KINYARWANDA TTS: USING A MULTI-SPEAKER DATASET TO BUILD A KINYARWANDA TTS MODEL

Samuel Rutunda
Digital Umuganda, Rwanda
samuel@digitalumuganda.com

Kleber Kabanda
Digital Umuganda, Rwanda
kleber@digitalumuganda.com

Adriana Stan
Technical University of Cluj-Napoca, Romania
adriana.stan@com.utcluj.ro

ABSTRACT

The field of text-to-speech (TTS) technology has been rapidly advancing in recent years, and has become an increasingly important aspect of our lives. This presents an opportunity for Africa, especially in facilitating access to information to many vulnerable socio-economic groups. However, the lack of availability of high-quality datasets is a major hindrance. In this work, we create a dataset based on recordings of the Bible. Using an existing Kinyarwanda speech-to-text model we were able to segment and align the speech and the text, and then created a multi-speaker Kinyarwanda TTS model.

1 INTRODUCTION

The vast majority of Africa throughout history was traditionally an oral culture, with its history, culture, and values being transmitted orally from one generation to another. While this was a rich and valuable tradition, it was also prone to loss and misinterpretation over time. The introduction of writing allowed African societies to preserve their history, culture, and knowledge in a more reliable and permanent form. The era of writing brought mostly through education was a milestone, however, the literacy rate in Africa is still around 67%, with some areas having less than 30%. This hinders the socio-development of the area, and the most affected by this problem are the marginalized citizens of the society, such as people with disabilities, women, people from rural areas, etc. Technology Inclusion, especially with the rise of Artificial intelligence (AI), can be a driving force in mitigating some of the problems and a major driver to benefit those left behind.

Text to speech (TTS), the ability of the computer to read out loud written text, is one of those AI tools that can help. With the advent of end-to-end TTS models, it became easier to build speech synthesis, without the need for much more complicated processes, such as feature engineering. However, an end-to-end model still requires a lot of data, which low-resource languages do not possess. A good TTS requires studio-quality recordings devoid of background noise, and segmented into medium length audio chunks. Such data is hard to come by for African languages.

In our work, we build a TTS model leveraging the existing Kinyarwanda Audio bible, using the existing Kinyarwanda[1] Speech to Text (STT) and Nemo[2] CTC-Segmentation to align the audio and text datasets. Following this process, we obtained utterances with a duration ranging from 3 to 47 seconds amounting to a total of 67.84 hours of studio quality dataset. The TTS model is then trained using the Coqui[3] framework.

[1] https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_rw_conformer_ctc_large

[2] https://github.com/NVIDIA/NeMo

[3] https://github.com/coqui-ai/TTS

2 RELATED WORK

2.1 ADVANCES IN TEXT TO SPEECH

With the development of deep learning and artificial intelligence, neural network-based TTS has significantly improved the quality of synthesized speech in recent years. In the near 2010s, as neural network and deep learning have achieved rapid progress, some early neural models such as deep neural network (DNN) based Qian et al. (2014) and recurrent neural network (RNN) based Zen (2015), Fan et al. (2014) are adopted in Statistical Parametric Speech Synthesis (SPSS), an earlier TTS method, to replace HMMs for acoustic modeling.

Neural TTS usually consists of three components: a text analysis module, a parameter prediction module (acoustic model), and a waveform generation module (vocoder). The text analysis module first processes the text, including text normalization Sproat et al. (2001), grapheme-to-phoneme conversion Bisani & Ney (2008), word segmentation, etc. It then extracts the linguistic features, such as phonemes, syllabification, and POS tags at different granularities. The acoustic models are trained with these linguistic features to generate acoustic features including fundamental frequency, phone duration, spectrum, or cepstrum Fukada et al. (1992). The vocoder synthesize speech from the predicted acoustic features.

