Speech Recognition Corpus of the Khinalug Language for Documenting Endangered Languages

Zhaolin Li[1], **Monika Rind-Pawlowski**[2], **Jan Niehues**[1]
[1]Karlsruhe Institute of Technology, Germany
[2]Goethe University Frankfurt, Germany
{zhaolin.li, jan.niehues}@kit.edu, monikapawlowski@live.de

Abstract

Automatic Speech Recognition (ASR) can be a valuable tool to document endangered languages. However, building ASR tools for these languages poses several difficult research challenges, notably data scarcity. In this paper, we show the whole process of creating a useful ASR tool for language documentation scenarios. We publish the first speech corpus for Khinalug, an endangered language spoken in Northern Azerbaijan. The corpus consists of 2.67 hours of labeled data from recordings of spontaneous speech about various topics. As Khinalug is an extremely low-resource language, we investigate the benefits of multilingual models for self-supervised learning and supervised learning and achieve the performance of 6.65 Character Error Rate (CER) points and 25.53 Word Error Rate (WER) points. The benefits of multilingual models are further validated through experimentation with three additional under-resourced languages. Lastly, this work conducts quality assessments with linguists on new recordings to investigate the model's usefulness in language documentation. We observe an evident degradation for new recordings, indicating the importance of enhancing model robustness. In addition, we find the inaudible content is the main cause of wrong ASR predictions, suggesting relating work on incorporating contextual information.

Keywords: speech corpus, automatic speech recognition, endangered language

1. Introduction

There are approximately 7,000 languages spoken today, but a large portion of them is endangered and being spoken by less than 1,000 people. Linguists are actively engaged in documenting these languages, but manual documentation is both time-consuming and expensive, and it can be difficult to maintain accuracy and consistency.

Automatic Speech Recognition (ASR) has been a popular approach to facilitate endangered language documentation by transcribing the recordings (Godard et al., 2018; San et al., 2023, 2022; Liu et al., 2022; Rodríguez and Cox, 2023). Given the increased research on documenting endangered languages with ASR, there is a growing demand for speech corpora.

One of the main challenges in building ASR models for language documentation is data scarcity because numerous labeled data are needed for training, which is not practical for acquiring endangered languages. Current research shows Multilingual Representation Learning (MRL), coupled with self-supervised learning, significantly boosts model performance, particularly in low-resource scenarios that endangered languages have (Conneau et al., 2020; Guillaume et al., 2022; Sikasote and Anastasopoulos, 2021). This approach leverages the shared knowledge learned from both the labeled and unlabeled data of other languages to learn building ASR models for the target language. The multilingual models are currently pre-trained with at most 1,406 languages (Pratap et al., 2023). However, the multilingual model trains with linguistically similar and dissimilar languages to the target language. To our best knowledge, further research is needed to examine the impact of increasing the number of languages on ASR performance while neglecting language similarity (Conneau et al., 2020).

In this work, we present the creation and application of an ASR system for language documentation on the example of Khinalug, a Nakh-Dagestanian language spoken by around 2,300 people in Azerbaijan. We demonstrate how current achievements in ASR can benefit linguists engaged in language documentation. As a first step, we develop a speech corpus for Khinalug and publish it for further research[1]. Being the first speech corpus of this language family, the corpus consists of recordings of spontaneous speech collected by the linguists during fieldwork. Due to the high cost of manual annotation, the corpus has a total of 1,230 labeled samples with an average duration of 7.80 seconds.

The main challenge when creating the ASR system is the data limitation. To address this, we investigate the effectiveness of MRL in self-supervised learning with models pre-trained with different numbers of languages. Besides, we explore the strength of MRL in supervised learning

[1]https://huggingface.co/datasets/AI4LT/Khinalug_ASR

LREC-COLING 2024, pages 15171–15180
20-25 May, 2024.

by training together with languages from the same language family. The motivation is feeding more labeled data of speech recognition task might enhance performance.

Developing ASR models aims to assist language documentation, and this work conducts a quality assessment with linguists on untranscribed recordings to evaluate the usefulness of the ASR models. To imitate language documentation scenarios, the recordings are from the speakers already in the corpus and speakers not in the corpus.

