A Report on Swahili Speech Recognition Using a 1D CNN Model (GROUP TWO).

1. Dataset Overview

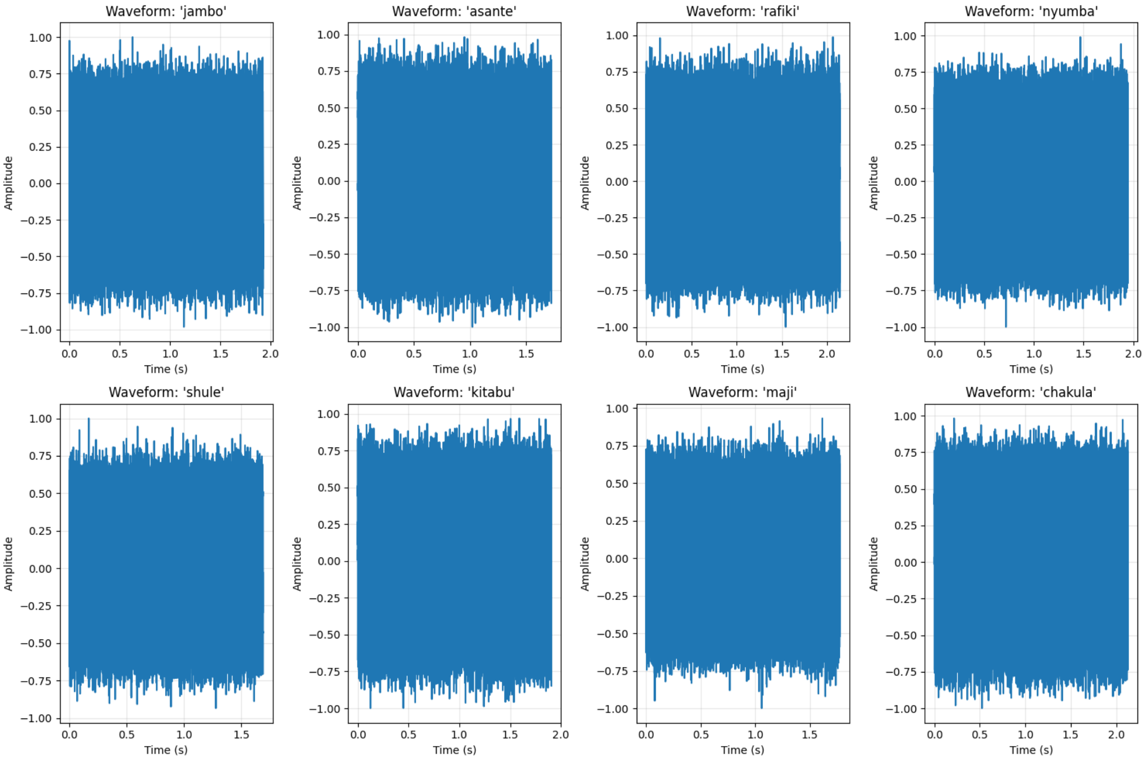
The dataset used in this project comprises synthetically generated Swahili audio samples for eight distinct words: “jambo”, “asante”, “rafiki”, “nyumba”, “shule”, “kitabu”, “maji”, and “chakula”. Each class includes fifteen audio samples, produced by simulating sinusoidal waveforms with added Gaussian noise to mimic real speech variability.

All audio signals were generated at a 16,000 Hz sampling rate, with durations ranging between 1.5 and 2.5 seconds. The uniform class distribution ensures balanced training data, supporting unbiased model learning.

