Centre for High Performance Computing 2021 National Conference

Contribution ID: **38** Type: **Student Micro-talk**

Multilingual Acoustic Modelling for improving ASR performance of Under-resourced Malian Languages

Automatic speech recognition (ASR) is a field of research that use statistical computing and machine learning to convert human speech into text. The United Nations is currently actively using ASR systems developed in our previous work to support their humanitarian relief programmes in rural Africa. These systems monitor community radio broadcasts in local languages to obtain information that can inform and support humanitarian decision making. In this work, we develop ASR systems for two additional and also severely under-resourced African languages, Bambara and Maasina Fulfulde, both of which are spoken in the country of Mali. Since no speech resources were available in these languages, corpus compilation had to begin from scratch. However, the accurate transcription of speech is time-consuming and expensive, since it requires specialist annotators with linguistic expertise. To date, we have compiled 10 hours of transcribed speech in both languages, which, in terms of the state of the art, is still considered severely under-resourced. As a dataset augmentation strategy, we take advantage of six similarly under-resourced datasets in other languages for multilingual training. These languages are Luganda, Acholi, Ugandan English and Tunisian Modern Standard Arabic, each of which has between 5.6 to 10 hours of data. We investigate whether different ways of combining this data as a training set can yield improved acoustic models for the target languages Bambara and Maasina Fulfulde. Our criterion for choosing the best combination of languages in the multilingual training pool is the word error rate (WER) performance metric.

All acoustic models were trained and evaluated using the Kaldi ASR. We followed the standard practice of training a hidden Markov model/Gaussian mixture model (HMM/GMM) system to obtain alignments for deep neural network (DNN) training. The DNN architecture used for the multilingual acoustic modelling follows the Librispeech recipe and consists of six convolutional neural network (CNN) layers followed by twelve factorised timedelay deep neural network (TDNN-F) layers. The hidden layers and the output layer are shared between languages and the lattice free maximum mutual information (LF-MMI) criterion is used as training objective. A SpecAugment layer precedes the CNN layers as it has proved beneficial in all experiments. Threefold speed perturbation is also applied during training.

The SRILM toolkit was used to train trigram language models, since higher order models did not provide any further benefit. First, a baseline language model is trained using the training transcriptions of the target domain data. Thereafter, language models trained on the additional text resources are interpolated with the baseline to obtain the final language model that is used in the ASR experiments. The interpolation weights were chosen to optimise the development set perplexity.

From our experiments we find that, although maximising the size of the training pool by including all six additional languages provides improved speech recognition in both target languages, substantially better performance

can be achieved by a more judicious choice. In fact, our experiments show that the addition of just one language provides the best performance. For Bambara, this additional language is Maasina Fulfulde, and its introduction leads to a relative word error rate reduction of 6.7%, as opposed to a 2.4% relative reduction achieved when pooling all six additional languages. For the case of Maasina Fulfulde, best performance was achieved when adding only Luganda, leading to a relative word error rate improvement of 9.4% as opposed to a 3.9% relative improvement when pooling all six languages. We conclude that careful selection of the out-of-language data is worthwhile for multilingual training even in highly under-resourced settings, and that the general assumption that more data is better does not always hold.

Student?

Yes

Supervisor name

Thomas Niesler

Supervisor email

trn@sun.ac.za

Primary authors: PADHI, Trideba (Stellenbosch University); VAN DER WESTHUIZEN, Ewald (Stellenbosch University); NIESLER, Thomas (University of Stellenbosch)

Presenter: PADHI, Trideba (Stellenbosch University)

Session Classification: Micro-talks

Track Classification: Cognitive Computing and Machine Learning