IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 6, June 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306287|

Speaker Identification using Machine Learning

M Sai Akash, S Praveena

I M. Tech, DECE, Department of ECE, MGIT, Hyderabad, India

Associate Professor, Department of ECE, MGIT, Hyderabad, India

ABSTRACT: Speaker identification determines the identity of a speaker based on their unique voice characteristics. This study explores a multifaceted approach that leverages four key audio features: pitch, Mel Frequency Cepstral Coefficients (MFCC), short-time energy, and zero-crossing rate.Pitch provides information on the fundamental frequency of a speaker's voice, which is unique to individuals. MFCC captures the timbral texture of speech signals, offering a compact representation of their spectral properties. Short-time energy measures variations in the audio signal's energy over short segments, reflecting the speaker's vocal intensity. The zero-crossing rate quantifies the frequency of zero-amplitude crossings in the waveform, characterizing speech patterns. By combining these features, we aim to enhance the accuracy of speaker identification systems. The proposed approach extracts these features from speech samples and employs machine learning techniques to classify and identify speakers. Experiments conducted on a diverse dataset demonstrate that this method significantly improves identification performance compared to conventional approaches. Our findings suggest that integrating multiple audio features enhances the robustness and reliability of speaker identification systems, making this approach invaluable for applications such as forensic analysis, secure access systems, and personalized voice interfaces.

KEYWORDS: Speech Recognition, KNN, Forensics, Machine Learning, Sound waves, Smart surroundings, Monitoring

I. INTRODUCTION

Forensics, surveillance, and remote computer access are among the many applications of general speech recognition. Other uses include securing credit card transactions, protecting confidential information, and verifying customer identity in telephone banking systems. Voice recognition combines both what individuals say and how they say it, providing a robust two-factor authentication mechanism. As a reliable and secure identification method, voice recognition is increasingly essential in today's smart world.

Voice recognition plays a pivotal role in various domains, such as voice-activated devices, home automation, and banking. In science and engineering, speaker identification falls under the broader category of "pattern recognition." The aim of pattern recognition is to classify objects of interest into predefined classes or categories. In the context of speaker identification, "objects of interest" (or "patterns") refer to acoustic vectors extracted from input speech using methods described earlier. These patterns correspond to different speakers and are processed to assign them to specific classes.

When the classes in a set of patterns are predefined and known, the process is referred to as supervised pattern recognition. A classification algorithm is trained using these patterns, which form the training set. The remaining patterns constitute the test set and are used to evaluate the classification algorithm's performance. The accuracy of the approach can be measured if the correct class labels for all patterns in the test set are available.

The speech recognition process begins with voice feature extraction, typically using the Mel Frequency Cepstrum Coefficient (MFCC) technique. MFCC is designed to capture signal characteristics that reflect human speech. The method requires converting analog signals to digital form. Building on this, the study proposes a novel technique for managing outlier data using a double-distance measurement combined with a K-Nearest Neighbor (KNN) classifier. The double-distance method calculates the distance between each data point and the KNN cluster center, which is used during the recognition phase to enhance accuracy.

The experiment evaluates the method's performance by focusing on the following objectives:\n\n- **Dataset Collection:** Gather speech data from reliable databases.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306287|

Preprocessing: Split the dataset into training and testing subsets while extracting features for input to KNN classifiers. **Training and Testing:** Use the extracted features to train the KNN classifier. The classifier is then tested with unseen speech signals from each speaker to assess its generalization performance.

The remainder of this study is organized as follows:\n\n- **Section II:** Discusses previous research related to the topic. - **Section III:** Provides a detailed explanation of the proposed methodology and algorithm. **Section IV:** Presents the results of the proposed method, along with a comparative analysis. **Section V:** Concludes the study and suggests directions for future research.

II. RELATED WORK

K-Nearest Neighbors (KNN) is a classification method that categorizes objects based on their proximity to learning data. It classifies new input items by referencing the closest samples from the training dataset. This concept aligns with clustering algorithms, where newly entered data is categorized by its similarity to neighboring data points. In this study, the KNN technique was employed to classify speech sample data collected from various speakers. The classification process involves determining the parameter KKK, calculating the distance between the object and the training data, and sorting data based on proximity to the nearest neighbors. The closest neighbors are then used to classify the trial data [1].

The Hidden Markov Model (HMM) is another widely used text-dependent approach. Initially, a microphone records the user\u2019s voice as a .wav file. The analog voice signal is then converted into a digital stream using an Analog-to-Digital (A2D) converter. During the training phase, all spoken words are transformed into the cepstrum domain. The user\u2019s feature parameters are compared against a reference voice sample to calculate a likelihood ratio, which determines the probability of the data fitting one model over another. Based on this likelihood ratio, the system decides whether to accept the user\u2019s voice or reject it as an impostor. The HMM also supports automated training and operation. It isolates unwanted noise inputs and models the spectral representation as a collection of Gaussian functions. The spectral envelope is then fitted to these Gaussian functions during system setup, resulting in cepstral coefficients [2].

