ASR Models for Traditional Emirati Arabic: Challenges, Adaptations, and Performance Evaluation

Maha AlBlooki
Mohamed bin Zayed University of AI
Abu Dhabi, UAE
maha.alblooki@mbzuai.ac.ae

Kentaro Inui
Mohamed bin Zayed University of AI
Abu Dhabi, UAE
kentaro.inui@mbzuai.ac.ae

Shady Shehata
University of Waterloo
Ontario, Canada
shady.shehata@uwaterloo.ca

Abstract

Traditional Emirati Arabic, a culturally rich and linguistically distinct dialect, remains underrepresented in modern automatic speech recognition (ASR) systems. This paper addresses the gap by introducing a curated speech corpus derived from heritage broadcasts and literary sources, and by evaluating the performance of state-of-the-art ASR models on this low-resource dialect. We examine the zero-shot and fine-tuned performance of five pre-trained models—Wav2Vec2, XLS-R, Whisper, and Massively Multilingual Speech (MMS)—on our traditional Emirati Arabic dataset. Our results show that fine-tuning improves both Word Error Rate (WER) and Character Error Rate (CER), with MMS achieving the best results post-adaptation. Through detailed error analysis, we highlight challenges posed by dialectal morphology, phonology, and lexical variation, and propose targeted adaptations for dialect-specific ASR. This work establishes a foundational benchmark for traditional Emirati ASR and contributes to the broader goal of preserving linguistic heritage through speech technology.

1 Introduction

Automatic Speech Recognition (ASR) technologies have achieved remarkable performance in high-resource languages such as English and Mandarin. However, their effectiveness diminishes sharply for low-resource languages and dialects, particularly those with significant phonological and morphological variation. Arabic presents unique challenges in this regard, being a highly diglossic language with numerous regional dialects, many of which are underserved by current ASR systems.

Traditional Emirati Arabic is one such dialect. Rooted in the oral traditions of the United Arab Emirates, it retains linguistic features from Bedouin, coastal, and mountain communities that are increasingly overshadowed by Modern Standard Arabic (MSA) and urban Gulf variants. This dialect is not only linguistically distinct but also culturally significant, encoding idiomatic expressions, heritage knowledge, and regional identity.

Table 1: Linguistic Features of Traditional Emirati Dialect

Feature Type	**Example from Transcript**	**Description**
Phonological	جيه (Chaih), مب (mub)	ج pronounced as /ch/ (instead of /j/); consonant reduction from ما هو (ma huwa)
Morphological	بنتعاون, بنخبر, يسوونه	Prefix ب for future tense; Gulf-specific plural verb conjugation
Lexical	هلنا, اللّوليين, ربعنا, يارنا	Heritage terms for "our family", "elders", "neighbors", "our friends"
Syntactic	يوم بتكبر, هالشي	Use of يوم (yawm) for conditionals; contracted demonstrative هالشي
Discourse Markers	على طول, يعني	Filler word يعني ; Gulf expression على طول meaning "immediately"

Despite its value, traditional Emirati Arabic has been largely ignored in computational linguistics. Existing ASR systems are ill-equipped to handle its unique phonetic and lexical traits. To address this gap, we develop a dedicated speech dataset sourced from the *Alsanaa* (Dalmook, 2021) program and related literary content, and evaluate how modern ASR models perform on this data.

In this paper, we present:

- A curated traditional Emirati speech corpus with standardized transcription and preprocessing
- A comparative evaluation of five leading ASR models (Wav2Vec2, XLS-R, Whisper Small, Whisper Medium, MMS) in zero-shot and fine-tuned scenarios
- Insights into model-specific strengths and limitations for dialectal Arabic ASR.

2 Related Work

2.1 ASR for Arabic and Dialectal Variants

ASR systems have achieved remarkable progress for major world languages, yet robust solutions for Arabic dialects-particularly traditional Emirati Arabic-remain limited due to unique linguistic features and data scarcity. The Emirati dialect, with its distinct phonological and grammatical characteristics, poses significant challenges for ASR, especially given the lack of dedicated speech resources. Addressing such dialectal diversity is crucial for both technological inclusion and cultural preservation.

