

---

# CLSRIL-23: CROSS LINGUAL SPEECH REPRESENTATIONS FOR INDIC LANGUAGES

---

Anirudh Gupta<sup>\*</sup>, Harveen Singh Chadha<sup>†</sup>, Priyanshi Shah<sup>1</sup>, Neeraj Chimmwal<sup>1</sup>, Ankur Dhuriya<sup>1</sup>, Rishabh Gaur<sup>1</sup>, and Vivek Raghavan<sup>2</sup>

<sup>1</sup>Thoughtworks  
<sup>2</sup>Ekstep Foundation

## ABSTRACT

We present a CLSRIL-23, a self supervised learning based audio pre-trained model which learns cross lingual speech representations from raw audio across 23 Indic languages. It is built on top of wav2vec 2.0 which is solved by training a contrastive task over masked latent speech representations and jointly learns the quantization of latents shared across all languages. We compare the language wise loss during pretraining to compare effects of monolingual and multilingual pretraining. Performance on some downstream fine-tuning tasks for speech recognition is also compared and our experiments show that multilingual pretraining outperforms monolingual training, in terms of learning speech representations which encodes phonetic similarity of languages and also in terms of performance on down stream tasks. A decrease of 5% is observed in WER and 9.5% in CER when a multilingual pretrained model is used for finetuning in Hindi. All the code models are also open sourced. CLSRIL-23<sup>3</sup> is a model trained on 23 languages and almost 10,000 hours of audio data to facilitate research in speech recognition for Indic languages. We hope that new state of the art systems will be created using the self supervised approach, especially for low resources Indic languages.

## 1 Introduction

Speech recognition has made remarkable progress in the past few years especially after the advent of deep learning Hinton et al. [2012]. End to end (E2E) models have surely simplified the modelling process but they are also notoriously known for huge amount of data requirements. Especially more so for low resource languages Besacier et al. [2014].

This is of particular importance for countries with many languages and dialects such as India which has 22 official languages with an additional 1500 minor languages/dialects. Apart from a few major languages, most of the languages are low resource, thereby making it more difficult to develop speech related technologies Bourlard et al. [2011].

(E2E) networks become an attractive choice for multilingual ASRs since they combine the acoustic model, pronunciation and lexicon model into a single network. One way to tackle with multiple languages using a single model would be to train a multilingual ASR model we can take a union over all the language characters and jointly train a model on all the languages. But even in that approach huge amounts of data is needed per language.

In recent years self supervised learning has emerged as a new paradigm in which representations are learnt from the data itself and then fine tuning is done on several other down stream tasks. This approach has been widely successful in natural language processing (NLP) applications Devlin et al. [2019]Peters et al. [2018] and is active area of research in other fields.

In the past few years self supervised learning has been actively studied for speech recognition. In Jiang et al. [2019] the authors perform unsupervised pretraining with masked predictive coding using a transformer model. Most of work

---

<sup>\*</sup>anirudh.gupta@thoughtworks.com

<sup>†</sup>harveen.chadha@thoughtworks.com

<sup>3</sup><http://github.com/Open-Speech-EkStep/vakyansh-models>

in this space as well has been in monolingual speech recognition Chung and Glass [2018]Tjandra et al. [2019]Jiang et al. [2019]Harwath et al. [2020]. Our approach is based on the wav2vec 2.0 Baevski et al. [2020] the details of which are explained in the coming sections. An approach which uses multiple languages in pre-training and fine-tuning is described in Conneau et al. [2020]. It also shows that cross lingual pre-training outperforms monolingual pre-training. We extend the work by Riviere et al. [2020] and Conneau et al. [2020] by pretraining only on Indic languages so that speech recognition tasks have a better performance on Indic languages.

Languages spoken in the South Asian region belong to at least four major language families: Indo-European (most of which belong to its sub-branch Indo-Aryan), Dravidian, Austro-Asiatic, and Sino-Tibetan. Almost one third of our mother-tongues in India (574 languages) belong to the Indo-Aryan family of languages - spoken by 73.30% of Indians. The Dravidian languages, 153 in number, form the second major linguistic group of the country (24.47% )<sup>4</sup>. Since most of the 23 we have used are in common language families we aim to utilise language similarity to aid representation learning for low resource languages.

## 2 Modeling Approach

The method we use masks the speech input in the latent space and solves a contrastive task defined over a quantization of the latent representations which are jointly learned Baevski et al. [2020] , and shows that powerful representations can be learnt from speech audio alone. The approach encodes raw speech audio via a multi layer convolutional network and then masks resulting latent speech representations, similar to Devlin et al. [2019]. The latent representations are fed to a Transformer network to build contextualized representations and the model is trained via a contrastive task where the true latent is to be distinguished from distractors. The discrete speech units are learnt by a Gumbel softmax Jang et al. [2017] to represent the latent representations in the contrastive task.

