The Problem of Forming a Feature Space for Speaker Recognition

Nomaz Mirzaev[1][0000-0002-0657-2654], Johongir Urinboev[2][0009-0000-2305-8887],

Sayyora Ibragimova[2][0000-0002-8118-8919], Gulmira Mrzaeva[1][0000-0002-0412-3951],

Azizbek Tillavoldiev[2][0009-0004-7579-2185]

[1] Tashkent university of information technologies named after Muhammad al-Khwarizmi Tashkent, Uzbekistan
[2] Digital Technologies and Artificial Intelligence Research Institute Tashkent, Uzbekistan
jahongir8010@gmail.com

Abstract. This work examines the issue of person identification based on voice signals. Speech-based biometric systems are widely used in information security, human-computer interaction, and intelligent information systems. However, noise, channel effects, and high dimensionality of features can negatively affect the system's accuracy. This study investigated the development of noise-resistant and computationally efficient recognition algorithms based on a voice dataset consisting of 50 English speakers. Acoustic features such as MFCC, LPC, PLP, spectral centroid, energy, and entropy were extracted from speech signals, and a single feature vector was formed through feature fusion of these features. In this research, person recognition methods were thoroughly studied, with the main focus on two important approaches: feature fusion and feature space reduction using Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Extensive experiments were conducted on various speech datasets characterized by different noise levels and numbers of speakers. The study yielded good results for a single dataset and classifiers. To reduce redundant and mutually correlated parts of features, dimensionality reduction methods - PCA and ICA - were applied to the feature space. The resulting features were evaluated using K-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA) classifiers. These results indicate a significant increase in computational speed and speaker recognition efficiency due to the feature fusion and feature space reduction approaches.

Keywords Speaker identification, speaker verification, feature extraction, feature fusion, feature space reduction, PCA, ICA, KNN, LDA.

1 Introduction

This work presents an extended analysis of speaker recognition (SR) systems, focusing on feature fusion methods applied to speech datasets of various sizes [1]. In the approach proposed in this study, the model integrates 18 distinct features via feature

fusion, including mel-frequency cepstral coefficients (MFCC), linear predictive coding (LPC), perceptual linear prediction (PLP), root mean square (RMS) value, centroid and entropy features, as well as their corresponding delta (Δ) and delta-delta (ΔΔ) feature vectors. Experimental results demonstrated that feature fusion of different features together with their associated delta and delta-delta components leads to a significant increase in speaker recognition accuracy. This resulted in improved recognition rates for noise-free voice data using linear discriminant (LD) and K-nearest neighbor (KNN) classifiers. Furthermore, the Equal Error Rate (EER) for speaker verification was reduced compared to using a single feature.

Feature fusion of numerous features may appear beneficial for capturing diverse aspects of the data; however, it can introduce dimensionality-related challenges, including increased computational complexity, overfitting, and reduced efficiency. In the effective design of speaker recognition systems, it is crucial to balance the number of features with the desired performance indicators and computational constraints. To address this issue, methods for reducing the feature space are employed, including Principal Component Analysis (PCA) and Independent Component Analysis (ICA). These techniques compress data by reducing dimensionality while preserving the most important features. Consequently, the training process accelerates and computational efficiency improves [2,3].

The main achievements of this work are as follows:

• We have conducted a detailed study of methods for reducing the feature space, specifically designed to address the complexities of high-dimensional data in speaker recognition systems.

• Overall, the research emphasizes the importance of feature fusion in speaker recognition systems and demonstrates the advantages of feature space reduction methods.

• The aim is to achieve an optimal feature fusion strategy that not only improves recognition accuracy but also reduces the dimensionality of the feature space and enables faster computation.

• Our proposed models provide reliable solutions for improving the efficiency of recognizing various numbers of speakers in noisy environments, using English-language speech datasets containing 50 speakers. Our models are adaptable and can be applied to a wide range of application domains and datasets with different scales and characteristics.

2 Problem Statement

The primary objective of the research is to enhance the accuracy and computational efficiency of speaker recognition algorithms by integrating various acoustic features extracted from voice signals and reducing their dimensionality using PCA and ICA methods.

- To achieve this goal, the following tasks were established:
- extracting MFCC, LPC, PLP, and additional spectral features from speech signals;
- fusing the features into a single feature-level vector;
- reducing the dimensionality of features using PCA and ICA methods;

- evaluating identification accuracy using KNN and LDA classifiers;
- comparing and analyzing the obtained results.

3 Proposed methodology

3.1 Exploration of the problem

Our primary goal is to develop the most optimal set of features that enhance accuracy, reduce errors, and improve computational efficiency, thereby conserving computational resources. Moreover, we aim to create a universal approach that can be applied to datasets of varying sizes. In the field of speaker recognition, relatively few studies have focused on computational time. Our proposed work addresses this gap by providing detailed information about computational time, which is a crucial factor alongside accuracy and error rates. To achieve this objective, we present two distinct approaches for speaker identification (SI).

1. Methodology for feature fusion: Feature fusion, which transforms features from various sources or databases into a single, enriched feature set, is considered a fundamental approach in speaker recognition systems. Spectral features, encompassing frequency, power (energy), and other signal properties such as MFCC, LPC, PLP, spectral centroid, and entropy, provide a stable foundation. On the other hand, prosodic features include auditory characteristics such as stress, pitch variation, and intonation (e.g., RMS). Although spectral features, especially MFCC, have demonstrated higher performance compared to pitch-based systems (e.g., pitch, RMS), their combined use provides unparalleled stability, which is crucial for recognition systems. Furthermore, feature derivatives enable the quantitative assessment of subtle changes in voice signals. We aimed to increase accuracy by extracting six features and calculating their derivatives and second-order derivative values [6-9].