We summarize the main findings and contributions as follows:

- **Endangered Language Speech Corpus**: We introduce the Khinalug speech corpus with the aligned speech and transcript. Khinalug is a language facing endangerment. The collection of this corpus follows language documentation scenarios, making it a realistic resource for the preservation of endangered languages.
- **Effectiveness of Multilingual Representation Learning**: We find multilingual representation in self-supervised learning benefits ASR performance, and increasing the number of languages, mainly dissimilar languages, in pre-training brings no clear impact. However, we find multilingual representation in supervised learning harms ASR performance, although having access to more labeled data of languages within the same language family to learn to perform ASR tasks.
- **Speech Recognition Analysis**: We perform quality assessments with linguists to explore the usefulness of ASR models. We find that inaudible content is the primary error resource of the current ASR model, and we observe the ASR model has performance degradation on new recordings and speakers.

2. Khinalug Corpus

2.1. Khinalug Language

Khinalug (ISO code kjj) is one language spoken by approximately 2,300 people in the Khinalug village in Northern Azerbaijan (Rind-Pawlowski, 2023a,b). All villagers are at least bilingual in Khinalug and Azerbaijani, and the older generation speaks Russian as well. In addition, Khinalug is spoken by a diaspora of at least 10,000 people in Azerbaijan and Russia, with a decreasing level of fluency.

Khinalug belongs to the Nakh-Dagestanian (also known as East Caucasian) family. As a severely endangered language recognized by (UNESCO, 2023) and (Ethnologue, 2024), documentation work has been done to investigate language documentation since 2011. With contributions from linguists, the annotated corpus of natural speech has been developed and consistently extended and revised.

The identification of the phoneme inventory and the best suitable transcription orthography for Khinalug are still in development. As for now, Khinalug has 9 phonemic vowels and 40 phonemic consonants (Rind-Pawlowski, 2023b). Since the distinction between phonemes and allophones is still a research question for Khinalug, the transcription aims at distinguishing allophones by different graphemes, wherever these are clear enough to identify and wherever there are enough symbols available. Therefore, in the current transcription, one grapheme represents one specific sound, either a phoneme (if the phoneme has no identifiable allophones) or single allophones (if a phoneme has identifiable allophones). In total, the orthography of Khinalug has 49 graphemes.

	#Sample	#Hour	A.audio	A.text
Train	1107	2.41	7.83	61.24
Test	123	0.26	7.50	59

Table 1: Dataset statistic of the Khinalug corpus. *A.audio* indicates the average duration of samples in second; *A.text* indicates the average transcript length.

2.2. Corpus Creation

The recordings of this corpus are spontaneous speeches of native speakers about various topics. With one consultant who repeated every word slowly, linguists wrote down the transcription. After that, the linguists proofread the transcription for correction and morphological glossing as the writing system of Khinalug is in development.

As shown in Table 1, this corpus consists of 2.57 hours of labeled data. In data collection, each recording is a long audio for one talk. Therefore, each audio consists of multiple sentences. As this corpus is developed for speech recognition, the recording is too long to train a model. Thus, the recordings are segmented into shortcuts with the timing information that linguists manually add.

After segmentation, the shortcuts are shuffled and partitioned into a 90-10 ratio for the training and test sets. Consequently, the corpus has 2.32 hours of training data and 0.25 hours of testing data.

2.3. Challenges

This corpus consists of spontaneous speech recordings, inherently presenting the following challenges for building speech recognition models.

Unintelligible Content

During proofreading, there are speaking parts that are not understandable because of background noise, unclear pronunciation, and various other reasons. To address this challenge without involving misleading information, a placeholder $ is added in the transcript to indicate the non-understandable content. For a better understanding of the challenge, the sentence with unintelligible content is less than 3% as shown in Appendix A).

The unintelligible content is a natural challenge for building a speech recognition model. This placeholder would highlight the task for the speech recognition model to recognize unintelligible content.

Speaking Disfluency

All recordings of this corpus are long audios of free monologues. The speaker behaves naturally in the conversation and, therefore, has disfluency in the speech, such as stammering, hesitation, and thinking. To keep originality, this work neglects the speaking disfluency in building the corpus, leaving room for research to address and handle this challenge.

3. Automatic Speech Recognition

After collecting the speech corpus, we aim to build an ASR system for Khinalug. Since the amount of data is insufficient to train a successful ASR model from scratch, we leverage MRL to enhance ASR performance by training with corpora of other languages.

