Karlsruhe Institute of Technology

Speech Recognition Corpus of the Khinalug Language for Documenting Endangered Languages

Zhaolin Li, Monika Rind-Pawlowski, Jan Niehues

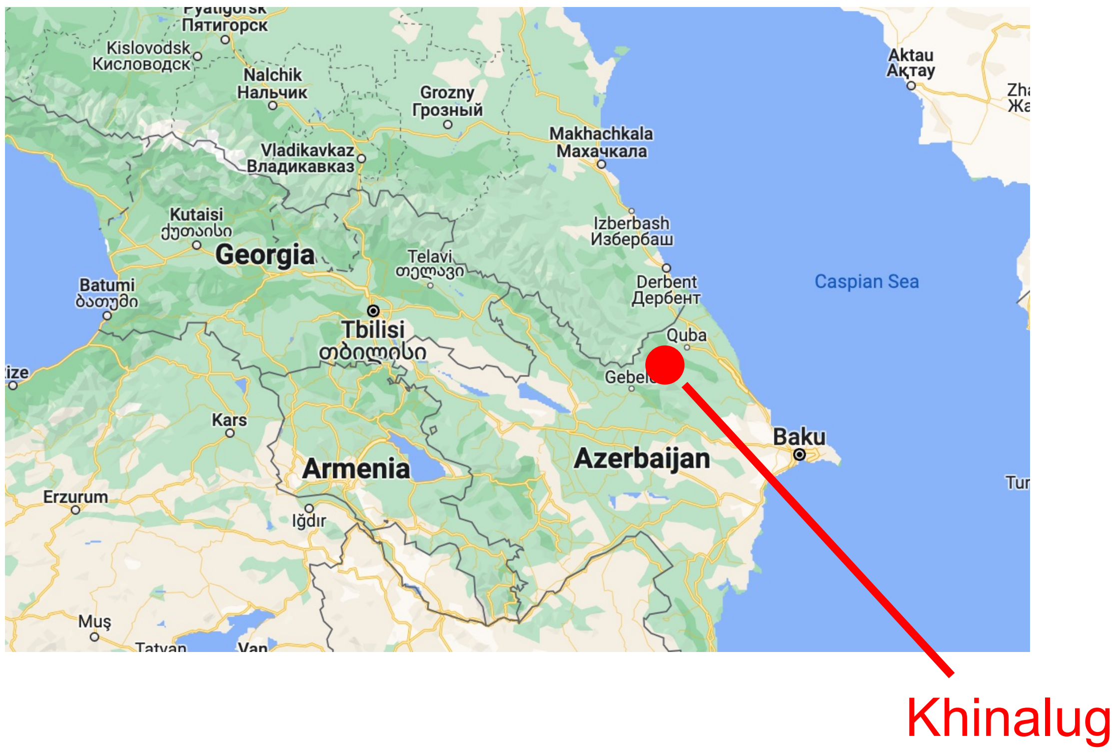
Language documentation

- >7000 languages & 2/3 endangered
- Manual documentation
  - Time consuming
  - Expensive
  - Consistency and accuracy
- Documenting language with ASR
  - Low resourced scenarios

https://vasco-translator.com/articles/languages/how-many-languages-are-there-in-the-world/

Khinalug

- Northern Azerbaijan with 2,300 speakers
- bilingual in Khinalug and Azerbaijani
- Recognized as a severely endangered language



