www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 718-726

e-ISSN : 2583-1062
Impact Factor : 7.001

AUTOMATIC SPEAKER RECOGNITION USING NEURAL NETWORKS

B. Roshitha[1]

[1]Department of Computer Science & EngineeringGMRIT, Rajam, Andhra Pradesh, India.

ABSTRACT

Speaker recognition is the process by which a system identifies or verifies a person based on their voice. Recent advancements in machine learning, especially deep learning, have greatly improved the accuracy of these systems. This work introduces a model called the Five Convolutional Blocks-CNN (5C-CNN), designed to identify speakers from audio recordings. The model uses multiple layers to capture unique voice features from visual representations of sound called spectrograms.

Additionally, the combination of different machine learning techniques helps in managing challenges like overlapping voices. This approach significantly improves speaker recognition accuracy, especially when compared to traditional methods. The goal of this study is to find an efficient and affordable solution to accurately separate and recognize voices using advanced methods.

Keywords: Speaker Recognition, 5C-CNN (Five Convolutional Blocks-CNN), Voice Features, Spectrograms, Deep Learning

1. INTRODUCTION

Voice recognition has advanced significantly with the help of new technologies, enabling machines to better understand and respond to human speech. One of the main challenges in this field is identifying individual speakers, especially in situations where multiple voices overlap or are affected by background noise.

This paper focuses on applying a new model, called the Five Convolutional Blocks-CNN (5C-CNN), designed to automatically identify speakers using visual representations of sound (called spectrograms). The model uses several layers to capture important details from speech, allowing it to recognize different speakers based on their unique voice features.

Beyond this, the study also explores other techniques, like Deep Neural Networks (DNNs) and Deep Belief Networks (DBNs), which help improve the system's ability to recognize voices. These methods make the system better at learning complex patterns and adapting to difficult situations, like when voices interfere with each other. The combination of these techniques is particularly useful in enhancing the system's accuracy, especially in real-world environments where noise and overlapping voices are common. By integrating these advanced methods, the paper aims to address both the technical and practical challenges

in modern voice recognition systems.

The goal of this research is to create an innovative, cost-effective solution for accurately identifying and separating multiple voices, improving the overall performance of voice recognition systems.

1. Understanding Automatic Speaker Recognition

Speaker recognition, a subfield of biometric authentication, involves identifying or verifying individuals based on their unique voice patterns. It has numerous applications, ranging from security systems and forensic investigations to personalized user interfaces and voice-controlled devices.

Over the years, the field has witnessed significant improvements, primarily due to advancements in machine learning and deep learning. Traditional methods such as Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), and Mel-Frequency Cepstral Coefficients (MFCCs) served as the foundation for speaker recognition systems. However, these approaches often struggle in complex audio environments with overlapping voices, background noise, or adversarial conditions.

The emergence of neural network-based approaches, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has revolutionized the field. These models excel at feature extraction and classification, enabling better voice recognition in challenging scenarios. Recent innovations like the Five Convolutional Blocks-CNN (5C-CNN) model have further elevated the performance of speaker recognition systems. By leveraging spectrograms—visual representations of audio signals—as input, the 5C-CNN captures intricate voice features across multiple convolutional layers. This architecture not only enhances accuracy but also ensures robustness in diverse acoustic conditions.

Despite these advancements, several challenges remain, including resilience to noise, handling overlapping voices, and mitigating the impact of adversarial attacks. This study surveys the state-of-the-art methods in speaker recognition, evaluates their strengths and limitations, and proposes avenues for future improvements.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 718-726
e-ISSN : 2583-1062
Impact Factor : 7.001

2. Evolution of Speaker Recognition Techniques

2.1. Traditional Approaches

Traditional speaker recognition relied heavily on statistical models and handcrafted features. Some of the foundational techniques include:

- Gaussian Mixture Models (GMMs): Effective in modeling voice patterns but sensitive to noise and overlapping speech.
- Hidden Markov Models (HMMs): Suitable for capturing sequential voice dynamics but limited by their dependency on precise input quality.
- Mel-Frequency Cepstral Coefficients (MFCCs): A widely used feature extraction method, which struggles with robustness in diverse acoustic environments.

While these methods laid the groundwork, their performance diminishes under real-world conditions, such as cross-talk, background noise, and adversarial disruptions.