Speaker identification involves determining which speaker from a predefined set the system has been trained to recognize. This process, often referred to as closed-set speaker identification, evaluates one voice from a known group of speakers. In this scenario, the test speaker must be part of the enrolled database. In contrast, open-set identification addresses cases where the test speaker may not be included in the training database, potentially belonging to the broader, unregistered public. Open-set identification is more complex and requires additional effort compared to the simpler and more accurate closed-set identification approach [3].

Preprocessing in speaker identification encompasses multiple steps, including noise isolation and file conversion. To mitigate distortion in the incoming audio stream, Additive White Gaussian Noise (AWGN) is filtered out using Adaptive Recursive Least Squares filtering for effective noise cancellation. The audio data is then represented as a spectrogram to facilitate feature extraction. A spectrogram is a two-dimensional visualization that displays the spectral composition of a signal over time and frequency. In this representation, the axes denote time and frequency, while brightness or color intensity indicates the amplitude of specific frequency components at each time interval [6].

The rapidly expanding internet has opened up new avenues for development across various fields, including home automation. Over the past few years, the home automation industry has witnessed significant growth, capturing the interest of individuals worldwide.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306287|

III. METHODOLOGY

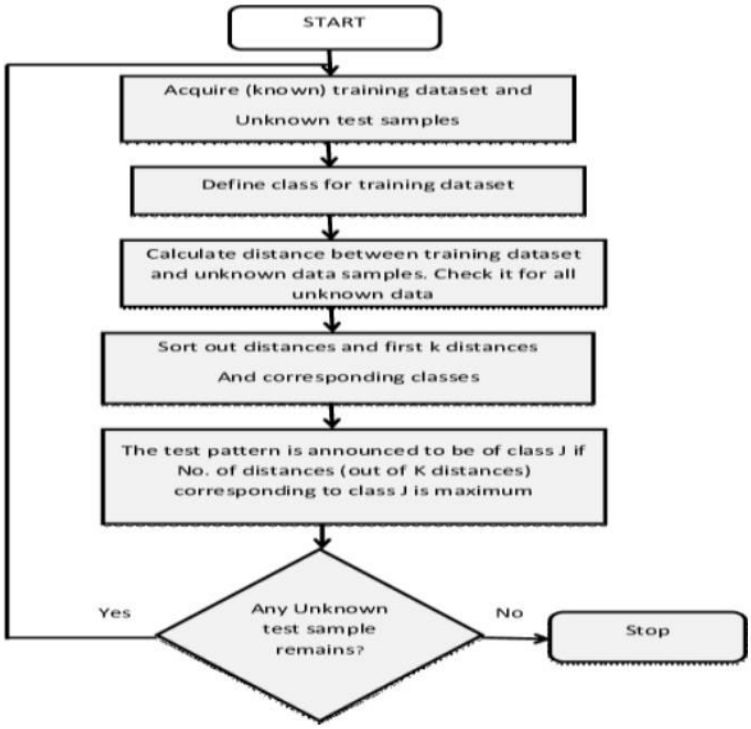

Figure 1

The flowchart illustrates the step-by-step process of implementing the K-Nearest Neighbors (KNN) classification algorithm. The process begins with acquiring the dataset, which includes a set of labeled training samples and a collection of unknown test samples that need to be classified. The training dataset is pre-defined with specific classes, which serve as references during the classification phase. For each test sample, the algorithm calculates the distance between the test sample and all training samples. This distance, which could be computed using metrics such as Euclidean or Manhattan, helps determine the proximity of the test sample to the training data.

Once the distances are calculated, they are sorted in ascending order, and the first kkk smallest distances are selected along with their corresponding classes. The test sample is then assigned to a class based on majority voting, where the most frequent class among the kkk-nearest neighbors determines the classification. If multiple test samples are present, the process is repeated for each one. This iterative process continues until all test samples have been classified. Finally, the process terminates when no unknown test samples remain. This approach highlights the simplicity and effectiveness of the KNN algorithm, which relies on proximity and majority voting to achieve accurate classification results.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306287|

