations generated from the Wav2Vec 2.0 variants along with Whisper provides a wide range of data for analysing the distinct phonemic and lexemic variances between the various Jordanian Dialects.

4) Experimental Setup: In order to evaluate the feasibility of the proposed method, 400 speech samples were collected in initial experiments for training and early testing prior to conducting large-scale testing using the entire dataset. These initial samples were collected exclusively from the Irbid dialect in order to conduct controlled preliminary experiments before extending the study to other Jordanian dialect regions.

There were two models of Wav2Vec 2.0 specifically tailored for Arabic, the first being jonatasgrosman/wav2vec2-large-xlsr-53-arabic and the second being mohammed/wav2vec2-large-xlsr-arabic. For each model, two different configurations of training were performed to evaluate how hyperparameter settings affect training behaviour and performance. Both models were fine-tuned using the CTC training objective and optimised with AdamW, with different settings of learning rates, batch sizes, and gradient accumulations used for the training process.

Regularization techniques, such as dropout, were applied to improve generalization and reduce the risk of overfitting. During the training phases of both models, two distinct evaluation strategies were employed. One evaluation assessed the model's performance at the end of each training epoch. The other evaluation occurred at pre-decided intervals. These evaluations allowed us to assess the convergence of the two model-training processes under different training settings. These initial experiments were all conducted in a controlled laboratory using Google Colab (GPU-accelerated). These initial experiments help us test ideas and improve the design before moving on to the full-scale experiments.

5) Evaluation Metrics: To improve the performance of these ASR models, WER and CER were used as the main evaluation metrics.

WER measures the percentage of word-level errors between the predicted transcription and the reference text. It is widely used as a standard measure in speech recognition experiments. The formula for WER is:

$$\text{WER} = \frac{S + D + I}{N} \times 100 \tag{1}$$

$$\begin{aligned} \text{Where:} \quad & S \text{ is the number of substitutions,} && (2) \\ & D \text{ is the number of deletions,} && (3) \\ & I \text{ is the number of insertions,} && (4) \\ & N \text{ is the total number of words in the reference text.} && (5) \end{aligned}$$

CER was used to detect character-level errors, giving a more detailed evaluation of the model's performance. This is important for Arabic dialectical speech, where spelling and pronunciation differences are common. The formula for CER is:

$$\text{CER} = \frac{S + D + I}{N} \times 100 \tag{6}$$

$$\begin{aligned} \text{Where:} \quad & S \text{ is the number of substitutions at the character level,} && (7) \\ & D \text{ is the number of deletions at the character level,} && (8) \\ & I \text{ is the number of insertions at the character level,} && (9) \\ & N \text{ is the total number of characters in the reference text.} && (10) \end{aligned}$$

Using both WER and CER covers both lexical accuracy and fine-grained character-level errors.

Throughout training, the models were evaluated on a validation set, and the configuration that achieves the lowest WER was selected as the best-performing setup. The model was finally tested on unseen test data to check its real performance and generalization ability. These metrics were applied across all experiments to allow fair and reliable comparison between these models with training configurations.

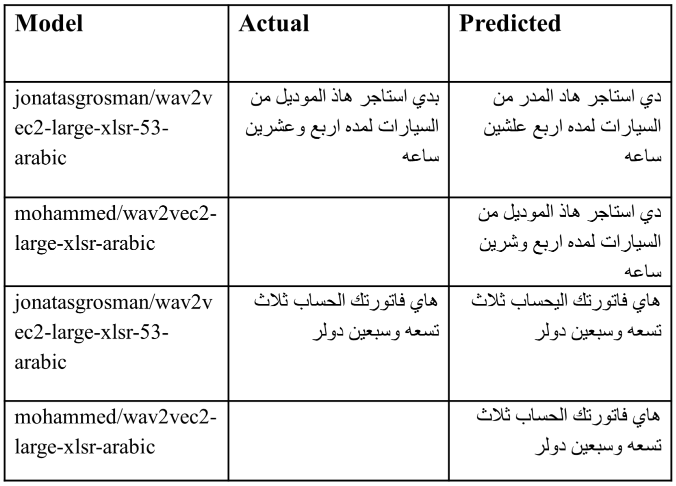
6) Experimental Procedure: The experimental procedure began by preprocessing the collected audio and text datasets. To ensure consistency, all audio samples were normalized and converted to a consistent format. The textual data were normalized to remove diacritics and punctuation, as well as to unify common Arabic letter variants, ensuring unified and clean text for model training.

