Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Full Length Research Article

A CNN-BASED APPROACH FOR SPEAKER IDENTIFICATION USING MFCC FEATURES IN NOISY AND REAL-WORLD ENVIRONMENTS

[1]Philip O. Odion, *[1]Pasha R. Adeiza, [2]Tijjani Abdullahi

[1]Department of Computer Science, Nigerian Defence Academy (NDA), Kaduna, Nigeria
[2]Department of Computer Science, Baze University, Abuja, Nigeria

*Corresponding Author Email Address: peebaby055@gmail.com

ABSTRACT
Speech extraction and recognition have become essential components in modern intelligent systems, especially in applications requiring accurate speaker identification under real-world conditions. Thus, this study represents a deep learning-based approach to speech recognition for speaker identification using Convolutional Neural Networks (CNNs) trained on Mel-Frequency Cepstral Coefficients (MFCCs). This research integrates both locally collected speech data and an external benchmark dataset (Mikhailava et al., 2022) to evaluate model performance under varying acoustic conditions. A total of 630 audio samples were collected from 21 participants across diverse environments, including both clean and noisy recordings. The proposed model achieved training accuracy exceeding 95%, validation accuracy of approximately 73%, and test accuracy of 75.82% on the local dataset. Evaluation on the benchmark dataset produced a test accuracy of 100%, indicating strong model learning under controlled conditions. The results showed that while high accuracy can be achieved with clean data, performance declines in real-world noisy environments due to variability in speech patterns, recording quality, and background interference. This study demonstrates that CNN-based models can effectively support speaker identification tasks, while highlighting the need for improved generalization, larger datasets, and enhanced noise-handling techniques for practical deployment.

Keywords: Speech extraction, Speaker recognition, Deep learning, Convolutional Neural Network (CNN), Mel-Frequency Cepstral Coefficients (MFCC).

INTRODUCTION

The human voice is a primary medium of communication that enables the rapid and efficient exchange of ideas through spoken language. The integration of speech (Sounds) with computing systems has attracted significant research attention, especially with advances in Artificial Intelligence (AI) and machine learning techniques. Speech recognition systems are designed to convert spoken language into machine-readable formats that enable computers to interpret, process, understand, and respond to human speech (Ali et al., 2023). Over time, these systems have become increasingly relevant in applications such as virtual assistants, biometric authentication, healthcare services, security systems, and automated customer support (Swietlicka et al., 2022). The development of speech extraction and recognition technology dates back several decades, beginning with early systems such as "Audrey," developed at Bell Laboratories in the 1950s, which was capable of recognizing spoken digits with limited accuracy (Laing, 2024). Subsequent advancements led to other systems such as IBM's Shoebox, which expanded vocabulary recognition capabilities (Maneesh & Sharda, 2021). In the 1980s and 1990s, statistical models, particularly Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs), were the dominant approaches for speech recognition tasks due to their ability to model temporal and probabilistic characteristics of speech signals (Kheddar et al., 2023). While all these methods provided a solid foundation, their performance was often limited in complex and noisy environments. The emergence of deep learning has significantly enhanced the performance of speech extraction recognition systems. Deep neural networks, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated strong capabilities in learning complex patterns from large-scale speech data (Rahul et al., 2023). CNNs, in particular, have proven effective at extracting local spectral features from speech representations such as spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs), which are widely used in speech processing tasks (Prabhu & Seethalakshmi, 2025). These advancements have enabled improved accuracy and robustness in speech recognition systems, especially when dealing with high-dimensional and unstructured data.

Despite these improvements, speech extraction and speaker recognition remain challenging tasks. Speech extraction and recognition involve isolating the target speaker's voice from a mixture of multiple speakers and background noise, while speaker recognition focuses on identifying or verifying the speaker's identity based on vocal characteristics (Islam et al., 2024). In real-world environments, speech signals are often affected by noise, reverberation, overlapping voices, and variations in recording devices, which complicate accurate recognition (Khazaleh & Khrais, 2024). Additionally, differences in accents, pitch, tone, and speaking styles introduce further variability, reducing the performance of recognition systems (Balachandran et al., 2025). Feature extraction plays a critical role in speech recognition systems. Several techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Coding (LPC), are commonly used to represent speech signals in a compact and informative form (Rahul et al., 2023). MFCCs in particular capture perceptually relevant spectral characteristics of speech, making them effective for distinguishing between speakers. However, these features are often very sensitive to noise and environmental distortions, which can degrade system performance in practical applications (Sujatha et al., 2025). Recent studies have explored the use of deep learning models to improve robustness in speech extraction recognition tasks. For instance, CNN-based models trained on spectrogram representations have demonstrated

superior performance compared to traditional feature-based methods (Deka & Kumari, 2025). Similarly, hybrid architectures combining CNNs with recurrent layers have been used to capture both spatial and temporal features of speech signals, achieving high accuracy in speaker identification tasks (Hema & Marquez, 2023). Nevertheless, many existing studies rely on clean, well-curated datasets, which, in turn, do not fully represent real-world conditions in which noise and variability are prevalent (Hamsa et al., 2023). Another important consideration in speech extraction recognition research is the availability and diversity of datasets. Large-scale datasets such as VoxCeleb and RAVDESS have contributed significantly to model development and evaluation (Taha et al., 2024). However, locally collected datasets often present additional challenges, including limited size, variability in recording quality, and privacy concerns related to voice data collection (Hussein et al., 2025). These challenges can affect model generalization and highlight the need for robust data preprocessing, augmentation, and ethical data handling practices. This study focuses on the development of a deep learning-based speech and speaker extraction and recognition system using CNNs trained on MFCC features. Unlike many existing works that rely primarily on clean datasets, this research incorporates both locally collected speech data and an external benchmark dataset (Mikhailava et al., 2022) to evaluate performance under varying acoustic and environmental conditions. The locally collected dataset includes recordings from diverse speakers in different educational environments, capturing real-world variability in speech signals. This approach provides a more realistic assessment of the model's performance and highlights the challenges associated with deploying speech extraction recognition systems in practical settings. The main contribution of this research is to demonstrate the effectiveness of CNN-based models in handling both clean and noisy speech data for speaker identification. The study also provides insights into the impact of data quality, environmental overlapping noise, and speaker variability on model performance. Furthermore, it emphasizes the importance of data augmentation, feature extraction techniques, and model optimization in improving recognition accuracy. While deep learning has significantly advanced the field of speech extraction recognition, challenges related to noise, data diversity, and generalization remain critical. This research addresses these issues by proposing and evaluating a CNN-based model for speech extraction and speaker recognition using both controlled and real-world datasets, intending to improve robustness and applicability in practical environments.

Deep Learning-Based Speaker Recognition Performance

Several recent studies have demonstrated the effectiveness of deep learning techniques in improving speaker extraction recognition accuracy, particularly with Convolutional Neural Networks (CNNs). For instance, Reddy et al. (2023) evaluated a CNN-based speaker identification system using spectrogram representations and reported an accuracy of 98.79%, significantly outperforming traditional Mel-Frequency Cepstral Coefficient (MFCC)-based approaches, which achieved lower identification rates. Also, Singh (2023) compared Artificial Neural Networks (ANNs), Recurrent Neural Networks (RNNs), and CNNs, concluding that CNNs achieved superior performance due to their ability to capture spatial hierarchies in speech features. Balachandran et al. (2025) further explored deep CNN architectures such as ResNet and VGG for speaker extraction recognition using the VoxCeleb dataset. These findings indicated that ResNet-34 achieved the highest accuracy across both speaker identification and verification tasks. In another study, Deka & Kumari (2025) incorporated self-attention mechanisms into CNN architectures and reported improved classification accuracy compared to traditional CNN models, achieving up to 90.80% top-1 accuracy. Such findings confirm that CNN-based architectures are highly effective for speaker recognition tasks, particularly when trained on structured and high-quality datasets.

Impact of Feature Extraction Techniques

Feature extraction is a critical factor influencing the performance of speech extraction recognition systems. Rahul et al. (2023) emphasized the importance of discriminative features such as MFCC and Linear Prediction Cepstral Coefficients (LPCC), demonstrating that several feature fusion techniques improve recognition performance by combining complementary speech characteristics. Similarly, Arpita et al. (2025) proposed a method that uses specialized characteristic spectrograms combined with MFCC and LPCC features, achieving a recognition rate of 96.7% and outperforming individual feature-based approaches. (2025) investigated the use of MFCC features with a feedforward neural network and reported accuracies between 83.50% and 92.90%, depending on training configurations. However, their findings also indicated that MFCC features are highly sensitive to environmental noise, which can further reduce system reliability in real-world conditions. Singh (2024) demonstrated that amplitude mel-spectrograms achieve better performance for accent classification than conventional MFCC features, with accuracies between 0.964 and 0.987. This suggests that while MFCC remains widely used, alternative spectral representations may offer improved robustness in certain tasks.