Recent advances in self-supervised learning (SSL) have enabled substantial improvements in ASR for low-resource languages and dialects. Models such as wav2vec2, HuBERT, and WavLM have demonstrated strong performance gains when fine-tuned on limited labeled data (Zhao and Zhang, 2022). Cross-lingual models, including XLS-R and Meta's MMS, further extend these capabilities, with XLS-R achieving impressive results even with as little as five minutes of training data in Indonesian language experiments (Sakti and Titalim, 2023). For Arabic, multilingual SSL models generally outperform monolingual approaches, as shown by Younis and Mohammad (2023), who report that fine-tuned XLS-R and MMS models achieve lower word error rates (WER) compared to monolingual baselines.

End-to-end models such as Whisper have also gained traction for their ability to generalize across languages. Talafha et al. (2023) benchmarked Whisper on multiple Arabic dialects, finding that while zero-shot performance often surpasses fully fine-tuned XLS-R models, significant drops occur for previously unseen dialects, including Emirati. The VoxArabica system further demonstrates the potential of SSL-based models for both dialect identification and ASR across a wide range of Arabic varieties (Waheed et al., 2023).

Hybrid approaches that combine deep learning with traditional phonetic modeling have also been explored. Dhouib et al. (2022) provide a systematic review of Arabic ASR research, highlighting the predominance of MSA-focused studies and the underrepresentation of dialectal variants. Novel architectures, such as CNN-LSTM with attention mechanisms, have shown promise for dialectal ASR, with Alsayadi et al. (2022) reporting improved WER on SASSC and MGB-3 datasets.

2.2 Low-Resource ASR Techniques

Transfer learning is a key strategy for improving ASR in low-resource settings. Elmahdy et al. (2014) utuilize MSA data to enhance recognition of under-resourced Arabic dialects, achieving notable WER reductions for Qatari Arabic. Data augmentation methods, including SpecAugment, synthetic speech, and self-training, have also proven effective. Bartelds et al. (2023) demonstrate that self-training and TTS-based augmentation consistently reduce WER for minority languages. Similarly, Khudhair and Talib (2022) show that combining data augmentation with language modeling yields competitive results for Arabic ASR.

Innovative data creation pipelines further address resource scarcity. Yeroyan and Karpov (2024) introduce a workflow for generating ASR datasets from audiobooks, enabling practical ASR development for languages with limited training data.

2.3 Datasets and Benchmarking

The development of high-quality datasets is foundational for Arabic ASR research. The Casablanca dataset covers eight Arabic dialects, including Emirati, and provides comprehensive annotations for benchmarking (Talafha et al., 2024). Mixat offers Emirati-English code-switching data, highlighting the challenges of bilingual ASR (Ali and Aldarmaki, 2024). SADA (Alharbi et al., 2024) and QASR (Mubarak et al., 2021) further expand resources for Gulf and multi-dialect Arabic speech, supporting supervised training and a range of speech and NLP tasks.

Efforts to benchmark code-switching ASR are exemplified by Hamed et al. (2022), who introduce a new Egyptian Arabic-English corpus and demonstrate the benefits of combining DNN-hybrid and Transformer approaches. Despite these advances,

challenges remain in achieving consistent evaluation and broad dialectal coverage.

2.4 Gaps and Motivation

While recent work has advanced ASR for Arabic and its dialects, systematic evaluation and adaptation of state-of-the-art pre-trained models for traditional Emirati Arabic remain largely unexplored. This study addresses this gap by benchmarking and fine-tuning leading ASR models on Emirati speech, aiming to identify effective strategies for robust dialectal ASR and contribute to the broader field of low-resource speech technology.

