SPEAKER VERIFICATION USING GEMINI DF RESNET WITH INTEGRATED TRANSFORMATION MODULE

Vinoodhini D, Ajai Ram and Arockia Xavier Annie R

Anna University, India

ABSTRACT

Audio verification is a key biometric authentication method used to confirm an individual's identity based on their voice.This research addresses challenges such as dynamic acoustic conditions (e.g., background noise, reverberation, and microphone variability) and diverse vocal traits to enhance speaker verification robustness.Existing approaches are ineffective in practical situations where security demands necessitate reliable performance under unpredictable environments.Leveraging the DF-ResNet architecture,which integrates a transformation module with depth-first search, our approach optimizes voice feature extraction and analysis.The model was tested on real-world datasets simulating environments like crowded public spaces, quiet offices, and reverberant halls.Its ability to increase accuracy while preserving low computational complexity is demonstrated by experimental results, which makes it a workable option for contemporary biometric identification systems.

KEYWORDS

Biometric Authentication, Depth First Resnet, Transformation Module, Speaker verification.

1. INTRODUCTION

Speaker verification, which verifies a person's identity using their distinctive speech traits, is the foundation of biometric identification.Improving the accuracy, reliability, and efficiency of voice based systems is crucial as they are increasingly used for secure access in applications like mobile devices and banking services.Traditional speaker verification approaches may not work well in real world contexts due to background noise, unpredictability in speech patterns, and fluctuating acoustic circumstances. The Depth-First ResNet (DF-ResNet) architecture has shown promise in improving feature extraction for speaker verification tasks. The well knownResNet architecture is creatively extended by DF-ResNet, which uses residual learning to solve the disappearing gradients issue in deep networks.

By using a depth-first search approach, DF ResNet goes one step further and minimizes computing complexity while enabling the model to concentrate on significant aspects of the voice data. DF-ResNet incorporates a depth-first search strategy to extract essential voice features efficiently. The system is evaluated under various real world conditions to demonstrate its effectiveness against traditional approaches. This research aims to develop a scalable, real time speaker verification solution that enhances security.

The main contributions of this research are as follows:

David C. Wyld et al. (Eds): InWeS, NSEC, VLSIA, NLAIM, CSIP – 2025
pp. 09-23, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2025.151102

1. Developed a DF-ResNet-based speaker verification system that enhances feature extraction by incorporating a depth-first search approach within the ResNet framework, leading to improved accuracy and computational efficiency.
2. Integrated a transformation module that adapts input voice signals to various formats and acoustic conditions, increasing robustness against background noise, speaker variability, and other real-world challenges.
3. Conducted a comprehensive performance evaluation, comparing the proposed system with traditional speaker verification methods across diverse acoustic environments.
4. Optimized the model for real-time efficiency and scalability, ensuring its applicability in large-scale deployments such as mobile devices, banking services, and virtual assistants.

The remainder of this paper is organized as follows: Section 2 provides a brief review of related work on speaker verification. Section 3 describes the dataset used in this study.The general system architecture is described in Section 4, and the suggested system's detailed design is shown in Section 5. Section 6 discusses the experimental results, followed by Section 7, which provides a comprehensive analysis of test cases and performance evaluation. Section 8 concludes the study by summarizing the key findings and discussing possible directions for future research.

2. RELATED WORK

2.1. Self Supervised Learning Approaches

Bing Han et al. presented self-supervised learning with cluster-aware distillation for high-performance robust speaker verification [4], where they proposed a self-supervised learning architecture that replaces labeled data with self-distillation. Their approach incorporates a cluster-aware training technique, significantly enhancing speaker verification performance. Experimental results demonstrate notable improvements in equal error rates (EER) across multiple test sets, highlighting the framework's effectiveness in building a robust speaker verification system without extensive human annotation.

Cai et al. (2023) [9] proposed a self-supervised system that integrates visual and auditory information for speaker recognition tasks. Their approach leverages clustering to generate pseudo-labels, allowing the model to learn without human annotations. This multi-modal framework enhances performance in both single-modal audio and multi-modal audio-visual tasks, demonstrating its adaptability across various scenarios. By incorporating visual data, the system improves recognition accuracy and robustness, making it a practical solution for real-world speaker recognition.

2.2. Uncertainity and Attention Mechanism

Qiong Wang and Kong Aik Lee [2] introduced an uncertainty-aware cosine scoring system for speaker verification.Despite being computationally efficient, typical cosine similarity scoring ignores speaker embedding variability or uncertainty.To address this, the authors propose assessing uncertainty at the front-end embedding stage and incorporating it into the cosine scoring back-end. This method strengthens the system's resilience under trying circumstances and increases its capacity to manage speaker variability.

Zhu and Mak (2023) [6] developed a Bayesian self-attention model to address redundancy in multi-head attention mechanisms for speaker verification. Their method enhances the discriminative power of speaker embeddings by reducing redundant attention heads. By employing probabilistic modeling to optimize attention distribution, the model better captures distinct speaker attributes. This technique significantly reduces the Equal Error Rate (EER) on benchmark datasets such as VoxCeleb and Speakers In The Wild.