Speech Recognition in Noisy and Real-World Environments

A major limitation in many speech extraction recognition systems is their reduced performance in noisy environments. Khazaleh & Khrais (2024) investigated the impact of environmental conditions on speaker extraction recognition systems. They found that while CNN-based models perform well under controlled conditions, their accuracy declines in the presence of background noise and recording distortions. Despite this, the models retained some ability to distinguish speakers, indicating partial robustness. Hamsa et al. (2023) proposed a deep learning framework for speaker identification under emotional and noisy speech conditions. Their model achieved average identification rates of 85.2% to 87.0% across multiple datasets, demonstrating improved performance in challenging acoustic conditions. However, the study highlighted the difficulty of extracting stable features from speech signals affected by noise and emotional variability. Similarly, Balachandran et al. (2025) developed a convolutional recurrent neural network (CRNN) for speaker identification across emotional states and languages, achieving accuracy rates above 97% in controlled environments. However, performance declined in cross-language settings, indicating limited generalization. These studies highlight that while deep learning models perform well in controlled settings, their effectiveness in real-world environments remains constrained by noise, variability, and data quality.

Accent Variability and Speaker Diversity

Accent variability presents a significant challenge in speech recognition systems. Mikhailava et al. (2022) investigated accent

detection using CNNs trained on sparse, crowd-sourced datasets and reported an accuracy of approximately 75%. While their model successfully captured accent-related features, the study was limited to a small number of European accent groups, restricting generalization. Lesnichaia et al. (2022) further explored accent classification and demonstrated that deep learning models can achieve high accuracy when trained on carefully designed spectral features. However, both studies relied on relatively controlled datasets, limiting their applicability to diverse real-world populations. Taha et al. (2024) proposed a multifunctional architecture that integrates speaker identification, gender classification, and emotion detection, achieving accuracy above 96% across tasks. Their findings suggest that incorporating multiple speech characteristics can enhance model performance, although such systems require large and diverse datasets.

Dataset Limitations and Generalization Challenges

The availability and quality of datasets play a crucial role in the performance of speech recognition systems. Many studies rely on benchmark datasets such as VoxCeleb and RAVDESS, which provide clean and well-annotated audio samples (Taha et al., 2024). While these datasets support high model accuracy, they do not fully capture the variability present in real-world environments. Hutiri & Aaron (2022) highlighted issues related to bias and fairness in automatic speaker recognition systems, particularly when datasets lack diversity in terms of demographics and recording conditions. Similarly, Hamsa et al. (2023) noted that obtaining large, representative datasets remains a major challenge, especially for multilingual and emotionally expressive speech. Dwijayanti et al. (2022) also observed overfitting in CNN models when trained on limited datasets, even with regularization techniques such as dropout. This suggests that model performance is strongly influenced by dataset size and diversity. These findings emphasize the need for locally collected, diverse datasets that reflect real-world conditions while also addressing ethical concerns such as privacy and consent in voice data collection.

Deep Embeddings, Attention, and Transformer Models in Speaker Recognition

While Convolutional Neural Networks (CNNs) have demonstrated strong performance in speaker recognition tasks, recent advancements have shifted toward deep embedding approaches and attention-based architectures that offer superior robustness and generalization.

One of the most influential developments in this area is the x-vector architecture, which employs a time-delay neural network (TDNN) with statistics pooling to project variable-length utterances into fixed-dimensional speaker embeddings. Building upon the x-vector paradigm, the ECAPA-TDNN (Emphasized Channel Attention, Propagation and Aggregation in TDNN) introduced several enhancements including Res2Net modules with squeeze-and-excitation blocks, multi-layer feature aggregation, and channel-dependent frame attention. Recent work by Toth et al. (2025) demonstrated that fine-tuning ECAPA-TDNN models on augmented datasets significantly improves robustness to noise, reducing equal error rates by half in challenging classroom environments compared to baseline models. Transformer architectures have emerged as powerful alternatives for speaker recognition tasks. Unlike CNNs, which rely on local receptive fields, transformers employ self-attention mechanisms to capture global dependencies across entire speech sequences. A systematic review by Hema & Marquez (2023) confirmed that transformer models have revolutionized speech recognition technology, especially in challenging acoustic environments, by efficiently modeling the dynamic and context-rich nature of speech. However, as noted by Tu et al. (2025), applying vanilla transformers directly to speaker verification has not yielded the same success observed in natural language processing, primarily due to their inferior locality modeling capability. They argue that speaker characteristics are often reflected in local speech dynamics that standard self-attention fails to capture effectively. To address this limitation, they proposed ConFusionformer, a multi-resolution attention fusion that upsamples low-resolution attention score maps and fuses them with standard attention maps to explicitly enhance locality modeling. Their results on VoxCeleb, CNCeleb, SRE21, and SRE24 demonstrate that ConFusionformer substantially outperforms conventional Conformers and transformers for speaker verification. Another significant advancement is the integration of channel-temporal attention mechanisms with Conformer architectures. Researchers have proposed lightweight end-to-end TD-SV models based on convolution-augmented transformer (Conformer) architectures enhanced with channel-temporal attention modules that specifically target speaker-discriminative patterns. Experimental results on challenging far-field and noisy datasets demonstrate that such approaches achieve a competitive equal error rate of 2.04% while enabling real-time deployment on edge devices via INT8 quantization. Self-supervised speech models pre-trained on large-scale unlabeled data have also shown remarkable potential for speaker identification. Models such as Wav2Vec 2.0, XLS-R, and Whisper learn rich speech representations that can be fine-tuned for downstream speaker tasks. Research by Stuhlmann (2025) revealed that Wav2Vec 2.0 and XLS-R capture speaker-specific features effectively in their early transformer layers, with fine-tuning improving stability and performance, while Whisper shows better performance in deeper layers. Further research by Vaessen et al. (2025) demonstrated that continued pre-training of Wav2Vec 2.0 XLS-R models on archival datasets achieves state-of-the-art performance for specific languages, confirming the value of domain-adaptive pre-training for speaker representation learning. The robustness of these deep embedding models under degraded conditions has been systematically evaluated. Studies examining the impact of environmental noise and GSM coding on pre-trained speaker embedding models revealed that ECAPA-TDNN achieves a minimum equal error rate (EER) of 1% in clean conditions, but performance drops notably under white noise at 10 dB SNR, with EER increasing to 6.4%. These findings highlight that while deep embeddings offer significant advantages, environmental degradation remains a critical challenge.

Attention-based feature fusion is another promising direction, in which complementary information from different feature domains is dynamically integrated. Karthikeyan et al. (2025) proposed an attention-based multidimensional fused-feature convolutional neural network (AMDF-CNN) that combines features from 1D raw waveforms, 2D Mel-spectrograms, and 3D dynamic time-space features through an attention-based fusion mechanism. Their approach achieved an identification rate of 97.59% and demonstrated reliability exceeding 85% under various noise conditions. Despite these advances, several limitations are identified in the application of deep embedding and transformer-based models to real-world speaker recognition. First, these models typically require large-scale training data (e.g., VoxCeleb

contains over 1 million utterances from 7,000+ speakers), which is not feasible for locally collected datasets or resource-constrained research settings. Second, the computational demands of transformer-based architectures, particularly the quadratic complexity of the self-attention mechanism, pose challenges for deployment in resource-constrained environments. Third, as noted in recent forensic voice comparison studies, most evaluations have been conducted on clean controlled datasets, leaving questions about robustness in noisy, real-world conditions largely unanswered. This gap is significant because environmental noise and recording variability remain primary obstacles to practical system deployment.