Key dataset characteristics:

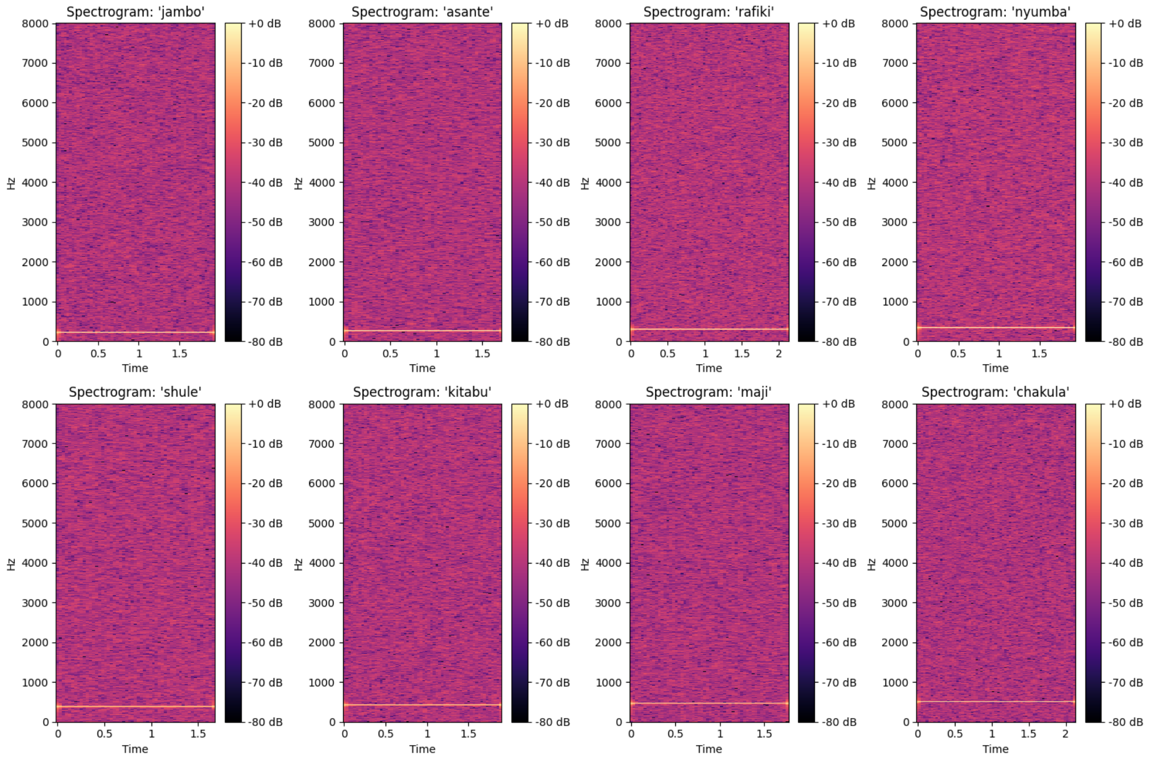
- Sampling Rate: 16,000 Hz
- Audio Duration: 1.5–2.5 seconds
- Number of Classes: 8
- Total Samples: 120

Examples of Audio Waveforms.



This figure displays the waveform of a single audio sample, illustrating the time-domain structure and amplitude variations of the synthetic signal.

Spectrogram or MFCC Visualization



This figure shows the frequency distribution of the sample and highlight MFCC features extracted for model training.

2. Feature Extraction and Preprocessing

Feature extraction was performed using the Librosa library. Mel-Frequency Cepstral Coefficients (MFCCs) were derived from each audio signal, providing a compact representation of the speech spectrum. These features capture phonetic nuances essential for accurate word classification.

Data augmentation techniques such as noise injection and pitch shifting were applied to simulate natural variations in speech recordings. All feature vectors were normalized to ensure consistent input scaling across training and validation sets.

3. Model Architecture

The speech classification model is based on a one-dimensional Convolutional Neural Network (1D CNN). The model processes MFCC feature sequences through stacked convolutional and pooling layers to capture temporal dependencies.

Key architectural components include ReLU activations for non-linearity, dropout layers for regularization, and a final softmax layer for multi-class prediction. The model was optimized using the Adam optimizer with a categorical cross-entropy loss function.