3 Dataset

To develop and evaluate ASR models for traditional Emirati Arabic, we curated a dialect-specific speech corpus sourced from *Alsanaa* (Dalmook, 2021) program, broadcast by *Aloula* station and supported by the Hamdan bin Mohammed Heritage Center. The dataset includes 102 MP3 audio files (approximately 4 hours) and their corresponding transcriptions, extracted from *alsanaa* book, a heritage literature book authored by Abdullah Bin Dalmook. These recordings capture authentic Emirati Arabic speech, preserving the dialectal nuances and linguistic patterns unique to the region.

Given the lack of existing Emirati ASR corpora, we aligned the audio and text manually, converting them into structured plain-text pairs. The dataset was partitioned into 80% training, 10% validation, and 10% test splits. Notably, the audio is spoken by a single male speaker, limiting speaker diversity but preserving dialectal authenticity.

من الأشياء الجميله إللي شفناها
عند هلنا الأوليين إنهم يصالحون
بين العرب. يعني إذا اثنينه متزاعلين
ولاّ اثنينه متضاربين ولاّ شي،
ساروا وصالحوا بينهم.

Figure 1: Sample transcription

Preprocessing included:

- Diacritics and punctuation removal to standardize transcriptions
- Audio cropping (removal of non-speech intro/outro segments)
- Mono conversion and resampling to 16 kHz
- Normalization to standardize amplitude levels

This dataset captures phonological, morphological, and lexical features unique to traditional Emirati Arabic and serves as a foundational resource for dialect-specific ASR. The full dataset and preprocessing pipeline are available online.[1]

4 Models and Training

We adopt a comparative experimental framework to evaluate the performance of state-of-the-art ASR models on traditional Emirati Arabic. Our approach consists of two main stages: zero-shot evaluation and fine-tuning.

4.1 Model Selection

We evaluated five pre-trained ASR architectures, each fine-tuned or adapted for Emirati Arabic or closely related dialects:

Wav2Vec 2.0 (`eabayed/wav2vec2emiratidialict`$_1$) Wav2Vec 2.0 is a self-supervised learning framework for speech recognition that learns audio representations via a contrastive task, enabling strong performance with limited labeled data (Baevski et al., 2020). Its architecture combines a convolutional feature encoder with a Transformer network, allowing effective modeling of phonetic and lexical features in low-resource settings. The model used here is further adapted to Emirati Arabic using audio from regional media, resulting in 315 million parameters and improved recognition of dialectal nuances.

XLS-R (`jonatasgrosman/wav2vec2-large-xlsr-53-arabic`) XLS-R extends Wav2Vec 2.0 to the multilingual domain, pre-trained on over 436,000 hours of speech in 128 languages (Babu et al., 2021). This enables robust cross-lingual transfer and strong performance on low-resource dialects. The Arabic-adapted variant, with 315 million parameters, is fine-tuned on Common Voice 6.1 and the Arabic Speech Corpus, making it well-suited for Emirati Arabic (Babu et al., 2021).

Whisper Small (`ayoubkirouane/whisper-small-ar`) Whisper is a transformer-based encoder-decoder ASR system trained on diverse multilingual data

[1] https://github.com/MahaAlBlooki/alsanaa-emirati-dataset

(Radford et al., 2022). The small Arabic model (241M parameters) is fine-tuned on the Mozilla Common Voice v11 dataset for Arabic, and further adapted for Emirati speech, balancing efficiency with accuracy.

Whisper Medium (`Seyfelislem/whisper-medium-arabic`) This variant of Whisper, with 763 million parameters, is optimized for Arabic speech recognition. Fine-tuning on Emirati data enhances its ability to transcribe dialectal speech, leveraging the robust encoder-decoder architecture of Whisper.

MMS (`facebook/mms-1b-all`) Massively Multilingual Speech (MMS) is a self-supervised model trained on over 1,000 languages, including Arabic dialects (Zhang et al., 2023). With 965 million parameters, MMS is designed for broad language coverage and demonstrates strong zero-shot and few-shot ASR capabilities. While not specifically fine-tuned for Emirati Arabic, its multilingual training enables generalization to underrepresented dialects.

