Journal of Computing and Communication Vol.4, No.2, PP. 45-62, 2025

Improving Spoken Language Identification in Noisy Environment Based on Feature Reduction Using PCA

Mona Nagy ElBedwehy*[,a], Kholoud Mayyalou[b], G. M. Behery[c], Reda Elbarougy[d]

[a]Department of Computer Science, Faculty of Computer and Artificial Intelligence, Damietta University, Egypt. Email: monaelbedwehy@du.edu.eg.
[b]Department of Computer Science, Faculty of Computer and Artificial Intelligence, Damietta University, Egypt.
[c]Department of Computer Science, Faculty of Computer and Artificial Intelligence, Damietta University, Egypt. Email: gbehery@du.edu.eg.
[d]Department of Artificial Intelligence and Data Science, College of Computer Science and Engineering, University of Ha'il, Ha'il 81481, Saudi Arabia. Email: reda.Elsayed@uoh.edu.sa.
[d]Department of Information Technology, Faculty of Computer and Artificial Intelligence, Damietta University, Egypt. Email: elbarougy@du.edu.eg.

*Corresponding Author: Mona Nagy ElBedwehy [monaelbedwehy@du.edu.eg]

ARTICLE DATA

Article history:
Received 20 june 2025
Revised 23 July 2025
Accepted 30 July 2025
Available online

Keywords:
Spoken Language Identification; Noisy Environment; PCA; LDA.

ABSTRACT

Automatic Spoken Language Identification (ASLID) is essential for effective multilingual communication, especially in real-world environments characterized by noise and acoustic variability where noise significantly impacts performance. This research introduces a robust ASLID framework that highlights the significance of feature reduction via principal components analysis (PCA) integrated with linear discriminant analysis (LDA) to enhance classification performance in noisy environments. The system utilizes OpenSMILE to extract extensive audio features, capturing diverse speech characteristics necessary for accurate language discrimination. To address the high dimensionality and redundancy inherent in the feature set, PCA is employed to reduce the feature space, preserving the most significant variance and enhancing computational efficiency. Following PCA, LDA is applied to maximize class separability, further refining the feature space for effective language classification. The proposed approach is evaluated on a benchmark dataset under various noise levels and test set proportions. Extensive experiments conducted on the IIIT-H Indic speech dataset demonstrate that the proposed PCA-LDA approach outperforms traditional methods, achieving an accuracy of up to 99.92% in noisy conditions, even with reduced feature dimensions. Experimental results demonstrate that integrating PCA with LDA significantly improves accuracy and robustness, outperforming conventional feature selection and classification techniques. The findings affirm that the combined PCA-LDA strategy effectively enhances the resilience of ASLID systems in challenging acoustic environments, making it a promising solution for practical multilingual speech processing applications.

1. Introduction

In an increasingly interconnected world, the ability to accurately identify spoken languages from audio signals plays a vital role in facilitating seamless communication across diverse linguistic borders. ASLID systems are essential for numerous applications, including multilingual voice assistants, automated transcription services, security systems, and cross-cultural communication platforms [1,2,3]. The core challenge in ASLID lies in extracting meaningful features from raw speech data that can reliably discriminate between languages, even under adverse environmental conditions such as background noise [4,5,6].

Despite significant advancements, one of the prominent challenges faced by ASLID systems is the degradation of performance in noisy environments [7,8]. Background noise, reverberations, and other acoustic interferences distort speech signals, making it difficult for traditional feature extraction and classification methods to accurately identify the spoken language. This challenge is particularly critical in real-world scenarios where controlled, noise-free conditions are rarely present. In particular, high-dimensional feature spaces extracted from speech signals tend to include irrelevant or correlated information, increasing computational load and potentially degrading classifier performance. Therefore, enhancing system robustness in such conditions is imperative for practical deployment.

The ASLID involves several stages, each critical to the successful identification of spoken language from audio input, particularly when leveraging machine learning techniques. The first stage of ASLID is data collection, which entails gathering diverse audio samples across multiple languages. This phase is essential for training robust models capable of generalizing well across different accents, dialects, and speaking styles. High-quality datasets contribute significantly to the model's performance, as they provide a rich variety of linguistic features for analysis.

The second stage is feature extraction, where specific characteristics of the audio signals are identified and quantified distinguish between different languages. Linear Predictive Analysis (LPC) is one of the most commonly utilized feature extraction approaches [9], Mel-Frequency Cepstral Coefficients (MFCCs) [10], Perceptual Linear Predictive Coefficients (PLPC) [11] that are used to capture distinctive speech characteristics. These features enable machine learning (ML) algorithms to distinguish between languages. However, their high dimensionality often introduces redundancy and noise, which can degrade classification accuracy, particularly in noisy environments.

OpenSMILE (Open-source Media Interpretation by Large feature-space Extraction) [12] is a robust toolkit made to extract features from speech and audio inputs developed by the International Audio Laboratories Erlangen. It is particularly renowned for its application in emotion recognition, speech analysis, and other multimedia processing tasks. Because the toolkit is open-source, researchers and developers are free to alter and expand its features to suit their needs. OpenSMILE offers a comprehensive set of features that cater to various audio processing requirements. Some of the primary feature types include: MFCCs, spectral features, prosodic features, formant frequencies, energy and amplitude features, voice quality features and emotional features. OpenSMILE extracts a substantial number of features, totaling 6,373. Due to the numerous and diverse features extracted by OpenSMILE, which include those obtained from various feature extraction methods, this research chose to utilize it for feature extraction.

Some of these features may be irrelevant, potentially impacting the accuracy of ASLID. Therefore, Feature reduction stage is necessary to identify the most relevant features while eliminating redundant or irrelevant ones. Feature reduction techniques, such as PCA, present effective solutions to these challenges by transforming high-dimensional feature sets into lower-dimensional subspaces that retain the most significant variability in the data. PCA reduces computational complexity while simultaneously improving feature discrimination by reducing redundancy and noise components. This transformation enables classifiers to more effectively distinguish between languages, especially in noisy environments where irrelevant variations can obscure critical features.

The final stage is classification, where ML algorithms, PCA is highly effective, but its combination with categorization algorithms such as LDA, Support Vector Machines (SVM), and deep learning (DL) [13] models can further enhance ASLID performance. This integrated approach ensures that noise-reduced features are optimally exploited for discriminative classification, making the system more resilient to environmental distortions. Prior research suggests that combining feature reduction with robust classifiers significantly improves accuracy in challenging acoustic environments.

This paper emphasizes the critical role of feature reduction through PCA in refining the correctness and strength of spoken language identification systems under challenging acoustic conditions. By systematically applying PCA prior to classification, the proposed approach reduces feature space dimensionality, enhances class separability, and mitigates the adverse effects of noise. This emphasis on feature modification and reduction seeks to create a computationally efficient, noise-resistant ASLID system that can retain high accuracy across a wide range of settings and resource restrictions.

The present work emphasizes the application of PCA for feature reduction in an ML-based speech language identification system and evaluates its effectiveness under noisy conditions. By systematically reducing the feature space dimensionality, the proposed methodology aims to improve classification accuracy and computational efficiency simultaneously. The research leverages the large-scale IIIT-Hindustan dataset, which encompasses diverse Indic languages and offers a comprehensive testbed for evaluating robustness in noisy scenarios.

In practical settings, speech signals are rarely clean, necessitating systems that can maintain high recognition performance despite acoustic disturbances. The globally diverse linguistic landscape, especially in regions like India,

further underscores the need for resilient language identification systems capable of functioning reliably in various environmental contexts. Implementing effective feature reduction strategies like PCA directly addresses these challenges by distilling salient speech features and enhancing model robustness.

The study opens with a detailed review of comparable research in the field of speech and language processing, emphasizing the strengths and limitations of existing systems. Subsequently, it describes the proposed methodology, including the feature extraction process using openSMILE, PCA-based feature reduction, and classification via Linear Discriminant Analysis. The experimental setup, including datasets, noise simulation, and evaluation metrics, is then discussed, followed by the presentation and analysis of experimental results.

