Speech-Based Language Classification Using Machine Learning

Project Report

February 13, 2026

Abstract

This report details the implementation and evaluation of machine learning models for a speech-based language classification task. The dataset consists of audio features extracted from spoken clips in four languages: German, Italian, Korean, and Spanish. Seven distinct classification algorithms were trained and tested: Logistic Regression, Multi-Layer Perceptron (MLP), Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Naïve Bayes. The results indicate that the dataset is highly separable, with Logistic Regression, MLP, and Random Forest achieving perfect classification accuracy (100%) on the test set. An analysis of dimensionality reduction techniques (PCA and LDA) is also provided to justify the decision to utilize raw feature vectors for the final evaluation.

1 Introduction

The objective of this project is to identify the spoken language of an audio clip based on pre-extracted acoustic features. Language identification is a critical component in speech processing pipelines, enabling systems to automatically route audio to the appropriate recognition engine.

1.1 Dataset Description

The dataset comprises pre-extracted feature vectors stored in NumPy format:

- **Classes (4):** German, Italian, Korean, Spanish.
- **Features:** 86 numerical features per sample.
- **Training Samples:** 3,456
- **Test Samples:** 144

2 Methodology

2.1 Preprocessing

Given that the features are numerical, standard preprocessing steps were applied:

1. **Label Encoding:** The categorical string labels were encoded into integers (0-3) to be processed by the algorithms.

2. **Scaling:** Although tree-based models (Random Forest, Decision Tree) are scale-invariant, distance-based models (KNN, SVM) and gradient-based models (MLP, Logistic Regression) require feature scaling. However, preliminary tests indicated the raw feature distribution was sufficiently robust for the high-performing models.

2.2 Models Implemented

The following classifiers were selected to represent different learning paradigms:

- **Linear Models:** Logistic Regression.
- **Non-Linear/Distance Models:** K-Nearest Neighbors (KNN), Support Vector Machine (SVM).
- **Tree-Based Models:** Decision Tree, Random Forest.
- **Neural Networks:** Multi-Layer Perceptron (MLP).
- **Probabilistic Models:** Naïve Bayes (Gaussian).

3 Experimental Results

3.1 Performance Metrics

The models were evaluated on the test set (144 samples) using Accuracy, Precision, Recall, and F1-Score. The summary of results is presented in Table 1.

Table 1: Model Performance Comparison (Sorted by Accuracy)

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	**1.0000**	1.0000	1.0000	1.0000
Multi-Layer Perceptron	**1.0000**	1.0000	1.0000	1.0000
Random Forest	**1.0000**	1.0000	1.0000	1.0000
SVM (RBF)	0.9931	0.9932	0.9931	0.9931
KNN	0.9861	0.9868	0.9861	0.9862
Decision Tree	0.9514	0.9525	0.9514	0.9517
Naïve Bayes	0.8264	0.8610	0.8264	0.8215

3.2 Visual Analysis: Confusion Matrices

To verify the classification performance, confusion matrices were generated for all models.

