Exploring the Reusability of Northern Kurdish Resources for Badini Speech Recognition

Mohammad Mohammadamini[1], **Aveen Jalal Mohammed**[2]
Barzan Hussein Mohammed[3], **Dezheen H. Abdulazeez**[2]
Imad Saeed Sadeeq[3], **Dilgash Mohammed Salih Tayib**[3]
Amera Ismail Melhum[2], **Abuobaida Abdullah Dheyab**[3]
[1] LIUM, Le Mans Université, France
[2] Department of Computer, University of Duhok, Kurdistan, Iraq
[3] Department of Kurdish, College of Languages, University of Duhok, Kurdistan, Iraq
first.last@univ-lemans.fr, first.last@uod.ac

Abstract

Badini is a variant of the Kurdish language spoken in the Duhok province of the Kurdistan Region of Iraq. It is written mainly in a modified version of the Arabic script. Although it shares the same script as Central Kurdish (CKB), it is linguistically classified under the Northern Kurdish (KMR) branch. In this paper, we explore the potential and limitations of Northern Kurdish ASR resources for the Badini variant. Firstly, we transliterate the Common Voice 18 dataset from the Latin script into the modified Arabic script and revise it to align with the orthographic conventions of the Badini variant. Additionally, we introduce the first text collection for the Badini variant, containing 14.22 million tokens, which serves as a source for speech synthesis. The third resource developed in this research is a standard speech recognition benchmark recorded by 5 speakers which includes 2 hours and 46 minutes of multi-domain read speech. Results show that combining transliterated and synthetic data significantly improves recognition accuracy, achieving a 6.8% CER and 34% WER. All three resources curated during this research will be made available under the CC BY-NC-ND 4.0 license.

Keywords: Northern Kurdish, Badini Variant, Speech Recognition, Common Voice, Low-resource ASR

1. Introduction

Kurdish population with more than 35 million people speak a multi-dialect continuum categorized into several groups: Central Kurdish (CKB), Northern Kurdish (KMR), Southern Kurdish (SDH) (Matras, 2019), and Hewrami-Zazaki (Todd, 1985). Despite its wide usage, the Kurdish language is considered a low-resource language (Veisi et al., 2020). Its use across four countries, the coexistence of both Arabic and Latin writing systems, and the wide range of dialects have made the development of an inclusive Kurdish automatic speech recognition (ASR) system challenging.

In recent years, there have been significant efforts to develop speech recognition resources for Kurdish dialects. Among the most prominent of these resources, we can point to Common Voice (Ardila et al., 2020), which in its 22nd version includes 136 hours of Central Kurdish, 71 hours of Northern Kurdish, and 2 hours of validated Zazaki speech. [1] Other notable efforts include the Asosoft Speech Corpus (Veisi et al., 2022), which in its first version introduced 46 hours of designed and phonetically balanced audio for Central Kurdish and has been used in several studies with remarkable results (Veisi et al., 2022; Mohammadamini et al., 2025). In other significant endeavors to provide speech recognition resources, (Hameed et al., 2025) has provided ASR resource for Hawrami and Southern Kurdish. FLEURS is another resource that includes Central Kurdish for speech recognition and translation (Conneau et al., 2023).

The scope of the current research is the Badini variant of Northern Kurdish, for which, to the best of our knowledge, there are no specific studies or ASR resources to date. Northern Kurdish is mainly written in the Latin script in Turkey and Syria, while the Badini variant from this group is primarily written in the Arabic script in the Kurdistan Region of Iraq.

In the current research, we explore the reusability of ASR resources of Northern Kurdish in the Badini variant. Since Badini is written in the same script as Central Kurdish, we firstly show the limitations of Central Kurdish ASR for Badini. In fact, Badini is a subgroup of Northern Kurdish but the codified variant of Northern Kurdish is mainly written in the Latin script. In order to show the potential of reusing Northern Kurdish resources for this variant, we transliterated Common Voice of Northern Kurdish into Arabic script and revised it manually to adapt it with the orthographic conventions of Kurdish in the Arabic script (Team, 2025). After demonstrating the potential and limitations of speech corpora from these dialects for the Badini dialect, we propose a text-only data augmentation approach using Badini text and a Northern Kurdish TTS system (Pratap et al., 2024). The synthetic speech is generated from a text corpus which is collected during the current research. The curated Badini text collection includes 14.22 million space

[1] https://commonvoice.mozilla.org/en/datasets

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 110–115
16 May 2026.

separated tokens mainly collected from websites. The third effort in this study is developing a multi-domain ASR benchmark for Badini dialect which is used to evaluate the developed systems. The developed benchmark includes 2 hours and 46 minutes of read speech validated manually.