Later, WaveNet Oord et al. (2016) is proposed to directly generate waveforms from linguistic features, which can be regarded as the first modern neural TTS model. Other models like DeepVoice Arik et al. (2017) still follow the three components in statistical parametric synthesis but upgrade them with the corresponding neural network-based models. Furthermore, some spectrogram models (e.g: Tacotron 1/2 Wang et al. (2017), Deep Voice 3 Ping et al. (2018), and FastSpeech 1/2 Ren et al. (2019; 2022)) are proposed to simplify text analysis modules and directly take character/phoneme sequences as input and simplify acoustic features with Mel-spectrograms. Fully end-to-end TTS systems were developed to directly generate waveforms from text, such as ClariNet Ping et al. (2019), FastSpeech 2 Ren et al. (2022), and EATS Donahue et al. (2021). TTS has continued to evolve with multilingual and multi-speaker models Casanova et al. (2022) enabling the training of multiple speakers, and multiple languages, giving advantages to low-resourced languages. The other end of the spectrum is training on a huge dataset Wang et al. (2023) you are able to generate voice using a clone of a 3-second recording.

Compared to previous TTS systems based on concatenative synthesis and statistical parametric synthesis, the advantages of neural network-based speech synthesis include higher quality in terms of both intelligibility and naturalness, and fewer requirements of human preprocessing and feature development.

2.2 TEXT-TO-SPEECH FOR AFRICAN LANGUAGES

There are several works that have focused on building speech synthesis datasets and models for African languages. Bible TTS, Josh, et al. Meyer et al. (2022), leverages the open licensed Bible. They created Text to Speech datasets for 10 African languages, they used Montreal force Alignment to align the text with the audio. (Ogayo et al., 2022) uses the CMU Flite to build TTS models for 12 African languages. CMU Flite is based on a random forest and does not require any GPU for computation, and a small dataset. The drawback is the quality of the TTS where the naturalness metric is low. Gamayun Öktem et al. (2020) is a data collection platform for machine translation and voice data collection used to collect voice across Africa, easing the creation of new datasets. (Gutkin et al., 2020) created an open clean Yoruba dataset. (Gakuru et al., 2005) created a parametric Swahili TTS based on the Festival speech synthesizer [4].

2.3 KINYARWANDA

Kinyarwanda is part of the Bantu Nurse & Philippson (2003) language family spoken in Central-eastern Africa, part of the JD subgroup of Bantu languages alongside Kirundi, Fuliro, Ha, Havu, and others. It is spoken by more than 13 million speakers in Rwanda, and it is one of the official languages of Rwanda, used in administration, education, and as the lingua franca of Rwanda. Kin-

[4] https://www.cstr.ed.ac.uk/projects/festival/

yarwanda is a tonal language Muhirwe (2010), although it is written without the tones, as they are implicitly added by the speaker. In Kinyarwanda, two or words words may be written the same but pronounced differently, e.g. family [umuryaango] (family) and door [umuryângo]; the reader must extract the meaning depending on the context. Note, this paper does not cover the scope of the tones.

With respect to available digital data, Kinyarwanda still lacks the resources needed to create speech-enabled applications. One of the most notable efforts in this direction has been the CommonVoice [5] endeavor which enabled the collection of over 2387 hours of data from 1106 speakers. This resource is extremely important for developing speech-to-text systems, but does not correspond to the requirements of a high-quality dataset for text-to-speech.

3 DATASET

To create a viable dataset; we used two separate data sources. We obtained the Kinyarwanda bible audio recordings from Faith Comes By Hearing[6] website and the text from bible.com (BIR: Bibiliya Ijambo ry'Imana)[7] through web-scraping. The audio covers only the 39 books of the Old Testament; consequently, we only used the Old Testament books from the scraped text. We converted numerals to their corresponding Kinyarwanda words using hand-crafted rules.

