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

Anonymous authors
Paper under double-blind review

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. 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 [*link removed to preserve anonymity*].

1 INTRODUCTION

In recent years, there has been a surge in the development of text-to-speech (TTS) systems, with several works proposing various deep learning architectures for speech synthesis from text Wang et al. (2017), Arık et al. (2017), Li et al. (2019). 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 (the use of multiple languages within a single sentence 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").

In this paper, we present the SALT v2 dataset, an extension of the original multi-way text SALT dataset, which incorporates Luganda and (Ugandan) English speech data. The speech data includes 4,850 utterances and 5 hours of speech data in both languages, recorded by a professional voice-over artist in a studio setting.

A TTS model trained on this data achieves good quality speech synthesis, obtaining a mean opinion score (MOS) of 3.37. We were also able to obtain less natural, but still intelligible, generated speech from secondary models in which only crowd-sourced Luganda speech recordings from Common Voice were used for training (MOS: 2.5/2.33 for female/male voices). To our knowledge, this is the first TTS dataset and production-quality model for Luganda, which additionally is designed to address the challenges posed by code-switching. The model, code and data resources are all publicly available to aid further research and development in this direction.

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. The second approach was the utilization of crowd-sourced speech data from the Common Voice platform. The combination of these two data sources provided a diverse set of speech data to train the TTS models, enabling us to perform a comprehensive evaluation of the TTS models' performance.

2.1 STUDIO SPEECH RECORDINGS

For the studio data, 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 professional speech/voice actor. We found

	Ground truth	Studio data	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 for models created based on the indicated data with 95% confidence intervals.

available TTS data collection software difficult to apply because of the studio setup in which the microphone was inside the recording booth, connected to a machine outside at the mixing desk. Initially we devised a prompt application with which the voice-over artist could view the text to be spoken on a mobile phone screen, but due to issues with network cutouts we prepared the majority of the dataset by manually exporting each sentence with the existing studio software. Following the recording, we employed voice activity detection to trim off the silent portions from the beginning and end of each speech sample.

2.2 Crowd-sourced speech recordings

In addition to the studio recordings, we utilized secondary data from the Common Voice corpus. 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, but carried out no further filtering or selection of samples.

3 Models

We applied the Tacotron2 architecture Shen et al. (2018), using the SpeechBrain implementation Ravanelli et al. (2021), to generate speech. Beginning with a US English model, we fine-tuned on each of the datasets prepared. When training the model on studio data, we used Luganda and English data simultaneously in order to yield multilingual TTS. We found that Tacotron2 convergence in each case was lengthy, and required several weeks of training on a single P100 GPU.

4 Results

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, and evaluates how natural the speech sounds. Our MOS evaluations followed the *Absolute Category Rating scale* ITU-T Recommendation (1996) with rating scores from 1 (bad) to 5 (excellent). Results in Table 1 show that the model performs best when trained on studio-based datasets that are clear with less background noise. Performance is significantly better than the models trained on Common Voice speech data. It also performs close to the ground truth which in this case is the professional speech from the studio. For Common Voice data, we observe the model based on the female speech data outperforms the male model, possibly because the pre-trained checkpoint used was a female US English voice.

5 Conclusion

All the resources arising from this work (TTS training data, training code and trained models) are available publicly at [*link withheld to preserve anonymity*]. As well as providing studio data, we add to the evidence in the literature that crowd-sourced speech recordings can be used to train usable TTS, even when not collected for that purpose. We believe that speech interfaces are particularly important in the African context to help make NLP-based systems accessible to the majority.

Acknowledgements

[*Withheld for anonymity.*]

REFERENCES

Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*, pp. 195–204. PMLR, 2017.

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.

Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6706–6713, 2019.

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.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017.

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.