

Establishing High-Fidelity Tone-Aligned Resources for Tonal Neural Text-to-Speech: A Validated Pipeline for Automatic Tone Annotation and Feature Analysis in Yorùbá

[Title in Malay Language — to be provided by author]

Charity Ojochogwu Egbunu^{1,2*}, Fadhilah Rosdi¹, Abdul Hadi Abd Rahman¹ & Mohammed Faizul Nasrudin¹

¹Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Malaysia

²Department of Computer Science, Joseph Sarwuan University Makurdi 970101, Benue State, Nigeria

*Corresponding author: Charity Ojochogwu Egbunu (e-mail: P119562@siswa.ukm.edu.my)

ABSTRACT

Neural text-to-speech (TTS) systems increasingly rely on prosodic conditioning to produce intelligible and natural speech. For a tonal language such as Yorùbá, this requirement is not merely stylistic because high, mid, and low tones distinguish lexical and grammatical meanings. However, most forced-alignment workflows generate word and phone boundaries without producing a dedicated tone tier, while Yorùbá is not supported by a readily available pretrained Montreal Forced Aligner model. This extended abstract presents a reproducible front-end pipeline to derive time-aligned tone annotations from tone-marked Yorùbá orthography and validate them for downstream pitch-conditioned TTS. A multi-speaker corpus of 6,746 read utterances was processed using a custom tone-aware lexicon, TextGrid post-processing, acoustic normalization, and lightweight predictive modelling. The resulting annotations achieved 100% tone coverage and 100% lexicon consistency across 122,622 aligned vowel tokens. Acoustic validation showed systematic normalized F0 separation for high (+1.46 semitones), mid (−0.19 semitones), and low (−1.03 semitones) tones, while Random Forest classification reached a macro-F1 of approximately 0.724 under grouped evaluation. Ablation results further indicate that dynamic pitch contours and neighbouring tone context contribute more to tone recoverability than mean F0 alone. The contribution is therefore a compact, auditable, and reusable annotation-validation pipeline that can supply high-fidelity training assets for tone-aware ASR and pitch-conditioned TTS in Yorùbá and related under-documented tonal languages.

Keywords: Yorùbá; tone annotation; forced alignment; TextGrid; acoustic phonetics; neural text-to-speech; low-resource languages.

INTRODUCTION

Lexical tone is central to intelligibility in many languages and is especially important in Yorùbá, where high (H), mid (M), and low (L) level tones participate in word meaning and grammatical interpretation. Typological and phonological studies consistently describe tone as a major prosodic system across the world's languages, and Yorùbá is often treated as a prototypical tone language because pitch contrasts are phonemic rather than optional (Hyman 2016; Michaud & Sands 2020; Shittu 2022). In continuous Yorùbá speech, the task is further complicated by downstep, terracing, vowel identity, local context, and speaker-specific pitch range (Adeniyi 2020; Agbeye 2021; Laniran & Clements 2003). These properties mean that a neural TTS system that produces an incorrect pitch contour may generate speech that sounds unnatural or, more seriously, changes the intended lexical meaning.

Recent neural TTS models address the one-to-many mapping between text and acoustics by conditioning generation on prosodic variables such as F0, duration, and energy. FastSpeech 2 and FastPitch, for example, explicitly model pitch and duration to improve controllability and acoustic naturalness (Bak et al. 2021; Łańcucki 2020; Ren et al. 2020). Transformer-based and cross-utterance prosody models also show that pitch contours and durational variation are valuable for robust synthesis (Liu, Jiang & Yang

2024; Zhang et al. 2023). For tonal languages, however, these techniques require reliable tone supervision: the model must know where tone-bearing units occur, what tone each unit carries, and how those categories are realized acoustically.

The main bottleneck is therefore not only the scarcity of Yorùbá speech data, but the scarcity of auditable, time-aligned tone annotations suitable for machine learning. Forced alignment tools such as the Montreal Forced Aligner (MFA) can produce word and phone intervals, but they do not natively output a dedicated tone tier (McAuliffe et al. 2017). This creates a practical gap for Yorùbá TTS research: even when tone-marked transcripts are available, researchers still need a reproducible method for converting orthographic tone into aligned vowel-level or syllable-level training labels. This study addresses this gap by proposing and validating a reproducible pipeline for automatic tone-tier generation and acoustic feature analysis, providing reliable tone-aligned resources for tone-aware neural text-to-speech systems.