2.3. Efficient and Lightweight Models

Wang, Lin, and Zhang (2023) [5] proposed a hybrid model that combines a lightweight Convolutional Neural Network (CNN) with Conformer blocks for efficient speaker verification in resource-constrained environments, such as smartphones. The model enhances feature extraction through channel-frequency attention while improving efficiency by replacing shallow Conformer blocks with depth-wise separable convolutions. This design reduces parameters by 60.6% and Floating Point Operations Per Second (FLOPS) by 36.8%. With an Equal Error Rate of just 0.61% on the VoxCeleb-O dataset, the proposed technique maintains strong performance while keeping computational demands low—ideal for deployment on devices with tight memory and power constraints.

Liu et al. (2023) [7] introduced an adaptive neural network quantization technique for lightweight speaker verification. Their method employs mixed-precision quantization, optimizing accuracy and efficiency by applying different precision levels to network layers.By drastically shrinking the model’s footprint and cutting memory demands without compromising performance, this method is perfectly suited for deployment on resource-constrained edge hardware—such as smartphones and embedded systems.The study demonstrates the advantages of dynamic quantization in achieving high compression with minimal accuracy loss

2.4. Domain Adaptation and Ensemble Methods

Lin and Mak (2022) [3] proposed an approach to mitigate domain shifts in speaker verification, which can significantly impact accuracy and robustness in real-world applications. Their method leverages a deep weight space ensemble, combining the strengths of both base and fine-tuned models. By integrating multiple models trained on different subsets or conditions and merging their outputs, the ensemble approach enhances accuracy and adaptability. The study demonstrates the effectiveness of this strategy in handling mismatches between training and test conditions, improving the resilience of speaker verification systems for real-world deployment

3. DATASET DESCRIPTION

In this study, we employed the English Multi-Speaker Voice Cloning Toolkit (VCTK) corpus [10], which is officially maintained by the Centre for Speech Technology Research (CSTR) at the University of Edinburgh.. This dataset consists of 400 sentences read by 109 English speakers with various accents, providing diverse speech data. The text sources include The Rainbow Passage, a standardized excerpt from the International Dialects of English Archive, an Elicitation Paragraph, and newspaper articles from The Herald Glasgow. While The Rainbow Passage and the Elicitation Paragraph were consistent across all speakers, newspaper excerpts were selected individually for each speaker using a greedy algorithm to optimize contextual and phonetic coverageTable 1 presents the dataset's partitioning and key statistics:

Table 1. Voice cloning toolkit corpus

CSTR DATASET	No. of sentences	No. of Speakers	Utterances	Duration
	400	109	43600	36hrs

4. SYSTEM ARCHITECTURE

The proposed speaker verification system using the Gemini DF ResNet model is shown in Fig. 2. It comprises of several interconnected modules that process, analyze, and evaluate speech-related data for a comprehensive application. The system employs a structured pipeline to ensure efficient handling of audio input via multiple preprocessing approaches, feature extraction methodologies, and validation loops. The system model begins with the Data Acquisition and Preprocessing Module, where raw speech signals from the English Multi-speaker Corpus for Voice Cloning Toolkit are transformed to extract useful features that capture the unique vocal attributes like spectrograms that are the visual depictions of frequency variations over time and Short-Time Fourier Transforms (STFT) that provide detailed time-frequency representations essential for capturing both transient and sustained signal components. Methods such as Mel-Frequency Cepstral Coefficients (MFCCs) compress the spectral envelope into a concise, perceptually meaningful representation, while delta and delta-delta coefficients quantify temporal dynamics by measuring how that spectrum changes over time. Preprocessing includes noise removal, normalization, where audio signals are normalized to maintain uniform amplitude levels across different recordings; and feature extraction, such as Spectrogram, which are visual representations of frequency variations over time domain These extracted characteristics serve as input for the Model Initialization and Training Module, where deep learning techniques such as Resnet are used to learn patterns in voice signals. Through iterative learning, the Training and Optimization Loop guarantees that the model attains high accuracy. During enrollment, each speaker's embeddings are calculated and stored in a database. In the verification phase, test embeddings are compared to enrolled ones using Cosine Similarity, providing accurate speaker verification through similarity scores.

This entails validation-based performance enhancement, loss function minimization, and parameter adjustment. By evaluating the model on unseen data and adjusting hyperparameters as needed, the Validation Loop Module further refines its prediction accuracy.Additionally, the system incorporates a Decision-Making Module that generates a final output after classifying the voice data according to taught patterns.These modules are integrated to form a unified processing pipeline. Preprocessing directly affects model accuracy by refining input data, whereas training and validation loops improve prediction performance. The final decision making module ensures real-world application by making relevant classifications or forecasts.

