Automatic Speech Recognition for Low-Resourced Middle Eastern Languages

Razhan Hameed[1,*], *Sina Ahmadi*[2,*], *Hanah Hadi*[3], *Rico Sennrich*[2]

[1]Vox AI, Netherlands
[2]Department of Computational Linguistics, University of Zurich, Switzerland
[3]Paitaxt Technical Institute, Kurdistan Regional Government, Iraq

Correspondence: sina.ahmadi@uzh.ch

Abstract

Despite significant advancements in language and speech technologies, many languages in the Middle East remain underserved, leading to a technological disparity that negatively impacts these languages. This paper presents a pioneering effort to address this issue by focusing on speech technologies for low-resourced languages in the Middle East. We introduce a community-driven volunteer-based initiative to collect audio recordings for six languages spoken by an estimated population of 30 million speakers. Through this initiative, we collect over 40 hours of speech data, with 75% of utterances based on multilingual parallel corpora. In our experiments, we demonstrate the impact of data collection and fine-tuning models on the performance of speech technologies for these languages. This research serves as a crucial step towards preserving and promoting linguistic diversity in the Middle East while ensuring equal access to speech technologies for all language communities.

Index Terms: audio data collection, automatic speech recognition, low-resourced languages

DOLMA-NLP/ASR

1. Introduction

Known for its historical and geopolitical importance, the Middle East[1] has been a nexus of cultural exchange and linguistic innovation for millennia. Beyond the officially recognized languages that dominate public discourse, namely Arabic, Persian, Turkish, and Hebrew, this region of over 400 million inhabitants encompasses a remarkable spectrum of linguistic diversity spanning the Afroasiatic, Indo-European, Caucasian, and Turkic language families. Each of these languages represents distinct linguistic traditions that have evolved over centuries of cultural interaction and exchange. However, the vitality of many of these languages faces severe challenges in the modern era due to restrictive language policies, coupled with systematic marginalization and assimilation campaigns [1, 2, p.1187]. As such, numerous varieties have been pushed to the brink of extinction, with some communities experiencing complete language loss within a single generation. In fact, UNESCO has classified many of these languages as endangered, with particularly concerning trends showing accelerated decline as younger generations increasingly shift away from their heritage languages [3]. This situation is further exacerbated by the digital divide, where the absence of language technologies for these varieties reinforces existing patterns of sociolinguistic inequality and cultural marginalization, creating a self-perpetuating cycle of linguistic disadvantage.

The development of language and speech technologies presents a crucial opportunity for under-represented languages, offering potential solutions to these pressing challenges. Most of such languages lack standardized written forms and formal documentation, while their speakers often have limited literacy in their native languages due to historically restricted educational access and systematic discrimination [4, 5]. This complex sociolinguistic situation makes automatic speech recognition (ASR) particularly relevant as it can enable digital participation without requiring advanced written literacy, a significant barrier for many speakers. ASR systems can facilitate digital inclusion by allowing speakers to interact with technology through speech, creating pathways for participation that align with oral communication practices while preserving traditional modes of interaction. Additionally, these languages present distinct challenges for ASR development, including severe data scarcity, lack of standardization, and the complexities of unwritten languages [6]. Despite these challenges, the development of such technologies could serve as a vital tool for language preservation and revitalization, potentially reversing decades of linguistic marginalization [7].

This paper contributes to expanding speech technologies for Middle Eastern languages, promoting fair and inclusive speech science and technology in a region where such developments are critically needed. We present pioneering work in collecting speech data for six under-represented languages in the region, namely **Gilaki**, **Laki Kurdish**, **Hawrami**, **Mazandarani**, **Southern Kurdish** and **Zazaki**, through three complementary approaches: professional studio recordings with native speakers, a community-driven method that engages local contributors through a Telegram bot to enrich utterances in a multilingual parallel corpus with audio recordings, and finally, transcribing radio and TV programs. Our methodology specifically addresses the challenges of data collection in politically sensitive contexts while ensuring high-quality, community-driven contributions. Through extensive experimentation, we demonstrate the limitations of current state-of-the-art models in handling these languages and present comprehensive results of fine-tuned models, highlighting both achievements and areas requiring further development. Our work not only addresses immediate technological gaps but also establishes a foundation for future developments, including direct speech-to-speech translation and more sophisticated language preservation tools. By documenting both our methodologies and challenges, we aim to facilitate further research and development in this crucial yet under-served area of language and speech technology, potentially serving as a model for similar efforts in other linguistically diverse regions facing comparable challenges.

