Leveraging Cross Lingual Speech Representations To Build ASR For Under-resourced Languages

Sougata Mukherjee
Dept. of Electrical, Electronics and Communication Engineering
Indian Institute of Technology Dharwad
Dharwad, India
211022004@iitdh.ac.in

Prashant Bannulmath
Dept. of Electronics and Communication Engineering
Indian Institute of Information Technology
Dharwad, India
prashantb@iiitdwd.ac.in

Deepak K.T.
Dept. of Electronics and Communication Engineering
Indian Institute of Information Technology
Dharwad, India
deepak@iiitdwd.ac.in

S.R.Mahadeva Prasanna
Dept. of Electrical, Electronics and Communication Engineering
Indian Institute of Technology
Dharwad, India
prasanna@iitdh.ac.in

2023 26th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA) | 979-8-3503-4402-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/O-COCOSDA60357.2023.10482936

***Abstract*—Automatic speech recognition (ASR) technology can help document and preserve under-resourced tribal languages by converting spoken words into text. But building robust Automatic Speech Recognition(ASR) models for tribal languages is challenging due to scarcity of annotated data. This work shows the building of text and speech corpora for two tribal languages namely Lambani and Kui and attempts to develop an ASR system for those. Wav2vec2.0 is known to learn higher dimensional powerful representation directly from raw audio data and boost the performance of under-resource data. In order to overcome the limitation of scarce annotated data for under-resourced languages, wav2vec2.0 framework has been used along with data augmentation techniques like Gaussian noise addition, Pitch variation and Reverberation which increases the data synthetically. As Kui and Lambani are both indic languages, so in the wav2vec2.0 framework a model which has been pre-trained on 10k hours of data and 23 indic languages has been used to leverage the cross lingual representations for the development of ASR for these languages. Using wav2vec2.0 framework a WER of 41.55% and 49.57% and a CER of 12.49% and 12.52% has been achieved for 5 hours of Lambani and Kui respectively. The data augmentation techniques gives a relative improvement of 8.64% and 3.73% in terms of WER for Lambani and Kui. Besides this, this work also compares the performance of English language with the other two tribal languages which has already been used during pre-training with same amount of data setting and framework.**

***Index Terms*—Wav2vec2.0, Kui, Lambani, Mini-Librispeech, CLSRIL-23**

I. Introduction

Nearly half of the world's estimated 7000 languages will go extinct by next century. In India at least 197 languages are in various stages of extinction which is highest in the world.Most of these languages are spoken by the tribal groups spread across India. Digitising the audio and textual evidence of these tribal languages will help in preserving the languages and cultures of the tribal people who speak them. Automatic speech recognition (ASR) has the potential to support such digital preservation of these tribal languages,but building robust ASR models for such zero resource languages presents challenges.

In this paper, we present automatic speech recognition for such zero-resourced tribal language called Lambani and Kui. Lambani is the oral language which does not have its own script. It belongs to indo-aryan languge family [1]. It is spoken by a Banjara tribe spread across different part of India i.e.Karnataka, Andhra Pradesh, Telangana State, Maharastra and Rajasthan.We are mainly focussing on lambani spoken in karnataka state. In karnataka they have an estimated population of more than 40 millions mostly settled in the northern part of karnataka as depicted in the figure 1. Kui is Dravidian language [2] spoken by Khandas tribes from odisha with estimated population of 10 million.The major districts where these tribes live is shown in the figure 1. Due to modernization and influence of outer societies around them the tribal language and culture are in the danger of depletion. Therefore, there is an urgent need to protecting the linguistic identity of these tribal language

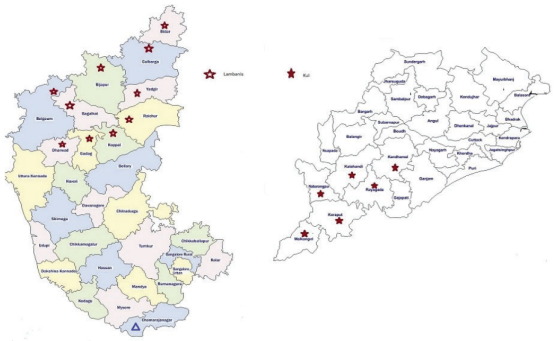

Fig. 1. The figure shows a map of Karnataka and Odisha state indicating the areas containing significant Lambani and Kui speakers

979-8-3503-4402-8/23/$31.00 ©2023 IEEE

The contributions of the proposed work are:

- Corpus created with Lambani and Kui speakers
- Character and word level transcription of 4760 Lambani and Kui sentences sentences
- This proposed ASR system of Lambani may become an initial step to preserve these zero-resource indigenous languages from the risk of extinction.

