Ekstep MUCS Approach

/thoughtworks

Abstract

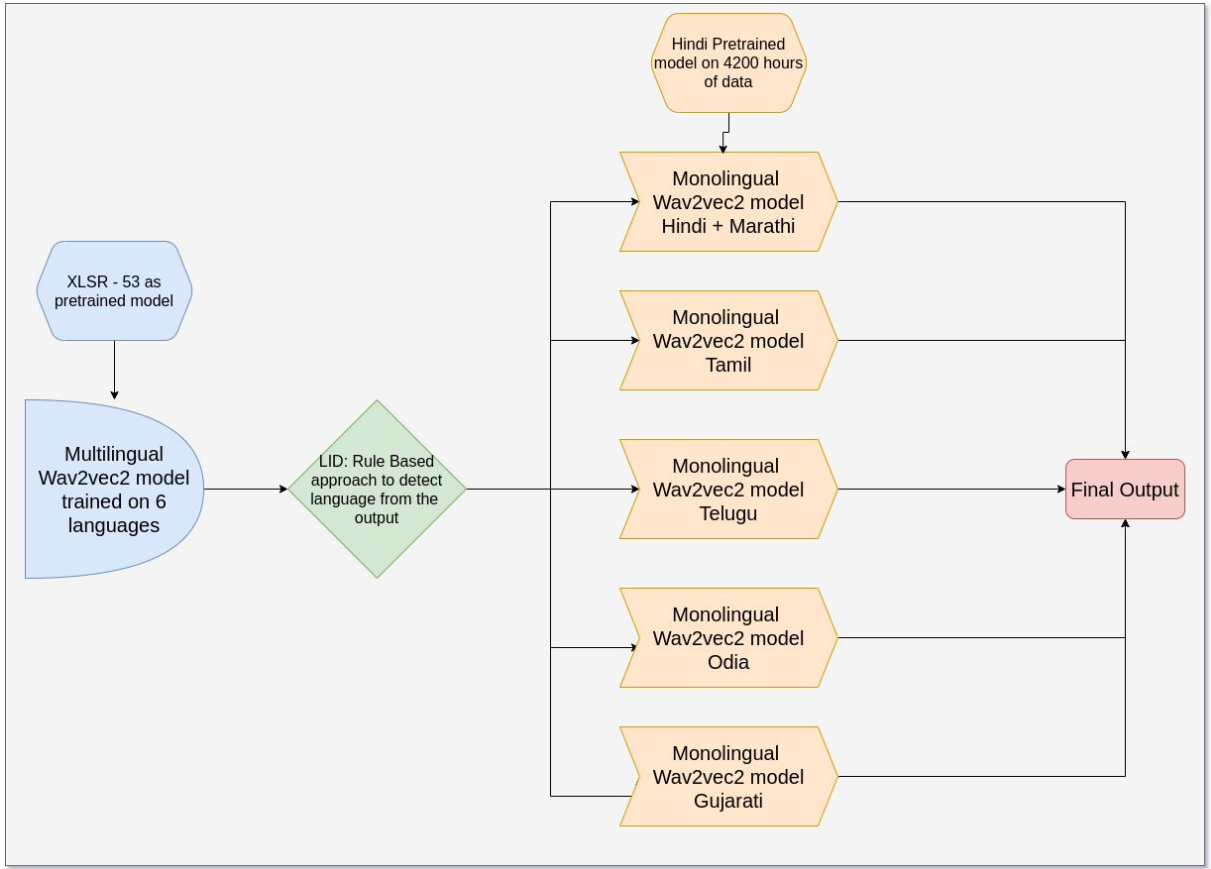
Training multilingual automatic speech recognition (ASR) systems is challenging because acoustic and lexical information is typically language specific. In this study we compare the performance of end to end (E2E) speech recognition systems trained on multiple languages to the performance of individual models conditioned on language identification (LID). We use LID and ASR in a complementary manner. The decoding information from a multilingual model is used for language identification.