[*]These authors contributed equally.

[1]While acknowledging that the term "Middle East" lacks precise geographical boundaries and its definition varies across different contexts, we use it in this paper to broadly refer to the region of Western Asia.

2. ASR for Middle Eastern languages

ASR has been a focus of research in the Middle East for over two decades, primarily concentrated on the region's dominant languages as for Turkish [8] and Arabic [9]. Early work utilized traditional frameworks such as the Carnegie Mellon University Sphinx engine [10] and KALDI Speech Recognition [11], while recent advances have seen the emergence of state-of-the-art multilingual approaches like Whisper [12] and Seamless [13]. Despite the region's rich linguistic landscape, ASR development remains severely limited in scope, with recent attention expanding to include Northern and Central Kurdish [14, 15].

A fundamental challenge across languages in the Middle East is data scarcity, though the severity varies significantly [16, 17]. Previously, diverse strategies have been employed for data collection, each with distinct advantages and limitations. Traditional studio recording, while ensuring the highest quality, is resource-intensive and requires professional speakers. Alternative approaches include utilizing broadcast news reports [18], extracting content from YouTube [19], and implementing various crowd-sourcing methods. Platforms like Mozilla Common Voice [20] have democratized ASR development by providing robust collection and validation functionalities. Similarly, the proliferation of smartphones has enabled novel data collection methods [21], while messaging platforms like Telegram have proven effective for languages such as Kazakh [22], Uzbek [23], and Central Kurdish [24].

Our work addresses critical gaps in ASR development for under-represented Middle Eastern languages. Building on our previous community-driven initiative to create parallel corpora discussed in [25], we now expand these resources by incorporating audio recordings. Among our target languages, only Zazaki has any presence in Common Voice, with a mere three hours of recorded speech. The remaining languages in our study lack any substantial speech resources for ASR development. We contribute to filling this void by combining professional studio recordings with community-driven contributions. Furthermore, we demonstrate the effectiveness of fine-tuning pretrained Whisper models for these languages, establishing baseline and identifying areas for improvement.

3. Dataset

3.1. Volunteer-based Community Recordings via Telegram

Our data collection approach builds upon our previous efforts in creating parallel corpora for under-resourced Middle Eastern languages [25]. In that work, we successfully collected over 35,000 translations across eight languages by leveraging a high-resource language (English) paired with another language (Persian), resulting in trilingual corpora suitable for evaluating and fine-tuning machine translation models. To extend these resources with speech data, we leveraged the same community network we had established during the translation phase.

While studio recording offers optimal audio quality, financial constraints made it impractical for widespread data collection across our target languages. Instead, we develope a Telegram bot that implements a systematic data collection workflow. The process begins with users selecting their language and specifying their gender to keep track of demographic balance in the dataset. The bot then guides users through a structured recording process, presenting sentences from the parallel corpora for audio capture. Although we initially targeted all the languages in the corpus, i.e. Luri Bakhtiari, Gilaki, Hawrami, Laki Kurdish, Mazandarani, Southern Kurdish, Talysh and Zazaki, only Gilaki, Hawrami, Laki Kurdish, Mazandarani, Southern Kurdish and Zazaki received attention from the community. We run the data collection campaign over a period of five weeks with constant support and daily monitoring throughout the process.

The recording interface was designed with quality control in mind. Users can record each sentence multiple times if needed, replay their recordings for verification, and must explicitly approve their submissions before moving forward. Before beginning any recordings, users are presented with a clear agreement window outlining privacy protections and terms of use, which emphasizes that while contributed recordings will be released as open-source resources, all personal information remains strictly confidential. Upon successful submission, the bot automatically presents the next sentence, maintaining a smooth and efficient recording flow. To further ensure recording quality, automated heuristic checks verify each submission's length, filtering out recordings that are suspiciously short or long. Additionally, we conduct random manual reviews of recordings from each contributor as an extra quality assurance measure.

