IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14204

Dialect Harmonization Using Text-To-Speech-Audio Technology

Oryina K. Akputu[1*], Msugh Ortil[2], Koko G. Twaki[2], Ikechukwu J. Onubogu[1]

Department of Computing Sciences, Admiralty University of Nigeria, Ibusa, Nigeria[1]

Department of Computer Science, FCT College of Education, Zuba, Nigeria[2]

Abstract: Text-to-Speech (TTS) systems for low-resource languages like Igbo face significant challenges due to dialectal diversity. This research presents a dual-language TTS system for English and Igbo, specifically designed to harmonize the diverse pronunciations and linguistic features across different Igbo dialects. Leveraging a custom model trained on a curated dataset, the system aims to generate natural-sounding speech for both languages. The system is implemented as a web application, providing a user-friendly interface with features like pitch and rate adjustment for English. The system's performance is evaluated using Word Error Rate (WER) on diverse sentences, demonstrating its ability to handle various linguistic complexities within Igbo. This research contributes to enhancing accessibility for Igbo speakers, promoting language preservation, and advancing TTS technology for low-resource languages.

Keywords: Igbo TTS, Dialectal TTS, Low-Resource Language TTS, Multilingual TTS, TTS for Igbo.

I. INTRODUCTION

Text-to-Speech (TTS) systems have become indispensable in modern communication, enabling text-to-audio conversion for applications ranging from assistive technologies for the visually impaired[1], [2] to educational tools and language learning platforms[3], [4]. While advancements in TTS have led to significant improvements in speech quality and naturalness, challenges remain in effectively synthesizing speech for languages with high dialectal diversity, such as Igbo. The Igbo, a major language spoken in southeastern Nigeria, exhibits significant variation across its numerous dialects. These variations encompass phonological, lexical, and morphological differences, impacting pronunciation and posing challenges for accurate speech synthesis. Existing TTS systems for Igbo often struggle to capture these nuances, leading to unnatural or inaccurate speech output.

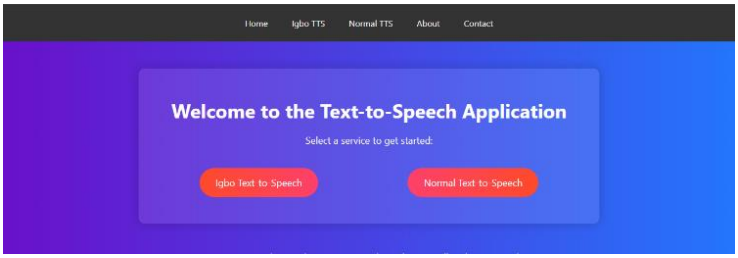

Figure 1. Screen shot of the TTS System

This paper addresses this challenge by developing a dual-language TTS system that supports both English and Igbo, with a specific focus on harmonizing the diverse dialects of Igbo. Figure 1 reflects a screen shot of the system. The system leverages a combination of advanced machine learning techniques and a carefully curated dataset to generate high-quality speech that accurately reflects the linguistic characteristics of Igbo. By addressing the limitations of existing systems, this research aims to improve the accessibility of written content for Igbo speakers, promote the preservation of the language, and contribute significantly to the field of TTS for low-resource languages.

II. RELATED WORKS

There are some existing works (e.g.) that could form the foundation of existing work in Text-to-Speech (TTS) systems. Early research focused on rule-based approaches, where handcrafted rules governed the conversion of text to speech [5], [6]. While these systems provided a foundational understanding of speech synthesis, they often lacked the flexibility and naturalness of more recent data-driven approaches.

The advent of deep learning has revolutionized TTS, leading to the development of powerful neural network-based models. [7], for example, demonstrated remarkable progress in generating high-quality audio, but its sample-level self-

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14204

correction mechanism can lead to slow inference times. DeepVoice [8] introduced a fully neural approach, replacing traditional components with neural networks, but it requires separate training for each component, increasing complexity.

