INNOVATIONS IN INFORMATION TECHNOLOGY
https://www.worldlinkpub.com/journals/innovations-in-information-technology/
ISSN: 3141-6004

ACOUSTIC FEATURE OPTIMIZATION FOR TONE CLASSIFICATION UNDER LOW-RESOURCE CONDITIONS: A YORÙBÁ SPEECH STUDY

[*1] Charity O. Egbunu, [2] Gabriel S. Iorundu and [3] Awaziko W. Ambani
[1] Department of Computer Science, Joseph Sarwuan Tarka University, Makurdi
[2] Department of Computer Science, Joseph Sarwuan Tarka University, Makurdi.
3 Department of Computer Science, Joseph Sarwuan Tarka University, Makurdi.
Corresponding Author: charityakowe@gmail.com

Received: 25th May 2026
Accepted for publication: 15 June 2026
Published: 01 July 2026

ABSTRACT

This study presents an optimized acoustic feature selection framework for Yorùbá tone recognition using both machine learning and deep learning approaches. Yorùbá is a low-resource tonal language in which pitch variations carry lexical meaning, making accurate tone modelling essential for speech technologies such as text-to-speech synthesis and automatic speech recognition. However, the absence of large annotated corpora and robust alignment tools continues to limit progress in tone-aware speech processing for the language. To address this challenge, a syllable-level speech corpus was developed and used to extract 63 acoustic features, including Mel-frequency cepstral coefficients (MFCCs), pitch statistics, spectral contrast, chroma features, RMS energy, and harmonics-to-noise ratio (HNR). Three feature optimization techniques; Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Random Forest Feature Importance (RFFI) were investigated to identify the most discriminative tonal features. The selected features were evaluated using Random Forest, Support Vector Machine (SVM), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and CNN–LSTM hybrid classifiers. Experimental results showed that supervised feature selection methods significantly outperformed unsupervised dimensionality reduction. Using only two LDA-selected features, the LSTM model achieved 94.19% classification accuracy, while Random Forest attained the same accuracy using the top 10 RFFI-ranked features. In contrast, PCA-based representations produced comparatively lower and less stable performance across models. The findings demonstrate that compact, tone-sensitive acoustic representations can effectively support Yorùbá tone classification even with limited training data. This study further highlights the importance of pitch-related and harmonic features for tonal discrimination and provides a practical foundation for future tone-aware Yorùbá speech synthesis systems.

Keywords: Acoustic feature optimization, Linear Discriminant Analysis, low-resource speech processing, Random Forest feature selection, speech signal processing, tonal language processing, Yorùbá tone classification.

1.0 INTRODUCTION

Lexical tone is a fundamental linguistic feature in many African and Asian languages, where variations in pitch contribute directly to word meaning and grammatical interpretation (Best, 2019; Connell, 2000; Creel *et al*., 2023; Hyman and Leben, 2017; Mehler *et al*., 2011). In tonal languages such as Yorùbá, tone functions as a phonemic element, meaning that syllables with identical segmental structures may convey entirely different meanings depending on their tonal realization. For example, the Yorùbá syllable *ję* may represent different lexical meanings depending on whether it is produced with High, Mid, or Low tone patterns (Boco and Dagba, 2022; Mehler *et al*., 2011; Niekerk and Barnard, 2013). Accurate tone representation is therefore essential for speech technologies such as automatic speech recognition (ASR), speech synthesis, and text-to-speech (TTS) systems because tonal errors can significantly reduce intelligibility and naturalness in synthesized speech (J. Li and Hasegawa-Johnson, 2022; Niekerk and Barnard, 2013) . Recent advances in deep learning have considerably improved speech processing systems for high-resource languages. Neural architectures such as Tacotron 2, FastSpeech, Transformer-TTS, and diffusion-based speech synthesis models have demonstrated remarkable performance in generating natural and expressive speech (Jeong *et al*., 2021; Pamisetty and Sri Rama Murty, 2023; J. Yang *et al*., 2024). Similarly, tone recognition studies in Mandarin Chinese have achieved near-human classification accuracy using convolutional neural networks (CNNs), bidirectional recurrent neural networks, and self-supervised learning models trained on large annotated corpora (Gao *et al*., 2019; W. Li *et al*., 2019). However, these approaches rely heavily on extensive speech datasets, robust phonetic aligners, and large-scale computational resources that are generally unavailable for most low-resource tonal languages.

Yorùbá, one of the major Niger-Congo languages spoken across West Africa, remains underrepresented in speech technology research despite being spoken by millions of native speakers (Bengono Obiang *et al*., 2024; Niekerk and Barnard, 2013). Existing Yorùbá speech processing studies have primarily focused on tone restoration, speech-to-text systems, and conventional machine learning approaches for isolated tone recognition (Adetunmbi *et al*., 2016; Sosimi *et al*., 2019). Although recent studies have explored the use of self-supervised representations such as wav2vec 2.0 for Yorùbá tone recognition, tonal distinctions are often weakened during representation learning and vector quantization processes, particularly when multilingual models are trained predominantly on