This community-driven approach through Telegram proved particularly effective given the platform's widespread use in the region and its robust audio recording capabilities. The integration with our existing parallel corpora ensures that the collected audio data aligns with our previously validated textual resources, creating a rich multi-lingual and multi-modal dataset for these under-resourced languages. The resulting dataset comprises 24,500 total records, including utterances recorded more than once, among which 17,600 are unique. Given that utterances are aligned with translations in English, we further enrich the dataset with English recordings generated automatically via open-source Kokoro text-to-speech (TTS) model.[2] Finally, each entry in the dataset includes metadata such as dialect, recorder's numeral ID, and link to the English TTS counterparts, enabling potential future use of the dataset for projects like speech-to-speech and inter-language translation.

3.2. Studio Recording

In a parallel effort, we record Hawrami speakers over a five-week period in controlled environments including professional studios and quiet office spaces. 102 native Hawrami, mostly university students, participated in this phase who represent diverse demographics, with ages ranging from 4 to 90 years. Each session was carefully monitored for audio quality, with noise levels measured using Adobe Audition and Audacity. The recordings are standardized to WAV format (16,000 Hz, 16-bit mono). This methodical approach yielded approximately four hours of speech data, comprising 4,977 sentences.

3.3. Radio and TV Transcription

Additionally for Hawrami, we collect and transcribe broadcast content from various media sources. This portion of the corpus includes 830 utterances, totaling 100 minutes of speech recorded from local public radio and TV programs. The content spans diverse genres including news, religious texts, poetry, daily conversations, and cultural programs, with novels and poetry being the most represented categories. Speakers in these recordings represent various regions in Iraq and Iran with the largest speaker groups coming from Balkha (36.27%), Kharpani (21.57%), and Tawela (17.65%) in Iraqi Kurdistan.

The two previous approaches could not be explored for all languages due to limited on-site access to contributors.

[2] https://huggingface.co/hexgrad/Kokoro-82M

Table 1: *Dataset statistics per language, showing sentence counts, lengths (Len.), duration (Dur.), speaker demographics, and average (ave.) utterance characteristics for train and test splits in character (c), second (s) and hour (h).*

Language	Split	Sentences	Duration (h)	Speakers	Male (%)	Female (%)	Avg. Len. (c)	Avg. Dur. (s)
Gilaki	train	2,961	5.38	17	70.6	29.4	49.10	6.55
	test	625	0.92	14	71.4	28.6	38.40	5.29
Hawrami	train	10,166	18.30	17	82.4	17.6	50.70	6.48
	test	1,263	1.91	13	84.6	15.4	42.10	5.44
Hawrami (Studio)	train	4,773	3.89	99	56.6	43.4	33.40	2.94
	test	204	0.16	19	68.4	31.6	31.80	2.78
Hawrami (TV & Radio)	train	830	1.40	11	63.6	36.4	76.73	6.08
Laki Kurdish	train	755	0.86	5	80.0	20.0	38.90	4.11
	test	313	0.37	4	75.0	25.0	39.40	4.25
Mazanderani	train	875	0.95	6	66.7	33.3	37.80	3.91
	test	249	0.24	5	60.0	40.0	34.10	3.53
Southern Kurdish	train	5,912	6.92	14	85.7	14.3	37.40	4.21
	test	757	0.89	8	87.5	12.5	38.90	4.25
Zazaki	train	201	0.22	3	100.0	0.0	44.60	3.94
	test	50	0.05	3	100.0	0.0	41.30	3.68
All	train	25,643	36.52	161	76.6	23.4	43.79	5.11
	test	3,461	4.54	66	82.2	17.8	39.27	4.72

4. Experimental Setup

4.1. Dataset

To ensure compatibility between our speech corpus and the existing parallel corpora, we use the test sets introduced in [25] as the foundation for our data splits ensuring uniform representation across dialects while maintaining consistent orthography. Specifically, we identify audio recordings that correspond to sentences in the parallel corpora's test sets and designate these as our test set. All remaining audio-text pairs were allocated to the training set, with validation data merged into training to maximize the limited data available in this low-resource scenario. This approach maintains consistent evaluation benchmarks across both speech and text tasks while making optimal use of the available data. We preprocess all text by removing extraneous whitespace, standardizing punctuation marks, and normalizing various Unicode characters to their canonical forms using KLPT [26] and regular expressions.