Corpus

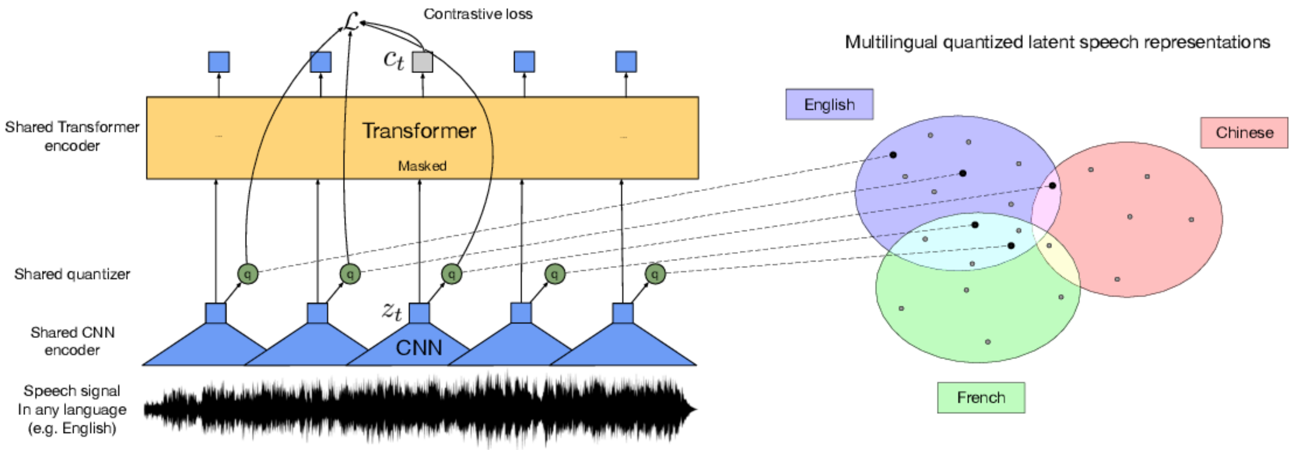
- Spontaneous speeches of native speaker
- Joint work with consultant and linguists
- Long audio segmentation
- 1,230 samples & 2.6 hours labelled data
- Challenges
  - Speaking Disfluency
  - Unintelligible Content $

	#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.

ASR for low-resource languages

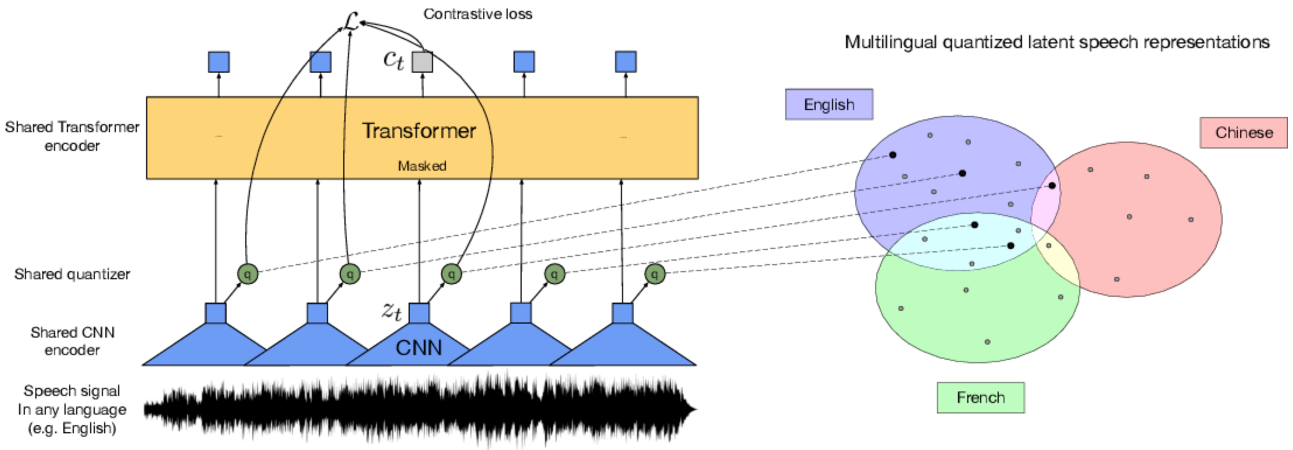
- Self-supervised Learning (SSL)
  - Pre-training
  - Fine-tuning
- Multilingual Representation Learning (MRP)
- Wav2vec2+CTC



Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning for speech recognition." *arXiv preprint arXiv:2006.13979* (2020).