4.1. MelSpectrogram Generation Module

Mel spectrograms shown in Fig 1 provide a time-frequency representation of speech, highlighting vocal characteristics crucial for text-independent speaker verification (TISV). Time is plotted along the horizontal axis and mel-scaled frequency components along the vertical axis, with color intensity reflecting the distribution of energy. High-energy regions correspond to voiced segments, while darker areas signify silence or unvoiced phonemes. Consistent spectral structures across utterances reinforce their role in speaker verification. In this work, raw audio from the Voice Cloning Toolkit corpus is normalized, resampled, and segmented. A Hamming window minimizes spectral leakage, and the Short-Time Fourier Transform (STFT) converts signals into the frequency domain. Mel-scaled filters simulate human auditory perception, followed by a logarithmic transformation. The resulting Mel spectrograms, stored as numpy arrays, serve as model inputs, capturing speech variations effectively.

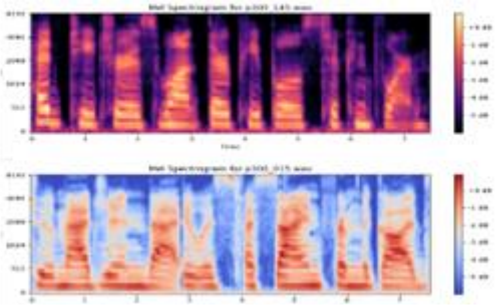

Fig. 1. Mel Spectrogram Visualization

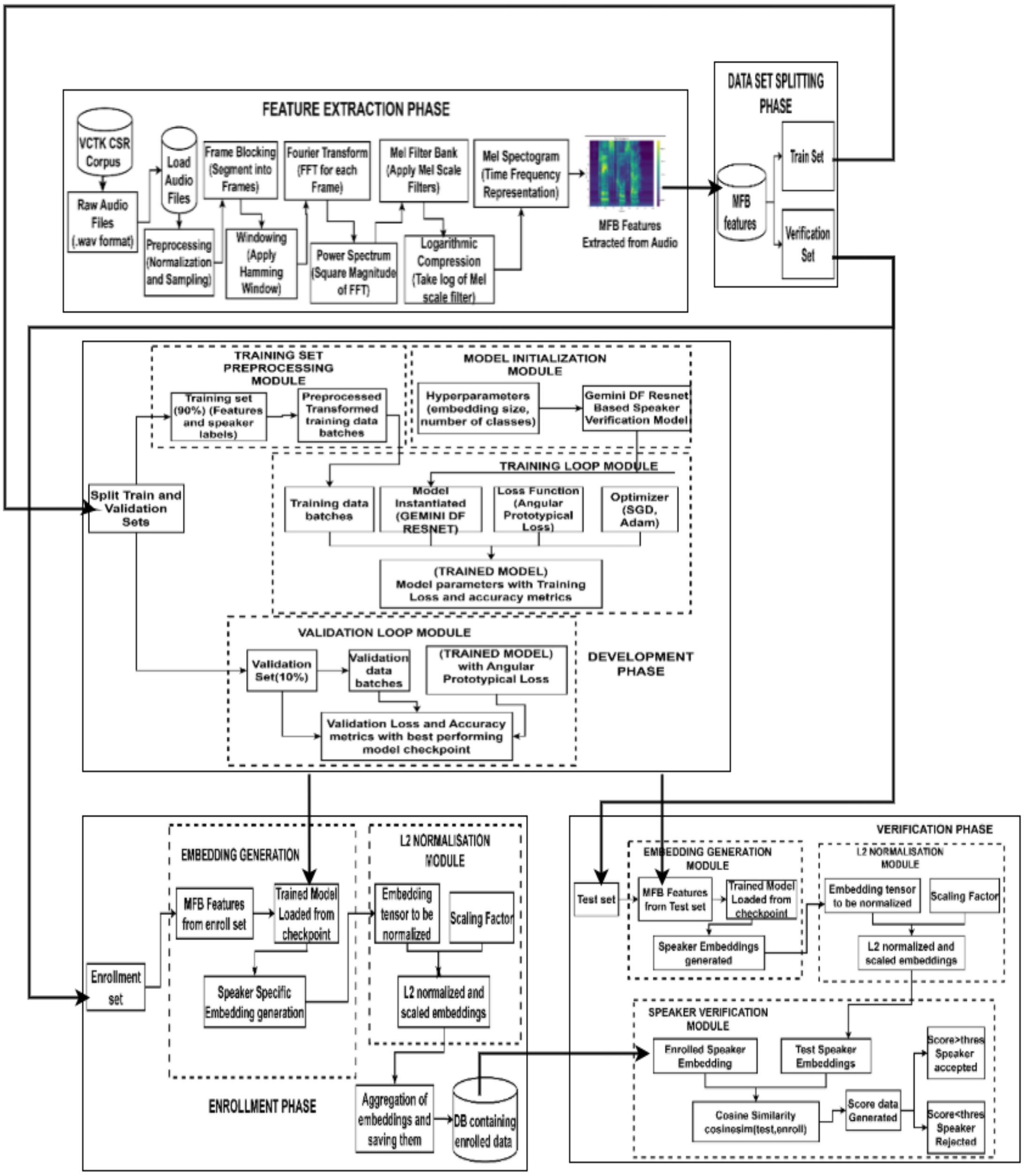
FEATURE EXTRACTION PHASE
VCTK CSR Corpus
Raw Audio Files (.wav format)
Load Audio Files
Preprocessing (Normalization and Sampling)
Frame Blocking (Segment into Frames)
Windowing (Apply Hamming Window)
Fourier Transform (FFT for each Frame)
Power Spectrum (Square Magnitude of FFT)
Mel Filter Bank (Apply Mel Scale Filters)
Logarithmic Compression (Take log of Mel scale filter)
Mel Spectogram (Time Frequency Representation)
MFB Features Extracted from Audio
DATA SET SPLITTING PHASE
MFB features
Train Set
Verification Set
TRAINING SET PREPROCESSING MODULE
Training set (90%) (Features and speaker labels)
Preprocessed Transformed training data batches
MODEL INITIALIZATION MODULE
Hyperparameters (embedding size, number of classes)
Gemini DF Resnet Based Speaker Verification Model
TRAINING LOOP MODULE
Split Train and Validation Sets
Training data batches
Model Instantiated (GEMINI DF RESNET)
Loss Function (Angular Prototypical Loss)
Optimizer (SGD, Adam)
(TRAINED MODEL) Model parameters with Training Loss and accuracy metrics
VALIDATION LOOP MODULE
Validation Set(10%)
Validation data batches
(TRAINED MODEL) with Angular Prototypical Loss
Validation Loss and Accuracy metrics with best performing model checkpoint
DEVELOPMENT PHASE
EMBEDDING GENERATION
MFB Features from enroll set
Trained Model Loaded from checkpoint
Speaker Specific Embedding generation
Enrollment set
L2 NORMALISATION MODULE
Embedding tensor to be normalized
Scaling Factor
L2 normalized and scaled embeddings
ENROLLMENT PHASE
Aggregation of embeddings and saving them
DB containing enrolled data
VERIFICATION PHASE
Test set
EMBEDDING GENERATION MODULE
MFB Features from Test set
Trained Model Loaded from checkpoint
Speaker Embeddings generated
L2 NORMALISATION MODULE
Embedding tensor to be normalized
Scaling Factor
L2 normalized and scaled embeddings
SPEAKER VERIFICATION MODULE
Enrolled Speaker Embedding
Test Speaker Embeddings
Cosine Similarity cosinesim(test,enroll)
Score data Generated
Score>thres Speaker accepted
Score<thres Speaker Rejected