Char2Wav [3] further advanced the field by introducing an end-to-end model that directly generates speech from character inputs. However, it still relies on a separate vocoder for the final audio generation. Tacotron [4] significantly simplified this process by directly predicting raw spectrograms, enabling more efficient and direct speech synthesis.

Commercial TTS systems, such as IBM Watson Text-to-Speech [5] and Google Text-to-Speech [6], have achieved high-quality speech synthesis and are widely used in various applications. These systems often incorporate advanced techniques, such as WaveNet-based models, to enhance naturalness and clarity. However, they may have limitations in handling the diverse linguistic needs of low-resource languages, particularly those with significant dialectal variations.

This work distinguishes itself from the previous studies in three aspects. First, it focuses on dialectal variation of the Igbo, thus addressing the unique challenges posed by the diverse dialects, a critical aspect often overlooked in TTS research for low-resource languages. Secondly, it aims to promote inclusivity for all Igbo speakers by developing a unified speech model that can generate speech that represents the various dialects of Igbo within a single system. Thirdly, its user-friendly and customized features prioritize accessibility for persons with visual impairments, reading disabilities and language learners.

III. METHODOLOGY

This paper follows a data-driven methodology to develop the suggested TTS system for Igbo, addressing the challenges of dialectal diversity. Contextually, a data-driven approach is suitable for the following three reasons: first, recent advances in TTS have been majorly data-driven and proven effective in generating high-quality, natural-sounding speech, surpassing traditional rule-based systems in terms of both accuracy and naturalness [9], [10] Secondly, the adaptability and flexibility behaviour of data-driven models noted in [11] is in fact crucial for addressing the evolving needs of the Igbo-speaking community and ensuring that the TTS system remains relevant and effective over time. The third reasons for choosing the data-driven approach methodology is due to their ability to allow the model to learn the intricate nuances and variations across different Igbo dialects directly from the data. The methodology encompasses key stages, beginning from data collection and preparation to model development and implementation.

A. Data Collection and Preparation

Data collection for low-resource languages like Igbo presents significant challenges. Building upon the work of [12], [13], we focused on collecting data from three major dialects within the Delta zone: Enuani, Ukwuani, and Ika. Data sources included YouTube videos, audio recordings from market interactions, call center dialogues, and religious services. The aim of diversifying the data collection sources is to capture the natural variability and nuances of the spoken Igbo across different contexts and speaking styles, following best practices outlined in [14], [15].To ensure data quality and consistency, several steps were undertaken. First is data Cleaning, where the audio recordings were carefully cleaned to remove noise, background interference, and any extraneous sounds that could negatively impact the training process. This step aligns with the recommendations of [16], [17] to ensure data quality and improve model performance. Next, transcriptions were created for the audio recordings ensuring high fidelity in representing the spoken language. Further on, the transcript audio and their corresponding text were carefully aligned to ensure accurate match between the acoustic and linguistic features domains. This important step, as noted in [18], [19], allowing model to learn the complex mapping between spoken and written language there by impacting on their performance and robustness. Thus, several techniques, including forced alignment and dynamic time warping, were explored to achieve optimal alignment accuracy.

The resulting cleaned and aligned data (refers to as dataset1) from the three dialects were then combined to create a comprehensive dataset for training the TTS model. This approach aimed to capture the inherent variability within the Igbo language while also addressing the challenges of limited data availability for individual dialects.

B. Model Development

The paper employed Tacotron [20] for the Igbo model. The Tacotron is a widely recognized and effective model for end-to-end text-to-speech synthesis. It is the suitable model in context due to its end-to-end nature [21], ability to generate high-quality spectrograms and flexibility to allow for some customization and adaptation. The Tacotron 2 model was fine-tuned on the curated Igbo dataset, incorporating features such as data augmentation [22] and regularization to improve model performance and generalization capabilities. These techniques have been shown to enhance the robustness and generalization ability of TTS models in previous research. For the English language, we leveraged a pre-trained model, providing a strong foundation for high-quality speech synthesis. This pre-trained model was further enhanced by pitch and rate adjustment to allowing users to customize the synthesized speech to their preferences.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14204