4.2. Baseline

In our study, we rely on two variants of Whisper [12]: the `Base` variant containing 74 million parameters and the `Small` variant containing 244 million parameters. To establish baseline performance, we evaluate untuned Whisper on our languages using orthographically related high-resource languages: Turkish for Zazaki as they both use a Latin-based script and Persian (Farsi) for other languages which use an Arabic-based script. This zero-shot baseline achieves WERs between 90-96%, confirming the need for language-specific adaptation.

4.3. Fine-tuning

Given the limited size of our datasets, training an ASR model from scratch is not feasible. Therefore, we fine-tune Whisper in the following two distinct configurations:

1. **Monolingual**: language-specific fine-tuning
2. **Multilingual**: multilingual fine-tuning

Model-specific decisions regarding language tokens and vocabulary were necessary for our experiments. Initial tests showed that vocabulary expansion with new language tokens doubles convergence time compared to using existing language tokens, leading us to adopt the latter approach. Experiments with English and language-family-specific tokens yield equivalent results, primarily because the tokenizer mapped individual characters to single tokens regardless of linguistic relationships.

4.4. Evaluation Metric

For evaluation, we use Word Error Rate (WER) [27] and Character Error Rate (CER) [28]. CER, preferred for cursive orthographies [29], is more sensitive to morphological segmentation errors, crucial for our languages (all except Zazaki use Arabic-based scripts). Additionally, it better captures orthographic accuracy with grapheme-level discrepancies between merged affixes, common in handwritten and casual writing, leading to inconsistent affix tokenization [30], compounded by training data irregularities from orthographic variations [31].

4.5. Hyper-parameters

The fine-tuning process is implemented using the Hugging Face Transformers library[3] with a learning rate of 5e-5 and 100 warmup steps, training the models for 5 epochs. The batch sizes are 192 and 128 for the `Base` and `Small` models, respectively, achieved through gradient accumulation. Mixed precision training (fp16) and gradient check-pointing are enabled for memory efficiency, with evaluation and checkpoint saving occurring at the end of every epoch. Training is conducted on a single A100 40 GB GPU.

[3] `https://github.com/huggingface/transformers`

Table 2: *ASR performance based on WER with CER in parenthesis–both in percentage; lower is better ↓. Fine-tuned monolingual models outperform the baseline and multilingual models in both* `Base` *and* `Small` *setups, with the later outperforming* `Base`*. Δ shows the difference of the best model (monolingual) vs. the baseline per language (larger negative values show greater improvement).*

Language	Whisper `Base` ↓				Whisper `Small` ↓			
	Baseline	**Multilingual**	**Monolingual**	**Δ**	**Baseline**	**Multilingual**	**Monolingual**	**Δ**
Gilaki	96.8 (92.1)	96.1 (38.2)	93.2 (37.5)	-3.6 (-54.6)	95.1 (89.8)	92.5 (35.7)	**89.5 (34.8)**	-5.6 (-55)
Hawrami	95.3 (77.8)	41.5 (8.7)	40.1 (8.6)	-55.2 (-69.2)	93.1 (85.2)	38.2 (7.9)	**37.9 (7.5)**	-55.2 (-77.7)
Laki Kurdish	93.7 (84.9)	60.4 (16.5)	58.5 (16.0)	-35.2 (-68.9)	91.2 (82.4)	56.1 (15.5)	**54.3 (15.0)**	-36.9 (-67.4)
Mazanderani	94.1 (88.5)	68.4 (25.1)	66.1 (24.5)	-28 (-64)	92.3 (86.9)	64.1 (23.5)	**62.1 (22.9)**	-30.2 (-64)
Southern Kurdish	90.6 (83.1)	56.1 (17.2)	54.4 (16.8)	-36.2 (-66.3)	88.9 (81.4)	52.1 (15.5)	**50.4 (14.9)**	-38.5 (-66.5)
Zazaki	91.4 (85.2)	76.2 (31.8)	74.0 (31.2)	-17.04 (-54)	89.7 (83.1)	71.2 (29.9)	**69.0 (29.1)**	-20.7 (-54)
Mean	93.5 (85.3)	66.4 (22.9)	64.4 (22.4)	-29.1 (-62.9)	91.7 (84.8)	62.2 (21.3)	**60.5 (20.5)**	-31.2 (-64.3)