Cite this article: Egbunu, C.O., Iorundu, G.S. and Ambani, A.W. (2026). Acoustic feature optimization for tone classification under low-resource conditions: a Yorùbá speech study. *Innovations in Information Technology,* 1(1):21-28

non-tonal languages (Bengono Obiang *et al*., 2024; Osakuade and King, 2024). This limitation highlights the continued importance of explicit acoustic-prosodic modelling for tonal language processing. Acoustic feature engineering remains one of the most important components of tone classification systems, especially in low-resource environments where training data are limited. Features such as fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCCs), spectral contrast, harmonics-to-noise ratio (HNR), chroma features, and energy-related descriptors have been widely used to capture tonal variations in speech signals. Among these, pitch-related features are generally regarded as the most discriminative because lexical tone is primarily encoded through variations in fundamental frequency (Connell, 2000; Hyman and Leben, 2017). Nevertheless, relying on large unoptimized feature sets may introduce redundancy, increase computational complexity, and reduce model generalization performance, particularly when working with small syllable-level corpora(Lee and Kreiman, 2023). Consequently, attention has increasingly shifted toward feature optimization and dimensionality reduction techniques. Principal Component Analysis (PCA) reduces feature dimensionality by preserving global variance within the dataset, while Linear Discriminant Analysis (LDA) seeks projections that maximize inter-class separability (Afrad *et al*., 2025). Random Forest Feature Importance (RFFI) has also been widely adopted for ranking discriminative acoustic features based on their contribution to classification performance(W. Li *et al*., 2019). Previous studies have shown that supervised dimensionality reduction techniques often outperform unsupervised approaches in classification tasks because they explicitly incorporate class-label information during feature projection. However, comparative studies evaluating these techniques specifically for Yorùbá tone classification remain limited.

Another challenge in Yorùbá speech processing is the limited availability of publicly standardized alignment resources and large-scale syllable-level annotated corpora suitable for tonal modelling and speech synthesis research. Although tools such as the Montreal Forced Aligner (MFA) can be adapted for Yorùbá through custom lexicons and acoustic modelling, the language still lacks widely adopted alignment pipelines and richly annotated benchmark corpora comparable to those available for high-resource languages such as Mandarin and English (Adetunmbi *et al*., 2016; Bengono Obiang *et al*., 2024). This has led many studies to rely on manually segmented or isolated syllable datasets for tone recognition and acoustic-prosodic analysis (ỌDélọBí, 2008; Sosimi *et al*., 2019). Nevertheless, syllable-level analysis remains highly important because lexical tone in Yorùbá is primarily realized at the syllabic level, where variations in pitch contour and harmonic structure contribute directly to lexical distinction (Connell, 2000; Hyman and Leben, 2017).

Tone classification has been extensively investigated in high-resource languages such as Mandarin Chinese, where the availability of large datasets and pretrained models facilitates the development of sophisticated modeling strategies ((Hou and Huang, 2020; Shen *et al*., 2024). Deep learning models, particularly convolutional neural networks (CNNs) applied to mel-spectrogram inputs, have achieved near-perfect accuracy in tone classification of Mandarin syllables, demonstrating significant robustness even in noisy environments. For instance, the ToneNet model developed by Gao *et al*., trained on Mandarin mel-spectrograms, achieved over 99% accuracy and F1-score, underscoring its resilience to background noise and spectral variation(Gao *et al*., 2019). Beyond speech recognition, phonological attributes such as tone motifs and rime structures have been utilized in stylometric analysis. These features have proven highly effective in authorship attribution, illustrating the versatility of tone-based cues across linguistic tasks. By employing tone and rime motifs as feature vectors, classifiers such as SVMs and Random Forests have successfully identified authorship in Chinese literary texts (Hou and Huang, 2020). However, tone classification accuracy often declines when transitioning from isolated syllables to continuous speech. Liu *et al*. demonstrated that Multi-Space Distribution Hidden Markov Models (MSD-HMMs) achieved 88.8% accuracy in continuous Mandarin tone recognition, which is notably lower than performance on isolated syllables. This highlights the importance of context-aware models in real-world tonal processing tasks (Cooper-Leavitt, 2016). Self-supervised learning (SSL) models such as HuBERT and XLS-R can encode tone information even when trained on non-tonal or multilingual corpora. However, when these representations are discretized—typically via k-means clustering—tone-relevant cues are often lost. (Osakuade and King, 2024) showed that even Mandarin-HuBERT embeddings experienced tonal degradation after quantization, raising concerns about tonal fidelity in downstream applications (Osakuade and King, 2024; Shen *et al*., 2024).

Feature selection and optimization are essential for improving model performance, particularly in environments with limited data availability. Principal Component Analysis (PCA) is a traditional unsupervised method that reduces dimensionality by projecting features onto orthogonal directions of maximal variance, thereby simplifying the feature space while preserving structural information (He *et al*., 2013; D. Yang *et al*., 2024). Although PCA does not directly incorporate class-label information, it has been demonstrated to provide significant benefits in tone and emotion classification contexts. For example, (Afrad *et al*., 2025) showed that integrating PCA prior to an Artificial Neural Network (ANN) classifier enhanced text emotion classification performance. In contrast, Linear Discriminant Analysis (LDA) projects features to maximize inter-class variance, making it suitable for tone classification tasks. Random Forest Feature Importance (RFFI) provides a supervised ranking of features based on their impurity reduction, allowing for the selection of high-impact acoustic cues (W. Li *et al*., 2019; J. Yang *et al*., 2024). Several hybrid approaches have also been proposed. For example, (Dasare *et al*., 2023) combined filter-based and wrapper methods to improve tone classification in Urdu and Nigerian Pidgin. (Ekpenyong and Inyang, 2016) employed non-negative matrix factorization and vowel-only features for tonal clustering. In a related context,(Glocker and Georges, n.d.) demonstrated that multitask learning can enhance phoneme and tone recognition in multilingual, low-resource settings.