RELATED WORK

Prior work confirms that tone information improves speech technology for tonal languages. In ASR, tone-dependent phone modelling, F0-derived features, and explicit tone transcription have improved recognition accuracy in Mandarin, Vietnamese, Bribri, and Yorùbá (Bengono Obiang et al. 2024; Cao et al. 2025; Coto-Solano 2021; Lei et al. 2006; Nguyen et al. 2022). Pitch features designed for speech recognition, including Kaldi-style pitch and probability-of-voicing parameters, have also become widely used because they encode information that segmental features alone may miss (Ghahremani et al. 2014). In TTS, tone and prosody are equally important: Yorùbá synthesis has shown the value of pitch targets for tone realization, while modern neural systems increasingly use pitch predictors, tone embeddings, and prosody conditioning to improve intelligibility and naturalness (Niekerk & Barnard 2013; Ren et al. 2020; Wang, Takaki & Yamagishi 2018; Yang et al. 2024). It is evident that low-resource tonal-language research often depends on custom data pipelines (De Jesus & Nunes 2024; Romanovsky et al. 2021; Wells, Richmond & Lamb 2023). Public and community-oriented corpora such as Common Voice, OpenSLR Yorùbá resources, ÌròyìnSpeech, and YORÜLECT have expanded African-language speech resources, but these resources do not by themselves solve the problem of validated time-aligned tone supervision (Ahia et al. 2024; Ardila et al. 2020; Gutkin et al. 2020; Ogunremi et al. 2023). Existing computational studies on Yorùbá tone recognition and tone restoration have demonstrated the usefulness of acoustic and orthographic information, but many remain limited to smaller datasets, isolated classification tasks, or pipelines that are not fully documented as reusable alignment-to-tone workflows (Adetunmbi, Obe & Iyanda 2016; Bengono Obiang et al. 2024; Oladiipo Asahiah, Ajadi Odejobi & Rotimi Adagunodo 2017; Sosimi, Adegbola & Fakinlede 2019). The current work therefore positions its contribution at the infrastructure level: it does not propose a new TTS architecture, but supplies the validated tone-aligned corpus and feature evidence required for such architectures to learn tonal realization more reliably.

METHODOLOGY

The proposed research constructs a lightweight, reproducible pipeline divided into two interconnected modules: alignment tier augmentation and acoustic-predictive validation, as seen in Figure 1. The proposed pipeline consists of four connected stages: corpus preparation, tone-aware lexicon construction, forced alignment with tier augmentation, and acoustic-predictive validation. The corpus contains 6,746 read Yorùbá utterances drawn from publicly available speech resources, including Lagos NWU/OpenSLR materials and a cleaned Common Voice subset. The audio was standardized as 16 kHz mono WAV, while transcripts were normalized to preserve Yorùbá orthographic tone marks and characters such as ẹ, ọ, and ẹ.

A custom lexicon mapped each word to a phone sequence in which tone-bearing vowels were marked with `_H`, `_M`, or `_L` according to the orthography. MFA was then used to generate word and phone boundaries. Because MFA does not directly create tone tiers, a TextGrid post-processor inserted two synchronized tiers: a Tone tier assigning H/M/L labels to vowel-nucleus intervals and a `phones_tone` tier

combining vowel identity and tone tag. Intrinsic checks confirmed tone coverage and lexicon consistency.

For acoustic validation, F0 values were extracted using Praat/Parselmouth and normalized into semitones relative to each file's median F0, reducing cross-speaker pitch-range differences (Boersma 2001; Jadoul, Thompson & De Boer 2018). The feature matrix included mean F0, dispersion measures, start-mid-end windows, five-point contour landmarks, slope, curvature, duration, intensity, vowel identity, and neighbouring tone context. Logistic Regression and Random Forest classifiers were then evaluated using grouped splits to reduce file-level leakage.