2. Feature space reduction: This approach focuses on reducing the dimensionality of the feature set, optimizing computational processes, while preserving crucial information for accurate speaker recognition. Principal Component Analysis (PCA) and Independent Component Analysis (ICA) are applied in our work to reduce the feature space. However, it should be noted that feature fusion is a complex process that may slow down the computational process. Therefore, carefully considering the balance between PCA and ICA is important to ensure equilibrium between computational efficiency and information retention [13-16].

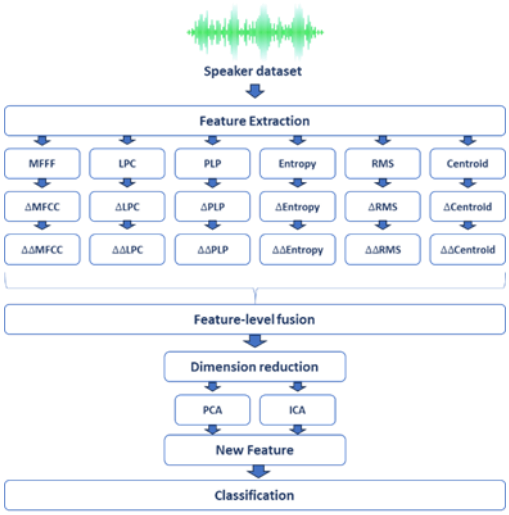

Fig. 1. Proposed approach for the speaker recognition model.

3.2 The approach to feature fusion (Approach 1)

Feature fusion is the process of harmonizing features extracted from various sources or databases to create a comprehensive and enriched set of features. This integration enhances the distinctiveness of speech features, thereby facilitating more accurate classification and speaker recognition [1]. Figure 2 illustrates the approach to feature fusion applied to voice data.

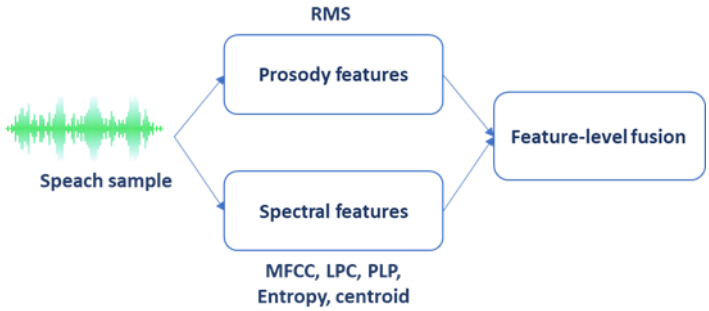

Fig. 2. Method of feature fusion

Methods of feature extraction. Mel-frequency cepstral coefficients (MFCC) [17; 18] are often used in automatic speech recognition [19] and are generated by correlating filter banks based on a Type 2 discrete cosine transform. They were introduced in the 1980s by Davis and Mermelstein, with the idea stemming from an attempt to preserve important phonetic information from speech that may have syntactic changes and variations in duration. MFCC extracts features that are more independent of such variations [20]. Mel-scale filter banks have a high degree of correlation, as triangular filters partially overlap each other. Applying the discrete cosine transform to each filter bank yields the Mel-frequency cepstral coefficients (MFCC). Usually, half of them are discarded, as higher-numbered filter banks indicate rapid changes in high

energy. Mel-frequency cepstral coefficients are less susceptible to additive interference. Based on psychoacoustic theory [11], MFCC features are widely used in speaker recognition systems. The i-vector speaker recognition system proposed by Dehak et al. [12] considers MFCCs as the primary source of speech features.

The selection of the primary 13 DCT coefficients as mel-frequency cepstral coefficient (MFCC) features by converting mel-scale frequencies to a logarithmic scale and applying discrete cosine transform (DCT) is established in these works [10,11,21-23]. Linear Predictive Coding (LPC) is a method of computing the current sample by linearly combining previous samples, where inverse filtering is applied to eliminate formants from speech signals, resulting in a residual signal. To calculate LPC coefficients, the VQ-LBG algorithm is used, in which vector quantization (VQ) of LPC coefficients is performed in the Linear Spectral Frequency (LSF) domain to reduce the data transmission rate.

Understanding the autoregressive (AR) version of speech is crucial for determining LPC. The sound signal is modeled as a p-order AR technique, which is expressed by equation (1), where each sample at the n-th moment is influenced by p previous samples and $u(n)$ Gaussian noise. The LPC coefficients are denoted by a and are determined using the Yule-Walker equations described in equations (2) - (4), with the solution given in expression (5). For simplicity, only the first 13 LPC coefficients are used. More detailed information about the proposed LPC features, VQ-LBG algorithm, and calculation steps can be found in works [24,25]. Linear predictive coding (LPC) is indeed a powerful method, but it may have limitations, particularly sensitivity to changes in sampling rate and reliance on assumptions about speech signal characteristics. Therefore, in our proposed work, we complement LPC with a number of additional features to enhance the robustness and effectiveness of the speaker recognition system.

$$s(n) = -\sum_{k=1}^{p} a_k \times (n-k) + u(n) \tag{1}$$

$$D(i) = a_0 + \sum_{n=1}^{N} \left(s(n) \times (n-i) \right) \tag{2}$$

The final form of the Yul-Walker expressions is represented by equations (3) and (4).

$$\sum_{k=1}^{p} a_k D\left(|i-k|\right) = -D(i) \tag{3}$$

$$\begin{bmatrix} \mathrm{D}(0) & \cdots & \mathrm{D}(\mathrm{p}-1) \\ \vdots & \ddots & \vdots \\ \mathrm{D}(\mathrm{p}-1) & \cdots & \mathrm{D}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ \vdots \\ a_k \end{bmatrix} = - \begin{bmatrix} \mathrm{D}(\mathrm{i}) \\ \vdots \\ \mathrm{D}(\mathrm{K}) \end{bmatrix} \tag{4}$$

Expression (5) provides the final solution for determining the LPC coefficients.

$$a = -D^{-1} r \tag{5}$$

In this research, PLP markers are selected due to their effectiveness in noise reduction, reverberation suppression, and echo cancellation, which increases overall performance efficiency. Extracting PLP features involves several steps, including pre-

emphasis, cube root compression, and removal of unnecessary speech data [5,26,27]. Thirteen PLP features are determined by calculating the average of all PLP features for each voice, which effectively reduces system complexity using PYTHON software [26,27]. To match the dimensions of PLP features with the dimensions of other features, the average value of each frame is calculated, resulting in 13×1 feature vectors in each frame. Figure 2 provides a visual representation of the feature extraction process for MFCC, LPC, and PLP features.

