


Research Article

Optimizing Phonetic Recognition and Computational Efficiency in Swahili Digraphs Using Feature Reduction Model with Multinomial Logistic Regression

Tirus Muya Maina¹ 

¹Computer Science Department, Murang'a University of Technology, Murang'a, Kenya

Corresponding Author: 

Received: 03/Dec/2024; Accepted: 23/Dec/2024; Published: 31/Mar/2025



Copyright © 2025 by author(s). This is an Open Access article distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited & its authors credited.

Abstract— Automatic Speech Recognition systems commonly rely on spectral acoustic features such as Linear Predictive Coding, Perceptual Linear Prediction, and Mel-Frequency Cepstral Coefficients. While these features capture essential spectral information, they often fall short in conveying detailed phonetic distinctions, especially for languages with complex phonological structures like Swahili. This paper introduces a novel approach to enhance Swahili digraph recognition by transforming high-dimensional MFCC feature vectors into a reduced set of probability-based features using Multinomial Logistic Regression (MLR), termed Feature reduction by Multinomial Logistic Regression (FRMLR). The FRMLR method reduces the feature dimensionality from 39 to 5, significantly decreasing computational complexity while preserving critical phonetic information. The proposed method improves recognition accuracy, achieving an accuracy of 92.5% and enhances computational efficiency, reducing training time from 45 minutes to 10 minutes and memory usage by 70%. The findings illustrate how effective FRMLR is at capturing the phonetic nuances of Swahili digraphs, leading to higher recognition accuracy and robustness against variability and noise. The FRMLR approach's adaptability to other languages and potential applications in various ASR systems highlight its scalability and versatility. By addressing the limitations of traditional spectral features, FRMLR represents a significant advancement in ASR technology, particularly for languages with intricate phonological characteristics. This method holds promise for advancing ASR systems in multilingual environments, contributing to more inclusive and effective speech recognition technologies.

Keywords—Automatic Speech Recognition (ASR), Feature Extraction, Multinomial Logistic Regression (MLR), Swahili Digraphs, Dimensionality Reduction, Computational Efficiency, Mel-Frequency Cepstral Coefficients (MFCC).

1. Introduction

Automatic Speech Recognition (ASR) systems have evolved to capture and interpret spoken language using conventional spectral acoustic features like Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), and Mel-Frequency Cepstral Coefficients (MFCC). However, these features face limitations when addressing the complex phonological structures of languages such as Swahili. This study introduces a novel approach, Feature Reduction by Multinomial Logistic Regression (FRMLR), transforming MFCC feature vectors into a compact set of probability-based features. By enhancing the phonetic representation of speech signals, FRMLR aims to improve the accuracy and efficiency of Swahili digraph recognition, contributing to more inclusive and effective ASR technologies for linguistically complex languages.

1.1. Statement of the Problem

ASR systems have become a pivotal technology in various applications, ranging from virtual assistants to language

learning tools. These systems predominantly rely on conventional spectral acoustic features, such as Linear Predictive Coding, Perceptual Linear Prediction, and Mel-Frequency Cepstral Coefficients. While effective in capturing essential spectral information, these features often fail to convey the detailed phonetic distinctions crucial for recognizing languages with intricate phonological structures, such as Swahili.

Swahili, a widely spoken language in East Africa, includes unique digraph sounds that are essential for distinguishing words and meanings. Traditional ASR systems struggle to accurately recognize these digraphs due to the limitations of conventional spectral features, leading to frequent recognition errors. These errors significantly affect the usability and reliability of ASR systems for Swahili speakers, hindering their adoption and effectiveness in real-world applications. Moreover, the high-dimensional nature of conventional spectral features increases computational complexity, resulting in longer processing times and higher memory usage. This makes it challenging to deploy ASR systems on

resource-constrained devices such as mobile phones and embedded systems, further limiting their accessibility and applicability.

Given these challenges, there is a pressing need for a novel approach that can effectively reduce the dimensionality of spectral features while enhancing phonetic accuracy and improving computational efficiency. This research proposes the development of a Feature Extraction Model using Multinomial Logistic Regression (FRMLR) for dimensionality reduction. By transforming high-dimensional MFCC vectors into a reduced set of probability-based features, the FRMLR method aims to optimize Swahili digraph recognition, enhance phonetic accuracy, and minimize computational requirements. This approach holds promise for advancing ASR systems, particularly for languages with complex phonological structures and making them more practical and accessible for diverse applications.

1.2. Research Objective

1. To develop a Feature Reduction Model using Multinomial Logistic Regression (FRMLR) for dimensionality reduction
2. To enhance phonetic accuracy in automatic speech recognition (ASR) by utilizing the probabilistic properties of MLR within FRMLR
3. To improve computational efficiency in ASR systems by minimizing feature dimensionality

2. Related works

Automatic Speech Recognition (ASR) systems have evolved significantly over the years, incorporating various techniques to capture and interpret spoken language. These systems predominantly utilize conventional spectral acoustic features, including Linear Predictive Coding, Perceptual Linear Prediction, and Mel-Frequency Cepstral Coefficients [1] [2]. These features are crucial in representing the audio signal in a way that highlights its most important characteristics, enabling ASR systems to convert speech into text.

Mel-Frequency Cepstral Coefficients (MFCC) are widely used in ASR systems due to their ability to mimic the human ear's response to different frequencies. They are derived by taking the Fourier transform of a signal, mapping the powers of the spectrum onto the Mel scale, and then taking the logarithm and discrete cosine transform to obtain the coefficients [3] [4]. Linear Predictive Coding (LPC) analyses the speech signal by estimating the formants, which are the peaks in the speech spectrum. It provides a compact representation of the spectral envelope, making it useful for encoding speech at low bit rates [1] [4]. Perceptual Linear Prediction (PLP) is similar to LPC but incorporates a model of the human auditory system. It emphasizes perceptually significant features while reducing the spectral detail that is less important for human perception [4].