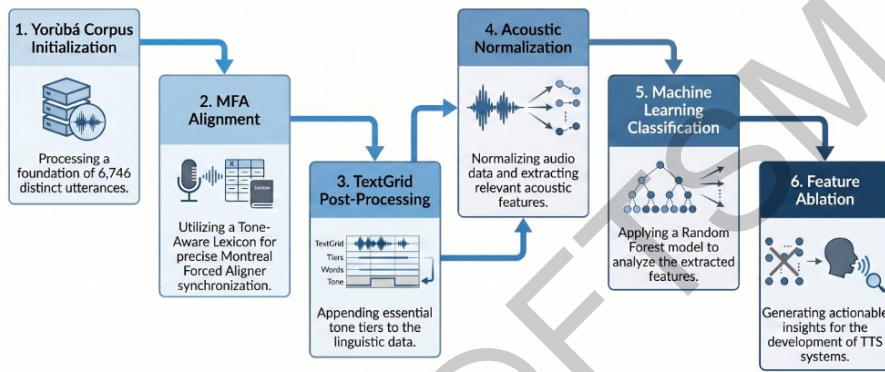


FIGURE 1. Automated Annotation and Feature Engineering Pipeline.

RESULTS AND DISCUSSION

The pipeline produced complete intrinsic supervision. All 122,622 aligned vowel tokens received tone labels, and the final lexicon achieved 100% consistency with the aligned tone sequence after dictionary rebuilding. This confirms that tone-marked orthography can be converted into reliable machine-readable tone tiers when alignment and post-processing are carefully controlled.

The acoustic results also support the validity of the generated labels. Speaker/file-normalized F0 values separated the three tone classes in the expected direction: high tones were above the normalized baseline (+1.46 semitones, 95% CI [1.42, 1.49]), mid tones were close to baseline (−0.19 semitones, 95% CI [−0.21, −0.16]), and low tones were below baseline (−1.03 semitones, 95% CI [−1.07, −0.99]). Effect sizes were stronger for H-M separation ($d = 0.59$) than for M-L separation ($d = 0.29$), indicating that mid and low tones are more acoustically confusable in continuous speech.

Predictive validation showed that tone labels were recoverable from lightweight acoustic and contextual features. A Random Forest model obtained approximately 0.733 accuracy and 0.723 macro-F1 on a grouped 80/20 hold-out split, with five-fold GroupKFold macro-F1 of 0.724 ± 0.003 and grouped-bootstrap 95% CI [0.722, 0.727]. Feature ablation indicated that mean F0 alone is insufficient for robust modelling. The largest gains came from dynamic contour landmarks and contextual cues, showing that TTS front-ends should preserve contour shape, duration, vowel identity, and neighbouring tonal environment. A summary of the outputs is displayed in Table 1.

The practical value of this work lies in its direct compatibility with tone-aware neural TTS development. The phones_tone tier can serve as a discrete linguistic input that allows a text encoder to distinguish tone-bearing vowels before acoustic generation. The Tone tier can support syllable- or vowel-level supervision for pitch prediction, duration modelling, and tonal error analysis. The normalized semitone targets also provide interpretable relative pitch anchors for multi-speaker synthesis, while the ablation evidence identifies which features should be prioritized in future prosody modules.

TABLE 1. Key Quantitative Outputs.

Validation dimension	Output
Intrinsic validity	100% tone coverage and 100% lexicon consistency over 122,622 vowel tokens
Acoustic validity	H = +1.46 st; M = -0.19 st; L = -1.03 st after file-level normalization
Predictive validity	Random Forest macro-F1 ≈ 0.724 under grouped evaluation
Feature insight	Dynamic contours and contextual cues add more predictive value than mean F0 alone

This makes the proposed pipeline useful as a front-end training asset rather than merely an annotation exercise. It supplies ground truth for pitch-conditioned models, supports objective evaluation of tonal fidelity, and provides a reproducible pathway for other low-resource tonal languages where orthography encodes tone but alignment tools do not directly expose it. Figure 2 demonstrates the connection with neural TTS.