Fig 2. Detailed System Architecture

4.2. Development Module

The Gemini DF ResNet model illustrated in Fig 3 is trained to learn speaker-independent parameters using a large sample of speakers. The Mel spectrogram is reshaped to a size of (N * M, mel_features, frames), where N is the number of speakers, M is the number of utterances per speaker, and mel_features is typically 80. The Gemini DF ResNet34-based SpeechEmbedModel generates 512-dimensional embeddings for each syllable, resulting in a tensor of size (N, M, 512) for loss calculation. The speaker verification model, shown in Fig. 5, leverages residual blocks to learn deep representations without the vanishing gradient issue. The model is optimized using Angular Prototypical Loss to enhance speaker embedding separation. Parameters such as learning rate, batch size, and epochs are tuned for effective training, and an optimizer like SGD or Adam is used to minimize the loss function.

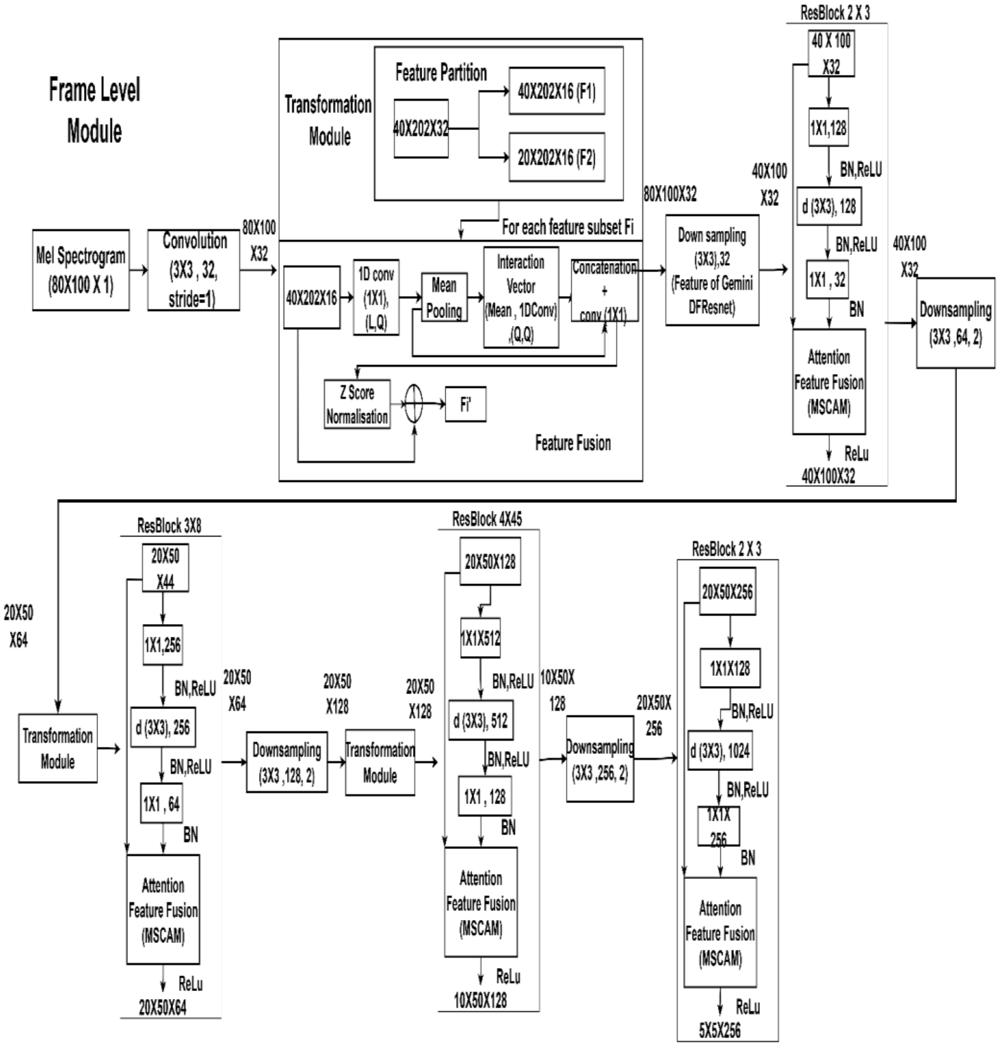

Fig 3. Depth first Resnet with transformation module

4.2.1. Transformation Module

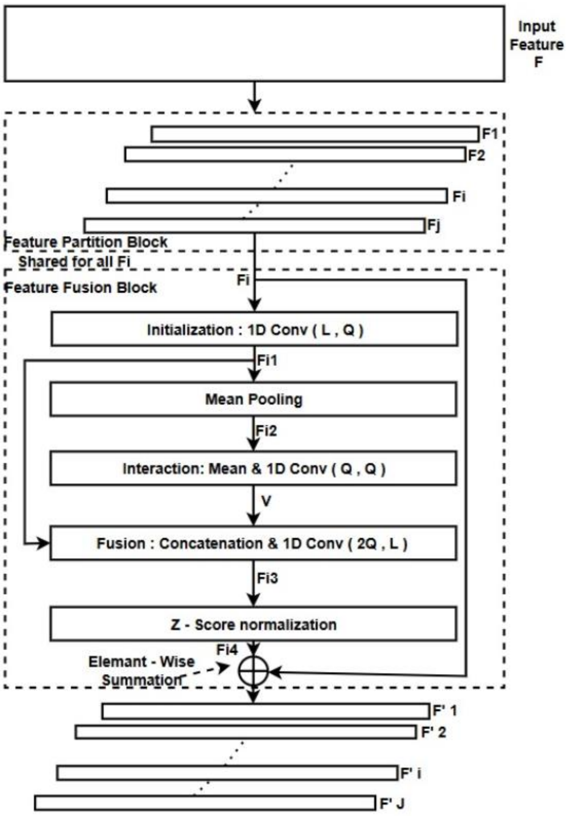

Fig. 4. Transformation Module

The Transformation Module (TM) for speaker verification depicted in Fig 4 enhances speaker verification through a structured approach to multi-scale feature integration. It processes spectrograms by first segmenting them into frequency-channel subsets, isolating local patterns (e.g., transient phonetic elements) to minimize feature redundancy. Each subset undergoes hierarchical processing stages: localized feature extraction via convolutional operations, followed by pooling to retain salient attributes. A fusion block then aggregates these local features into a unified interaction vector, which is recombined with the original subsets. This process preserves fine-grained spectral details and global speaker traits (e.g., vocal identity cues), enabling the model to learn robust, context-aware embeddings. By balancing localized discriminative features with holistic representations, the TM improves verification accuracy under variable acoustic conditions, ensuring reliable speaker discrimination without overfitting to transient artifacts or channel-specific noise.

