SABC2TXT - Mobile Audio Transcription for Low-Resourced South African Languages

Fardoza Tohab
Computer Science
University of Cape Town
thbfar002@myuct.ac.za

ABSTRACT

Computational and statistical models need electronic documents in languages with limited resources and minority languages. The accuracy of statistical machine learning and natural language processing (NLP) models depends on a vast amount of language data. We can gather and analyze vast volumes of data by having digitized documents in low-resourced languages, which may then be used to create and enhance statistical machine learning algorithms and NLP models. In turn, this can enhance our capacity to process these languages digitally and help us better comprehend their distinctive characteristics. One way to produce these electronic resources is by using mobile devices as a means to transcribe audio in these low-resourced languages into text and then use that to support studies in search engines, machine learning, etc. The review of the literature indicated that mobile audio transcription does pose some issues, which included its inability to filter out background noises, inability to transcribe everyday words, colloquial words, and slang, which can alter the accuracy of the data gathered. However, mobile devices are a good stepping stone towards creating electronic documentation for low-resourced languages because mobile device transcription is more accessible, effective, and accurate.

CCS CONCEPTS

• Information systems → Multimedia content creation; • Computing methodologies → Natural language processing; • Computing methodologies → Speech recognition; •Applied computing → Document management and text processing

KEYWORDS

Speech/language, automatic speech recognition, mobile devices, natural language processing, low-resourced languages

1. INTRODUCTION

South Africa's official public broadcaster is called the South African Broadcasting Corporation (SABC). The provision of services in all of South Africa's languages is one of its duties, guided by the public broadcasting service model (PBS) [17]. Since South Africa has eleven official languages, ten of which lack adequate electronic linguistic resources, this is a considerable amount of content [18]. There aren't many articles in most South African languages on Wikipedia, which is bad for computational and statistical systems such as language models that rely on these articles. The development of speech technology is closely related to resource gathering since the statistical models that now dominate Text-to-speech (TTS) and Automated Speech Recognition (ASR) systems depend on the availability of suitable resources for the determination of their parameters [19].

ASR, or Automatic Speech Recognition, is a separate, mechanical method of decoding and transcribing oral speech. A typical ASR system listens to a speaker through a microphone, analyses the audio data using some kind of pattern, model, or algorithm, and then generates an output, typically in the form of text [15, 16]. It's critical to distinguish speech recognition from speech understanding, which is the process of figuring out an utterance's meaning as opposed to transcribing. Speech recognition is distinct from voice recognition in that the former entails a machine's capacity to identify the words that are said (i.e., what is said), whereas the latter requires a machine's capacity to identify speaking style (i.e., who said something) [16]. ASR for low-resourced languages has drawn a lot of interest from the speech research community during the past ten years [20, 21]. According to Berment [2004], a language is considered to be low-resourced if it possesses some (but not all) of the following characteristics: no distinctive writing system, no linguistic knowledge, few web resources, and few electronic resources for speech and language technology. Low-data-density, resource-poor, low-data, and less-resourced languages are all names for the same idea. It is crucial to emphasize that it differs from a minority language, which is a language spoken by a minority of a territory's people [13]. Only a small portion of the world's approximately 7, 000 languages [22] provide the materials needed for the application of human language technologies.

The goal of our project is to see if a computer system can automatically convert audio from SABC broadcasts (which aren't in English or Afrikaans; we could also utilize other audio sources, like YouTube) into text, which we can then use to support studies in search engines, machine learning, etc. To transcribe the material, we will employ publically accessible speech technology tools (through SADILAR), but we also plan to evaluate how practical text segmentation and language identification are.

In this literature review, I will be exploring the use of mobile devices to transcribe audio, which includes uncontrolled audio for example a casual conversation between two people in a noisy environment, and what quality we can get when we do an automatic transcription of uncontrolled audio. Mobile device transcription is important since it can aid in the creation of electronic resources for low-resourced languages, including the majority of South African languages. The percentage of people in sub-Saharan Africa who use mobile phones has increased significantly over the past ten years, reaching 60% of the total population [23]. This means that even if they do not have access to more expensive equipment, speakers of low-resourced languages can capture and transcribe audio using their mobile devices. Overall, because mobile device transcription is more accessible, effective, and accurate, it can be a great tool for developing electronic resources for languages with few resources.

