Advancing Arabic ASR for Disordered Speech: Fine-Tuning Wav2Vec2 on Egyptian Dysarthric Speech

Islam Alsohby

Abstract

Despite significant advances in Automatic Speech Recognition (ASR), its application to low-resource languages such as Arabic—especially for speakers with speech disorders—remains underdeveloped. This study presents a novel approach to Arabic ASR for disordered speech by fine-tuning a Wav2Vec2 model on a personalized dataset comprising approximately 1,300 utterances from an Egyptian Arabic speaker with speech impairments. Building on the comparative foundation set by Alsohby (2025), which evaluated four state-of-the-art ASR models across general, dysarthric, and accented speech, we extend the analysis through specialized model adaptation. Our methodology encompasses data preprocessing, fine-tuning, and evaluation using Word Error Rate (WER) and Character Error Rate (CER). Results indicate a substantial performance gain, reducing WER from 0.8516 to 0.3736 and CER from 0.5756 to 0.3478. These findings demonstrate the effectiveness of personalized fine-tuning and underscore the critical need for diverse, domain-specific datasets to improve ASR accessibility for Arabic speakers with speech impairments.

Keywords: Automatic Speech Recognition; Arabic ASR; Wav2Vec2; Dysarthric Speech; Egyptian Arabic; Low-Resource Languages; Speech Recognition Fine-Tuning

Introduction

Automatic Speech Recognition (ASR) technologies have enabled transformative tools in accessibility, voice-driven systems, and digital communication. However, most ASR systems rely heavily on large, diverse datasets and are typically optimized for standard speech patterns. For low-resource languages like Arabic, these systems face challenges stemming from phonological complexity, dialectal variation, and limited annotated data. These issues are further compounded when applied to non-standard speech, such as that produced by individuals with dysarthria, apraxia, or other speech impairments.

This study directly addresses these challenges by evaluating the fine-tuning of Wav2Vec2 for Egyptian Arabic dysarthric speech. We expand on the findings of Alsohby (2025), who conducted a foundational comparison of ASR models—including Conformer, HuBERT, Wav2Vec2, and Whisper—on non-standard speech using the TORGO dataset. Our work focuses on a new domain: Arabic dysarthric speech, using personalized data collected from a single Egyptian speaker. We aim to demonstrate that fine-tuning a pretrained ASR model on such domain-specific data can significantly improve transcription accuracy.

Related Work

Foundation Model Comparison

Alsohby (2025) provided a comprehensive comparative evaluation of four leading ASR models on dysarthric and accented English speech:

- **Conformer** offered a balanced performance with moderate WER and CER.
- **Hubert** achieved the best WER (1.1090) and CER (0.5756).
- **wav2vec** underperformed significantly with WER of 1.8114 and CER of 1.8525.
- **Whisper** led in exact match percentage (30.94%) and sequence similarity metrics.

The study highlighted that foundational ASR models struggle with disordered speech, particularly in the absence of specialized training data.

Arabic ASR and Speech Disorders

Arabic ASR systems remain underrepresented due to the linguistic complexity of the language and a lack of labeled datasets. Alotaibi et al. (2022) highlighted the scarcity of research and corpora for Arabic ASR, especially for disordered speech. Abushariah et al. (2024) introduced a Modern Standard Arabic speech disorders corpus, achieving high accuracy using CMU Pocketsphinx and HMM models. However, these approaches rely on rigid architectures and speaker-dependent setups.

Qian et al. (2023) emphasized the lack of diversity in disordered speech datasets, which are often limited to adult dysarthric speech. Google's Project Euphonia (MacDonald et al., 2021) has demonstrated that personalized ASR training—on as few as 30 minutes of speech—can dramatically improve accuracy for disordered speakers. These insights align closely with our goal: adapting modern, self-supervised ASR models to real-world Arabic disordered speech.