Synthesis of Research Gap

The preceding literature review reveals a clear trajectory of advancement in speaker recognition, from traditional statistical models through CNN-based architectures to contemporary deep embedding and transformer-based approaches. However, several critical gaps remain unaddressed. First, while deep embedding architectures (x-vectors, ECAPA-TDNN) and transformer-based models (ConFusionformer, Wav2Vec 2.0) have achieved state-of-the-art performance on benchmark datasets like VoxCeleb, their evaluation has been predominantly confined to clean, controlled acoustic conditions. Even studies that specifically address noise robustness, such as Toth et al. (2025), demonstrate that performance degrades significantly, with equal error rates increasing from 1% to 6.4% under white noise at 10 dB SNR, confirming that environmental robustness remains an unsolved challenge. Second, existing studies on robust speaker recognition have primarily relied on artificially augmented clean datasets or benchmark noise databases rather than locally collected, ecologically valid speech data that captures the full complexity of environmental variability, device diversity, and demographic representation. The use of synthetic noise augmentation, while methodologically convenient, may not fully replicate the unpredictable acoustic challenges encountered in real-world settings (Karthikeyan et al., 2025). Third, there is a notable scarcity of research that systematically compares model performance across datasets of varying quality, from clean, curated benchmarks to locally collected noisy samples, using consistent architectures and evaluation metrics. Such comparative analysis, as advocated by Jakubec et al. (2024) in their review of deep speaker embeddings, is essential for establishing realistic expectations for system deployment. Fourth, many existing studies emphasize accent classification or general speech recognition rather than speaker-specific identification, which requires capturing subtler and more complex vocal attributes. As noted by Stuhlmann (2025), even advanced transformer models such as Wav2Vec 2.0 and Whisper exhibit varying effectiveness across layers when fine-tuned for speaker identification, indicating that optimal architecture design for speaker-specific tasks remains an open research question. Fifth, there is limited engagement with the practical challenges of speech data collection in underrepresented populations, including participant privacy concerns, consent procedures, and data security issues that directly affect the feasibility of developing diverse, representative datasets for robust model training.

This study addresses these interconnected gaps by: (1) developing a CNN-based speaker identification system trained on a locally collected dataset that incorporates real-world environmental variability, device diversity, and demographic representation; (2) conducting a comparative evaluation against a clean benchmark dataset (Mikhailava et al., 2022) to quantify the performance degradation associated with noisy conditions; (3) providing transparent reporting of dataset characteristics, preprocessing decisions, and model limitations; and (4) offering a realistic assessment of CNN+MFCC performance under practical conditions, thereby establishing a baseline against which more sophisticated architectures (embeddings, transformers) can be meaningfully compared on locally collected, noisy speech data.

MATERIALS AND METHODS

This section presents the methodological framework adopted in developing the speech extraction and speaker identification system. The approach follows a structured deep learning pipeline, focusing on processing speech signals and applying Convolutional Neural Networks (CNNs) for accurate speaker classification. Emphasis is placed on handling real-world speech data, which includes variations in recording environments, background noise, and device quality, all of which influence system performance. The section is structured to provide a clear progression from data acquisition to model evaluation. It begins with an overview of the research design and system workflow, followed by detailed descriptions of dataset development, preprocessing techniques, feature extraction methods, and data augmentation strategies. The chapter then outlines the proposed system architecture, the model development process, the evaluation metrics, and the ethical considerations that guide the study.

Research Design and System Overview

The research adopts a data-driven experimental design focused on developing a deep learning model for speaker identification. The system is structured as a sequential processing pipeline, where raw audio data is transformed into meaningful representations and subsequently used for model training and evaluation. This design enables systematic handling of speech signals while ensuring consistency in data preparation and model input (Jahangir et al., 2021; Kheddar et al., 2023).

The overall workflow of the system is organized into five key stages as depicted in Figure I:

Figure. I High Methodological Figure of the Study

Dataset Development and Collection:

Speech data is collected from a diverse group of participants across different environments using multiple recording devices. This stage ensures variability in speech patterns, recording conditions, and demographic representation, which is essential for building a model capable of handling real-world scenarios.

Preprocessing:

The collected audio data undergoes preprocessing to standardize formats, remove unwanted noise, and ensure consistency in

sampling rates. This step improves the quality of the input data and enhances the reliability of subsequent feature extraction processes (Alasadi et al., 2019).

Feature Extraction:
Relevant acoustic features are derived from the speech signals, primarily using MFCCs and spectral representations. These features capture important frequency and temporal characteristics necessary for distinguishing between speakers (Islam et al., 2024).

Model Development:
A Convolutional Neural Network (CNN) is designed to learn patterns from the extracted features. The model processes the input data through multiple layers to identify distinguishing characteristics associated with individual speakers.

Evaluation:
The model's performance is assessed using standard evaluation metrics, including accuracy, loss, and confusion matrices. This stage determines the model's ability to generalize to unseen data and identifies areas requiring improvement.

Dataset Development and Collection
The development of a reliable speech recognition and speaker identification system depends largely on the quality, diversity, and representativeness of the dataset used for training and evaluation. In this study, a locally sourced dataset was developed to reflect real-world speech conditions, incorporating variations in speaker characteristics, recording environments, and device quality. This approach supports the development of models capable of handling practical deployment scenarios in which speech signals are often affected by noise and other environmental factors (Khazaleh & Khrais, 2024). The dataset development process involved participant recruitment, controlled recording procedures, the use of multiple recording devices and formats, and the organization of the collected data into structured categories suitable for machine learning tasks.

Participant Selection
The dataset was compiled from a diverse group of participants to ensure adequate representation of different speech patterns and vocal characteristics. A total of 50 individuals, aged between 18 and 60 years, were initially recruited from various institutional backgrounds, including the College of Agriculture and Animal Science, Ahmadu Bello University, Zaria, and the Air Force Institute of Technology, Kaduna. This diversity was necessary to capture variations in speech influenced by age, educational background, and social exposure, which are critical for robust speaker identification (Hema & Marquez, 2023). Participation in the study was voluntary, and ethical considerations were strictly observed throughout the data collection process. Consent forms were provided to all participants to ensure transparency about the study's purpose and the intended use of their voice data. Despite these measures, several challenges were encountered, including reluctance to participate due to privacy concerns and fears regarding potential misuse of recorded speech data. These concerns align with findings in existing studies, which highlight privacy and trust as significant barriers in speech data collection (Hutiri & Aaron, 2022).

Recording Procedure
The recording process was designed to ensure consistency while still capturing natural variations in speech. Each participant was required to read a standardized text passage during recording sessions. This approach ensured uniformity in spoken content across all participants, allowing the model to focus on speaker-specific characteristics rather than differences in linguistic content. Recordings were conducted in a variety of environments, including classrooms, lecture halls, homes, offices, libraries, and open spaces such as aircraft hangars. These settings introduced varying levels of background noise and acoustic conditions, intentionally incorporated to reflect real-world scenarios. Environmental factors such as ambient noise, human activity, and external disturbances were present in many recordings, which contribute to dataset variability. Pre-recording checks were conducted to test the recording setups and ensure that audio signals were captured clearly. However, the complete elimination of environmental noise was not always possible, thereby further enhancing the dataset's realism and relevance for robust model training (Hamsa et al., 2023).

Recording Devices and Formats
Multiple recording devices were used during data collection to introduce variability in audio quality and to evaluate the system's adaptability to different input conditions. A smartphone with an inbuilt microphone was initially used due to its accessibility and convenience. However, recordings from the smartphone exhibited noticeable distortion and reduced clarity, particularly in noisy environments. To improve audio quality, a Dictaphone was later introduced for recording sessions. The device demonstrated higher sensitivity and captured clearer audio signals, although it also recorded subtle background sounds such as ticking clocks, distant conversations, and environmental noise. This sensitivity contributed to both improved signal capture and increased noise presence. Two primary audio formats were used: MP3 and WMA. The MP3 format provided compressed audio with reduced file size while maintaining acceptable quality, whereas the WMA format allowed better preservation of audio fidelity at lower bitrates. The use of multiple formats ensured compatibility and enabled comparative evaluation of model performance across different audio encoding conditions.