2. NATURAL LANGUAGE PROCESSING (NLP)

The study of how computers can comprehend and use natural language text or speech for beneficial purposes is known as natural language processing, or NLP [14]. NLP researchers work to learn more about how people interpret and utilize language so that the right tools and methods may be created to help computers comprehend and manipulate natural languages in order to carry out the necessary tasks. Computer and information sciences, linguistics, mathematics, electrical and electronic engineering, artificial intelligence and robotics, psychology, and other fields form the basis of NLP. Machine translation, natural language text processing and summarization, user interfaces, multilingual and cross-linguistic information retrieval (CLIR), speech recognition, artificial intelligence, and expert systems, among other fields of study, are a few examples of applications of NLP [14].

For academics utilizing computer methods to analyse and study languages, language resources are crucial. These tools are required to develop text analysis, machine learning, information retrieval, and natural language processing (NLP) [24]. Many NLP technologies have been created by researchers to automatically analyse, parse, and annotate various languages. In these activities, language resources serve two purposes. The first is the use of extensively annotated corpora to power statistical NLP algorithms. The second is the requirement for test collections in order to evaluate results in comparison to a gold standard. With the help of initiatives like the Linguistic Resources and Evaluation (LRE) Map, such NLP resources are documented [25]. But, there aren't many of these resources for some languages like the majority of South African languages. In our situation, the ability to transcribe audio in South African languages without gold-standard resources lead us to investigate methods for developing them. One method investigated in this literature review is the use of mobile audio transcription to produce electronic resources for low-resource South African languages.

3. MOBILE DEVICES FOR SPEECH DATA COLLECTION AND ASSISTIVE TECHNOLOGY

There are numerous studies that have been conducted on how mobile devices can aid in the creation of electronic resources and documents, which can then be used to support studies in search engines, machine learning, and other language models. Some of the studies include the use of mobile devices as all-purpose recorders to collect speech data for automatic speech recognition (ASR) systems, how deaf or hard-of-hearing people can use transcription on mobile devices as a form of assistive technology, speech recognition toolkits, language documentation, and speech corpora creation using mobile devices.

3.1 Data collection using mobile devices for ASR

Using mobile devices as all-purpose recording devices to gather speech data is a relatively interesting data collection technique. Seeing mobile devices as all-purpose recording devices is drastically altering the way ASR data collection is done, especially for languages with limited resources. Mobile device availability and cost are falling fast, even in developing countries. With the advent of these mobile computers, it is now possible to dynamically pick and present previously assembled prompts to be recorded on the mobile screen, record the actual speech utterances at the broadband quality (usually to the SD card), and collect metadata related to each prompt [2].

This trend was led by a smartphone application introduced by Hughes et al. in 2010, which is a technique for quickly and simply assembling speech corpora that have been reliably transcribed and contain utterances from several speakers in varied acoustic environments. The system consists of an Android mobile device running a client application connected to a server over an unstable Internet connection. The speaker's demographic information is gathered by the client application, which also sends the speaker text prompts from the server to read, records the speaker's voice, and uploads the audio file and any relevant metadata to the server. The method has so far been used to collect around 3000 hours of audio that have been transcribed in 17 different languages. However, the device does not compensate for speaker errors such as speech mistakes, repeats, unnecessary comments, and UI mistakes. Hence, the prompt phrase can only roughly represent the actual transcript [5].

A similar approach was taken by a smart-phone based solution that enables speakers to record prompted speech directly on their phones, which was introduced by Lane et al. [2010]. Three steps make up the speech data collection process: obtaining gender, age, and language information to identify individual speakers; gathering speech data offline by asking users to read sentences from an on-board database while holding down a push-to-talk button; and uploading the speech data one utterance at a time to an FTP server, allowing for pausing and restarting of the upload process. During the collection stage, users can also listen to older recordings and record them again. Lack of training, however, resulted in a much-decreased output of high-quality recordings [8].