4.2.2. MultiScale Channel Attention Mechanism

Multi-Scale Channel Attention Mechanism (MSCAM) shown in figure 5, dynamically adjusts thechannel relevance across different feature-map resolutions improving speaker verification. It suppresses noise while giving greater weight to speaker-specific characteristics like harmonic patterns and discriminative auditory characteristics like formant structures and pitch fluctuations using multi-scale processing. Strong angular separations in embedding space are made possible by MSCAM's integration of hierarchical representations, which capture both localized details (phoneme-level cues) and more general prosodic patterns (speaking rhythm). The MSCAM framework processes multi-dimensional feature maps (C × F × T, C is channels, F is frequency bins, and T is temporal frames) to enhance speaker verification. First, Global Average Pooling (GAP) extracts global spatial features, reducing dimensionality while retaining key information. The pooled features pass through a Conv1+BN+ReLU layer, where convolution (Conv1) extracts patterns, Batch Normalization (BN) stabilizes activations, and ReLU introduces non-linearity. The refined features undergo a second convolutional layer (Conv2+BN) for further enhancement, while the original feature representation bypasses these operations for multi-scale fusion. To optimize computational efficiency, a channel compression operation (C / r × 1 × 1) r being the compression factor reduces feature dimensions, and a C × 1 × 2 convolution kernel aggregates time-frequency information. MSCAM then generates two attention representations: G(X) (global features) and L(X) (local details), which are combined via element-wise summation. Finally, a Sigmoid activation normalizes attention weights, dynamically adjusting channel importance. This structured flow refines speaker embeddings, improving verification accuracy and ensuring robust identity recognition across diverse acoustic conditions.