The spectral centroid (SC) determines the center of gravity of the spectrum, which is considered important, and represents a unique value that expresses the frequency domain characteristic of the speech signal. A high SC value indicates that the signal is strong [28], which is calculated using expression (6). The variable x_i represents the i -th part of the speech signal, while $x_i k$ denotes the amplitude value in the k-th frequency interval of the i-th part of the speech signal.

$$C(i) = C_1(i) / C_2(i) \tag{6}$$

$$C_1(i) = \sum_{k=0}^{N-1} k \left| x_i(k) \right|, C_2(i) = \sum_{k=0}^{N-1} \left| x_i(k) \right| \tag{7}$$

Spectral entropy (SE). To calculate entropy as a feature, the following steps are carried out:

• We calculate the power spectral density $s(f)$ for the signal $x(t)$.

• Based on the selected frequency range, we calculate the power of the spectral range. After calculating the power of the spectral range, we normalize the power in the selected range.

• We calculate the spectral entropy using expression (8) [29].

$$SE = \sum s(f) \times \ln \frac{1}{s(f)} \tag{8}$$

The root mean square (RMS) is a measure that indicates the amplitude of an audio signal. It is calculated by taking the square root of the sum of the mean squares of the sound sample amplitudes. The RMS formula is given in expression (9) [30]. Here, $x_1, x_2, \ldots x_n$ represent n observations, and x_{rms} denotes the RMS value for these n observations.

$$x_{rms} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} \tag{9}$$

Delta features are crucial for reflecting the rate of change in voice features, especially the power dynamics of speech signals relative to noise. Valuable dynamic information is extracted from speech signals using delta (Δ) and delta-delta (ΔΔ) features [4,9,23]. The calculation of the Δz feature is carried out by subtracting the previous feature f_{z-1} from the current feature f_z, as shown in equation (10).

$$\Delta z = f_z - f_{z-1} \tag{10}$$

Similarly, the feature $\Delta\Delta z$ is obtained by subtracting the current feature Δz from the previous feature $\Delta\Delta_{z-1}$, as shown in equation (11).

$$\Delta\Delta z = \Delta z - \Delta_{z-1} \tag{11}$$

Feature fusion method. The following steps were taken to select a model for feature fusion:
1. All 18 features were tested separately using speech data from 50 speaking English.
2. Out of the 18 features, the 2 best features with the highest SI accuracy and the lowest average EER value were selected.
3. When determining the best model, the average accuracy and average EER values across three classifiers were taken into account. LPC and PLP were identified as the first and second best features, respectively. Equation (12) shows the method for calculating the average accuracy based on the results of all three classifiers.

$$Average\,accuracy\,resault = \frac{\mathrm{LD} + \mathrm{KNN}}{2} \tag{12}$$

4. In the second stage, the best LPC and PLP features are separately fused with the remaining 17 features. Two more best models are selected from this process. The two best models identified at this stage were the combination model of MFCC and LPC, and the combination model of PLP and LPC.
5. In the third stage, the remaining 16 features are separately combined with the two best models selected from the second stage, resulting in a three-feature fusion configuration.
6. Feature fusion configurations involving 4 to 18 features are performed in the same manner, and at each stage, the two best-performing models are selected. A total of 315 models are tested for the dataset.

3.3 Methods of reducing the feature space (Approach 2)

To address the computational challenges associated with feature fusion in Approach 1, specifically to simplify the time-consuming and labor-intensive process of identifying the best models, a more efficient solution is needed. To overcome these issues, feature space reduction algorithms have been developed, aimed at simplifying calculations by reducing the size of training samples.

Principal Component Analysis (PCA). One of the widely used methods in this regard is Principal Component Analysis (PCA). PCA addresses the complexity problem arising in high-dimensional space by reducing the dimensions while preserving the components that represent the maximum variance of the data. In high-dimensional spaces, data points are usually sparse, making them difficult to generalize and study. PCA identifies the directions (principal components) in which the data vary the most and projects the data onto a lower-dimensional subspace covered by these components. This reduces the dimensionality of the space while retaining most of the important information in the data.

The steps performed in PCA are as follows [31]:

1. Loading input data: The feature fusion model, which serves as the input dataset, is fed into the PCA algorithm.

2. Subtracting the mean: The mean value of the data is subtracted from each feature in the initial dataset. This step ensures the centralization of data around the origin of the coordinate system.

3. Calculating the covariance matrix: The covariance matrix of the dataset is calculated. This matrix reflects the relationships and variations between different features.

4. Determining eigenvectors: The eigenvectors associated with the largest eigenvalues of the covariance matrix are identified. These eigenvectors represent the directions of maximum variance in the dataset.

5. Projecting the dataset: The initial dataset is projected onto the eigenvectors obtained in the previous step. This projection transforms the data into a lower-dimensional subspace spanned by the eigenvectors.

6. By following these steps, PCA effectively reduces the dimensionality of the dataset while preserving the most important information and capturing the most significant variations in the data [31,32].