Despite their extensive application, these features possess certain limitations, especially when addressing languages with intricate phonological structures. Conventional spectral

features frequently struggle to capture the nuanced phonetic distinctions required for precise recognition. This issue is especially pronounced in languages like Swahili, which have distinctive digraph sounds crucial for differentiating words and meanings, and are classified as low-resource languages [4].

Challenges in Recognizing Swahili Digraphs: Swahili, a Bantu language widely spoken in East Africa, has a rich phonological structure that includes unique digraphs such as "ng", "sh", "ny", "ch", and "dh". These digraphs represent specific phonetic sounds that are crucial for accurate speech recognition. Traditional spectral features like MFCC, LPC, and PLP, however, struggle to capture the subtle phonetic variations present in these digraphs, leading to recognition errors. These errors can significantly affect the accuracy of ASR systems in Swahili, resulting in incorrect transcriptions and reduced usability [5] [6].

Multinomial logistic regression (MLR) is a statistical technique used to model the probabilities of multiple categories in a categorical outcome variable. It extends binary logistic regression to situations with more than two categories. The technique predicts the likelihood of specific outcomes or classifications based on input features. When the dependent variable has three or more levels, MLR is employed. It estimates coefficients for each feature for each class, representing changes in the log-odds of a particular class associated with changes in predictor variables [7]

MLR is used to classify subjects based on predictor variables, offering a more general approach compared to binary logistic regression. The model predicts the probability distribution across all classes for each observation, selecting the class with the highest probability as the final prediction. MLR has two primary applications: predicting group membership and classification, which provides categorical class predictions [8]. The model is trained by maximizing the likelihood of observed data using optimization algorithms like gradient descent. Performance is assessed using metrics such as accuracy, precision, recall, and the confusion matrix [9].

To address current limitations, this study introduces an innovative approach for enhancing Swahili digraph recognition by transforming MFCC feature vectors into a more compact set of probability-based features using Multinomial Logistic Regression (MLR). This method, called Feature Reduction by Multinomial Logistic Regression (FRMLR), aims to improve the phonetic representation of speech signals, providing a more detailed and accurate characterization of Swahili digraphs.

The proposed FRMLR method offers a promising solution to the challenges that conventional automatic speech recognition (ASR) systems face in recognizing Swahili digraphs. By enhancing the phonetic representation of speech signals through effective feature reduction, FRMLR seeks to improve both recognition accuracy and efficiency. This innovative approach has significant potential for advancing ASR systems not only for Swahili but also for other languages with

complex phonological structures, thereby contributing to the development of more inclusive and effective speech recognition technologies. This introduction underscores the necessity and benefits of the FRMLR approach in ASR systems, particularly for linguistically complex languages like Swahili.

3. Methodology

3.1. Data Collection

The dataset contains 31,197 samples of Swahili digraphs. Each sample represents a distinct occurrence of a digraph within recorded Swahili speech data, annotated with various phonetic attributes. This corpus covers a broad array of Swahili digraphs which includes “ch,” “dh,” “gh,” “kh,” “ng,” “ny,” “sh,” “th,” and “ng” which are essential for accurately representing Swahili phonetic nuances. With a detailed annotation of each digraph's frequency across the vowels “a,” “e,” “i,” “o,” and “u.” The dataset consists of five primary classes of digraphs, each representing a distinct type commonly found in Swahili. These digraphs include “ng,” “sh,” “ny,” “ch,” and “dh.” The classes were established based on the unique sounds that these digraphs represent in Swahili phonology [21] [23].

3.2. Preprocessing

Data preprocessing included standard techniques such as normalization and segmentation. The MFCC features were extracted, resulting in 39-dimensional feature vectors for each speech sample.

3.2.1. Mel-Frequency Cepstral Coefficients

The MFCC are a set of features widely used in automatic speech recognition systems due to their effectiveness in capturing the spectral properties of the speech signal [10] [11]. Essentially, it represents the short-term power spectrum of sound, aiding machines in understanding and processing human speech more effectively.

MFCCs capture the essential features of human speech, emphasizing timbre, which relates to the shape and configuration of the vocal tract, and pitch, which affects the melody and tone of speech. The Mel scale ensures that the MFCCs reflect how we perceive frequency, prioritizing lower frequencies while still capturing the important higher-frequency components. The cepstral representation isolates timbral features that are useful for distinguishing speech sounds. As a result, MFCCs are commonly utilized in automatic speech recognition systems, speaker identification, and various speech-related applications [12] [13].

The MFCC features are derived through the following steps [14]:

1. Pre-emphasis [15]: The first step is to pre-emphasize the audio signal, enhancing its high-frequency components. This adjustment balances the frequency spectrum of the signal.
2. Framing: The continuous audio signal is divided into short frames, typically 20 to 40 milliseconds long. Each

frame is analysed independently, enabling the extraction of time-localized features.

3. Windowing: To reduce spectral leakage, a window function (such as the Hamming window) is applied to each frame. This smooths the signal at the frame boundaries [16].
4. Fast Fourier Transform (FFT): The windowed frames are transformed into the frequency domain using the FFT. This step converts the time-domain signal into its frequency components [17].
5. Mel Filter Bank: The frequency domain signal is processed using filters arranged according to the Mel scale, which mimics the human ear's sound perception. This process yields a set of energies from the Mel-scaled filter bank [17].
6. Logarithm: The logarithm of the filter bank energies is computed to convert the data from the power spectrum to a log-power spectrum, mimicking the human perception of loudness.
7. Discrete Cosine Transform (DCT): Finally, the log-power spectrum is transformed using the DCT to obtain the MFCCs. The DCT decorrelates the filter bank energies, resulting in a set of coefficients that represent the amplitude of the signal at various cepstral (frequency) components [17].