Dataset Composition
The locally collected dataset used in this study comprises 1,500 audio recordings from 50 participants, with approximately 30 samples per speaker and individual recording durations ranging from 10 to 50 seconds. Of these, approximately 600 samples were classified as clean (with minimal background noise), while 900 contained varying degrees of environmental interference. Noise types present in the dataset include ambient sounds (e.g., fans, traffic), human activity (conversations, footsteps), device-induced artifacts (compression noise, microphone hiss), and reverberation from large recording spaces such as aircraft hangars and lecture halls. Based on qualitative assessment of waveforms and spectrograms, the estimated signal-to-noise ratio (SNR) ranged from approximately 5 dB (high noise) to 30 dB (clean). The dataset was partitioned into training (70%, 1,050 samples), validation (15%, 225 samples), and test (15%, 225 samples) splits, with stratification by speaker to ensure representation across all partitions.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Data Preprocessing

Data preprocessing is a critical stage in developing a speech recognition system, as it ensures that raw audio data is transformed into a consistent, usable format for feature extraction and model training. Given the variability in recording environments, devices, and audio quality, preprocessing was necessary to enhance signal clarity and maintain uniformity across the dataset (Prabhu & Seethalakshmi, 2025).

a) **Audio Loading:**
All recorded audio files were loaded into the processing environment using the Librosa library in Python. This enabled efficient handling of different audio formats and facilitated subsequent signal processing operations. The audio signals were converted into time-series representations suitable for analysis.

b) **Sampling Rate Standardization:**
To ensure consistency across all recordings, the audio files were standardized to a uniform sampling rate, typically between 16 kHz and 22 kHz. Standardizing the sampling rate is essential in speech processing, as variations in sampling frequency can lead to inconsistencies in feature representation and negatively impact model performance (Hussein et al., 2025).

c) **Noise Reduction:**
Given the presence of environmental noise in many recordings, noise reduction techniques were applied to improve signal quality. Basic filtering methods were used to suppress background noise while preserving the integrity of the speech signal. This step enhanced the reliability of extracted features, particularly for recordings obtained in uncontrolled environments (Hamsa et al., 2023).

d) **Data Cleaning:**
The dataset was further refined through a cleaning process that involved removing corrupted, inaudible, or excessively noisy recordings. This ensured that only samples with sufficient clarity were used for training and evaluation. Data cleaning also involved trimming silent segments and ensuring that audio clips met the required duration and quality thresholds.

Feature Extraction

Feature extraction involves transforming raw audio signals into compact and informative representations that can be effectively processed by machine learning models. In this study, both Mel-Frequency Cepstral Coefficients (MFCCs) and spectral features were utilized to capture relevant characteristics of speech signals for speaker identification. These features provide insight into the frequency and temporal properties of speech, which are essential for distinguishing between speakers (Latiff et al., 2025).

MFCC Extraction

MFCCs were employed as the primary feature representation due to their effectiveness in modeling human auditory perception and capturing speech-specific characteristics, as shown in Figure II.

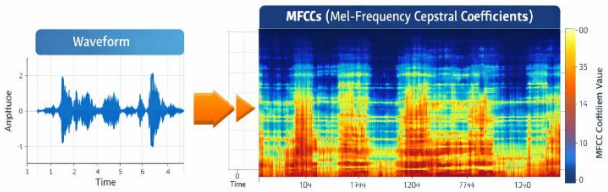

Figure. II Example of MFCC Feature Representation of a Speech Signal

i. **Frame Segmentation:**
Each audio signal was divided into short overlapping frames to capture the quasi-stationary nature of speech. This segmentation allowed the analysis of speech signals over small time intervals, during which frequency characteristics remain relatively stable.

ii. **Coefficient Generation:**
For each frame, a set of MFCCs was computed, typically extracting 40 coefficients. These coefficients represent the spectral envelope of the speech signal on a perceptual scale, enabling the model to capture important features such as pitch and timbre. MFCCs are widely used in speaker identification tasks due to their ability to represent distinctive vocal characteristics (Chen et al., 2025).

Feature Normalization and Reshaping

To ensure compatibility with the CNN model and improve training efficiency, the extracted features underwent normalization and reshaping processes.

Padding/Truncation:

Since audio recordings varied in length, the resulting MFCC feature matrices were standardized to a fixed size. Shorter sequences were padded with zeros, while longer sequences were truncated to a predefined length (e.g., 100 frames). This ensured uniform input dimensions across all samples.

Channel Expansion:

The two-dimensional MFCC feature matrices were expanded to include an additional channel dimension, converting them into a format suitable for convolutional layers. This transformation allows the CNN to process the input in a manner similar to image data, enabling effective learning of spatial and temporal patterns in the speech signal.

Data Augmentation Techniques

Data augmentation was applied to increase the dataset's size and variability, particularly to address the limitations associated with the relatively small number of locally collected samples. This process introduces controlled modifications to existing audio data, enabling the model to learn more generalized patterns and improve its performance on unseen data (Kheddar et al., 2023).

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Noise Addition: Artificial noise was introduced into selected audio samples to simulate real-world environmental conditions. This includes background sounds such as ambient noise and low-level disturbances. The inclusion of noisy samples during training enhances the model's ability to distinguish speech signals under varying acoustic conditions, thereby improving robustness (Hamsa et al., 2023).

Pitch Variation: Pitch-shifting techniques were applied to alter the frequency characteristics of the speech signal without changing its duration. This process helps the model become less sensitive to variations in vocal pitch across different speakers, thereby improving its ability to generalize across diverse voice profiles (Sujatha et al., 2025).

Environmental Simulation: Environmental simulation techniques were used to replicate various recording conditions, including reverberation and echo effects. These transformations mimic real-life acoustic environments and expose the model to a wider range of speech variations during training.

Proposed System Architecture
The proposed system architecture is designed to process speech signals and perform speaker identification using a Convolutional Neural Network (CNN). The architecture integrates feature extraction outputs with deep learning techniques to learn distinctive patterns associated with individual speakers.

System Framework
The system follows a structured pipeline in which preprocessed audio data is transformed into feature representations and then passed into the CNN model for classification, as depicted in Figure III. The input to the model consists of normalized MFCC feature matrices, which are treated as two-dimensional inputs with an additional channel dimension. The CNN architecture comprises multiple convolutional layers that extract hierarchical features from the input data. These layers are responsible for identifying local patterns in the spectral representation of speech, such as frequency variations and temporal dependencies. Pooling layers are incorporated to reduce dimensionality and computational complexity while retaining essential features. Following the convolutional stages, fully connected layers are used to perform classification based on the learned feature representations. The final output layer produces probability scores corresponding to each speaker class, enabling the model to assign an input audio sample to a specific speaker.

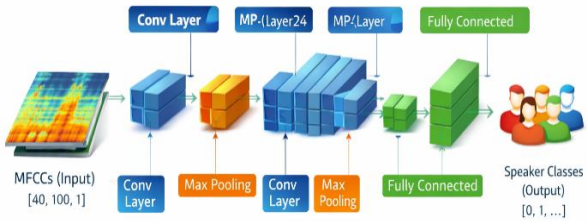

Figure III: Architecture of CNN Model for Speaker Identification

System Enhancements
The CNN model accepts MFCC feature matrices of dimension 40 × 100 × 1, where 40 denotes the number of MFCC coefficients, and 100 denotes the number of time frames after padding or truncation. The architecture begins with a Conv2D layer with 32 filters and a 3×3 kernel, using the same padding and ReLU activation, followed by a MaxPooling2D layer with a 2×2 pool size to reduce the spatial dimensions. A second Conv2D layer with 64 filters and a 3×3 kernel is then applied, again followed by MaxPooling2D. A third Conv2D layer with 128 filters and a 3×3 kernel, followed by MaxPooling2D, extracts higher-level feature representations. The resulting feature maps are flattened into a one-dimensional vector of length 7,680. To mitigate overfitting, a Dropout layer with a rate of 0.5 is applied, followed by a fully connected dense layer with 256 units and ReLU activation, another Dropout layer with a rate of 0.3, and finally an output dense layer with 50 units (one per speaker) and Softmax activation, which produces a probability distribution across all speaker classes. The total number of trainable parameters in this architecture is approximately 2.07 million.

Noise Robustness
The integration of noise reduction during preprocessing and the inclusion of noisy samples during training improved the model's resilience to environmental interference. This enables more reliable speaker identification even in non-ideal recording conditions.

Dataset Diversity:
The use of a locally collected dataset with varied participant demographics, recording environments, and device types enhances the model's ability to generalize across different speech patterns and conditions.

Augmentation Integration:
Data augmentation techniques were embedded within the training pipeline to expand the effective dataset size and introduce variability. This contributed to improved model stability and reduced the likelihood of overfitting.

Model Development
This section presents the development of the Convolutional Neural Network (CNN) model used for speaker identification. The model was designed to learn discriminative patterns from extracted speech features, particularly MFCC representations, and to classify input audio samples into their corresponding speaker categories. The development process involved defining the architecture, selecting appropriate training strategies, and implementing the model using suitable computational tools (Kheddar et al., 2023).

CNN Architecture Design
The CNN architecture was structured to effectively process two-dimensional MFCC feature inputs and learn both spectral and temporal patterns associated with individual speakers.

