ISSN No:-2456-2165

Survey on Speech to Text Modelling for the Shona Language

B Mupini[1,] S Chaputsira[1], Bk Sibanda[2]
Computer Science Department, School of Information Science and Technology,
Harare Institute of Technology, Harare, Zimbabwe[1]
Independent Researcher[2]

Abstract:- Conversion of speech to text (STT) for various applications is of huge interest, which involves technological approaches which are innovative that should be applied to accommodate spoken languages in Africa. However, African countries are falling behind on the embracing of STT technologies, with Automatic Speech Recognition (ASR) having been done for popular East African languages. This has always kept transcription at a minimum and has also resulted in a retard in the use of many African languages on a world-wide scale, with another problem being that a single African language may encompass several dialects. This research looks at modern technologies and models that have been implemented to construct ASR and STT models for African languages and existing datasets, with particular interest to the Shona language spoken by the people of Zimbabwe. A survey has been done on STT for the Shona language and it uncovers techniques present which can be used to achieve effective STT for this language. An example of such a technique is accounting for procedures taken to convert spoken words into actual text that can be displayed. The usage of ASR techniques can help in many application areas such as assisting individuals with hearing impairment, transcription services, use in voice commands and control, dictation and notes taking, language learning and translation, customer service and support and also voice search and content indexing. ASR is dominating together with other technologies such as STT conversion, Text to Speech (TTS) conversion and language translation. Cumulatively, these technologies have aided in bridging the gap between people who speak different languages especially tourists and language enthusiasts. In African countries most of which are underdeveloped, many spoken African languages are underrepresented and lowly resourced, which has hampered the advancement of ASR technology on these low resource languages. Bridging this gap will result in African languages, especially Shona, being recognized more in the world and finding use in everyday applications and technologies.

***Keywords*:-** *Transcribe, Dataset, Models, Dialect, Conversion.*

I. INTRODUCTION

Speech to text (STT) technology, also called speech recognition or automatic speech recognition (ASR) is basically synthesizing speech and producing text as output. There is a myriad of application areas for this technology especially for the Shona language including driving voice assistants which use machine learning (ML) and other artificial intelligence (AI) techniques to automatically interpret and understand spoken languages[1]. STT technology is imperative for improved language use and development especially for a developing country. However, many African languages are low resourced, meaning that there has not been much ASR work done for transcribing these languages. The major reason for this is basically the lack of resources such as readily prepared and cleaned datasets, baselines on which to build a model from and published work that is already in existence for particular languages. Moreover, African languages such as Shona have several dialects that are encompassed within the Shona language itself that sound similar in tone but with variations in word use and derived word meanings from one dialect to another. In contrast to other well-resourced languages such as English, Japanese and Korean just to mention but a few, there is significant work that has been done including research, model building and applications for ASR, some of which can be accessed online. Under developed countries lag because of scarcity of funds as well as resources to conduct research, build and clean datasets and eventually implement working models. ASR technologies can be expensive to construct, but with better funding, African languages may be better represented as well.

Instead of focusing on the many dialects of a particular language, it may be pertinent to consider the dialects in general, and consider the language as a whole. Shona for instance has several dialects including Karanga, Zezuru, Ndau just to mention a few. Speech-to-text technology has been primarily developed and trained on widely spoken languages, often those with larger user bases and wide-ranging linguistic resources. ASR has proved to be more successful on highly resourced languages (HRL) equated to lowly resourced languages (LRL) which are underrepresented[1]. The study of Shona phonotactics is essential to guide the making of Shona Text-To-Speech (TTS) synthesized voices for use on e-infrastructure[2].

It is important to note that the availability and quality of STT technology for African languages is still limited compared to widely used languages like English. It is also worth noting that existing work on using speech directly typically relies on already available off-the-shelf STT systems [3]. However, researchers, organizations and communities are working towards addressing this gap and promoting the development of speech-to-text technologies for African languages in-cooperating these in learning devices for children as well. Even so, research has exposed that children who have had a fascinating mode of learning a certain concept will hold it in memory and use it better than those who have been pressured to learn it[4].