Each scraped Bible page is stored in its file. A newline in the page's file separates the elements of a page (verses and headers). Since each audio covers a page in the bible, all audio files were longer than thirty seconds, thirty seconds being around the expected audio clip duration needed to train our model. We used the CTC-Segmentation technique Kürzinger et al. (2020) to generate and align audio corresponding to each line on every page of the scraped bible text with their matching audio segment in the audio representing that page. In the background, the CTC-Segmentation algorithm uses a speech-to-text model; for this purpose, we used NVIDIA's Kinyarwanda conformer-CTC based Automatic Speech Recognition model, trained on the Mozilla Common Voice(v9.0) Kinyarwanda dataset with an 18.22 WER. After the segmentation, the CTC-segmentation algorithm generates a confidence score(in the log space). We used the confidence score to measure the overlap and alignment between the audio's content and their corresponding transcript. After finding all audio and transcripts pairs with low alignment confidence scores(confidence score below -5); we choose which ones to manually correct (add, or remove text in the transcript to match the audio) and which to dismiss depending on the amount of mismatch. It should be noted that the mismatch in the resulting segmentation was mostly caused by contradictions between spoken words in the audio and their corresponding transcriptions as they come from separate sources. Additionally, we ignored all audio files of less than three seconds, since most of them were erroneous. The result is sixty-seven hours, multi-speaker dataset, and its brief summary is shown in Table 1.

Table 1: Dataset specification

Specification	**value**
Total number of speakers	11
Duration of a single clip (seconds)	3 - 47
Total number of clips	39951
Total number of hours	67.84

4 MODEL

YourTTS Casanova et al. (2022) is a multilingual, multi-speaker TTS model with zero-shot capabilities, it is based on the VITS Kim et al. (2021) model. The VITS model is a single-stage end-to-end text-to-speech model (see Figure 1 YourTTS builds upon the work on VITS and adds multi-speakers and multilingual capabilities. YourTTS uses raw text instead of phonemes as input, a transformer-based text encoder, and for multi-lingual training it employs trainable language embeddings. A

[5] https://commonvoice.mozilla.org/rw

[6] https://www.faithcomesbyhearing.com/audio-bible-resources/recordings-database

[7] https://www.bible.com/bible/395

HIFI-GAN version 1 is used as the vocoder and it is connected to the TTS model using a variational autoencoder, in this case, a Posterior Encoder. The Posterior Encoder receives linear spectrograms and outputs latent variables that are fed to the vocoder; this configuration allows for the learning of intermediate representations; a contributing factor in the model's improved results compared to the two-stage models. A flow-based decoder takes the latent variable(from the Posterior Decoder) and speaker embeddings and conditions them with respect to a prior distribution P, to align P with the output from the text encoder a Monotonic Alignment Search(MAS) is used. A stochastic duration predictor is used to generating human-like rhythms of speech, it receives as input the language embedding, speaker embeddings, and duration obtained through MAS. A pretrained speaker encoder is used to generate speaker embeddings; these embeddings give the model zero-shot capabilities by conditioning the flow-based decoder, the posterior encoder, and the vocoder, they are also used to find the Speaker Consistency Loss(SCL) which is the cosine similarity between the generated audio and the ground truth. To synthesize a waveform at inference time; MAS is not used, the distribution P is predicted by the text encoder, and the inverted stochastic duration predictor given a noise as input, is used to sample the duration. The latent variable is sampled from the distribution P and with the speaker embeddings fed to the inverted flow-based decoder. The output of the flow-based decoder is passed to the vocoder which in turn generates the waveform.

In this work, we trained a YourTTS model using sixty-seven hours of bible data on a single A-100 GPU. We used raw text instead of phonemes, since we did not have a grapheme-to-phoneme converter. Although the data contained speaker information, this was not used during training. At inference time, a reference sample of the target speaker was used to condition the output identity of the TTS system. After a hundred epochs and sixteen hours of training, we reached a reasonable voice quality. The model is available on Huggingface.[8]