Input Shape (MFCC-based):
The input to the CNN model consists of MFCC feature matrices extracted from audio signals. Each sample is represented as a two-dimensional array (coefficients × time frames), typically standardized to a fixed size (e.g., 40 × 100). An additional channel dimension is introduced, resulting in an input shape suitable for convolutional operations.

Convolutional Layers
The convolutional layers extract local feature patterns from the MFCC inputs. These layers apply filters (kernels) that slide across the input matrix to produce feature maps. The convolution

operation can be expressed in equation 1 as:

$$y(i,j) = \sum_m \sum_n x(i+m, j+n) \cdot w(m,n) + b \quad (1)$$

where $x(i,j)$ represents the input feature matrix, $w(m,n)$ denotes the convolutional kernel, b is the bias term, and $y(i,j)$ is the resulting feature map. This operation enables the model to capture important local dependencies within the speech signal, such as frequency transitions and temporal variations (Reddy et al., 2023).

Pooling layers are typically introduced after convolutional layers to reduce dimensionality and retain dominant features, thereby improving computational efficiency and reducing overfitting.

Activation Functions:

Non-linear activation functions are applied after each convolutional layer to introduce non-linearity into the model, allowing it to learn complex patterns. The Rectified Linear Unit (ReLU) is commonly used and is defined in equation 2 as:

$$f(x) = \max(o, x) \quad (2)$$

ReLU helps mitigate the vanishing gradient problem and accelerates training convergence (Arpita et al., 2025). In the final layer, a Softmax activation function is applied to produce probability distributions over the speaker classes.

Training Strategy

The model was trained for 60 epochs using a batch size of 64. The Adam optimizer was employed with a fixed learning rate of 0.001, and categorical cross-entropy served as the loss function, suitable for multi-class speaker identification with 50 output classes. The dataset was partitioned with 70% of samples used for training and 30% for validation. No early stopping or model checkpointing callbacks were used, allowing the model to complete all 60 epochs; the final model weights were retained for evaluation. Training was performed on an 8GB NVIDIA GPU, with a total training time of approximately 45 minutes.

Implementation Tools

The model was implemented in Python for its flexibility and extensive support for machine learning and signal processing tasks.

Python:

Python was the primary programming language for developing and implementing the system. Its simplicity and wide adoption in the research community make it suitable for rapid prototyping and experimentation.

Libraries (Librosa, TensorFlow/Keras):

The Librosa library was used for audio processing tasks, including loading audio files and extracting MFCC features. TensorFlow and Keras were used to design, train, and evaluate the CNN model. These frameworks provide efficient tools for building deep learning architectures and handling large-scale data processing tasks (Deka & Kumari, 2025).

Model Evaluation Metrics

The performance of the proposed CNN-based speaker identification model was assessed using standard evaluation metrics that provide insight into its classification accuracy, error rate, and overall reliability. These metrics enable a comprehensive understanding of how well the model performs on both seen and unseen data (Jahangir et al., 2022).

Accuracy:

Accuracy measures the proportion of correctly classified audio samples relative to the total number of samples. It provides a direct indication of the model's effectiveness in identifying speakers.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where *TP* represents true positives, *TN* true negatives, *FP* false positives, and *FN* false negatives. In the context of speaker identification, accuracy reflects the percentage of speech samples correctly assigned to their respective speakers.

Loss:

Loss quantifies the error between the predicted output and the actual target values. It is used during training to guide the optimization process. In classification tasks, categorical cross-entropy loss is commonly applied: (see equation 4)

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (4)$$

where y_i represents the true label and $\hat{y}_i$ denotes the predicted probability. A lower loss value indicates better model performance and more accurate predictions (Hussein et al., 2025).

Confusion Matrix:

The confusion matrix provides a detailed breakdown of classification performance across all speaker classes. It shows how many instances were correctly or incorrectly classified for each class. The diagonal elements represent correct predictions, while off-diagonal elements indicate misclassifications. This metric is useful for identifying patterns of confusion between similar speakers.

Generalization Performance:

Generalization performance refers to the model's ability to perform well on unseen data. It is assessed by comparing training, validation, and test results. A significant gap between training and validation accuracy indicates overfitting, where the model learns training-specific patterns rather than generalizable features (Kheddar et al., 2023). The inclusion of validation and test datasets ensures a reliable evaluation of real-world performance.

Reproducibility Statement

To facilitate independent replication of this study, the following implementation details are provided. A fixed random seed (42) was set for NumPy, TensorFlow, and Python's random module prior to model training to ensure deterministic behavior across runs. The model was implemented in Python 3.9 using the following key libraries: TensorFlow 2.13 (Keras API) for deep learning model construction and training; Librosa 0.10.1 for audio loading, preprocessing, and MFCC feature extraction; NumPy 1.24 for numerical operations; and Scikit-learn 1.3 for evaluation metrics. Model training was performed on an 8GB NVIDIA GPU, with a total training time of approximately 45 minutes for 60 epochs; CPU-only execution remains possible, though with substantially longer training times. The source code, preprocessed dataset, and trained model weights are available from the corresponding author upon reasonable request.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

Ethical Considerations

Ethical considerations were integral to this research due to the sensitive nature of voice data, which can be considered personally identifiable information. Appropriate measures were implemented to ensure that data collection and usage adhered to ethical standards and respected participant rights (Hutiri & Aaron, 2022).

Privacy Preservation:

To protect participant identity, all collected speech data was anonymized. Personal identifiers were removed, and recordings were labeled using coded identifiers rather than names. This ensured that the data could not be directly traced back to individual participants.

Consent:

Informed consent was obtained from all participants prior to data collection. Each participant was provided with clear information regarding the purpose of the research, the type of data being collected, and how the data would be used. Participation was entirely voluntary, and individuals had the option to withdraw at any stage of the study.

Data Security:

All recorded audio data was securely stored to prevent unauthorized access. Measures such as restricted access and controlled storage environments were implemented to safeguard the dataset. These practices ensured compliance with ethical research standards and reinforced trust between the researcher and participants.

RESULTS AND DISCUSSION

This section presents the results obtained from the developed Convolutional Neural Network (CNN)-based speaker identification system. The evaluation is conducted using both the locally collected dataset and the anchor dataset from prior work. The results are reported using key performance metrics, including accuracy, loss, and confusion matrix analysis. Interpretation is integrated within each subsection to provide a clear understanding of model behavior, learning patterns, and classification performance under varying data conditions.

Results on Locally Collected Dataset

In this research, Mel-frequency cepstral coefficients (MFCCs) were exclusively used as features. The architecture chosen to process these MFCCs is a Convolutional Neural Network (CNN).

Training and Validation Performance

The training and validation performance of the model is illustrated in Figure IV, which shows the progression of accuracy and loss over 60 epochs. The training curve shows steady improvement in training performance, rising from approximately 8% at the initial stage to over 98% by the final epoch. This indicates that the model effectively learned discriminative speech features from the training data.

Figure IV Training and Validation Curve of the model with the local dataset

The validation accuracy follows a similar upward trend, rising from about 6% to approximately 75% by the end of training. The close alignment between the training and validation curves suggests stable learning behavior, with no significant divergence observed during training. This indicates that the model maintains reasonable consistency when exposed to unseen validation data. The loss curves further support this observation. The training loss decreases significantly from an initial value of approximately 2.8 to a minimal level, reflecting improved prediction confidence. Similarly, the validation loss declines from about 2.7 to approximately 1.07, although with minor fluctuations across epochs. These fluctuations are expected due to variability in the validation samples and recording conditions.

Test Performance

Evaluating the model on the test dataset provides an unbiased measure of its performance on completely unseen data. The model achieved a test accuracy of 75.82% and a corresponding test loss of approximately 1.0716. This result indicates that the model generalizes reasonably well beyond the training and validation datasets. The performance level reflects the model's ability to distinguish between different speakers under real-world recording conditions, where variations in noise, recording devices, and speaking patterns are present. Although the test accuracy is lower than the training accuracy, the difference remains within an acceptable range, indicating a slight overfitting and that the model does not rely solely on memorized patterns from the training data. Instead, it captures meaningful speech characteristics that contribute to reliable speaker identification. From a practical perspective, the achieved performance demonstrates that the system can operate in realistic environments, though improvements in robustness and feature representation could further enhance accuracy.