Each model was evaluated in two modes:

- **Zero-shot inference**: Direct evaluation without further training on our dataset.
- **Fine-tuning**: Models were adapted to the Emirati dataset using transfer learning.

4.2 Fine-Tuning Strategy

Fine-tuning involved freezing most pretrained layers and training only the final layers (e.g., projection heads and classification layers). The following configuration was used:

- **Optimizer**: AdamW with weight decay
- **Learning rate schedule**: Linear warm-up followed by decay
- **Batch size**: Adjusted per model based on memory constraints
- **Epochs**: Trained until validation loss convergence (early stopping applied)
- **Data augmentation**: Speed perturbation and SpecAugment to improve generalization
- **Gradient accumulation**: Enabled to simulate larger batch sizes on limited hardware

5 Evaluation

5.1 Metrics

We use two standard ASR metrics:

- **Word Error Rate (WER)**: Percentage of word-level errors (insertions, deletions, substitutions).
- **Character Error Rate (CER)**: Measures character-level discrepancies; useful for morphologically rich languages and dialects.

Both metrics were calculated on the validation and test splits of our Emirati dataset.

5.2 Evaluation Protocol

All models were tested directly on the test set without any adaptation to measure out-of-the-box generalization. After training, models were evaluated on the same test set to assess improvements in recognition accuracy.

5.3 Qualitative Analysis

Beyond quantitative metrics, we conducted a qualitative error analysis focused on the recognition of dialect-specific lexical items, morphological transformations (e.g., future tense prefixes), and common phonological shifts (e.g., hamza deletion, /j/ $\rightarrow$ /ch/ substitutions).

6 Results

We evaluated the zero-shot and fine-tuned performance of several state-of-the-art pre-trained ASR models-Wav2Vec 2.0, XLS-R, Whisper (small and medium), and MMS-on traditional Emirati Arabic speech. Performance was measured using WER and CER, providing insight into both word-level and subword recognition accuracy.

6.1 Baseline Performance

In the zero-shot setting in Table 2, Wav2Vec 2.0 achieved the best results among all models, with a WER of 46.50% and CER of 17.13%. This suggests that its self-supervised pre-training enables effective generalization to unseen dialects, capturing phonetic patterns even when word-level recognition is challenging. MMS ranked second (WER 67.21%, CER 24.56%), likely benefiting from its broad multilingual training and explicit support for Arabic dialects. XLS-R, despite its cross-lingual design, performed poorly (WER 88.26%, CER

Model	WER (%)	CER (%)
Wav2Vec 2.0	46.50	17.13
XLS-R	88.26	40.37
Whisper Small	93.06	81.02
Whisper Medium	86.10	75.01
MMS	67.21	24.56

Table 2: Average WER and CER on the whole dataset in baseline inference

40.37%), indicating potential limitations in its coverage of Gulf Arabic and a significant domain gap when applied to Emirati speech. Whisper models showed the weakest zero-shot performance, with Whisper Small reaching 93.06% WER and 81.02% CER, and Whisper Medium slightly better at 86.10% WER and 75.01% CER. The high CER values for Whisper indicate substantial difficulties at the character level, likely due to mismatches between the pre-training data and the phonological characteristics of Emirati Arabic.

Qualitative analysis of model errors revealed that models often misrecognized dialect-specific vocabulary and morphemes, with XLS-R and Whisper in particular producing transcriptions influenced by other Arabic dialects. For example, XLS-R frequently substituted Emirati morphemes with those more typical of Egyptian or Levantine Arabic, reflecting gaps in dialectal representation in the pre-training corpus.

These results highlight the challenges of recognizing traditional Emirati Arabic with existing ASR models and underscore the importance of dialect-specific adaptation. The findings establish a benchmark for future work and inform model selection and adaptation strategies for low-resource dialectal ASR, with broader implications for Arabic speech technology research

6.2 Fine-Tuned Performance

Table 3 summarizes the impact of fine-tuning each ASR model on the Emirati *Alsanaa* dataset. Fine-tuning led to substantial performance gains for some architectures, while others showed limited or even negative adaptation.

