Language Identification in Indonesian Speech Using Acoustic Features and Machine Learning

Andi Fayza Maharani*, Michael Satpelin Williamtu*, Mario Aloysius Lukman,

Muhammad Razan Alamudi, Satwika Nino Wandhana*

Department Computer Science and Electronics, Universitas Gadjah Mada
Gedung C, Lantai 4, Sekip Utara, Bulaksumur, Sendowo, Sinduadi,
Kec. Mlati, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281

***Abstract*: Speech recognition technologies have significantly advanced communication accessibility, yet challenges remain in adapting these systems to linguistically diverse regions like Indonesia. This paper presents a method for separating voiced and unvoiced speech segments in Indonesian, Javanese, and Sundanese languages using Zero Crossing Rate (ZCR) and Short-Time Energy (STE). The study involves data collection from native speakers, preprocessing to standardize and enhance audio quality, and feature extraction to distinguish speech components. Classification is performed using both rule-based and machine learning approaches, with performance evaluated through accuracy, precision, recall, and F1-score. Additionally, data augmentation techniques such as speed perturbation, pitch shifting, and noise injection are employed to improve robustness. The results demonstrate the system's effectiveness in handling linguistic variability and environmental noise, contributing to the development of more inclusive speech recognition technologies for Indonesian languages.**

***Keyword:* speech recognition, voiced-unvoiced separation, Zero Crossing Rate (ZCR), Short-Time Energy (STE), Indonesian languages, feature extraction, data augmentation**

I. Introduction

Speech recognition technology is increasingly important. It allows for hands-free device use and makes technology more accessible, particularly for people with physical or learning challenges. However, current systems often struggle in places with many languages, as they are not always able to adapt to different sound patterns and language structures. This is a notable concern in Indonesia, home to over 700 languages and dialects, where commonly used speech recognition systems which are often trained on languages like English face significant hurdles.

Developing speech technology that truly serves Indonesian users requires a deep understanding of the sound characteristics of major regional languages like Javanese and Sundanese, alongside standard Indonesian. An important early stage in this work is correctly distinguishing between voiced (like the 'aah' sound) and unvoiced (like the 'sss' sound) parts of speech. This accuracy is vital for later analysis. Methods like Zero Crossing Rate (ZCR) and Short-Time Energy (STE) are effective and efficient for this task, particularly when computing resources are limited.

Our work explores a combined method for identifying Javanese, Sundanese, and Indonesian speech. We use established signal processing techniques such as ZCR, STE, and pitch analysis, along with Mel-Frequency Cepstral Coefficients (MFCCs), and apply machine learning to differentiate the languages. We tested this system using publicly available speech recordings, preparing and enhancing the data to improve its reliability in various scenarios. Our aim is to show that by combining different types of sound features, we can achieve strong language identification, which is a step towards creating more inclusive and flexible speech recognition systems for Indonesia's many languages.

II. Literature review

The classification of speech signals into voiced and unvoiced segments is a foundational step in many speech processing applications, including language identification. Several acoustic features have been widely used for this purpose, such as short-time energy (STE), zero-crossing rate (ZCR), and Mel-Frequency Cepstral Coefficients (MFCCs).

Madiha et al. proposed a methodology for classifying voiced and unvoiced segments of speech using short-time energy, magnitude, zero-crossing rate, and autocorrelation. They found that voiced segments typically exhibit higher short-time energy and magnitude, while unvoiced segments have higher zero-crossing rates and non-periodic autocorrelation patterns [2]. These features are crucial because the accurate discrimination of voiced/unvoiced parts directly influences the quality of subsequent feature extraction and classification stages in a speech recognition or language identification system. This observation highlights the importance of carefully selecting window functions and lengths to achieve stable and meaningful measurements [2].

Similarly, Hanifa et al. applied short-time energy and zero-crossing rate to separate voiced and unvoiced signals in Malay continuous speech [1]. Their findings aligned with previous studies, confirming that voiced speech segments display higher energy and lower ZCR, whereas unvoiced segments show the opposite pattern. This study further emphasizes that even in morphologically similar languages such as Malay and Indonesian, acoustic segmentation through these fundamental features can enhance speech preprocessing for tasks like language identification [1].

On the other hand, more advanced acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) have shown significant success in speech classification tasks. Vimal et al. utilized MFCCs alongside energy features to classify

emotional speech using machine learning algorithms, including Random Forest, SVM, and Decision Tree classifiers [3]. Their approach achieved high accuracy, especially using the Random Forest classifier, suggesting that MFCCs can effectively represent the phonetic structure of speech necessary for emotion or language classification tasks [3]. Although their work focused on emotion recognition, the pipeline—comprising feature extraction using MFCCs followed by classification using machine learning—mirrors typical strategies in language identification systems.