Despite these advances, comparative investigations of supervised and unsupervised feature optimization strategies for Yorùbá tone classification remain limited. Consequently, there is still a need for interpretable and data-efficient approaches capable of supporting robust tonal classification under low-resource conditions. This study therefore investigates the effectiveness of acoustic feature optimization and comparative machine learning and deep learning approaches for Yorùbá tone recognition, with the broader aim of supporting future tone-aware speech technologies for Yorùbá and related tonal languages.

This study was guided by the following objectives:

1. To identify the most discriminative acoustic features for distinguishing High, Mid, and Low tones in Yorùbá speech.
2. To evaluate the effectiveness of supervised and unsupervised dimensionality reduction techniques for tonal classification.

3. To develop a compact, adaptable, and data-efficient framework for tone classification in low-resource tonal languages.

By combining feature optimization with comparative model evaluation, this study provides empirical insight into compact acoustic representations capable of supporting efficient tonal classification for Yorùbá and related tonal languages.

2.0 Materials and Method

This study adopted a syllable-level tonal classification framework for Yorùbá speech involving corpus preparation, acoustic feature extraction, feature optimization, and comparative evaluation using machine learning and deep learning models (Egbunu *et al*., 2025). Acoustic-prosodic features associated with tonal variation were extracted from Yorùbá speech recordings and subsequently processed using supervised and unsupervised dimensionality reduction techniques. The optimized feature representations were then evaluated across multiple classification models to determine their effectiveness for tonal discrimination under low-resource conditions. The overall workflow adopted in this study is illustrated in Figure 1.

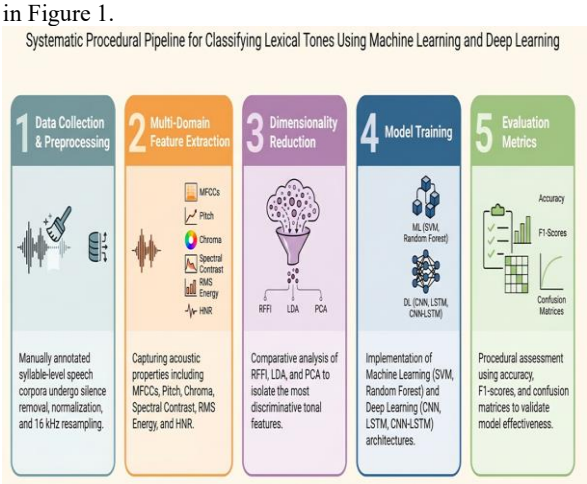

Figure 1: Research Framework

2.1 Speech Corpus and Data Preparation

This study utilized a syllable-level Yorùbá speech corpus adapted from an earlier concatenated Yorùbá speech synthesis framework developed for syllable-based speech generation and acoustic analysis (Egbunu *et al*., 2026). The corpus consisted of 772 manually segmented Yorùbá syllables representing the three lexical tone categories: High, Mid, and Low, with class distributions of 256, 244, and 272 syllables, respectively. Because tonal distinctions in Yorùbá are primarily realized at the syllabic level, manual segmentation and annotation were employed to preserve accurate tonal boundary representation and improve tonal consistency across the dataset. The speech recordings were produced by a native speaker of Yorùbá and underwent several preprocessing procedures prior to feature extraction. Silent regions and background noise at the beginning and end of each recording were removed to improve signal quality, after which amplitude normalization was applied to maintain consistent recording intensity across the corpus. All recordings were subsequently resampled to 16 kHz to ensure uniformity during acoustic analysis. Each syllable was manually verified and annotated according to its corresponding tonal category before the classification experiments were conducted.

2.2 Acoustic Feature Extraction

Acoustic feature extraction was carried out to capture tonal, spectral, harmonic, and energy-related properties associated with Yorùbá speech. A total of 63 acoustic features were extracted from each syllable using standard speech processing techniques. Feature extraction was performed using standard speech signal processing procedures implemented in Python-based audio analysis libraries. The extracted features included Mel-frequency cepstral coefficients (MFCCs), pitch-related statistics, chroma features, spectral contrast descriptors, Root Mean Square (RMS) energy, and harmonics-to-noise ratio (HNR). Pitch-related features were included because lexical tone in Yorùbá is strongly associated with variations in fundamental frequency (F0). MFCCs were used to capture spectral envelope characteristics associated with speech articulation, while spectral contrast and chroma descriptors were included to represent harmonic distribution and spectral variations within the speech signal. RMS energy was utilized to estimate signal intensity, whereas HNR provided information regarding the periodicity and harmonic quality of the speech signal. Together, these features provided a multidimensional representation of tonal and acoustic properties relevant to Yorùbá tone discrimination. The features extracted from each syllable are listed in Table 1.

Table 1: Acoustic Features Extracted

Feature Type	Description
MFCCs (Mel-Frequency Cepstral Coefficients)	Captures timbral and spectral characteristics of speech signals.
Pitch Mean and Std Dev	Represents variations in fundamental frequency (F0) associated with tonal changes.
Chroma Features	Represents harmonic content and pitch-class energy distribution associated with tonal variation.
Spectral Contrast	Measures variations in energy distribution across spectral frequency bands.
RMS Energy	Estimates the overall intensity or loudness of the speech signal.
HNR (Harmonics-to-Noise Ratio)	Measures signal periodicity and differentiates harmonic and non-harmonic speech components.