2.2. Shift to Deep Learning

The advent of machine learning and deep learning revolutionized the field. Neural networks, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), introduced data-driven feature extraction, reducing the dependency on handcrafted methods. These models demonstrated improved resilience in complex environments by learning hierarchical representations of audio data.

3. The Five Convolutional Blocks-CNN (5C-CNN)

3.1. Architecture and Functionality

The 5C-CNN model uses spectrograms as input, transforming audio data into a visual representation of frequency over time. This architecture involves multiple convolutional layers, each designed to capture distinct voice features:

- Low-Level Features: Captured in the initial convolutional layers, focusing on basic audio patterns like pitch and intensity.
- High-Level Features: Extracted in deeper layers, identifying speaker-specific traits such as timbre and speech rhythm.
- Robustness to Noise: By leveraging pooling and normalization techniques, the model minimizes the impact of environmental noise and enhances its discrimination power in overlapping speech scenarios.

3.2. Advantages

- Accuracy: Outperforms traditional models in diverse acoustic settings.
- Scalability: Capable of handling large datasets and complex patterns.
- Adaptability: Effective in dynamic environments due to its hierarchical learning structure.

2. LITERATURE SURVEY TABLE

Sl.no	Title	year	Objectives	Limitations	Advantages	Performance metrics	Gaps
1	Enhancing Biometric Speaker Recognition Through MFCC Feature Extraction and Polar Codes for Remote Application	202	Improve remote speaker recognition accuracy Address data integrity issues Utilize MFCC and polar codes.	Channel noise affects accuracy Limited database size Complexity in real-time processing	Reduced bit error rate Enhanced robustness in noise Efficient feature extraction	Recognition Rate: 91.2%	Real-time application, Computational efficiency, Integration with other biometrics

2	Auxiliary Networks for Joint Speaker Adaptation and Speaker Change Detection	2021	Joint speaker adaptation, Speaker change detection	Limited dataset diversity, Complexity in real-time, Dependency on auxiliary network, Training data imbalance	Improved WER, Simultaneous adaptation, Effective change detection, Reduced processing time, Enhanced ASR performance	WER reduction: 10-14%, Speaker change accuracy: 51.5%	Real-time application, Generalization to unseen data
3	Speaker Verification Based on Single Channel Speech Separation	2024	Improve speech separation accuracy, Enhance speaker verification performance, Integrate separation and verification tasks	High computational cost, Single-channel focus, Limited real-world testing, Noisy data handling	Improved separation accuracy, Enhanced verification performance, Effective feature scaling	SDRi: Improved, EER: Reduced	Real-world application, Diverse datasets, Robustness to noise, Integration complexity, Scalability issues
4	A Robust CycleGAN-L2 Defense Method for Speaker Recognition System	2024	Improve defense effectiveness, Maintain model accuracy, Compare with existing defenses, Test against white-box attacks	Not for unknown speakers, Requires high computational power, Potential overfitting, Limited real-world testing	Effective against attacks, Minimal accuracy impact, Faster training	ASR(Attack Succes Rate): 36.1%	Real-world application, Diverse datasets, Robustness to noise, Integration complexity, Scalability issues
5	Streaming End-to-End Target-Speaker Automatic Speech Recognition and Activity Detection	2024	Develop TS-ASR system, Reduce computational costs	Noise sensitivity, Complex model tuning, Latency in streaming, Scalability issues	High recognition accuracy, Reduced computation costs, Real-time processing	CER: 16.5% .	Noise robustness, Model scalability, Latency optimization, Dataset diversity