2. Related Works

Robust ASLID in noisy environments remains a challenging task, prompting researchers to explore advanced feature extraction and dimensionality reduction techniques [8]. While many studies address SLID in clean conditions, fewer focus on noisy scenarios. PCA has emerged as an effective method for reducing feature dimensionality, eliminating redundancy, and enhancing classification accuracy.

Fathoni et al. [14] demonstrated that combining feature extraction methods (e.g., MFCC, GFCC, LFCC) for capturing relevant acoustic information. While feature combination enhances recognition accuracy, it also leads to increased feature dimensionality. To address this, PCA is applied to reduce redundancy features and determine the optimal feature set, followed by classification and evaluation using SVM. PCA can maintain high recognition accuracy (99.38%). Similarly, Ramoji et al. [15] showed that supervised i-vector modeling, when followed by PCA, significantly improves discriminative power.

Noteworthy, the effective feature extraction is critical in noisy environments. Thimmaraja et al. [16] proposed a noise-resilient LPC encoding method enhanced with spectral subtraction and VAD. Experiments are executed using various noisy speech data types affected by musical noise, factory noise, car noise, and F16 noise.

Nassif et al. [17] improved speaker identification by integrating a noise reduction module based on Computational Auditory Scene Analysis (CASA) with a GMM-CNN classifier, achieving high accuracy on emotionally and environmentally varied datasets. Kantamaneni et al. [18] introduced a DNN-based Kalman filter for better LPC estimation in noisy conditions. Biswas et al. [19] addressed multilingual identification with noise-augmented datasets. They designed a model to identify both foreign and Indian languages. To enhance the robustness of the system in noisy environments, various types of everyday noise were incorporated into the datasets. Following the extraction of macro-level features from the supplemented dataset's MFCC time series, the FRESH (Feature Extraction based on Scalable Hypothesis Tests) algorithm was used to pick features. An artificial neural network (ANN) was then trained using the chosen features. When the model was tested on three datasets, it demonstrated remarkable accuracy: Eight languages from the VoxForge dataset had a 98.43% success rate, six languages from the IIT-M IndicTTS speech database had a 99.93% success rate, and seven languages from the IIIT-H Indic speech database had a 99.94% success rate. The time-series feature extraction process is very difficult, although the model performs better than many earlier methods in this area. Notably, it took almost four days to extract the final features from the 240-hour audio sample using an Intel i7 processor and 32 GB of RAM. In contrast, both training and testing on the GPU were relatively quick, taking only about 6–7 hours. Nonetheless, the feature extraction phase remained computationally intensive.

These studies highlight the critical role of PCA and robust preprocessing techniques in enhancing ASLID performance under noisy conditions. Building on this, this research aims to develop an ASLID system that leverages PCA to achieve high accuracy in noisy environments while also prioritizing computational efficiency by reducing processing and training time.

3. Methodology

Automatic spoken language identification in noisy environments presents major challenges in speech processing, primarily due to the degradation of audio quality. The ASLID involves several stages—*data collection*, *feature extraction*, *feature reduction*, and *classification*—that are essential for accurately identifying the spoken language from audio input, particularly when using machine learning techniques. These stages, illustrated in Figure 1, form a comprehensive framework for building robust ML- based ASLID systems. Optimizing each stage is critical for enhancing ASLID performance and supporting more effective multilingual communication and accessibility in a globalized world. While ASLID systems typically

perform well under clean acoustic conditions, this paper proposes a system specifically designed to maintain high language recognition accuracy even in the presence of significant background noise.

3.1. Data Collection

The proposed model begins with collecting speech samples across multiple languages. The quality and diversity of this data plays a critical role in determining the performance of the identification system. These recordings may be from public multilingual corpora or custom datasets comprising diverse speakers and dialects. This stage is vital for training robust models that can generalize well across various accents, dialects, and speaking styles.

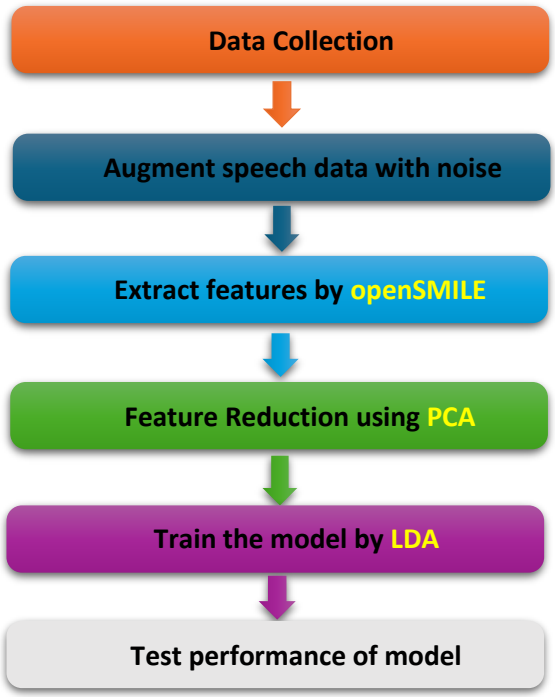


Figure 1 A Flowchart for the proposed model

3.2. Data Augmentation

To simulate real-world conditions, clean audio samples are artificially augmented with different types of environmental noise, including white noise, babble noise, street noise, or Gaussian noise [20] at different Signal-to-Noise Ratios (SNRs) [21]. This augmentation enables the model to generalize well in noisy environments.

In this research, background noise was artificially generated using Gaussian white noise with a mean of zero, applied through Ocenaudio software, and adjusted to an intensity of -24 dB, as illustrated in Figure 2. This approach effectively simulates real-world noisy environments while allowing the system to maintain its language identification accuracy under adverse acoustic conditions. In ASLID or other speech processing applications, Gaussian white noise is often added to audio samples to simulate real-world noise conditions, including background conversation, mechanical vibrations, or environmental noise. The mathematical expression for the discrete-time Gaussian white noise signal is as follows [20]:

$$x[n] = \mu + \sigma . z[n]$$

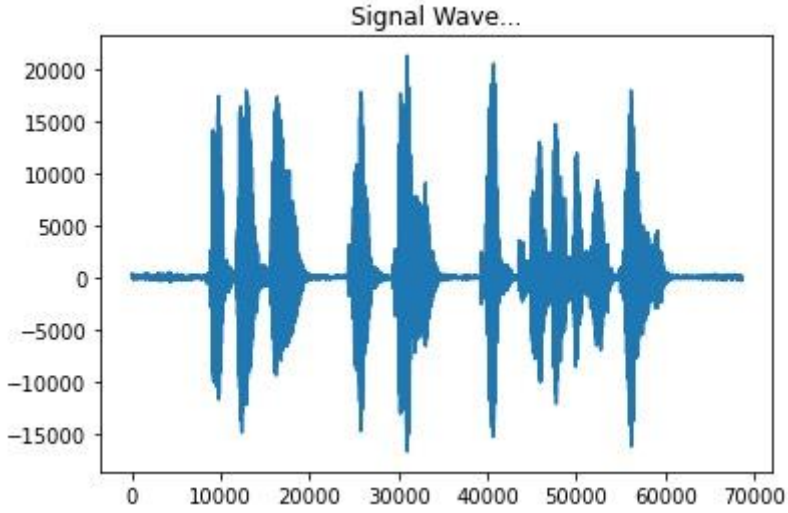
where:
$x[n]$ represents the noise level at time step n.
μ is the mean (often zero).
σ is the standard deviation, which controls the amplitude's spread.