Confusion Matrix Analysis

The confusion matrix presented in Figure V provides a detailed breakdown of the model's classification performance across individual speakers. Each row represents the actual speaker class, while each column represents the predicted class. The diagonal elements correspond to correct classifications, while off-diagonal elements indicate misclassifications.

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

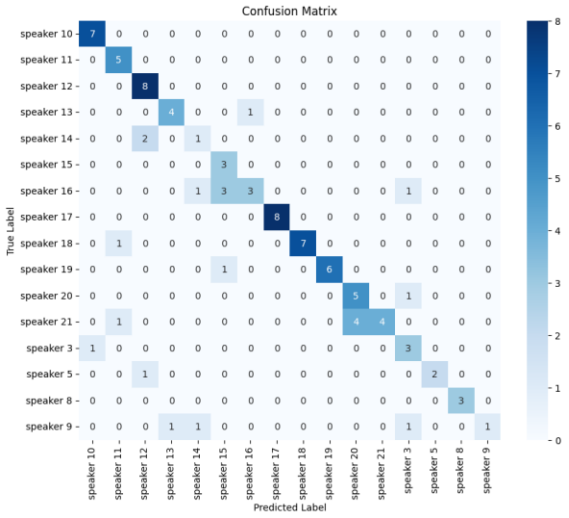

Figure V Confusion matrix of the model on the Local Dataset

An examination of the matrix shows that several speakers achieved relatively strong correct classification rates. For instance, the model recorded up to 6 correct predictions, indicating that it was able to learn and effectively distinguish the speaker's vocal characteristics. Similarly, the speaker 13 shows '4' correct classifications and just '1' instances, demonstrating moderate recognition performance. Further analysis of the off-diagonal elements reveals recurring misclassification patterns. For example, samples belonging to speaker 13 were incorrectly classified as speaker 16 in '1' instances, suggesting a similarity in their speech characteristics. This pattern highlights overlapping acoustic features among these speakers. These results indicate that the model's performance varies depending on the distinctiveness of each speaker's voice. Speakers with clearer or more unique vocal traits tend to achieve higher classification accuracy, while those with similar pitch, tone, or speaking style are more prone to misclassification.

From a practical standpoint, the confusion matrix reveals that the model is effective at identifying certain speakers but has difficulty distinguishing between acoustically similar individuals. This suggests that while the current MFCC-based feature representation captures general speech characteristics, it may not fully separate fine-grained speaker-specific variations.

Performance Under Varying Noise Conditions

To assess the model's robustness to environmental degradation, the test set (225 samples) was stratified into two categories based on qualitative assessments of the original recordings' noise. Clean samples (n = 95) were defined as those with minimal background noise and an estimated SNR greater than 20 dB, while noisy samples (n = 130) exhibited noticeable environmental interference, such as ambient sounds, human activity, or device-induced artifacts, with an estimated SNR below 20 dB. The model achieved an accuracy of 86.3% on clean test samples compared to 68.5% on noisy test samples, representing a performance drop of approximately 18 percentage points. This decline is attributable to the sensitivity of MFCC features to additive noise. Background noise introduces spurious frequency components that mask the fine spectral structure of the speech signal, reducing the discriminability of speaker-specific features. The effect is particularly pronounced for speakers with naturally softer voices or those recorded in high-reverberation environments such as aircraft hangars, where echo further degrades feature clarity. These results are consistent with the findings of Khazaleh & Khrais (2024), who observed similar degradation patterns in CNN-based speaker recognition under environmental noise, and with Toth et al. (2025), who reported that ECAPA-TDNN equal error rates increased from 1% to 6.4% under white noise at 10 dB SNR. The substantial performance gap between clean and noisy samples underscores the need for more aggressive noise-reduction preprocessing, larger and more diverse training data, or the integration of noise-robust feature representations, such as spectrogram-based features or deep embeddings.

Results on Anchor Dataset (Mikhailava et al., 2022)

Training and Testing Performance

The performance of the CNN model on the anchor dataset is illustrated in Figure VI. The training process shows rapid convergence, with accuracy improving steadily across epochs and reaching a perfect score of 1.0000 on the test dataset. Similarly, the loss consistently reduced to a final value of 0.0000, indicating highly confident predictions.

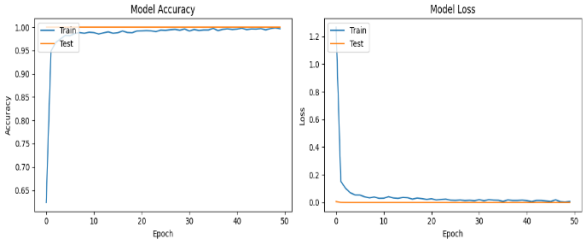

Figure VI Training and Validation Curve of the Model with Anchor Datasets

The alignment between training and testing performance suggests that the model learned the distinguishing features within the dataset effectively. Unlike the locally collected dataset, where variability in recording conditions introduced complexity, the anchor dataset exhibits more consistent acoustic characteristics, contributing to high classification performance.

The near-perfect accuracy demonstrates that the model is highly capable of learning structured and well-represented speech patterns. However, this level of performance is strongly influenced by the dataset's uniformity, recording conditions, and feature consistency across samples. From an evaluation perspective, the results confirm that the CNN architecture, when applied to well-structured data, can achieve very high recognition performance with minimal classification error.

Confusion Matrix of Anchor Datasets

The confusion matrix shown in Figure VII further confirms the model's strong performance on the anchor dataset. The matrix shows a clear concentration of values along the main diagonal, indicating that most samples were correctly classified into their respective speaker categories. Unlike the locally collected dataset, the off-diagonal values are absent, suggesting that misclassifications are absent. Each speaker's samples are consistently mapped to the correct class, reflecting high separability between speaker features within the dataset. This strong diagonal dominance indicates that the model captured highly distinctive patterns for each speaker without confusion

Science World Journal Vol. 21(No 2) 2026 https://dx.doi.org/10.4314/swj.v21i2.13
www.scienceworldjournal.org
ISSN: 1597-6343 (Online), ISSN: 2756-391X (Print)
Published by Faculty of Science, Kaduna State University

between classes. The absence of overlapping classifications suggests that the dataset contains well-defined and distinguishable speech characteristics.

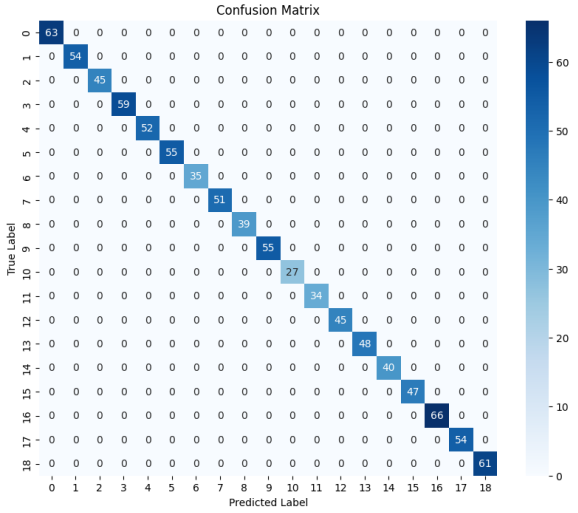

Figure VII Confusion Matrix of the model with Anchor datasets

From a practical standpoint, these results demonstrate that the model performs exceptionally well under controlled conditions where noise levels are minimal and speech features are consistent. However, compared with the locally collected dataset, the performance difference highlights the influence of environmental variability and recording diversity on model behavior.

Comparative Analysis of Results

A comparison between the locally collected dataset and the anchor dataset (Mikhailava et al., 2022), as presented in Table I, reveals clear differences in model performance across multiple evaluation metrics, including accuracy, loss, robustness, and generalization behavior.