2.3 Feature Selection and Dimensionality Reduction

To reduce feature redundancy and improve tonal discriminability, three feature optimization and dimensionality reduction techniques were investigated: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Random Forest Feature Importance (RFFI). Principal Component Analysis (PCA) was employed as an unsupervised dimensionality reduction technique to transform the original feature space into a smaller set of orthogonal components while preserving the majority of the variance within the dataset. Components accounting for approximately 95% of the total variance were retained for subsequent classification experiments. Linear Discriminant Analysis (LDA) was utilized as a supervised feature projection method designed to maximize class separability between High, Mid, and Low tonal categories. Unlike PCA, LDA incorporates class-label information during projection, thereby producing feature representations optimized specifically for

classification tasks. Random Forest Feature Importance (RFFI) was applied to rank acoustic features according to their contribution to tonal discrimination. Feature importance scores were computed based on impurity reduction across the ensemble decision trees. The most discriminative features identified through RFFI were subsequently selected for classification experiments. This approach enabled the selection of compact feature subsets while preserving tone-discriminative information.

2.4 Classification Models

Both conventional machine learning and deep learning models were investigated to evaluate the effectiveness of the optimized acoustic representations for Yorùbá tone classification. The machine learning models included Support Vector Machine (SVM) and Random Forest (RF). The SVM classifier employed a Radial Basis Function (RBF) kernel to model nonlinear decision boundaries within the acoustic feature space. Random Forest classification was implemented using an ensemble of decision trees to improve robustness and capture nonlinear feature interactions associated with tonal variation. The deep learning approaches included Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), and a hybrid CNN–LSTM architecture. The CNN model was designed to learn localized acoustic feature patterns associated with tonal distinctions, whereas the LSTM model was utilized to model sequential dependencies within the feature representations. The hybrid CNN–LSTM architecture combined convolutional feature extraction with sequential modelling in an attempt to capture both local and contextual acoustic relationships relevant to tone classification. All models were trained and evaluated using the optimized feature representations generated through PCA, LDA, and RFFI in order to compare the influence of different feature optimization strategies on classification performance.

2.5 Experimental Setup and Evaluation

The experimental setup was designed to ensure reliable evaluation of tonal classification performance across different feature optimization strategies and classification models. Prior to model training, the extracted acoustic features were standardized using z-score normalization to minimize scale differences among the feature dimensions and improve model convergence during training. This preprocessing step ensured that features with larger numerical ranges did not dominate the learning process. The dataset was divided into training, validation, and testing subsets using stratified sampling to preserve the proportional distribution of the three tonal classes (High, Mid, and Low) across all subsets. Approximately 70% of the dataset was allocated for model training, while 15% each were reserved for validation and testing purposes. The validation set was used for hyper-parameter tuning and monitoring model generalization during training, whereas the testing set was used exclusively for final performance evaluation.

To improve experimental reliability and reduce the influence of random initialization effects, each experiment was repeated multiple times using different random seeds, and the average performance was reported. This approach helped ensure that the observed classification results were stable and not dependent on a single random data split or model initialization. For the Random Forest classifier, experiments were conducted using an ensemble of decision trees with optimized parameters selected through validation-based tuning. The Support Vector Machine (SVM) classifier utilized a Radial Basis Function (RBF) kernel because of its effectiveness in handling nonlinear decision boundaries within acoustic feature spaces. For the deep learning experiments, the CNN, LSTM, and CNN–LSTM models were trained using the Adam optimizer and categorical cross-entropy loss function. Early stopping and dropout regularization were applied during training to reduce overfitting, particularly because of the relatively limited size of the syllable-level dataset. Model performance was evaluated using multiple classification metrics, including accuracy, precision, recall, and F1-score. Accuracy was used as the primary evaluation metric because the dataset maintained relatively balanced tonal class distributions. Precision and recall were included to assess the consistency of individual tone predictions, while the F1-score provided a balanced measure of classification performance by combining both precision and recall. In addition to overall classification accuracy, confusion matrix analysis was performed to examine tone-specific classification behaviour and identify dominant confusion patterns among High, Mid, and Low tones. Particular attention was given to Mid–Low tonal confusions because previous studies have shown that these categories often exhibit overlapping acoustic characteristics in continuous speech environments. Comparative evaluations were subsequently conducted across all feature optimization techniques and classification models to determine the most effective acoustic representation strategy for Yorùbá tone recognition under low-resource conditions.

3.0 RESULTS

3.1 Feature Importance Analysis

Feature selection analysis revealed that pitch-related and harmonic features were the most discriminative acoustic representations for Yorùbá tone classification. Using the Random Forest Feature Importance (RFFI) approach, the mean pitch achieved the highest importance score, followed by pitch standard deviation, multiple chroma features, and spectral contrast descriptors, as presented in Figure 2. These findings align with the linguistic characteristics of Yorùbá, where lexical tone is primarily conveyed through variations in fundamental frequency (F0). The dominance of pitch mean confirms that tonal height remains the most important acoustic cue for distinguishing High, Mid, and Low tones. Similarly, the prominence of chroma-based features suggests that harmonic energy distribution also contributes significantly to tonal discrimination. High-tone syllables generally exhibited stronger harmonic concentration within specific chroma bands, whereas low tones displayed comparatively lower harmonic distributions. Spectral contrast features further contributed to tonal separation by capturing differences in energy distribution across frequency bands. In contrast, Mel-frequency cepstral coefficient (MFCC)-based features showed relatively lower importance scores compared with pitch and harmonic descriptors. This observation suggests that broad spectral envelope characteristics contribute less directly to tonal discrimination than pitch-related acoustic information. Overall, the feature importance analysis demonstrated that compact acoustic representations emphasizing pitch and harmonic structure are highly effective for Yorùbá tone classification.