$z[n]$ consists of a series of independent random variables selected from a standard normal distribution. If $y[n]$ represents the original speech signal and $x[n]$ is the noise value at time step n, then the speech signal with the expression for Gaussian white noise is:

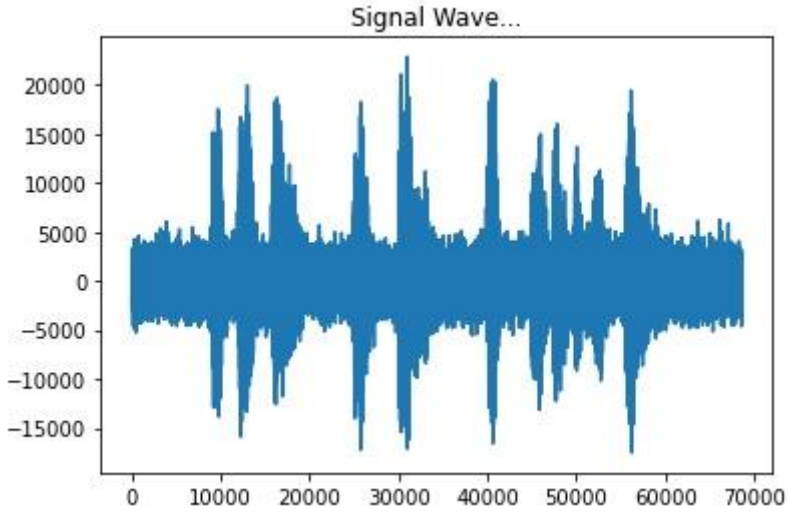
$$y'[n] = y[n] + x[n]$$

3.3. Feature Extraction using openSMILE

Acoustic features are extracted using the **openSMILE** [12] toolkit, a widely-used feature extractor in speech processing. The toolkit robust low-level descriptors (LLDs) such as MFCCs, pitch**,** formants**,** energy, spectral features, voice quality features, emotion-related features, prosodic features, which serve as the input feature vectors for further processing, and other high-level features derived from the speech signal [12,22]. The openSMILE supports three standard feature sets for audio feature extraction: **ComParE 2016**, **GeMAPS**, and **eGeMAPS** [12] as illustrated in Table 1. OpenSMILE was chosen as the feature extraction tool in this research for its capability to extract a diverse set of features using various signal processing techniques.



(a)



(b)

Figure 2 (a) Original speech signal. (b) Speech signal with gaussian white noise.

Table 1. Number of features according to feature sets for each level [12]

Feature sets	Number of features		
	LDD	LLD	Functionals level
ComParE_2016	65	65	6373
GeMAPSv01a	18	-	62
GeMAPSv01b	18	-	62
eGeMAPSv01a	23	-	88
eGeMAPSv01b	23	-	88
eGeMAPSv02	25	-	88

3.4. Feature Reduction using PCA

PCA [23] is a commonly used statistical method for dimensionality reduction. It transforms a high-dimensional feature space into a lower-dimensional subspace while preserving the most significant variability present in the original data. To do this, PCA finds a set of orthogonal basis vectors, or *principle components* (PCs), that are linear combinations of the original variables and are arranged based on how much variance they represent. In speech and language processing tasks, PCA helps reduce noise, minimize computational complexity, and enhance classifier performance by eliminating redundant or less informative features. PCA not only enhances the robustness of the system in noisy environments but also reduces computational overhead during training and inference. Figure 3 illustrate a flowchart for PCA algorithm.

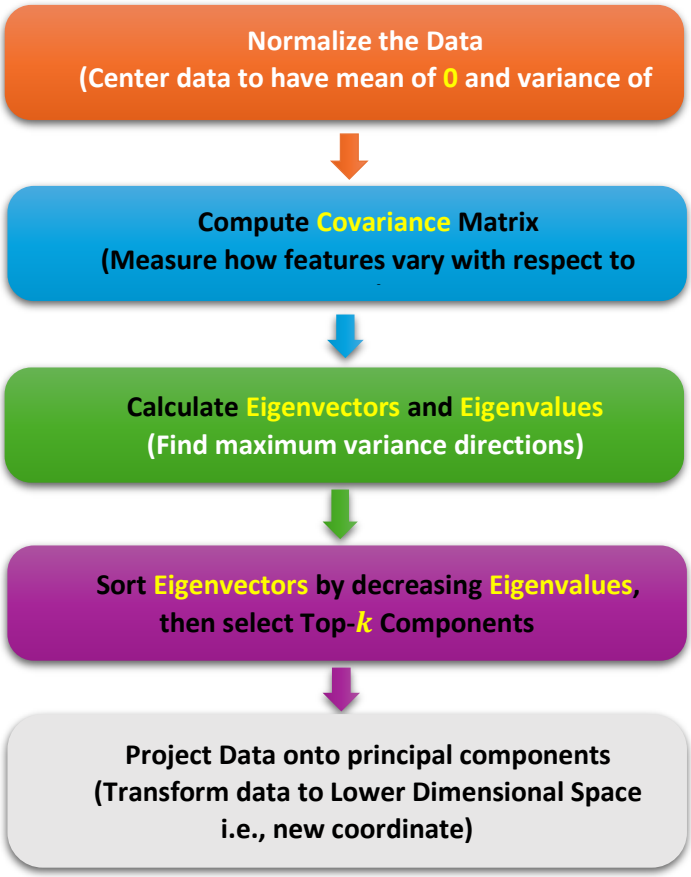


Figure 3 A Flowchart for PCA algorithm

3.5. Training the Model using LDA

The reduced features are used to train a LDA classifier. When training an ASLID model, selecting the appropriate machine learning model is crucial for achieving accurate and efficient language classification. LDA is chosen for its efficiency in maximizing class separability in a low-dimensional space and minimizing the within-class variance that making classes more compact and well-separated and making it suitable for language classification tasks with limited data [24,25]. So, this research considers the LDA classifier to be an effective machine learning model for achieving high performance in noisy environments.

Algorithm 1. Linear Discriminant Analysis (LDA) Classifier

Given a dataset with N samples and M features, where each sample x_i belongs to one of K classes, LDA computes the following:

1. **Determine the Mean Vectors.**

   Calculate the mean vector corresponding to each class as well as the overall mean vector of all samples.

2. **Within-Class Scatter Matrix ($\mathbf{S_W}$).**

   Measures how much the samples within each class deviate from the class mean.

$$S_W = \sum_{i=1}^{c} \sum_{x \epsilon D_i} (x - \mu_i)\,(x - \mu_i)^T$$

   where:

   c refers to the quantity of classes,
   D_i represents the class i data,
   x is a data point in class i,
   μ_i is the mean vector of class i.

3. **Between-Class Scatter Matrix (S_B).**

   Calculates how far apart the class means are from the overall mean.

$$S_B = \sum_{i=1}^{c} N_i\,(\mu_i - \mu)(\mu_i - \mu)^T$$

   Where:

   N_i is the number of samples in class i,
   μ is the overall mean of all classes.

4. **Calculate the Linear Discriminants (Optimal Projection).**

   The goal is to find a projection matrix that maximizes the ratio of the between-class variance to the within-class variance by solving the following generalized eigenvalue problem.

$$S_W^{-1}\,S_B\,\omega = \lambda\omega$$

   Where:

   ω is the matrix of linear discriminants (i.e., the directions that maximize class separability),
   λ is the eigenvalue.

5. **Projection of Data.**

   Project the data points onto the new axes defined by the eigenvectors corresponding to the largest eigenvalues. For two classes, this will be a single line, while for multiclass problems, it may be several dimensions.

6. Classify a new observation.

In order to classify a new observation x, LDA evaluates the discriminant score corresponding to each class and allocates x to the class exhibiting the maximum score. The discriminant function for every class k is articulated as:

$$g_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln(P_k)$$

Where:

Σ is the common covariance matrix,
$\mu_k$is the mean vector for class k,
P_k is the prior probability of class k.

4. Results and Discussion

The task of ASLID in noisy environments poses a significant challenge for automatic systems. It is crucial to differentiate between languages despite interference from background noise, which can severely impact the accuracy of the system. The experimental findings of the suggested system are shown in this part along with an analysis of how well it performs in noisy environments. The IIIT-H dataset was used to start the experiment.