Table I: Comparative Analysis of Results

Metric	Locally Collected Dataset	Benchmark Dataset
Training Accuracy	>95%	100%
Validation Accuracy	~73%	~100%
Test Accuracy	75.82%	100%
Test Loss	1.0716	0.0000
Data Quality	Mixed (clean + noisy)	Clean and controlled
Recording Environment	Real-world (classrooms, homes, etc.)	Controlled
Dataset Diversity	High (devices, speakers, noise)	Limited variability
Generalization Ability	Moderate	Very high (within dataset)
Misclassifications	Present (visible in confusion matrix)	Minimal to none
Robustness to Noise	Moderate	Low (not tested under noise)

The locally collected dataset achieved a test accuracy of 75.82% with a corresponding test loss of 1.0716, indicating moderate performance under real-world conditions. These conditions include background noise, variability in recording environments, and differences in recording devices such as smartphones and dictaphones. In contrast, the anchor dataset achieved a test accuracy of 1.0000 and a test loss of 0.0000, indicating near-perfect classification performance under clean, controlled recording conditions. Further insights from Table 4.1 show that while both datasets achieved high training accuracy, the locally collected dataset recorded a validation accuracy of approximately 73%, compared to near-perfect validation performance in the benchmark dataset. This difference highlights the influence of dataset quality and consistency on model learning and validation behavior. The disparity in performance can be attributed to differences in data characteristics. The locally collected dataset comprises both clean and noisy recordings captured across diverse environments, including classrooms, homes, and offices. These variations introduce acoustic distortions, including background noise and reverberation, which affect feature extraction and make speaker discrimination more challenging. As a result, the model exhibits moderate generalization ability, with noticeable misclassifications. In contrast, the anchor dataset consists of clean, uniform recordings, enabling clearer separation of speaker-specific features. This contributes to the model's high generalization performance within the dataset, as reflected in the absence of classification errors. The confusion matrix analysis further supports these observations. The locally collected dataset shows several off-diagonal values, indicating misclassification between speakers with similar vocal characteristics. This reflects the model's sensitivity to overlapping speech patterns under variable conditions. On the other hand, the confusion matrix for the benchmark dataset demonstrates strong diagonal dominance, confirming consistent and accurate classification across all speaker classes. Additionally, the comparison highlights differences in robustness. The locally collected dataset provides a more realistic evaluation of the model's ability to operate under noisy and unpredictable conditions. In contrast, the benchmark dataset does not adequately assess performance under such variability.

Error Analysis of Miscalculations

Examination of the confusion matrix (Figure V) reveals several recurring misclassification patterns that provide insight into the model's limitations. For instance, samples belonging to Speaker 13 were incorrectly classified as Speaker 16, while similar cross-confusion was observed between Speaker 8 and Speaker 21. These errors suggest acoustic similarity between the affected speaker pairs. Several factors likely contribute to these misclassifications. First, speakers of the same gender and similar ages often exhibit overlapping pitch and formant frequencies, which MFCC features may not fully separate. Second, variations in recording conditions—such as one speaker recorded in a quiet office and another in a noisy classroom—introduce environment-specific artifacts that the model may mistakenly associate with speaker identity. Third, speakers with less distinct or more monotone prosodic patterns are more frequently misclassified than those with highly characteristic vocal features, such as distinctive pitch variation or speaking rate. These findings align with observations in prior work (Jahangir et al., 2021; Khazaleh & Khrais, 2024), which similarly report that speaker similarity and environmental variability remain primary challenges for CNN-based speaker identification systems.

Comparison with Existing Studies

The findings of this research, as illustrated in Table II, align with existing studies that highlight the effectiveness of Convolutional Neural Networks in speech-based classification tasks.

Table II Comparison with the existing study

Study	Task	Dataset Type	Features Used	Model	Accuracy (%)	Key Observations
Mikhailava et al., (2022)	Accent classification	Crowd-sourced (European L1 speakers)	Mel-spectrograms	CNN	75%	Limited accent coverage; controlled dataset
Dwijayanti et al., (2022)	Speaker identification	Clean dataset	MFCC	CNN	~85%	High performance under controlled conditions
Hamsa et al., (2023)	Speaker identification (noisy/emotional speech)	Benchmark dataset	Spectrogram + VGG features	Deep CNN (VGG-based)	~90%	Strong robustness with advanced architecture
Proposed Study (Local Dataset)	Speaker identification	Locally collected (noisy + diverse)	MFCC	CNN	75.82%	Real-world variability; moderate generalization
Proposed Study (Benchmark Dataset)	Speaker identification	Clean benchmark dataset	MFCC	CNN	100%	Near-perfect classification; possible data similarity

The work by Mikhailava et al. (2022) reported an accuracy of approximately 75% in accent classification using CNNs trained on sparse and crowd-sourced datasets. Similarly, the present study achieved a test accuracy of 75.82% on the locally collected dataset, indicating comparable performance under conditions involving variability and noise—other studies, such as Dwijayanti et al. (2022) and Hamza et. Al. (2023) have also demonstrated that CNN-based models can achieve high accuracy in speaker identification tasks when trained on well-structured datasets. These studies report improved performance when the training data is extensive, clean, and acoustically consistent, which aligns with the near-perfect results observed on the anchor dataset in this research. However, the current study extends existing work by incorporating a locally collected dataset that reflects real-world recording conditions. Unlike many prior studies that rely on controlled datasets, this research evaluates model performance under varying environmental conditions, including background noise and device variability. This approach provides a more realistic assessment of system reliability. Furthermore, the observed misclassification patterns in the confusion matrix support findings from Jahangir et al. (2021), which emphasize that speaker similarity and feature overlap remain key challenges in speaker identification systems. The difficulty in distinguishing acoustically similar speakers observed in this study reinforces the need for more advanced feature extraction techniques and improved model architectures. The results confirm that CNN-based approaches remain effective for speaker identification, while also highlighting the limitations imposed by data quality and environmental variability. The study contributes to existing literature by demonstrating that meaningful performance can still be achieved using locally sourced data, thereby supporting the development of practical and adaptable speech recognition systems.

Recommendations

Based on the findings of this study, the following recommendations are proposed to improve the performance and applicability of CNN-based speaker identification systems:

First, future research should focus on expanding the dataset size and diversity. Increasing the number of speakers and incorporating a wider range of accents, age groups, and speaking styles will improve the model's ability to generalize across different populations.

Second, advanced noise-reduction and speech-enhancement techniques should be integrated into the preprocessing. This will help mitigate the impact of environmental noise and improve the quality of extracted features, particularly for real-world applications.

Third, adopting more robust feature extraction techniques is recommended. Combining MFCCs with additional features such as pitch-based or prosodic features may enhance the model's ability to distinguish between speakers with similar vocal characteristics.

Fourth, data augmentation techniques should be further explored and refined. Controlled variations such as noise injection, pitch modification, and environmental simulation can help improve model robustness and reduce overfitting.

Fifth, future work should consider using more advanced deep learning architectures. Models such as hybrid CNN-RNN or attention-based networks may provide improved performance by capturing both spatial and temporal dependencies in speech signals.

Finally, strict attention should be paid to ethical considerations, particularly regarding privacy and data security. The implementation of anonymization techniques and secure data handling practices will help ensure that voice data is protected and used responsibly.

Limitations of the Study

Despite the contributions of this research, several limitations were identified during the study.

One limitation relates to the size of the dataset. Although efforts were made to collect diverse speech samples, the number of participants and recordings remains relatively limited, which may restrict the model's ability to generalize across a wider population.

Another limitation is the variability in recording conditions. While this was intentional to simulate realistic environments, differences in background noise, recording locations, and device quality introduced inconsistencies that affected feature extraction and model performance.

The use of different recording devices, including smartphones and a dictaphone, also contributed to variations in audio quality. These inconsistencies may have affected the reliability of the extracted features and, in turn, the classification results.

Additionally, relying solely on MFCC features may limit the model's ability to capture more subtle speaker-specific characteristics. Incorporating additional or alternative feature representations could potentially improve performance.

Finally, the model exhibited moderate generalization performance, as indicated by the gap between training and test accuracy. While the results are promising, further improvements are required to enhance robustness in more complex scenarios.

Conclusion and Future Work

The findings of this study demonstrate that a CNN-based approach, combined with MFCC feature extraction, is effective for speaker identification tasks. The model successfully learned to distinguish between speakers and achieved a test accuracy of 75.82% on the locally collected dataset, reflecting its capability to operate under realistic conditions. The evaluation further showed that model performance improves significantly when trained and tested on structured datasets with consistent recording conditions. This confirms that data quality plays a central role in determining classification accuracy. At the same time, the study revealed challenges associated with environmental noise, speaker similarity, and dataset variability. These factors affect the model's ability to fully generalize across different conditions, indicating areas where further refinement is needed. Future work should focus on expanding the dataset to include a larger and more diverse group of speakers. Increasing the volume and variability of training data would improve the model's ability to generalize across different populations and environments. There is also a need to explore more advanced feature extraction techniques that can capture finer details of speech characteristics beyond MFCCs. Techniques such as hybrid feature representations or deep feature learning approaches may enhance classification performance.