ASR for low-resource languages

- Self-supervised Learning (SSL)
  - Pre-training
  - Fine-tuning
- Multilingual Representation Learning (MRP)
- Wav2vec2+CTC
- Questions
  - How good the ASR system is for Khinalug?
  - Dissimilar languages in MRP?
  - SSL or SL?



Conneau, Alexis, et al. "Unsupervised cross-lingual representation learning for speech recognition." *arXiv preprint arXiv:2006.13979* (2020).

ASR with language model

- Decoding with shallow fusion
- Ngram language model
- pyctcdecode package
- Word level

$$\log P_{\mathrm{LM}}(\text{text}) = \log P(\text{text}) + LM(\text{text})$$

Datasets

- Khinalug
- Three other languages
  - Endangered
  - Low-resourced

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.

Results – multilingual SSL

	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.

Results – multilingual SL

- Too little supervised data to fully fine-tuning the ASR model
- ASR corpus from similar languages
- Necessary to have multilingual supervised training?

	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.

Results – multilingual SL

	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.

Quality assessment

- New recordings
- New speakers
  - 21/24 are from one speaker

Quality assessment

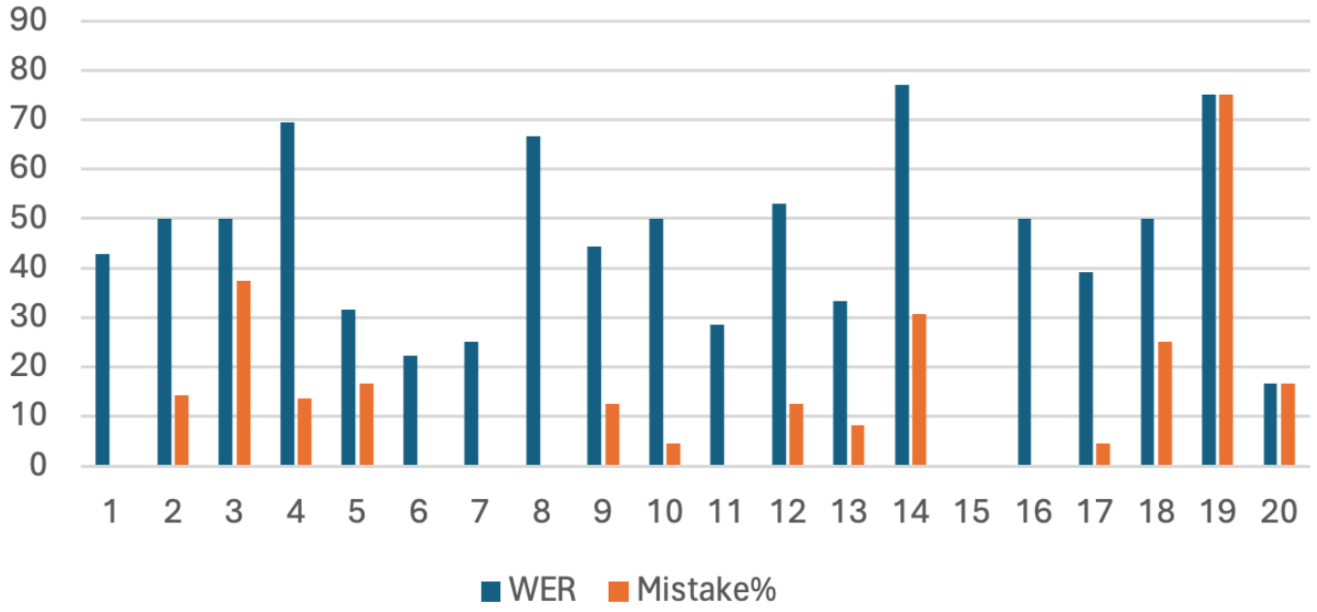
- New recordings
- New speakers
  - 21/24 are from one speaker

	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.

Linguistic Analysis

- Automatic evaluation
  - Potential gap to the actual corrections
- Linguist evaluation
  - Audible mistakes



Linguistic Analysis

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

Conclusion

- Speech corpus for Khinalug and ASR system
- MRL in low-resource languages
- Quality assessment
  - Robustness
  - Contextual information