The significant challenges of working with these tribal language are:

- Lambani and Kui are zero-Resource Indigenous Language so does not have any proper script, dictionary, or writing system and speech files in digital domain.
- Generally, native speakers encounter difficulties in reading the Roman transcripts of Lambani, which leads to data collection being more complicated.
- Data collection takes lot of time and manual efforts.
- It becomes difficult to achieve high-performance such scarce amount of data.
- Building phoneme spoken language technologies and its analysis becomes difficult as these languages doesn't have pronunciation dictionary.

The remainder of the paper is structured as follows:- Section II shows the related works that has explored for the development of ASR for under-resourced languages. Section III describes the creation of Lambani and Kui corpora. Section IV explains the entire wav2vec2.0 framework used for the experimentation. The experimental details has been given in section V. This section also shows the performance obtained using wav2vec2.0 framework with and without data augmentation and does analysis of it. Section VI concludes the work and states the future scope.

II. Related Works

In early nineties, automatic speech recognition systems are built using statistical models based on Hidden Markov Model (HMM) [3]. These Conventional speech recognition frameworks uses acoustic model and pronunciation dictionary to build automatic speech recognition system [4]. In past decades, End-to-End deep learning have improved automatic speech recognition task remarkably [5]. End-to-End deep learning model uses large amounts of data and single architecture by eliminating complex process of creating pronunciation dictionary. End-to-End deep learning models have outperformed statistical based models in the speech recognition task over many datasets. The use of deep learning to build ASR acoustic models has resulted in substantial reductions in word error rate (WER) for high-resource languages. However, deep learning frameworks typically require very huge amounts of data making them hard to train for the under-resource scenarios typically encountered with many tribal languages. Speech recognition performance can be improved through transfer learning by leveraging knowledge from rich resource languages. In transfer learning pre-training is done using high resourced languages and finetuning is done on the target under resourced language. In recent years self-supervised learning is popular technique to do pre-training [6]. In self-supervised learning model will learn general purpose task agnostic speech representation which will be the starting point for downstream specific task. Contrastive Predictive Coding (CPC) is a popular self-supervised training criterion which uses contextual representations to predict the latent future representation [7]. Wav2vec use contrastive predictive coding to generate speech representation to achieve better performance [8]. Wave2vec2.0 [9] uses mask prediction with contrastive predictive coding to achieve better performance with 10 minutes of transcribed speech on Librispeech [10] test set. In this work, we are trying to build an ASR on indic under-resourced tribal language by leveraging the speech representations learnt from 23 indic languages. Our work is extended to data augmentation techniques with wav2vec2.0 framework.

III. Building Lambani and Kui corpora

A. Text data collection

Lambani and Kui are 2 languages which has spoken form, but no written form. So,written form had to be prepared with lot of manual efforts and time. We had to make sure that the native Lambani and Kui speakers can articulate the words easily. So, initially 1000 English sentences containing swadesh list [11] of words was prepared taking help from Linguist. These are 4 to 10 word sentences. Examples of short and long sentences from the swadesh list includes "All kids want sweet" and "Before I went to her house I changed my clothes'. An ASR system requires several hours of speech data to train. The duration of the recording of the sentences containing Swadesh list of words vary from 2 to 4 seconds. So, the number of sentences had to be increased following the same procedure as discussed above. During the increment in the number of sentences text had to be extracted from several sources. Major sources of English sentences were NCERT and Wikipedia. Among the books published by NCERT, we focused on English language textbooks meant for the students of the lower, middle, and higher secondary schools. For retrieving text, we used text extractors written by us using Python language. We used optical character recognition to extract sentences from publicly available scanned versions of the books on Lambani and languages using Adobe Reader's API.