4.1. Performance Evaluation Metrics

The language identification system performance is assessed using the following [26]:

1. **Accuracy:** The percentage of correct language identifications out of all test samples.

$$Accuracy = \frac{\text{Number of correct language identifications}}{\text{Total number of test samples}}$$

2. **Confusion Matrix:** To analyze the misclassifications and understand which languages are most commonly confused by the system.

4.2. IIIT-H Indic Speech Dataset

The IIIT-H Indic voice dataset is a useful resource for advancement in the field of speech processing, especially in the context of the Indian subcontinent's linguistic diversity. By providing a large and varied dataset across multiple languages, it provides an opportunity to build more accurate, robust, and comprehensive speech recognition systems, and supporting the development of AI-driven applications that can cater to the diverse languages spoken in India.

The dataset contains speech samples across a diverse range of Indic languages, including Telugu, Tamil, Hindi, Bengali, Kannada, Malayalam, and Marathi [27]. Theses languages were selected based on two main criteria: each language had over 10,000 Wikipedia articles, and native speakers were readily available on campus. The text corpus consisted of Wikipedia articles in Indian languages, from which 1,000 sentences were selected to cover the 5,000 most frequently occurring words in each language's corpus. The text data is provided in two formats: IT3 (a transliteration technique) and Unicode (UTF-8). Native speakers of each language recorded their speech in a studio setting with a typical headset microphone attached to a Zoom handy recorder. The choice of a handy recorder was driven by its portability and ease of use, while the headset microphone ensured that the distance between the mouth and microphone remained constant, helping maintain consistent recording levels.

4.2.1. Results and Performance of IIIT-H Dataset in a Clean Environment

Table 2 and Figure 4 present the classification performance of the proposed system in a clean environment under varying numbers of principal components and test set proportions 0.2 to 0.5. The *optimal range of principal components* is from 1000 to 2500 because of the proposed system consistently achieves near-

perfect performance (≥99.9%) across all test proportions. This provides that the essential variance required for accurate classification is captured within the first 2500 components. As the number of principal components increases beyond 2500, a gradual degradation in classification accuracy is observed, particularly at higher test proportions 0.4 and 0.5. This indicates that including too many components may reintroduce noise or overfit the model to the training data, especially in smaller training sets.

For the impact of test proportion, Table 2 shows that lower test proportions 0.2 or 0.3 are more resilient to increases in dimensionality, maintaining high accuracy even with up to 4000 components. Whilst, higher test proportions 0.4 or 0.5 show significant sensitivity, with accuracy sharply declining as the number of components increases. For instance, at 5000 components and a test proportion of 0.5, performance drops to 9.31%, indicating severe overfitting and loss of generalization.

In short, achieving robust performance for the proposed model necessitates the careful selection of an optimal number of principal components. Empirical results indicate that retaining approximately 1500 to 2500 components offers an effective trade-off between preserving discriminative information and suppressing noise across different test proportions. This range consistently yields high classification accuracy while minimizing the risk of overfitting. Therefore, dimensionality must be judiciously tuned in relation to the available training data to ensure strong generalization and reliable model performance.

Table 2. The accuracy of the proposed system in a clean environment at various test set proportions and various number of PCs

	Test proportions			
Number of principal components	**0.2**	**0.3**	**0.4**	**0.5**
1000	**100%**	**100%**	**100%**	**100%**
1500	**100%**	**100%**	**100%**	99.97%
2000	**100%**	**100%**	99.96%	99.94%
2500	**100%**	**100%**	99.96%	99.88%
3000	**100%**	99.90%	99.89%	99.48%
3500	**100%**	99.90%	99.78%	41.08%
4000	99.78%	99.76%	95.82%	24.25%
4500	99.85%	98.57%	28.14%	12.25%
5000	99.28%	32.76%	12.42%	9.31%
5500	81.14%	12.85%	8.25%	8.97%

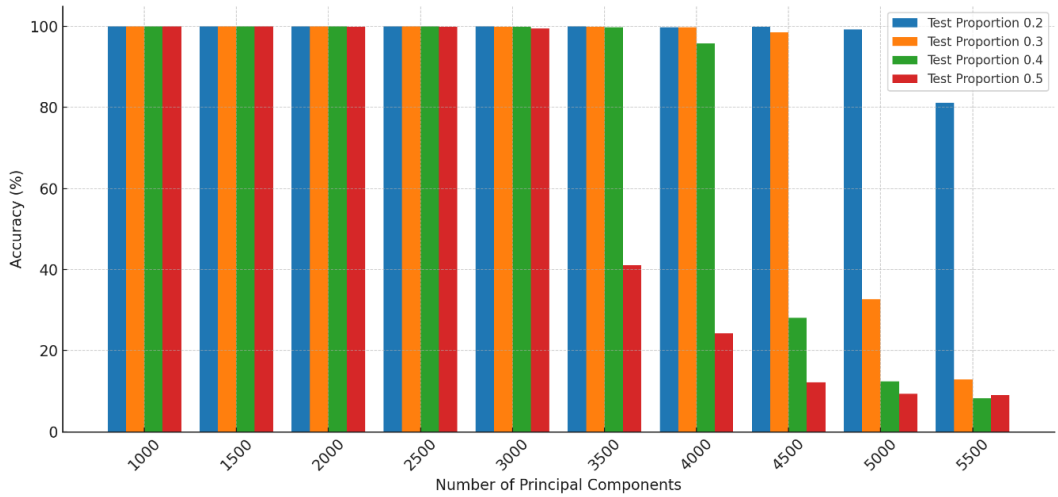


Figure 4 Accuracy for the proposed model in a clean environment vs number of PCs at various test proportions

Figures 5 to 8 show the confusion matrices generated by the proposed system in a clean environment for identifying 7 languages, using different test set proportions and numbers of principal components.

4.2.2. Impact of Noise on Language Identification

After incorporating background noise into the audio files, it is essential to reiterate the methodologies employed, beginning with feature extraction and progressing through classification, followed by a comprehensive analysis of the impact of noise on the system performance.

Table 3 and Figure 9 illustrate the classification accuracy of the proposed language identification system under noisy conditions, evaluated at various test set proportions 0.2 to 0.5 and numbers of principal components ranging from 1000 to 5500. The results demonstrate that the system maintains high accuracy when the number of principal components is kept between 1000 and 2500, with accuracy consistently exceeding 99% across all test proportions. This indicates strong robustness of the system to noise within this dimensionality range.

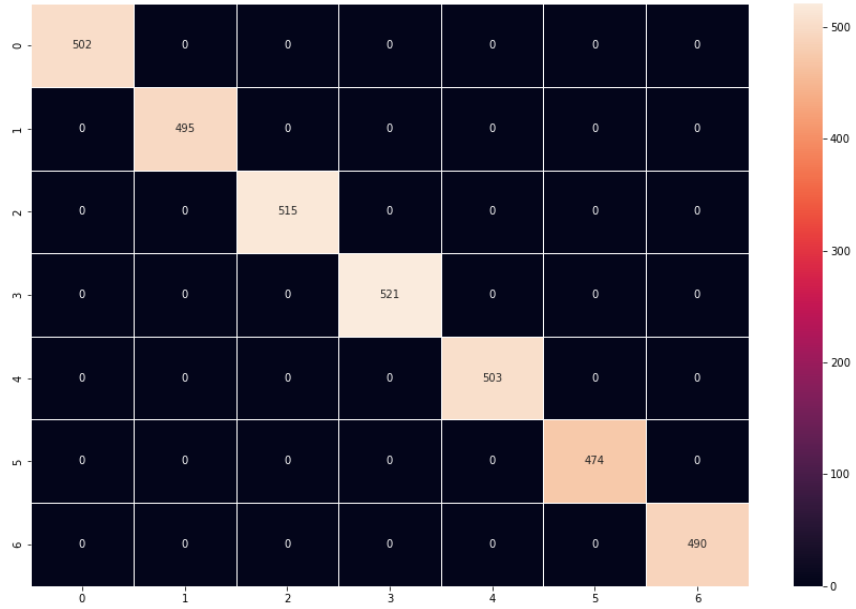


Figure 5 Confusion matrix of the proposed system in a clean environment for identifying 7 languages, with a test set proportion of 0.5 and 1000 PCs used