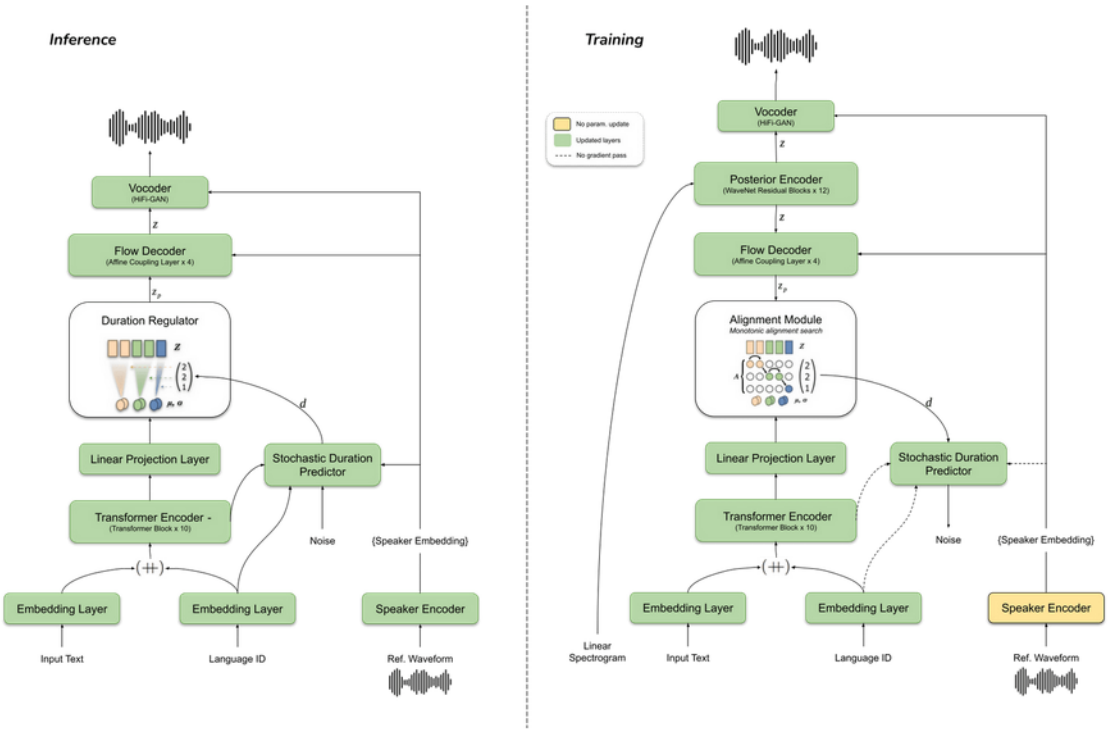

Figure 1: YourTTS training and inference architecture[9](Casanova et al., 2022)

[8] https://huggingface.co/DigitalUmuganda/Kinyarwanda_YourTTS_v1

[9] https://coqui.ai/blog/tts/yourtts-zero-shot-text-synthesis-low-resource-languages

Table 2: Training Hyperparameters

Hyperparameter	Value
Epochs	100
Sampling rate	22050
Learning rate	0.001
Weight decay	0.01
Betas	[0.8,0.99]
Batch size	12
No. of CPUs	72
No. of GPUs	1
No. of mels	80
Use phonemes	False

5 Results and Analysis

A first analysis looked into the objective intelligibility measure. We synthesized five-hundred samples using in-domain bible data, and transcribed them using Nvidia's ASR CTC model. The ASR's WER on natural data is 10.03%. The resulting transcriptions from synthesized audio files gave a 30.09% WER.

In a separate analysis, the output speech was evaluated by Kinyarwanda native speakers, in terms of naturalness and intelligibility. The naturalness was evaluated on a MOS scale of 1 [very unnatural] to 5 [natural - human speech]. The listeners were presented with both synthesised and natural samples. A boxplot of the results is shown in Figure 2. The average rating for the synthesised speech was 2.3. This is below the general levels of subjective evaluation, but this was to be expected as the data used to train the TTS system was not standard. In the intelligibility section, the users were asked to rate if the speech was intelligible, partly intelligible or not intelligible at all. A piechart of the results for the synthesised samples is shown in Figure 3. It can be noticed that around 87% of the samples were rated as intelligible or partially intelligible.