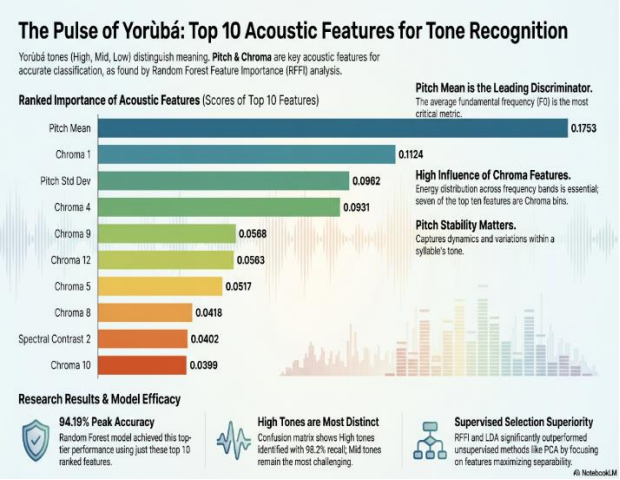

Figure. 1: Top 10 Feature Extraction Scores

3.2 Comparison of Feature Optimization Techniques

The comparative evaluation of the feature optimization techniques demonstrated clear differences between supervised and unsupervised dimensionality reduction approaches. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Random Forest Feature Importance (RFFI) each produced distinct feature representations with varying levels of discriminative effectiveness. PCA retained approximately 95% of the total variance within the original feature space, resulting in 43 principal components. Although this approach successfully reduced feature redundancy, the transformed feature space still contained substantial variance unrelated to tonal distinctions. Consequently, PCA-based representations produced comparatively lower classification performance across several models. In contrast, LDA generated a highly compact two-dimensional feature space optimized specifically for tonal discrimination. Figure 3 illustrates the resulting LDA projection, where High, Mid, and Low tone categories formed relatively distinct clusters with minimal overlap. High-tone syllables occupied a clearly separated region, while Low tones formed another compact cluster. Mid-tone syllables generally appeared between the High and Low tone regions, although slight overlap with Low tones remained observable. This overlap reflects the acoustic similarity often observed between Mid and Low tones in Yorùbá speech.

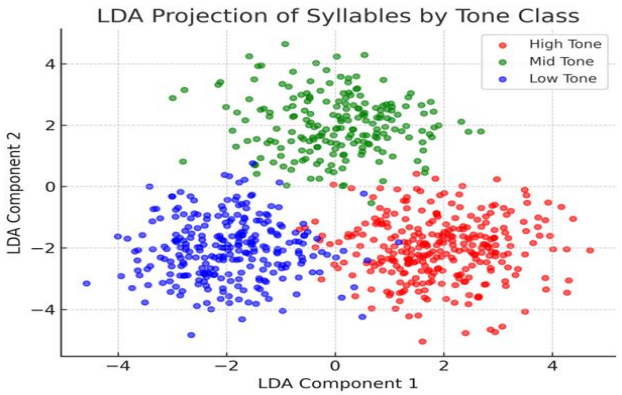

Figure 3: Two-dimensional Linear Discriminant Analysis (LDA) Feature Space

The Yorùbá tone classes are depicted in a simulated scatter plot for illustration.

Unlike PCA, LDA incorporated class-label information during feature projection, thereby maximizing inter-class separability while minimizing intra-class variance. The two most discriminative features identified through LDA were MFCC 1 and Spectral Contrast 7, both of which contributed strongly to tonal separation. The resulting compact representation significantly simplified the classification task and enabled highly consistent performance across both machine learning and deep learning models. The Random Forest Feature Importance approach also produced highly effective feature representations by selecting only the most discriminative acoustic descriptors. The top-ranked features consisted primarily of pitch-related and harmonic features, reinforcing the importance of F0 and harmonic structure in Yorùbá tone modelling. Compared with PCA, both LDA and RFFI demonstrated stronger tonal discriminability and superior classification consistency.

3.3 Comparative Classification Performance

The classification accuracies obtained across the different feature optimization strategies and classification models are presented in Figure 4 and Table 2. The results revealed several important performance trends. Among all evaluated approaches, the Random Forest classifier trained using the RFFI-selected top 10 features achieved one of the highest classification accuracies of 94.19%. This result demonstrates that a compact subset of carefully selected acoustic features can effectively preserve the essential tonal information required for accurate classification. The Support Vector Machine (SVM) also produced strong performance on the RF-selected features, achieving an accuracy of 89.68%. For the LDA-based feature representation, classification performance remained consistently high across nearly all models. The LSTM model achieved the highest overall accuracy of 94.19%, while the Random Forest, CNN, and CNN–LSTM models each achieved approximately 93.55%. The SVM classifier also maintained strong performance with an accuracy of 92.90%. These findings indicate that the supervised LDA projection effectively condensed tone-discriminative information into a highly separable feature space that could be learned efficiently by both classical and deep learning models.

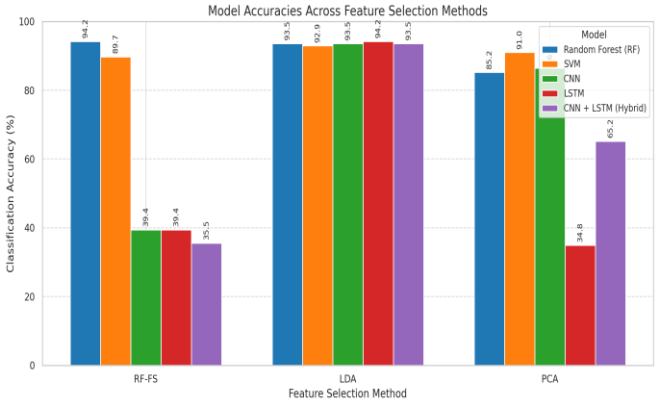

Figure 4: Bar chart showing model accuracies across the different feature selection methods