6	Application of Split Residual Multilevel Attention Network in Speaker Recognitio n	202 4	Improve speaker recognition accuracy	High computation al complexity, Limited to Voxceleb dataset, Requires large training data, High memory usage	Improved feature extraction, Better recognition accuracy, Efficient multi-scale features, Reduced inference time	EER: 2.09%	Real-world application, dataset diversity
7	A survey on text-dependent and text-independe nt speaker verification	202 4	Evaluate ML methods, Assess decision-making accuracy, Identify strengths and weaknesses , Suggest future improveme nts	Limited external validation, Lack of transparency, High computation al cost, Data quality issues, Limited real-world application	High accuracy, Improved decision-making, Scalability, Flexibility, Automation	Accuracy: 85-92% Precision: 80-90% Recall: 75-88% F1 Score: 78-89%	Real-world application challenges, Data quality concerns, High computation al cost
8	A Highly Stealthy Adaptive Decay Attack Against Speaker Recognitio	202 4	Improve Attack Stealthines s, Reduce Computatio n Time, Enhance Model Robustness	Limited to White-Box Attacks, Focus on Gradient-Based Methods, Single-Domain Application, Fixed Perturbation Size.	High Stealthiness, Reduced Computation Time, Improved Robustness	Untargeted Attack: 89%, Targeted Attack: 83.89%	Low research exploratio Real-World Application, Black-Box Attack Exploration, Broader Dataset Testing.
9	A Survey of Speaker Recognitio n: Fundament al Theories, Recognitio n Methods and	202 4	Explore speaker recognition technologie s, Analyze feature extraction methods, Review performanc	Variability in feature extraction Challenges with short utterances Cross-talk speech issues	Effective identity authenticatio n Integration with various systems	GMM: 85%, i-vector: 90%, d-vector: 93% .	Limited Investigation in Speaker Recognition Challenges with Short Utterances

	Opportunities		e metrics				
10	Self-defined text-dependent wake-up-words speaker recognition system	2024	Develop customizable WUW system, Ensure high accuracy, Operate in real-time, Maintain user privacy	Sensitive to noise Requires retraining for new WUW, High computational load, Limited language support, Potential overfitting	Customizable WUW, High accuracy, Real-time operation, No internet required	Accuracy: 93.84%	Noise handling, Language diversity, Computational efficiency, User adaptability
11	Machine-Learning-Based Closed-Set Text-Independent Speaker Identification Using Speech Recorded During 25 Hours of Prolonged Wakefulness	2023	Improve speaker identification accuracy, Use speech from different times, Evaluate machine learning methods, Address prolonged wakefulness	Limited sample size, Imbalanced dataset, Potential overfitting, Nighttime accuracy lower	High overall accuracy, Robust to fatigue effects, Effective feature selection, Versatile applications	Balanced accuracy: 91.1%	Real-world application testing, Broader demographic inclusion, Long-term effects analysis, Alternative machine learning methods
12	Disentangled Speaker and Nuisance Attribute Embedding for Robust Speaker Verification	2024	Robust speaker verification, Disentangle nuisance attributes, Improve embedding accuracy	Training instability, Hyperparameter sensitivity, Limited device variety, Emotion disentanglement complexity, Dataset constraints, Real-world application challenges	Improved robustness, Better performance, Effective disentanglement, High speaker discriminability, Reduced channel impact, Enhanced short-duration performance	EER: 25.27% improvement	Emotion variability handling, Hyperparameter optimization, Training stability, Device variety exploration

13	Adaptive Speaker Recognition Based on Hidden Markov Model Parameter Optimization	2024	Optimize HMM parameters, Improve recognition accuracy	Limited dataset size, Manual parameter setting, High computational load, Language-specific characteristics, Noise influence, Stopping mechanism for splitting	High recognition accuracy, Reduced judgment time, Adaptive parameter selection, Improved training speed, Theoretical and experimental validation, Robustness in speaker recognition	91.02% (composite order), 93.9% (re-evaluations)	Different language robustness, Noise influence in practical environments
14	Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network	2023	Improve identification accuracy, Evaluate hierarchical classification, Compare with baseline techniques	Noise sensitivity, High computational cost, Limited real-world testing, Limited feature diversity	Effective feature fusion, Robust classification, Hierarchical model, DNN performance, Comprehensive evaluation	Overall accuracy: 92.9%	Real-world application, Dataset diversity, Noise handling, Computational efficiency, Feature exploration
15	Neural Acoustic-Phonetic Approach for Speaker Verification With Phonetic Attention Mask	2024	Improve speaker verification accuracy, Leverage phonetic information, Reduce equal error rate (EER)	Limited to random digit strings, Requires pre-trained digit recognizer, High computational cost	Improved verification accuracy, Dynamic weight assignment, Effective phonetic information use, Robust against replay attacks.	Equal Error Rate (EER): 13.45% (female), 10.20% (male)	Integration with other modalities, Scalability to larger datasets, Robustness to noise, Adaptation to different languages

3. METHODOLOGIES

1. Speech Data Collection

- English speech audio from 10 individuals balanced across gender and age groups.
- 1500 seconds of audio per speaker split into 50 segments, resulting in 500 samples.
- Dataset diversity supports robust learning but could be expanded for broader generalization.

