UNIVERSITAS
GADJAH MADA

# Language Identification in Indonesian Speech Using Acoustic Features and Machine Learning


Group 5

Satwika Nino Wandhana 23/516202/PA/22066

Michael Satpelin Williamtu 23/517103/PA/22151

Andi Fayza Maharani 23/516238/PA/22074

Mario Aloysius Lukman 23/516320/PA/22091

Muhammad Razan Alamudi 23/511396/PA/21784


LOCALLY ROOTED,

GLOBALLY RESPECTED

ugm.ac.id

# Introduction

Speech recognition brings wonderful accessibility, but this help often doesn't reach people in diverse places like Indonesia because current systems aren't built to understand their many unique languages. This project aims to improve language recognition in speech technology by identifying local languages and accents — critical for inclusivity in a linguistically rich country like Indonesia.

# Methodology - Overview

UNIVERSITAS GADJAH MADA

The system detects Indonesian, Javanese, and Sundanese by analyzing speech features and accents.

**The process includes**

- **Data Acquisition**
- **Pre - processing**
- **Feature Extraction**
- **Classification**
- **Performance Evaluation**

# Methodology - Data collection

UNIVERSITAS GADJAH MADA

## DATASETS

- OpenSLR 35 - *Javanese* Speech Corpus
- OpenSLR 36 - *Sundanese* Speech Corpus
- Nexdata.ai - *Indonesian* speech Data

## CONTENT

Audio recordings (.wav/.flac) with transcripts

## SPEAKER DIVERSITY

Varying age, gender, and dialect

# Methodology - Preprocessing

UNIVERSITAS GADJAH MADA

**RESAMPLING**

All audio is converted to a sampling rate of 16,000 Hz, the standard rate for speech recognition systems.

**ERROR HANDLING**

Audio files that fail to load due to corruption or format mismatch are automatically skipped.

**SILENCE REMOVAL VIA PITCH THRESHOLDING**

Segments with negligible pitch content (i.e., less than 5% of frames contain periodic components) are excluded.

# Methodology - Frame Segmentation

UNIVERSITAS GADJAH MADA

To perform localized analysis of the speech signal, each audio file is segmented into short overlapping frames using the following parameters

| FRAME LENGTH | FRAME SHIFT |
| --- | --- |
| 25 milliseconds (400 samples at 16kHz) | 10 milliseconds (160 samples at 16kHz) |

This sliding window technique allows the capture of temporal dynamics in short segments while ensuring sufficient overlap.

# Methodology - Feature Extraction

UNIVERSITAS GADJAH MADA

## The Process Includes

**ZERO CROSSING RATE (ZCR)**

High ZCR values indicate aperiodic (noisy) content typical of unvoiced sounds, while low ZCR suggests voiced speech.

**SHORT-TIME ENERGY**

Measures the average power of the frame. Voiced speech typically has higher energy due to sustained vocal fold vibrations.

**PITCH FREQUENCY**

Estimated using autocorrelation. The peak in the autocorrelation function (excluding the zero lag) indicates the fundamental frequency.

Each of these features is computed for all frames in an utterance.
Aggregated statistics (mean, standard deviation, min, max) are then extracted to represent the entire utterance.

# Methodology - Feature Extraction

UNIVERSITAS GADJAH MADA

## MFCC - BASED FEATURES

**MFCC (Mel-Frequency Cepstral Coefficients)**

13 MFCCs are extracted from the full audio. For each coefficient, both the mean and standard deviation are computed, resulting in 26 MFCC-derived features per utterance.

## FINAL FEATURE VECTOR

- **2 ZCR stats (mean, std)**
- **2 STE stats (mean, std)**
- **4 pitch stats (mean, std, min, max)**
- **1 voiced frame ratio (percentage of frames with pitch > 0)**
- **26 MFCC stats**
- **Total: 35+ features per utterance**

# Methodology - Classification

UNIVERSITAS GADJAH MADA

## LABEL ENCODING AND NORMALIZATION

Labels (Javanese, Sundanese, Indonesian) are encoded using LabelEncoder

Feature normalization is performed using StandardScaler to ensure zero mean and unit - variance.

## PERFORMANCE METRICS

**For Each Model =**

Accuracy is Computed

Classification reports include precision, recall, and F1-score per class.

A bar chart visualizes accuracy comparisons across models.

# Methodology - Classification

UNIVERSITAS GADJAH MADA

## MODEL TRAINING & EVALUATION

**Data is split into 70% training and 30% testing using stratified sampling.**
**Four classifiers are trained and evaluated:**

**Support Vector Machine** (SVM) - Linear SVM for separating languages based on maximal margin.

**Decision Tree** - Interpretable rules based on feature thresholds.

**Random Forest** - Ensemble of decision trees with bagging to reduce variance.

**XGBoost** - Gradient boosting framework with strong generalization performance.

# Results

UNIVERSITAS GADJAH MADA

| Classifier | Accuracy | Javanese | | | Sundanese | | | Macro F1-Score |
|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1-Score | Precision | Recall | F1-Score | |
| SVM | 0.8889 | 0.91 | 0.87 | 0.89 | 0.87 | 0.91 | 0.89 | 0.89 |
| Decision Tree | 0.7556 | 0.75 | 0.78 | 0.77 | 0.76 | 0.73 | 0.74 | 0.76 |
| Random Forest | 0..8444 | 0.86 | 0.83 | 0.84 | 0.83 | 0.86 | 0.84 | 0.84 |
| XGBoost | 0.7667 | 0.78 | 0.76 | 0.77 | 0.76 | 0.77 | 0.76 | 0.77 |

# Limitations

UNIVERSITAS GADJAH MADA

The amount of speech data we used was somewhat limited,
which can make it harder to develop even more advanced systems.

The system might not work as well with casual, everyday conversations,
in noisy environments, or with dialects not included in our initial recordings.

The sound characteristics we analyzed, while clear,
may not capture every subtle nuance in speech.

# Limitations

UNIVERSITAS GADJAH MADA

The amount of speech data we used was somewhat limited, which can make it harder to develop even more advanced systems.

The system might not work as well with casual, everyday conversations, in noisy environments, or with dialects not included in our initial recordings.

The sound characteristics we analyzed, while clear, may not capture every subtle nuance in speech.

# Future Research

UNIVERSITAS GADJAH MADA

Expanding the collection of speech recordings to include a wider variety of speakers, dialects, and background noises.

Exploring advanced deep learning techniques could allow the system to identify more complex and subtle patterns in sound.

Analyzing speech based on language structure or individual sound units (phonemes) may improve accuracy, particularly for instances where people mix languages.

# Conclusion

Our research demonstrates that combining known sound analysis methods with machine learning effectively identifies different regional languages in Indonesia.

This system is a significant step towards developing speech recognition that can better serve the many people and cultures across the country.

UNIVERSITAS
GADJAH MADA

Thank You!
Any Questions?