Building TTS systems for low resource languages under resource constraintss

Perez Ogayo[†]*, Graham Neubig*[†‡]*, Alan W Black*[†]

[†]Language Technologies Institute, Carnegie Mellon University
[‡]Inspired Cognition
Pittsburgh, PA, USA

{aogayo, gneubig, awb}@cs.cmu.edu

Abstract

The field of speech synthesis has advanced to remarkable levels of producing natural-sounding speech given sufficient high-quality data. As a result, speech synthesis applications are increasingly becoming ubiquitous for high resource languages. However, support for low resource languages is limited by the lack of data. This project aims to democratize text-to-speech systems and datasets for African languages. Through a participatory approach, we curate data from existing "found" sources and record datasets using more affordable equipment. We build Flite-based voices that can be easily deployed to mobile phones and require less expensive compute to train so that the work can be accessible. We release the speech data, code, and trained voices for 16 African languages to help researchers and developers. In addition, through our website users can interact with the synthesizers and provide feedback for iterative improvement of the synthesizers. Finally, we show that we can develop synthesizers that generate intelligible speech with 25 minutes of created speech, even when recorded in suboptimal environments.

Index Terms: Speech Synthesis, Text-to-Speech, African Languages, Language Resources

1. Introduction

Text to Speech (TTS) technologies for low-resource languages lag far behind as current TTS techniques mostly require high-quality single-speaker recordings with text transcription for at least 2 hours of speech [1, 2]. For many languages in developing countries, the audio and text transcripts necessary to produce a deployable TTS engine are difficult to obtain. Data collection is not just a matter of finding and recording speakers but is influenced by literacy levels and the availability of recording devices, which are influenced by the level of development in the regions where the language is spoken. In contrast to high resource languages with lots of text and audio resources, this type of high-quality speech synthesis data is both readily available and relatively easy to create when it does not already exist.

In this paper, we focus on African languages, which tend to lack high-quality speech synthesis data despite often having a relatively large number of speakers, This state of resource scarcity across the continent is not limited to speech synthesis but extends across the entire field of language technologies [3, 4]. This is arguably because many languages are overlooked due to greater economic incentives for other languages and the lack of researchers and technologists from Africa in industry and academia [5]. We extend the work presented in [6] to include more languages.

We describe the development of an initiative, AfricanVoices, that attempts to change this status quo through a participatory methodology to create and curate single speaker speech synthesis datasets for African languages. The following principles guide us:

Accessibility: An important consideration when creating TTS engines for low resource languages is the ease of access for those who need it. To this end, we make all developed data and trained speech synthesizers publicly available under easy-to-use licenses. We also focus on underlying technology that is easy to train and deploy in low-resource environments, such as low-powered Android smart phones. To do so, we build on top of the long-standing FestVox project [7] to build deployable TTS engines. Our generated TTS models use the CMU Flite [8] framework, using the random forest based statistical synthesizer [9] for voices that are directly deployable on all Android phones through the open Google TTS API. In addition, data created during this process is also suitable for a wide range of current and future corpus-based synthesis techniques, thus better synthesizers may be created with this data in future iterations.

Quality: To maintain maximal quality, we curate high quality data for a few languages using low-cost methods. Importantly, to allow data creation to be a participatory process we also encourage contributions from the community, providing a comprehensive set of directions that also cover issues like data collection and licensing.

Breadth: To cover many languages, we also include found data from the web in the AfricanVoices dataset. Specifically, we follow the CMU Wilderness project [10], which bootstraps datasets from found data (any long form audio with related text) using initial cross-lingual acoustic models to get the initial alignment, then in-language acoustic models to improve the dataset.

We open-source AfricanVoices which includes a speech synthesis corpora for 16 languages (including 3 that we record) and accompanying ready to use speech synthesizers on the AfricanVoices website available at https://www.africanvoices.tech/. We release alignment code, and number dictionaries which are accessible at https://github.com/neulab/AfricanVoices.

2. Focus Languages

AfricanVoices strives for both breadth and depth in its eventual goal of having high-quality voices for all African languages. The languages covered in the current iteration of the dataset are spoken in Southern, Western, Central and Eastern Africa. Figure 1 shows the current language coverage.

2.1. Created Data

We curated text and recorded high-quality speech data for 3 languages: Luo, Suba and Kenyan English.

Luo (luo) or Dholuo is a Nilo-Saharan language spoken by about 4.2 million speakers in Kenya and Tanzania.