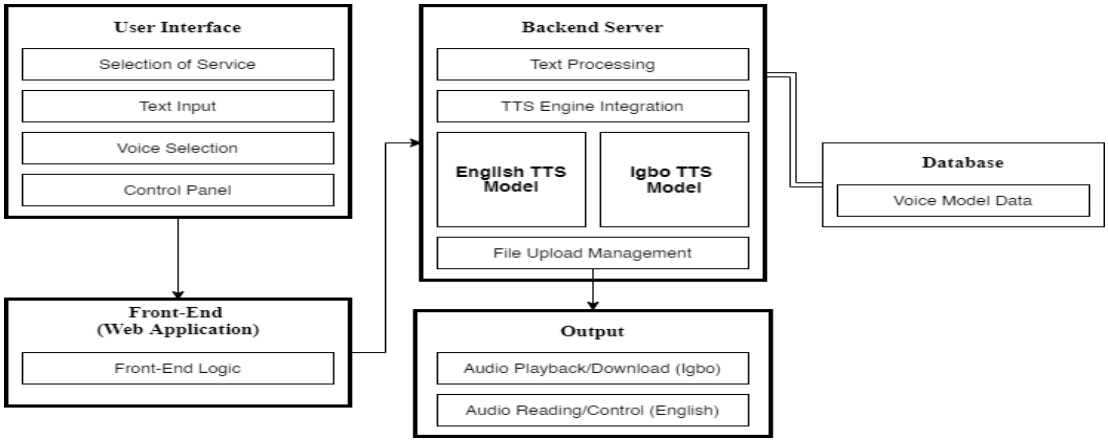

Figure 2. Block diagram of the system.

C. System Implementation

The TTS system was implemented as a web application, aimed to provide a more user-friendly interface for seamless interaction. Figure 2 reflects the block diagram of the system.The system has a front-end implemented using core web technologies viz, HTML, CSS, and JavaScript. Notably the front-end provides a visually appealing and intuitive interface for text input, language selection, and output control. There is also a back end which developed using Python programming allowing core functionalities such as text processing, model interaction, audio generation, and user input or output management.

IV. EXPERIMENT AND EVALUATION

A comprehensive evaluation framework was designed to evaluate the performance the system. The framework utilizes Word Error Rate (WER), a primary metric widely adopted in speech recognition domain. The WER metric q measures the percentage of words incorrectly transcribed compared to the original text. WER provides a quantitative measure of the system's accuracy in generating speech that accurately reflects the input text.

A. Evaluation Datasets

The paper uses the two distinct datasets (dataset 1 and dataset 2) for the evaluation.

The Dataset 1, discussed in section 3.1 is curated and specifically created for this research, encompassing a diverse range of text styles, including sentences with heavy use of Igbo alphabets, simpler sentences, and a mix of both.

The Dataset 2 on the other hand is a second and traditional Igbo dataset sourced from Common Voice[23], [24]. Both datasets were processed through further steps noted earlier in section 3.1, providing an independent benchmark to evaluate the system's performance on unseen data.

B. Evaluation Procedure

The paper selected a set of sentences from each dataset as input for the TTS system. Based on these inputs, the system generated audio output for each input sentence. The generated audio was then processed using an external Automatic Speech Recognition (ASR) system specifically trained for Igbo. The result (the transcribed text) from the ASR system was compared to the original input text, followed by calculating the WER using standard techniques. This mode of evaluation is aimed to mimic real-world usage scenarios and provide a more realistic assessment of the system's performance in converting text to speech and its subsequent intelligibility.

V. RESULTS AND DISCUSSION

Table 1 reflect the WER results across different sentences and the datasets. Specifically, on Dataset 1, the system achieved an average WER of 26.19%, with significant variation across sentences. Sentences with complex linguistic structures, such as those featuring heavy use of Igbo alphabets and intricate grammatical constructions, exhibited higher error rates. On the Dataset 2, the system achieved an average WER of 25.76%, showcasing similar trends in performance variability. The results on this independent dataset provided valuable insights into the system's generalizability and its ability to handle unseen data. It is important to note various factors likely to contribute to the observed WER variations. One of these the presence of diverse dialects within the dataset which is the posed challenge to study behavior of the model, as it needed to adapt to different pronunciation patterns and linguistic nuances. Another factor is sentence complexity as more complex sentences, with intricate grammatical structures and multiple clauses, presented greater challenges for both

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025
DOI: 10.17148/IJARCCE.2025.14204

the TTS system and the ASR system. Other factors hinge on variations speaking characteristics, recording conditions, and environmental noise which possibly impact the accuracy of the system.

