SLTU-2014, St. Petersburg, Russia, 14-16 May 2014

# The NCHLT Speech Corpus of the South African languages

Etienne Barnard[1], Marelie H. Davel[1], Charl van Heerden[1], Febe de Wet[2] and Jaco Badenhorst[2]

1Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa

2Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa

{etienne.barnard, marelie.davel, cvheerden}@gmail.com, {fdwet, jbadenhorst}@csir.co.za

## Abstract

The NCHLT speech corpus contains wide-band speech from approximately 200 speakers per language, in each of the eleven official languages of South Africa. We describe the design and development processes that were undertaken in order to develop the corpus, and report on associated materials such as orthographic transcriptions and pronunciation dictionaries that were released as part of the corpus. In order to benchmark speech recognition performance on the corpus, we have also developed both phone-recognition and word-recognition systems for all eleven languages; we find that high accuracies can be achieved for these speaker-independent but vocabulary-dependent recognition tasks in all languages.