In the following, in Section 2 we give a brief description of Badini dialect, in Section 3 the proposed methodology of reusing the available resources is presented, in Section 4 the curated resources are described and Section 5 presents the obtained results.

2. Kurdish language and Badini variant

The Kurdish language is an Indo-European language spoken by an estimated 30 to 46 million native speakers across the Kurdistan regions of Iran, Iraq, Turkey, and Syria (Sheyholislami, 2015). In Turkey and Syria, the Kurdish dialects are mainly written in the Latin script, while in Iran and Iraqi Kurdistan, almost all dialects use a modified version of the Arabic script. The scope of our study is Badini, which is a variant of Northern Kurdish. While the majority of Northern Kurdish speakers live in Turkey and Syria and use the Latin script, a significant number of Northern Kurdish speakers live in Iraqi Kurdistan. The spoken Northern Kurdish sub-dialect in these regions is called *Badini*. More precise linguistic classification studies place Badini in the *Southeastern* subgroup of Northern Kurdish, which is spoken in the Hakkari province in Turkey and the Duhok province of Iraqi Kurdistan (Öpengin and Haig, 2014) (Haig, 2018) (Figure. 1).

Both the Latin and Arabic scripts used for Kurdish dialects are almost phonemic (Veisi et al., 2020); there is almost one-to-one correspondence between letters and phonemes. However, this rule is relative. In Table 2, the list of letters in the Badini variant and the Latin script used for Northern Kurdish are presented. As shown in the table, there are six letters (ح، ع، خ، غ، ڵ، ێ) in the Arabic script that are not represented in the Latin script. All of these six letters correspond to phonemes in the Badini variant, but not all of them occur in every subgroup of Northern Kurdish (Öpengin and Haig, 2014), and therefore they are not consistently represented in the script. The /ə/ phoneme is the only phoneme represented in the Latin script that does not have a corresponding letter in the Arabic script. We will return back to this mismatch between the phonemes and its impact on the performance of ASR systems developed for standard variant of Northern Kurdish on Badini variant (see Section 5).

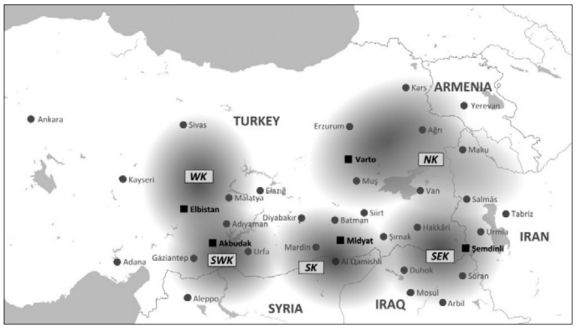

Figure 1: The geography of Badini shown by SEK label as a part of the Southeastern subgroup of Northern Kurdish (Öpengin and Haig, 2014).

Badini Letter			KMR	IPA
Init	Med	Fin		
بـ	ـبـ	ـب	B b	b
د	ـد	ـد	D d	d
جـ	ـجـ	ـج	C c	ʤ
گـ	ـگـ	ـگ	G g	g
ڤـ	ـڤـ	ـڤ	V v	v
ز	ـز	ـز	Z z	z
ژ	ـژ	ـژ	J j	ʒ
غـ	ـغـ	ـغ	–	ɣ
عـ	ـعـ	ـع	–	ʕ
تـ	ـتـ	ـت	T t	t
چـ	ـچـ	ـچ	Ç ç	ʧ
کـ	ـکـ	ـک	K k	k
پـ	ـپـ	ـپ	P p	p
قـ	ـقـ	ـق	–	q
ئـ	ـئـ	ـئ	–	ʔ
هـ	ـهـ	ـه	H h	h
سـ	ـسـ	ـس	S s	s
شـ	ـشـ	ـش	Ş ş	ʃ
فـ	ـفـ	ـف	F f	f
خـ	ـخـ	ـخ	X x	x
حـ	ـحـ	ـح	–	ħ
ر	ـر	ـر	R r	r
ڕ	ـڕ	ـڕ	RR rr	r
لـ	ـلـ	ـل	L l	l
ڵـ	ـڵـ	ـڵ	–	ɬ
مـ	ـمـ	ـم	M m	m
نـ	ـنـ	ـن	N n	n
یـ	ـیـ	ـی	Y y	j
و	ـو	ـو	W w	w
ی	ـی	ـی	Î î	iː
ا	ـا	ـا	A a	aː
ێـ	ـێـ	ـێ	Ê ê	ɛ
ۆ	ـۆ	ـۆ	O o	o
و	ـو	ـو	U u	u
ووـ	ـووـ	ـوو	Û û	uː
ه	ـه	ـه	E e	æ
-	-	-	I i	ə