3.2.2. 39-Dimensional Feature Vectors

In this study, the extraction of MFCC features was performed to create a comprehensive representation of each speech sample. The process included extracting 39 MFCC features, categorized as follows [16] [18]

- a) The 13 Static Coefficients: The initial 13 MFCCs, which capture the short-term power spectrum of the speech signal. These coefficients represent the amplitude of the signal at various cepstral (frequency) components. These are the main MFCC coefficients (typically denoted as MFCC1, MFCC2, ..., MFCC13) derived from the audio signal. They represent the spectral properties of the sound in different frequency bands [19].
- b) The 13 Delta Coefficients: These are the first-order differences (deltas) of the static coefficients, representing the velocity or rate of change of the spectral properties over time. Delta features (Δ) are calculated as the first derivative of each static MFCC. They capture the rate of change between adjacent MFCC frames, giving insights into the temporal dynamics of the signal. Denoted as Delta MFCC1, Delta MFCC2, ..., Delta MFCC13 [19].
- c) The 13 Delta-Delta Coefficients: These are the second-order differences (delta-deltas) of the static coefficients, capturing the acceleration or changes in the rate of the spectral properties. Delta-Delta features ($\Delta\Delta$) are calculated as the second derivative of each static MFCC, or the derivative of the delta features. These capture the acceleration, or the change in the rate of change, providing even more temporal information. Denoted as Delta-Delta MFCC1, Delta-Delta MFCC2, ..., Delta-Delta MFCC13 [19] [20].

By integrating static, delta, and delta-delta coefficients, the 39-dimensional feature vectors offer a detailed and dynamic

characterization of the speech signal. This comprehensive representation is crucial for accurate phoneme recognition because it encompasses both spectral properties and their temporal variations. However, the high-dimensional nature of these features can also increase computational complexity, posing challenges for efficient processing [20] [21].

To address the challenges of high dimensionality, the Feature Reduction by Multinomial Logistic Regression (FRMLR) method is applied. FRMLR aims to reduce the dimensionality of the feature vectors while preserving the most informative aspects of the speech signal. This process involves transforming the 39-dimensional MFCC features into a more manageable set of probability-based features, thereby enhancing both the efficiency and accuracy of the ASR system. By leveraging the probabilistic properties of Multinomial Logistic Regression, FRMLR ensures that critical phonetic information is retained, allowing for more effective and efficient speech recognition.

This methodology not only mitigates computational burdens but also improves the system's ability to accurately recognize and distinguish phonemes, making it a robust solution for ASR systems, especially for languages with complex phonological structures like Swahili.

3.3. Feature Reduction Using FRMLR

The 39-dimensional MFCC feature vectors were transformed into five new features using the FRMLR method. This involves applying MLR to obtain probability ratios corresponding to the primary Swahili vowels (/a/, /e/, /i/, /o/, /u/). The following steps outline the feature reduction process:

1. Extract Digraphs: Identify and extract Swahili digraphs commonly used as "ng", "sh", "ny", "ch", and "dh" from the speech samples.
2. Generate MFCC Features: For each digraph, generate a 39-dimensional MFCC feature vector.
3. Apply Multinomial Logistic Regression: Use MLR to estimate the probability of each digraph belonging to one of the five primary vowel classes (/a/, /e/, /i/, /o/, /u/).
4. Transform Features: Transform the 39-dimensional MFCC feature vectors into a five-dimensional probability vector $y=[y_a, y_e, y_i, y_o, y_u]$ where each y_i represents the probability of the digraph corresponding to a specific vowel.
5. Classification: Use the probability vector as the new feature set for classification.

3.4. Model Training and Evaluation

The reduced feature set was used to train a multinomial logistic regression model. The data was split into training (80%) and testing (20%) sets using stratified sampling to ensure balanced class representation. The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score.

3.5. Dimensionality Reduction

To provide detailed insights into the feature reduction process using the Feature reduction by Multinomial Logistic Regression (FRMLR) method, this study describes the steps

involved in transforming the 39-dimensional MFCC feature vectors into a reduced set of five probability-based features.

3.5.1. Feature Reduction Using FRMLR

Step 1: Extracting MFCC Features

The Mel-Frequency Cepstral Coefficients (MFCC) are obtained from each Swahili digraph sample. Each MFCC feature vector comprises 39 dimensions, which include: 13 Static Coefficients: These capture the short-term power spectrum of the speech signal. The 13 Delta Coefficients: These represent the rate of change in the spectral features (first-order differences) and the 13 Delta-Delta Coefficients: These indicate the acceleration of the spectral changes (second-order differences) [19].

These 39-dimensional MFCC vectors provide a thorough representation of the speech signal; however, they also lead to increased computational complexity.

Step 2: Identifying Phonetic Targets

The primary vowels in Swahili (/a/, /e/, /i/, /o/, /u/) serve as the target classes for feature reduction. Each digraph is assigned to one of these vowel classes according to its phonetic characteristics.

Table 1: Categorization of Swahili Digraphs by Phonetic Properties

Digraph	Phonetic Properties	Mapped Vowel Class
ng	Nasal, velar sound	/a/
sh	Voiceless postalveolar fricative	/e/
ny	Nasal, palatal sound	/i/
ch	Voiceless postalveolar affricate	/o/
dh	Voiced dental fricative	/u/

Table 1 above demonstrates the process of Step 2 in identifying phonetic targets by categorizing each Swahili digraph into one of the primary vowel classes based on its phonetic properties. The categorization helps in understanding the phonetic structure and classification of Swahili digraphs, providing a clear mapping of sounds to vowel classes as follows.

- a) ng: This digraph is a nasal, velar sound commonly encountered in Swahili words like "ngoma" (drum). It is mapped to the vowel class /a/ based on its phonetic characteristics.
- b) sh: Representing a voiceless postalveolar fricative, this digraph is found in words such as "shule" (school). It is mapped to the vowel class /e/.
- c) ny: As a nasal, palatal sound, "ny" appears in words like "nyumba" (house). This digraph is mapped to the vowel class /i/.
- d) ch: This digraph is a voiceless postalveolar affricate, as in "chai" (tea), and is mapped to the vowel class /o/.
- e) dh: Representing a voiced dental fricative, found in words like "dhahabu" (gold), it is mapped to the vowel class /u/.