IV. EXPERIMENTAL RESULTS

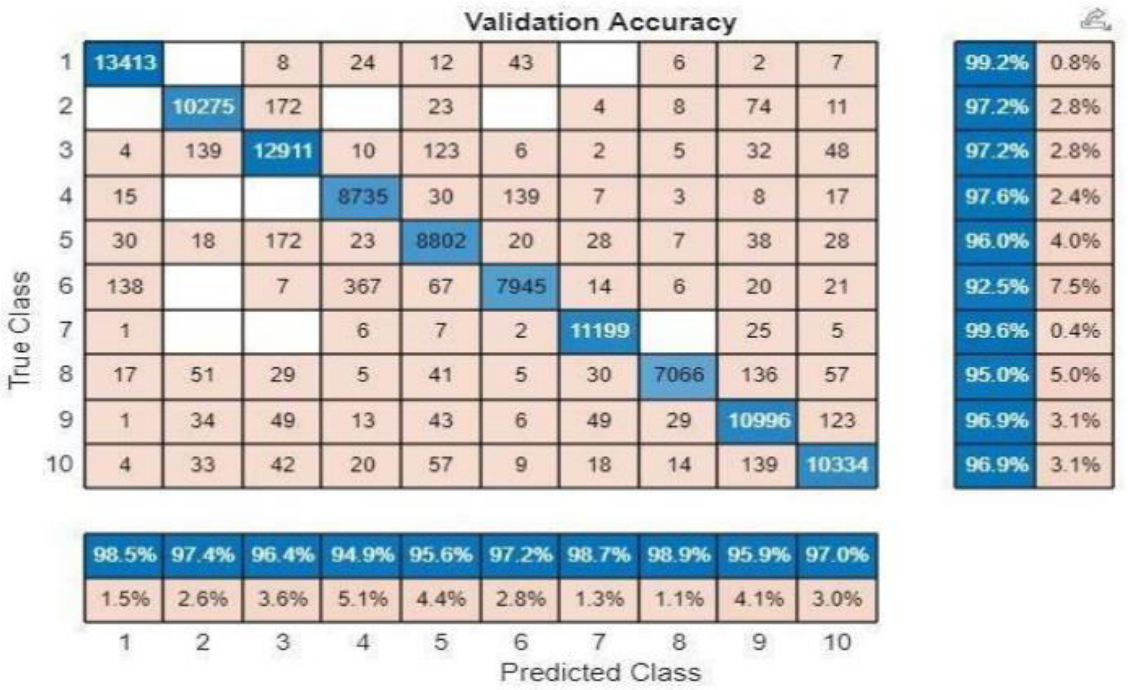

FIGURE 2: VALIDATION ACCURACY

A **confusion matrix** is a tool used to evaluate the performance of a classification model. In this figure, it is applied to a speaker identification dataset with 10 distinct classes, each representing a different speaker. The rows represent the true classes, while the columns represent the predicted classes. Diagonal values indicate correct classifications, with high counts showing the model's strong performance. Off-diagonal values indicate misclassifications, where the model confused one speaker with another. The rightmost column shows the class-wise accuracy, reflecting how accurately the model identifies samples for each speaker. Most accuracies exceed 97%, although Class 7 shows a relatively lower accuracy of 92.5%. The bottom row displays precision, which measures the percentage of predictions for each class that are correct. Overall, the model demonstrates high accuracy and precision, although occasional misclassifications suggest that certain classes (e.g., Class 7 and Class 9) may require further refinement in data or features. This analysis highlights the model's strengths and areas for improvement in speaker identification tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306287|