Table 2: Accuracies of the different Models

Feature Selection Method	Random Forest (RF)	SVM	CNN	LSTM	CNN + LSTM (Hybrid)
Random Feature Selection (Top 10 Features)	94.19	89.68	39.35	39.35	35.48
LDA (2 Features)	93.55	92.9	93.55	94.19	93.55
PCA (Principal Components)	85.16	90.97	86.45	34.84	65.16

In contrast, the PCA-based feature representations produced comparatively lower and less stable classification performance. Although the SVM classifier achieved a relatively strong accuracy of 90.97%, the Random Forest and CNN models produced lower accuracies of 85.16% and 86.45%, respectively. The LSTM model demonstrated particularly weak performance on PCA features, achieving only 34.84% accuracy, while the CNN–LSTM hybrid achieved 65.16%. The comparatively lower performance of the deep learning models on PCA features suggests that the retained principal components did not adequately preserve tone-discriminative information despite capturing substantial global variance. Additionally, the relatively limited size of the syllable-level corpus may have restricted the ability of the deep learning models to learn stable representations from high-dimensional transformed feature spaces. In contrast, the classical machine learning approaches, particularly Random Forest and SVM, demonstrated greater robustness under limited-data conditions. Overall, the results indicate that supervised feature optimization techniques provide more effective tonal representations than unsupervised variance-preserving approaches for Yorùbá tone classification. The findings further demonstrate that compact and interpretable acoustic representations can support highly accurate tonal discrimination even with relatively limited training data.

3.4 Confusion Matrix and Error Analysis

The confusion matrix for the best-performing Random Forest classifier using RF-selected features is presented in Figure 5. The results showed strong classification performance across all tonal categories, although certain confusion patterns remained observable. High-tone syllables achieved the strongest classification performance, with 54 out of 55 syllables correctly identified, corresponding to a recall rate of approximately 98.2%. Only one High-tone syllable was misclassified as Low tone, while no High tones were confused with Mid tones. This result indicates that High tones possess highly distinctive pitch-based acoustic characteristics that can be reliably captured by the optimized feature representations. Low-tone syllables also demonstrated strong performance, with 55 out of 61 syllables correctly classified, corresponding to approximately 90.2% recall. Most of the classification errors for Low tones involved confusion with Mid tones rather than High tones. This suggests that Low and Mid tones exhibit partially overlapping acoustic characteristics, particularly in natural speech conditions where tonal realizations may vary because of speaker articulation and contextual variation.

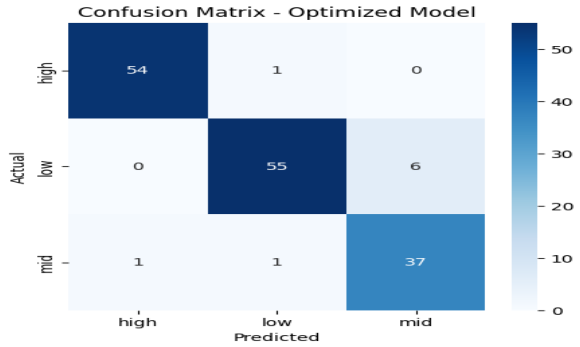

Figure 5: Confusion Matrix for RF - FI Optimized

Overall, the confusion matrix analysis demonstrated that the proposed feature optimization strategies effectively preserved the acoustic distinctions among the three Yorùbá tone categories. The remaining classification errors were dominated primarily by Mid–Low confusions, which is consistent with observations reported in previous tonal speech studies.

4.0 DISCUSSION

The findings of this study demonstrate that acoustic feature optimization plays a critical role in improving tonal classification performance for low-resource tonal languages such as Yorùbá. The results consistently showed that compact and carefully selected acoustic representations were capable of achieving highly accurate tone discrimination across the High, Mid, and Low tonal categories. In particular, the strong performance achieved using only two LDA-derived features and the top 10 Random Forest-selected features indicates that tonal information in Yorùbá can be effectively represented using a relatively small set of discriminative acoustic descriptors.

The first objective of this study was to identify the most discriminative acoustic features for distinguishing Yorùbá tones. The feature importance analysis demonstrated that pitch-related and harmonic descriptors contributed most significantly to tonal discrimination. Mean pitch emerged as the strongest individual feature, followed by pitch standard deviation, chroma features, and spectral contrast descriptors. These findings reinforce the linguistic understanding that lexical tone in Yorùbá is primarily conveyed through variations in fundamental frequency and harmonic structure. The relatively lower contribution of MFCC-based spectral envelope features further suggests that tonal classification depends more strongly on pitch dynamics than on broad spectral characteristics associated with vowel quality. Consequently, the study provides empirical evidence that compact pitch-sensitive feature representations can effectively capture tonal distinctions in Yorùbá speech.

The second objective focused on evaluating supervised and unsupervised dimensionality reduction techniques for tonal classification. The results clearly demonstrated that supervised feature optimization approaches produced superior performance compared with unsupervised variance-preserving methods. Linear Discriminant Analysis (LDA), which explicitly incorporated tonal class information during projection, generated highly separable feature representations that enabled consistently strong classification performance across nearly all models. In contrast, Principal Component Analysis (PCA), despite retaining most of the dataset variance, preserved substantial non-discriminative information

unrelated to tonal separation. This finding highlights an important distinction between variance preservation and class discriminability in tonal speech modelling. The study therefore confirms that supervised dimensionality reduction techniques are more effective for tonal classification tasks where labelled tonal categories are available.