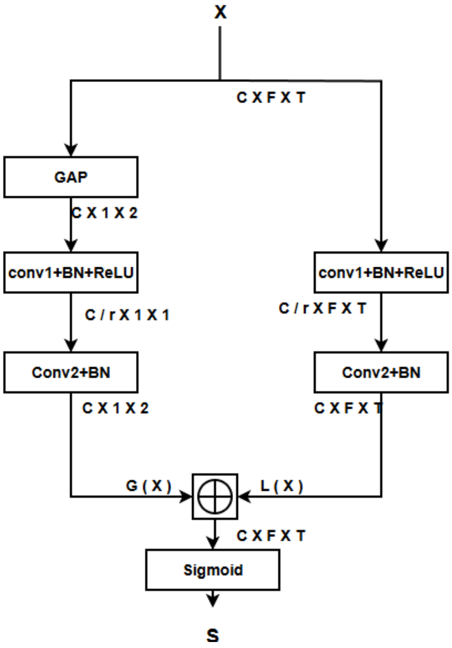

Fig. 5. Multiscale Channel Attention Mechanism

4.2.3. Enrollment and Verification Module

The enrolment module processes raw audio into Mel spectrograms and extracts speaker embeddings using a DF-ResNet model. These embeddings undergo L2 normalization to standardize vector magnitudes, ensuring similarity comparisons depend solely on angular differences between vectors (direction) rather than amplitude variations. During verification, test audio undergoes identical processing. The system calculates cosine similarity between the test embedding and enrolled templates, producing scores between -1 (dissimilar) and 1 (identical). Authentication is considered successful when the score rises above an empirically determined threshold of 0.5, ensuring an optimal balance between security and user convenience.This threshold corresponds to a ≤60° angular difference between embeddings, indicating strong correlation while minimizing false acceptances/rejections.

5. EXPERIMENTAL SETUP

All speaker verification models in this study were trained on a GPU T4X2. To make model training and evaluation easier, the dataset is split into three sets. Partition and statistics of VCTK CSTRcorpus dataset is given in Table 2
Training Set (80%): This section is used to teach the speaker verification model patterns and characteristics unique to each speaker. In order to assess how well the model performs during training, a subset of the training set is reserved for validation (30% of training set). This keeps an eye on measures like accuracy and validation
Test Set (20%): This set is saved for the last assessment of the model's functionality and capacity to accurately identify speakers in data that hasn't been seen yet.

Table 2. Dataset Partition and Statistics

Category	Train	Valid	Test	Total
No of Speakers	61	27	21	109
Utterances	24400	10800	8400	43600
Duration	20 hrs	9hrs	7hrs	36hrs

During training, two-second audio segments were randomly extracted and converted into 40-dimensional Mel spectrograms to improve robustness against speech variations and prevent overfitting. The DF ResNet model employed Angular Prototypical Loss, optimizing speaker embedding discriminability by emphasizing angular separations (over Euclidean distances). Training batches included all 87 speakers (300 utterances each), repeated across three seeds to ensure consistency. For evaluation, ten uniformly sampled segments per recording were processed into L2-normalized embeddings. Cosine similarity scores between test and enrolled embeddings were averaged to mitigate noise-induced variance, with final performance measured using Equal Error Rate (EER).

6. EXPERIMENTAL RESULTS

The performance of the proposed unimodal speaker verification system was evaluated on the VCTK CSR Corpus dataset.

6.1. Test Case Analysis

1. Single Speaker Multiple Utterances: Tests if multiple utterances from the same speaker map to the correct centroid with high similarity. It is done by calculating cosine similarity for each embedding with centroids and verifies if the true speaker matches the predicted speakers.
2. Multiple Speakers, Distinct Utterances: Ensures distinct speakers are not confused with one another. The centroids represent each speaker, and similarity scores are computed for every utterance. The predicted label (based on maximum similarity) is checked against the true label.
3. Genuine and Impostor Pairs Validates that the system can differentiate between genuine (same speaker) and impostor (different speaker) pairs. The similarity score matrix (sim_matrix) indirectly checks impostor cases when scores for non-matching speakers are lower than for matching ones.

Table 3 shows the output of the speaker verification system using cosine similarity (CosSim). It represents whether the model correctly verifies the speaker based on the similarity between their query embedding and the stored reference embedding.