This paper offers an intuition into diverse methods that can be used in the development of automatic speech recognition and speech to text conversion for the Shona language. The paper also looks at some of the work that has been done on the Shona language and how the work has contributed to building a corpus for this low resourced language.

- *Problems with STT and ASR on African Languages*

There has been some extensive work done by other authors to cater for the African languages. Some authors postulate that currently, approximately 31 out of about 2 000 African languages that re known have been covered in present language models[5]. The authors went on to develop SERENGETI[5], which is a hugely multilingual language model that covers 517 African languages and their varieties. This language model also covers the Shona language as well.

Table 1 Encoder-Only Models with African Languages Represented [5]

Language Model	**African Languages Represented**
MBERT	Afrikaans, Malagasy, Swahili, Yoruba
XLM-R	Afrikaans, Amharic, Hausa, Oromo, Somali, Swahili, Xhosa.
KinyarBERT	Kinyarwanda
AfriBERTA	Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya and Yoruba
Afro-XLMR	Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Nigerian Pidgin, Kinyarwanda, Kirundi, **Shona**, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu
AfroLM	Amharic, Afaan Oromoo, Bambara, Ghomala, Ewe, Fon, Hausa, Igbo, Kinyarwanda, Lingala, Luganada, Luo, Moore, Chewa, Nigerian Pidgin, Shona, Swahili, Setswana, Akan Twi, Wolof, Xhosa, Yoruba, IsiZulu
SERENGETI	Includes 517 African languages.

There have been a lot of problems that researchers have encountered in their attempts to make STT models and transcription better and more accurate. Some of the issues are described below.

- *Background Noise*

This is the noise that can be heard together with the speech recording, which can distort or overshadow the actual recording. Some authors discern that the consequence of background noise is shown by testing the models under test with noisy data to observe the bearing of training on noisy data[6]. Several ways can be employed to combat background noise. One of the ways is to be sure that there is no background noise when recording audio. This can be achieved by using a noise-cancelling headset if one is working at home or at the office. Another method is increasing the computer's recording volume so that the microphone picks up only what is being said and nothing else.

- *Different Accents*

A person's accent can make it hard to make out the words that are being said, and this can also make it tedious to grasp the meaning of the utterances. The linguistic inconsistency of accents presents tough challenges for ASR and STT systems in both data collection and modelling strategies[1]. This is because language is comprehended by listening to context as well as understanding individual words. Most speech recognition software is not good at distinguishing between different voices. One way of combatting strong accents is by building accent-specific models, but this is limited because of the low availability of data for most accents which are mostly not well sourced[7].

- *Numerous Voices in a Single Recording*

Noise in the background affects speech-to-text accuracy largely due to the quality of recording apparatus used. A single speaker gives more accurate results. If there are multiple voices in a recording, for example a conversation, then each voice should be separated from the other using a physical barrier. Another sure way is to use multiple microphones, and thus instructing the participants into which microphone they should speak.

- *Clipping and Distortion*

The effectiveness of speech-to-text or voice-to-text conversion can be greatly reduced by clipping and distortion of the recording. Clipping occurs when there is too much audio all at once in a recording. This usually occurs when the audio is too soft. These issues can make it difficult for a speech-to-text program to understand what is being said, and there may be need to adjust the settings of recording devices or software.

- *Use of Technical Words/Jargon*

This occurs relentlessly in documents which are technical such as instruction manuals. Most African languages, including Shona, do not have native words for most of the technical words and phrases since these are usually borrowed from other languages. Language of this type can be difficult to comprehend for transcribers, particularly if it is not used consistently throughout the entire document. A transcriber may have to carry out additional research or pay particular attention to contextual clues in order to acquaint themselves with the technical words so that they may fully understand what is being said.

➢ *Why Speech to Text Transcription for the Shona Language?*

It is important to note that the Shona language is native to Zimbabwe only. It is largely a Bantu language spoken by the Shona people of Zimbabwe. The Bantu family contains the largest number of languages spoken in Africa[8], with Shona being one of them with its many dialects. The name Shona describes jointly all the central Shona diversities comprising of Zezuru, Manyika, Korekore and Karanga[9]. The Shona language is in turn closely related to Ndau which is Eastern Shona and Kalanga which is Western Shona[9].