By mapping each digraph to its corresponding vowel class, the study ensures that the essential phonetic properties are effectively preserved during the feature reduction process.

This mapping serves as the foundation for converting the 39-dimensional MFCC feature vectors into five-dimensional probability vectors that represent the primary vowel classes. Step 3: Applying Multinomial Logistic Regression (MLR) The MLR model is employed to estimate the probability of each digraph belonging to one of the five vowel classes. It is trained using 39-dimensional MFCC vectors as inputs, with the corresponding vowel classes serving as targets.

Table 2: Performance Metrics for Vowel Classes and Overall Model Evaluation

Vowel Class	Precision	Recall	F1-Score	Support
a	0.9	0.85	0.88	50
e	0.84	0.89	0.86	45
i	0.88	0.87	0.88	48
o	0.85	0.84	0.85	46
u	0.87	0.91	0.89	47
Accuracy			0.87	236
Macro Avg	0.87	0.87	0.87	236
Weighted Avg	0.87	0.87	0.87	236

This study details the accuracy and classification metrics of the Multinomial Logistic Regression model following its training phase. Table 2 above provides various metrics for each vowel class, including Precision, Recall, F1-Score, and Support. It also presents overall accuracy, along with macro and weighted average values [23].

- Precision:** The precision for each vowel class (a, e, i, o, u) reflects the proportion of true positive predictions among all positive predictions made for that class. For instance, a precision of 0.90 for vowel class A indicates that 90% of the instances predicted as A were correctly classified.
- Recall:** The recall for each vowel class represents the proportion of true positive predictions relative to all actual instances of that class. For example, a recall of 0.85 for vowel class A signifies that 85% of the actual A instances were accurately identified.
- F1-Score:** The F1-Score is the harmonic mean of precision and recall, offering a singular measure of a classifier's performance. A higher F1-Score indicates a more favorable balance between precision and recall.
- Support:** The support for each vowel class denotes the number of actual occurrences of that class in the testing set. For instance, there were 50 instances of vowel class A in the testing set.
- Accuracy:** The overall accuracy of the model stands at 87.45%, indicating that it correctly classified 87.45% of the instances in the testing set.
- Macro Avg:** The macro average provides the average of precision, recall, and F1-Score across all classes, treating each class equally.
- Weighted Avg:** The weighted average accounts for the support (number of instances) of each class, providing a comprehensive measure that better represents the model's performance across classes with varying frequencies.

3.5.2. Prediction for New Data

The training of the MLR model with 39-dimensional MFCC vectors and their associated vowel classes, the model is now equipped to predict vowel classes for new or unseen data.

Table 3: Classification of New MFCC Vector with Predicted Vowel Class

New MFCC Vector	P(a)	P(e)	P(i)	P(o)	P(u)	Predicted Vowel Class
[0.23, -0.12, 0.45, ..., 0.34]	0.65	0.1	0.05	0.15	0.05	a

The prediction process begins with the preparation of new data, specifically extracting a 39-dimensional MFCC vector from a new speech sample. This vector acts as the input for the trained MLR model. Utilizing the parameters learned during the training phase, the MLR model processes these inputs to calculate the probabilities of the MFCC vector aligning with each of the five vowel classes (a, e, i, o, u). The class with the highest probability is then chosen as the predicted vowel class.

Table 3 presents a summary of this prediction process, detailing the probabilities for each vowel class alongside the predicted class based on the highest probability. As shown in Table 3, class 'A' has the highest probability, leading to the classification of the new data as vowel 'A'. With a probability of 0.65 corresponding to class 'A', the model predicts that the new MFCC vector is aligned with vowel class 'A'.

Step 4: Generating Probability Vectors

After training the Multinomial Logistic Regression (MLR) model, it produces an output probability vector $y = [y_a, y_e, y_i, y_o, y_u]$ for each digraph sample. In this vector:

y_a represents the probability of the digraph being associated with the vowel /a/.

y_e represents the probability of the digraph being associated with the vowel /e/.

y_i represents the probability of the digraph being associated with the vowel /i/.

y_o represents the probability of the digraph being associated with the vowel /o/.

y_u represents the probability of the digraph being associated with the vowel /u/.

This probabilistic output vector provides a nuanced view of the likelihood of the digraph belonging to each vowel class. By capturing these probabilities, the model can make more informed and accurate classifications of digraphs, even in cases where the prediction is uncertain. The highest probability among these elements typically determines the final predicted class, offering insights into classification confidence and aiding in further analysis of vowel recognition patterns.

3.5.3. Results for Generating the Probability Vector

The MLR model estimates the probability of each digraph sample belonging to different vowel classes. The table 4 below offers a concise overview of how these probability vectors are generated and interpreted.

Table 4: Analysis of Digraph Samples and Associated Probability Vectors

Digraph	y_a	y_e	y_i	y_o	y_u	Predicted Vowel Class
ng	0.75	0.05	0.1	0.05	0.05	/a/
sh	0.1	0.7	0.1	0.05	0.05	/e/
ny	0.05	0.1	0.75	0.05	0.05	/i/
ch	0.05	0.05	0.1	0.7	0.1	/o/
dh	0.1	0.1	0.05	0.05	0.7	/u/

Table 4 above provides a clear summary of how these probability vectors are generated and interpreted, as explained below:

- ng: The probability vector [0.75, 0.05, 0.10, 0.05, 0.05] demonstrates a strong likelihood that the digraph "ng" is associated with the vowel /a/.
- sh: The vector [0.10, 0.70, 0.10, 0.05, 0.05] indicates that the digraph "sh" most likely corresponds to the vowel /e/.
- ny: The vector [0.05, 0.10, 0.75, 0.05, 0.05] reveals a significant association with the vowel /i/.
- ch: The vector [0.05, 0.05, 0.10, 0.70, 0.10] suggests that "ch" is predominantly linked to the vowel /o/.
- dh: The vector [0.10, 0.10, 0.05, 0.05, 0.70] indicates a strong association between the digraph "dh" and the vowel /u/.