The integration of classical signal processing techniques like ZCR and STE with modern feature extraction methods such as MFCCs and machine learning classifiers provides a robust framework for speech-based classification tasks. Given the similarities in vocal characteristics between Malay and Indonesian languages, and the established efficacy of these techniques, they are highly relevant and adaptable to the problem of Indonesian language identification.

III. Methodology

The proposed system for Indonesian language identification to differentiate between Javanese, Sundanese and Indonesian is composed of five main stages: data acquisition, preprocessing, frame segmentation, feature extraction, and classification. This section provides a detailed explanation of each component, following the structure and logic of the Python implementation.

A. Data Acquisition

Three open-source Indonesian speech corpora are used:

1) OpenSLR 35 (Javanese Corpus)
2) OpenSLR 36 (Sundanese Corpus)
3) 359 Hours Indonesian Speech by Nexdata (Indonesian Corpus)

Each dataset contains speech recordings in .flac or .wav formats, with diverse speakers in terms of age, dialect, and gender. To ensure manageable processing, we randomly select up to 10,000 audio files per language.

The OpenSLR datasets are downloaded via a custom script using requests and extracted using zipfile. The Nexdata corpus is cloned from its GitHub repository using subprocess.

B. Preprocessing

To maintain consistency across datasets and prepare the audio for analysis, the following preprocessing steps are applied:

1) Resampling
All audio is converted to a sampling rate of 16,000 Hz, the standard rate for speech recognition systems. Resampling is done using *librosa.load(sr=16000)*.

2) Silence Removal via Pitch Thresholding
Segments with negligible pitch content (i.e., less than 5% of frames contain periodic components) are excluded. This ensures that the dataset focuses on meaningful voiced segments.

3) Error Handling
Audio files that fail to load due to corruption or format mismatch are automatically skipped, preventing system crashes during batch processing.

C. Frame Segmentation

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

1) Frame Length: 25 milliseconds (400 samples at 16kHz)
2) Frame Shift: 10 milliseconds (160 samples at 16kHz)

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

D. Feature Extraction

Each frame is processed to extract both time-domain and frequency-domain features. These are then aggregated per utterance to form the final feature vector.

1) Frame-Level Features
   a) Zero Crossing Rate (ZCR)
      Computed as the rate at which the signal changes sign. High ZCR values indicate aperiodic (noisy) content typical of unvoiced sounds, while low ZCR suggests voiced speech.
   b) Short-Time Energy (STE)
      Measures the average power of the frame. Voiced speech typically has higher energy due to sustained vocal fold vibrations.
   c) Pitch (Fundamental Frequency)
      Estimated using autocorrelation. The peak in the autocorrelation function (excluding the zero lag) indicates the fundamental frequency. A valid pitch is expected to be within 50–400 Hz.

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.

2) MFCC-Based Features
   a) 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. These coefficients capture spectral shape and are widely used in speech classification.

3) Final Feature Vector
   The final feature vector includes:
   a) 2 ZCR stats (mean, std)
   b) 2 STE stats (mean, std)
   c) 4 pitch stats (mean, std, min, max)
   d) 1 voiced frame ratio (percentage of frames with pitch > 0)
   e) 26 MFCC stats

   Total: 35+ features per utterance

E. Classification

The extracted feature vectors are passed into a supervised classification pipeline.

1) Label Encoding and Normalization
    a) Labels (Javanese, Sundanese, Indonesian) are encoded using LabelEncoder
    b) Feature normalization is performed using StandardScaler to ensure zero mean and unit variance.
2) Model Training and Evaluation
   Data is split into 70% training and 30% testing using stratified sampling.
   Four classifiers are trained and evaluated:
    a) Support Vector Machine (SVM): Linear SVM for separating languages based on maximal margin.
    b) Decision Tree: Interpretable rules based on feature thresholds.
    c) Random Forest: Ensemble of decision trees with bagging to reduce variance.
    d) XGBoost: Gradient boosting framework with strong generalization performance.
3) Performance Metrics
   For each model:
    a) Accuracy is computed.
    b) Classification reports include precision, recall, and F1-score per class.
    c) A bar chart visualizes accuracy comparisons across models.

IV. Results and Discussion

mean_zcr	std_zcr	mean_ste	std_ste	mean_pitch	std_pitch	min_pitch	max_pitch
0.1625	0.1348	0.0017	0.0030	17.73	75.07	0.00	363.64
0.1563	0.0820	0.0010	0.0017	13.96	59.97	0.00	296.30
0.1781	0.0860	0.0019	0.0028	17.41	63.24	0.00	285.71
0.1694	0.1371	0.0092	0.0174	32.96	95.45	0.00	355.56
0.1315	0.1096	0.0024	0.0035	14.93	61.37	0.00	296.30

Table 1: Sample of Extracted Audio Features

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

Table 2: Classifier Performance Metrics

A. Extracted Features

The audio preprocessing pipeline successfully extracted 36 features per audio sample, combining both time-domain and frequency-domain attributes. Table 1 illustrates a sample of the extracted features, including statistical metrics for zero-crossing rate (ZCR), short-time energy (STE), pitch-related features, voiced ratio, and 13 Mel-frequency cepstral coefficients (MFCCs) along with their means and standard deviations.