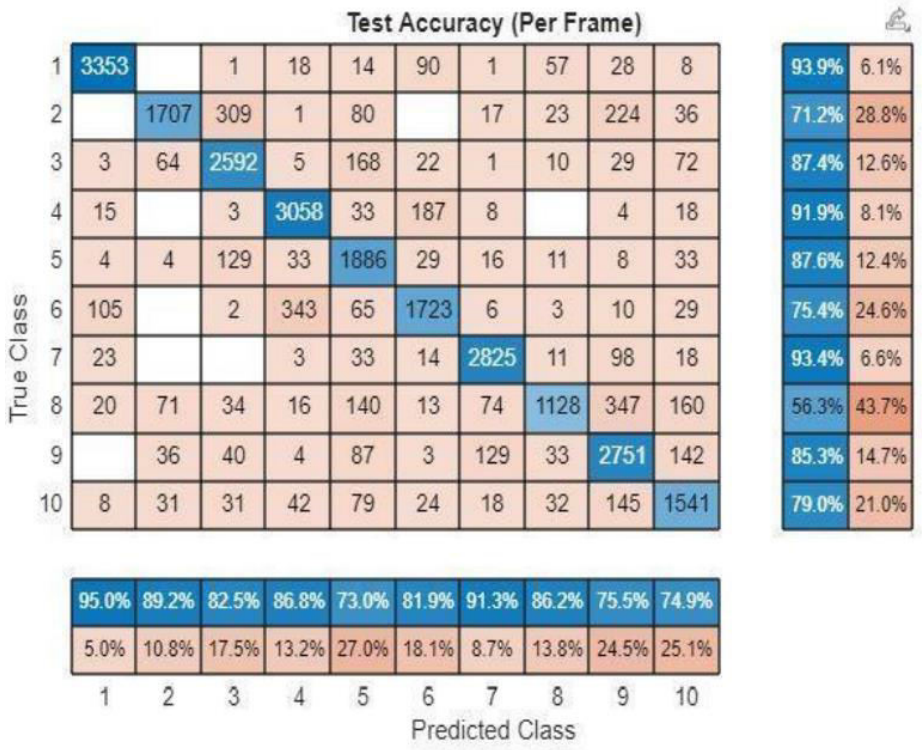

Figure 3

This confusion matrix represents the performance of a classification model, likely for speaker identification, focusing on test accuracy per frame across 10 classes. The rows correspond to the actual (true) classes, while the columns represent the predicted classes. Diagonal values show correct classifications, while off-diagonal values indicate misclassifications. The rightmost column shows the accuracy for each class, representing how often the model correctly identifies samples from that class. For example, Class 1 has high accuracy (93.9%), while Class 9 has the lowest (56.3%), indicating challenges in distinguishing this class. The bottom row shows the precision for each predicted class, reflecting the proportion of correct predictions. While some classes, such as Class 1 and Class 7, show strong performance with high accuracy and precision, others, like Class 6 and Class 9, exhibit significant misclassifications, suggesting potential overlaps in features or data issues. Overall, this matrix highlights the model's strengths in several classes while identifying areas for improvement in others.

V. CONCLUSION

The results obtained using the K-Nearest Neighbors (KNN) algorithm for speaker identification highlight the effectiveness of features such as Pitch, MFCC (Mel-Frequency Cepstral Coefficients), Short-Time Energy, and Zero Crossing Rate. These features capture unique characteristics of speech, enabling accurate classification for most speaker classes. The confusion matrices show strong performance for certain classes, with high accuracy and precision, such as Class 1 (93.9% accuracy) and Class 7 (93.4% accuracy). However, some classes, like Class 6 (75.4% accuracy) and Class 9 (56.3% accuracy), exhibit lower performance, indicating challenges in distinguishing speakers with similar vocal traits or insufficient feature representation. Misclassifications between neighboring classes suggest potential overlaps in feature space or the need for parameter optimization, such as fine-tuning the number of neighbors in KNN. While the selected features and algorithm perform well overall, incorporating additional features or exploring advanced classification techniques like SVM or deep learning could further enhance performance, particularly for underperforming classes.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com | A Monthly Peer Reviewed & Referred Journal | e-ISSN: 2319-8753; p-ISSN: 2347-6710|

|Volume 13, Issue 6, June 2024|

|DOI: 10.15680/IJIRSET.2024.1306287|

REFERENCES

[1] Umar, R., Riadi, I., Hanif, A., &Helmiyah, S. (2019). Identification of speaker recognition for audio forensic using k-nearest neighbor. International Journal of Scientific and Technology Research, 8(11), 3846-3850.

[2] Shah, Hairol& Rashid, M.Z. &Abdollah, MohdFairus&Kamarudin, Muhammad Nizam& Lin, C.K. &Kamis, Z. (2014). Biometric Voice Recognition in Security System.Indian Journal of Science and Technology. 7. 104-112. 10.17485/ijst/2014/v7i1.9.

[3] Pawar, R.V., Jalnekar, R.M. &Chitode, J.S. Review of various stages in speaker recognition system, performance measures and recognition toolkits. Analog IntegrCirc Sig Process 94, 247– 257 (2018). https://doi.org/10.1007/s10470-017-1069-1

[4] Z. Liu, Z. Wu, T. Li, J. Li and C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition," in IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 32443252, July 2018, doi: 10.1109/TII.2018.2799928.

[5] T. -H. Tsai, P. -C. Hao and C. -L. Wang, "Self-Defined TextDependent Wake-Up-Words Speaker Recognition System," in IEEE Access, vol. 9, pp. 138668-138676, 2021, doi: 10.1109/ACCESS.2021.3117602.

[6] Dhakal P, Damacharla P, Javaid AY, Devabhaktuni V. a Near RealTime Automatic Speaker Recognition Architecture for Voice-Based User Interface. Machine Learning and Knowledge Extraction. 2019; 1(1):504-520. https://doi.org/10.3390/make1010031

[7] l-Moneim, S.A., Nassar, M.A., Dessouky, M.I. et al. Text-independent speaker recognition using LSTM-RNN and speech enhancement. Multimed Tools Appl 79, 24013–24028 (2020). https://doi.org/10.1007/s11042-019-08293-7

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details