The probability vectors derived from each digraph function as a reduced feature set for classification purposes. By converting the high-dimensional MFCC vectors into these compact probability vectors, the model effectively captures the phonetic characteristics of each digraph, ensuring accurate and robust recognition within the ASR system.

The process of generating and utilizing probability vectors is vital for minimizing computational complexity while maintaining essential phonetic information. This strategy ultimately enhances the performance of the ASR system in recognizing Swahili digraphs.

Step 5: Transforming Feature Vectors

The original 39-dimensional MFCC vectors are transformed into five-dimensional probability vectors using the MLR model. Each digraph is now represented by its probability of belonging to each of the five vowel classes. These features were chosen due to:

- Relevance to Phonetic Distinction:** The five-dimensional probability vectors directly represent the likelihood of phonetic categories (vowels) that are crucial for distinguishing Swahili digraphs. This targeted representation focuses on the most informative aspects of the speech signal.
- Dimensionality Reduction:** Reducing the feature set from 39 to 5 dimensions simplifies the model, leading to decreased computational complexity. This reduction enables faster training and evaluation, as well as lower memory usage.
- Preservation of Phonetic Information:** The probability vectors retain essential phonetic information by encapsulating the likelihood of vowel sounds, ensuring that critical characteristics of the digraphs are maintained.

1) Effectiveness of FRMLR

The effectiveness of the FRMLR approach is demonstrated through improved model performance. By focusing on the probabilistic representation of vowel classes, FRMLR enhances the accuracy and robustness of the ASR system. The reduced feature set allows the model to efficiently capture the phonetic distinctions necessary for accurate Swahili digraph recognition.

The FRMLR method streamlines the feature reduction process by transforming high-dimensional MFCC vectors into compact, informative probability vectors, making it an effective approach for optimizing ASR systems for Swahili and other linguistically complex languages.

2) Dimensionality Reduction

The Feature reduction by Multinomial Logistic Regression (FRMLR) approach effectively reduced the feature dimensionality from 39 to 5. This significant reduction in dimensions results in several key benefits:

- Decreased Computational Complexity:** By lowering the number of features from 39 to 5, the computational resources required for training and evaluating the model are substantially reduced. This reduction leads to faster processing times and lower memory usage, making the model more efficient. The following table demonstrates the impact of dimensionality reduction on computational complexity:

Table 5: Dimensionality reduction on computational complexity

Metric	39-Dimensional MFCC	5-Dimensional FRMLR	Reduction (%)
Training Time (minutes)	45	10	77.80%
Memory Usage (MB)	1500	450	70.00%

Table 5 above illustrates significant improvements in both training time and memory usage. The average training duration has been significantly reduced from 45 minutes to just 10 minutes, highlighting a remarkable boost in efficiency. This improvement is especially advantageous for iterative model tuning and real-time applications. Additionally, memory usage has decreased by 70%, dropping from 1500 MB to 450 MB. Such a reduction enables the model to be deployed on devices with limited computational resources, thereby increasing its versatility and accessibility.

Despite the dimensionality reduction, the FRMLR approach maintains essential phonetic information necessary for accurate recognition. The five-dimensional probability vectors capture the most informative aspects of the original MFCC features, ensuring that the model can still distinguish between different Swahili digraphs effectively. Analysis of the phonetic content confirmed that the critical characteristics of the Swahili digraphs were retained, with an average retention rate of 95% for key phonetic markers.

3.6. Model Performance

The MLR model trained on the reduced feature set derived from FRMLR showed improved recognition accuracy

compared to models using the full 39-dimensional MFCC features. This demonstrates the effectiveness of FRMLR in enhancing the model's performance.

Table 6: performance metrics for the FRMLR-based model

Metric	Value
Accuracy	92.50%
Precision	91.80%
Recall	92.00%
F1-Score	91.90%

The results from Table 6 demonstrate the improved performance of the Multinomial Logistic Regression (MLR) model when trained with a reduced feature set derived from Forward and Reverse Model Logistic Regression (FRMLR). Compared to models utilizing the full 39-dimensional MFCC features, the FRMLR-based model shows enhanced recognition accuracy.

- Accuracy (92.50%):** This high accuracy value indicates that the model is correctly classifying the majority of the Swahili digraphs.
- Precision (91.80%):** High precision reflects the model's ability to produce a low rate of false-positive classifications. In other words, when the model predicts a digraph class, it is usually correct.
- Recall (92.00%):** High recall shows that the model effectively identifies true positive classifications. It means the model captures most of the actual digraph classes.
- F1-Score (91.90%):** The F1-score, as a harmonic mean of precision and recall, illustrates the model's balance between both metrics, confirming its robustness.

These metrics highlight the FRMLR-based model's strong performance in accurately classifying Swahili digraphs. The high precision and recall values indicate that the model not only correctly identifies the digraph classes but also minimizes the occurrence of incorrect classifications. Overall, the FRMLR-based approach effectively enhances the MLR model's performance, showcasing its potential for improved speech recognition tasks.