After preprocessing, the dataset was split into training, validation and test sets. The training dataset was used to fine-tune the models, while the validation set was used to monitor performance and guide adjustments to the hyperparameters. The test set was only employed during the final evaluation to measure the models' ability to perform well on unseen data. Each model selected was fine-tuned on the initial dataset. To evaluate the impact of various hyperparameters on accuracy and overall performance, several training setups were evaluated for every model. WER and CER were used to track performance during training, allowing for the identification of overfitting and guiding changes to increase accuracy. The hyperparameter set that produced the greatest results on the validation set was chosen as the ideal model after training. The model's capacity for generalization was then assessed using the unseen test set. Lastly, the efficacy of each model and configuration in identifying Jordanian dialectal speech was evaluated by comparing the findings.

7) Results and Discussion: Table II provides an overview of the chosen ASR models' performance on the Jordanian dialectal speech dataset. WER and CER for two distinct training configurations of each model are used to report the results. These findings show that Jordanian dialectal patterns can be learned by both Wav2Vec 2.0 variants. However, hyperparameter settings have a significant impact on model performance. Specifically, the Mohammed / wav2vec2-large-xlsr-arabic model's second configuration shows a clear improvement, attaining performance on par with the Jonatas-grosman variation. These findings emphasize how crucial it is to carefully adjust hyperparameters when fine-tuning ASR models for dialectal speech.

TABLE II: WER and CER Results for Wav2Vec 2.0 Models on Jordanian Dialect Speech

Model	**Configuration**	**WER**	**CER**
jonatasgrosman/ wav2vec2-large-xlsr-53-arabic	1	0.70	0.24
jonatasgrosman/ wav2vec2-large-xlsr-53-arabic	2	0.77	0.27
mohammed/ wav2vec2-large-xlsr-arabic	1	0.87	0.34
mohammed/ wav2vec2-large-xlsr-arabic	2	0.70	0.23



Model	Actual	Predicted
jonatasgrosman/wav2vec2-large-xlsr-53-arabic	بدي استاجر هاذ الموديل من السيارات لمده اربع وعشرين ساعه	دي استاجر هاد المدر من السيارات لمده اربع علشين ساعه
mohammed/wav2vec2-large-xlsr-arabic		دي استاجر هاذ الموديل من السيارات لمده اربع وشرين ساعه
jonatasgrosman/wav2vec2-large-xlsr-53-arabic	هاي فاتورتك الحساب ثلاث تسعه وسبعين دولر	هاي فاتورتك اليحساب ثلاث تسعه وسبعين دولر
mohammed/wav2vec2-large-xlsr-arabic		هاي فاتورتك الحساب ثلاث تسعه وسبعين دولر

Fig. 4: Examples of Actual vs Predicted Sentences by ASR Model

The qualitative findings show that the two ASR models differ significantly from one another. The jonatasgrosman/wav2vec2-large-xlsr-53-arabic model appears to favor Modern Standard Arabic. Its performance in identifying dialectal expressions is negatively impacted by this bias, which results in substitutions and less natural word selections. Additionally, the model has trouble with compound numbers, which leads to noticeable numerical inaccuracies. The majority of its errors affect word forms and grammatical precision and are phonetic and morphological in nature. However, the transcriptions generated by the mohammed/wav2vec2-large-xlsr-arabic model are more in sync with the ground truth. Even when minor pronunciation errors occur, it handles word boundaries better and produces outputs that are more semantically accurate. All things considered, this model shows a greater capacity to represent the features of dialectal speech.

References

[1] B. Talafha, K. Kadaoui, S. M. Magdy, M. Habiboullah, C. M. Chafei, A. O. El-Shangiti, H. Zayed, M. C. Tourad, R. Alhamouri, R. Assi, A. Alraeesi, H. Mohamed, F. Alwajih, A. Mohamed, A. El Mekki, E. M. B. Nagoudi, B. D. M. Saadia, H. A. Alsayadi, W. Al-Dhabyani, S. Shatnawi, Y. Ech-chammakhy, A. Makouar, Y. Berrachedi, M. Jarrar, S. Shehata, I. Berrada, and M. Abdul-Mageed, "Casablanca: Data and models for multidialectal Arabic speech recognition," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 21 745–21 758. [Online]. Available: https://aclanthology.org/2024.emnlp-main.1211/

[2] K. Abu Kwaik, M. Saad, S. Chatzikyriakidis, and S. Dobnik, "Shami: A corpus of levantine arabic dialects," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018, pp. 1–6. [Online]. Available: https://aclanthology.org/L18-1576/

[3] UNESCO, "Arabic language, a bridge between civilizations," https://unesdoc.unesco.org/ark:/48223/pf0000380347, 2021, accessed: 2026-01-03.

[4] A. Dhouib, A. Othman, O. El Ghoul, M. K. Khribi, and A. Al Sinani, "Arabic automatic speech recognition: A systematic literature review," *Applied Sciences*, vol. 12, no. 17, p. 8898, 2022.

[5] E.-S. Badawi, "Levels of contemporary arabic in egypt," 1996.