Table 1: List of Badini alphabet with initial, medial, and final Arabic forms, KMR equivalents, and IPA.

3. Methodology

For many multi-dialect languages, ASR systems are developed for one or a few standardized dialects, which tend to perform poorly on local variants (Blaschke et al., 2025). In our proposed approach, we explore the reusability of Northern Kurdish resources for Badini, which is a subgroup of this language. Our approach consists of two main parts. First, we use the Common Voice corpus of Northern Kurdish. The scripts are transliterated

using the AsoSoft library[2] into the Arabic script (Mahmudi and Veisi, 2021). In the Arabic script, there is a general tendency to write words together, whereas in the Latin script, words are usually written separately. This simple transliteration therefore causes a significant mismatch between the two scripts. After transliteration, the Arabic script is revised manually to adapt it to the standards commonly accepted in the Arabic orthography. For example, in the Arabic script, the verb "to be" is usually attached to the preceding word, and the second part of circumpositions is often joined with the previous word. During the revision, phonemes and words were not modified; only the rules regarding word joining and separation were adjusted.

In order to compensate for the dialectical differences between standard variation of Northern Kurdish and Badini variant, we use a text-only data augmentation by applying a TTS model available as a part of MMS (Pratap et al., 2024). The flowchart of the proposed approach is shown in Figure 2.

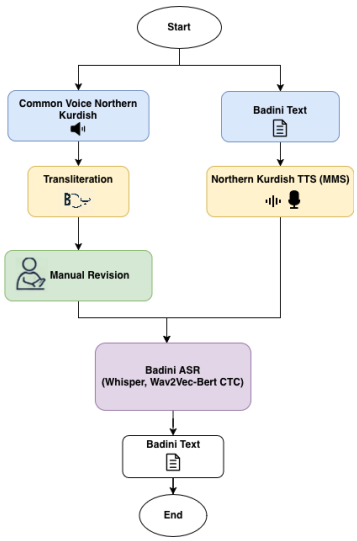

Figure 2: Reusbality of KMR tools and resources for Badini variant

Having the transliterated and synthesised speech we are using two state-of-the art models for Badini speech recognition. The first model is Whisper V3 Large model (Radford et al., 2022) fine-tuned for Badini and the second model is Whav2vec2.0-BERT model optimizes a CTC loss function (Chung et al., 2021). The Whisper model was fine-tuned for five epochs with a learning rate initialized at 1e5.

[2]https://github.com/AsoSoft/AsoSoft-Library

4. Dataset Description

4.1. Transliterated Common Voice

The first dataset used for training the models is the Common Voice 18 transliterated into Arabic script of Kurdish language [3]. This dataset includes 16,808 unique sentences, having several recordings for each sentence the total duration of the validated recorded audio in the transliterated common voice results in 68 hours of speech (Table 3).

4.2. Raw text corpora

During the current study we collected the first text collection for Badini variant from different resources such as news agencies, university websites etc. Using crawling and scraping we collected 70k documents from 15 websites which is equivalent to 14.22 million tokens[4]. The details of the collected text is shown in Table 2

Website Name	Tokens (M)	Documents
Gav	4.56	30964
Speda	0.99	5242
Bas news	1.01	6731
University of Duhok	0.36	2958
University of Zakho	0.06	478
University of Cihan	1.20	9389
Badinan	0.85	1099
Matin journal	1.28	608
Xani agency	0.41	3450
Badinpedia	1.31	1744
kurdax	0,51	1336
kurdislamic	0,54	2421
mzgaft	0,21	343
nrtTv	0,1	542
zebadinan	0,83	2642
Total	14.22	70k

Table 2: Badini Text collection dataset

4.3. Synthetic speech

From the text collection (Table 2), around 119k sentences containing between 3 and 20 tokens were extracted. The extracted sentences were synthesized using the MMS model for Northern Kurdish (Pratap et al., 2024). Among the preprocessing steps, we applied Unicode and number unification (Mahmudi et al., 2019), as well as a specific date and number-to-word conversion for Badini, implemented specifically for this purpose. The synthesized speech from this dataset is 207 hours (Table 3).