The third objective of this study was to develop a compact and data-efficient framework suitable for tone classification in low-resource tonal languages. The experimental results demonstrated that highly accurate classification performance could be achieved using relatively small feature subsets and limited syllable-level data. Notably, the Random Forest classifier trained on the top 10 RF-selected features and the LSTM model trained on the two-dimensional LDA representation both achieved classification accuracies of 94.19%. These findings indicate that efficient tonal classification does not necessarily require excessively high-dimensional feature spaces or highly complex deep learning architectures. Instead, the quality and discriminative relevance of the selected acoustic features appear to play a more important role in determining classification performance.

Another important observation from this study was the comparatively strong performance of classical machine learning approaches under limited-data conditions. Random Forest and Support Vector Machine (SVM) models consistently demonstrated robust and stable performance across different feature representations, whereas some deep learning architectures showed comparatively lower or unstable convergence behaviour, particularly when trained on PCA-transformed features. This suggests that classical machine learning models may provide more reliable solutions for tonal classification tasks when annotated datasets remain relatively small. The findings therefore challenge the assumption that increasingly complex deep learning architectures will always outperform conventional machine learning approaches in low-resource speech processing environments.

The confusion matrix analysis further provided important insight into the tonal behaviour of Yorùbá speech. High tones exhibited the strongest classification consistency, indicating that their pitch-related acoustic cues are highly distinctive. In contrast, Mid–Low confusions remained the dominant source of classification error across the experiments. This observation is consistent with the acoustic similarity often observed between Mid and Low tones in natural Yorùbá speech production, where tonal boundaries may vary because of contextual articulation and speaker variation. Despite these challenges, the optimized feature representations were still able to preserve strong tonal separability across all three categories.

Overall, this study contributes to the growing body of research on low-resource tonal speech processing by demonstrating the effectiveness of feature optimization and supervised dimensionality reduction for Yorùbá tone classification. The study further contributes a compact and interpretable tonal classification framework capable of supporting efficient acoustic modelling under limited-data conditions. Beyond classification performance, the findings provide important insight into the acoustic properties most relevant for Yorùbá tonal discrimination and establish a practical foundation for future tone-aware speech synthesis and speech recognition systems for Yorùbá and related tonal languages. Despite the strong classification performance achieved in this study, the experiments were conducted using a relatively small syllable-level corpus derived from a controlled single-speaker environment. Future work should therefore investigate the robustness of the proposed framework using larger multi-speaker continuous speech datasets under more diverse acoustic conditions.

5.0 CONCLUSION

This study investigated the effectiveness of acoustic feature optimization techniques for Yorùbá tone classification using both machine learning and deep learning approaches. The findings demonstrated that compact and discriminative acoustic representations can effectively support tonal classification under low-resource conditions. Among the evaluated features, pitch-related and harmonic descriptors emerged as the most important acoustic cues for distinguishing High, Mid, and Low tones, confirming the central role of fundamental frequency and harmonic structure in Yorùbá tonal realization. The comparative evaluation further showed that supervised dimensionality reduction techniques produced superior performance compared with unsupervised variance-preserving approaches. In particular, Linear Discriminant Analysis (LDA) generated highly separable tonal representations that enabled strong and consistent classification performance across multiple models. Similarly, the Random Forest Feature Importance (RFFI) approach demonstrated that a small subset of carefully selected acoustic features could preserve substantial tonal information while reducing feature redundancy and computational complexity.

The study also revealed that classical machine learning models, particularly Random Forest and Support Vector Machine (SVM), remained highly effective under limited-data conditions, often outperforming more complex deep learning architectures. These findings highlight the importance of feature quality and discriminative representation in low-resource tonal speech processing, rather than relying solely on increasing model complexity. Overall, this work contributes a compact and interpretable framework for Yorùbá tone classification and provides empirical insight into the acoustic properties most relevant for tonal discrimination. The findings further establish a practical foundation for future tone-aware speech technologies for Yorùbá and other low-resource tonal languages.

REFERENCES

Adetunmbi, O. A., Obe, O. O., and Iyanda, J. N. (2016). Development of Standard Yorùbá speech-to-text system using HTK. *International Journal of Speech Technology*, *19*(4), 929–944. https://doi.org/10.1007/s10772-016-9380-2

Afrad, M., Muljono, M., and Pujiono, P. (2025). Utilization Of Principal Component Analysis To Improve Emotion Classification Performance In Text Using Artificial Neural Networks. *Journal of Applied Intelligent System*, *9*(1), 8–18. https://doi.org/10.62411/jais.v9i1.9923

Bengono Obiang, S. G. B., Tsopze, N., Melatagia Yonta, P., Bonastre, J.-F., and 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*(12), 1–11. https://doi.org/10.1145/3690384

Best, C. T. (2019). The Diversity of Tone Languages and the Roles of Pitch Variation in Non-tone Languages: Considerations for Tone Perception Research. *Frontiers in Psychology*, *10*, 364. https://doi.org/10.3389/fpsyg.2019.00364

Boco, C. A. C. Y., and Dagba, T. K. (2022). *An End to End Bilingual TTS System for Fongbe and Yoruba*. *1653 CCIS*, 294–304. Scopus. https://doi.org/10.1007/978-3-031-16210-7_24

Connell, B. (2000). The Perception of Lexical Tone in Mambila. *Language and Speech*, *43*(2), 163–182. https://doi.org/10.1177/00238309000430020201

Cooper-Leavitt, J. E. (2016). A computational classification of Thai lexical tones. *Journal of the Acoustical Society of America*, *139*(4_Supplement), 2216–2216. https://doi.org/10.1121/1.4950631

