Mobile Application for Improving Speech and Text Data Collection Approach

Sarah Samson Juan[1] and Jennifer Fiona Wilfred Busu[1,2]
[1]*Institute of Social Informatics and Technological Innovations, Universiti Malaysia Sarawak, Sarawak, MALAYSIA.*
[2]*Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, MALAYSIA.*
sjsflora@unimas.my

***Abstract*—This paper describes our work in developing a mobile application for collecting language speech and text data. The application is built to assist linguists or researchers in simplifying their tasks in data collection who of native speakers living in remote interiors. Researchers rely on numerous apparatus to carry out their tasks to capture audio or text from far to reach places, but with this mobile application, they would only need to carry one device, which can ease their logistics troubles. The mobile app, named as Kalaka, is designed for users to store details of native speakers, record speech and insert speech transcripts all in one platform. Kalaka is built on the Android platform, which allows data stored in the mobile device to be transferred to a cloud storage using WiFi networks. Usability tests performed in respondents shows, all participants in the evaluation are able to use the application to record their voices and save texts. We also received positive feedbacks on the mobile application from our survey, with more than half of the respondents gave their confidence using Kalaka and they would use the system frequently.**

***Index Terms*—Mobile Application; Data Collection Tools; Corpus Development.**

I. INTRODUCTION

Language documentation is a process of recording linguistic properties which could help in preserving an identity of a language. This process is tedious for linguists and researchers. They may need to travel to rural areas to collect data from native speakers. Most of the time, researchers may use several tools for recording and transcribing speech data such as speech recorder, papers, or computers. However, keeping speech and transcripts in separate tools (recorder and notebook, for example) during a data collection trip has high risks. Researchers could lose valuable data if one or both tools broken or damaged. Moreover, storing or labelling data poorly could cause data loss.

Mobile devices can store audio and digital text data. They are lightweight and easy to carry around when traveling in rural areas. Nowadays, open source operating system such as Android OS, enables developers to create simple applications for low-cost mobile devices. This is cost-effective for researchers who are constantly on the move and need to gather a lot of data from native speakers in rural areas with limited connectivity. Furthermore, with the availability of WiFi and cloud technologies support in mobile platform, developers can build an application for sharing data to cloud storage or real-time database. Hence, researcher can do a backup whenever it is necessary.

Thus, this paper reports our first steps in developing an open source mobile application called *Kalaka*, which can store speech and digital text. Among other functionalities of the application are, store speaker details, edit entries to list of languages, categories or origins, and synchronize data from device to cloud. The latter can be done when a WiFi network is available.

The flow of the paper is as follows. In Section II, we describe our motivation for building creating digital tools to collect speech data and Section III explains the development of *Kalaka* for mobile devices. Subsequently, Section IV presents the interfaces of the system and Section V reports results of our usability evaluation. Finally, Section VI concludes our paper and describes our next steps.

II. COLLECTING DATA FOR RESEARCH IN SPEECH TECHNOLOGY

In speech technology, researchers have developed many methods for building speech applications that could help in human-computer interaction. For example, automatic speech recognition (ASR) systems are used to convert human speech to readable texts. Currently, there are many well-known ASR systems such as Apple's Siri, Microsoft's Cortana and Google's Google Now.

To build ASR, the following data are needed [1-3]:

i) Speech and speech transcripts for acoustic model – typically, a minimum of 20 hours of speech from female and male speakers of each language
ii) Text for language model – a minimum of 500MB of text for training a language model. The language model is used to define the grammar rule and it helps to select the best ASR output.

The presence of Sarawak languages in ASR research is still very low. Juan [4] has published a thesis on exploring resources for building ASR for under-resourced languages in Malaysia. The author's work focused on developing ASR for Iban, a language that is largely spoken in Sarawak. The Iban ASR was built using 7 hours of transcribed speech and text data with 2 million words. Due to inconsistent spellings found in Iban text and very low amount of transcribed speech, the Iban ASR achieved 85% accuracy [5]. This performance is considered low compared to other state-of-the-art ASR systems ([6] - 90-95% accuracy).

Thus, doing research in Speech Technology in Sarawak languages is a challenge as we need to collect large amount or data for building speech applications. There are several related issues such as:

i) Native (original) speakers live in rural areas
   - Travelling cost is high for researchers to meet native speakers
ii) The low amount of electronic text data available
   - Not many digital texts in target language.