Speech Communication 56 (2014) 119–131

A smartphone-based ASR data collection tool for under-resourced languages

Nic J. de Vries [a,⇑], Marelie H. Davel [b,*], Jaco Badenhorst [a,b], Willem D. Basson [a,b], Febe de Wet [a,c], Etienne Barnard [b], Alta de Waal [a]

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

b Multilingual Speech Technologies, North-West University, Vanderbijlpark 1900, South Africa

c Department of Electrical and Electronic Engineering, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa

Abstract

Acoustic data collection for automatic speech recognition (ASR) purposes is a particularly challenging task when working with underresourced languages, many of which are found in the developing world. We provide a brief overview of related data collection strategies, highlighting some of the salient issues pertaining to collecting ASR data for under-resourced languages. We then describe the development of a smartphone-based data collection tool, Woefzela, which is designed to function in a developing world context. Specifically, this tool is designed to function without any Internet connectivity, while remaining portable and allowing for the collection of multiple sessions in parallel; it also simplifies the data collection process by providing process support to various role players during the data collection process, and performs on-device quality control in order to maximise the use of recording opportunities. The use of the tool is demonstrated as part of a South African data collection project, during which almost 800 hours of ASR data was collected, often in remote, rural areas, and subsequently used to successfully build acoustic models for eleven languages. The on-device quality control mechanism (referred to as QC-on-the-go) is an interesting aspect of the Woefzela tool and we discuss this functionality in more detail. We experiment with different uses of quality control information, and evaluate the impact of these on ASR accuracy. Woefzela was developed for the Android Operating System and is freely available for use on Android smartphones.