One possible drawback of the above approaches is the fact that read speech (speech or presentation that has been written out ahead of time and then read aloud by the speaker) is the only speaking style suitable for such a collection. Nonetheless, it has been discovered that these data are particularly valuable as the initial and final application corpora for the development of ASR systems [2].

Reitmaier et al. [2022] developed the Voice Notes Android app, which can process audio data from other apps like WhatsApp. When users long-press on a voice message and click the share icon in a WhatsApp group or conversation, they have the option to select the Voice Notes app from a list of Share Targets that normally includes other well-known apps like Email, SHAREit, Bluetooth, etc. The program creates a duplicate of the voice message's content and sends it to the cloud service. The cloud service then creates an asynchronous transcription request utilizing the English (South African) language model using the Google Cloud Speech-to-Text service. The app is then provided with the transcription results. However, the ASR probe has certain drawbacks, such as the fact that only the most frequently used words are accurately transcribed and that some less common phrases are not included in the app's list of words. Comparing short instructions to a longer voice message that included a recitation from a poem that was unsuccessfully transcribed, the transcriptions were of higher quality for the shorter instructions. The metadata that is lost when transferring a voice message between WhatsApp and the Voice Notes app cannot be recovered, either [6].

Inspired by the Google approach, an open-source program called Woefzela was created with the intention of gathering speech data for low-resourced languages in the developing world. The platform was made specifically for gathering automated speech recognition (ASR) data, which is a crucial stage in developing ASR systems. Users of the Woefzela platform, introduced by de Vries et al. [2014], can record their speech and upload it to a central server via a smartphone app. Many features are provided by the application, such as automatic speech segmentation, tagging, and recording quality checks. Woefzela does provide some basic quality control checks, such as recording volume and background noise detection, however, the quality of the acquired data can have certain restrictions. This could have an effect on the accuracy of ASR systems developed using these data [3]. Although the recording setting was carefully regulated, responder errors and unintentional and/or uncontrolled ambient noise nevertheless posed a significant barrier. Simple quality checks do not catch all incorrect recordings, such as background speech, transcription mistakes, or stuttering. More sophisticated quality checks based on utterance length estimates were introduced by Badenhorst et al. [2012]. These checks target transcription mismatches to find additional incorrect recordings that were missed by the initial on-device validation phase. The fundamental quality criteria were proven to be helpful, and the extra criteria that were suggested appeared to have promising potential [4].

3.2 Using mobile device transcription as assistive technology

Individuals who are deaf or hard of hearing may have trouble understanding spoken words from others and may struggle to be aware of auditory events in their surroundings. This is particularly true in public settings, when there might not be readily available methods of disseminating announcements and other audio events [1]. Some of the approaches used to help improve this are discussed below.

In order to help the deaf and hard of hearing understand spoken conversations, Matthews et al. [2006] designed Scribe4Me, a mobile sound transcription tool. Using speech recognition software, the gadget converts spoken words into text that can be seen in real-time on the user's mobile device. A study was done on a group of deaf and hard-of-hearing people to determine the efficacy of Scribe4Me. 6 participants, ranging in age from 25 to 51, utilizing the tool in the study in a range of contexts, including one-on-one interactions, group discussions, and lectures. The system's real-time transcribing feature allows those who are deaf or hard of hearing to follow discussions as they are happening. Although the tool's ASR technology attempted to accurately translate spoken words into text, it occasionally failed. As a result, there was uncertainty and misunderstanding between the speaker and the deaf user. The tool had trouble eliminating background noise, which made it difficult for the ASR technology to reliably record spoken phrases. In noisy settings, such as packed public venues, this proved particularly challenging. The technology also struggled to distinguish between various voices and accents, which further reduced the transcription's accuracy. However, Scribe4Me may have trouble accurately transcribing some accents, dialects, or speech patterns even if it can record spoken language in real time. When numerous persons are speaking at once or in noisy locations, some Scribe4Me users have complained that the tool's accuracy can be unpredictable. Also, since slang and colloquial words are not part of standard English, the tool can have trouble translating them [1].