In the first step, we explain the ASR system with the acoustic and language models (§ 3.1). Then, we explore MRL in self-supervised learning to study the effectiveness of MRL and the impact of dissimilar languages on the target language in pre-training (§ 3.2). Finally, we investigate the effectiveness of MRL in supervised learning (§ 3.3).

3.1. System Overview

Given the limited amount of data, even unlabeled data in the target language, self-supervised learning is a highly promising approach. This approach enables us to leverage large amounts of unlabelled data in other languages, reducing the data requirement for the target language.

The ASR system has two steps to predict transcription from speech. In the first step, self-supervised trained models are employed to generate a sequence of latent acoustic representations. In the second step, a classification head is incorporated into the model to map the representations to the vocabulary elements. During inference, there is an option to directly predict the transcription or incorporate an additional language model.

The self-supervised model is pre-trained with the speech-only data and then fine-tuned with the labeled data. Instead of pre-training from scratch, this work makes efficient and effective use of the existing strong Wav2vec2.0 (Baevski et al., 2020) pre-trained models. This process involves initializing the acoustic model with the parameters from the pre-trained model, adding the classification head, and fine-tuning the model with the labeled data using the Connectionist Temporal Classification (CTC) loss function.

The acoustic models learned with CTC loss focus on prediction at the character level. Integrating a language model is promising to bring contextual information at the word and sentence levels to the ASR system and improve model performance. In inference, we exploring combine the predictions from the acoustic model and a 5-gram language model that is developed with the training data. The combination is implemented with the package pyctcdecode [2].

3.2. Multilingual Self-supervised Learning

The first research question was to investigate the effect of the multilingual representation learned in self-supervised training. This question arises from the possibility that linguistically diverse languages, which is the majority of pre-training languages, could have an adverse effect on the speech recognition performance of the target language(Conneau et al., 2020), and solely increasing the number of languages in the pre-training may not yield beneficial results.

In our experiments, we conducted a comparative analysis involving a monolingual model that is pre-trained with English and various multilingual models that are pre-trained with different numbers of languages.

3.3. Multilingual Supervised Learning

As a second research question, we investigated whether multilingual data is also helpful in the supervised phase of the training. To explore this, we extended the model's training beyond just the Khinalug training corpus to include a language closely related to it. We select a similar language in supervised learning to avoid the negative impact of

[2] https://github.com/kensho-technologies/pyctcdecode/tree/main

dissimilar languages. The choice of similar languages is Azerbaijani according to the sociolinguistic and diglossic situations of Khinalug (Clifton et al., 2005; GARIBOVA and ZEYNALOV, 2023) and the corpus availability.

4. Experimental Setups

This section provides the experimental setups for building ASR systems. The training details are shown in 4.1 for the sake of reproducibility. After that, we introduce one endangered and two low-resource languages for experiments 4.2 to validate our findings about MRL. In the final part 4.3, we explain the evaluation metrics used to assess the performance of the ASR system.

4.1. Training Details

Following the pre-processing approach of wav2vec2.0 (Baevski et al., 2020; Pratap et al., 2023), we lowercase all characters and remove the punctuation. The vocabulary of the ASR model consists of all distinct characters in the corpus. Pre-processing reduces the size of vocabulary and mitigates the complications inherent in model developments.

During training, the feature extractor of the wav2vec 2.0 model is frozen as it is well-trained in pre-training (Baevski et al., 2020; Conneau et al., 2020; Babu et al., 2021). SpecAugment (Park et al., 2019) is implemented as data augmentation. The optimizer is AdamW with 0.9 beta1 and 0.9999 beta2. With 500 warm-up steps, the learning rate starts at 1e-4 and has a linearly decreased schedule. The weight decay ratio is 0.005.

Previous research indicates increasing the capacity of the pre-trained model benefits performance for low-resource scenarios, even with the potential need for more data for training the extra parameters (Babu et al., 2021; Pratap et al., 2023). Therefore, considering the availability of pre-trained models, this work experiments with two model configurations that have the same model architecture and transformer setups but different numbers of transformer blocks. The small model contains 24 transformer blocks, and the large model contains 48 transformer blocks. Each transformer block has a model dimension of 1,024, an inner dimension of 4,096, and 16 attention heads.

4.2. Approach Effectiveness across Languages

In addition to building the model and exploring the effectiveness of MRL on Khinalug, we validate our findings on one other endangered language Mboshi, and two low-resource languages Dhivehi and Danish.