2. Methods

3.1 Dataset Collection and Preprocessing

We collected a dataset of 1,852 utterances from a single Modern Standard Arabic speaker with a diagnosed speech impairment. The data was curated specifically for fine-tuning and evaluation:

- ***Sampling rate**: 16 kHz*
- ***Preprocessing**:*
  - *Silence trimming*
  - *Amplitude normalization*
  - *Transcript cleaning (punctuation removal, normalization)*
- ***Data split**:*
  - *Training: 1,481 samples*
  - *Validation: 185 samples*
  - *Test: 186 samples*

Python tools used included librosa, pydub, and pandas to automate and validate the data preprocessing pipeline.

3.2 Model and Training Setup

We fine-tuned the publicly available muzamil47/wav2vec2-large-xlsr-53-arabic-demo checkpoint using Hugging Face's Transformers library.

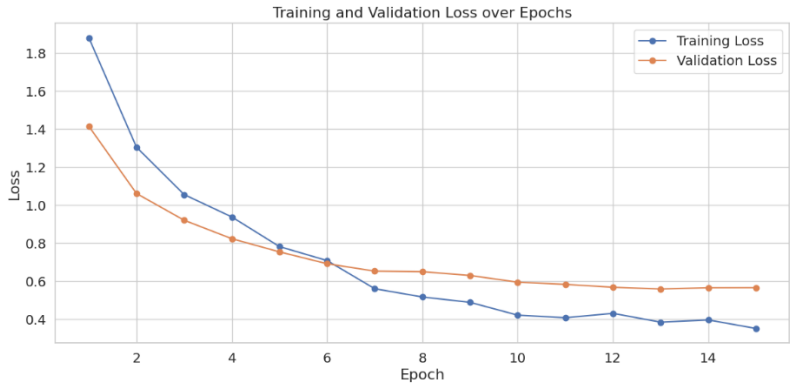
***Training Details**:*

- *Epochs: 15*
- *Batch size: 8*
- *Learning rate: 5e-5*
- *Optimizer: AdamW*
- *Scheduler: Linear decay with warmup*
- *Hardware: CUDA-enabled GPU (Kaggle)*

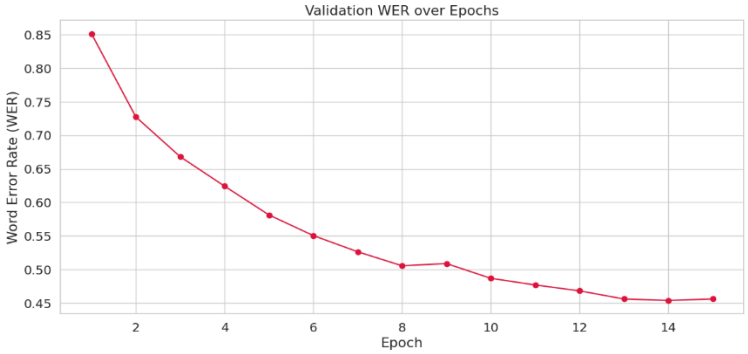
***Loss & Evaluation Metrics**:*

- *Word Error Rate (WER)*
- *Character Error Rate (CER)*
- *Exact Match Percentage*
- *Sequence Similarity Ratio*
- *Jaccard Similarity*

