SPEECH RECOGNITION PROJECT REPORT

AUTOMATIC SPEECH RECOGNITION FOR THE NATIVE ZAMBIA LANGUAGE BEMBA

Mamadu Gueye[1], Claytone Sikasote[2]
{mgueye[1], csikasote[2]}@aimsammi.org

Abstract

In this report we present the results of the work of the in-class speech recognition project whose aim was to build an automatic speech recognition (ASR) for the native low resource languages. In this project, the speech dataset was recorded in the Zambian Language called Bemba. The contrastive predictive coding (CPC) pretrained model was fine tuned by training with on a 1hr train set and tested on another 1hr of test set. From the experiments, we obtained the character error rate (CER) of 0.92 -/+ 0.01.

1. Introduction

In this in-class speech recognition project, we were tasked to create a speech dataset of atleast two hours and build an automatic speech recognition on it.

2. Approach

2.1. Data collection

In this project we used a two (2) hours speech dataset that was obtained by eliciting texts gathered of the Zambian native/local language called Bemba. For recordings, we used the Infinix X680 phone running Android v10 on which the Lig-Aikuma software program was installed. As per requirements of this project, the dataset was split into 1:08hr train set and 1hr test set. Further, in order to train , the training set was split into 21.5 minutes validation set and 40.3 minutes training set.

2.1. Training settings

The training environmental settings for this project included using the pretrained CPC model to train and fine-tune the model with CTC loss using Adam optimizer, learning rate of 0.0002, batch size of 8 and window size of 20480. The model was trained for more than 50 epochs.

3. Results

As per project guidelines, the model was evaluated using the character error rate (CER) as the metric. Upon completing training and fine tuning the model we obtained the CER of 0.92 -/+ 0.01 on test dataset.

4. Challenges, limitations and lessons learnt

Despite the project being interesting, it was not short of challenges. Firstly, It was difficult to find literature on the internet of the bemba language of Zambia. Secondly, the performance of the Lig-Aikuma software program could crush during recording sessions. Thirdly, It was difficult to read some of the words in the native language. This could negatively have impacted the quality of the recordings.

We argue that we could have obtained better results had we tried out many things to reduce the CER. For example training a multilingual model, trying different model architectures, data augmentation as well as training on a big sized speech dataset. However, this was hampered by the fact that there was not enough time and the small size of the dataset. The other aspect that we did not do is carry out error analysis.

All in all, it was a very interesting project. It was interesting to understand the practically aspect of the course. We hope to scale the project up by building atleast good sized datasets of our local languages and train a multilingual model on them.

Github project repository link: https://github.com/csikasote/speech-recognition-project