Independent Component Analysis (ICA). ICA was first proposed in the 1980s by J. Herault, C. Jutten, and B. Ans, who developed an iterative algorithm for real-time processing [33]. Independent Component Analysis (ICA) is a method of dimensionality reduction in feature space, aimed at extracting independent components from a dataset. It is an extended version of PCA, allowing the identification of latent factors or sources that contribute to the formation of observed data. ICA is widely used in signal processing and data analysis due to its ability to separate mixed signals into their original sources.

The main steps performed in ICA can be summarized as follows:

1. Preprocessing: As in PCA, data is usually preprocessed - centered to provide a common reference point, and features are scaled to ensure equal contribution from each feature.

2. Whitening: Whitening is performed to transform the data into a new representation, where features are uncorrelated with each other and their variance is set to 1. This step helps eliminate any linear dependencies between features.

3. Determination of non-Gaussianity criterion: The ICA method aims to find components that are statistically as independent as possible. Various non-Gaussianity measures, such as kurtosis or negentropy, are used to quantify the degree of deviation from Gaussianity and to guide the process of separating independent components.

4. Optimization: The main task of ICA is to maximize the non-Gaussianity criterion for each component. This is achieved through an optimization process that involves finding weights or mixing matrices that maximize the non-Gaussianity criterion.

5. Iterative estimation: ICA often involves an iterative estimation process to better separate independent components.

6. Reconstruction: After independent components are obtained, they can be used to reconstruct the original dataset or further analyzed for specific purposes, such as feature extraction or signal separation [32-34].

These approaches to reducing the feature space help identify the best models in the integration process more quickly and effectively.

3.4 Classification

Linear Discriminant Analysis (LDA) Classifier. The Linear Discriminant Analysis (LDA) classifier is a statistical classification method based on finding the linear projection that best reveals the differences between classes. Its principle is to project data from a given feature space in such directions that the within-class scatter is minimized while maximizing the between-class separation. In practice, LDA relies on the mean vector and pooled covariance for each class, and the decision boundary takes the form of a hyperplane (i.e., the classification is "linear"). Consequently, LDA is computationally efficient, performs robustly on small and medium-sized datasets, and yields particularly good results when classes are approximately normally distributed with similar covariances. However, LDA's capabilities may be limited if the data are strongly nonlinearly separable or if class covariances differ significantly. Comprehensive information about LDA is provided in reference [35].

K-Nearest Neighbors (KNN) Classifier. The K-Nearest Neighbors (KNN) approach is a method of classifying unknown data points based on their similarity to neighboring data points. In this approach, the parameter K indicates the number of dataset elements that contribute to the classification process, and it is set to K=5 for this specific work. The KNN algorithm can be summarized in the following steps [36]:

1. We select the value of K, which represents the number of neighbors.
2. We calculate the Euclidean distance between the unknown given point and its K nearest neighbors.
3. We classify the K nearest neighbors according to the calculated Euclidean distances.
4. We count the number of data points in each class.
5. We assign the new data point to the class with the highest count.

4 Evaluation

4.1 Database preparation

In the proposed work, we used a dataset consisting of 50 English speakers. The dataset comprises 2,511 speech samples from 50 speakers in total. For classification studies, the dataset was divided into training (70%) and testing (30%) subsets using stratified sampling. We used 1,758 samples for the training set and 753 samples for the testing set. The average duration of each sample was approximately 59 seconds, with the total duration of speech being approximately 41 hours. All audio files are stored in WAV format and recorded in mono. The recordings have sampling frequencies of 16 kHz and 22.05 kHz; however, to ensure accuracy in feature extraction, all signals were resampled to 16 kHz.

4.2 Evaluating the Effectiveness of Speaker Identification (SI)

Speaker recognition accuracy evaluates the system’s ability to correctly identify different speakers based on their voices. This is crucial as it provides security, personali-

zation, efficiency, and convenience for applications that utilize voice communication or authentication. The accuracy of Speaker Identification (SI), calculated using PYTHON's classification learning library, encompasses all proposed models with both classifiers. Simplified results tables only reflect the models with the highest accuracy. Accuracy is determined by the ratio of correctly identified speakers voices to the total number of voices in the dataset (equation 13).

$$Accuracy = \frac{\text{Number of voices correctly identified}}{\text{Total number of audio files}} \tag{13}$$

4.3 Evaluating the Effectiveness of Speaker verification (SV)

For the SV problem, EER values are calculated using PYTHON for each model. The EER is determined by the intersection point of the receiver operating characteristic (ROC) curve and the diagonal axis from (0, 1) to (1, 0), which corresponds to the false positive rate [37, 38].

Figure 3 shows the ROC curve and EER point for individual MFCC and LPC features, as well as feature fusion configurations—including fusion of all features and Δ-fusion—constructed by isolating and pairing their feature vectors in various combinations. The figure also reports the results of feature-space reduction using PCA and ICA, with all experiments evaluated using the KNN classifier. We observe that the ROC curve closer to the point (1, 0) yields the lowest EER, indicating better performance compared to other applied methods. Thus, based on this curve, we can conclude that among all the methods used, the feature space reduction method using the ICA technique produces the best results.

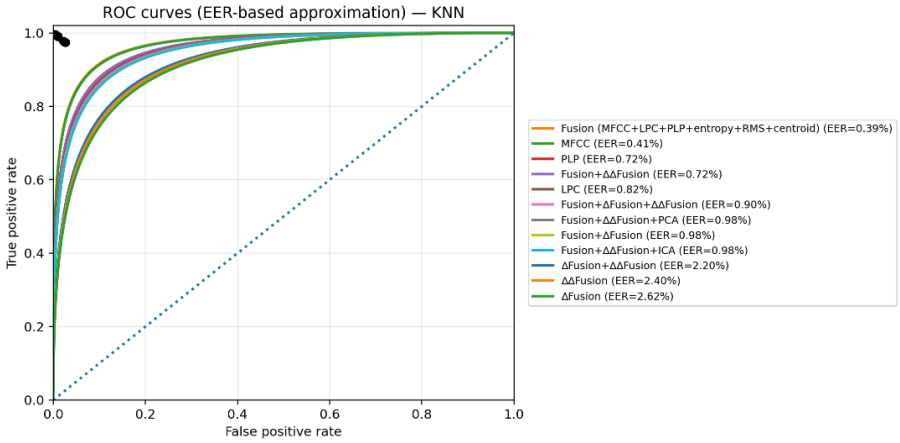

Fig. 3. ROC curve for English speech data from 50 speakers using the KNN classifier