Table 3. Speaker Verification – test cases

Enrolled Speaker	Predicted Speaker	Verification Result	Reason
speaker_p228	speaker_p228	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p276	speaker_p276	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p233	speaker_p233	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p347	speaker_p347	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p333	speaker_p333	Not Authenticated	Similarity score is not within the given threshold the user must exit
speaker_p243	speaker_p243	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p315	speaker_p307	Not Authenticated	Similarity score is not within the given threshold the user must exit
speaker_p297	speaker_p297	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p228	speaker_p347	Not Authenticated	Similarity score is not within the given threshold the user must exit
speaker_p276	speaker_p243	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p276	speaker_p276	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p243	speaker_p243	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified
speaker_p276	speaker_p239	Not Authenticated	Similarity score is not within the given threshold the user must exit
speaker_p239	speaker_p239	Authenticated	Similarity score is greater than the threshold (0.5) the user is verified

6.2. Visualization

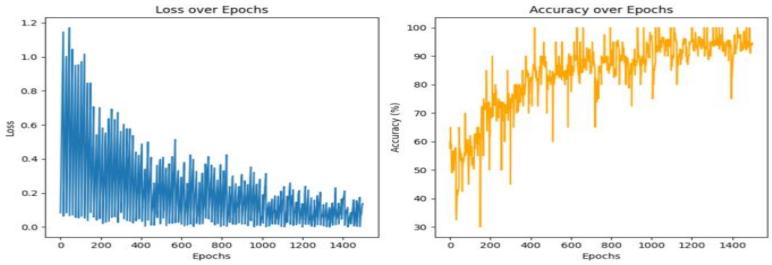

Fig. 6. Accuracy-Loss curve for DFResnet Model Training

The plots shown in Fig 6 illustrates how the model's performance evolves over 100 training epochs, The accuracy-loss curve illustrates the relationship between model accuracy and training loss over multiple epochs. Initially, high loss and low accuracy indicate misaligned predictions. As training progresses, the model refines its parameters, reducing loss and improving accuracy. Eventually, the curve stabilizes, signaling convergence with minimal further improvement. Although the metrics fluctuate, the overall trajectory shows the model converging on an optimal solution, with loss minimized and accuracy maximized by the end of training.

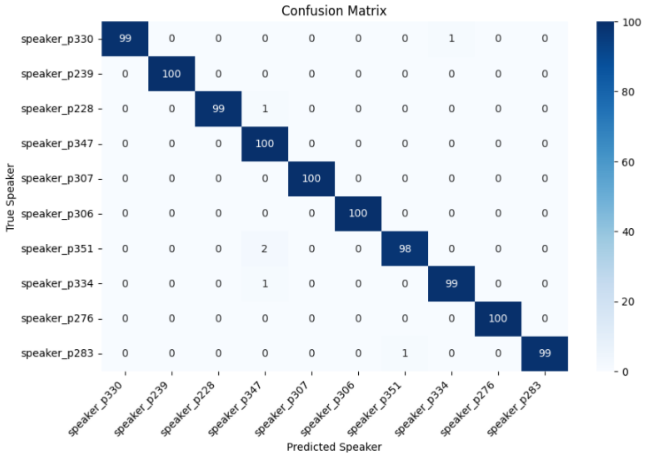

Fig. 7. Confusion Matrix for the DF-ResNet Based Speaker Verification System.

The confusion matrix displayed in Figure 7 demonstrates the system's performance in accurately identifying distinct speakers. In this matrix, the rows correspond to the actual speaker identities, while the columns reflect the labels predicted by the classification system.The near-100 counts on the major diagonal demonstrate that the model typically correctly identifies each speaker, as indicated by the high diagonal values. Off-diagonal entries show few misclassifications because they stay at or close to zero. The effectiveness of the system is reinforced by the colour gradient, darker diagonal cells signify frequent correct matches, whereas lighter off-diagonal tones indicate minimal or no errors in speaker recognition.Overall, the matrix shows that the DF-ResNet-based model achieves excellent accuracy with little misclassification errors, efficiently differentiating between the various speakers.

Table 4. Similarity measurement among enrolled and test embeddings

OBJ	**Enrolled Speaker 1**	**Enrolled Speaker 2**	**Enrolled Speaker 3**	**Enrolled Speaker 4**
Test Speaker 1	0.9825	0.9623	0.9552	0.9791
Test Speaker 2	0.9863	0.8265	0.9947	0.9867
Test Speaker 3	0.9678	0.8127	0.9811	0.9623
Test Speaker 4	0.9662	0.8079	0.9813	0.9614
Test Speaker 5	0.9629	0.9813	0.9234	0.9578
Test Speaker 7	0.9943	0.9031	0.9899	0.9932
Test Speaker 8	0.8609	0.9871	0.7979	0.8528
Test Speaker 9	0.8853	0.9912	0.8236	0.8756
Test Speaker 10	0.9777	0.9622	0.951	0.977
Test Speaker 11	0.9903	0.9201	0.9818	0.9908
Test Speaker 12	0.8562	0.9847	0.7921	0.8466