5. Experiment Results

Table 2 presents ASR performance across six languages using Whisper models in baseline, multilingual, and monolingual configurations. Monolingual models consistently outperform others, achieving mean WER reductions of 29.1-31.2 percentage points with a clear hierarchy: baseline (88-96% WER) < multilingual (38-76% WER) < monolingual (37-74% WER). Hawrami achieves the best performance (37.9% WER) with the largest improvement (55.2 percentage points), while languages using an Arabic-based script, i.e. all except Zazaki, show more dramatic gains than Latin-script Zazaki. That said, Gilaki remains most challenging despite using an Arabic-based script, showing only 3.6-5.6 percentage point improvements and the highest WER (89.5-93.2%), indicating that factors beyond script complexity affect performance. While multilingual training offers deployment advantages, monolingual fine-tuning proves essential for optimal performance on under-resourced languages with complex writing systems.

To validate the reliability of our results, we perform statistical significance testing through bootstrap resampling with 1000 samples (except 200 for Zazaki). The improvements from baseline to fine-tuned models demonstrate statistical significance ($p < 0.01$) across all languages. The performance differences between monolingual and multilingual configurations achieve significance ($p < 0.05$) for languages using Arabic-based scripts, with Hawrami exhibiting the highest significance level ($p < 0.01$). Additionally, we observe a statistically significant performance gap ($p < 0.05$) between Whisper `Base` and `Small` models, where `Small` consistently outperforms `Base`. These findings underscore the importance of model capacity selection in processing low-resource languages and validate our architectural choices in both model selection and training configuration.

5.1. Effectiveness of Data Sources

We analyze the relative performance of Hawrami ASR using two distinct data sources: community-contributed Telegram conversations and studio-recorded along with TV transcriptions. The Whisper `Base` model achieves notably better performance with Telegram data (WER: 40.13%, CER: 8.56%) compared to the other sources (WER: 62.21%, CER: 26.08%). While studio recordings offer higher acoustic quality, several factors may contribute to this performance difference. Since our test set predominantly consists of Telegram-sourced conversations, the improved performance could be attributed to the similarity in acoustic conditions and speech patterns between training and testing data. Additionally, the informal nature of Telegram conversations may better represent natural speech patterns in the community. These findings align with [31] in demonstrating the viability of community-contributed data for endangered language documentation, though further investigation with balanced test sets would be needed to draw broader conclusions.

5.2. Error Analysis

Through manual inspection of 500 incorrect transcriptions, we identify several distinct categories of errors and their relationships to data characteristics. The most prominent challenge stems from the multiplicity of valid sound representations in Arabic script, leading to orthographic variation errors; for instance, Hawrami's <v>/<w> distinction as in وێش' (*wêş*) versus ڤێش' (*vêş*) cause multilingual model errors that monolingual training resolves. Affix merging errors show a significant disparity between languages using Arabic-based scripts and Latin-based writing systems, affecting 38% of Gilaki errors compared to only 9% in Zazaki. This pattern directly correlates with the lack of standardized orthographies. Environmental factors in community-contributed recordings manifest in truncated word endings, occurring in 19% of these cases versus 6% in studio recordings. Dialectal variations emerge as another significant factor, particularly evident in Southern Kurdish where 63% of such errors occur in the underrepresented Khanaqin variety. This systematic analysis reveals two critical areas for improvement: the need for consistent orthographic conventions and the importance of balanced dialectal representation in training data collection.