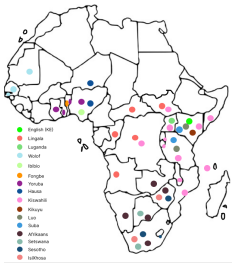

Figure 1: *Our focus languages*

Suba (sxb) is an endangered Bantu language spoken by the Suba people of Kenya and Tanzania, who live on the shores and islands of Lake Victoria. There are approximately 157,787 Kenyans who identify as ethnically Suba [11]; however, the number of native Suba speakers is less than 10,000 [12].

Kenyan English (en-ke) is the primary language of the government, media and schools and the official language alongside Kiswahili. English was introduced in Kenya through colonialism. Its vocabulary is similar to the British English while its phonology has been heavily influenced by local languages [13].

2.2. Found Data

In addition, we used existing found data for each where available under a permissive license. These languges below all use a modified version of Latin script.

Lingala (lin) is a Bantu language spoken by about 45 million native and lingua franca speakers mainly in the Democratic Republic of the Congo and to a lesser degree in Angola, the Central African Republic and Southern South Sudan [14].

Kikuyu (kik) is a highly agglutinative Bantu language spoken in the central region of Kenya by about 8 million people as a first language [11].

Yorùbá (yor) is a Niger-Congo language spoken by about 34 million speakers in Nigeria and other countries on the West African coast [15].

Hausa (hau) is an Afro-Asiatic language of the West Chadic branch with 60 million native speakers [16]. It is spoken in Nigeria and Niger.

Ibibio (ibb) is a Benue–Congo language with about 10 million speakers in Nigeria.

Wolof (wol) is a Niger-Congo language spoken by about 10 million speakers in total [14]. It is spoken in Senegal, Gambia and Mauritania.

Luganda (lug) is a Bantu language spoken in Uganda by about 20 million speakers.

Fongbe (fon) is a Niger-Congo language spoken in Benin by about 4.1 million speakers [17].

Afrikaans (afr) is a West-Germanic language spoken in South Africa, Namibia, Botswana, Zambia and Zimbabwe with an estimated 15-23 million speakers.

Kiswahili (swa) is a Niger-Congo language with about 200 million speakers [18]. It is a lingua franca in the Africa Great Lakes region and an official language for the East African Community states.

Sesotho (sot), Setswana (tsn), isiXhosa (xho) are southern Bantu language spoken in with about 5.6, 8.2, 8.7 million native speakers respectively. Sesotho is spoken in Lesotho, South Africa and Zimbabwe; Setswana is spoken in South Africa and Botswana and IsiXhosa is spoken in South Africa and Zimbabwe.

3. Dataset Creation

We encourage a participatory approach [4], enabling interested parties to easily create voices for the languages they speak or are interested in. We have created a comprehensive set of guidelines to allow even those without extensive experience to develop data and voices and contribute them back to the dataset if they wish. Below we briefly outline the processes and illustrate tricks, difficulties or pitfalls we encountered with our focus languages.

3.1. Creating Speech Data

Developing Prompt Sets We collect text data in the target languages by scraping various websites. Most of the data was created and published originally in our target languages. The text data sources for Luo and Suba included website copy, folklore and stories, research papers and theses, grammar and instructional books, and social media text. Most of Suba text data was obtained from [19], a website part of a project to revitalize the language. Luo was obtained from several sites, including translated data from an English news corpus developed from news articles by Kenyan media houses. It was easier to collect data for Luo, where we obtained 13,879 raw utterances than for Suba, where only 2,078 utterances were collected.

The lack of a standardized orthography for both languages posed challenges, as there were several instances of alternate spellings. Some Suba utterances also contained a mixture of dialects. For the Luo prompt set, we used Festvox tools [7] to select 1500 utterances that represent the phonetic and prosodic contexts of the language from the entire text corpus. This selection was not possible for Suba as its textual dataset was small, and thus we used the entire text corpus.

English prompt was obtained from CMU ARCTIC databases which consist of around 1150 phonetically balanced utterances [20].

Speaker selection When creating a corpus for TTS, a speaker (voice talent) should be fluent, literate, trained and familiar with voice recording [21]. Speaker selection was crucial for Suba as few people are fluent and literate in Suba. We advertised the position on social media groups whose participants were learning the Suba language. We received interest from participants who sent sample writings and recordings, but we ended up assigning the talent recommended by an authority on the language. The voice talents were paid in cash and kind.