MMS exhibited the most pronounced improvement, with WER dropping from 67.21% to 41.04% and CER from 24.56% to 13.34%. This 26.17 and 11.22 percentage point reduction in WER and CER, respectively, highlights the effectiveness of MMS's multilingual pre-training in facilitating rapid adaptation to low-resource dialects. After fine-tuning, MMS outperformed all other models, establishing a new benchmark for Emirati Arabic ASR.

Wav2Vec 2.0 also benefited from fine-tuning, achieving a modest reduction in WER (from 46.50% to 44.30%) and CER (from 17.13% to 15.96%). The relatively small improvement suggests that the model's self-supervised representations already captured much of the dialectal variation present in the dataset, resulting in stable performance before and after adaptation.

In contrast, XLS-R's performance deteriorated after fine-tuning, with WER rising from 88.26% to 89.78% and CER from 40.37% to 42.31%. This decline may indicate overfitting to the limited training data or challenges in adapting broad cross-lingual representations to specific dialectal features, a phenomenon noted in low-resource ASR adaptation literature.

The Whisper models showed mixed results. Fine-tuning Whisper Small led to further degradation, with WER exceeding 100% (100.04%); it seemed like the model was encountering a repetition or loop behavior at the end of some transcriptions. For instance, in one of the transcriptions, تومیه, a gibberish prediction of what is supposed to be توايهوا (*cheek-kissed*) is repeated many times consecutively, which isn't in the original text. This type of repetition has artificially inflated the WER of Whisper Small model. On the other hand, the CER increased to 75.53%, suggesting substantial insertion errors and a mismatch between model architecture and the Emirati dialect under data-scarce conditions. Whisper Medium showed only marginal change, with WER shifting from 86.10% to 88.60% and persistently high CER, indicating that additional data or specialized adaptation techniques may be required for effective dialectal ASR with Whisper.

Overall, these results underscore the importance of model selection and adaptation strategy for low-resource dialectal ASR. While MMS demonstrates strong adaptability to Emirati Arabic, other architectures may require more sophisticated fine-tuning or larger datasets to achieve competitive performance.

6.3 Error Analysis

A detailed error analysis reveals notable differences in how each model adapts to traditional Emirati Arabic, highlighting both architectural

كالشي يعدنا هنيه له سلام خاصبه
و اليوم انت اندخلت على عرب
و ريته و يتغدون يحتاي
ان تقول لهم هنهم
و هي دعوه من الله ان يهنيهم بالأكل
و هم يحتاي يردون عليهك
و يقولون و منهم و المعنى انه انت
قربوا يا نا و تغدوا يا نا
و تانه هم ممتغدّي
و اللي مبماكلت را بيتغدّه
يا عمو تاني متغدّي ما بيرو مزيد
لكنه الاصل انه هذا هو الشلام
و السلام عقب الغداء انه تغدّه و خلّص
و انه ما تغدّه و خلّص
و انه جافضه من الغداء نشو
توميه و توميه و توميه و توميه
و توميه و توميه و توميه و توميه
و توميه و توميه و توميه و توميه
و توميه و توميه و توميه و توميه

Figure 2: Sample transcription with decoding repetition

strengths and persistent challenges. Models based on self-supervised pre-training, such as MMS and Wav2Vec 2.0, consistently outperformed Whisper variants, suggesting that phonetic representation learning is more effective for dialectal ASR than multitask training approaches.

A key observation is the relationship between zero-shot and fine-tuned performance: models with strong zero-shot results (e.g., Wav2Vec 2.0) exhibited only modest improvements after fine-tuning, while models with moderate zero-shot performance (e.g., MMS) showed substantial gains. This pattern

Model	**WER (%)**	**CER (%)**
Wav2Vec 2.0	44.3	15.96
XLS-R	89.78	42.31
Whisper Small	100.04	75.53
Whisper Medium	88.60	72.03
MMS	41.04	13.34

Table 3: Average WER and CER on test set after fine-tuning

underscores the importance of evaluating both generalization and adaptability when selecting ASR architectures for low-resource dialects.