Figure 4 depicts the ROC curve and EER point for various feature vectors using the LDA classifier. These include individual MFCC, PLP, and LPC features, as well as feature fusion configurations—such as fusion of all features, Δ-fusion, and their various combinations—together with feature-space reduction methods based on PCA and ICA. We observe that the ROC curve closer to the point (1, 0) yields the lowest EER, indicating better performance compared to other applied methods. Thus, based on this

curve, we can conclude that among all the methods used, the feature space reduction method using ICA technique produces the best results.

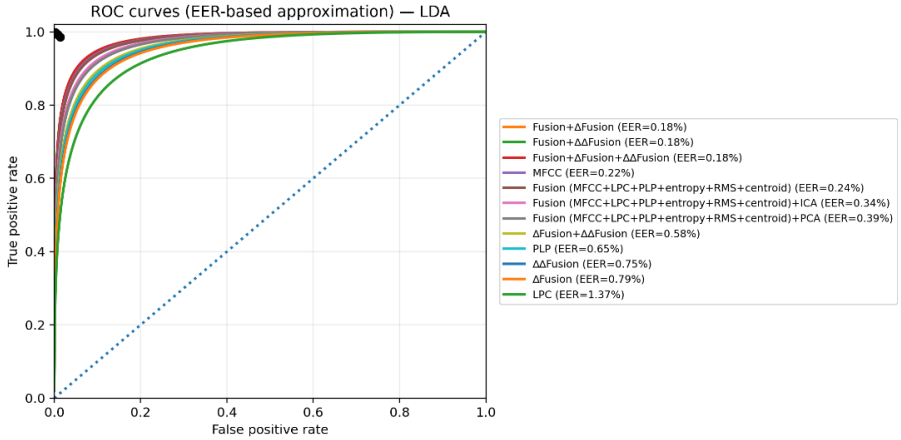

Fig. 4. ROC curve for English speech data from 50 speakers using the LDA classifier

5 Results Discussion

5.1 Optimal results were achieved through feature fusion (Approach 1)

Tables 1-2 reflect the results of speaker recognition using individual features such as MFCC and LPC, as well as their associations and performance indicators of the most effective feature fusion model. This model achieves the highest accuracy and the lowest equivalent error level (EER) in experiments conducted on a dataset of 50 speakers. The presented data demonstrates the advantages of feature fusion, which is consistently preferable to relying on a single feature or a limited number of fusion configurations. Tables 1-2 include only the classification method that yielded the best results.

In the experiment, the results achieved through the feature fusion approach showed that the best Speaker Identification (SI) accuracy and Equal Error Rate (EER) values obtained using the Linear Discriminant (LD) classifier were 97.4% accuracy with an EER value of 0.18% for 50 speakers. Similarly, the best SI accuracy and EER values obtained using the K-Nearest Neighbors (KNN) classifier were 95.82% accuracy with an EER value of 0.39% for 50 speakers. Below, we present the classification results using the feature fusion approach for the dataset of 50 speakers (Tables 1-2).

Table 1. Comparison of results using the feature fusion approach, with classification implemented in the LDA classifier.

Feature used (Model)	Classifier Model	Training Time (s)	Testing Time (s)	SI Accuracy (%)	SV EER (%)
MFCC	LDA	0.026	0.0004	97.01	0.22
LPC	LDA	0.023	0.0003	94.22	1.37

PLP	LDA	0.065	0.006	95.8	0.65
Fusion (MFCC+LPC+PLP+ +entropy+RMS+centroid)	LDA	0.028	0.0004	97.01	0.24
ΔFusion	LDA	0.029	0.0003	93.03	0.79
ΔΔFusion	LDA	0.029	0.0004	93.43	0.75
Fusion+ΔFusion	LDA	0.047	0.0005	97.21	0.18
Fusion+ΔΔFusion	LDA	0.046	0.0005	97.21	0.18
ΔFusion+ΔΔFusion	LDA	0.045	0.0005	93.83	0.58
Fusion+ΔFusion+ΔΔFusion	LDA	0.082	0.0005	97.4	0.18

Table 2. Comparison of results using the feature fusion approach, with classification implemented in the KNN classifier.

Feature used (Model)	Classifier Model	Training Time (s)	Testing Time (s)	SI Accuracy (%)	SV EER (%)
MFCC	KNN	0.002	1.355	95.82	0.41
LPC	KNN	0.002	0.015	95.42	0.82
PLP	KNN	0.004	0.08	94.9	0.72
Fusion (MFCC+LPC+PLP+ +entropy+RMS+centroid)	KNN	0.002	0.013	96.81	0.39
ΔFusion	KNN	0.0021	0.016	85.28	2.62
ΔΔFusion	KNN	0.0021	0.017	84.48	2.40
Fusion+ΔFusion	KNN	0.0034	0.048	94.62	0.98
Fusion+ΔΔFusion	KNN	0.0034	0.017	94.42	0.72
ΔFusion+ΔΔFusion	KNN	0.0032	0.015	87.27	2.20
Fusion+ΔFusion+ΔΔFusion	KNN	0.0037	0.019	93.43	0.90

Figure 5 illustrates how speaker recognition accuracy varies with different feature fusion sizes (i.e., different numbers of fused features) across multiple classification methods. Analysis of the graphs shows that increasing the number of features generally improves accuracy. However, it should be emphasized that randomly adding features does not guarantee better results. Furthermore, different feature fusion configurations exhibit varying performance across classifiers. These methods aim to identify the most informative features and reduce redundancy, maximizing the discriminative ability of the feature set. By reducing the feature space, we can improve the feature fusion process, thereby enhancing the robustness and effectiveness of speaker recognition systems under various noise conditions and across datasets of different sizes.