Another system introduced by Agbeyangi et al. [2019] is a speech-to-text system that can listen to human speech using a mobile phone's microphone, convert it to text using an Arduino program and a hardware display system through Bluetooth, and then show the text of the speech on an LCD. The system was designed to aid deaf people to better partake in conversations. However, the system does not cater to mistakes made during pronunciations and only works in a room with little background noise [7].

These findings are particularly important since they highlight the difficulty of transcribing conversations in uncontrolled environments and the difficulty of transcribing conversations that contain different accents, dialects, and colloquial words. It is particularly relevant in understanding these issues and overcoming them when conducting our experiments.

3.3 Using mobile devices to create speech corpora and language documentation

Bird et al. [2014] introduced Aikuma, a mobile app that is designed to put the key language documentation tasks of respeaking, translating, and recording in the hands of a speech community. Aikuma is a smartphone application made for group language documentation. Aikuma is a useful resource for linguists, language students, and anybody else interested in the documentation and

preservation of languages. It is the perfect option for community-based language documentation projects due to its simplicity of use and collaborative capabilities. To build a searchable collection of linguistic information, the app lets users record and translate conversations in several languages. The program also has tools for documenting and reviving languages, like the capacity to make courses and distribute recordings to other users. Aikuma's capacity to support collaborative language documentation is one of its distinctive qualities. Users have the option to invite other people to take part in the documentation process, which enables a wider variety of voices to be heard and captured. Support for low-resource languages is another standout aspect of Aikuma. The app has built-in support for several minority and endangered languages and is intended to operate offline, without an internet connection. Based on automatic voice recognition technology, Aikuma's transcription and annotation functions might not be accurate for all languages and dialects. It can be necessary for users to manually edit or add to transcriptions, which can take time [11].

Gauthier et al. [2016] further reported on the ongoing efforts to collect speech data in under-resourced or endangered languages of Africa and introduced an improved version of the Android application Aikuma - developed by Steven Bird and colleagues [11]. The app now has features that make it easier to collect parallel speech data following the guidelines of the French-German ANR/DFG BULB (Breaking the Unwritten Language Barrier) project. The end product is an app called LIG-AIKUMA that works on a variety of smartphones and tablets and offers a variety of voice-gathering modes (recording, respeaking, translation, and elicitation). Almost 80 hours of speech were recorded using it for field data collection in Congo-Brazzaville [12].

3.4 Speech recognition toolkits

Lakdawala et al. [2018] presented an offline voice-to-text transcription system for healthcare companies. It can be used by counselors and non-governmental groups to capture talks during surveys, convert them to text, and then save the messages. This system includes an open-source application. The CMUSphinx toolkit is utilized for speech recognition. The system can recognize multiple languages. The language model, phonetic dictionary, and acoustic model are all utilized by the CMUSphinx toolset. The user captures their voice using the mobile application, and the CMUSphinx toolkit analyzes and transcribes it. The transcription file will be saved as a text file in the device's memory, from which the user can use the application to upload and download data to and from the database server. Although the CMU Sphinx speech recognition toolset is dependable and accurate, mistakes can nevertheless happen. Background noise, speaker accents, and speech speed, among other things, can affect how accurately a transcript is made [9].

4. DISCUSSION

Past studies have shown promising results when employing mobile devices for ASR data collection, assistive technology, and speech documentation. The use of mobile cellphones as all-purpose recording devices to capture speech data is a very intriguing data collecting technique that has fundamentally altered the way ASR data collection is done, especially for languages with limited resources. Mobile devices are becoming more widely available and affordable, even in developing countries, making it possible to dynamically select and present previously assembled prompts to be recorded on the mobile screen, record the actual speech utterances at broadband quality, and collect metadata related to each prompt. Nevertheless, this approach is limited to read speech and could result in inaccurate recordings.