Across all models, CER was consistently lower than WER, indicating that character-level recognition is more robust than word-level recognition. This discrepancy, especially pronounced in Wav2Vec 2.0 and MMS, suggests that while phonetic patterns are captured effectively, models struggle with accurate word segmentation and lexical reconstruction. Integrating language models during post-processing may help mitigate these issues.

Architectural differences also affected data efficiency. MMS demonstrated high data efficiency, achieving significant improvements with limited Emirati data, whereas XLS-R and Whisper required more extensive adaptation to yield comparable results. Notably, fine-tuned Whisper Small frequently truncated longer utterances, omitting culturally salient content and narrative details. Additionally, repetition errors were observed, with the model generating nonsensical word sequences, artificially inflating the WER.

Dialectal specialization remains a significant challenge. Even after fine-tuning, high error rates persisted-particularly for Whisper Small and XLS-R, which are primarily pre-trained on Egyptian or MSA data. These models often substituted Emirati morphemes with forms from other dialects, reflecting insufficient representation of Gulf Arabic in the pre-training corpus. Furthermore, inconsistent diacritization in XLS-R outputs, despite ground-truth normalization, introduced additional errors.

These findings emphasize the need for careful model selection, larger dialectal datasets, and potentially pre-training strategies tailored to Gulf Arabic. The persistent performance gaps highlight the ongoing challenge of developing inclusive ASR technologies for underrepresented dialects, underscoring the importance of both technical innovation and investment in dialectal language resources.

7 Limitations

This work faces several limitations, one of which is dataset diversity. The dataset includes a single speaker (male), limiting phonetic and demographic diversity. This may bias model performance toward that speaker's vocal and dialectal traits. Another limitation is duration. With only 4 hours of audio, the dataset is relatively small, constraining model generalization. Additionally, dialect cover-

age is limited. While rich in traditional features, the dataset does not fully represent all sub-dialectal varieties across the UAE (e.g., eastern vs. western tribal variants). Moreover, the evaluation scope of WER and CER focused on transcription accuracy, without assessing downstream tasks, such as speaker identification or sentiment analysis.

Future work should explore speaker diversity, cross-dialectal robustness, and larger-scale datasets.

8 Conclusion

This paper presents the first ASR benchmark for traditional Emirati Arabic, a linguistically and culturally significant but technologically underserved dialect. By compiling a novel dataset and evaluating state-of-the-art ASR models in both zero-shot and fine-tuned settings, we demonstrate the value of transfer learning and domain-specific adaptation.

Our results demonstrate that self-supervised models with strong multilingual pre-training, particularly MMS, achieve superior adaptability and performance after fine-tuning, while other architectures exhibit varying degrees of success. The persistent gap between character- and word-level accuracy underscores the need for improved modeling of dialectal lexical and phonological features.

This work contributes to Arabic dialectal ASR research and highlights the role of speech technology in preserving oral heritage. We release our dataset and preprocessing tools to encourage further research on Gulf Arabic ASR.

9 Ethics Statement

This research adheres to the ACL Ethics Policy. All audio recordings used in this study were publicly available and sourced from cultural heritage broadcasts and literary materials produced by the Hamdan bin Mohammed Heritage Center. Proper credit has been given to the original author, Abdullah Bin Dalmook, whose work was used with the intent of preserving linguistic and cultural heritage.

The dataset features speech from a single speaker who is a public broadcaster and author. No personally identifiable or sensitive information is included. The goal of this research is to support inclusive technology and cultural preservation, not surveillance or misuse.

We acknowledge the potential risks of dialectal ASR systems being misused for sociolinguistic profiling or discrimination. To mitigate this, our work is released with a cultural preservation focus, encouraging ethical use in academic and heritage documentation contexts.