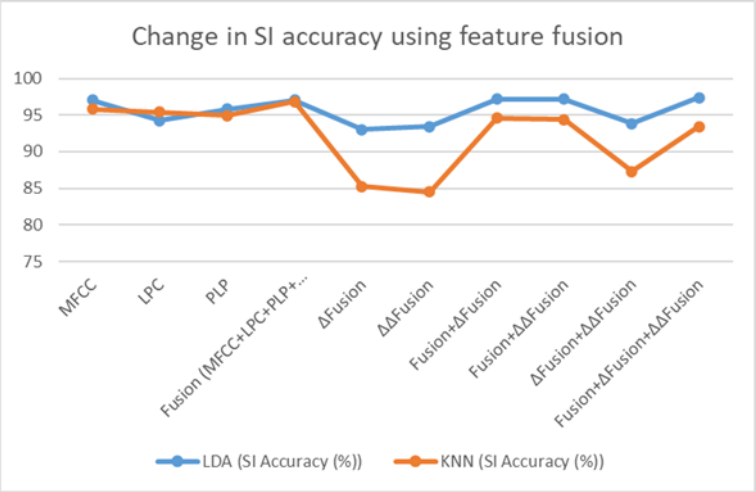

Fig. 6. Graph showing the change in SI accuracy when using feature fusion.

5.2 Optimal results using the feature space reduction approach (Approach 2)

Tables 3-6 present the best results achieved when applying the PCA and ICA methods for feature space reduction, including the computation times for both the training and testing phases.

Table 3. Comparison of results using the PCA method for feature space reduction. Classification was performed using the LDA classifier.

Feature used (Model)	Classifier Model	Training Time (s)	Testing Time (s)	SI Accuracy (%)	SV EER (%)
Fusion (MFCC+LPC+PLP+ +entropy+RMS+centroid)+PCA	LDA	0.02	0.0003	98.60	0.39
ΔFusion+PCA	LDA	0.02	0.0003	96.22	0.41
ΔΔFusion+PCA	LDA	0.03	0.0003	96.02	0.44
Fusion+ΔFusion+PCA	LDA	0.03	0.0003	97.81	0.37
Fusion+ΔΔFusion+PCA	LDA	0.03	0.0003	97.81	0.46
ΔFusion+ΔΔFusion+PCA	LDA	0.03	0.0003	96.42	0.37
Fusion+ΔFusion+ΔΔFusion+PCA	LDA	0.02	0.0003	97.61	0.22

Table 4. Comparison table of results using the PCA method for feature space reduction, with classification performed using the KNN classifier.

Feature used (Model)	Classifier Model	Training Time (s)	Testing Time (s)	SI Accuracy (%)	SV EER (%)
Fusion (MFCC+LPC+PLP+ +entropy+RMS+centroid)+PCA	KNN	0.001	0.01	96.62	1.72

ΔFusion+PCA	KNN	0.001	0.014	94.23	1.18
ΔΔFusion+PCA	KNN	0.001	0.014	93.43	1.11
Fusion+ΔFusion+PCA	KNN	0.001	0.012	95.82	1.28
Fusion+ΔΔFusion+PCA	KNN	0.001	0.013	97.01	0.98
ΔFusion+ΔΔFusion+PCA	KNN	0.001	0.013	95.62	0.97
Fusion+ΔFusion+ΔΔFusion+PCA	KNN	0.001	0.012	96.42	0.79

Table 5. Comparison of results using the ICA method for feature space reduction. Classification was performed using the LDA classifier.

Feature used (Model)	Classifier Model	Training Time (s)	Testing Time (s)	SI Accuracy (%)	SV EER (%)
Fusion (MFCC+LPC+PLP+ +entropy+RMS+centroid)+ICA	LDA	0.03	0.0003	98.40	0.34
ΔFusion+ICA	LDA	0.03	0.0003	96.42	0.40
ΔΔFusion+ICA	LDA	0.03	0.0003	96.02	0.41
Fusion+ΔFusion+ICA	LDA	0.03	0.0003	97.21	0.22
Fusion+ΔΔFusion+ICA	LDA	0.03	0.0003	97.61	0.47
ΔFusion+ΔΔFusion+ICA	LDA	0.03	0.0003	96.81	0.55
Fusion+ΔFusion+ΔΔFusion+ICA	LDA	0.03	0.0003	97.61	0.22

Table 6. Comparison of results using the ICA method for feature space reduction. Classification was performed using the LDA classifier.

Feature used (Model)	Classifier Model	Training Time (s)	Testing Time (s)	SI Accuracy (%)	SV EER (%)
Fusion (MFCC+LPC+PLP+ +entropy+RMS+centroid)+ICA	KNN	0.001	0.012	96.62	1.72
ΔFusion+ICA	KNN	0.001	0.015	93.24	1.19
ΔΔFusion+ICA	KNN	0.001	0.012	94.23	1.01
Fusion+ΔFusion+ICA	KNN	0.001	0.015	95.82	1.28
Fusion+ΔΔFusion+ICA	KNN	0.001	0.013	97.01	0.98
ΔFusion+ΔΔFusion+ICA	KNN	0.001	0.013	95.02	0.89
Fusion+ΔFusion+ΔΔFusion+ICA	KNN	0.001	0.011	96.42	0.79

Tables 3-6 present the calculated accuracy of speaker identification (SI) and the equal error rate (EER) for PCA and ICA methods in reducing the highest-dimensional feature space. According to these results, the PCA method yielded the best outcomes. When using the LD (linear discriminant) classification method, the best SI accuracy was 98.6% with an EER of 0.22%. Additionally, when employing the KNN (K-nearest neighbor) classifier, the best SI accuracy was recorded at 97.01% with an EER of 0.79%.