6. Conclusion

This paper presents pioneering work in developing speech recognition capabilities for six severely under-resourced Middle Eastern languages. Through a community-driven initiative, we collect over 40 hours of speech data using three complementary approaches. Our experiments with Whisper models demonstrate that monolingual fine-tuning consistently outperforms multilingual training across all languages. However, the moderate performance difference suggests multilingual models remain a practical choice for supporting multiple low-resource languages simultaneously. For Hawrami, our experiments with different data sources show better performance with Telegram-sourced conversations compared to studio recordings, though this may be influenced by acoustic similarity between training and testing conditions. While our work establishes important baselines for six severely under-resourced languages, the current error rates (ranging from 37.9% to 89.5% WER with Whisper `Small`) underscore the challenges in developing robust ASR systems. These challenges particularly manifest in handling dialectal variations and non-standardized orthographies. Our methodology and findings provide valuable insights for future efforts in speech technology for under-resourced languages.

7. Acknowledgments

This work was supported by the Swiss National Science Foundation (MUTAMUR project no. 213976) and the Stanford Initiative on Language Inclusion and Conservation in Old and New Media (SILICON). We extend our heartfelt gratitude to the volunteers who participated in the data collection campaign.

8. References

[1] S. Moradi, "Languages of Iran: Overview and critical assessment," *Handbook of the changing world language map*, 2020.

[2] T. Skutnabb-Kangas and D. Fernandes, "Kurds in Turkey and in (Iraqi) Kurdistan: A comparison of Kurdish educational language policy in two situations of occupation," *Genocide Studies and Prevention*, vol. 3, no. 1, pp. 43–73, 2008.

[3] C. Moseley and A. Nicolas, *Atlas of the World's Languages in Danger*, ser. Memory of Peoples. UNESCO, 2010, vol. 30.

[4] J. Rosenhouse, "Bilingualism/Multilingualism in the Middle East and North Africa: a focus on cross-national and diglossic bilingualism/multilingualism," *The handbook of bilingualism and multilingualism*, pp. 899–919, 2012.

[5] I. G. Or, *Language Policy and Education in the Middle East and North Africa*. Springer International Publishing, 2016, pp. 1–13.

[6] T. Reitmaier, D. K. Raju, O. Klejch, E. Wallington, N. Markl, J. Pearson, M. Jones, P. Bell, and S. Robinson, "Cultivating spoken language technologies for unwritten languages," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 2024, pp. 614:1–614:17.

[7] O. Scharenborg, L. Ondel, S. Palaskar, P. Arthur, F. Ciannella, M. Du, E. Larsen, D. Merkx, R. Riad, L. Wang, E. Dupoux, L. Besacier, A. W. Black, M. Hasegawa-Johnson, F. Metze, G. Neubig, S. Stüker, P. Godard, and M. Müller, "Speech technology for unwritten languages," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 964–975, 2020.

[8] H. Polat, A. K. Turan, C. Koçak, and H. B. Ulaş, "Implementation of a Whisper architecture-based Turkish automatic speech recognition (ASR) system and evaluation of the effect of fine-tuning with a low-rank adaptation (lora) adapter on its performance," *Electronics*, vol. 13, no. 21, p. 4227, 2024.

[9] 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, p. 8898, 2022.

[10] X. Huang, F. Alleva, M.-Y. Hwang, and R. Rosenfeld, "An overview of the sphinx-ii speech recognition system," in *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*, 1993.

[11] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. Proceedings of Machine Learning Research, 2023.