Language	Split	#Sample	#Hour
Mboshi	Train	4616	3.38
	Test	514	0.37
Dhivehi	Train	2677	3.83
	Validation	2227	3
	Test	2212	3.04
Danish	Train	2746	2.92
	Validation	2222	2.66
	Test	2160	2.57

Table 2: Dataset statistics for other low-resource languages to explore the effectiveness of multilingual representation learning. *#Sample* indicates the number of samples and *#Hour* indicates the number of hours.

Table 2 shows the statistics of three testing languages. Mboshi is a typical Bantu language spoken in Congo-Brazzaville by around 130,000 people. Like Khinalug, Mboshi is an endangered language with a short documentation history; Dhivehi is an Indo-Aryan language mainly spoken in the South Asian island country of Maldives. As the official language of Maldives, the language is spoken by around 340,000 people; Danish is a North-Germanic language spoken by around 6 million people, mainly in and around Denmark. The Mboshi corpus from (Godard et al., 2018), and the Dhivehi and Danish datasets are from Common Voice version 13.0.

Khinalug and these other languages exhibit clear geographical distinctions and pose challenges in terms of documentation and language development history. Therefore, we regard these three languages as good test cases for exploring approach effectiveness.

4.3. Evaluation Metrics

This work evaluates speech recognition performance with Word Error Rate (WER) and Character Error Rate (CER). WER and CER measure the percentage of inaccuracies in predicted words and characters when compared to the reference. A lower score indicates a better model performance.

5. Results analysis

5.1. Multilingual Self-supervised Learning

Effectiveness of Multilingual Representation Learning

As shown in Table 3, we experiment with monolingual and multilingual pre-trained models with

	Khinalug	Mboshi	Dhivehi	Danish	Average
Mono-small	9.88/41.11	7.81/28.83	100/100	100/100	54.42/67.49
+ LM	8.64/34.56	7.47/26.06	96.78/99.33	96.25/98.63	52.29/64.65
Multi-53-small	**7.58/34.65**	6.70/25.10	10.51/55.94	11.88/38.98	9.17/38.67
+ LM	8.82/37.6	6.46/23.05	**10.34/56.3**	11.94/39.58	9.39/39.13
Multi-128-small	7.96/34.19	6.63/24.82	**10.45/55.52**	**10.27/33.82**	**8.83/37.09**
+ LM	7.43/33.26	6.51/23.96	10.55/59	**10.52/35.48**	**8.75/37.93**
Multi-1406-small	7.70/33.55	**6.27/24.12**	11.42/58.07	11.98/39.24	9.34/38.75
+ LM	**7.4/32.07**	**6.09/22.72**	11.19/59.02	11.55/35.84	10/41.62
Multi-128-large	7.92/35.30	7.13/26.09	12.25/59.65	13.08/41.15	10.10/40.55
+ LM	7.68/33.64	6.96/24.92	13.84/75.88	15.31/52.67	10.25/43.36
Multi-1406-large	7.63/32.07	6.57/24.12	11.64/57.94	12.69/41.34	9.63/38.87
+ LM	7.76/32.35	6.77/24.19	11.89/62.16	12.51/38.99	9.49/38.64

Table 3: Experiments about self-supervised learning with different pre-trained models; The models pre-trained with 1, 53, 128, and 1,406 languages are from (Baevski et al., 2020), (Conneau et al., 2020), (Babu et al., 2021), and (Pratap et al., 2023), respectively; *small* and *large* mean the model configurations with 24 and 48 transformer blocks; *Average* represents the average of experimental results of the four languages; *+LM* means integrating the 5-gram language model with the acoustic model; The results are displayed in the format of CER/WER, and the smaller value indicates a better performance. The overall best models of experiment with and without language model are marked as bold. This work simply sets the experiment with the smallest sum of CER and WER as the best model.

small and large configurations. Note that Khinalug and Mboshi are not included in the pre-training languages of all multilingual models.

Among experiments, we observe that utilizing a multilingual pre-trained in building ASR models is beneficial to Khinalug and Mboshi and essential to Dhivehi and Danish. The experiment with the best pre-trained multilingual model outperforms that of the monolingual model by 2.30 and 1.54 CER points, as well as 6.56 and 4.71 WER points for Khinalug and Mboshi, respectively. Hence, we conclude that multilingual representation learning is an effective approach, whether or not the target language is included in the pre-training languages.