Creel, S. C., Obiri-Yeboah, M., and Rose, S. (2023). Language-to-music transfer effects depend on the tone language: Akan vs. East Asian tone languages. *Memory* and *Cognition*, *51*(7), 1624–1639. https://doi.org/10.3758/s13421-023-01416-4

Dasare, A., Chowdhury, A. R., Menon, A. S., Anand, K., Deepak, K. T., and Prasanna, S. R. M. (2023). Bridging the Gap: Towards Linguistic Resource Development for the Low-Resource Lambani Language. In A. Karpov, K. Samudravijaya, K. T. Deepak, R. M. Hegde, S. S. Agrawal, and S. R. M. Prasanna (Eds.), *Speech and Computer* (Vol. 14339, pp. 127–139). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48312-7_10

Egbunu, C. O., Rosdi, F., Nasrudin, M. F., and Rahman, A. H. A. (2025). Enhancing Naturalness in Text-to-Speech Synthesis: Optimizing TD-PSOLA with Hybrid Pitch Detection and Cross-Fade Technique. *2025 International Conference on Electrical Engineering and Informatics (ICEEI)*, 1–7. https://doi.org/10.1109/ICEEI68459.2025.11331223

Ekpenyong, M. E., and Inyang, U. G. (2016). Unsupervised mining of under-resourced speech corpora for tone features classification. *2016 International Joint Conference on Neural Networks (IJCNN)*, 2374–2381. https://doi.org/10.1109/IJCNN.2016.7727494

Gao, Q., Sun, S., and Yang, Y. (2019). ToneNet: A CNN Model of Tone Classification of Mandarin Chinese. *Interspeech 2019*, 3367–3371. https://doi.org/10.21437/Interspeech.2019-1483

Glocker, K., and Georges, M. (n.d.). *Hierarchical Multi-Task Transformers for Crosslingual Low Resource Phoneme Recognition*.

He, J., Tan, E.-L., and Gan, W.-S. (2013). Time-shifted principal component analysis based cue extraction for stereo audio signals. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 266–270. https://doi.org/10.1109/ICASSP.2013.6637650

Hou, R., and Huang, C.-R. (2020). Robust stylometric analysis and author attribution based on tones and rimes. *Natural Language Engineering*, *26*(1), 49–71. https://doi.org/10.1017/S135132491900010X

Hyman, L. M., and Leben, W. R. (2017). Word prosody II: Tone systems. *UC Berkeley Phonology Lab Annual Reports*, *13*. https://doi.org/10.5070/P7131040752

Jeong, M., Kim, H., Cheon, S. J., Choi, B. J., and Kim, N. S. (2021). Diff-TTS: A Denoising Diffusion Model for Text-to-Speech. *Interspeech 2021*, 3605–3609. https://doi.org/10.21437/Interspeech.2021-469

Lee, Y., and Kreiman, J. (2023). Within- versus between-speaker acoustic variability in Thai. *The Journal of the Acoustical Society of America*, *153*(3_supplement), A295–A295. https://doi.org/10.1121/10.0018911

Li, J., and Hasegawa-Johnson, M. (2022). Autosegmental Neural Nets 2.0: An Extensive Study of Training Synchronous and Asynchronous Phones and Tones for Under-Resourced Tonal Languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1918–1926. https://doi.org/10.1109/TASLP.2022.3178238

Li, W., Chen, N. F., Siniscalchi, S. M., and Lee, C.-H. (2019). Improving Mispronunciation Detection of Mandarin Tones for Non-Native Learners With Soft-Target Tone Labels and BLSTM-Based Deep Tone Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2012–2024. https://doi.org/10.1109/TASLP.2019.2936755

Mehler, A., Walraven, K., and Melber, H. (Eds.). (2011). Sub-Saharan Africa. In *Africa Yearbook Volume 7* (pp. 1–16). Brill. https://doi.org/10.1163/ej.9789004205567.i-550.7

Niekerk, D. R. V., and Barnard, E. (2013). Generating fundamental frequency contours for speech synthesis in yorùbá. *Interspeech 2013*, 1027–1031. https://doi.org/10.21437/Interspeech.2013-112

ỌDélọBí, Ọdétúnjí Àjàdí. (2008). Recognition of Tones in YorÙbÁ Speech: Experiments With Artificial Neural Networks. In B. Prasad and S. R. M. Prasanna (Eds.), *Speech, Audio, Image and Biomedical Signal Processing using Neural Networks* (Vol. 83, pp. 23–47). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-75398-8_2

Osakuade, O., and King, S. (2024). *Do Discrete Self-Supervised Representations of Speech Capture Tone Distinctions?* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2410.19935

Pamisetty, G., and Sri Rama Murty, K. (2023). Prosody-TTS: An End-to-End Speech Synthesis System with Prosody Control. *Circuits, Systems, and Signal Processing*, *42*(1), 361–384. https://doi.org/10.1007/s00034-022-02126-z

Shen, G., Watkins, M., Alishahi, A., Bisazza, A., and Chrupała, G. (2024). *Encoding of lexical tone in self-supervised models of spoken language* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2403.16865

Sosimi, A. A., Adegbola, T., and 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*(5), 895. https://doi.org/10.4314/jasem.v23i5.20

Yang, D., Liu, S., Huang, R., Weng, C., and 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*, 2913–2925. https://doi.org/10.1109/TASLP.2024.3402088

Yang, J., Nyima, T., and Qi, J. (2024). *DAEPK:Domain-Adaptive Text Feature Enhancement Technology Integrating Prior Knowledge Domain In Text Classification*. In Review. https://doi.org/10.21203/rs.3.rs-4004271/v1