4.4. Evaluation Benchmark

The third resource developed during this study is an evaluation benchmark designed for Badini ASR.

[3]https://huggingface.co/datasets/BadiniSpeechNLP/commonvoicebadini

[4]https://huggingface.co/datasets/BadiniSpeechNLP/badinitextcorpora

Model	Utterances	Duration
Common Voice KMR	61.60k	68h
Synthetic data	119k	207h

Table 3: Train speech datasets

The benchmark includes sentences from eight topics. We aimed to create a relatively realistic evaluation set that reflects the main challenges faced by ASR systems, such as code-switching. The scientific category is mainly adapted from textbooks and contains many English scientific terms, while the religion category primarily includes Arabic loanwords. In the evaluation benchmark, we have 623 unique sentences recorded by 5 speakers (2 male and 3 female). The total number of recordings is 1,028 files, resulting in 2 hours and 46 minutes of speech [5]. All recordings in this benchmark were manually validated (Table 4).

Topic	#Samples	Duration	Tokens
art	112	21.48	2173
economy	150	27.70	2944
health	87	14.20	1502
humanities	94	9.86	863
literature	131	10.40	891
politics	48	9.19	998
religion	181	25.42	2945
science	225	48.35	5277
overall	1028	2h46mins	17593

Table 4: Badini ASR evaluation benchmark

5. Experiments and Results

As the first experiment, we evaluated a Whisper-large model fine-tuned on Central Kurdish Common Voice 18 (CV-CKB), achieving 28.22 CER and 76.65 WER, which reflects the high dialectical difference between Central Kurdish and Badini (Table 5). In the second experiment, we fine-tuned both Whisper and Wav2Vec-BERT using the transliterated Northern Kurdish Common Voice (CV-KMR) dataset, where we observed an 11.33 CER with Wav2Vec-BERT. These results show that, although KMR data can lead to significantly better performance than CKB resources due to linguistic similarity, we are still relatively far from a practical ASR model. In the third experiment, we used synthetic Badini speech, obtaining a 9.25 CER. Our results indicate that data generated by TTS models can generalize better than real transliterated Northern Kurdish Common Voice data. In the final experiment, we combined both transliterated Common Voice and synthetic data, achieving a 6.80 CER and 34 WER. Leveraging the potential of both approaches, we achieved a 39% CER relative improvement compared to using only real Northern Kurdish speech.

[5] https://huggingface.co/datasets/BadiniSpeechNLP/badini-asr-benchmark

One reason behind this improvement of using the Badini synthetic data is the compensation of existing lexical difference between standard Northern Kurdish and Badini variant.

Training	Whisper		Wav2Vec-BERT	
	CER	WER	CER	WER
CV-CKB	28.22	76.65	-	-
CV-KMR	24.61	64.30	11.33	50.01
TTS-Badini	18.85	58.46	9.25	39.11
TTS-Bad-CV-KMR	9.00	39.38	6.80	34.41

Table 5: CER and WER on the Badini ASR benchmark for models trained on Central Kurdish, Northern Kurdish, and synthesized Badini data

Based on the results obtained by Wav2Vec-BERT, we present the detailed results for each topic in Table 6. The best performance was achieved in the politics domain, which may be attributed to the fact that a large portion of the synthesized speech originates from this domain. In contrast, the scientific domain showed the poorest performance, likely due to the presence of more challenging sentences containing a higher proportion of non-Kurdish words.