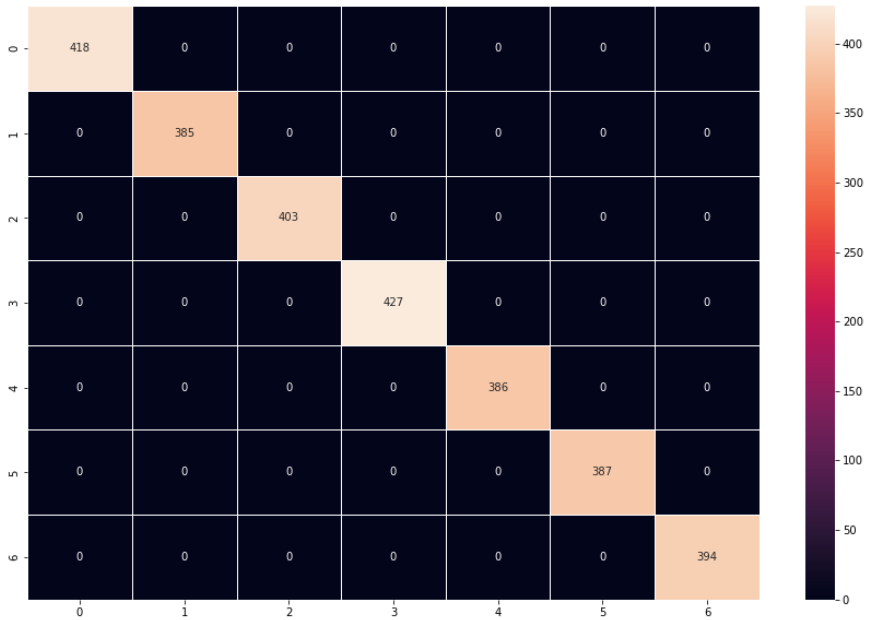


Figure 6 Confusion matrix of the proposed system in a clean environment for identifying 7 languages, with a test set proportion of 0.4 and 1000 PCs used

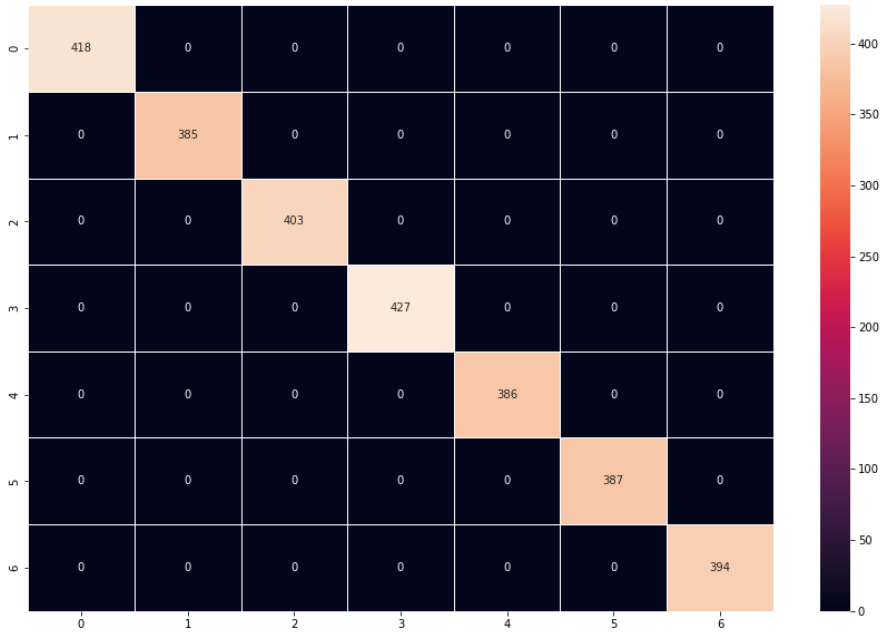


Figure 7 Confusion matrix of the proposed system in a clean environment for identifying 7 languages, with a test set proportion of 0.4 and 1500 PCs used

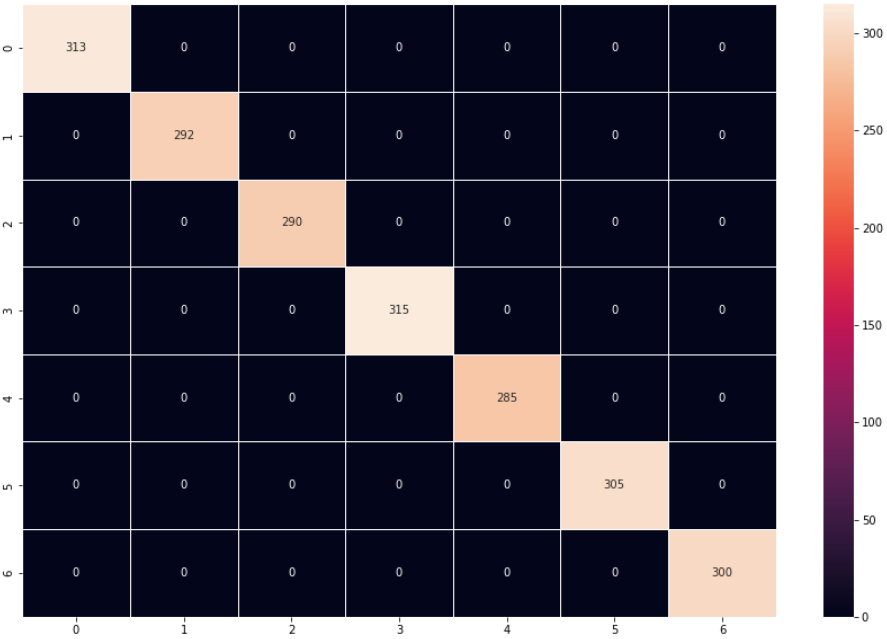


Figure 8 Confusion matrix of the proposed system in a clean environment for identifying 7 languages, with a test set proportion of 0.3 and 2500 PCs used

In summary, the experimental results underscore the critical role of dimensionality reduction and precise parameter optimization in enhancing system performance under noisy conditions. The optimal number of principal components lies in the range of 1000 to 2500, within which the system consistently achieves high and stable classification accuracy across varying test set proportions. Beyond this range, the performance degrades markedly, particularly in the presence of increased noise and larger test partitions, likely due to the adverse effects of overfitting and the inclusion of noise-sensitive features. Notably, the same optimal range of principal components was observed to be effective in both clean and noisy environments. This consistency highlights the robustness and generalizability of the proposed model, which demonstrates comparable

performance regardless of environmental conditions, thereby confirming its efficacy in noise-resilient speaker recognition.

Figures 10 to 13 show the confusion matrices generated by the proposed system in a noisy environment for identifying 7 languages, using different test set proportions and numbers of principal components.