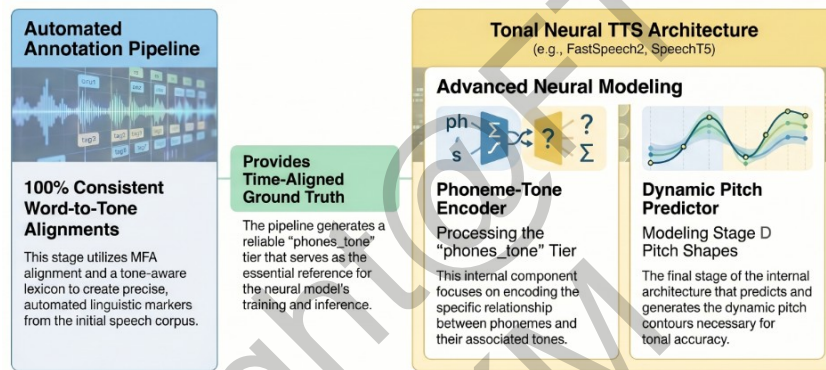


FIGURE 2. Tonal Neural TTS Integration Architecture.

CONCLUSION

This extended abstract presents a validated pipeline for converting tone-marked Yorùbá orthography into time-aligned tone annotations suitable for neural speech technology. The pipeline achieved complete coverage and lexicon consistency across a multi-speaker corpus, produced acoustically meaningful H/M/L pitch distributions, and demonstrated that dynamic contours and context are central to computational tone recoverability. The main contribution is a reusable tone-annotation and validation framework that can strengthen the front-end data foundation for tone-aware ASR and pitch-conditioned Yorùbá TTS. Future work will integrate these tone-aligned resources into neural TTS training and evaluate whether explicit tone supervision reduces tonal substitution, improves perceived intelligibility, and enhances naturalness in synthesized Yorùbá speech. The findings demonstrate that reliable tone supervision can be derived automatically from tone-marked Yorùbá orthography and that dynamic contour information provides a stronger foundation for tone-aware speech modelling than static pitch statistics alone.

REFERENCES

- Adeniyi, K. 2020. Lexicalisation of tonal downstep in Yoruba, *Canadian Journal of Linguistics/Revue canadienne de linguistique* 65, No. 4 (2020): 535–555.
- Adetunmbi, O. A., Obe, O. O., & Iyanda, J. N. 2016. Development of Standard Yorùbá speech-to-text system using HTK, *International Journal of Speech Technology* 19, No. 4 (2016): 929–944.
- Agbeye, M. O. 2021. An autosegmental analysis of tone usage among Yoruba-speaking CVA-aphasic adults, *Path of Science* 7, No. 9 (2021): 2001–2013.

- Ahia, O., Aremu, A., Abagyan, D., Gonen, H., Adelani, D. I., Abolade, D., Smith, N. A., & Tsvetkov, Y. 2024. Voices unheard: NLP resources and models for Yorùbá regional dialects, *arXiv preprint*.
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. 2020. Common Voice: A massively-multilingual speech corpus, *arXiv preprint*.
- Bak, T., Bae, J.-S., Bae, H., Kim, Y.-I., & Cho, H.-Y. 2021. FastPitchFormant: Source-filter based decomposed modeling for speech synthesis, *Interspeech 2021* (2021): 116–120.
- Bengono Obiang, S. G. B., Tsopze, N., Melatagia Yonta, P., Bonastre, J.-F., & Jiménez, T. 2024. Improving tone recognition performance using Wav2vec 2.0-based learned representation in Yoruba, a low-resourced language, *ACM Transactions on Asian and Low-Resource Language Information Processing* 23, No. 12 (2024): 1–11.
- Boersma, P. 2001. Praat, a system for doing phonetics by computer, *Glott International* 5, No. 9 (2001): 341–345.
- Cao, Y., Yu, Y., Li, Y., & Wang, L. 2025. AMNet: An acoustic model network for enhanced Mandarin speech synthesis, *arXiv preprint*.
- Coto-Solano, R. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri, *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas* (2021): 173–184.
- De Jesus, G., & Nunes, S. S. 2024. Data collection pipeline for low-resource languages: A case study on constructing a Tetun text corpus, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (2024): 4368–4380.
- Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., & Khudanpur, S. 2014. A pitch extraction algorithm tuned for automatic speech recognition, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014): 2494–2498.
- Gutkin, A., Demirşahin, I., Kjartansson, O., Rivera, C., & Túbòsún, K. 2020. Developing an open-source corpus of Yoruba speech, *Interspeech 2020* (2020): 404–408.
- Hyman, L. 2016. Lexical vs. grammatical tone: Sorting out the differences, *5th International Symposium on Tonal Aspects of Languages (TAL 2016)* (2016): 6–11.
- Jadoul, Y., Thompson, B., & De Boer, B. 2018. Introducing Parselmouth: A Python interface to Praat, *Journal of Phonetics* 71 (2018): 1–15.
- Łańcucki, A. 2020. FastPitch: Parallel text-to-speech with pitch prediction, *arXiv preprint*.
- Laniran, Y. O., & Clements, G. N. 2003. Downstep and high raising: Interacting factors in Yoruba tone production, *Journal of Phonetics* 31, No. 2 (2003): 203–250.
- Lei, X., Siu, M., Hwang, M.-Y., Ostendorf, M., & Lee, T. 2006. Improved tone modeling for Mandarin broadcast news speech recognition, *Interspeech 2006* (2006).
- Liu, M., Jiang, R., & Yang, H. 2024. Using transfer learning to realize low resource Dungan language speech synthesis, *Applied Sciences* 14, No. 14 (2024): 6336.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi, *Interspeech 2017* (2017): 498–502.