The last step of decoding is done using monolingual E2E models for 6 Indian languages. For monolingual models we start finetuning on a particular language from a pretrained model in a high resource language. We see that the average WER across all languages decreases from 55.165 to 26.56 when we use LID information and language specific speech recognition models

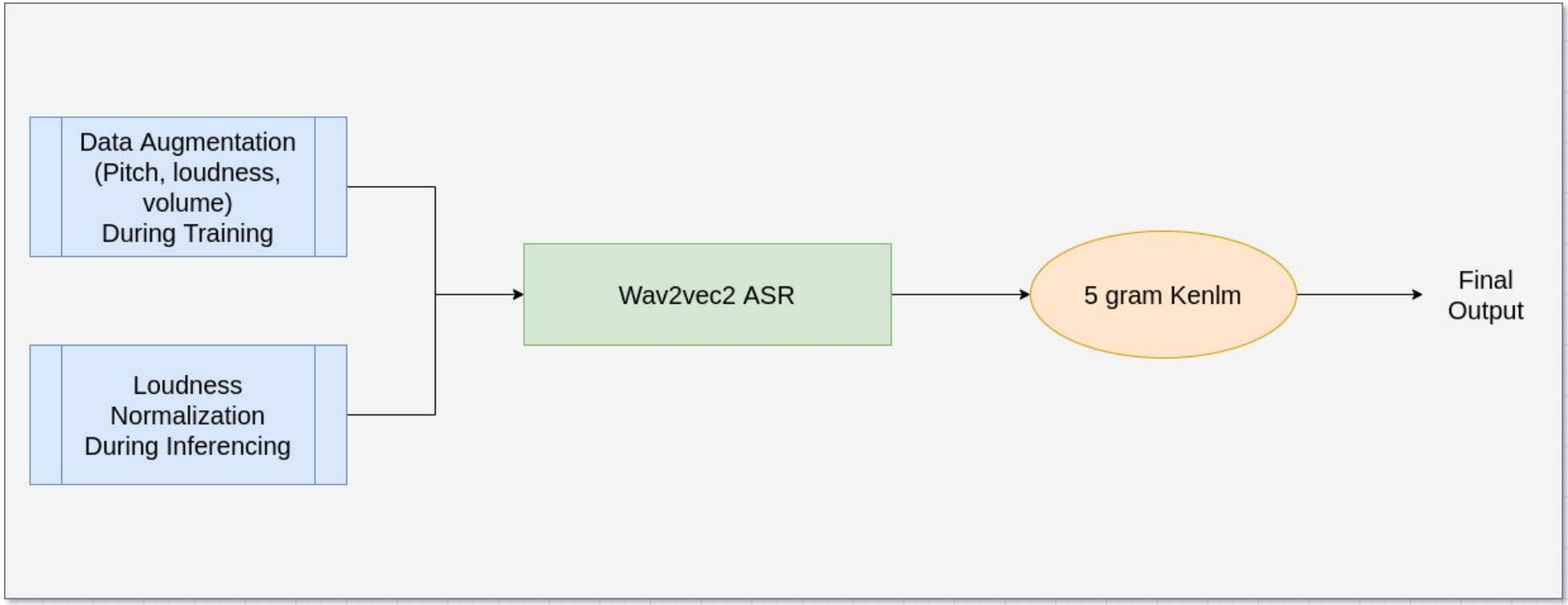
System Architecture

- We used self supervised learning approach to this problem
- The algorithm used is wav2vec 2.0
- Self supervised learning = unsupervised learning (pretraining with unlabelled data) plus supervised learning (finetuning with labelled data)

System Architecture



System Architecture



Results

Only using the multilingual wav2vec (XLSR-53 as pretrained)

Language	WER - Dev	WER - Test
Hindi	24.34	18.14
Marathi	16.02	103.68
Odia	31.71	29.44
Tamil	29.72	63.53
Telugu	28.7	41.33
Gujarati	17.2	74.87
Average	24.615	55.165

Results

Using the combination of multilingual as LID and monolingual

Language	WER - Dev	WER - Test
Hindi	21.51	12.24
Marathi	15.66	39.74
Odia	33.79	27.1
Tamil	28.61	27.2
Telugu	26.39	22.43
Gujarati	15.79	30.65
Average	23.625	26.65

Thank you