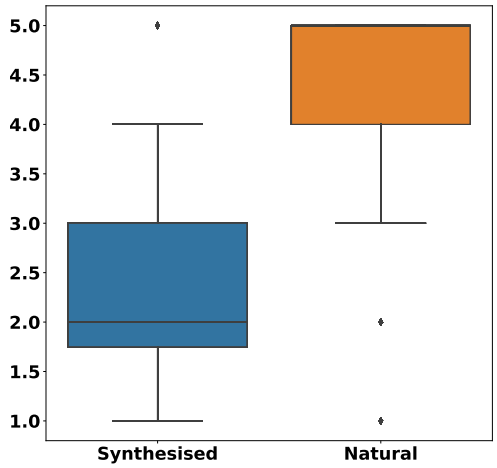

Figure 2: Naturalness - Synthesized - Natural - Common voice synthesized

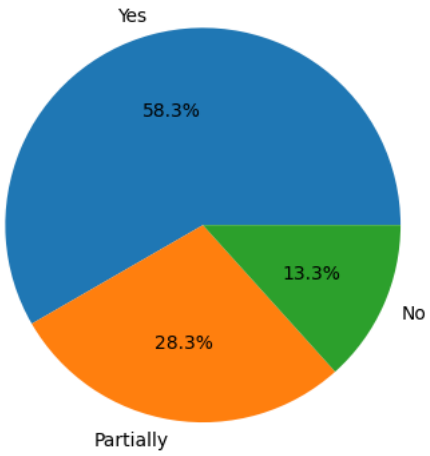

Figure 3: Intelligibility piechart.

6 CONCLUSIONS

In this paper, we created and used a Bible Kinyarwanda dataset to build a TTS model, leveraging an existing Kinyarwanda STT model; we used the CTC-Segmentation technique to align the voice and the text from different sources. We generated 67.84 hours dataset and trained it on the YourTTS model, which we tested on the WER , the naturalness (MOS) and the intelligibility. The result was 30.09% WER on generate audio, The MOS was 2.3 and 87% of the samples were rated as intelligible or partially intelligible.

Future work will involve training a multilingual model using other Bantu languages such as Luganda and Swahili, investigating ways to improve the tonal quality, using techniques such as creating a tonal dictionary, and adding a studio-quality dataset that uses modern and frequently used sentences.

ACKNOWLEDGMENTS

We would like to express our gratitude to the FAIR Forward - Artificial Intelligence for All[10] and DFKI (German Research Center for Artificial Intelligence)[11] for their generous support of this research project. Special thanks to GIZ for providing us with the necessary resources to carry out this work and to DFKI for providing us with the GPU infrastructure that was used to train our text-to-speech model. This research would not have been possible without their valuable contribution. We also thank the mbaza-nlp community[12] for their contribution in testing and providing feedback on how to improve the model.

REFERENCES

Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep Voice 2: Multi-Speaker Neural Text-to-Speech, September 2017. URL http://arxiv.org/abs/1705.08947. arXiv:1705.08947 [cs].

Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434–451, May 2008. ISSN 0167-6393. URL https://www.sciencedirect.com/science/article/pii/S0167639308000046.

Edresson Casanova, Julian Weber, Christopher Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir Antonelli Ponti. YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice

[10]FAIR Forward: https://www.giz.de/expertise/html/61982.html
[11]DFKI: https://www.dfki.de/en/web
[12]Mbaza NLP community: https://mbaza.org/

Conversion for everyone, February 2022. URL http://arxiv.org/abs/2112.02418. arXiv:2112.02418 [cs, eess].

Jeff Donahue, Sander Dieleman, Mikołaj Bińkowski, Erich Elsen, and Karen Simonyan. End-to-End Adversarial Text-to-Speech, March 2021. URL http://arxiv.org/abs/2006.03575. arXiv:2006.03575 [cs, eess].

Yuchen Fan, Yao Qian, Fenglong Xie, and F. Soong. TTS synthesis with bidirectional LSTM based recurrent neural networks. 2014. URL https://www.semanticscholar.org/paper/TTS-synthesis-with-bidirectional-LSTM-based-neural-Fan-Qian/c217905bc98f00af747e8e9d5f6b79fb89a90886.

T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai. An adaptive algorithm for mel-cepstral analysis of speech. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 137–140 vol.1, March 1992. doi: 10.1109/ICASSP.1992.225953. ISSN: 1520-6149.