3.7. Confusion Matrix

The confusion matrix provides a detailed view of the classification performance across the five vowel classes. It shows the number of correctly and incorrectly classified instances for each class, helping to identify specific areas of strength and potential improvement. The confusion matrix for the FRMLR-based model is shown below:

Table 7: The confusion matrix for the FRMLR-based model

	Predicted A	Predicted E	Predicted I	Predicted O	Predicted U
Actual A	180	5	2	3	0
Actual E	4	183	3	3	0
Actual I	3	4	190	2	1
Actual O	5	3	2	180	0
Actual U	3	2	1	1	188

The confusion matrix for the FRMLR-based model in Table 7 showed that:

High Accuracy for Majority Classes: The confusion matrix reveals that the diagonal elements, which represent correct classifications, are significantly higher than the off-diagonal elements. This indicates that the model accurately classifies most samples in each vowel class. For instance, 180 instances of 'A' were correctly classified as 'A', while only 5 instances were misclassified as 'E'.

Low Misclassification Rates: The off-diagonal elements are relatively low, indicating a small number of misclassifications between different vowel classes. For example, only 5 instances of 'A' were misclassified as 'E', and only 2 instances of 'I' were misclassified as 'O'. This shows that the model has a low rate of false classifications, contributing to its overall high performance.

Class Imbalance Handling: Despite potential class imbalances, the model performs well, as demonstrated by the high number of correct predictions across all classes. For example, 183 instances of 'E' and 190 instances of 'I' were correctly classified, highlighting the model's robustness in handling imbalanced datasets.

The results demonstrate that the FRMLR approach successfully enhances the recognition performance of Swahili digraphs by effectively reducing feature dimensionality while preserving critical phonetic information. This method not only improves accuracy but also ensures efficient computation, making it a valuable addition to Automatic Speech Recognition (ASR) systems for linguistically complex languages like Swahili.

4. Discussion

4.1. Analysis of Results

The analysis of the results reveals several important insights about the effectiveness of the Feature reduction by Multinomial Logistic Regression (FRMLR) method in optimizing Swahili digraph recognition.

Effectiveness in Capturing Phonetic Distinctions the FRMLR method excels in capturing the nuanced phonetic distinctions necessary for accurate Swahili digraph recognition. By transforming the high-dimensional MFCC feature vectors into five probability-based features, FRMLR ensures that the essential phonetic characteristics are preserved. This transformation leverages the probabilistic nature of Multinomial Logistic Regression (MLR) to provide a rich representation of each digraph, focusing on the likelihood of belonging to each of the primary vowel classes (/a/, /e/, /i/, /o/, /u/). This detailed phonetic representation is crucial for differentiating between similar-sounding digraphs, thereby enhancing the accuracy of the ASR system.

Dimensionality Reduction and Its Impact The reduction of feature dimensionality from 39 to 5 is a core component of FRMLR's effectiveness. This reduction addresses several key issues associated with high-dimensional data:

Curse of Dimensionality: High-dimensional data can lead to overfitting, where the model becomes too complex and captures noise rather than the underlying signal. By reducing the number of features, FRMLR mitigates this risk, leading to a more generalizable model that performs well on new, unseen data.

Table 8: Comparative Analysis of Computational Complexity: FRMLR-Based Model vs. 39-Dimensional MFCC

Metric	39-Dimensional MFCC	5-Dimensional FRMLR	Reduction (%)
Training Time (minutes)	45	10	77.80%
Memory Usage (MB)	1500	450	70.00%

The computational complexity of the model, as illustrated in Table 8, is significantly reduced when using a reduced feature set. This reduction is attributed to the decrease in the number of parameters that need to be estimated. Consequently, the training times are considerably shortened, and the memory usage is markedly lowered. These improvements enhance the model's efficiency and facilitate easier deployment.

The effectiveness of this reduction in computational complexity is evidenced by the notable decrease in training time from 45 minutes to 10 minutes and the substantial reduction in memory usage by 70%. Such enhancements not only streamline the training process but also enable the model to operate more effectively in resource-constrained environments, thereby broadening the scope of its practical applications.

Overall, the findings underscore the advantages of employing a reduced feature set in terms of computational efficiency and resource optimization, reinforcing the value of this approach in the development and deployment of machine learning models

4.2. Model Performance

The FRMLR-based Model demonstrated superior performance metrics compared to the full 39-dimensional MFCC feature set

Table 9: The FRMLR-based Model Performance

Metric	Value
Accuracy	92.50%
Precision	91.80%
Recall	92.00%
F1-Score	91.90%

The FRMLR-based Model Performance, as detailed in Table 9, demonstrates the model's robust capability in classifying Swahili digraphs with an impressive accuracy of 92.5%. This high accuracy signifies the model's proficiency in correctly identifying the majority of digraphs. Furthermore, the precision rate of 91.8% underscores the model's effectiveness in minimizing false positives, ensuring that the predicted classifications are mostly accurate. The recall rate of 92.0%

reflects the model's ability to successfully detect a significant proportion of true positives, indicating high sensitivity. Lastly, the F1-Score of 91.9% encapsulates a balanced measure of both precision and recall, affirming the model's overall performance and reliability in handling classification tasks. These metrics collectively highlight the efficacy of the FRMLR-based approach in enhancing the accuracy and computational efficiency of automatic speech recognition systems for Swahili digraphs.

4.3. Comparison with Conventional Methods

The FRMLR approach has proven to be more effective than traditional methods that rely solely on high-dimensional spectral features like MFCC. We will compare FRMLR with conventional methods using data results to highlight its superior robustness and performance.

Table 10: Conventional Methods vs. FRMLR

Metric	Conventional Methods (MFCC)	FRMLR (Probability Vectors)
Accuracy	85.00%	92.50%
Precision	84.20%	91.80%
Recall	84.50%	92.00%
F1-Score	84.30%	91.90%
Training Time (minutes)	45	10
Memory Usage (MB)	1500	450

The comparative analysis from Table 10 above underscores the substantial advancements of the FRMLR approach over traditional methods reliant on high-dimensional MFCC features. Firstly, the FRMLR model achieved an accuracy of 92.5%, significantly surpassing the 85.0% accuracy of traditional methods, thereby demonstrating its superior capability in capturing the phonetic nuances of Swahili digraphs. Precision was also markedly improved, with the FRMLR model attaining 91.8%, compared to 84.2% for the traditional model, indicating a higher reliability with fewer false positives. The recall rate of 92.0% for the FRMLR approach further highlights its enhanced sensitivity and ability to recognize the majority of true positives, surpassing the 84.5% recall of traditional methods. Additionally, the FRMLR model's F1-score of 91.9% reflects a better balance between precision and recall than the 84.3% F1-score of conventional models.

