Novel semi-supervised learning methods for indigenous languages.

Monica Romero Ferron[1], Ivan G. Torre[2],
1 romeroferronmonica@gmail.com Politecnic University of Madrid **2** Polytecnic University of Madrid, University of the Basque Country

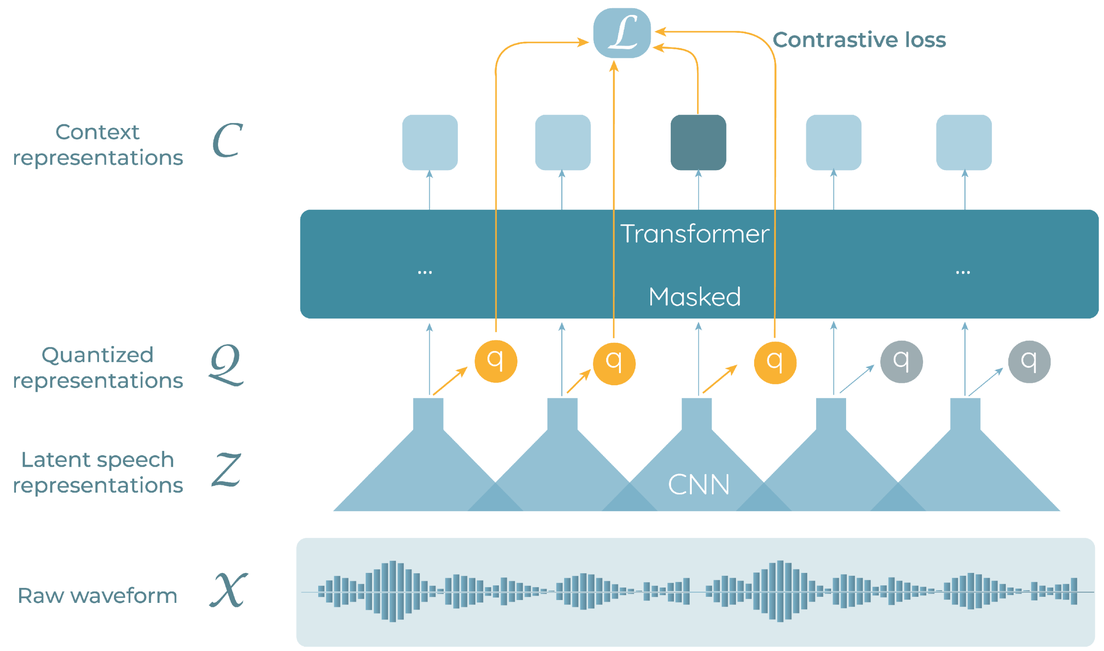
Introduction

There are 600 indigenous languages approximately in Latinamerica. Many of them are considered low resource languages due to the lack of information registered or conserve about them. This fact bias the studies and research done to this languages for instance the training of algorithms to recognice them as they are in more spoken languages as Chinese, Spanish or English.

This project applies semisupervised end to end learning methods for automatic speak recognition in 5 indigenous languages from Brazil, Colombia, Peru, Costa Rica and Paraguay.

Methodology

We explore the application of novel semi-supervised end-to-end (E2E) learning methods on automatic speech recognition (ASR), specifically wav2vec2.0 architecture. We have finetuned wav2vec2 XLS-R 300M model for all cases but for the Kotiria where the best results were achieved by finetuning an XLS-R 1B model.



Dataset

In addition to the dataset available from America's challenge, we have collected transcribed speech for three indigenous languages: Quechua, Kotiria, and Bribri. The acoustic model benefited by adding speed augmentation techniques, so the original audio speed was modified with a factor of 0.9 and 1.1 to produce two alternative versions. Additionally, spec augmentation was applied online during finetuning to increase the robustness and generalization of the models. On the decoding, we explored several techniques that included greedy decoding, Beam search with LM trained on transcribed acoustic data and Beam Search with LM trained on externally collected data. However, given the lack of transcribed data, LM did not generalize correctly for four of the language where the best decoding option was to apply a greedy decoding followed by some heuristic corrections.

Language	Number of hours in training dataset			
	Challenge dataset	External dataset	Augmentation dataset	Total
Bribri	0.49	0.91	0.98	2.38
Guaraní	0.32	-	0.65	0.97
Kotiria	2.69	21.8	5.43	29.92
Wai'khana	1.45	-	4.66	6.11
Quechua	1.67	7.04	3.38	12.09

Final Results

	Bribri	Guaraní	Kotiria	Wai'khana	Quechua	Total Average
Character Error Rate (CER)	**0.3470**	**0.1559**	**0.3659**	**0.3523**	**0.1214**	**0.2685**