Mucemi Gakuru, Frederick Iraki, Roger Tucker, Ksenia Shalonova, and Kamanda Ngugi. Development of a kiswahili text to speech system. pp. 1481–1484, 09 2005. doi: 10.21437/Interspeech.2005-522.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E. Rivera, and Kólá Túbòsún. Developing an open-source corpus of yoruba speech. In *Proc. of Interspeech 2020*, pp. 404–408, October 25–29, Shanghai, China, 2020., 2020. URL http://dx.doi.org/10.21437/Interspeech.2020-1096.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech, June 2021. URL http://arxiv.org/abs/2106.06103. arXiv:2106.06103 [cs, eess].

Ludwig Kürzinger, Dominik Winkelbauer, Lujun Li, Tobias Watzel, and Gerhard Rigoll. CTC-segmentation of large corpora for german end-to-end speech recognition. In *Speech and Computer*, pp. 267–278. Springer International Publishing, 2020. doi: 10.1007/978-3-030-60276-5_27. URL https://doi.org/10.1007%2F978-3-030-60276-5_27.

Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iroro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus, July 2022. URL http://arxiv.org/abs/2207.03546. arXiv:2207.03546 [cs, eess].

Jackson Muhirwe. Morphological Analysis of Tone Marked Kinyarwanda Text. In Anssi Yli-Jyrä, András Kornai, Jacques Sakarovitch, and Bruce Watson (eds.), *Finite-State Methods and Natural Language Processing*, Lecture Notes in Computer Science, pp. 48–55, Berlin, Heidelberg, 2010. Springer. ISBN 978-3-642-14684-8. doi: 10.1007/978-3-642-14684-8_6.

Derek Nurse and Gérard Philippson. *The Bantu Languages*. Routledge, London, July 2003. ISBN 978-0-203-98792-6. doi: 10.4324/9780203987926.

Perez Ogayo, Graham Neubig, and Alan W. Black. Building African Voices, July 2022. URL http://arxiv.org/abs/2207.00688. arXiv:2207.00688 [cs, eess].

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio, September 2016. URL http://arxiv.org/abs/1609.03499. arXiv:1609.03499 [cs].

Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning, February 2018. URL http://arxiv.org/abs/1710.07654. arXiv:1710.07654 [cs, eess].

Wei Ping, Kainan Peng, and Jitong Chen. ClariNet: Parallel Wave Generation in End-to-End Text-to-Speech, February 2019. URL http://arxiv.org/abs/1807.07281. arXiv:1807.07281 [cs, eess].

Yao Qian, Yuchen Fan, Wenping Hu, and Frank K. Soong. On the training aspects of Deep Neural Network (DNN) for parametric TTS synthesis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3829–3833, May 2014. doi: 10.1109/ICASSP.2014.6854318. ISSN: 2379-190X.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech: Fast, Robust and Controllable Text to Speech, November 2019. URL http://arxiv.org/abs/1905.09263. arXiv:1905.09263 [cs, eess].

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, August 2022. URL http://arxiv.org/abs/2006.04558. arXiv:2006.04558 [cs, eess].

Richard Sproat, Alan W. Black, Stanley Chen, Shankar Kumar, Mari Ostendorf, and Christina D. Richards. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333, July 2001. ISSN 0885-2308. doi: 10.1006/csla.2001.0169. URL https://www.sciencedirect.com/science/article/pii/S088523080190169X.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers, January 2023. URL http://arxiv.org/abs/2301.02111. arXiv:2301.02111 [cs, eess].

Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. Tacotron: Towards End-to-End Speech Synthesis, April 2017. URL http://arxiv.org/abs/1703.10135. arXiv:1703.10135 [cs].

Heiga Zen. Acoustic Modeling in Statistical Parametric Speech Synthesis - From HMM to LSTM-RNN. In *Proc. MLSLP*, 2015.

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, and Grace Tang. Gamayun - language technology for humanitarian response. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 1–4, 2020. doi: 10.1109/GHTC46280.2020.9342939.