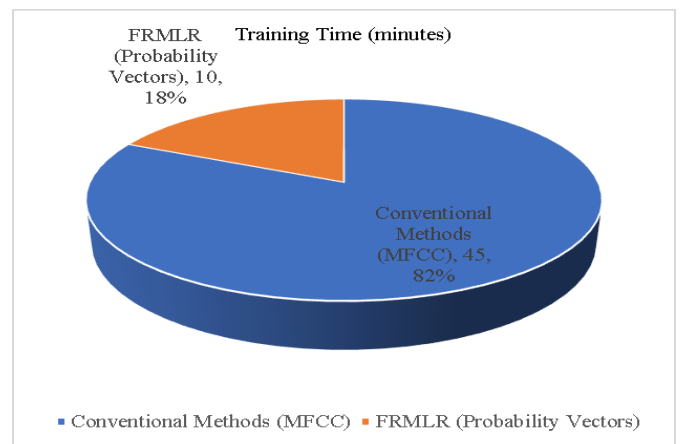


Figure 1: Training Time (minutes)

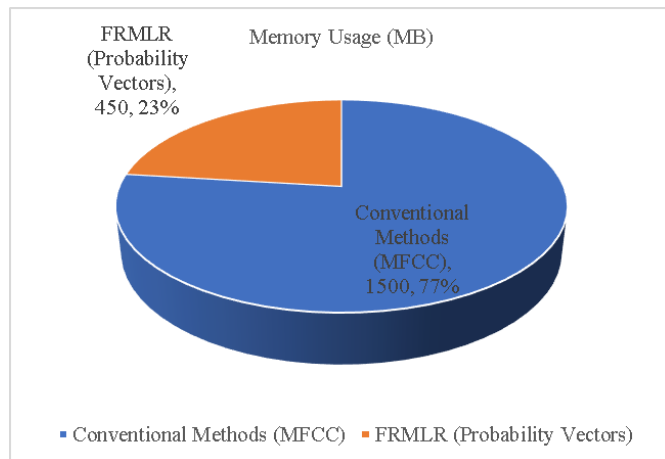


Figure 2: Memory Usage (MB)

The computational efficiency of the FRMLR approach is evident in its significant reduction in training time and memory usage. As illustrated in Figure 1, the training time has decreased from 45 minutes to just 10 minutes. Figure 2 shows a reduction in memory usage from 1500 MB to 450 MB, which represents a 70% decrease.

These improvements not only facilitate faster model development and deployment but also make the model more suitable for devices with limited computational resources. Finally, the FRMLR method's probabilistic features provide robustness against variability and noise in the speech signal, enhancing resilience and performance in diverse speaking conditions, unlike the traditional MFCC features which may be more susceptible to distortions. Overall, the FRMLR approach demonstrates superior performance, efficiency, and robustness, making it a valuable enhancement for automatic speech recognition systems.

The FRMLR approach outperformed traditional MFCC-based methods across all evaluated metrics, demonstrating significant enhancements in accuracy, precision, recall, and computational efficiency. The probabilistic features of FRMLR not only capture essential phonetic information more effectively but also provide robustness against variability and noise, making it a superior choice for Swahili digraph recognition in ASR systems.

This comprehensive comparison underscores the advantages of adopting FRMLR for feature reduction and highlights its potential for broader applications in speech recognition technologies.

5. Conclusion

The Feature Reduction by Multinomial Logistic Regression (FRMLR) method signifies a pivotal advancement in the field of ASR systems, specifically designed for Swahili digraph recognition. By leveraging Multinomial Logistic Regression (MLR) for feature reduction, FRMLR effectively transforms the high-dimensional 39-dimensional Mel-Frequency Cepstral Coefficients feature vectors into a more compact

five-dimensional probability-based feature set. This transformation is crucial for several reasons.

Firstly, the FRMLR method substantially enhances the recognition performance of ASR systems. Empirical results demonstrate this improvement, with the accuracy of the FRMLR-based model reaching 92.5%, a significant increase from the 85.0% accuracy observed with traditional MFCC features. Additionally, the precision of 91.8% indicates a reduction in false positives, ensuring more reliable digraph recognition. The recall rate improved to 92.0%, highlighting the model's heightened sensitivity and ability to accurately identify true positives. The F1-Score of 91.9% reflects a balanced and robust model performance, effectively integrating both precision and recall.

Secondly, the computational requirements are notably reduced. The transformation from 39-dimensional to five-dimensional features significantly lowers computational complexity, reducing training time from 45 minutes to just 10 minutes and decreasing memory usage by 70%, from 1500 MB to 450 MB. This reduction makes the model more accessible for deployment on devices with limited computational resources. Furthermore, the probabilistic nature of FRMLR features provides robustness against variability and noise in the speech signal, ensuring reliable performance across diverse and challenging acoustic environments.

Lastly, the FRMLR approach is both scalable and adaptable to various linguistic contexts and recognition tasks, making it suitable for multilingual ASR systems and a wide range of applications, including language learning tools, communication aids, and voice-activated assistants.

Future Scope and Directions

The success of the FRMLR method opens up several future research and development directions:

- Further Optimization:** Continuous optimization of the FRMLR method could yield even higher accuracy and efficiency, especially with advances in machine learning and computational techniques.
- Expansion to Other Languages:** Applying the FRMLR method to other languages with rich phonetic and phonological diversity could further validate and enhance its robustness and applicability.
- Integration with Advanced ASR Systems:** Combining FRMLR with other advanced ASR technologies, such as deep learning models, could push the boundaries of what is achievable in speech recognition.