B. Text processing

The text extracted from various sources quite often contains incomplete sentences, semantically incorrect sentences, and long sentences which may be difficult to speak for illiterate people. The following preprocessing steps were applied to the raw text extracted from various electronic sources.

- The passages of extracted text were processed to derive a set of sentences. The sentences containing fewer than 3 words were eliminated. Sentences longer than 10 words were removed as they will be difficult to utter for illiterate or older tribal people.

- Incomplete sentences, syntactically or semantically incorrect sentences and sentences containing symbols and characters not present in the Roman script were removed.
- Sentences containing words that may be too complex for a tribal person to speak were discarded. Text containing controversial statements including political statements was removed from the set of sentences.

The English sentences which successfully passed through the above-mentioned preprocessing steps qualify to be a part of the sentence corpus. The selected English sentences were converted to kannada and Odia languages(contact languages) using the Kannada and Odia script as it was the formal language in their respective area. Then, the Kannada and Odia sentences were translated to Lambani and Kui languages using the Kannada and Odia script respectively by the corresponding native speakers. The Lambani and Kui text data in kannada and Odia script respectively were preserved in digital format by writing them in a spreadsheet. This project is supported by Ministry of electronics and information Technology. So, the project involves building of several spoken language technologies. So,the amount of text data is increasing continously.

C. Speech files recording

These sentences collected as text data were spoken by multiple native Lambani speakers which was recorded using Laptop. Graphical User Interface (GUI) was designed to collect data through Laptop. The Lambani sentence spoken is displayed on the GUI. The recorded voice is replayed to asses its quality. If necessary, the GUI offers a feature to re-record the current sentence. Every speaker will record about one hour of data in seven sessions, which means almost 100 recordings per session.

The entire process of data collection strategy can be summarized as a flowchart given in figure 2

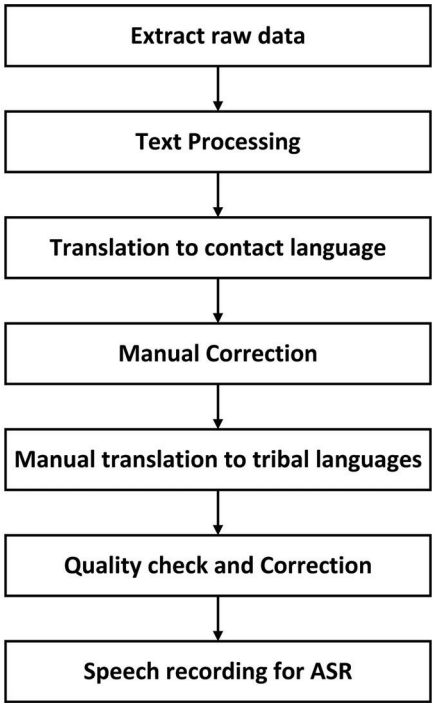

Fig. 2. ASR corpus creation flowchart

IV. Self-Supervised Learning for Acoustic Model Building

Self-supervised learning is a learning process which learns general representations from unlabelled data. In speech recognition task, speech representations are learned from unlabelled speech data. In this work, we are considering wav2vec2.0 [9] self-supervised learning framework for the development of ASR for under-resourced tribal languages. Self-supervised learning is also called pre-training. After this pre-training the network is fine-tuned with labelled data of that particular language in which ASR is to be developed.

A. Pre-training

The four models of operation in pre-training are: (a) raw waveform normalization (b) feature extraction using convolutional neural network (CNN) (c) masking of latent representation generated from CNN (d) learning of sequential information by passing through transformer network. (e) using contrastive loss and diversity loss as a training objective. As shown in figure 3 the pre-training task starts with normalization of raw waveform to zero mean and unit variance. After normalization, it is fed to a seven layered 1-D CNN network which encodes the speech to feature vectors Z for every 20 ms of the waveform which is continous in nature. It has a receptive field of 400 samples(sampling rate = 16kHz) and 25 ms. These feature vectors are discretized via product quantization to generate targets Q during contrastive task. The product quantization [12] involves choosing of quantized representations from multiple codebooks (groups) and concatenating them. The latent feature vectors Z generated from the feature encoder is fed to transformer encoder [13] which consists of 12 transformer blocks after masking a portion of the features. The transformer network learns the contextualized speech representations C for those masked portions of latent features by solving a contrastive task which tries to identify the true targets generated from quantization module from a set of distractors. Then the diversity loss is added along with the contrastive loss so that the codebook entries are used equally.