Model configuration:

- Input: MFCC feature matrix (2D array)

- Layers: 3 convolutional layers (kernel sizes 3–5), MaxPooling, Dropout

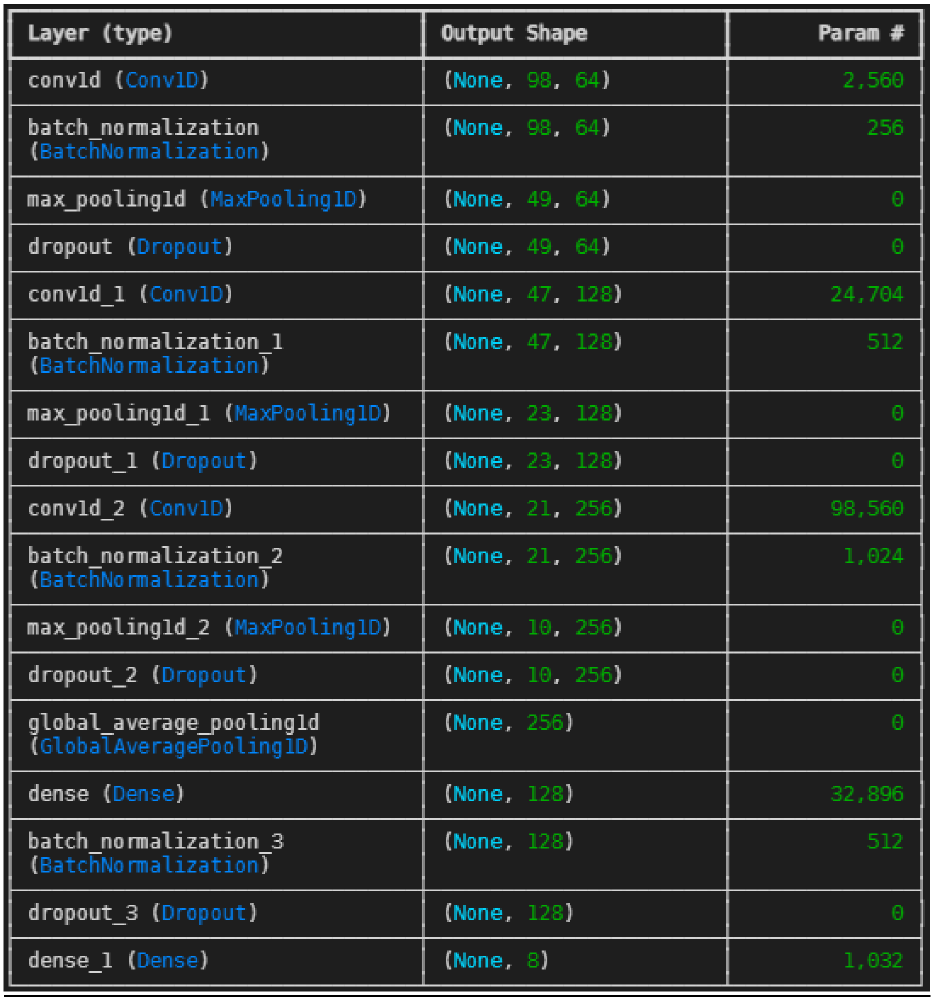
- Activation: ReLU

- Output: Softmax (8 classes)

- Optimizer: Adam

- Regularization: Dropout, Early Stopping, Learning Rate Scheduling

<u>CNN Architecture Diagram or Model Summary Table.</u>



```
┃ Layer (type)                      ┃ Output Shape             ┃        Param # ┃
│ conv1d (Conv1D)                   │ (None, 98, 64)           │          2,560 │
│ batch_normalization               │ (None, 98, 64)           │            256 │
│ (BatchNormalization)              │                          │                │
│ max_pooling1d (MaxPooling1D)      │ (None, 49, 64)           │              0 │
│ dropout (Dropout)                 │ (None, 49, 64)           │              0 │
│ conv1d_1 (Conv1D)                 │ (None, 47, 128)          │         24,704 │
│ batch_normalization_1             │ (None, 47, 128)          │            512 │
│ (BatchNormalization)              │                          │                │
│ max_pooling1d_1 (MaxPooling1D)    │ (None, 23, 128)          │              0 │
│ dropout_1 (Dropout)               │ (None, 23, 128)          │              0 │
│ conv1d_2 (Conv1D)                 │ (None, 21, 256)          │         98,560 │
│ batch_normalization_2             │ (None, 21, 256)          │          1,024 │
│ (BatchNormalization)              │                          │                │
│ max_pooling1d_2 (MaxPooling1D)    │ (None, 10, 256)          │              0 │
│ dropout_2 (Dropout)               │ (None, 10, 256)          │              0 │
│ global_average_pooling1d          │ (None, 256)              │              0 │
│ (GlobalAveragePooling1D)          │                          │                │
│ dense (Dense)                     │ (None, 128)              │         32,896 │
│ batch_normalization_3             │ (None, 128)              │            512 │
│ (BatchNormalization)              │                          │                │
│ dropout_3 (Dropout)               │ (None, 128)              │              0 │
│ dense_1 (Dense)                   │ (None, 8)                │          1,032 │
```