### 2.1 Pre-training

The main body of the architecture consists of a CNN based feature encoder, a Transformer based sequence network and a quantization module. The feature encoder maps the raw waveform  $X$  to latent speech representations  $Z$ . These representations are then fed to a transformer block to generate context representations  $C$ , capturing the information in the entire sequence. The quantization module is used to discretize latent speech representations  $Z$  into  $Q$ . Given  $G$  codebooks with  $V$  entries where dimension of each codebook is  $\mathbb{R}^{V \times d/G}$ . Row corresponds to an entry in the codebook. We choose one entry from each codebook and concatenate the resulting vectors  $e_1, e_2, \dots, e_G$ . This concatenated vector is then mapped to  $q$  using a linear transform. The Gumbel softmax enables choosing discrete codebook entries in a fully differentiable way and probabilities for choosing the  $v$ -th codebook entry of group  $g$  are -

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^V \exp(l_{g,k} + n_k)/\tau} \quad (1)$$

where  $\tau$  is a non-negative temperature,  $n = -\log(-\log(u))$  and  $u$  are uniform samples from  $\mathcal{U}(0, 1)$ .

The speech representations during pre-training are learnt by a contrastive task  $\mathcal{L}_m$ . This is augmented by a codebook diversity loss  $\mathcal{L}_d$ .

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d \quad (2)$$

where  $\alpha$  is a tuned hyperparameter. Where:

$$\mathcal{L}_m = -\log \frac{\exp(\text{sim}(q_t, c_t)/\kappa)}{\sum_{\tilde{q} \sim Q_t} \exp(\text{sim}(\tilde{q}_t, c_t)/\kappa)} \quad (3)$$

$$\mathcal{L}_d = \frac{1}{GV} \sum_{g=1}^G \sum_{v=1}^V \bar{p}_{g,v} \log \bar{p}_{g,v} \quad (4)$$

$\mathcal{L}_m$  is the contrastive loss to make the model distinguish true representations from latent distractors  $\tilde{q}$ . In equation 3  $\text{sim}$  is the cosine similarity. The diversity loss  $\mathcal{L}_d$  is designed to increase use of the quantized notebook representations.

<sup>4</sup><https://www.education.gov.in>

	Assamese	Bengali	Bodo	Dogri	English	Gujarati	Hindi	Kannada	Total
<i>train</i>	254.9	331.3	26.9	17.1	819.7	336.7	4563.7	451.8	6802.1
<i>valid</i>	1.2	1.7	0.13	0.07	3.7	1.7	23.52	2.14	34.16
	Maithili	Konkani	Malayalam	Manipuri	Marathi	Nepali	Odia	Punjabi	
<i>train</i>	113.8	36.8	297.7	171.9	458.2	31.6	131.4	486.05	1727.45
<i>valid</i>	0.6	0.19	1.56	0.9	2.2	0.18	0.63	2.53	8.79
	Santali	Tamil	Telugu	Urdu	Kashmiri	Sanskrit			
<i>train</i>	6.56	542.6	302.78	259.68	67.8	58.8			1238.22
<i>valid</i>	0.03	2.67	1.41	1.20	0.37	0.30			5.98

Table 1: Language wise duration of *train* and *valid* sets in hours

## 2.2 Fine-tuning

Pre trained models are fine-tuned by adding a fully connected layer on top of the context network with the size of output layer equal to the vocabulary of the task. Models are optimized using a CTC loss Graves et al. [2006]. During training only weights of the transformer module are updated but not of the feature encoder module.

## 3 Training Data

All our data has been processed through the open sourced framework called *Vakyansh*<sup>5</sup>. The basic steps of the process are -

- Download and convert audio to *wav* format with sample rate 16000, number of channels 1 and bit rate per sample of 16.
- We split an audio into voiced chunks using voice activity detection<sup>6</sup>. We make sure that all the voiced chunks lie between 1 and 30 seconds.
- To detect and reject noisy samples we use a signal to noise ratio (SNR) approach described by Kim and Stern [2008]. We consider any audio sample below a SNR value of 25 as noise and do not include them in training data.
- We perform speaker and gender identification on our audio data. A high level representation of voice is learnt using a voice encoder based on Wan et al. [2020]. For each audio sample the voice encoder creates a 256 dimensional encoding that summarizes characteristics of the spoken voice. For gender identification we train a support vector machine algorithm on the embedding with manually labelled data. Our goal for speaker identification was to get a sense of the number of speakers in a particular audio source. To estimate we use a hierarchical clustering approach to cluster *similar* embeddings in the sense of cosine similarity. The number of speakers are thus the number of clusters.

From table 1 it can be seen that we have a total of 9816.7 hours of audio data out of which 9767.77 hours is training data and 48.93 hours is validation data in 23 Indic languages overall. To compare cross language exchange we also pretrain a model using only 4563.7 hours of Hindi data.