Shona is regarded as a low resource language, with not much work or publications available that discuss its prevalence in STT and ASR models. There exist very few published speech datasets for the Shona language. The main troubling issues are the unavailability of cleaned ready datasets for Shona and the lack of pretrained models to build on the Shona language. In Africa, low resource languages are not being met in terms or funding to encourage researchers to build datasets and construct speech models. ASR models can prove to be invaluable especially for use in business transactions and learning institutions. However, due to the underrepresentation of these languages like Shona on public online platforms for use, the LRL remain underutilized and frowned upon, up to an extent that they are not used in official areas such as court proceedings and drafting of business contracts.

The current available research on ASR and STT models for Shona is outdated and do not speak to present transcription problems and needs. In this light, there are no real public baselines to build upon or compare with, hence research in LRL is novel and al results obtained are aiding in building comparative baselines. So basically, for researchers when it comes to the Shona language, there are no starting points to conduct effective research. Some of the models being used in conjunction with the few publicly available speech datasets for Shona are multilingual, whereas to be able to construct and train a viable model, Shona needs to be in cooperated as a monolingual language, to better train the model.

➢ *How to Improve ASR and STT for Shona Language*

Numerous papers have been issued on the subject of speech to text transcription of African languages. Diverse techniques have been looked at with regards to STT and some techniques have resulted in models being built. The suggested methods for improvement discussed below have been employed on most African languages and therefore are expected to work just as well for Shona

There is an issue that most STT and ASR models tend to incorrectly transcribe names of entities. There is need to investigate the performance of state-of-the-art (SOTA)[1] ASR models on named African entities such as people, places etc.[1]. This may be achieved by designing a comprehensive and effective strategy to evaluate STT and ASR models on Shona speech datasets that do not have prior Named Entity Recognition (NER) annotations[1]. The fine-tuned datasets and ASR models can then be made opensource so that other researchers may benefit from the work.

There is general sticky issue is that there are no publicly available Shona speech datasets that researchers can make use of. Platforms such as Mozilla Common Voice presently have not covered the Shona language. The only available dataset is available on Google Fleurs together with a myriad of other languages. [8]cited that the issue of data availability has been the major push factor that has seen researchers either collecting their own data to create viable datasets, or making use of available alternative data sources. Radio archives are a great source of speech data.

Some authors suggest making use of Deep Learning (DL) models for employment in the area of ASR[10]. Authors also suggested implementing large-scale training datasets paired with resources that have high computation and storage[10]. According to the author, there is no work present that looks into detail and reviewing critically Deep Transfer Learning models on ASR. This could be another approach that may be used to research on ASR models for Shona as well. ASR and STT models also encompass aspects such as acoustic information, lexical information, syntactic information and semantic knowledge[10]. Processing of an acoustic model (AM) comprises speech coding, speech enhancement, source separation, speech security such as steganography and watermarking, and other know-hows that are made use of in audio analysis[10].

The general procedure for building a voice for a new language was well explained in[11]. The authors were making a corpus for Akan, a spoken language in Ghana with many dialects (12 in total) like Shona. The proposed process includes the following five steps; definition of a phone set, setting Letter-to-Sound or Grapheme-to-Phoneme rules, creation of a phonetic lexicon, development of a text analysis module including text normalization and other processing steps, and building a waveform synthesizer[11]. The authors went on to site that the lack of various data and linguistic resources in under-resourced and/or under-described languages, may cause difficulties at each step[11]. The same similar steps, with adequate resources, may well be employed for the Shona corpus as well.

In another paper, the authors researched on how vowel length contrast affected the modelling of ASR models. Their main focus was on two languages, Hausa and Wolof[12]. Shona also has variations in the length of vowels when words are said. The authors characterised that two versions of short or long vowel length of the same vowel persist in the phoneme list of the languages[12], the same can be said for Shona. Their common feature is to appear with vowel length

II. SUMMARY OF RECOMMENDED IMPROVEMENTS

The table below summarises the different techniques and technologies that can be employed in order to improve STT and ASR models, with particular emphasis on the Shona language. The techniques have been discussed in the literature and the recommendations also encompass the general perspective of author.