Recording Environment In many cases, it is not possible to obtain studio quality data for resource-constrained languages. Luo and English were recorded in a residential house using a smartphone and in-ear microphone. It was possible to obtain use of recording facilities at a local radio station for Suba. While we could have recorded in any other studio, it was essential to relocate the voice talent to Mfang'ano island, predominantly occupied by Subas, to use the local radio station facilities which enabled us to receive feedback from the radio station on the quality of recordings and text.

Speech Recording The voice talents recorded audio with the selected prompts. Suba was recorded over one night session, and Luo was recorded in 4-hour sessions over 8 days.

Quality Control Before recording, we removed utterances with mixed dialects or changed to the selected dialect. After recording, we modified the prompt-set to reflect the actual content of the speech. We power-normalized the recordings to minimize the variation resulting from recording in different sessions and prosodic inconsistencies to ensure consistent volume.

Since most African languages exist in multi-lingual environments, everyday speech contains many borrowed words and constant code-switching. We faced the challenge of using a word's foreign or adapted pronunciation. For example, Luo speakers would pronounce the English word 'fish' as /fis/ instead of /fiʃ/ because of the absence of /ʃ/ in Dholuo phonology. In this case, we let the speaker use the pronunciation that was most natural to them.

3.2. Aligning Found Data

For many low-resourced languages, the Bible is a major source of text and audio. In this project, we used the New Testament from Bible.is [1] for Suba and Open.bible[2] for the rest.

3.2.1. Text preparation

To preprocess the text, we added chapter introductions and subtitles present in the audio but missing in the script. We also normalized numbers. To this end, we release **number dictionaries** for all languages we focus on that can be used to normalize numbers.

3.2.2. Speech preparation

Audio obtained from the Bible sources were saved as chapters in mp3 format. To segment and align it to utterance level, we followed the CMU Wilderness project's [10] segmentation and alignment process. Alignment for New Testament data, which is $\approx$ 20 hours of speech, took a maximum of 5 days per language on a 16-CPU machine. Table 1 shows the resulting utterances.

Table 1: *Data from found sources.*
***Data is not released*

Language	*source*	*No. utterances*	*hrs*
Luo	Open.Bible	11263	15.92
Lingala	Open.Bible	12957	27.52
Kikuyu	Open.Bible	10877	17.72
Yoruba	Open.Bible	10978	18.04
Hausa-M	CommonVoice	518	0.62
Hausa-F	CommonVoice	1938	2.3
Luganda	CommonVoice	2942	4.52
Ibibio	LLSTI	125	0.32
Kiswahili	LLSTI	426	0.53
Wolof	ALFFA	1000	1.2
Fongbe	ALFFA	542	0.33
Suba**	Bible.is	11971	24.82
Afrikaans	[21]	2927	3.30
Sestwana	[21]	2378	3.51
Sesotho	[21]	2096	3.22
IsiXhosa	[21]	2420	3.11

3.3. Found Data Sources

In addition to Open.Bible and Bible.is data that we aligned, we obtained data from the following sources in utterance format:
LLSTI: The Local Language Speech Technology Initiative project developed TTS datasets for localization of speech technology. We obtained Ibibio [22] and Kiswahili [23] by converting the publicly distributed lpc and res files to wav using Festvox tools.

[1]https://www.faithcomesbyhearing.com/audio-bible-resources/bible-is

[2]https://open.bible/resources/

Mozilla CommonVoice : We selected data from a single speaker with the most utterances for Luganda and Hausa.
ALFFA: ALFFA project [24] developed TTS and ASR technologies and data for Kiswahili, Fongbe, Wolof and Amharic. We selected a single speaker subset of the data for each language.
[21]: This project built TTS datasets for for South African languages: Afrikaans, IsiXhosa, Sesotho and Setswana.

3.4. Corpus License

We limited our collection to sources that are distributable to allow for free redistribution of the AfricanVoices data and works. The data from open.bible is distributed under CC by SA license which allows for sharing and adaptation as long as they are attributed, distributed under the same license and no additional restrictions are put on the derivative works. We cannot redistribute Suba data from Bible.is. We release all works under CC by SA license or the original permitted license for found data.