Table 3. The accuracy of the proposed system in a noisy environment at various test set proportions and various number of PCs

	Test proportions				
Number of principal components		**0.2**	**0.3**	**0.4**	**0.5**
	1000	99.78%	99.61%	99.75%	**99.62%**
	1500	99.71%	**99.85%**	99.71%	99.60%
	2000	**99.92%**	**99.85%**	**99.78%**	99.31%
	2500	**99.92%**	**99.85%**	99.71%	99.08%
	3000	99.85%	99.57%	99.17%	96.57%
	3500	99.78%	99.42%	98.03%	25.34%
	4000	99.64%	98.47%	80.50%	25.20%
	4500	99.50%	92.90%	29.89%	13.57%
	5000	97.50%	32.80%	13.64%	9.28%
	5500	66.14%	13.19%	9.03%	7.51%

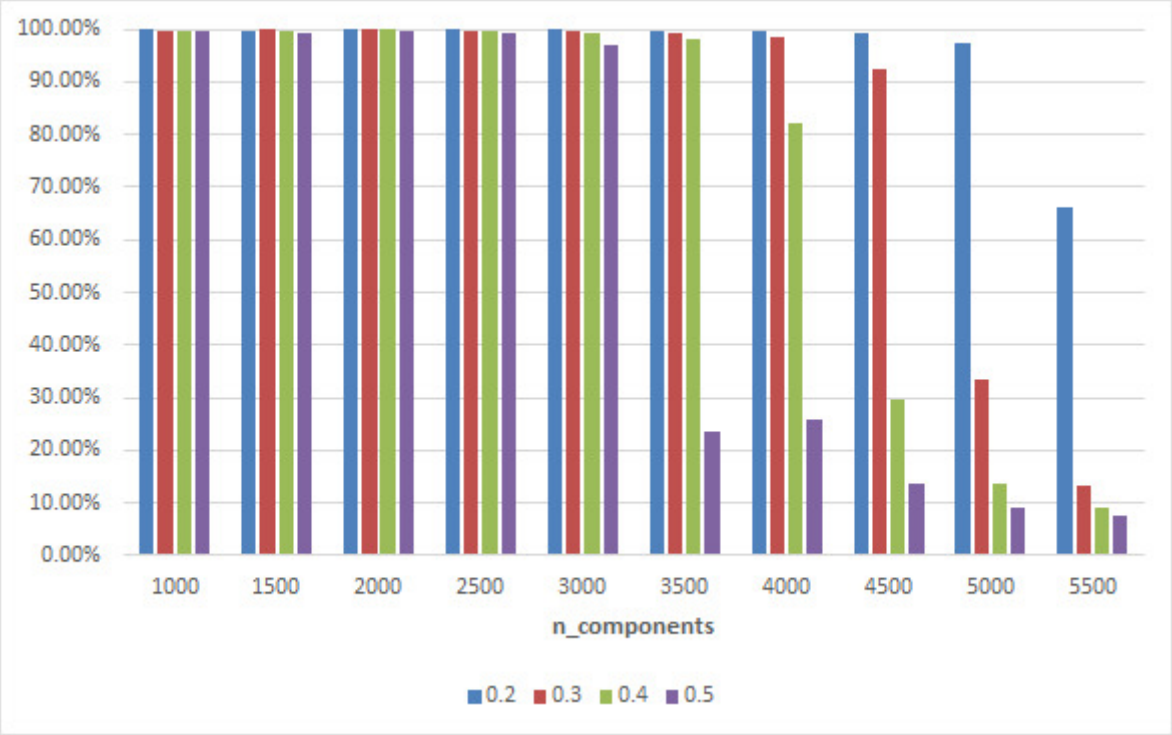


Figure 9 Accuracy for the proposed model in a noisy environment vs number of PCs at various test proportions

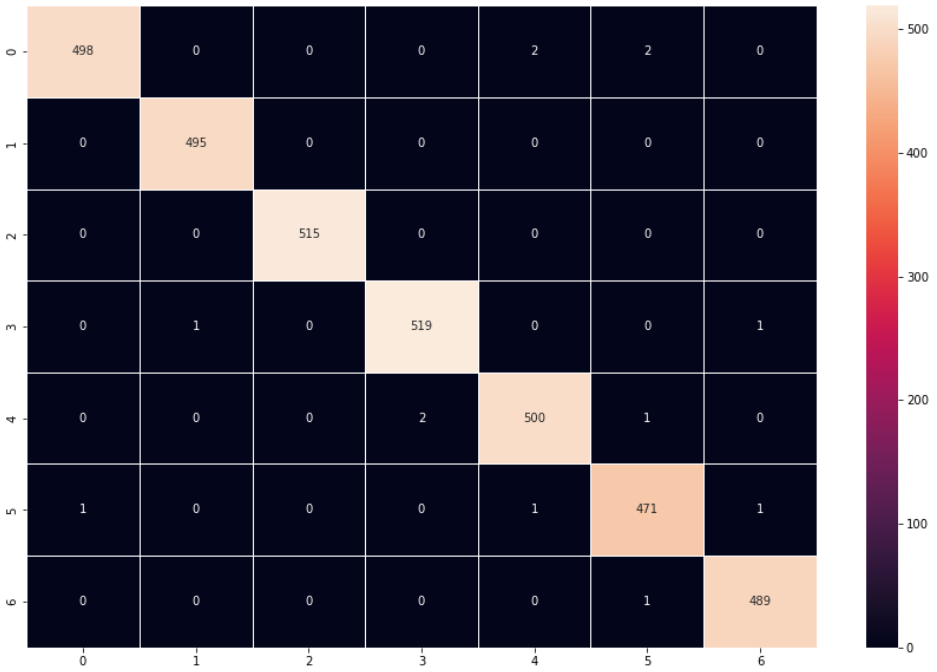


Figure 10 Confusion matrix of the proposed system in a noisy environment for identifying 7 languages, with a test set proportion of 0.5 and 1000 PCs used

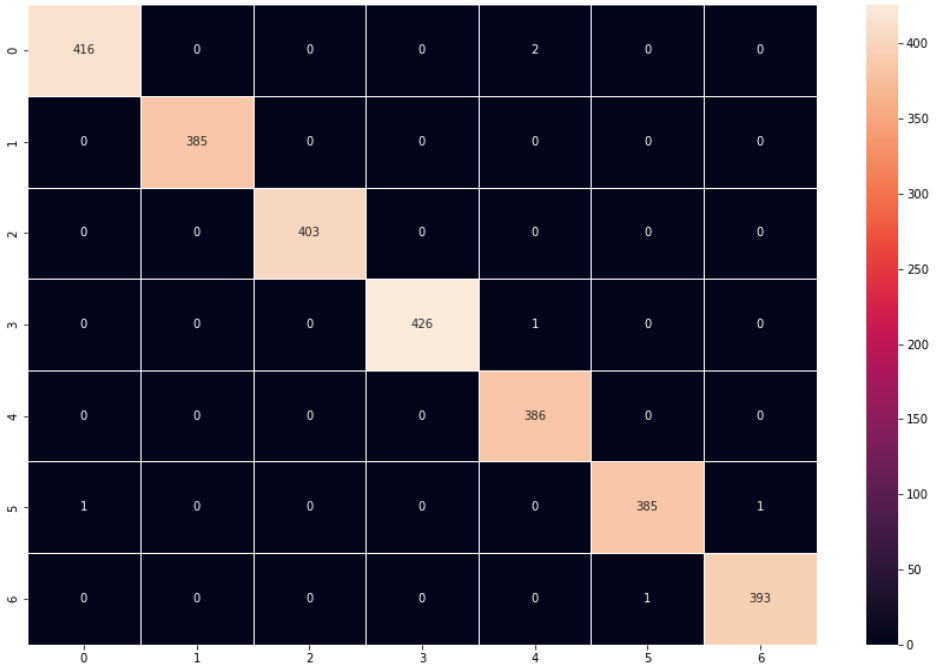


Figure 11 Confusion matrix of the proposed system in a noisy environment for identifying 7 languages, with a test set proportion of 0.4 and 2000 PCs used