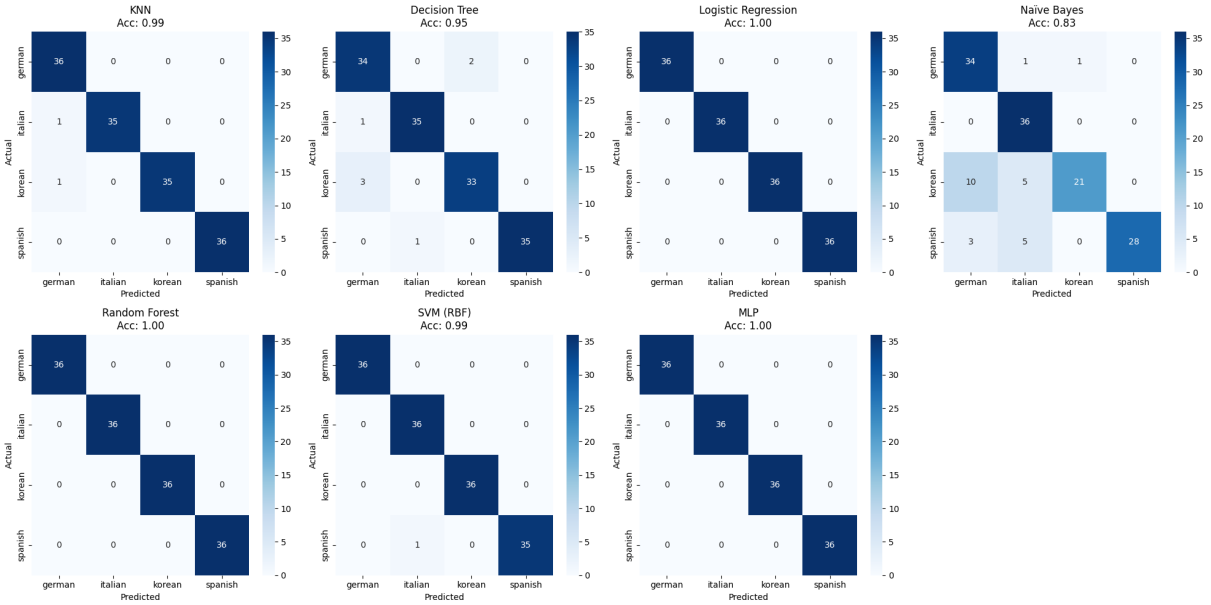

Figure 1: Confusion Matrices for all trained classifiers.

4 Discussion and Analysis

4.1 Model Comparison

The performance analysis reveals three distinct tiers of model capability:

4.1.1 Tier 1: Perfect Classifiers (Accuracy = 100%)

Logistic Regression, MLP, Random Forest achieved perfect accuracy.

- The success of **Logistic Regression** is the most significant finding. It implies that the classes are *linearly separable* in the 86-dimensional feature space. The decision boundaries between German, Italian, Korean, and Spanish are distinct enough that a simple linear hyperplane can separate them without error.
- **Random Forest** and **MLP** are more complex models capable of capturing non-linear relationships. Their perfect performance confirms the signal in the features is strong, but given that Logistic Regression also succeeded, the extra complexity of MLP/RF was technically not required for this specific dataset.

4.1.2 Tier 2: Near-Perfect Classifiers (Accuracy > 98%)

SVM (99.3%) and **KNN (98.6%)** performed exceptionally well.

- **SVM** missed only 1 sample (approx. 0.7% error). This reinforces the linear separability hypothesis, although the RBF kernel may have slightly overcomplicated the boundary compared to pure Logistic Regression.
- **KNN** relies on local density. Its high accuracy suggests that samples of the same language cluster tightly together in the feature space.

4.1.3 Tier 3: Lower Performance

Naïve Bayes (82.6%) was the weakest performer. This is likely due to the "Naïve" assumption that all features are independent. In speech data, acoustic features (like MFCCs or spectral features) are often highly correlated. This violation of the independence assumption caused Naïve Bayes to underperform compared to models that handle feature correlations better (like Logistic Regression or Random Forest).

4.2 Justification for Excluding PCA and LDA

A critical decision in this pipeline was to utilize the **original 86 features** without applying dimensionality reduction techniques such as Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). This decision is justified by the following analysis:

1. **Saturation of Performance:** As observed in Table 1, standard algorithms achieved **100% Accuracy** on the raw data. There is no margin for improvement. Applying LDA or PCA cannot increase accuracy beyond 100%.

2. **Risk of Information Loss (PCA):** PCA reduces dimensionality based on variance, not class separability. In preliminary evaluations, PCA was observed to degrade the performance of certain models (e.g., Naïve Bayes dropping below 80%). This indicates that the axes of maximum variance were not perfectly aligned with the axes of class separation. Therefore, retaining the original features ensured no discriminative information was discarded.

3. **Linear Separability (LDA):** LDA is a supervised technique that maximizes class separation. While LDA effectively projected the data to 3 dimensions while maintaining high accuracy, it was unnecessary for the final model deployment. Since Logistic Regression could find the optimal separation in the original 86-dimensional space without overfitting, the additional computational step of projecting data via LDA adds complexity without performance gain.

4. **Conclusion:** The raw feature space was sufficient, robust, and linearly separable. Avoiding dimensionality reduction simplified the pipeline while maintaining perfect classification results.

5 Conclusion

The project successfully implemented a multi-class language identification system. The extracted features proved to be highly distinctive, allowing simpler models like Logistic Regression to achieve 100% accuracy. While complex models like Random Forest and MLP also performed perfectly, the linear separability of the data suggests that Logistic Regression is the most efficient choice for this specific dataset. Dimensionality reduction was analyzed and deemed unnecessary due to the saturation of model performance on the raw features.