### 3.1 Finetuning data and Language Model

For our current work we are using labelled data only for Hindi. The labelled data is a combination of purchased data and transcripts generated from commercial speech to text engines. We normalize the text before doing any finetuning. Any punctuation is removed and numbers are converted to word format. For language modelling we use a statistical language model based on KenLMHeafield [2011]. We use a 5-gram language model with a beam size of 128. The text for language model consists text in the transcribed speech and Hindi data open sourced here <sup>7</sup>.

<sup>5</sup><https://open-speech-ekstep.github.io/>

<sup>6</sup><https://webrtc.org/>

<sup>7</sup><https://indianlp.ai4bharat.org/corpora/>

## 4 Model Training

All models are implemented in fairseq Ott et al. [2019] library.

### 4.1 Pretraining

We use the *base* architecture of the wav2vec 2.0 framework. It has 12 blocks with a model dimension of 768 and 8 attention blocks. The pretraining is restored from a checkpoint which is trained on 960 hours of librispeech data. We chose the base architecture over large since it is faster to train as it has almost half the parameters of large architecture. Base architecture also has a much lesser inference time when finetuned for speech recognition. We crop audio samples at 250,000 audio frames or 15.6 seconds and use a dropout of 0.1. The model is trained for almost 300,000 steps and start with a learning rate of 0.0005. We optimize using Adam Kingma and Ba [2017] where the first 32,000 steps are used as warmup updates for the learning rate after which it is linearly decayed. A weight of  $\alpha = 0.1$  is used for diversity loss  $\mathcal{L}_d$  in equation 2. We use  $G = 2$  codebooks with  $V = 320$  entries each for the quantization module. We train on 16 Nvidia A100 GPUs when performing pretraining on 23 languages and on 8 Tesla V100 GPUs when training on Hindi. It took around 100 hours to reach a stage where we did not see any gain in code perplexity. More details about training can be seen in the training logs<sup>8</sup>.

### 4.2 Finetuning

To finetune a model on speech recognition downstream task, a fully connected layer is added on top of the transformer block in which the output labels are the characters for the respective language. While finetuning the weights of the feature encoder are fixed. We finetune until we get the lowest WER on the valid set. Some features of the feature encoder are masked for data augmentation. It is a technique similar to SpecAugment and detailed out in Baevski et al. [2020]. All finetuning is performed on 8 Tesla V100 GPUs.



Figure 1: Language specific pretraining loss on valid set

<sup>8</sup><https://wandb.ai/harveenchadha/EKSTEP-PRETRAINING?workspace=user-agupta12>

## 5 Results

We firstly demonstrate that multilingual pretraining outperforms monolingual pretraining by calculating language specific loss for all 23 languages. Our experiments also show that there is a decrease in WER when multilingual pretraining model is used instead of monolingual.

### 5.1 Effectiveness of Cross Lingual Representation Learning

We calculate the language wise loss (contrastive and codebook) for audio training in both scenarios: when we have 23 languages in pretraining and when we have just Hindi. Figure 1 shows that for all languages apart from Hindi, the loss is lesser in the multilingual pretraining case. This is expected since low resource languages benefit from multilingual pretraining. A lower loss also indicates that more meaningful speech representations are being learnt as had been shown in Conneau et al. [2020].

We also analyze shared discrete speech representations for different languages. For each language, we sample 200 utterances and extract the quantized representations of the pretrained model. These vectors are normalized for each language to obtain vectors of size  $V \times G$ . K-means clustering is performed on these vectors and then the dimensions are reduced by PCA.