References

Sadeen Alharbi, Areeb Alowisheq, Zoltán Tüske, Kareem Darwish, Abdullah Alrajeh, Abdulmajeed Alrowithi, Aljawharah Bin Tamran, Asma Ibrahim, Raghad Aloraini, Raneem Alnajim, Ranya Alkahtani, Renad Almuasaad, Sara Alrasheed, Shaykhah Alsubaie, and Yaser Alonaizan. 2024. Sada: Saudi audio dataset for arabic. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10286–10290.

Maryam Al Ali and Hanan Aldarmaki. 2024. Mixat: A data set of bilingual emirati-english speech.

Hamzah A. Alsayadi, Salah Al-Hagree, Fahd A. Alqasemi, and Abdelaziz A. Abdelhamid. 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)*, pages 1–8.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation.

Abdullah Hamdan Bin Dalmook. 2021. *Alsanaa*. Hamdan bin Mohammad Heritage Center, Dubai. Paper Cover; Dimensions: 24x17 cm.

Amira Dhouib, Achraf Othman, Oussama El Ghoul, Mohamed Koutheair Khribi, and Aisha Al Sinani. 2022. Arabic automatic speech recognition: A systematic literature review. *Applied Sciences*, 12(17).

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2014. Development of a TV broadcasts speech recognition system for qatari Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3057–3061, Reykjavik, Iceland. European Language Resources Association (ELRA).

Injy Hamed, Pavel Denisov, Chia-Yu Li, Mohamed Elmahdy, Slim Abdennadher, and Ngoc Thang Vu.

2022. Investigations on speech recognition systems for low-resource dialectal arabic–english code-switching speech. *Computer Speech Language*, 72:101278.

Mohanad Khudhair and Ahmed Talib. 2022. Improving low resources arabic speech recognition using data augmentation. In *2022 Fifth College of Science International Conference of Recent Trends in Information Technology (CSCTIT)*, pages 60–65.

Hamdy Mubarak, Amir Hussein, Shammur Absar Chowdhury, and Ahmed Ali. 2021. Qasr: Qcri aljazeera speech resource – a large scale annotated arabic speech corpus.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Sakriani Sakti and Benita Angela Titalim. 2023. Leveraging the multilingual indonesian ethnic languages dataset in self-supervised models for low-resource asr task. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Bashar Talafha, Karima Kadaoui, Samar Mohamed Magdy, Mariem Habiboullah, Chafei Mohamed Chafei, Ahmed Oumar El-Shangiti, Hiba Zayed, Mohamedou cheikh tourad, Rahaf Alhamouri, Rwaa Assi, Aisha Alraeesi, Hour Mohamed, Fakhraddin Alwajih, Abdelrahman Mohamed, Abdellah El Mekki, El Moatez Billah Nagoudi, Benelhadj Djelloul Mama Saadia, Hamzah A. Alsayadi, Walid Al-Dhabyani, Sara Shatnawi, Yasir Ech-Chammakhy, Amal Makouar, Yousra Berrachedi, Mustafa Jarrar, Shady Shehata, Ismail Berrada, and Muhammad Abdul-Mageed. 2024. Casablanca: Data and models for multidialectal arabic speech recognition.

Bashar Talafha, Abdul Waheed, and Muhammad Abdul-Mageed. 2023. N-shot benchmarking of whisper on diverse arabic speech recognition.

Abdul Waheed, Bashar Talafha, Peter Sullivan, Abdel-Rahim Elmadany, and Muhammad Abdul-Mageed. 2023. Voxarabica: A robust dialect-aware arabic speech recognition system.

Ara Yeroyan and Nikolay Karpov. 2024. Enabling asr for low-resource languages: A comprehensive dataset creation approach.

Hiba Adreese Younis and Yusra Faisal Mohammad. 2023. Arabic speech recognition based on self supervised learning. In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, pages 528–533.

Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, Jungdam Won, Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, Jessica K. Hodgins, Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, Christopher Ré, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Jing Zhao and Wei-Qiang Zhang. 2022. Improving automatic speech recognition performance for low-resource languages with self-supervised models. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1227–1241.