Impact of Dissimilar Pre-training Languages

The pre-training languages are linguistically similar and diverse to the target languages. Current work focuses on increasing the number of pre-training languages while neglecting language similarity. To examine the impact, this work experiments with multilingual models pre-trained with 53, 128, and 1,406 languages.

Results show slight differences among languages. The models leading to the best performance for Khinalug, Mboshi, Dhivehi, and Danish are pre-trained with 53, 1,406, 128, and 128, respectively. With the 5-gram language model, the best experiments for Khinalug and Dhivehi change from 53 to 1,406, and from 128 to 53. On average, the model pre-trained with 128 languages exhibits a minor performance advantage over other models for the target languages.

Considering the corpus differences in domains, speakers, recording conditions, etc., the performance is too minor to conclude which multilingual pre-trained model is better. Therefore, we conclude that the linguistically diverse language in pre-training has no clear impact on ASR performance.

Impact of Model Capacity

Previous research (Babu et al., 2021; Pratap et al., 2023) shows that a larger capacity of pre-trained multilingual models might result in better speech recognition performance, including low-resource scenarios. However, the extra parameters might need more labeled data in training, therefore leading to inferior performance. Therefore, this work examines the impact of model capacity on the Khinalug corpus by experimenting with small and large model configurations with 24 and 48 transformer blocks.

As Table 3 shows, the models with small configurations lead to the best performance of most experiments, except one experiment of Khinalug. However, the performance differences between the small and large configurations are slight. On average of four languages, we observe that increasing model capacity decreases model performance, especially in the experiments of Multi-128. Therefore, we conclude that using a larger capacity of the pre-trained model has no benefits to speech recognition of endangered and low-resource languages.

5.2. Multilingual Supervised Learning

Refer to Section 3.3, this work experiments with supervised learning using multilingual annotated datasets. Because of corpus availability, we could not select the languages closest to the target language. Hence, we choose the language for multi-

lingual supervised learning based on the sociolinguistic and diglossic situations for Khinalug, and the language family division for other languages. Considering the significant language diversity in multilingual self-supervised learning, we assume the impact of language dissimilarity is less in this experiment.

Specifically, we select Azerbaijani for Khinalug, Basaa for Mboshi as they are both in the Bantu language family, Hindi for Dhivehi as they are both in the Indo-Iranian language family and Swedish for Danish as they are both North Germanic languages. The training dataset of the target language and the linguistically similar language are concatenated, and other data splits of the target language are mixed[3]. The details of the datasets are available in Appendix B. Throughout this section, the multilingual models employed are initialized with parameters pre-trained with 53 languages.

The effectiveness of MRL in supervised learning is based on the assumption that the labeled data is insufficient to fully train a model. We experiment with different portions of the training data to validate this assumption and keep the test data unchanged. As shown in Table 4, increasing training data leads to improved performance, which confirms our assumption that the training data is insufficient to fine-tune the model entirely.

	Full	Half	Quarter
Mono	6.70/25.10	9.72/35.91	12.85/46.90
Multi	7.30/27.44	11.52/42.50	13.96/50.79

Table 4: Experiments about data sufficiency in supervised learning on Mboshi. The results are displayed in the format of CER/WER; *Mono* represents monolingual training data with only Mboshi, and *Multi* represents multilingual training data with Mboshi and Basaa; *Full, Half, and Quarter* represents using different portions of Mboshi training data.

Effectiveness of Multilingual Representation Learning.

Table 5 shows the experimental results of involving linguistically similar languages in self-supervised learning. With additional training data, the speech recognition performance for all languages decreases. We conclude that training with mixed labeled data harms speech recognition performance in the target language, even with access to more training datasets to perform speech recognition tasks. We assume that the difference between languages (even from the same language family) is significant and negatively impacts the training on performing speech recognition tasks directly.

[3]We also have experiments about balancing the training data from the target and the similar language. The results show that using all training data of a similar language gives better results.

	CER	WER
Khinalug	7.58	34.65
Khinalug + Azerbaijani	9.95	43.32
Mboshi	6.70	25.10
Mboshi + Basaa	7.49	28.11
Dhivehi	10.51	55.94
Dhivehi + Hindi	15.62	45.83
Danish	11.88	38.98
Danish + Swedish	16.80	52.31

Table 5: Experimental results of multilingual supervised learning. For clarity, adding a new language means training with data of both language and testing on data of the target language.