Further improvements can be achieved by using more advanced model architectures, including deeper networks or hybrid models that combine CNNs with other learning approaches. These models may provide better discrimination between similar speakers. In addition, future research should consider real-world deployment and testing of the system to evaluate its performance in practical applications. This would provide deeper insight into system reliability and user acceptance.

REFERENCES

Ali, A. A., Gehad, A. A., Abed, S. A., Aymen, M., Al-Hejri, R. M., Hudhaifa, A. S., & Naqed, S. (2023). Development of Multilingual Speech Recognition and Translation Technologies for Communication and Interaction. *Proceedings of the First International Conference on Advances in Computer Vision and Artificial Intelligence Technologies (ACVAIT 2022)* 176: 711–723. DOI: 10.2991/978-94-6463-196-8_54

Almarshady, N. M., Alashban, A. A., & Alotaibi, Y. A. (2023). Analysis and investigation of speaker identification problems using deep learning networks and the YOHO English speech dataset. Applied Sciences, 13(19), 9567. https://doi.org/10.3390/app13179567

Chen, C., & Huang, L. (2025). A survey of semantic extraction for speech semantics communications: Metrics, approaches, perspectives, and challenges. Engineering Applications of Artificial Intelligence, 158(Part B), 111439. https://doi.org/10.1016/j.engappai.2025.111439

Dwijayanti, S., Putri, A. Y., & Suprapto, B. Y. (2022). Speaker Identification using a Convolution Neural Network. *Jurnal Resti* 6(1) 140-145 https://doi.org/10.29207/resti.v6i1.3795 Einar, H. D. (2023).The most spoken languages worldwide https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/

Encord. (2024, May 14). *Mastering supervised learning: A comprehensive guide*. https://encord.com/blog/mastering-supervised-learning-a-comprehensive-guide/

Farsiani, S., Izadkhah, H., & Lotfi, S. (2022). An optimum end-to-end text-independent speaker identification system using a convolutional neural network. *Computers and Electrical Engineering, 100*, 107882. https://doi.org/10.1016/j.compeleceng.2022.107882

Garvita, B., R., B., P., D., V., & Kakran, S. (2022). *Multidisciplinary research trends* (Vol. 3). Red'Shine Publication.DOI:10.25215/9395456140

Hamsa, S., Shahin, I., Iraqi, Y., Damiani, E., Bou Nassif, A., & Werghi, N. (2023). Speaker identification from emotional and noisy speech using learned voice segregation and speech VGG. *Expert Systems with Applications, 224:*119871. https://doi.org/10.1016/j.eswa.2023.119871

Hutiri, T. W., & Aaron, D. (2022). Bias in Automatic Speaker Recognition. *ACM Conference on Fairness, Accountability, and Transparency (FAccT' 22),* 21–24. Seoul, Republic of Korea

Ivanko, D., Ryumin, D., Kashevnik, A., Axyonov, A. & Karnov, A. (2022). Visual Speech Recognition in a Driver Assistance System. 30th European Signal Processing Conference (EUSIPCO) *1131-1135* DOI:10.23919/eusipco55093.2022.9909819

Jakubec, M., Jarina, R., Lieskovska, E., & Kasak, P. (2024). Deep speaker embeddings for Speaker Verification: Review and experimental comparison. *Engineering Applications of Artificial Intelligence*, 127(Part A), 107232. https://dblp.uni-trier.de/rec/journals/eaai/JakubecJLK24.html.

Khazaleh, O.R. & Khrais, L.A. (2024). An investigation into the reliability of speaker recognition schemes: analyzing the impact of environmental factors utilizing deep learning techniques. *Journal of Engineering. Applications Sciences.* 71(13): 1–30. https://doi.org/10.1186/s44147-023-00351-0

Kheddar, H., Himeur, Y., Al-Maadeed, S., Amira, A. & Bensaali, F. (2023). Deep transfer learning for Automatic Speech Recognition: Towards better Generalization, Science Direct, 277*,* 277,110851

Karthikeyan, V., Priyadharsini, S. S., & Balamurugan, K. (2025). Attention-based multidimensional fused-feature convolutional neural network framework for speaker recognition. *Multimedia Tools and Applications*, 84(31), 38927–38957. https://doi.org/10.1007/s11042-025-20694-5

Laing, Q. (2024). Automatic Speech Recognition Technology: History, Applications, and Improvements. *Applied and Computational Engineering, 65, 180–184.*Doi:10.54254/2755-2721/65/20240493

Latif, S., Cuayáhuitl, H., Pervez, F. (2023). A survey on deep reinforcement learning for audio-based applications. *Artif Intell Rev* 56, 2193–2240. https://doi.org/10.1007/s10462-022-10224-2

Lesnichaia, M., Mikhailava, V., Bogach, N., Lenhenin, I., Blake, J., & Pyshkin, E. (2022). Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms. *Interspeech* DOI:10.21437/Interspeech.2022-462

Liu, Y., & Ab Rahman, F. B. (2025). A systematic literature review of research on automatic speech recognition in EFL pronunciation. Cogent Education, 12(1), 1–15. https://doi.org/10.1080/2331186X.2025.2466288

Meftah, A. H., Mathkour, H., Kerrache, S., & Alotaibi, Y. A. (2020). Speaker identification in different emotional states in Arabic and English. *IEEE Access, 8,* 60070–60083. https://doi.org/10.1109/ACCESS.2020.2983029

Mikhailava, V., Lesnichaia, M., Bogach, N., Lezhenin, I, Blake, J. & Pyshkin, E. (2022). Language Accent Detection with CNN using Sparse Data from a Crowd-Source Speech Archive. *Mathematics* 10(16), 2913-2943 https://doi.org/10.3390/math10162913

Panarin, R. (2024). Semi-supervised learning explained: Techniques and real-world applications. Mad Devs. https://maddevs.io/blog/semi-supervised-learning-explained/

Polamuri, S. R., Kumbhkar, M., & Pon Daniel, D. A. (2022). *Introduction to deep learning* (1st ed.). AGPH Books (Academic Guru Publishing House). ISBN: 978-93-94339-21-7.

Rahul, M., Jha, S. K., Verma, S., Yadav, V., & Dellwar, D. K. (2023). An efficient Multilingual Speaker Recognition system using a fusion technique. *2nd International Symposium on Smart Cities Challenge Technologies and Trends, New Delhi, India*

Singh, M. K. (2023). A text-independent speaker identification system using ANN, RNN, and CNN classification techniques. Multimedia Tools and Applications, 83(16), 1–13. https://doi.org/10.1007/s11042-023-17573-2

Stuhlmann, L. (2025). Evaluating the Effectiveness of Transformer Layers in Wav2Vec 2.0, XLS-R, and Whisper for Speaker Identification Tasks. *arXiv preprint*, arXiv:2509.00230. https://arxiv.org/abs/2509.00230v1

Swietlicka, I., Kuniszyk-Jozkowiak, W., & Swietlicka, M. (2022). Artificial Neural Networks combined with principal component analysis for non-fluent speech recognition. *Multidisciplinary Digital Publishing Institute.* 22(1), 321: Doi.org/10.3390/s22010321

Taha, T. M., Messaoud, Z. B., & Frikha M. (2024*)*. Convolutional Neural Network Architectures for Gender, Emotional Detection from Speech and Speaker Diarization *International Journal of Interactivity Mobile Technologies 18(3)*:88-103 https://doi.org/10.3991/ijim.v18i03.43013

Toth, B., et al. (2025). FT-Boosted SV: Towards Noise Robust Speaker Verification for English Speaking Classroom Environments. *Interspeech 2025.* arXiv:2505.20222

Tu, Y., Mak, M. W., Lee, K. A., & Lin, W. (2025). ConFusionformer: Locality-enhanced Conformer through multi-resolution attention fusion for speaker verification. *Neurocomputing*, 644, 130429. https://doi.org/10.1016/j.neucom.2025.130429

Vaessen, N., Ordelman, R., & van Leeuwen, D. A. (2025). Self-supervised learning of speech representations with Dutch archival data. *arXiv preprint*, arXiv:2507.04554. https://arxiv.org/abs/2507.04554

Wang, Z., & Luo, Z. (2025). Speech separation using advanced deep neural network methods: A recent survey. Big Data & Cognitive Computing, 9(11), 289. https://doi.org/10.3390/bdcc9110289

Zaineb Ben Messaoud, T. M. T., & Frikha, M. (2024). Convolutional neural network architectures for gender, emotional detection from speech, and speaker diarization. International Journal of Interactive Mobile Technologies, 18(3). https://doi.org/10.3991/ijim.v18i03.4301