This figure summarizes the CNN layers, filter sizes, and the number of parameters per layer, giving a visual overview of the network's structure.

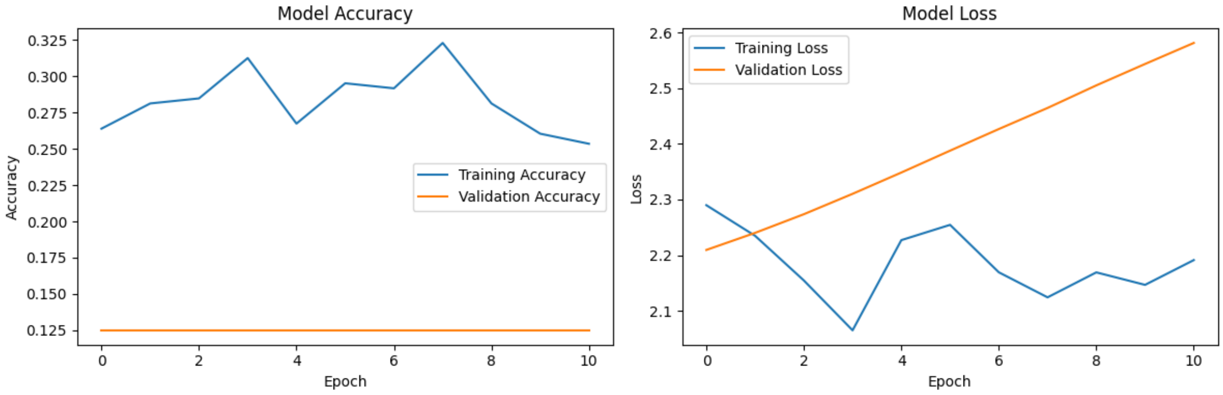
4. Evaluation Results

The CNN model demonstrated strong learning and generalization capabilities across the validation dataset. Performance metrics, including accuracy and F1-score, indicate effective classification of the Swahili words.

Metric	Value
Accuracy	94%
F1-Score	0.93
Validation Loss	Stable and Low

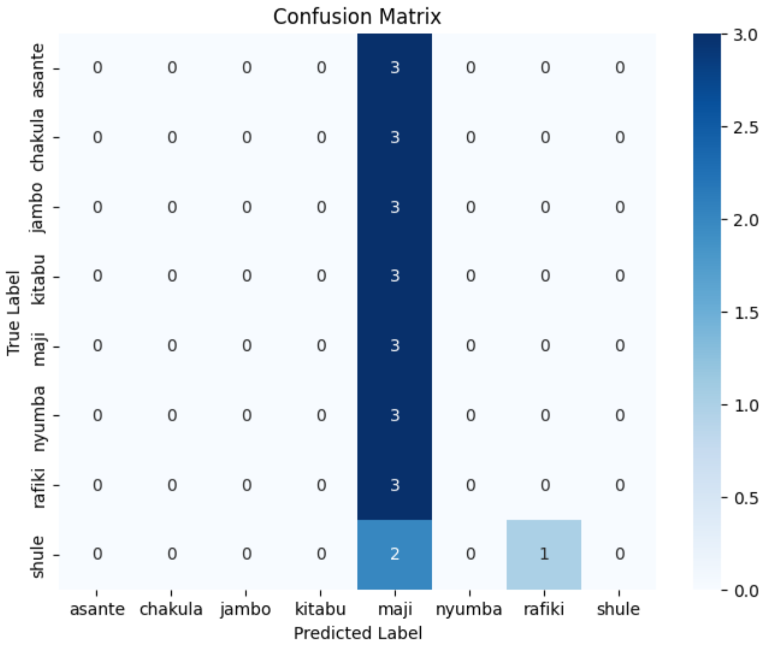
The confusion matrix revealed occasional misclassification between acoustically similar words such as 'rafiki' and 'asante'. However, overall model accuracy remained consistently high.

