Speech recognation

Amath Sow, Mbaye Babou, Sokhar Samb
African Master of Machine Intelligence(AMMI)

July 6, 2020

Introduction

The goal of this project is to collect data in our local language and fine tune a phone classification model. For our case we will be using wolof. Wolof is a language of Senegal, the Gambia and Mauritania, and the native language of the Wolof people.

Data Related Experiments

The first step of this project was to collect data using lig-Aikuma application. We use a text shared by the lecture to record the speech. The text has 2000 sentences that we split in 40 sessions of 50 sentences. After doing text elicitation, we obtain 2h:10mns of speech.

Algorithms related experiments

We first fine-tune a pre-trained model using a limited amount of labelled speech data. We are going to start with a simple evaluation setting where we have the phone labels for each time step corresponding to a CPC feature. We will work with a model already pre-trained on English data. We will use a 1h of our data.

However, aligned data are very practical, but in real life they are rarely available. That's why in the second part we will consider a fine-tuning with non-aligned phonemes. The model, the optimizer and the phone classifier will stay the same. However, we will replace our phone criterion with a CTC loss.

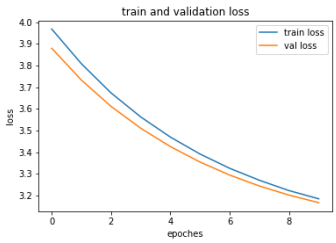
Results and discussion

When we train the cpc model in one epoch with aligned data, we got a low accuracy of 7%. So the next step was to fine tune the model. The following table show the performance of the model:

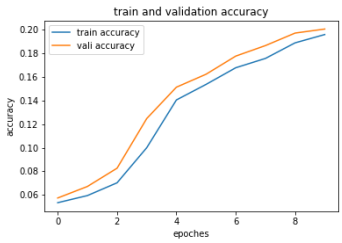
Model	**Accuracy**	**Loss**
fine tuning with aligned data + cross entropy loss	20%	3.1
fine tuning without aligned data +ctc$_{l}oss$	...	4.6

Table 1: This table show the loss and the accuracy we obtain by fine turning the cpc model.

We observe that the loss of the model fine tune with aligned data is less than the model fine tune with non aligned data. However all the two loss are high and the accuracy very small. This can be explain by the fact that the model was pre-trained in English with as very different phonemes with Wolof and also by the fact that we are . To have a overview of how our model perform, we visualized bellow the training and the validation loss both for the aligned phonemes and for the no aligned phonemes.



(a) Cross entropy loss



(b) Accuracy

This figure above show the loss and the accuracy of the fine turn model with aligned data. The flowing figure will gives the loss and the accuracy using the no aligned data.

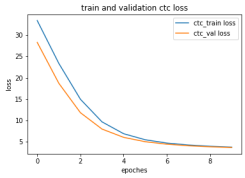


Figure 2: *ctc loss in training and validation*

In the figures above, we see that the loss training and validation are closed and both decreasing. Wich gives a good intuition about the performance of the model. We also evaluate the phones error(PER) rate and character error rate(CER) for the no aligned data in our test data which is show in the figure below.

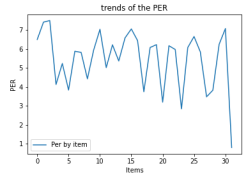


Figure 3: *Phones error rate with aligned phonemes*

We got an average phone error rate of 0.699 and an average character error rate of 0.68.

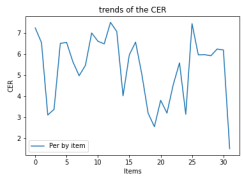


Figure 4: *character error rate with non aligned phonemes*

Conclusions

During our experiment, we realized that for the phone classifier we didn't get a high accuracy for the aligned data. Also for both model(model with aligned data and model with no aligned data), the loss are high. This can have many explanation. It can be due to the fact that there is a high difference between the phonemes in English and Wolof. It can be also due to the fact that we are not familial with reading Wolof, so it may have some miss pronunciation even if we have put a lot of effort to the data collection part to produce high quality data. However, we really enjoyed this project and we intend to collect more data and pursue research in this area. You can see the complete codes from from this github git and the data from google drive data