Table 1.The WER between the original input sentences and the transcribed text for the two datasets

Dataset	Original Sentence	Transcribed Sentence	WER (%)
Dataset 1	Onye nwoke amuru n'ala bekee dere ya.	Onye nwoke amụrụ n'ala Bekee dere ya.	28.57
	iri anọ na ise	iri anọ na ise	0
	oge ndị isi ọchịchị steeti ahụ nwere nzukọ.	Oge ndị isi ọchịchị steeti ha hu nwere nzuko.	50
	Average WER		26.19
Dataset 2	Owereree ha ụmụaka nwaanyị abụọ ahụ.	Owerere ha ụmụaka nwaanyi abụọ ahu.	50
	maọbụ gbanwee ákwà ya wee ruo abalịwe iri abụọ na asatọ.	maọbụ gbanwee ákwà ya wee ruo abaliwe iri abuo na asatọ.	27.27
	Martina Rabi lagoro mmụọ	Martina Rabi lagoro mmụọ.	0
	Average WER		25.76

VI. CONCLUSION

This paper has presented evaluation results on how the proposed TTS system demonstrated strengths in handling simpler sentences and achieving high accuracy in certain cases. This represents remarkable achievement in the attempt to design the TTS that can harmonize the diverse pronunciations and linguistic features across different Igbo dialects. However, the results also highlighted limitations in processing complex sentences and accurately capturing the nuances of all Igbo dialects. By addressing this limitation and incorporating the insights gained from this evaluation, further research efforts can further enhance the accuracy and naturalness of the TTS system, making it a more effective and valuable tool for the Igbo-speaking community.

REFERENCES

[1] Y. Galphat, B. Vaswani, C. Gangwani, and S. Dhekale, "EmoSpeak: An Emotionally Intelligent Text-to-Speech System for Visually Impaired," *International Conference on Advancements in Power, Communication and Intelligent Systems, APCI 2024*, 2024, doi: 10.1109/APCI61480.2024.10616666.
[2] M. Prabha, P. Saraswathi, J. Hailly, C. Sindhuja, and P. Udhaya, "Smart Glasses: A Visual Assistant for Visually Impaired," *International Conference on Emerging Trends in Engineering and Technology, ICETET*, vol. 2023-April, 2023, doi: 10.1109/ICETET-SIP58143.2023.10151485.
[3] M. Kambouri, H. Simon, and G. Brooks, "Using speech-to-text technology to empower young writers with special educational needs," *Res Dev Disabil*, vol. 135, p. 104466, Apr. 2023, doi: 10.1016/J.RIDD.2023.104466.
[4] A. Abduwali, P. Ghoji, and M. Husiyin, "Chinese Spoken Language Training System Based on Human-Computer Interaction Technology," *Procedia Comput Sci*, vol. 247, pp. 366–373, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.043.
[5] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A Survey on Neural Speech Synthesis," Jun. 2021, Accessed: Dec. 31, 2024. [Online]. Available: https://arxiv.org/abs/2106.15561v3
[6] Y. Ning, S. He, Z. Wu, C. Xing, and L. J. Zhang, "A Review of Deep Learning Based Speech Synthesis," *Applied Sciences 2019, Vol. 9, Page 4050*, vol. 9, no. 19, p. 4050, Sep. 2019, doi: 10.3390/APP9194050.
[7] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," Sep. 2016, Accessed: Dec. 31, 2024. [Online]. Available: http://arxiv.org/abs/1609.03499
[8] S. O. Arik *et al.*, "Deep Voice 2: Multi-Speaker Neural Text-to-Speech," *Adv Neural Inf Process Syst*, vol. 2017-December, pp. 2963–2971, May 2017, Accessed: Dec. 31, 2024. [Online]. Available: https://arxiv.org/abs/1705.08947v2
[9] Y. Kumar, A. Koul, and C. Singh, "A deep learning approaches in text-to-speech system: a systematic review and recent research perspective," *Multimed Tools Appl*, vol. 82, no. 10, pp. 15171–15197, Apr. 2023, doi: 10.1007/S11042-022-13943-4/METRICS.
[10] N. Oralbayeva, A. Aly, A. Sandygulova, and T. Belpaeme, "Data-driven Communicative Behaviour Generation: A Survey," *ACM Trans Hum Robot Interact*, vol. 13, no. 1, Jan. 2024, doi: 10.1145/3609235/SUPPL_FILE/3613529.SUPP.MP4.
[11] P. Liu, L. Wang, R. Ranjan, G. He, and L. Zhao, "A Survey on Active Deep Learning: From Model Driven to Data Driven," *ACM Computing Surveys (CSUR)*, vol. 54, no. 10, Sep. 2022, doi: 10.1145/3510414.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14204