[13] S. C. Team, "Joint speech and text machine translation for up to 100 languages," *Nature*, vol. 637, pp. 587–593, 2025. [Online]. Available: https://doi.org/10.1038/s41586-024-08359-z

[14] A. A. Abdullah, S. Tabibian, H. Veisi, A. Mahmudi, and T. A. Rashid, "End-to-end transformer-based automatic speech recognition for Northern Kurdish: A pioneering approach," *CoRR*, vol. abs/2410.16330, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2410.16330

[15] S. Ahmadi, D. Q. Jaff, M. M. I. Alam, and A. Anastasopoulos, "Language and speech technology for Central Kurdish varieties," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*. ELRA and ICCL, 2024, pp. 10 034–10 045.

[16] Y. Marmor, K. Misgav, and Y. Lifshitz, "ivrit.ai: A comprehensive dataset of Hebrew speech for AI research and development," *CoRR*, vol. abs/2307.08720, 2023.

[17] M. A. Kermanshahi, A. Akbari, and B. Nasersharif, "Transfer learning for end-to-end ASR to deal with low-resource problem in Persian language," in *26th International Computer Conference, Computer Society of Iran*. IEEE, 2021, pp. 1–5.

[18] P. Cardinal, A. Ali, N. Dehak, Y. Zhang, T. A. Hanai, Y. Zhang, J. R. Glass, and S. Vogel, "Recent advances in ASR applied to an Arabic transcription system for Al-Jazeera," in *15th Annual Conference of the International Speech Communication Association*. ISCA, 2014, pp. 2088–2092.

[19] S. Coats, "Dialect corpora from Youtube," *Language and linguistics in a complex world*, 2023.

[20] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, 2020, pp. 4218–4222.

[21] N. J. de Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. de Wet, E. Barnard, and A. de Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Speech Commun.*, vol. 56, pp. 119–131, 2014.

[22] S. Mussakhojayeva, A. Janaliyeva, A. Mirzakhmetov, Y. Khassanov, and H. A. Varol, "KazakhTTS: An Open-Source Kazakh Text-to-Speech Synthesis Dataset," in *22nd Annual Conference of the International Speech Communication Association*. ISCA, 2021, pp. 2786–2790.

[23] M. Musaev, S. Mussakhojayeva, I. Khujayorov, Y. Khassanov, M. Ochilov, and H. Atakan Varol, "USC: An open-source Uzbek speech corpus and initial speech recognition experiments," in *Speech and Computer: 23rd International Conference, SPECOM*. Springer, 2021, pp. 437–447.

[24] H. Veisi, H. Hosseini, M. MohammadAmini, W. Fathy, and A. Mahmudi, "Jira: a Central Kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon," *Language Resources and Evaluation*, pp. 917–941, 2022.

[25] S. Ahmadi, R. Sennrich, E. Karami, A. Marani, P. Fekrazad, G. Akbarzadeh Baghban, H. Hadi, S. Heidari, M. Dogan, P. Asadi, D. Bashir, M. A. Ghodrati, K. Amini, Z. Ashourinezhad, M. Baladi, F. Ezzati, A. Ghasemifar, D. Hosseinpour, B. Abbaszadeh, A. Hassanpour, B. Jalal Hamaamin, S. Kamal Hama, A. Mousavi, S. Nazir Hussein, I. Nejadgholi, M. Ölmez, H. Osmanpour, R. Roshan Ramezani, A. Sediq Aziz, A. Salehi Sheikhalikelayeh, M. Yadegari, K. Yadegari, and S. Zamani Roodsari, "PARME: Parallel corpora for low-resourced Middle Eastern languages," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. ACL, 2025.

[26] S. Ahmadi, "KLPT–Kurdish language processing toolkit," in *Proceedings of second workshop for NLP open source software (NLP-OSS)*, 2020, pp. 72–84.

[27] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763–786, 2007.

[28] A. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. INTERSPEECH*, 2004.

[29] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, Z. Kons, R. Hoory, and M. Picheny, "Effects of word frequency and morphology on speech recognition accuracy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1093–1104, 2020.

[30] M. Shafiei, "Persian ASR challenges: A lexical analysis of training data inconsistencies," in *Proceedings of the 13th Language Resources and Evaluation Conference*, 2022, pp. 1558–1565.

[31] L. Haddock, J. Barker, and H. Christensen, "A systematic study of noise in crowd-sourced data for ASR," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6099–6103.