5.3. Impact of Training Data Quality

Unintelligible content is the speaking parts that are non-understandable during proofreading, exhibiting one challenge in documenting endangered languages (§ 2.3). Previous work mainly discarded the samples with unintelligible content to keep training data quality high. However, that makes the limited training data less. This section investigates the impact of training data quality on speech recognition performance by experimenting with involving and not involving training samples having unintelligible content. Note that the test data is unchanged, leaving the model to see if it is possible to recognize the content.

As Table 6 shows, with access to 3% more samples, training with data including samples with unintelligible content leads to a 0.47 WER points increase but a 0.48 CER points decline. Incorporating a 5-gram language model slightly harms the model performance in terms of both CER and WER.

We observe the predictions have no unintelligible symbol $ regardless of whether there are unintelligible training samples. The results show that the unintelligible content is challenging to speech recognition, suggesting future work to address this challenge.

	#Train	Exp3	+ LM
With	1107	8.06/34.18	9.1/37.79
Without	1078	7.58/34.65	8.82/37.6

Table 6: Speech recognition performance with and without samples consisting of unclear content. The result is displayed as CER/WER. *With* means training with unclear data, while *Whthout* means training without unclear data; *#Train* means the number of training samples.

6. Quality Assessment by Linguists

The performance of a speech recognition model is evaluated with metrics measuring the difference between prediction and the ground truth transcription, while previous research (Guillaume et al., 2022) indicates the potential gap between evaluation metrics and the number of actual corrections that linguists have to make on the model prediction. To investigate the gap, this work assesses the usefulness of the ASR model by analyzing the prediction quality with linguists. The prediction is from the model corresponding to multi-1406-small in Table 3 with the language model.

6.1. Performance Analysis

The speech recognition model is developed with the labelled data of the corpus and subsequently employed to predict transcriptions for new recordings. Following this paradigm, this work selects one recording that is not included in the corpus for assessment. The new recording is from one speaker covered (Speaker 1 in later available Appendix) in the corpus to remove the impact of speaker variation, and the new recording includes a total of 20 sentences.

In assessment, the linguist identifies and quantifies the audible mistakes at the word level for each sentence. The audible mistake indicates the wrong prediction from the model where the speaking is clear to hear by the linguist, while the error in evaluation metrics neglects the audibility. For comparative analysis, we calculate WER for each sentence, measuring the percentage of error, and we normalize the number of audible mistakes as the percentage of mistakes for each sentence by dividing the wrong words by the total number of words.

As shown in Figure 1, for 20 sentences in the new recording from the covered speaker, the average WER is 43.74, and the average percentage of mistakes is only 14.29. The percentage of audible mistakes is lower than WER for every sentence. With the explanation of every wrong prediction from the linguist, we find the following potential reasons for the difference.

As Khinalug is an endangered language, it is not always possible to find speakers with clear and fluent speaking. For example, the speaker of this recording has a loose denture, leading to some mispronunciations. Besides, as the phoneme inventory and orthography of Khinalug are still in development, the pronunciations of some phonemes are not always clearly distinguishable, especially through listening recordings.

Figure 2 shows one example of the new recording. The prediction has a WER score of 69.56 points. In analysis, three words are assessed as mistakes, and eight words are assessed as inaudible words, which are not mistakes from the model. Therefore, the percentage of mistakes is 13.64%.

In language documentation, linguists can handle the inaudible content with contextual information and linguistic knowledge, while the speech recognition model can not. Even with the contextual information from the 5-gram language model, the speech recognition model barely recognizes the inaudible content.

6.2. Robustness Analysis

Language documentation often involves multiple speakers, and a key practical concern revolves around assessing the robustness of the ASR model to speaker variations. In this section, we select three recordings from three new speakers that are nonexistent in the corpus. We calculate the CER and WER of the model prediction and linguist's transcription.

	CER	WER
Test	7.4	32.07
Covered speaker	11.15	43.74
New speaker 1	46.55	81.82
New speaker 2	68.35	94.12
New speaker 3	31.78	82.36

Table 7: Speech recognition evaluation on test data and four new recordings.

With the experimental results in Table 7, we find the model has clear performance degradation on the recordings of the covered speaker and significant performance degradation on the recordings of the new speakers. The degradation for the new speakers is serious, leading to unusable model prediction assessed by the linguist.