B. Finetuning

Finetuning is done on annotated data for downstreaming task. As mentioned in the figure 4 the block 'X' that got trained in the pre-training stage has been used in the finetuning stage. During finetuning, a randomly initialized softmax linear layer is added on the top of the transformer network and trained using trained using Connectionist Temporal Classification (CTC) criterion [14] in order to predict the characters. The size of the linear layer is equal to the size of the vocabulary of the language. The feature encoder parameters are fixed and the transformer encoder weights are updated during finetuning.

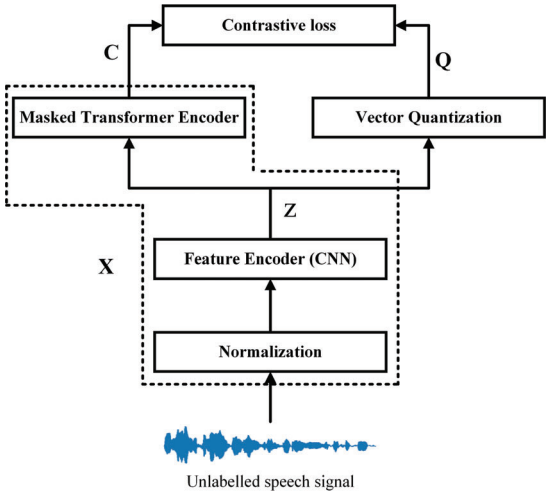

Fig. 3. Wav2vec2.0 pre-training architechture

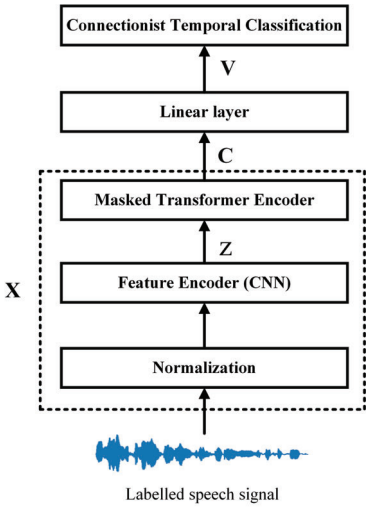

Fig. 4. Wav2vec2.0 finetuning architechture

V. Experimental Details

A. Data preprocessing

The mini-Librispeech dataset comes with 5hrs of training set named train-clean-5 and 2 hrs of development set named dev-clean-2. But, in our setting we have considered only 1 hr of testing data randomly choosing speaker files from the development set. The original sampling rate of Lambani and Kui dataset speech files were 44.1 kHz which was down-sampled to 16 kHz. The extra silence region in the audio files were removed by setting a threshold using end point detection algorithm. There was one text file which contained the entire text data of Lambani and Kui which was cleaned up by removing the punctuations present in the text files. These pruned text file was splitted into individual text file containing transcriptions corresponding to each audio file and naming them same as the names of the speech files with .txt extension. The detailed specification of the entire dataset created after prepocessing is given in the subsection V-B.

B. Dataset description

Initially, to build the framework, we considered an under-resourced(UR) setting of Mini-Librispeech data. Mini-Librispeech dataset is a subset of the Librispeech dataset [10]. The training set contains 15 female, 13 male speakers, and a total of 28 speakers. The testing set contains 7 male, 8 female speakers, and a total of 15 speakers. There are 1519 utterances, audio files and its corresponding text transcriptions. The training set has 534 utterances and audio files along with its corresponding transcriptions in testing set. That means there is one labelled audio file corresponding to each utterance.

The Lambani language dataset is also prepared in such a way so that it also contain 5 hrs of training data and 1 hr of testing data. It contains 9 male, 3 female speaker data and a total of 12 speakers in the training set. The dataset contains 2 female, 1 male and a total of 3 speakers data in the testing set. The training set contains 3990 utterances and annotated audio files. The testing set contains 770 annotated audio files and utterances. As Lambani language doesn't have its own script, so Kannada script is used from an ASR development point of view.