In a similar manner, past studies in the field of assistive technology have shown that persons who are deaf or hard of hearing may struggle to understand spoken words from others and may struggle to be aware of auditory occurrences in their surroundings. Previous research has led to the development of mobile sound transcription applications, such as Scribe4Me, to solve this issue by converting spoken words into text that can be seen in real-time on a user's mobile device through the use of speech recognition technology. Nevertheless, these techniques don't always accurately remove background noise, distinguish between various voices and accents, or translate certain accents, dialects, or speech patterns.

Moreover, smartphone apps like Aikuma have succeeded in language documentation by leaving the key tasks of respeaking, translating, and recording to a speech community. Aikuma can be useful for linguists, language students, and anybody else interested in the documentation and preservation of languages. Users may record and translate conversations in a number of languages using the program, which also offers tools for documenting and reviving languages. These capabilities include the ability to develop courses and exchange recordings with other users. Nevertheless, because automatic transcription and annotation features might not be precise for all languages and dialects, users may need to manually edit or add to transcriptions, which might take time.

Despite these drawbacks, prior research has paved the way for additional investigation into these areas and highlights how critical it is to recognize the difficulties associated with transcription of conversations in natural settings and the requirement to thoroughly examine them. Future work should focus on developing more sophisticated quality control procedures, improving transcription accuracy for different accents, dialects, and colloquial phrases, and devising workarounds for read speech data collection issues. These initiatives will provide more accurate and effective ASR systems, assistive technologies, and language documentation tools that can benefit a wider range of users.

5. SUMMARY

In summary, mobile devices have revolutionized the way speech data is gathered and transcribed. They have made it possible for researchers to swiftly, affordably, and dynamically capture speech data, particularly in low-resource languages. In order to preserve endangered languages, mobile devices have also been utilized to build speech corpora and linguistic documentation. However, there are still challenges to overcome, such as different accents or dialects, transcription errors due to noisy environments, and the use

of colloquial language. Although some of these issues have been addressed by speech recognition toolkits like CMUSphinx, additional work has to be done. Assistive technology for the deaf and hard of hearing can be developed on mobile devices, as well as speech corpora and language documentation. The collection of data is a key step in the development of speech technology, and mobile audio transcription can be a convenient and efficient way to obtain linguistic information for the statistical models that underpin automatic speech recognition and natural language processing systems.

REFERENCES

[1] Matthews, T., Carter, S., Pai, C., Fong, J. and Mankoff, J., 2006. Scribe4Me: Evaluating a mobile sound transcription tool for the deaf. In UbiComp 2006: Ubiquitous Computing: 8th International Conference, UbiComp 2006 Orange County, CA, USA, September 17-21, 2006 Proceedings 8 (pp. 159-176). Springer Berlin Heidelberg

[2] Nic J. de Vries, Marelie H. Davel, Jaco Badenhorst, Willem D. Basson, Febe de Wet, Etienne Barnard, and Alta de Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. Speech Communication 56, (January 2014), 119–131. DOI:https://doi.org/10.1016/j.specom.2013.07.001

[3] De Vries, NJ, J Badenhorst, MH Davel, E Barnard, and De Waal, A. 2017. Woefzela - An open-source platform for ASR data collection in the developing world. Csir.co.za (2017). DOI:http://hdl.handle.net/10204/5149

[4] Jaco Badenhorst. Quality Measurements for Mobile Data Collection in the Developing World. Retrieved March 9, 2023, from https://www.isca-speech.org/archive_v0/sltu_2012/papers/su12_139.pdf

[5] Thad Hughes, Kaisuke Nakajima, Linne Ha, Atul Vasu, Pedro Moreno, and Mike Lebeau. Building transcribed speech corpora quickly and cheaply for many languages. Retrieved March 11, 2023, from https://storage.googleapis.com/pub-tools-public-publication-data/pdf/36801.pdf

[6] Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. CHI Conference on Human Factors in Computing Systems (April 2022). DOI:https://doi.org/10.1145/3491102.3517639

[7] Abayomi O Agbeyangi and Adam B Olorunlomerue. 2019. A Smartphone-Based Multi-Functional Speech-To-Text Transcription System. ResearchGate. Retrieved March 11, 2023, from https://www.researchgate.net/publication/333507776_A_Smartphone-Based_Multi-Functional_Speech-To-Text_Transcription_System