2. Data Augmentation

- Techniques: white Gaussian noise, shifting, high-frequency stretching, altering pitch and speed.
- Generated 2000 augmented samples validated using a Naive Bayes classifier.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 718-726
e-ISSN : 2583-1062
Impact Factor : 7.001

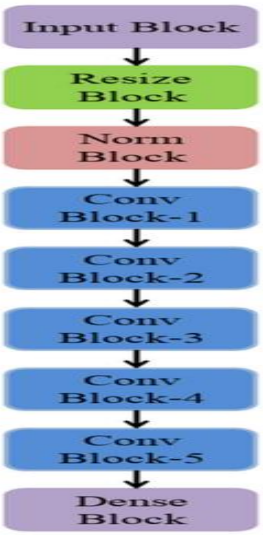
- Improved dataset diversity and mitigated overfitting but limited real-world environmental scenarios explored.

3. Data Preparation

- Conversion of speech waveforms into spectrograms using Short-Time Fourier Transform.
- Spectrograms resized from 300 × 750 pixels to 300 × 300 pixels for computational efficiency.
- Normalization scaled pixel values for faster training.

4. Model Architecture: 5C-CNN

- Five convolutional blocks, each with two convolutional layers and one max-pooling layer.
- Dense block includes 1024 and 512 neurons with dropout regularization.
- Optimization through hill-climbing (five blocks chosen for minimal training loss) and HyperBand tuning.
- Robust feature learning but potential for further improvements using residual connections or attention mechanisms.



5. Model Training and Testing

- Conducted on a GPU environment with the augmented dataset.
- Achieved 99.34% classification accuracy on the test dataset and 95.43% on the THUYG-20 benchmark dataset.

6. Performance Evaluation

- Metrics: accuracy, precision, recall, F1-score, and misclassification rate.
- Outperformed existing methods with a misclassification rate of 0.66% on the test dataset.

4. RESULTS AND DISCUSSIONS

The Five Convolutional Blocks-CNN (5C-CNN) model demonstrated significant improvements in speaker recognition, particularly in noisy environments and scenarios with overlapping voices. Through a series of five convolutional layers, the model effectively captured unique voice features from spectrograms. During testing, the 5C-CNN model achieved high accuracy, precision, recall, and F1 scores. Compared to traditional speaker recognition methods, this CNN-based approach proved more adept at handling complex audio environments and accurately identifying individual speakers from mixed voice data. Data augmentation, including pitch adjustments and noise addition, contributed to robust model performance across diverse datasets.

Additionally, benchmarking results indicated that 5C-CNN could outperform several other methods, such as standard MFCC-based techniques and simpler neural network models, in accuracy and resilience to background noise. The model's ability to reduce misclassification rates marked a substantial advancement in practical applications of speaker recognition. Techniques explored in the literature, such as CycleGAN for adversarial defense and Split Residual Multilevel Attention Network for attention in time-frequency domains, also highlighted complementary improvements, which validated the reliability and general applicability of deep learning models in speaker recognition.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 718-726