Table 2 General Recommendations for Improving STT for Shona

ISSUE	RECOMMENDATION
Incorrect transcription of Shona named entities.	Investigate performance of SOTA in named Shona entities. Create and fine-tune datasets and ASR models. Make ASR models and fine-tuned datasets public.
Unavailability of datasets.	Create datasets. Collect speech data from radio archive sources.
Inadequate resources.	Make use of big scale training datasets. Make use of high computational resources (online). Make use of large storage resources (online).
Poor model build quality.	Look into methods of improving models such as. • speech coding. • speech enhancement. • source separation. • speech security.
Difficulty classifying vowels.	Consider building the ASR system considering different vowel lengths for words in the language.

III. CONCLUSION

Numerous researchers are attempting to combine and amalgamate quite a lot of technologies to improve transcription. These include tools such as fine tuning, training larger datasets and extracting more features. This has been necessary since this seeks to improve training results and reduce the WER during transcription. However, increasing the number of features to extract is good but the overall training time may increase, which can be a drawback especially with lower specification GPUs. Picking a GPU with larger memory and overall performance can be a decent option to contemplate especially for large speech datasets. It is important to note that the first step is for African researchers, including those for Shona, is to build more speech datasets for the native languages. Most of the training has been based multilingual models, whereas there is need to focus on training with monolingual models to reduce the overall WER and improve the efficiency of the models.

REFERENCES

[1]. T. Olatunji et al., 'AFRINAMES: MOST ASR MODELS "BUTCHER" AFRICAN NAMES'. [Online]. Available: https://huggingface.co/datasets/tobiolatunji/afrispeech-200

[2]. P.-J. Farrugia and M. Rosner, 'Text to Speech Technologies for Mobile Telephony Services', 2005.

[3]. G. Y. Kebe, L. E. Richards, E. Raff, F. Ferraro, and C. Matuszek, 'Bridging the Gap: Using Deep Acoustic Representations to Learn Grounded Language from Percepts and Raw Speech', Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.13758

[4]. I. Mutamiri, G. Brooking, and V. Mugari, 'A Phonetic Transcription System for use in Shona Text-To-Speech'.

[5]. I. Adebara, ⋆ Abdelrahim Elmadany, M. Abdul-Mageed, and A. Alcoba, 'SERENGETI: Massively Multilingual Language Models for Africa'. [Online]. Available: https://github.com/UBC-NLP/serengeti

[6]. A. Pervaiz et al., 'Incorporating noise robustness in speech command recognition by noise augmentation of training data', Sensors (Switzerland), vol. 20, no. 8, Apr. 2020, doi: 10.3390/s20082326.

[7]. M. Masson and J. Carson-Berndsen, 'Investigating Phoneme Similarity with Artificially Accented Speech', 2023.

[8]. Y. Chanie, M. Elamin, P. Ewuzie, and S. Rutunda, 'MULTILINGUAL AUTOMATIC SPEECH RECOGNITION FOR KINYARWANDA, SWAHILI, AND LUGANDA: AD-VANCING ASR IN SELECT EAST AFRICAN LAN-GUAGES'.

[9]. R. K. Shosted, 'Just put your lips together and blow? The whistled fricatives of Southern Bantu'.

[10]. H. Kheddar, Y. Himeur, S. Al-Maadeed, A. Amira, and F. Bensaali, 'Deep Transfer Learning for Automatic Speech Recognition: Towards Better Generalization', Apr. 2023, doi: 10.1016/j.knosys.2023.110851.

[11]. H. S. Cho et al., 'A phone set of Asante-Twi defined in IPA and X-SAMPA A language resource for the first text-to-speech system of Akan'. [Online]. Available: http://streetlibraries.org/

[12]. E. Gauthier, L. Besacier, and S. Voisin, 'Automatic Speech Recognition for African Languages with Vowel Length Contrast', in Procedia Computer Science, Elsevier B.V., 2016, pp. 136–143. doi: 10.1016/j.procs.2016.04.041.