MULTILINGUAL MODEL AND DATA RESOURCES FOR TEXT-TO-SPEECH IN UGANDAN LANGUAGES

Isaac Owomugisha, Benjamin Akera, Ernest Mwebaze, John Quinn
Sunbird AI, Uganda

ABSTRACT

We present new resources for text-to-speech in Ugandan languages. Studio-grade recordings in Luganda and English were captured for 2,413 and 2,437 utterances respectively (totaling 4,850 utterances representing 5 hours of speech). We show that this is sufficient to train high-quality TTS models which can generate natural sounding speech in either language or combinations of both with code switching. We also present results on training TTS in Luganda using crowdsourced recordings from Common Voice. Additional data collection is currently underway for the Acholi, Ateso, Lugbara and Runyankole languages. The data we describe is an extension to the SALT dataset, which already contains multi-way parallel translated text in six languages. The dataset and models described are publicly available at `https://github.com/SunbirdAI/salt`.

1 INTRODUCTION

Speech interfaces are particularly important for languages which are primarily spoken rather than written. In recent years, there has been a surge in the development of text-to-speech (TTS) systems, with several works proposing effective deep learning architectures for speech synthesis from text. Despite this progress, TTS systems are not available for the majority of the world's languages, and those which exist often cannot support code-switching (Zhou et al., 2020). Code-switching is particularly relevant in multilingual regions such as Uganda, where local languages and English are combined within single sentences and even within single words (leading to the coining of hybrid words, e.g. "okutweeting", "to tweet").

We describe here an extension of the Sunbird African Language Technology (SALT) dataset (Akera et al., 2022) to include TTS training data in Ugandan languages. SALT already contains multi-way parallel text data sourced from the local community in English and five Ugandan languages, and the intention is to create a widely multi-lingual, multi-modal resource for training and evaluating a variety of language tasks. We describe here the Luganda and (Ugandan) English TTS data, totalling 4,850 utterances and 5 hours of speech data in both languages; collection is currently underway of TTS data in the Acholi, Ateso, Lugbara and Runyankole languages.

A TTS model trained on the Luganda and English data achieves good quality speech synthesis, obtaining a mean opinion score (MOS) of 3.37. We compare this to training TTS models in Luganda with crowd-sourced speech recordings from Common Voice (MOS: 2.5/2.33 for female/male voices). Whilst the Common Voice baseline was lower in quality, we noted that generated speech is intelligible enough to be usable in some practical applications, somewhat surprising given this data was not collected for the purposes of TTS.

Previous relevant work includes BibleTTS (Meyer et al., 2022), although Luganda recordings in this dataset were not yet aligned at the sentence level to make it possible to train or evaluate models. To our knowledge, SALT is the first TTS dataset and production-quality model for Luganda, which additionally is designed to address the challenges posed by code-switching. Data collection for other Ugandan languages is in progress. The model, code and data resources are all publicly available to aid further research and development in this direction.

	Gold standard	Studio	Common Voice (Male)	Common Voice (Female)
MOS	4.70 ± 0.27	3.37 ± 0.39	2.33 ± 0.34	2.50 ± 0.29

Table 1: MOS evaluations and 95% confidence intervals for each model on $N = 25$ audio samples. Gold standard refers to human speech by a professional voice-over artist reading the same test sentences.

2 Data Collection

In this study, two data collection approaches were employed. The first involved the collection of high-quality professional speech recorded in a studio setting. As a baseline, we also describe the utilization of crowd-sourced speech data from the Common Voice platform.

Studio speech recordings We collected a total of 4,850 phrases in a mixture of Luganda and English from the SALT multiway dataset, and had them recorded by a single professional speech/voice actor (female). Recordings were carried out in a studio setting. Following the recording, we employed voice activity detection to trim off the silent portions from the beginning and end of each speech sample.

Crowd-sourced speech recordings In addition to the studio recordings, we utilized secondary data from the Common Voice corpus, to understand the extent to which TTS could be trained with existing data. Although Common Voice data is collected with speech recognition, rather than speech generation, in mind, there are some findings in the literature than it can be used for TTS (Chien et al., 2021). We selected Luganda speech data from Common Voice for male and female speakers, in each case filtering for speakers within the age range 20-49 and then taking 15,000 samples. We again trimmed silences using voice activity detection.

3 Models and Results

We applied the Tacotron2 architecture (Shen et al., 2018), using the SpeechBrain implementation (Ravanelli et al., 2021), to generate audio. We fine-tuned to each dataset from a female US English model trained on the LJSpeech. When training the model on studio data, we used Luganda and English data simultaneously in order to yield multilingual TTS. We found that Tacotron2 was straightforward to train with studio data. Convergence was more difficult to achieve with the Common Voice dataset, perhaps due to the varying intonations of multiple speakers as well as background noise, requiring several weeks of training on a single P100 GPU (batch size 64).

To evaluate our TTS we used the Mean Opinion Score (MOS) metric which is based on subjective evaluations of the synthesized speech from survey participants. Scores were given by five individuals, native speakers of Luganda, for 25 generated samples per model. Our MOS evaluations followed the Absolute Category Rating scale (ITU-T Recommendation, 1996) with rating scores from 1 (bad) to 5 (excellent). We also evaluated MOS for gold standard audio, in this case the same 25 sentences recorded by the professional voiceover artist. Results in Table 1 show that the model performs well on studio data, as expected. Baseline models trained on Common Voice speech data are lower quality, though still clearly intelligible. We observe the model based on the female Common Voice speech data outperforms the male model, possibly because the pre-trained checkpoint used was a female US English voice.

4 Conclusion

All the resources arising from this work (TTS training data, training code and trained models) are available publicly at `https://github.com/SunbirdAI/salt`. As well as providing studio TTS data, we find interesting evidence that a baseline system trained on crowd-sourced speech recordings is also practical, yielding lower quality but intelligible synthesised speech. We believe that speech interfaces are particularly important in the African context to help make NLP-based systems accessible to the majority.

REFERENCES

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. Machine translation for african languages: Community creation of datasets and models in uganda. 2022.

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8588–8592. IEEE, 2021.

ITU-T ITU-T Recommendation. P. 800: Methods for subjective determination of transmission quality. *International Telecommunication Union, Geneva*, 1996.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, et al. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus. *arXiv preprint arXiv:2207.03546*, 2022.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al. Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*, 2021.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Xuehao Zhou, Xiaohai Tian, Grandee Lee, Rohan Kumar Das, and Haizhou Li. End-to-end code-switching tts with cross-lingual language model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7614–7618. IEEE, 2020.