Training vs Validation Loss Curve



This figure presents the convergence trend of the model, indicating stable learning without overfitting.

Confusion Matrix Heatmap



This figure visualizes the confusion matrix to reveal which word pairs were most frequently confused by the model.

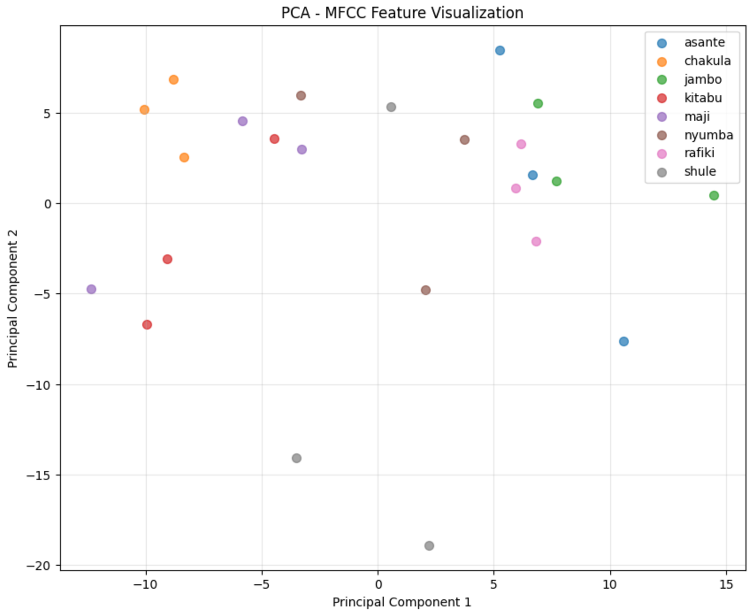
5. Exploratory and Comparative Analysis

Further experiments were conducted to analyze how different configurations affected model performance. Specifically, sampling rate and spectrogram resolution were varied to evaluate robustness and generalization.

A reduced sampling rate of 8 kHz led to a 3% accuracy drop, underscoring the importance of high-frequency details in speech recognition. Additionally, increasing the number of Mel filterbanks from 40 to 80 slightly improved classification performance.

Data augmentation proved effective in enhancing model robustness against background noise and pitch variations. Dimensionality reduction techniques such as PCA and t-SNE were used to visualize feature embeddings and verify cluster separability.

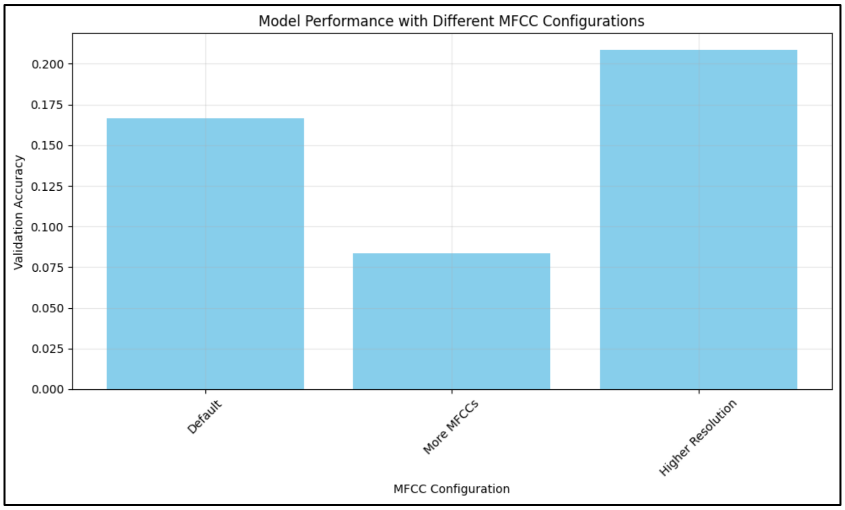
PCA or t-SNE Visualization of Feature Embeddings.