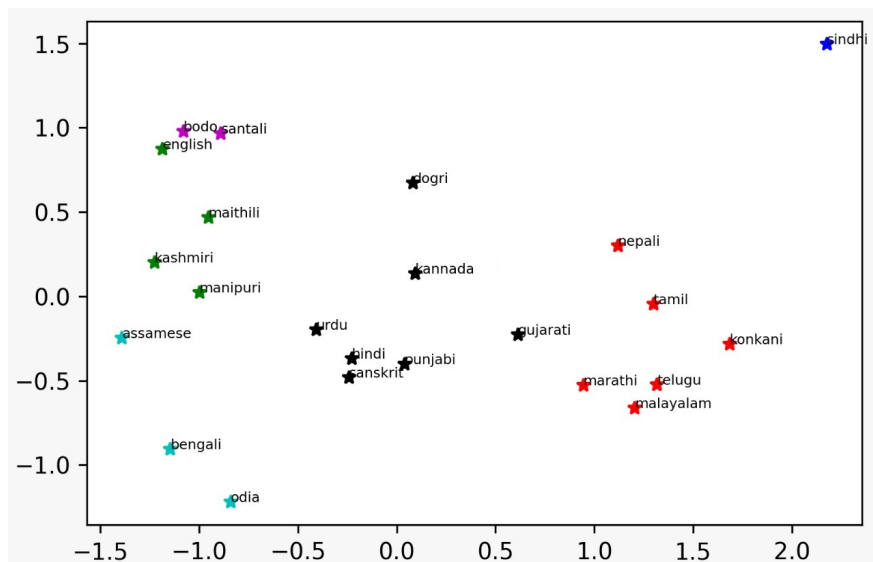


Figure 2: Quantized speech representations

From 2 we see that most phonetically similar languages are clustered together. The colours correspond to the clusters obtained by K-Means. We perform clustering before PCA to avoid any information loss. Assamese, Bengali and Odia are in one cluster. Hindi, Sanskrit, Urdu and Punjabi are also under one cluster. Most of the South Indian languages are clustered together as well, apart from Kannada. English in Indian accent is not a monolith. In our training data, English contains accents from different parts of India. As a result, it is far from many major languages and is together with other low resource languages. The purpose of this plot is not to recover underlying language families, but to check whether pretraining was able to learn phonetic information from a language.

### 5.2 Effect on finetuning

We finetune on 4563.7 using pretrained checkpoints from monolingual and multilingual pretraining. We see that even the high resource language, in our case Hindi, benefits from multilingual pretraining. On a separate 10 hour test we see a 5% decrease in WER and a 9.5% decrease in CER when decoding is done without a language model. It can be also seen from table 5.2 that the WER decrease by 3% and CER by 9% when using a language model while decoding.

Pretraining	Finetuning	Decoding	WER	CER
monolingual	Hindi	Viterbi	26.2	9.4
multilingual	Hindi	Viterbi	24.7	8.5
monolingual	Hindi	KenLM	16.26	8.9
multilingual	Hindi	KenLM	15.7	8.09

Table 2: Effect of multilingual and monolingual pretraining on WER and CER

## 6 Conclusion

In this work we present a multilingual pretrained model on 23 Indic languages in which representations are learnt from raw waveforms. Our results indicate that multilingual pretraining outperforms monolingual pretraining while learning speech representations while pre-training and also during finetuning performance in a downstream speech recognition task. We also show that model is able to encode phonetic similarity in speech representations. We hope this work kick starts development of high quality speech recognition Indic languages, especially low resource languages. All our code<sup>9</sup> and models<sup>10</sup> are open source. Our pretrained model can also be used for unsupervised speech recognition, especially for languages where no labeled text is present using the methods described in Baevski et al. [2021]. We plan to report finetuning results on all 23 Indic languages in the future.

## 7 Acknowledgements

All authors gratefully acknowledge Ekstep Foundation for supporting this project financially and providing infrastructure. A special thanks to Dr. Vivek Raghavan for constant support, guidance and fruitful discussions. We also thank Rajat Singhal, Heera Ballabh, Nireesh Kumar R, Sreejith V, Soujyo Sen and Amulya Ahuja for automated data pipelines and infrastructure support for data processing and model training.

## References

- G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012. doi:10.1109/MSP.2012.2205597.
- Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100, 2014.
- Herve Bourlard, John Dines, Mathew Magimai-Doss, Philip N Garner, David Imseng, Petr Motlicek, Hui Liang, Lakshmi Saheer, and Fabio Valente. Current trends in multilingual speech processing. *Sadhana*, 36(5):885–915, 2011.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi:10.18653/v1/N18-1202. URL <https://aclanthology.org/N18-1202>.
- Dongwei Jiang, Xiaoning Lei, Wubo Li, Ne Luo, Yuxuan Hu, Wei Zou, and Xiangang Li. Improving transformer-based speech recognition using unsupervised pre-training, 2019.
- Yu-An Chung and James Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech, 2018.
- Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. Vqvae unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019, 2019.
- David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech, 2020.

<sup>9</sup><https://github.com/Open-Speech-EkStep/vakyansh-wav2vec2-experimentation/tree/v2-hydra>

<sup>10</sup><https://github.com/Open-Speech-EkStep/vakyansh-models>

- 
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*, 2020.
- Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.
- Morgane Rivi re, Armand Joulin, Pierre-Emmanuel Mazar , and Emmanuel Dupoux. Unsupervised pretraining transfers well across languages, 2020.
- Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax, 2017.
- Alex Graves, Santiago Fern ndez, and Faustino Gomez. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *In Proceedings of the International Conference on Machine Learning, ICML 2006*, pages 369–376, 2006.
- Chanwoo Kim and Richard Stern. Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis. pages 2598–2601, 01 2008.
- Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification, 2020.
- Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/W11-2123>.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A fast, extensible toolkit for sequence modeling, 2019.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Unsupervised speech recognition, 2021.