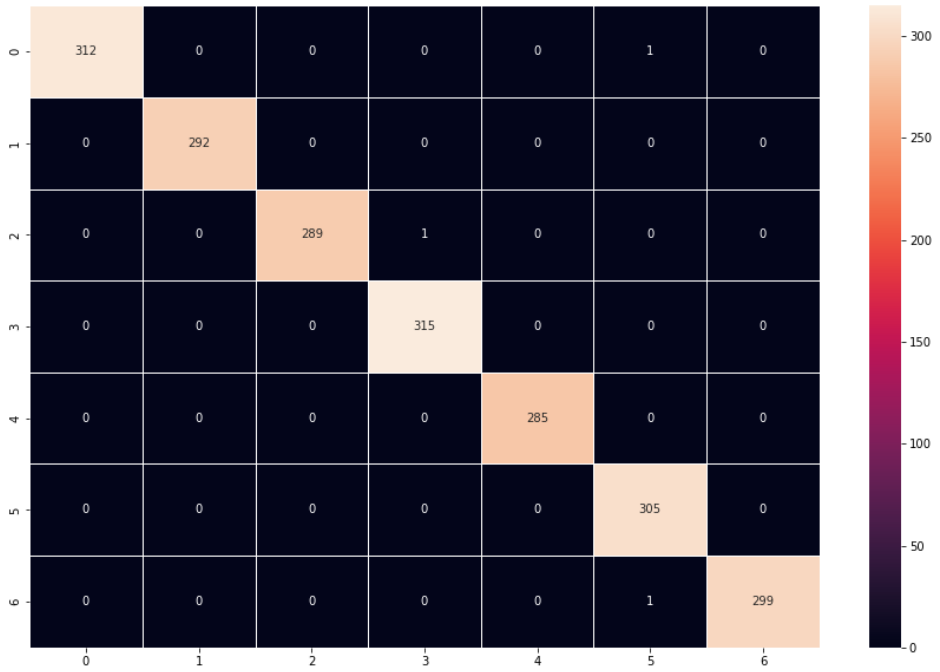


Figure 12 Confusion matrix of the proposed system in a noisy environment for identifying 7 languages, with a test set proportion of 0.3 and 2000 PCs used

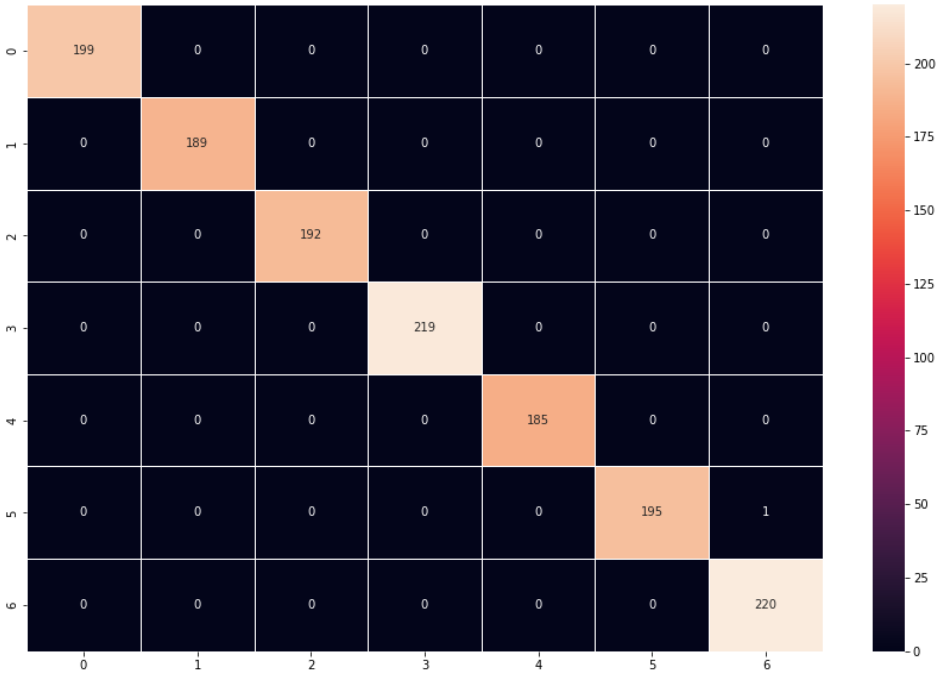


Figure 13 Confusion matrix of the proposed system in a noisy environment for identifying 7 languages, with a test set proportion of 0.2 and 2000 PCs used

4.2.3. Performance Evaluation

Table 4 presents a comparative analysis of various SLID techniques applied to the IIIT-H dataset in both clean and noisy environments. The table includes established methods from prior research, such as the use of traditional feature extraction techniques (e.g., LPC and MFCC) coupled with classical classifiers (e.g., SVM and Random Forest) by Gupta et al. [28], as well as more recent deep learning approaches, including CNN-based architectures [29], ensemble learning [30], and LSTM-based sequence classification models [31].

Among the models tested in clean environments, deep learning approaches generally demonstrate high accuracy, with results ranging from 99.50% to 99.94%. Notably, Paul et al. [31] achieved 99.80% using LSTM, and Biswas et al. [19] achieved 99.94% with MFCC-based time series features and neural networks in noisy conditions, representing the strongest performance among prior studies under adverse conditions.

The proposed model, which integrates PCA for dimensionality reduction and LDA for classification, outperforms all previously reported methods. It achieves a perfect accuracy of 100% in clean environments across multiple test proportions (0.20, 0.30, 0.40, and 0.50), demonstrating exceptional robustness and consistency. Even under noisy conditions, the proposed system achieves 99.92% accuracy at a 0.20 test proportion, surpassing other noise-robust systems such as that of Biswas et al. [19].

These results highlight several key insights. First, the proposed model effectively balances computational efficiency and classification performance through PCA-based feature reduction, which reduces redundancy and enhances discriminative capability. Second, LDA proves to be a highly suitable classifier for ASLID tasks, providing strong generalization in both clean and noisy environments. Finally, the model's superior performance across varying test proportions and noise levels confirms its robustness, scalability, and suitability for real-world applications where environmental variability is a critical factor.

Table 4. Performance measures obtained from various SLID techniques and the proposed model of IIIT-H Dataset in a clean and noisy environment

Reference	Technique	Test Proportional	Accuracy	Noise
Gupta et al. (2017) [28]	LPC and MFCCs for features extraction and SVM and Random Forest (RF) as classification techniques for language identification	0.30	92.60%	no
Athira et al. (2019) [29]	Deep Learning by CNN		99.50%	no
Mukherjee et al. (2020) [30]	Line Spectral Frequency (LSF) features combined with an ensemble learning-based classification approach		99.71%	no
Paul et al. (2021) [31]	MFCC and Pitch feature extraction methods and a Long Short-Term Memory (LSTM) sequence classification	0.20	99.80%	no
Biswas et al. (2023) [19]	MFCC based time series features and NN	0.05	99.94%	yes
Proposed Model	PCA for Feature reduction and LDA for language identification	0.20, 0.30, 0.40, 0.50	100.00%	no
Proposed Model	PCA for Feature reduction and LDA for language identification	0.20	99.92%	yes

Table 5 presents a comparative evaluation of language identification performance in a noisy environment using several CNN architectures—modified ResNet50, VGG16, and Inception-v3—combined with classical ML classifiers and synthetic voice data augmentation and PCA.

The application of PCA across all architectures contributes to improved accuracy in most cases by reducing feature dimensionality and enhancing class separability. Among the classical classifiers, LR, and KNN show notably high performance when paired with VGG16 and Inception-v3 features, reaching up to 94.80% and 97.00%, respectively. Modified versions of Inception-v3 and VGG16 achieve strong standalone results with 97.00% and 97.10% accuracy, indicating that deep convolutional features combined with PCA can be highly effective for language identification in noisy environments.

However, the modified ResNet50 model demonstrates comparatively lower performance, with most classifiers achieving below 70% accuracy, except for Random Forest 90.70% and its standalone use 93.30%. This suggests that ResNet50 features may be less suited to capturing language-specific acoustic patterns in noisy conditions, even with PCA applied.

Despite the strong performance of CNN-based models, the proposed model—utilizing PCA for feature reduction and LDA for classification— surpasses all CNN-based methods with an accuracy of 99.92%, significantly outperforming all tested configurations. This result underscores the effectiveness of the proposed approach, which achieves superior robustness and generalization without the computational complexity of DNNs.