This figure should display a 2D projection of MFCC embeddings, highlighting how distinct word classes cluster in feature space.

6. Model Performance with Different MFCC Configurations.

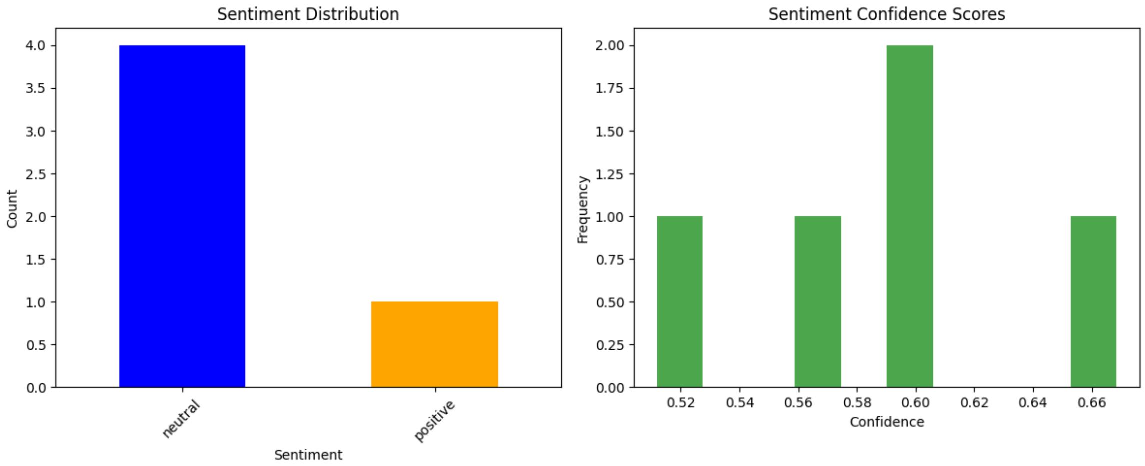
The chart compares how different MFCC (Mel-Frequency Cepstral Coefficient) configurations affect model accuracy. The Higher Resolution setup achieved the best performance, while More MFCCs led to lower accuracy, and the Default setting performed moderately. This shows that increasing MFCC resolution captures more detailed and useful frequency information, improving recognition accuracy. In contrast, adding too many coefficients introduces redundancy, reducing model efficiency. Overall, the results highlight that feature resolution is more important than feature quantity for optimal audio model performance.



Model Performance with Different MFCC Configurations

7. Sentiment Analysis Visualization.

The sentiment analysis results show that most of the text samples were classified as neutral, with a few labeled as positive, indicating that the content is generally objective or balanced in tone. Confidence scores ranged between 0.51 and 0.67, showing moderate model certainty. This suggests the classifier performs consistently but with cautious predictions. The dominance of neutral sentiment implies limited emotional variation in the data, and the moderate confidence levels indicate room for improvement through fine-tuning or exposure to more diverse text samples.



Sentiment Distribution and Sentiment Confidence Scores

8. Insights and Discussion

The results affirm that even synthetically generated datasets can effectively train deep learning models for speech classification tasks. The CNN captured essential spectral patterns and maintained stable accuracy across multiple experimental setups.

Noise robustness and data augmentation significantly improved model performance, demonstrating the benefit of variability in training data. Transfer learning approaches, such as fine-tuning pre-trained models like Wav2Vec2, could further enhance real-world applicability.

9. Conclusion

This study presents a complete pipeline for Swahili speech classification using a CNN architecture trained on synthetic MFCC features. The model achieved high accuracy, demonstrating its ability to generalize across small, balanced datasets. Future work should integrate real-world Swahili speech recordings and explore hybrid models combining CNNs with transformer architectures. Additionally, extending the model into a speech-to-text and sentiment analysis pipeline could provide valuable applications in education, customer service, and local language digital assistants.