4. Experiments

To evaluate the effectiveness of the above-mentioned creation and curation processes, we perform experiments with Luo and Suba seeking to answer the following questions:

- RQ1: Are the datasets sufficient to build a high quality synthesizer in the targeted languages?
- RQ2: How much curated data is necessary?
- RQ3: How does curated data compare with found data?

To answer the questions above, for both created and found data, we divided each into splits of 25 min, 50 min and 101 min (the largest amount of created data that we have which was 102 minutes for languages). As mentioned previously, to allow for those with less experience in speech technology to expand AfricanVoices to other languages, we built TTS systems using Festvox tools [7] due to their relative accessibility as they do not require expensive compute resources.

4.1. Results and Analysis

For objective evaluation, we used the mean Mel Cepstral Distortion (MCD) score [25].[3] Table 2 shows the results of the automatic evaluation.

Table 2: *Objective evaluation using MCD (lower is better).*

Lang	**Source**	**25**	**50**	**101**
Luo	Found	4.73	4.73	4.65
	Created	6.49	6.45	6.37
Suba	Found	4.67	4.40	4.37
	Created	5.15	4.58	4.80

We also conducted human evaluation, to obtain subjective scores. We advertised the call for evaluators on social media platforms such as WhatsApp and Slack workspaces. To conduct the listening tests, we used TestVox[4], an open source web-based framework for running subjective listening tests [26].
Preference test We did A/B test to compare the synthesizer created using found data and created data. In this task, the evaluators were asked to respond to *Listen to the two audio clips*

[3]Previous work reports that an improvement in MCD of 0.12 is significant and recognisable to listeners [25].

[4]https://bitbucket.org/happyalu/testvox/wiki/Home

below, and select the one you prefer. by selecting either *A, B, or No difference*. Tables 3 and 4 show the results for the A/B test.

Transcription test We asked the evaluators to transcribe synthesized audio. The lack of a standardized orthography for both languages was a major challenge for this task. The most common 'errors' made by evaluators were (i) whether to join an agglutinated word or not eg *kawuononi vs kawuono ni* and (ii) similar sounds especially the semi-vowel *w* and vowel *u* eg *dwe vs. due* and (iii) whether to use a double vowel or not eg *Mbeeri vs Mberi*. We found no significant difference in the results from the different splits of the data. Luo and Suba had an average CER of (5.85 found and 5.80 created) and (10.80 found and 13.78 created) respectively.

Our results answer the questions in section 4 as follows:

- RQ1: Both objective and subjective tests show that created and found data are sufficient to build a synthesizer.
- RQ2: We found that at least 25 minutes of curated data is needed. Recording less than 25 minutes might not be worth the effort and cost of preparing recording.
- RQ3: The A/B tests show that curated data is comparable to found data despite their recording conditions. The evaluators consistently preferred output from created Suba.

Table 3: *Preference test results for Luo*

Split	Evaluator	Found	Created	Same	Best
25 min	Evaluator1	10	10	0	Found
	Evaluator2	11	8	1	
50 min	Evaluator1	11	9	0	Found
	Evaluator2	13	5	2	
101 min	Evaluator1	7	13	0	Created
	Evaluator2	6	11	3	

Table 4: *Preference results for Suba*

Split	Evaluator	Found	Created	Best
25 min	Evaluator1	1	19	Created
	Evaluator2	10	10	
50 min	Evaluator1	0	20	Created
	Evaluator2	9	11	
101 min	Evaluator1	3	17	Created
	Evaluator2	0	20	

It is important to note that not all languages, voices, prerecordings are equal. Some found data may be especially good (a consistent speaker), and some found may not (e.g. the Bible.is data has quiet background music). Some speakers from created data may be better than others, so experiments may have to be done for the particular target language.

5. Related Work

Creating a high-quality speech synthesizer demands high-quality single-speaker corpus [27] unlike automatic speech recognition (ASR), which requires a diverse multi-speaker corpus to capture different accents, speaker characteristics, and acoustic environments. The voice talent who record the speech are usually highly trained, fluent, and have experience recording speech. In low-resource settings, finding such speakers is hard due to the low economic development levels. Furthermore, the cost required to record in studios and extensively collect textual data poses a significant challenge to building high-quality TTS corpora for low-resource languages. This necessitates innovative approaches for speaker selection, speech recording, and post- processing of recorded audio.