As the recordings are long audios of free monologues, the speech content might have apparent variations between recordings. Unlike speech recognition models for high-resource languages that are robust to speech content variation, speech recognition models for endangered languages might be biased on the speech content in the training data.

We investigate speech content variation by calculating Levenshtein Distance (LD) for every training sample. LD measures the similarity between two strings by quantifying the minimum number of single-character edits required to transform one string into another. For visualization, we use the ratio value, which is the normalized similarity between two strings, and the value 1 means the two strings being compared are identical.

As illustrated in Figure 3, the covered speaker and three new speakers to the training samples.

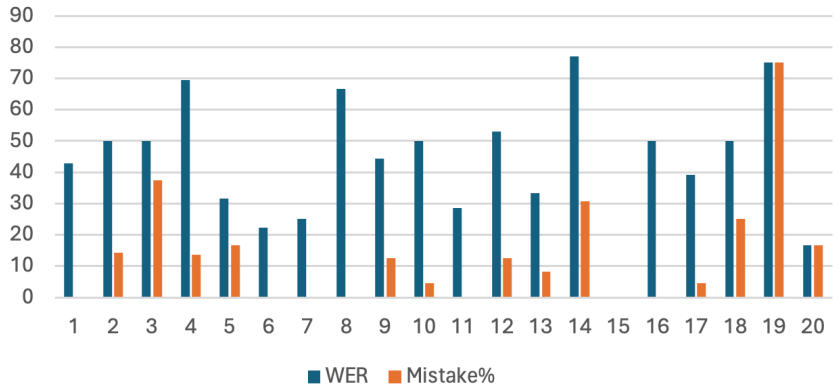

Figure 1: The WER and the percent of audible mistake for recordings of the covered speaker.

Prediction:
heç insanlış tərpəmiş tü xk̂olu sa kollatxunk̂oarişəvıżırıllı pşoa vıżırılli onğ vıżurılli heçĉu fi kank̂oarişəmə nəq̇ quba nə heş t̂u k̂oli

Transcription:
həç insanırzış tərpənmişf tü kolu sa kolu latxınk̂oarişəmə vıżırılli pşo vıżırılli onğ vıżırılli heĉĉu fi kank̂oarişəmə nə Quba nə heç t̂a koli

Figure 2: One example of the performance analysis. The green words indicate the prediction errors that are inaudible. The red words indicate the prediction errors that is heard by the linguist but predicted wrong by the ASR model, corresponding to the audible mistake in Section 6.1.

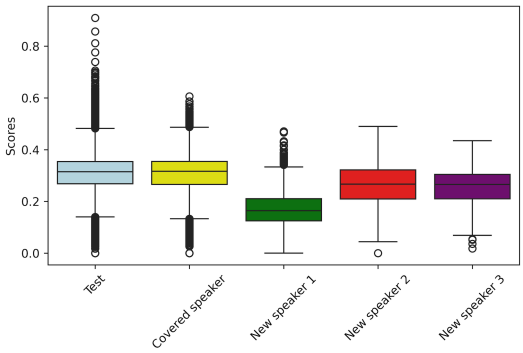

Figure 3: Visualization of speech content variations to the training samples.

We find the average LD ratios of the tested covered speakers are both 0.31, showing that there is no clear content difference between them. The average LD ratios for the three new speakers are 0.17, 0.27, and 0.26. However, the new speaker 1 has a lower average ratio but better performance than the new speaker 2. Hence, we conclude that the speech content variation is not the reason for performance degradation.

The performance degradation indicates that the ASR model needs improvement in robustness, suggesting future work to focus on model adaptation to new recordings, especially new speakers.

7. Conclusion

This work presents the Khinalug speech recognition corpus for exploring endangered language documentation. In addition, we show the effectiveness of multilingual representation learning in both self-supervised and supervised learning with models pre-trained with different numbers of languages. We also build ASR systems for this corpus. Lastly, we conduct a quality assessment with linguists to demonstrate the model's usefulness in language documentation. We observe a performance decline of the ASR model when applied to new recordings, and we find the model is inadequate to recognize inaudible content.

For future work, we consider integrating stronger language models with the acoustic model, such as the transformer-based language model. The integration aims to infuse contextual information into the ASR system, thereby enhancing its ability to handle inaudible content effectively. Additionally, we consider incorporating data augmentation techniques, such as synthesis data generation, to improve model robustness to new recordings.