The Kui data has also been prepared in such a way so that it contains 5 hrs of training and 1 hr of testing data. There are 7 female, 4 male and a total of 11 speakers in the training set. The testing set contains 1 female, 2 male speakers which sums upto 3 speakers in total. The training set has 4599 utterances , speech files along with its corresponding transcription. The testing set of this dataset contains 1200 utterance and transcribed speech files. So, there is one audio file corresponding to each utterance. Kui doesn't have any script of its own so, the Odia script is being used for its transcription from an ASR development point of view.

Each speech files in every dataset has a sampling rate of 16 kHz, Mono channel, 16-bit PCM encoding, and a Bit Rate of 256 kbits/sec. The entire dataset specification is summarized in the Table I

C. Data Augmentaion

In this work, we artificially increased the training data by almost three times the actually collected data with different data augmentation techniques like Adding Guassian noise,Pitch variation and Reverberation.In adding guassian noise we added noise signal uniformly choosen from collection of audio files at SNR level of 15 db to create artificial copy of the existing speech signal.In pitch variation we randomly varied pitch by shifting in the range of (-400 400) to create the artificial copy of the existing speech signal. In reverberation we create far-field speech by convolving existing speech files with an artificially generated room impulse response (RIR).

D. Data preparation

Data preparation step involved in creating Audio and Transcriptions data at the utterance level, Dictionary of tokens and Lexicon. Data preparation requires parallel speech and

TABLE I
SPECIFICATION OF THE DATASET AFTER PREPROCESSING

Parameters	Mini-Librispeech	Lambani	Kui
Amount of training data	5 hrs	5 hrs	5 hrs
Amount of test data	1 hrs	1 hrs	1 hrs
Training spkrs	28(15 females,13 males)	12(9 males,3 females)	11(7 females,4males)
Testing spkrs	15(7 females,8 males)	3(2 females,1 male)	3(1 female, 2 male)
Sampling rate	16KHz	16KHz	16KHz
Bit rate	256kbits/s	256kbits/s	256kbits/s
Channels	1	1	1
Utterances in training set	1519	3990	4599
Utterances in testing set	534	770	1200
No. of audio file in training set	1519	3990	4599
No. of audio file in testing set	534	770	1200
Text script used	English	Kannada	Odia

transcript in specific format to read from the pipelines. Each audio dataset split (test/valid/train) is written into a separate tsv files with audio file path and number of audio frames. The validation split is done by randomly selecting 10% of speech files and its corresponding transcription from the training data. Similarly corresponding transcripts split (test/valid/train) are written into separate files. Dictionary of tokens consists of a list of all subword units (Bytepairs / phonemes /characters) used in the training data.In our case we are using characters to build a token dictionary. Lexicon file contains a list of unique words represented in terms of tokens present in the dictionary. A space split token is used to separate each line in the lexicon.

E. Experimental setup

We have carried out all our experiments using vakyansh toolkit [15]. CLSRIL-23 BASE model has been used for finetuning. CLSRIL-23 [16] model is pre-trained on 10 khours of data consisting of 23 indic languages has been used for finetuning. The amount of sub-processes used for data loading is decided by the parameter Number of workers. This value is set to 6. We have used a batch size of 16. Each batch has 32×10^4 samples. The zero infinity parameter was set to a true value which means if the loss value goes to infinity in the ctc criterion then it will be capped to 0. The maximum number of update is 20K. We have set the learning rate to 5×10^{-5}. The sentence average was set to true value. Finetuning was performed using GPU with an update frequency of 4. Adam optimiser was used with adam betas value of 0.9 and 0.98. Adam eps was set to 1×10^{-8}. We have used tri-stage scheduler where the learning rate is warmed up for the first 10% of updates, held constant for the next 40% and then linearly decayed for the remainder. The final learning rate scale was set to 0.05. Mask probability, mask channel probability and mask channel length are 0.65, 0.5 and 64 respectively. Layer dropout and activation dropout is set to 0.05 and 0.1.