When creating a multi-speaker speech corpora for 11 South African languages, Niekerk et al. [21] recorded the audio in low-cost environments like university campus buildings using low-cost tools like laptops and cheaper microphones and applying audio processing techniques tools to control things like background noise. Common Voice [28] is a platform to crowdsource transcribed speech corpora including African languages. While the platform and resultant corpora is useful for speech technology research and development, most of the data is less useful for speech synthesis as it is multi-speaker and recorded in varied environments.

A majority of high quality speech corpora and consequently speech synthesizers for African languages are commercial. The available ones represent a small fraction of the 2000 languages, with South African languages dominating because of government support through SADiLaR [5]. We found the following speech synthesis resources for African languages:

Existing high-quality speech synthesis corpora Some of the high quality include: NCHLT speech corpus [29], Lwazi II corpus, Gamayun Coastal Swahili speech corpus[6]

"Found data" Data in this category include the ones available online as part of other projects eg audiobooks, entertainment and news, and data created/availed for ASR. Mozilla Common Voice, Gamayun Congolese Swahili [30], Open.bible, Bible.is, A Kiswahili Dataset for Development of Text-To-Speech System [31]

Publicly available speech synthesizers These are TTS systems that are available for free. An example is LLSTI [32]

Commercial synthesizers Microsoft Text-to-Speech, Google API, Ajala AI, Inclusive solutions.

6. Conclusion

In this paper, we describe creating a speech corpus from found and created data sources for low-resource languages with a limited budget. We outline the challenges of creating a voice for an endangered language and suggest ways to overcome them. We find that ≈1 hour of speech is sufficient for creating an average synthesizer, even when recorded in suboptimal conditions. We build TTS for 16 languages and open-source models and speech corpora.

AfricanVoices is an incremental project, and we invite contributors to cover more African languages. Future work should be done to increase the geographic diversity of the corpus, and it will be desirable to focus on languages that are at most risk of facing extinction. In addition, further work can be done to develop web and mobile applications that can be used to record voices by untrained voice talent, thus making the process more accessible.

7. Acknowledgements

We acknowledge the voice talents Reinhardt Odete (Luo) and Victor Warekwe (Suba), evaluators and editors. We appreciate BlackAIR, Lacuna Fund, Biblica and other funders who funded the resources used in this work. The National Science Foundation partly supported this work under grant #2040926 from the Singapore Defence Science and Technology Agency.

[5] https://sadilar.org/index.php/en/resources

[6] https://gamayun.translatorswb.org/data/

8. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," 2017. [Online]. Available: https://arxiv.org/abs/1712.05884

[2] A. Fazel, W. Yang, Y. Liu, R. Barra-Chicote, Y. Meng, R. Maas, and J. Droppo, "SynthASR: Unlocking Synthetic Data for Speech Recognition," 2021. [Online]. Available: https://arxiv.org/abs/2106.07803

[3] Arya McCarthy, "The New Digital Divide: Language is the Impediment to Information Access," Apr. 2017. [Online]. Available: https://hilltopicssmu.wordpress.com/2017/04/08/the-new-digital-divide-language-is-the-impediment-to-information-access/

[4] W. Nekoto, V. Marivate, T. Matsila, T. Fasubaa, T. Kolawole, T. Fagbohungbe, S. O. Akinola, S. H. Muhammad, S. Kabongo, S. Osei, S. Freshia, R. A. Niyongabo, R. Macharm, P. Ogayo, O. Ahia, M. Meressa, M. Adeyemi, M. Mokgesi-Selinga, L. Okegbemi, L. J. Martinus, K. Tajudeen, K. Degila, K. Ogueji, K. Siminyu, J. Kreutzer, J. Webster, J. T. Ali, J. Abbott, I. Orife, I. Ezeani, I. A. Dangana, H. Kamper, H. Elsahar, G. Duru, G. Kioko, E. Murhabazi, E. van Biljon, D. Whitenack, C. Onyefuluchi, C. Emezue, B. Dossou, B. Sibanda, B. I. Bassey, A. Olabiyi, A. Ramkilowan, A. Öktem, A. Akinfaderin, and A. Bashir, "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages," 2020.

[5] D. Blasi, A. Anastasopoulos, and G. Neubig, "Systematic disparities in language technology performance across the world's languagtes," in *ACL 2022*, 2022.

[6] P. Ogayo, G. Neubig, and A. W. Black, "Building African Voices," 2022. [Online]. Available: https://www.africanvoices.tech/

[7] G. K. Anumanchipalli, K. Prahallad, and A. W. Black, "Festvox: Tools for creation and analyses of large speech corpora," in *Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia*, 2011, p. 70.