Furthermore, the proposed model achieves this performance with significantly lower computational complexity than deep CNNs, which require substantial training time, data, and processing resources. This highlights a key strength of the proposed system: it provides state-of-the-art accuracy with greater efficiency and interpretability, making it highly suitable for real-time or resource-limited applications.

Overall, these results validate the efficacy of PCA as a preprocessing step and confirm the superiority of the proposed model framework in noisy environments, making it a highly competitive and computationally efficient alternative to deep learning-based ASLID systems.

Table 5. Performance measures for the modified RESNET50, VGG16, and Inception-v3 models with PCA [32] alongside the proposed model, evaluated on the IIIT-H Dataset in a noisy environment

Classifier	**Classical Augmentation with PCA**
Synthetic Voice Data Augmentation with modified VGG16	
RF	85.40%
Support Vector Machine	91.40%
Decision Tree (DT)	76.90%
K-Nearest Neighbors	92.90%
Logistic Regression (LR)	94.80%
Naïve Bayes (NB)	83.20%
modified VGG16	**97.10%**
Synthetic Voice Data Augmentation with modified RESNET50	
RF	90.70%
Support Vector Machine	69.80%
Decision Tree (DT)	69.80%
K-Nearest Neighbors	56.70%
Logistic Regression (LR)	62.70%
Naïve Bayes (NB)	69.80%
modified RESNET50	**93.30%**
Synthetic Voice Data Augmentation with modified Inception-v3	
RF	96.30%
Support Vector Machine	95.00%
Decision Tree (DT)	74.00%
K-Nearest Neighbors	**97.00%**
Logistic Regression (LR)	94.00%
Naïve Bayes (NB)	84.90%
modified Inception-v3	**97.00%**
Proposed Model	**99.92%**

5. Conclusion

This paper presents a robust framework for automatic spoken language identification in noisy environments by integrating PCA and LDA. The proposed approach leverages OpenSMILE for comprehensive feature extraction, followed by PCA to reduce the dimensionality and enhance computational efficiency. LDA is subsequently applied to maximize class separability, improving the system's ability to distinguish between languages even under challenging acoustic conditions. Experimental results demonstrate that the proposed model outperforms various classification techniques, achieving high accuracy across varying noise levels, test set proportions, and number of principal components. The proposed model achieves an accuracy of 99.92% on the IIIT-H Indic speech dataset in a noisy environment. These findings highlight the model's

effectiveness in maintaining high accuracy despite the presence of noise, reinforcing its applicability in real-world multilingual speech processing tasks. The paper establishes the proposed model as a promising solution for ASLID, with potential future enhancements through deep learning-based feature representations and advanced noise adaptation strategies to further enhance performance in challenging acoustic environments.

References

1. Bagi, R., Yadav, J., Rao, K. (2015). Improved recognition rate of language identification system in noisy environment. In Eighth International Conference on Contemporary Computing (pp.214-219), Noida, India.
2. Yu-bin, S., Jing, L., Hua, L., Yi-min, L. (2021). Language Identification in Real Noisy Environments. Journal of Beijing University of Posts and Telecommunications 44(6), 134.
3. Kilimci, H., Kilinc, H., Kilimci, Z. (2025). Automatic Language Identification from Speech using Transformer-Based Models. In 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) (pp.1-7).
4. Barnard, E., Cole, R. (1994). Reviewing automatic language identification. IEEE Signal Processing Magazine 11(4), 33–41.
5. O'Shaughnessy, D. (2025). Spoken language identification: An overview of past and present research trends. Speech Communication, 167.
6. Rai, M., Fahad, M., Yadav, J., Rao, K. (2016). Language identification using PLDA based on i-vector in noisy environment. In 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp.1014-1020).
7. Sáez, J., Luengo, J., Herrera, F. (2016). Evaluating the classifier behavior with noisy data considering performance and robustness: The equalized loss of accuracy measure. Neurocomputing, 176, 26-35.
8. H, M., Gupta, S., Dinesh, D., Rajan, P. (2021). Noise-Robust Spoken Language Identification Using Language Relevance Factor Based Embedding. In IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, IEEE.
9. Makhoul, J. (2005). Linear prediction: A tutorial review. In Proceedings of the IEEE (pp.561 - 580), vol. 63.
10. Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. Pattern Recognition and Artificial Intelligence, 374-388.
11. Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America 87(4), 1738–1752.
12. Eyben, F., Wöllmer, M., Schuller, B. (2010). OPENSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor. In Proceedings of the 9th ACM International Conference on Multimedia (pp.1459-1462).
13. Singh, G., Sharma, S., Kumar, V., Kaur, M., MohammedBaz, Masud, M. (2021). Spoken Language Identification Using Deep Learning. Computational Intelligence and Neuroscience, 1–12.
14. Fathoni, A., Hidayat, R., Bejo, A. (2022). Optimization of Feature Extraction in Indonesian Speech Recognition Using PCA and SVM Classification. In 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia.
15. Ramoji, S., Ganapathy, S. (2018). Supervised I-vector Modeling-Theory and Applications. INTERSPEECH, 1091-1095.
16. Thimmaraja, Y., Nagaraja, B., Jayanna, H. (2021). Speech enhancement and encoding by combining SS-VAD and LPC. International Journal of Speech Technology, 24, 165–172.
17. Nassif, A., Shahin, I., Hamsa, S., Nemmour, N., Hirose, K. (2021). CASA-based speaker identification using cascaded GMM-CNN classifier in noisy and emotional talking conditions. Applied Soft Computing, 103.
18. Kantamaneni, S., Charles, A., Babu, T. (2023). Speech enhancement with noise estimation and filtration using deep learning models. Theoretical Computer Science, 941, 14-28.
19. Biswas, M., Rahaman, S., Ahmadian, A., Subari, K., Singh, P. (2023). Automatic spoken language identification using MFCC based time series features. Multimedia Tools and Applications, 82, 9565–9595.
20. Salamon, J., Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. IEEE Signal Processing Letters, 24(3), 279-283.
21. Luo, Y., Mesgarani, N. (2019). Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. IEEE/ACM Transactions on Audio Speech and Language Processing, 27(8), 1256-1266.
22. audEERING. Available at: https://www.audeering.com/research/opensmile/.
23. Jolliffe, I. (2002). Principal Component Analysis 2nd edn. Springer , New York.
24. Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7(2), 179-188.
25. Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B. (2013). Linear Discriminant Analysis. In Robust Data Mining. Springer, New York.
26. Olson, D., Delen, D. (2008). Advanced Data Mining Techniques 1st edn. Springer.
27. Prahallad, K., Kumar, E., Keri, V., Rajendran, S., Black, A. IIIT-H Indic Speech Databases, IIIT Hyderabad, India. (Accessed 2024) Available at: http://festvox.org/databases/iiit_voices/.
28. Gupta, M., Bharti, S., Agarwal, S. (2017). Implicit language identification system based on random forest and support vector machine for speech. In 4th International Conference on Power, Control & Embedded Systems (ICPCES), Allahabad, India.

29. Athira, N., Poorna, S. (2019). Deep learning based language identification system from speech. In International Conference on Intelligent Computing and Control Systems (ICCS) (pp.1094-1097), Madurai, India.
30. Mukherjee, H., Das, S., Dhar, A., Obaidullah, S., Santosh, K., Phadikar, S., Roy, K. (2020). An ensemble learning-based language identification system. In Computational Advancement in Communication Circuits and Systems: Proceedings of ICCACCS, 2018.
31. Paul, B., Phadikar, S., Bera, S. (2021). Indian regional spoken language identification using deep learning approach. In Proceedings of the Sixth International Conference on Mathematics and Computing: ICMC 2020 (pp.263-274), Singapore.
32. AMBILI, A., ROY, R. (2023). The Effect of Synthetic Voice Data Augmentation on Spoken Language Identification on Indian Languages. IEEE Access, 11, 102391 - 102407.