Accent-Aware Text-to-Speech for Nigerian English: Building Inclusive Voice AI from Community-Curated Data

Voice technologies often fail to represent the linguistic diversity of emerging markets, particularly African accents and languages[4]. This work presents a multilingual text-to-speech (TTS) system tailored for Nigerian-accented English, built on the StyleTTS2 architecture[1] and trained on a community-curated dataset spanning the three major Nigerian ethnic groups: Yoruba, Igbo, and Hausa.

To construct the dataset, volunteers with technical backgrounds recorded readings from Nigerian-published texts across domains such as religion, politics, history, and education. Recordings ranged from 1 to 6 hours per speaker. Using Whisper[6] for transcription, audio was converted into timestamped SRT files and manually corrected by a four-person team. A custom script segmented the audio into variable-length clips (2–30 seconds), yielding over 4,000 paired samples. The dataset was split 80/20 for training and evaluation.

During preprocessing, transcriptions were converted into phonemes using the phonemizer Python package. The model learns the relationship between phoneme sequences and speaker-specific acoustic features, including pitch and prosody. At inference time, given a new text input and reference audio, the model mimics the speaker's vocal style by predicting pitch contours and generating expressive speech that reflects the speaker's accent and emotional tone.

As illustrated in Fig. 1, the system architecture integrates phoneme-level BERT embeddings, style and prosody encoders, and a diffusion-based decoder [2,3,5]. An informal evaluation was conducted using human raters, with three evaluators per sample. Approximately 87% of synthesized outputs were rated as accent-faithful and emotionally consistent with the reference audio. Additionally, Word Error Rate (WER) was used to assess intelligibility across test samples, with an average WER of 12.4% across the test set.

This work demonstrates the feasibility of building inclusive voice AI using modest resources and community participation, with potential applications in education, public services, and digital accessibility across Africa.

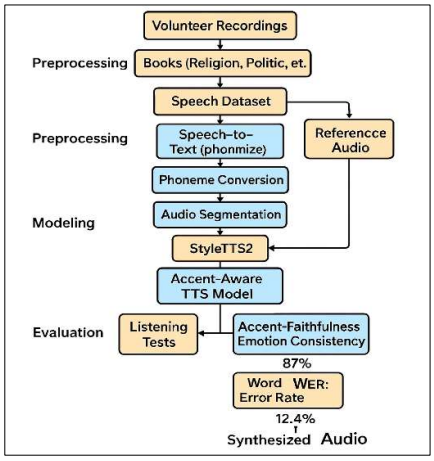

Fig. 1: Accent-Aware TTS Workflow

References

[1] Aaron, Y., Cong, L., Vinay, H., Raghavan, S., Mischler, G., & Mesgarani, N. (2023). StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. *Advances in Neural Information Processing Systems*, *36*, 19594–19621. https://styletts2.github.io/.

[2] Li, X., Song, C., Li, J., Wu, Z., Jia, J., & Meng, H. (2021). Towards Multi-Scale Style Control for Expressive Speech Synthesis. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *5*, 3461–3465. https://doi.org/10.21437/Interspeech.2021-947

[3] Li, Y. A., Han, C., Jiang, X., & Mesgarani, N. (2023). Phoneme-Level BERT for Enhanced Prosody of Text-to-Speech with Grapheme Predictions. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2023-June*. https://doi.org/10.1109/ICASSP49357.2023.10097074

[4] Ogun, S., Owodunni, A. T., Olatunji, T., Alese, E., Oladimeji, B., Afonja, T., Olaleye, K., Etori, N. A., & Adewumi, T. (2024). *1000 African Voices: Advancing inclusive multi-speaker multi-accent speech synthesis*. https://arxiv.org/pdf/2406.11727

[5] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., & Kudinov, M. (2021). Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. *Proceedings of Machine Learning Research*, *139*, 8599–8608. https://arxiv.org/pdf/2105.06337

[6] Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). *Robust Speech Recognition via Large-Scale Weak Supervision* (pp. 28492–28518). PMLR. https://proceedings.mlr.press/v202/radford23a.html