[12] A. Magueresse, V. Carles, and E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges," Jun. 2020, Accessed: Jan. 01, 2025. [Online]. Available: https://arxiv.org/abs/2006.07264v1

[13] T. Reitmaier *et al.*, "Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers," *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2022, doi: 10.1145/3491102.3517639/SUPPL_FILE/3491102.3517639-TALK-VIDEO.MP4.

[14] G. Friedland, "Data Collection and Preparation," *Information-Driven Machine Learning*, pp. 147–170, 2024, doi: 10.1007/978-3-031-39477-5_11.

[15] S. E. Whang, Y. Roh, H. Song, and J. G. Lee, "Data collection and quality challenges in deep learning: a data-centric AI perspective," *VLDB Journal*, vol. 32, no. 4, pp. 791–813, Jul. 2023, doi: 10.1007/S00778-022-00775-9/METRICS.

[16] P. O. Côté, A. Nikanjam, N. Ahmed, D. Humeniuk, and F. Khomh, "Data cleaning and machine learning: a systematic literature review," *Automated Software Engineering*, vol. 31, no. 2, pp. 1–75, Nov. 2024, doi: 10.1007/S10515-024-00453-W/METRICS.

[17] O. V. Girfanov and A. G. Shishkin, "Speech Enhancement with Generative Diffusion Models," *Automatic Documentation and Mathematical Linguistics 2023 57:5*, vol. 57, no. 5, pp. 249–257, Nov. 2023, doi: 10.3103/S0005105523050035.

[18] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, "ONE TTS ALIGNMENT TO RULE THEM ALL," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 6092–6096. doi: 10.1109/ICASSP43922.2022.9747707.

[19] B. Shridhar and B. M, "Autoregressive Speech-To-Text Alignment is a Critical Component of Neural Text-To-Speech (TTS) Models," *Int J Sci Res Sci Eng Technol*, vol. 9, no. 6, pp. 310–316, Dec. 2022, doi: 10.32628/IJSRSET229643.

[20] Y. Wang *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, pp. 4006–4010, Mar. 2017, doi: 10.21437/Interspeech.2017-1452.

[21] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-End Adversarial Text-to-Speech," *ICLR 2021 - 9th International Conference on Learning Representations*, Jun. 2020, Accessed: Dec. 31, 2024. [Online]. Available: https://arxiv.org/abs/2006.03575v3

[22] M. Lajszczak *et al.*, "DISTRIBUTION AUGMENTATION FOR LOW-RESOURCE EXPRESSIVE TEXT-TO-SPEECH," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 8307–8311. doi: 10.1109/ICASSP43922.2022.9746291.

[23] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," Dec. 2019, Accessed: Jan. 07, 2025. [Online]. Available: http://arxiv.org/abs/1912.06670

[24] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pp. 4218–4222, Dec. 2019, Accessed: Jan. 07, 2025. [Online]. Available: https://arxiv.org/abs/1912.06670v2

