_GRz.png _GRz.bb _GRz.png height

Contribution ID: **59**

Type: **Talk**

Speech data harvesting and factorized deep neural network development for the low-resource indigenous languages of South Africa

Wednesday, 4 December 2019 13:50 (20 minutes)

Speech-enabled human-machine interfaces are becoming increasingly prevalent. There is, however, a disparity between the languages these systems are available in and the languages spoken in South Africa. This is because a tremendous effort is required to collect and refine large speech corpora. For the majority of South Africa's languages, very little speech data is available. This creates a challenge due to the "data hungriness" of automatic speech analysis techniques, specifically those which yield the best performance such as deep learning. Once sufficient initial resources exist, automatic data harvesting strategies can be used to further increase the amount of available data in under-resourced languages. The aim of the work we report on is to improve existing ASR systems for the 11 official languages of the country. These systems are currently based on the recordings of the NCHLT (National Centre for Human Language Technology) corpus, which consists of an estimated 55 hours of speech obtained from 200 speakers for each language. During the NCHLT project additional speech data was collected but not released. To determine whether this additional data is useful, acoustic evaluation of the audio is required to detect both low signal-to-noise ratios (SNRs) and transcription errors. The decoding process with state-of-the-art Deep Neural Network-based models is more than an order of magnitude slower than real time on CHPC's Lengau Dell cluster's processors. It requires more than 10 CPU hours to process one hour of audio data. We therefore used CHPC to run the jobs required for processing one language in parallel. This allowed for approximately 200 hours of auxiliary data to be processed and used less than 50 GB of memory. The harvested data from this process was subsequently used to train factorized time-delay neural networks (TDNN-F). This model architecture was recently shown to yield good results in resource constrained scenarios (less than 100 hours of speech). These models significantly reduced phone error rates for the 11 languages. Each TDNN-F model trained in about 8-12 hours on CHPC's Lengau GPU cluster.

Supported Student

Primary authors: Dr BADENHORST, Jaco (CSIR Next Generation Enterprises and Institutions Cluster); DE WET, Febe (Stellenbosch University)

Presenter: Dr BADENHORST, Jaco (CSIR Next Generation Enterprises and Institutions Cluster)

Session Classification: HPC Applications

Track Classification: Cognitive Computing and Machine Learning