5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Code-switched English Pronunciation Modeling for Swahili Spoken Term Detection

Neil Kleynhans[a,b,*], William Hartman[c], Daniel van Niekerk[a,b], Charl van Heerden[a,b], Rich Schwartz[c], Stavros Tsakalidis[c], Marelie Davel[a,b]

aMultilingual Speech Technologies, North-West University,Vanderbijlpark, South Africa.
bCAIR, CSIR Meraka, South Africa
cRaytheon BBN Technologies, Cambridge, MA 02138, USA.

Abstract

We investigate modeling strategies for English code-switched words as found in a Swahili spoken term detection system. Code switching, where speakers switch language in a conversation, occurs frequently in multilingual environments, and typically deteriorates STD performance. Analysis is performed in the context of the IARPA Babel program which focuses on rapid STD system development for under-resourced languages. Our results show that approaches that specifically target the modeling of code-switched words, significantly improve the detection performance of these words.

Keywords: Spoken term detection; code switching; Swahili, pronunciation modeling