Topic	Whisper		Wav2Vec-BERT	
	CER	WER	CER	WER
art-culture	9.13	42.29	6.90	35.85
economy	8.02	32.03	5.59	30.29
health	9.13	38.56	7.72	35.11
humanities	9.13	42.06	7.81	37.43
literature	9.02	37.60	6.80	32.21
politics	6.40	31.99	5.29	30.89
religion	9.43	38.27	7.07	32.97
science	9.80	44.38	7.21	37.25
overall	9.00	39.38	6.80	34.41

Table 6: CER and WER per domain for Whisper and Wav2Vec-BERT trained on transliterated Common Voice and synthetic speech

One reason behind the low performance of using standard Northern Kurdish ASR data for Badini stems from the mismatch between the two alphabet (shown in Table 2) which cause a significant information loss during transliteration. We analyzed the percentage of these letters in the reference transcripts and predictions produced by the best model (Table 7). In all cases, the percentage of these letters in the predictions was two to three times lower than in the references. This indicates that a main limitation of our proposed approach stems from this issue. In order to have deeper explanation on this type of errors, we extracted the percentage of each letter from the evaluation benchmark. We observed that the frequency of four letters without having a corresponding letter in the Arabic script is very low. Since, the overall occurrence of these letters in the test set is less than one percent, the lexical and morphosyntactic differences (Öpengin and Haig, 2014) between Badini variant and standard Northern Kurdish are the second reason behind

the limited potential of reusing Northern Kurdish resources for Badini variant.

Characters	Reference	Prediction
ح	0.21%	0.07%
ع	0.13%	0.06%
غ	0.09%	0.03%
ڵ	0.16%	0.04%

Table 7: Information loss due to character mismatch between Latin and Arabic script

6. Conclusion

In this paper, we explored the reusability of standard Northern Kurdish ASR resources for the Badini variant. During this study, we prepared three resources: a transliterated and revised version of the Northern Kurdish Common Voice dataset in Arabic script, the first text collection for the Badini variant containing 14.22 million tokens, and a multi-domain ASR benchmark. In the proposed pipeline, which includes transliteration and data augmentation, we demonstrated how existing resources can be reused to improve ASR performance on local variants. However, our results also highlight the limitations of such resources. According to this study, it is necessary to develop dedicated resources for Badini dialect ASR, which will be the focus of our future work.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Mike Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4211–4215.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A Multi-Dialectal Dataset for German Dialect ASR and Dialect-to-Standard Speech Translation. In *Interspeech 2025*, pages 913–917.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Geoffrey Haig. 2018. 2.3. northern kurdish (kurmanjî). *The languages and linguistics of Western Asia: An areal perspective*, 6:106.

Razhan Hameed, Sina Ahmadi, Hanah Hadi, and Rico Sennrich. 2025. Automatic Speech Recognition for Low-Resourced Middle Eastern Languages. In *Interspeech 2025*, pages 733–737.

A Mahmudi, H Veisi, M MohammadAmini, and H Hosseini. 2019. Automated kurdish text normalization. In *The second international conference on kurdish and persian languages and literature*.

Aso Mahmudi and Hadi Veisi. 2021. Automated grapheme-to-phoneme conversion for central kurdish based on optimality theory. *Computer Speech Language*, 70:101222.

Yaron Matras. 2019. Revisiting Kurdish dialect geography: Findings from the Manchester database. In *Current Issues in Kurdish Linguistics*, pages 225–241. University of Bamberg Press, Bamberg.

Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, and Antoine Laurent. 2025. Scaling pseudo-labeling data for end-to-end low-resource speech translation (the case of Kurdish language). In *Interspeech 2025*, pages 898–902.

Ergin Öpengin and Geoffrey Haig. 2014. Regional variation in kurmanji: A preliminary classification of dialects. In *Kurdish Studies Archive*, pages 171–212. Brill.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25:1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Whisper: Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Jaffer Sheyholislami. 2015. *The Kurds: History, Religion, Language, Politics*, chapter Language Varieties of the Kurds. Austrian Federal Ministry of the Interior.

Badini Orthography Team. 2025. Renusi Badini (In Kurdish). University of Duhok.

Terry Lynn Todd. 1985. *A Grammar of Dimili, also known as Zaza*. Ph.d. dissertation, University of Michigan, Ann Arbor, MI. Near Eastern Studies: Languages and Literatures.

H. Veisi, H. Hosseini, M. MohammadAmini, et al. 2022. Jira: a central kurdish speech recognition system, designing and building speech corpus and pronunciation lexicon. *Language Resources and Evaluation*, 56:917–941.

Hadi Veisi, Mohammad MohammadAmini, and Hawre Hosseini. 2020. Toward kurdish language processing: Experiments in collecting and processing the asosoft text corpus. *Digital Scholarship in the Humanities*, 35(1):176–193.