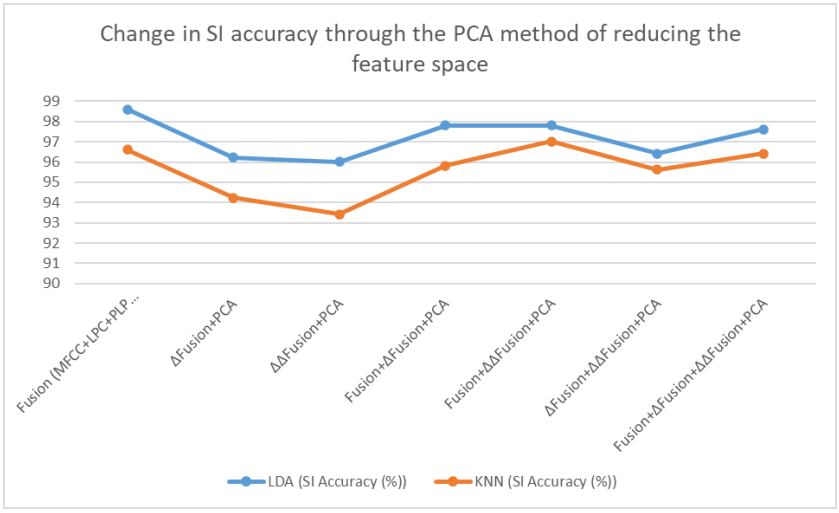

Fig. 6. Graph showing the change in SI accuracy through the PCA method of reducing the feature space.

The best results obtained using the ICA method were achieved when applying the LD (linear discriminant) classification method, with the highest SI accuracy of 98.4% and an equal error rate (EER) of 0.22%. Additionally, when using the KNN (K-nearest neighbor) classifier, the best SI accuracy was 96.62% with an EER of 0.79%.

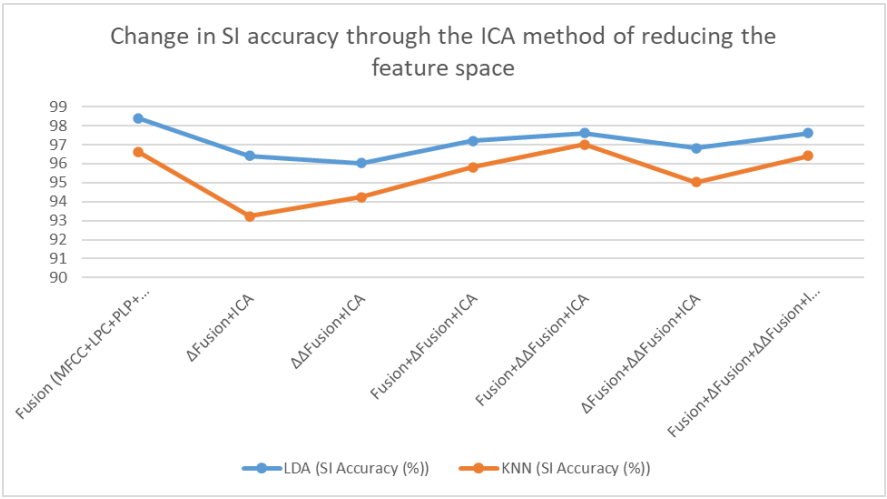

Fig. 7. Graph showing the change in SI accuracy through the ICA method of reducing the feature space

5.3 System Configuration

The research used PYTHON software and NVIDIA GeForce RTX 3090 Ti (24GB) GPUs for calculating training and testing times.

6 Conclusions

In this research, we attempted to solve the problem of speaker recognition through less informative learning through various approaches. We proposed a novel feature extraction method based on feature space reduction techniques. Our research encompassed an in-depth study of speaker recognition, feature fusion, and the application of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) methods for feature space reduction. Our newly proposed feature space reduction method, applied to reduced feature vectors, partially improved speaker recognition performance across various classification methods. Notably, we increased speaker recognition accuracy to 97.4% through feature fusion. Additionally, when applying PCA and ICA feature space reduction methods along with LD and KNN classifiers, we achieved an Equal Error Rate (EER) of 0.22% in speaker verification. In conclusion, the LDA classifier demonstrated high effectiveness when used in combination with feature space reduction methods across various noise levels and dataset sizes, indicating its potential for practical applications.

References

1. Chauhan, N., Isshiki, T., Li, D.: Enhancing Speaker Recognition Models with Noise-Resilient Feature Optimization Strategies. Acoustics 2024, 6(2), 439-469.
2. Mirzaev, N., Urinboev, J., Gafforov, N., Nugmanova, M.: Analysis of Speech Signal Processing Methods in Speech Recognition Systems. IEEE 6th International Conference on Problems of Cybernetics and Informatics (PCI), Baku, Azerbaijan, 25–28 august 2025, pp. 1-8,
3. Zamalloa, M., Bordel, G., Rodriguez, L., Penagarikano, M.: Feature selection based on genetic algorithms for speaker recognition. In Proceedings of the 2006 IEEE Odyssey—The Speaker and Language Recognition Workshop, San Juan, PR, USA, 28–30 June 2006, pp. 1–8.
4. Furui, S.: Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Trans. Acoust. Speech Signal Process. 1981, 29, 342–350.
5. Hermansky, H., Morgan, N., Bayya, A., Kohn, P.: Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech 1991), Genovo, Italy, 24–26 September 1991, pp. 1367–1370.
6. Adami, A.G., Mihaescu, R., Reynolds, D.A., Godfrey, J.J.: Modeling prosodic dynamics for speaker recognition. In Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 6–10 April 2003, pp. IV–788.
7. Sönmez, K., Shriberg, E., Heck, L., Weintraub, M.: Modeling dynamic prosodic variation for speaker verification. In Proceedings of the 5th International Conference on Spoken