- Michaud, A., & Sands, B. 2020. Tonogenesis, *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.
- Nguyen, T. D., Lam, Q. T., Do, D. H., Cai, H. T., Nguyen, H. S., Vo, T. H., & Nguyen, D. D. 2022. A linguistic-based transfer learning approach for low-resource Bahnar text-to-speech, *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)* (2022): 148–153.
- Niekerk, D. R. V., & Barnard, E. 2013. Generating fundamental frequency contours for speech synthesis in Yorùbá, *Interspeech 2013* (2013): 1027–1031.
- Ogunremi, T., Tubosun, K., Aremu, A., Orife, I., & Adelani, D. I. 2023. ÌròyìnSpeech: A multi-purpose Yorùbá speech corpus, *arXiv preprint*.
- Oladiipo Asahiah, F., Ajadi Odejebi, O., & Rotimi Adagunodo, E. 2017. Restoring tone-marks in Standard Yorùbá electronic text: Improved model, *Computer Science* 18, No. 3 (2017): 305.
- Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. 2020. FastSpeech 2: Fast and high-quality end-to-end text to speech, *arXiv preprint*.
- Romanovskyi, O., Iosifov, I., Iosifova, O., Sokolov, V., Kipchuk, F., & Sukaylo, I. 2021. Automated pipeline for training dataset creation from unlabeled audios for automatic speech recognition, in Z. Hu, S. Petoukhov, I. Dychka, & M. He (Eds.), *Advances in Computer Science for Engineering and Education IV* (2021): 25–36. Springer International Publishing: Cham.
- Shittu, S. 2022. Intrinsic fundamental frequency (IF0) and tonal gradient in Yorùbá, *International Journal of Linguistics, Literature and Translation* 5, No. 11 (2022): 46–53.
- Sosimi, A. A., Adegbola, T., & Fakinlede, O. A. 2019. Standard Yorùbá context dependent tone identification using Multi-Class Support Vector Machine (MSVM), *Journal of Applied Sciences and Environmental Management* 23, No. 5 (2019): 895.
- Wang, X., Takaki, S., & Yamagishi, J. 2018. Autoregressive neural F0 model for statistical parametric speech synthesis, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, No. 8 (2018): 1406–1419.
- Wells, D., Richmond, K., & Lamb, W. 2023. A low-resource pipeline for text-to-speech from found data with application to Scottish Gaelic, *Interspeech 2023* (2023): 4324–4328.
- Yang, D., Liu, S., Huang, R., Weng, C., & Meng, H. 2024. InstructTTS: Modelling expressive TTS in discrete latent space with natural language style prompt, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024): 2913–2925.
- Zhang, Y.-J., Zhang, C., Song, W., Zhang, Z., Wu, Y., & He, X. 2023. Prosody modelling with pre-trained cross-utterance representations for improved speech synthesis, *IEEE/ACM Transactions on Audio Speech and Language Processing* 31 (2023): 2812–2823.