8. Bibliographical References

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.

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

John Clifton, Laura Lucht, Gabriela Deckinga, Janfer Mak, and Calvin Tiessen. 2005. The sociolinguistic situation of the khinalug in azerbaijan. *SIL Electronic Survey Reports*, 2005-007.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

Ethnologue. 2024. Khinalugh | ethnologue free.

Jala GARIBOVA and Ildirim ZEYNALOV. 2023. Language change, language attrition and ethnolinguistic vitality of khinalug in azerbaijan: Is the quietly approaching threat reversible? *Tehlikedeki Diller Dergisi*, 13(22):27–54.

P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G-N. Kouarata, L. Lamel, H. Maynard, M. Mueller, A. Rialland, S. Stueker, F. Yvon, and M. Zanon-Boito. 2018. A very low resource language speech corpus for computational language documentation experiments.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of japhug (trans-himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2022. Enhancing documentation of hupa with automatic speech recognition. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192, Dublin, Ireland. Association for Computational Linguistics.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*. ISCA.

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. 2023. Scaling speech technology to 1,000+ languages.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Monika Rind-Pawlowski. 2023a. Some observations on the azerbaijani influence on khinalug. *Tehlikedeki Diller Dergisi*, 13(22):73 – 135.

Monika Rind-Pawlowski. 2023b. Verbal roots and verbal stems in khinalug. *Millenium*.

Lorena Martín Rodríguez and Christopher Cox. 2023. Speech-to-text recognition for multilingual spoken data in language documentation. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 117–123, Remote. Association for Computational Linguistics.

Nay San, Martijn Bartelds, Blaine Billings, Ella de Falco, Hendi Feriza, Johan Safri, Wawan Sahrozi, Ben Foley, Bradley McDonnell, and Dan Jurafsky. 2023. Leveraging supplementary text data to kick-start automatic speech recognition system development with limited transcriptions. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–6, Remote. Association for Computational Linguistics.

Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Simpson, and Dan Jurafsky. 2022. Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 41–51, Dublin, Ireland. Association for Computational Linguistics.

Claytone Sikasote and Antonios Anastasopoulos. 2021. Bembaspeech: A speech recognition corpus for the bemba language.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

UNESCO. 2023. Unesco's atlas of the world's languages in danger. Accessed on October 19, 2023.

A. Unintelligible Content Statistics

In this section, we provide statistics on the occurrence of unintelligible content in each long recording. As shown in Table 8, the proportion of the sentences with unintelligible content, which is marked as $, is 2.6%. After segmentation and dataset splitting, 29 out of 1107 training samples have unintelligible content, and 3 out of 123 test samples have intelligible content.

Recording	#Sent	#Sent with $
Speaker1.1	107	3
Speaker1.2	54	1
Speaker1.3	20	0
Speaker1.4	52	0
Speaker1.5	30	1
Speaker1.6	20	0
Speaker1.7	64	4
Speaker1.8	135	5
Speaker1.9	69	0
Speaker1.10	60	0
Speaker1.11	61	1
Speaker1.12	48	0
Speaker1.13	33	2
Speaker1.14	79	1
Speaker1.15	17	2
Speaker1.16	25	0
Speaker1.17	50	1
Speaker1.18	47	2
Speaker1.19	79	2
Speaker1.20	51	2
Speaker1.21	43	2
Speaker2.1	22	0
Speaker2.2	28	1
Speaker3.1	36	2
Total	1230	32

Table 8: Unintelligible Content Statistics. *Speaker1.1* means the first recording of speaker 1

Language	Split	#Sample	#Hour
Azerbaijani	Train	39	0.04
	Validation	21	0.04
	Test	27	0.04
Basaa	Train	763	0.82
	Validation	457	0.43
	Test	528	0.62
Hindi	Train	4479	4.66
	Validation	2281	2.62
	Test	2947	3.66
Swedish	Train	7407	7.15
	Validation	5114	4.96
	Test	5120	5.69

Table 9: Dataset statistic for multilingual supervised learning. *#Sample* indicates the number of samples and *#Hour* indicates the number of hours.

B. Dataset statistic for multilingual supervised learning

This section shows statistics of datasets used in multilingual supervised learning (Table 9). All datasets are from Common Voice version 13.0.