Language Processing (ICSLP 1998), Sydney, Australia, 30 November–4 December 1998, pp. 3189–9192.
8. Kumar, K., Kim, C., Stern, R.M.: Delta-spectral cepstral coefficients for robust speech recognition. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011, pp. 4784–4787.
9. Carey, M.J., Parris, E.S., Lloyd-Thomas, H., Bennett, S.: Robust prosodic features for speaker identification. In Proceedings of the Fourth International Conference on Spoken Language Processing, ICSLP 9‘6, Philadelphia, PA, USA, 3–6 October 1996, pp. 1800–1803.
10. Chauhan, N., Isshiki, T., Li, D.: Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database. In Proceedings of the 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), Singapore, 23–25 February 2019, pp. 130–133.
11. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Lang. Process. 2011, 19, 788–798.
12. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 1980, 28, 357–366.
13. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons: New York, NY, USA, (2001).
14. Rosca, J., Kopfmehl, A.: Cepstrum-like ICA representations for text independent speaker recognition. In Proceedings of the ICA’2003, Nara, Japan, 1–4 April 2003, pp. 999–1004.
15. Ding, P., Kang, X., Zhang, L.: Personal recognition using ICA. In Proceedings of the ICONIP (2001).
16. Cichocki, A., Amari, S.I.: Adaptive Blind Signal and Image Processing. John Wiley: Chichester, UK, (2002).
17. Mirzaev, N., Urinboyev, J.: Methods of Feature Extraction in Speaker Recognition and Their Effectiveness. International Scientific and Technical Conference Current State and Prospects for the Development of Digital Technologies and Artificial Intelligence, Bukhara, September 27-28, 2024, pp.102-108.
18. Mirzaev, N., Urinboyev, J, Nugmanova, M.: Main Methods of Speaker Identification Through Speech. Journal of Transport 2024, 1(4), 130-136.
19. Mirzaev, N., Urinboyev, J.: Methods of linear prediction and spectral centroid in the selection of features characterizing sound in the speaker recognition problem. Digital transformation and artificial intelligence: problems, innovations and trends 1st international scientific - practical conference, TSUE, Tashkent, September 11, 2024, pp. 29-30.
20. Mirzayev, N., Urinboyev, J.: Mel-frequency cepstral analysis method for extracting voice characteristics in speaker recognition. Computational models and technologies (CMT2024) Third International conference dedicated to the 90th birth anniversary of Professor M.I. Isroilov, Tashkent, April 24, 2024, pp. 199-201.
21. Chakroborty, S., Roy, A., Saha, G.: Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. In Proceedings of the 2006 IEEE International Conference on Industrial Technology, Mumbai, India, 15–17 December 2006, pp. 387–390.
22. Chauhan, N., Isshiki, T., Li, D.: Speaker Recognition using fusion of features with Feedforward Artificial Neural Network and Support Vector Machine. In Proceedings of the 2020 international conference on intelligent engineering and management (ICIEM), London, UK, 17–19 June 2020, pp. 170–176.

23. Ahmad, K.S., Thosar, A.S., Nirmal, J.H., Pande, V.S.: A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In Proceedings of the 2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR), Kolkata, India, 4–7 January 2015, pp. 1–6.
24. Wang, L., Chen, Z., Yin, F.: A novel hierarchical decomposition vector quantization method for high-order LPC parameters. IEEE/ACM Trans. Audio Speech Lang. Process. 2014, 23, 212–221.
25. Slifka, J., Anderson, T.R.: Speaker modification with LPC pole analysis. In Proceedings of the 1995 International Conference on Acoustics, Speech, and Signal Processing, Detroit, MI, USA, 9–12 May 1995, pp. 644–647.
26. Mirzaev, O, Radjabov, S, Mirzaev, N, Meliev, F, Tillavoldiev, A.: A family of recognition algorithms based on the construction of k-dimensional threshold rules. Procedia Computer Science, 2024 (237), 610-617.
27. Chauhan, N., Chandra, M.: Speaker recognition and verification using artificial neural network. In Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 22–24 March 2017, pp. 1147–1149.
28. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 1990 (87), 1738–1752.
29. Root-mean-square Value.: A Dictionary of Physics, 6th ed., Oxford University Press: Oxford, UK, (2009).
30. Ross, A.: Fusion, feature-level. In Encyclopedia of Biometrics, Li, S.Z., Jain, A., Eds., Springer: Boston, MA, USA, 2009, 597–602.
31. You, S.D., Hung, M.J.: Comparative study of dimensionality reduction techniques for spectral–temporal data. Information 2021, 12(1), 1-12.
32. Herault, J., Jutten, C., Ans, B.: Detection de grandeurs primitives dans un message composite par une architecture de calcul neuromimetique en apprentissage non supervise. In Proceedings of the GRETSI, Nice, France, 20–24 May 1985, pp. 536.
33. Opanasenko, V., Fazilov, Sh., Mirzaev, N., Kakharov, Sh.: A Model of Recognition Algorithms Based on Threshold Functions and Object Proximity Assessment. Cybernetics and Systems Analysis, 61(1), (2025), 148-160.
34. Tharwat, A.: Independent component analysis: An introduction. Appl. Comput. Inform. 2018, 17, 222–249.
35. Yao, Z., Ruzzo, W.L.: A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinform. 2006, 7, S11. 1-11
36. Subasi, A.: Machine learning techniques. In Practical Machine Learning for Data Analysis Using Python, Elsevier: Amsterdam, The Netherlands, 2020, pp. 91–202.
37. Tharwat, A.: Classification assessment methods: A detailed tutorial. Appl. Comput. Inform. 2020, 17, 168–192.
38. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 2015, 10 (e0118432), 1-21.