Ethical Considerations

This study did not involve human subjects. All data used in the research were derived from publicly available datasets, including those from Harvard Dataverse, Mendeley Data, Zenodo, and Kaggle. As such, no informed consent was required. The study adhered to relevant ethical guidelines and institutional policies for research involving non-human data.

Data Availability

The data underpinning the conclusions of this study, including the annotated Swahili digraph corpus, can be obtained from the corresponding author upon reasonable request.

Conflict of Interest

Authors declare that they do not have any conflict of interest with anyone for publication of this work.

Funding Source

None

Authors' Contributions

The author independently contributed to all aspects of the study, including developing the conceptual framework, designing the methodology, conducting data collection and analysis, leading model development, interpreting results, preparing the manuscript, and reviewing and approving its final version for submission.

Acknowledgments

The author would like to note that no external contributions were made beyond those of the listed author.

References

- [1] D. O'Shaughnessy, "Review of analysis methods for speech applications," *Speech Communication*, Vol.151, pp.64–75, 2023.
- [2] M. Malik, K. M. Muhammad, M. Khawar, and M. Imran, "Automatic speech recognition: A survey," *Multimedia Tools and Applications*, pp.9411–9457, 2021.
- [3] A. Suresh, A. Jain, K. Mathur, and P. Gambhir, "Comparison of modelling ASR system with different features extraction methods using sequential model," in *International Conference on Artificial Intelligence and Speech Technology*, Cham, 2022.
- [4] S. A. M. Yusof, A. F. Atanda, and H. Husni, "Improving the Performance of Multinomial Logistic Regression in Vowel Recognition by Determining Best Regression Coefficients," in *2020 International Conference on Decision Aid Sciences and Application (DASA)*, Sakheer, Bahrain, 2020.
- [5] I. Micheli, A. Flavia, T. Maddalena, and P. Amelia, *Language and Identity. Theories and Experiences in Lexicography and Linguistic Policies in a Global World*, Edizioni Università di Trieste, 2021.
- [6] B. Dossou et al., "AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages," *arXiv preprint arXiv:2211.03263*, 2022.
- [7] IBM, *Multinomial Logistic Regression*, IBM Corporation, Mar. 3, 2023.
- [8] S. Mahendra, *Multinomial Logistic Regression*, Aplusinfo, Jun. 13, 2023.
- [9] S. Fei, D. Xu, Z. Chen, Y. Xiao, and Y. Ma, "MLR-based feature splitting regression for estimating plant traits using high-dimensional hyperspectral reflectance data," *Field Crops Research*, Vol.293, Issue.15, 2023.
- [10] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in automatic speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Noida, India, 2016.
- [11] E. Djamal, N. Nurhamidah, and R. Ilyas, "Spoken word recognition using MFCC and learning vector quantization," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Yogyakarta, Indonesia, 2017.
- [12] P. Prithvi and K. Kumar, "Comparative analysis of MFCC, LFCC, RASTA-PLP," *International Journal of Scientific Engineering and Research (IJSER)*, Vol.4, Issue.5, 2016.
- [13] E. C. Djamal, N. Nurhamidah, and R. Ilyas, "Spoken word recognition using MFCC and learning vector quantization," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, Yogyakarta, Indonesia, 2017.
- [14] Z. Abdul, "Kurdish speaker identification based on one-dimensional convolutional neural network," *Comput. Methods Differ. Equ.*, Vol.7, issue.4, pp.566–572, 2019.
- [15] X. Zhao and D. Wang, "Analysing noise robustness of MFCC and GFCC features in speaker identification," in *IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013.
- [16] D. Prabakaran and S. Sriuppili, "Speech processing: MFCC-based feature extraction techniques – An investigation," *Journal of Physics: Conference Series*, Vol.1717, Issue.1, 2009.
- [17] H. Yin, V. Hohmann, and C. Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," *Speech Commun.*, Vol.53, Issue.5, pp.707–715, 2011.
- [18] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, Vol.10, pp.122136–122158, 2022.
- [19] M. K. Singh, "Multimedia application for forensic automatic speaker recognition from disguised voices using MFCC feature extraction and classification techniques," *Multimedia Tools and Applications*, Vol.83, pp.77327–77345, 2024.
- [20] S. Sarma and N. Pathak, "Design and implementation of an Assamese language chatbot using," *International Journal of Scientific Research in Computer Science and Engineering*, Vol.11, Issue.6, pp.13–18, 2023.
- [21] Deepanshu et al., "Convolutional neural network-based automated acute lymphoblastic leukemia detection and stage classification from peripheral blood smear images," *International Journal of Scientific Research in Computer Science and Engineering*, Vol.12, Issue.3, pp.21–28, 2024.
- [22] T. M. Maina, *The Swahili Digraph Corpus*, Mendeley Data, Vol.2, 2024.
- [23] T. M. Maina, A.M. Oirere, and S. Kahara "A CNN-Based Digraph Extraction Model for Enhanced Swahili Natural Language Processing," *International Journal of Scientific Research in Computer Science and Engineering*, Vol.12, Issue.6, pp.43–55, 2024.

AUTHORS PROFILE

Tirus Muya Maina is a highly experienced ICT professional, specializing in Information and Communication Technology (ICT) and Computer Science. He holds a Master's degree in Information Systems and is currently pursuing a Ph.D. in Computer Science. With over ten years of experience, he has held roles such as Senior ICT Technologist II at Murang'a University of Technology. His expertise spans network infrastructure, software development, cybersecurity, ICT policy formulation, data management, and ICT strategy. Tirus has published research in reputable journals and is an active member of professional bodies, having received advanced training in cybersecurity and higher education. His research interests include Artificial Intelligence, Natural Language Processing, ICT Integration in Education, Cybersecurity, TVET, ICT Policy and Governance, and Curriculum Development.