[8] A. Black and K. Lenzo, "Flite: a small fast run-time synthesis engine," in *4th ESCA Workshop on Speech Synthesis*, Scotland., 2001.

[9] A. Black and P. Muthukumar, "Random forests for statistical speech synthesis," in *Interspeech 2015*, Dresden, Germany., 2015.

[10] A. W. Black, "CMU Wilderness multilingual speech dataset," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5971–5975.

[11] Kenya National Bureau of Statistics, "2019," Kenya National Bureau of Statistics, Census report, Dec. 2019. [Online]. Available: https://www.knbs.or.ke/download/2019-kenya-population-and-housing-census-volume-iv-distribution-of-population-by-socio-economic-characteristics/

[12] Bernadine Racoma, "Olusuba Language of Africa on the Verge of Extinction," Mar. 2014. [Online]. Available: https://www.daytranslations.com/blog/olusuba-language-near-extinction/

[13] L. Nyaggah, *Cross-linguistic Influence in Kenyan English: The Impact of Swahili and Kikuyu on Syntax*. University of California, Los Angeles, 1990. [Online]. Available: https://books.google.com/books?id=DNwjQgAACAAJ

[14] Meeuwis, Michael, *A grammatical overview of Lingála : revised and extended edition*. Lincom, 2020.

[15] D. M. Eberhard, G. F. Simons, and C. D. Fennig (eds.), "Ethnologue: Languages of the World. Twenty-fifth edition," 2022, publisher: SIL International. [Online]. Available: https://www.ethnologue.com/language/yor

[16] ——, "Ethnologue: Languages of the World. twentieth edition," 2017, publisher: SIL International. [Online]. Available: https://www.ethnologue.com/language/yor

[17] "Building a database for Fongbe language in Africa | Knowledge 4 All Foundation Ltd." [Online]. Available: https://www.k4all.org/project/database-fongbe/

[18] The Conversation, "The story of how Swahili became Africa's most spoken language," Feb 2022. [Online]. Available: https://nation.africa/kenya/news/the-story-of-how-swahili-became-africa-s-most-spoken-language-3725834

[19] "Emirimo egia omunwa ogwa awasuba." [Online]. Available: https://subalanguage.com/sxb/amangana-amimpi

[20] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[21] D. van Niekerk, C. van Heerden, M. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, "Rapid development of TTS corpora for four South African languages," in *Proc. Interspeech 2017*, 2017, pp. 2178–2182. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1139

[22] D. Gibbon, E.-A. Urua, and M. Ekpenyong, "Problems and solutions in African tone language Text-To-Speech," *ISCA Publication*, 01 2006.

[23] M. Gakuru, F. Iraki, R. Tucker, K. Shalonova, and K. Ngugi, "Development of a Kiswahili text to speech system," 09 2005, pp. 1481–1484.

[24] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui, "Collecting resources in sub-Saharan African languages for automatic speech recognition: a case study of Wolof," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 3863–3867. [Online]. Available: https://aclanthology.org/L16-1611

[25] J. Kominek, T. Schultz, and A. W. Black, "Synthesizer voice quality of new languages calibrated with mean mel cepstral distortion." in *SLTU*, 2008, pp. 63–68.

[26] A. Wilkinson, A. Parlikar, S. Sitaram, T. White, A. Black, and S. Bazaj, "Open-source consumer-grade indic text to speech," 09 2016, pp. 190–195.

[27] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," *CoRR*, vol. abs/2106.08468, 2021. [Online]. Available: https://arxiv.org/abs/2106.08468

[28] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," 2020.

[29] E. Barnard, M. H. Davel, C. J. van Heerden, F. de Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *4th Workshop on Spoken Language Technologies for Under-resourced Languages, SLTU 2014, St. Petersburg, Russia, May 14-16, 2014*. ISCA, 2014, pp. 194–200. [Online]. Available: http://www.isca-speech.org/archive/sltu_2014/sl14_194.html

[30] A. Öktem, M. A. Jaam, E. DeLuca, and G. Tang, "Gamayun - language technology for humanitarian response," pp. 1–4, 2020.

[31] K. Rono, "A kiswahili dataset for development of text-to-speech system"," *Mendeley Data*, vol. V1, 2021.

[32] "Local Language Speech Technology Initiative." [Online]. Available: http://www.llsti.org/