F. Results and Discussion

The results obtained using wav2vec2.0 framework with data augmentation and without data augmentation for different UR settings has been tabulated in the table II. The data augmentation technique is giving relatively better performance because synthetically generated additional data is being used along with the original data for trainig the system. Here we can see that the performance of Mini-librispeech data is giving better performance in both the cases i.e with and without data augmentation as compared to other data because Mini-Librispeech dataset is based on English language which has been already used during pre-training. 819.7 hrs English language data has been used during the pre-training which is a huge amount as compared to the 5 hrs of under-resourced data.

TABLE II
THE ASR PERFORMANCE USING WAV2VEC2.0 FRAMEWORK WITH AUGMENTATION(W AUG) AND WITHOUT AUGMENTAIN(W/O AUG) OF DATA FOR DIFFERENT UR SETTINGS

Dataset		w/o aug	w aug
Mini-Librispeech	CER(%)	6.49	6.02
Lambani		12.49	11.58
Kui		12.52	11.12
Mini-Librispeech	WER(%)	18.22	16.80
Lambani		41.55	37.96
Kui		49.57	47.72

Here Word Error Rate is abbreviated as WER and Character Error Rate has been abbreviated as CER

VI. CONCLUSION AND FUTURE WORK

We have shown the data collection strategy of Lambani and Kui which are under-resourced languages. The focus was to show the effect of cross lingual representations learned from multiple indic languages for building an ASR on under-resourced languages. We have used a pre-trained model which was trained on 10 khours of data consisting of 23 indic languages. So, we have used the speech features from several rich languages to build an ASR for under-resourced languages. Our experiment also showed that if a language is already being

used during pre-training then it gives better performance when that particular language is used for finetuning. It has also been shown that data augmentation techniques improves the performance(WER) by 8.64% and 4.8% for Lambani and Kui respectively. As a part of future work, we can use finetuned wav2vec2.0 features to train other sate-of-art system to get better performance.

ACKNOWLEDGMENT

The authors would like to thank "Ministry of Electronics and Information Technology(MeitY)" for supporting us in the "Speech to Speech translation for tribal languages" project, which has helped us in collection of tribal languages data.

REFERENCES

[1] S. Madar, "Out-migration of lambani community: an observation," *Editorial Board*, vol. 5, no. 4, p. 175, 2016.
[2] W. W. Winfield, "A vocabulary of the kui language: Kui-english." Asiatic society of Bengal, 1929.
[3] B. H. Juang and L. R. Rabiner, "Hidden markov models for speech recognition," *Technometrics*, vol. 33, no. 3, pp. 251–272, 1991.
[4] F. Jelinek, B. Merialdo, S. Roukos, and M. Strauss, "A dynamic language model for speech recognition," in *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991*, 1991.
[5] W. Song and J. Cai, "End-to-end deep neural network for automatic speech recognition," *Standford CS224D Reports*, pp. 1–8, 2015.
[6] D. Jiang, X. Lei, W. Li, N. Luo, Y. Hu, W. Zou, and X. Li, "Improving transformer-based speech recognition using unsupervised pre-training," *arXiv preprint arXiv:1910.09932*, 2019.
[7] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[8] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.
[9] C. Yi, J. Wang, N. Cheng, S. Zhou, and B. Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.
[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
[11] M. Swadesh, "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos," *Proceedings of the American philosophical society*, vol. 96, no. 4, pp. 452–463, 1952.
[12] C.-C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," in *International Conference on Machine Learning*. PMLR, 2022, pp. 3915–3924.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[14] A. Graves and A. Graves, "Connectionist temporal classification," *Supervised sequence labelling with recurrent neural networks*, pp. 61–93, 2012.
[15] H. S. Chadha, A. Gupta, P. Shah, N. Chhimwal, A. Dhuriya, R. Gaur, and V. Raghavan, "Vakyansh: Asr toolkit for low resource indic languages," *arXiv preprint arXiv:2203.16512*, 2022.
[16] A. Gupta, H. S. Chadha, P. Shah, N. Chhimwal, A. Dhuriya, R. Gaur, and V. Raghavan, "Clsril-23: Cross lingual speech representations for indic languages," *arXiv preprint arXiv:2107.07402*, 2021.