[6] J. Dunn and S. Wong, "Stability of syntactic dialect classification over space and time," in *Proceedings of the 29th International Conference on Computational Linguistics*, N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, Eds. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 26–36. [Online]. Available: https://aclanthology.org/2022.coling-1.3/

[7] Q. Li, Q. Mai, M. Wang *et al.*, "Chinese dialect speech recognition: a comprehensive survey," *Artificial Intelligence Review*, vol. 57, no. 25, 2024. [Online]. Available: https://doi.org/10.1007/s10462-023-10668-0

[8] J. Lee, K. Kim, and M. Chung, "Korean dialect identification based on intonation modeling," in *2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2021, pp. 168–173.

[9] D. Alves, "Information theory and linguistic variation: A study of Brazilian and European Portuguese," in *Proceedings of the 12th Workshop on NLP for Similar Languages, Varieties and Dialects*, Y. Scherrer, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann, and M. Zampieri, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 9–19. [Online]. Available: https://aclanthology.org/2025.vardial-1.2/

[10] N. Sengupta, S. K. Sahu, B. Jia, S. Katipomu, H. Li, F. Koto, W. Marshall, G. Gosal, C. Liu, Z. Chen, O. M. Afzal, S. Kamboj, O. Pandit, R. Pal, L. Pradhan, Z. M. Mujahid, M. Baali, X. Han, S. M. Bsharat, A. F. Aji, Z. Shen, Z. Liu, N. Vassilieva, J. Hestness, A. Hock, A. Feldman, J. Lee, A. Jackson, H. X. Ren, P. Nakov, T. Baldwin, and E. Xing, "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models," 2023. [Online]. Available: https://arxiv.org/abs/2308.16149

[11] M. A. Menacer, O. Mella, D. Fohr, D. Jouvet, D. Langlois, and K. Smaili, "An enhanced automatic speech recognition system for Arabic," in *Proceedings of the Third Arabic Natural Language Processing Workshop*, N. Habash, M. Diab, K. Darwish, W. El-Hajj, H. Al-Khalifa, H. Bouamor, N. Tomeh, M. El-Haj, and W. Zaghouani, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 157–165. [Online]. Available: https://aclanthology.org/W17-1319/

[12] H. Alsayadi, A. Abdelhamid, I. Hegazy, and Z. Taha, "Data augmentation for arabic speech recognition based on end-to-end deep learning," *International Journal of Intelligent Computing and Information Sciences*, vol. 21, no. 2, pp. 50–64, 2021.

[13] M. M. Nasef, A. A. Elshall, and A. M. Sauber, "Arabrecognizer: modern standard arabic speech recognition inspired by deepspeech2 utilizing franco-arabic," *International Journal of Speech Technology*, vol. 27, no. 3, pp. 673–686, 2024.

[14] T. Zouhair, "Automatic speech recognition for low-resource languages using wav2vec2: Modern standard arabic (msa) as an example of a low-resource language," 2021.

[15] A. Babu, C. Wang, A. Tjandra, K. Gulati, F. Metze, and Y. Gu, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021. [Online]. Available: https://arxiv.org/abs/2111.09296

[16] A. Ali, S. Shon, Y. Samih, H. Mubarak, A. Abdelali, J. Glass, S. Renals, and K. Choukri, "The mgb-5 challenge: Recognition and dialect identification of dialectal arabic speech," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1026–1033.

[17] S. Shon, A. Ali, and J. Glass, "Convolutional neural networks and language embeddings for end-to-end dialect recognition," in *Proceedings of Odyssey 2018: The Speaker and Language Recognition Workshop*, 2018, pp. 98–104. [Online]. Available: https://arxiv.org/abs/1803.04567

[18] S. A. Chowdhury, A. Hussein, A. Abdelali, and A. Ali, "Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic ASR," in *Proceedings of Interspeech 2021*, 2021, pp. 2466–2470. [Online]. Available: https://arxiv.org/abs/2105.14779

[19] L. Grigoryan, N. Karpov, E. Albasiri, V. Lavrukhin, and B. Ginsburg, "Open automatic speech recognition models for classical and modern standard arabic," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, arXiv:2507.13977. [Online]. Available: https://arxiv.org/abs/2507.13977

[20] Z. Yang, Z. Wan, S. Li, C.-H. H. Yang, and C. Chu, "CoVoGER: A multilingual multitask benchmark for speech-to-text generative error correction with large language models," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, Eds. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 6313–6325. [Online]. Available: https://aclanthology.org/2025.emnlp-main.320/

[21] A. A. Safieh, I. A. Alhaol, and R. Ghnemat, "End-to-end jordanian dialect speech-to-text self-supervised learning framework," *Frontiers in Robotics and AI*, vol. 9, p. 1090012, 2022. [Online]. Available: https://doi.org/10.3389/frobt.2022.1090012

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.