[8] Ian Lane, Alex Waibel, Matthias Eck, and Kay Rottmann. 2010. Tools for Collecting Speech Corpora via Mechanical Turk. Association for Computational Linguistics. Retrieved March 11, 2023, from https://aclanthology.org/W10-0729.pdf

[9] Burhanuddin Lakdawala, Farhan Khan, Arif Khan, Yash Tomar, Rahul Gupta, and Ashfaq Shaikh. 2018. Voice to Text transcription using CMU Sphinx A mobile application for healthcare organizations. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (April 2018). DOI:https://doi.org/10.1109/icicct.2018.8473305

[10] Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. LIG-AIKUMA: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies. Hal.science (2016). DOI:https://hal.science/hal-01350062

[11] Steven Bird, Florian Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A Mobile App for Collaborative Language Documentation. Association for Computational Linguistics. Retrieved March 12, 2023, from https://aclanthology.org/W14-2201.pdf

[12] Vincent Berment, Methods to computerize “little equipped” languages and groups of languages, Theses, Universite Joseph-Fourier - Grenoble I, May 2004, ´ https://tel.archives-ouvertes.fr/tel-00006313.

[13] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. Speech Communication 56, (January 2014), 85–100. DOI:https://doi.org/10.1016/j.specom.2013.07.008

[14] Gobinda G. Chowdhury. 2005. Natural language processing. Annual Review of Information Science and Technology 37, 1 (January 2005), 51–89. DOI:https://doi.org/10.1002/aris.1440370103

[15] Lai, J., Karat, C.-M., & Yankelovich, N. 2008. Conversational speech interfaces and technologies. In A. Sears & J. A. Jacko (Eds.), The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications (2nd ed., pp. 381–91). New York, NY: Erlbaum

[16] Levis, J. and Suvorov, R., 2012. Automatic speech recognition. The encyclopedia of applied linguistics

[17] Gawie Botma. 2022. Representation of official languages on South African Broadcasting Corporation (SABC) television: A study of selected Tshivenda programmes By MPHO GIVEN RATHANDO. Retrieved March 22, 2023, from https://scholar.sun.ac.za/bitstream/handle/10019.1/126093/rathando_languages_2022.pdf?sequence=1

[18] Roald Eiselen and Martin Puttkammer. Developing Text Resources for Ten South African Languages. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1a7715e8d41a228f8fc35f7e3f2580988eea9b24

[19] Etienne Barnard, Davel, Marelie H, van Heerden, De Wet, Febe, and Jaco Badenhorst. 2014. The NCHLT Speech Corpus of the South African languages. Nwu.ac.za (2014). DOI:https://researchspace.csir.co.za/dspace/handle/10204/7549

[20] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, “Automatic speech recognition for underresourced languages: A survey,” Speech Commun., vol. 56, pp. 85–100, Jan. 2014.

[21] Mark J. F. Gales, Kate M. Knill, Anton Ragni, and Shakti P. Rath, “Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED,” in 4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014, 2014, pp. 16–23.

[22] “Ethnologue, Languages of the World”, https://www.ethnologue.com.

[23] Jenny C Aker and Isaac M Mbiti. 2010. Mobile Phones and Economic Development in Africa. Journal of Economic Perspectives 24, 3 (August 2010), 207–232. DOI:https://doi.org/10.1257/jep.24.3.207

[24] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2014. Creating language resources for under-resourced languages: methodologies, and experiments with ,Arabic. Language Resources and Evaluation 49, 3 (August 2014), 549–580. DOI:https://doi.org/10.1007/s10579-014-9274-3

[25] Calzolari, N., Soria, C., Gratta, R.D., Goggi, S., V.Q., Russo, I., Choukri, K., Mariani, J., & Piperidis, S. (2010). The LREC 2010 resource map. In The 7th international language resources and evaluation conference (LREC 2010), LREC 2010, Valletta, Malta, pp. 949–956.