The MFCC features, known for capturing perceptually relevant audio patterns, showed significant variation between samples. Pitch features such as mean and standard deviation, along with voiced ratios, also provided useful indicators for language classification, especially given the tonal and phonetic differences between Javanese and Sundanese.

B. Classification Performance

Four different machine learning classifiers—Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost—were trained and evaluated on the feature dataset. The classification task involved distinguishing between Javanese and Sundanese languages. Table 2 summarizes the performance of each classifier based on accuracy, precision, recall, and F1-score.

a. Support Vector Machine (SVM): SVM achieved the highest classification performance, with an overall accuracy of 88.89%. It maintained balanced precision and recall across both classes, with an F1-score of 0.89 for each language. This result indicates that SVM effectively leveraged the high-dimensional feature space for decision boundary optimization.
b. Decision Tree: The Decision Tree classifier performed relatively poorly, with an accuracy of 75.56%. While it managed reasonable precision and recall values, its tendency to overfit may have reduced generalization capability. This is reflected in lower consistency across metrics for both language classes
c. Random Forest: The ensemble-based Random Forest model outperformed the Decision Tree and XGBoost, achieving 84.44% accuracy. Its precision and recall scores for Javanese and Sundanese were closely matched, indicating good model balance. The aggregated decision-making process likely contributed to improved robustness over the single-tree approach.
d. XGBoost: XGBoost achieved an accuracy of 76.67%, slightly better than the Decision Tree but below Random Forest and SVM. Despite XGBoost's reputation for high performance, its relative underperformance here may be attributed to the small dataset size and possibly insufficient parameter tuning.

C. Discussion

The results suggest that SVM is the most suitable model for this classification task, likely due to its ability to handle high-dimensional data and define optimal

hyperplanes for separation. The balanced performance across Javanese and Sundanese classes further supports its robustness in this setting.

In contrast, Decision Trees showed limited performance, which is consistent with their known sensitivity to noise and overfitting on smaller datasets. Although Random Forest improved upon this through ensembling, it still fell short of the SVM's effectiveness. XGBoost, while generally a strong performer, may require further hyperparameter tuning or a larger dataset to realize its full potential in this context.

Overall, the experimental results validate the effectiveness of combining traditional signal processing features (e.g., ZCR, STE, pitch) with MFCCs for the task of regional language classification. Future work may consider deep learning approaches or more advanced feature selection to further enhance classification performance.

V. Conclusion and future research

A. Our Findings

We have developed a system capable of distinguishing Indonesian, Javanese, and Sundanese speech. This system analyzes features like Zero Crossing Rate (ZCR), Short-Time Energy (STE), pitch, and Mel-frequency Cepstral Coefficients (MFCCs) from audio. Among the various machine learning methods tested, the Support Vector Machine (SVM) provided the highest accuracy at 88.89%

B. Limitations

We recognize some limitations in our work. The amount of speech data used, although balanced for each language, is somewhat small for developing more advanced models, such as deep neural networks. Additionally, the system might not perform as well with everyday spoken language, background sounds, or significant dialect differences not present in our initial data. Lastly, the chosen speech features are clear and efficient but they may not identify all intricate speech patterns.

C. Future Research

In future research, increasing the variety of the speech data with more speakers, dialects, and noisy conditions will help the system work more reliably in different situations. Additionally, deep learning methods such as convolutional or recurrent neural networks might identify more complex sound patterns than the current features. Lastly, using language modeling or analyzing speech at the sound-unit (phoneme) level could improve its accuracy, particularly when people switch between languages which is a common occurrence in Indonesia.

D. Conclusion

This work shows that using established sound analysis techniques with machine learning is a successful approach for identifying different regional languages in Indonesia. Our system represents meaningful progress toward creating speech recognition that better serves Indonesia's diverse population. Further development in this area holds promise for improving digital access and supporting language preservation throughout the nation.

References

[1] Hanifa, R., et al. "Voiced and unvoiced separation in Malay speech using zero crossing rate and energy," in Indones. J. Electr. Eng. Comput. Sci, vol. 16, no. 2, pp. 775–780, 2019.

[2] M. Jalil, F. Butt, A. Malik, "Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals," in 2013 The international conference on technological advances in electrical, electronics and computer engineering (TAEECE), 2013, pp. 208–212.

[3] Vimal, B., et al, "MFCC Based Audio Classification Using Machine Learning," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-4.