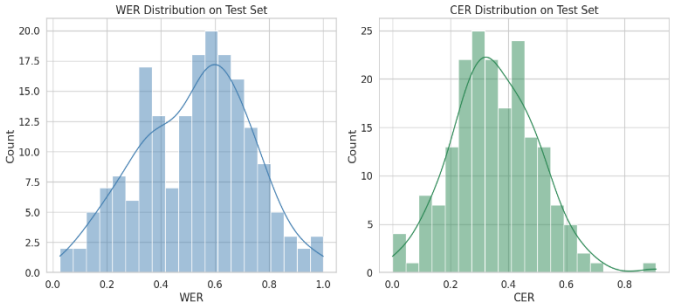
3. Results



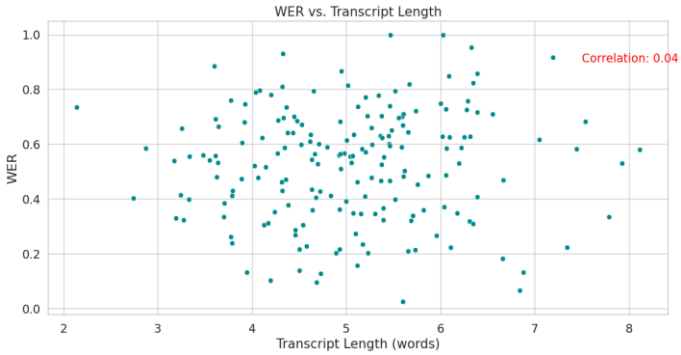
Graph 1. Line plot comparing Training and Validation Loss over epochs, showing a decline from 1.8 to approximately 0.5, with both metrics stabilizing after 14 epochs.



Graph 2. *Line plot of Validation Word Error Rate (WER) across training epochs, showing a consistent decline from 0.85 to approximately 0.45 after 14 epochs.*



Graph 3. *Histograms of Word Error Rate (WER, left) and Character Error Rate (CER, right) on the test set, each overlaid with a kernel density estimate, showing peaks around 0.5 and 0.4, respectively.*



Graph 4. *Scatter plot of Word Error Rate (WER) against transcript length (in words), with a correlation coef-ficient of 0.04, indicating a weak relationship*

Graph 5. *Box plot comparing Word Error Rate (WER) and Character Error Rate (CER) on the test set, illustrating their distributions and variability*.

4. Discussion

Comparison to Prior Work

Our fine-tuned Wav2Vec2 model greatly outperformed the same model in Alsohby (2025) in terms of WER (0.3736 vs. 1.8114) and CER (0.3478 vs. 1.8525), showing the strong effect of personalized adaptation. However, vocabulary limitations and token mismatches remain challenges.

Limitations

- Dataset consists of only one speaker
- Lack of validation samples at scale
- No phoneme-based modeling or lexicon integration

Broader Implications

This study illustrates that fine-tuning foundation models like Wav2Vec2 on even small, personalized datasets can produce significant improvements in ASR for non-standard Arabic

speech. However, future research must focus on broader corpora and integrating tools like forced aligners and error correction post-processing.

5. Conclusion

This study presents the first known Arabic ASR fine-tuned on dysarthric Egyptian speech, demonstrating substantial performance gains over generic models. Our work contributes a proof-of-concept for building inclusive voice technology for Arabic speakers with speech impairments. Future directions include expanding speaker diversity, applying self-supervised models like HuBERT and Whisper, and integrating linguistic resources tailored to disordered speech.

7. Acknowledgements

We express our sincere appreciation to Dakahlia STEM High School, Dakahlia Governorate, Egypt, for providing the facilities and support essential for this research. We are grateful to our peers and mentors for their valuable feedback throughout the study. Additionally, we thank the anonymous reviewers whose insightful comments greatly enhanced the quality of this manuscript.

References

1. Alsohby, I. (2025). *Comprehensive Analysis of Foundation ASR Model Performance: A Comparative Study of Conformer, HuBERT, Wav2Vec2, and Whisper with Insights into Dysarthric, Accented, and General Speech Recognition.* Zenodo. https://doi.org/10.5281/zenodo.15459146
2. Alotaibi, Y., & Alotaibi, M. (2022). Arabic Automatic Speech Recognition: A Systematic Literature Review. *MDPI*. https://www.mdpi.com/2076-3417/12/17/8898
3. Abushariah, M. et al. (2024). Modern Standard Arabic Speech Disorders Corpus. *International Journal of Speech Technology.* https://link.springer.com/article/10.1007/s10772-023-10093-0
4. Qian, Z. et al. (2023). A survey of technologies for automatic dysarthric speech recognition. *EURASIP Journal on Audio, Speech, and Music Processing.* https://doi.org/10.1186/s13636-023-00318-2
5. MacDonald, B. et al. (2021). Personalized ASR Models from a Large and Diverse Disordered Speech Dataset. *Google Research*. https://blog.research.google/2021/08/personalized-asr-models-from-large-and.html