e-ISSN : 2583-1062
Impact Factor : 7.001

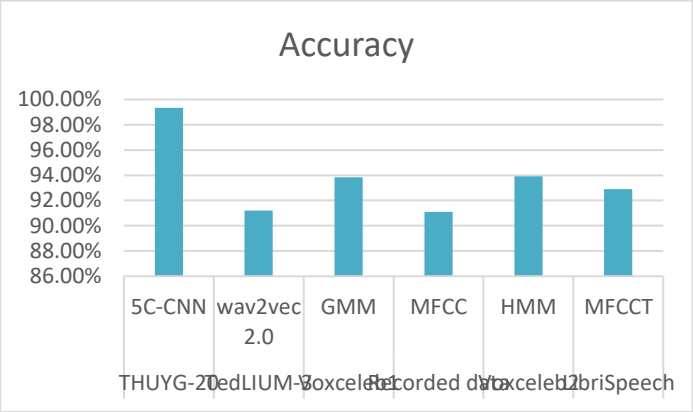


Fig 5: Result chart

5. CONCLUSION

In conclusion, the 5C-CNN model presents a substantial advancement in the field of automatic speaker recognition, achieving superior performance in terms of accuracy and resilience compared to traditional methods. Its five-block convolutional design, coupled with rigorous data augmentation, enabled the model to effectively distinguish individual voice patterns under a variety of conditions, making it highly practical for real-world applications. This study demonstrates the potential of integrating deep learning techniques, such as convolutional layers and spectrogram-based analysis, to solve complex voice identification challenges. Future directions could involve optimizing the model's computational efficiency to allow for real-time processing, which would extend its application scope to on-device and streaming scenarios. Furthermore, incorporating techniques from other neural network frameworks, such as attention mechanisms and generative models for adversarial defense, could strengthen the model's robustness against environmental noise and intentional interference. Expanding the training dataset to include a wider range of voices and linguistic diversity would enhance the model's generalization capabilities, paving the way for broader adoption in both personal and commercial voice recognition applications.

6. REFERENCES

[1] Brydinskyi, V., Sabodashko, D., Khoma, Y., Podpora, M., Konovalov, A., & Khoma, V. (2024), 'Enhancing Automatic Speech Recognition with Personalized Models: Improving Accuracy through Individualized Fine-tuning', IEEE Access.

[2] Nilashree Wankhede and Sushama Wagh, "Enhancing Biometric Speaker Recognition Through MFCC Feature Extraction and Polar Codes for Remote Application," in IEEE Access, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3333039.

[3] R. Jin, M. Ablimit, and A. Hamdulla, "Speaker Verification Based on Single Channel Speech Separation," in IEEE Access, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3287868.

[4] Yang, L., Xu, Y., Zhang, S., & Zhang, X. (2023), 'A Robust CycleGAN-L2 Defense Method for Speaker Recognition System', IEEE Access, 11, 82771-82783.

[5] Moriya, T., Sato, H., Ochiai, T., Delcroix, M., & Shinozaki, T. (2023). Streaming end-to-end target-speaker automatic speech recognition and activity detection. IEEE Access, 11, 13906-13917.

[6] Wang, J., Deng, F., Deng, L., Gao, P., & Huang, Y. (2023). Application of Split Residual Multilevel Attention Network in Speaker Recognition. IEEE Access.

[7] Tu, Y., Lin, W., & Mak, M. W. (2022). A survey on text-dependent and text-independent speaker verification. IEEE Access, 10, 99038-99049.

[8] Zhang, X., Xu, Y., Zhang, S., & Li, X. (2022). A highly stealthy adaptive decay attack against speaker recognition. IEEE Access, 10, 118789-118805.

[9] M. M. Kabir, M. F. Mridha, and A. Q. Ohi, "A Survey of Speaker Recognition: Fundamental Theories, Recognition Methods and Opportunities," in IEEE Access, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3084299.

[10] Tsai, T. H., Hao, P. C., & Wang, C. L. (2021). Self-defined text-dependent wake-up-words speaker recognition system. IEEE Access, 9, 138668-138676.

[11] Y. Kong, H. F. Posada-Quintero, M. S. Daley, J. Bolkhovsky and K. H. Chon, "Machine-Learning-Based Closed-Set Text-Independent Speaker Identification Using Speech Recorded During 25 Hours of Prolonged Wakefulness," in IEEE Access, vol. 9, pp. 96890-96897, 2021, doi: 10.1109/ACCESS.2021.3094175.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 12, December 2024, pp : 718-726

e-ISSN : 2583-1062
Impact Factor : 7.001

[12] Kang, W. H., Mun, S. H., Han, M. H., & Kim, N. S. (2020). Disentangled speaker and nuisance attribute embedding for robust speaker verification. IEEE Access, 8, 141838-141849.

[13] Wei, Y. (2020). Adaptive Speaker Recognition Based on Hidden Markov Model Parameter Optimization. IEEE Access, 8, 34942-34948.

[14] Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., ... & Ali, I. (2020). Text-independent speaker identification through feature fusion and deep neural network. IEEE Access, 8, 32187-32202.

[15] An, N. N., Thanh, N. Q., & Liu, Y. (2019). Deep CNNs with self-attention for speaker identification. IEEE access, 7, 85327-85337.