Test Speaker 13	0.9857	0.9542	0.9652	0.9858
Test Speaker 14	0.986	0.8656	0.9959	0.987
Test Speaker 15	0.9856	0.8654	0.9956	0.9865
OBJ	**Enrolled Speaker 1**	**Enrolled Speaker 2**	**Enrolled Speaker 3**	**Enrolled Speaker 4**
Test Speaker 17	0.9865	0.8675	0.9962	0.9867
Test Speaker 18	0.9876	0.8721	0.9962	0.9887
Test Speaker 19	0.9815	0.8536	0.9953	0.9832
Test Speaker 20	0.9926	0.9073	0.9882	0.9955
Test Speaker 21	0.986	0.8709	0.9933	0.9897
Test Speaker 22	0.9911	0.9367	0.9768	0.993
Test Speaker 23	0.9905	0.9351	0.9768	0.9928
Test Speaker 24	0.9919	0.9037	0.9877	0.9996
Test Speaker 25	0.991	0.9	0.9891	0.9963
EER: 8.2(Thres: 0.55 FAR :8.1 , FRR:8.3)				

The cosine similarity matrix in table 4, compares test speaker embeddings against enrolled speaker embeddings. The raw input has a shape of [4, 6, 160, 40], representing four speakers with six utterances, each with 160 frames of 40-dimensional features. After reshaping, the enrolled data has a shape of [4, 512], and the test data is [4, 6, 512]. The similarity matrix (4, 6, 4) shows how each of the six test utterances aligns with the four enrolled speakers. Values close to 1.0 indicate a strong match, while lower values suggest dissimilarity. The Equal Error Rate (EER) is calculated with a threshold of 0.99, along with the False Acceptance Rate (FAR) and False Rejection Rate (FRR) for performance evaluation.

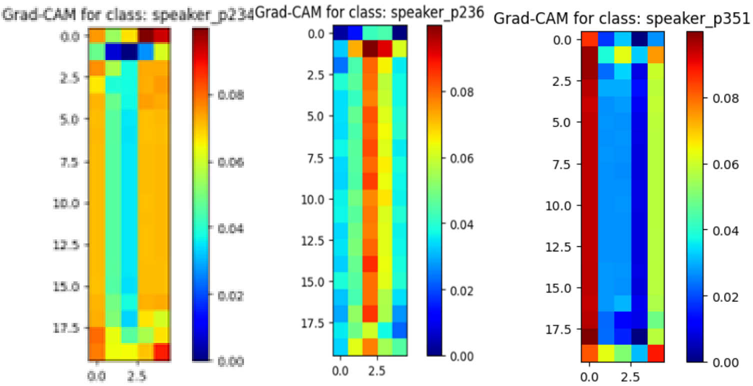

Fig. 8. Grad-CAM Visualization for DF-ResNet-Based Speaker Verification

The Grad-CAM implementation in Fig 8 captures forward activations and backward gradients for a target class by hooking into the DF-ResNet model's last convolutional layer. After loading a spectrogram, activations are preserved during the forward pass, and gradients are collected during the backward pass. Grad-CAM then applies a ReLU operation, weights the gradients by the corresponding activations, and normalizes the result to generate a heatmap. This visualization indicates that the network, places significant emphasis on key spectral regions particularly formant structures, harmonic components, and mid to high-frequency energy bands where speaker-specific characteristics are most pronounced. Additionally, abrupt intensity changes around phoneme transitions appear to be especially informative, as they capture subtle variations in vocal tract resonance and speech articulation. The Grad-CAM heatmap provides valuable insight into how DF-ResNet processes the spectrogram to differentiate among speakers.

6.3. Ablation Studies

6.3.1. Impact of Depth First Design

To investigate the impact of network scaling on speaker verification performance, we systematically varied the depth (*d*) and width (*w*) coefficients of ResNet18 while maintaining comparable computational budgets. Widening the network (increasing w*w*) led to rapid saturation in Equal Error Rate (EER), with performance degrading beyond w=1.4*w*=1.4 despite rising FLOPs. In contrast, deepening the architecture (increasing *d*) consistently improved accuracy, achieving a 35% relative EER reduction over the baseline at similar FLOPs. For instance, at *d*=1.8 (depth scaling), the system outperformed *w*=1.4 (width scaling) by 15% in EER, demonstrating that depth is more critical than width for robust speaker embedding learning.These results shown in Table 5 validate our hypothesis that prioritizing depth over width yields computationally efficient and high-performance models

Table 5.Impact of Depth vs Width Scaling on Resnet 18

Scaling Type	**Coefficient**	**Model variant**	**Params (M)**	**FLOPs (G)**	**EER (%)**
Baseline	d=1.0	ResNet18	4.11	2.22	10.5
Width (w)	w=1.4	ResNet18-Wide	6.82	3.10	10.1
Depth (d)	d=1.8	ResNet34	6.63	4.63	8.1

7. CONCLUSION

This research introduces a novel speaker verification approach using the Depth-First ResNet (DF-ResNet) architecture, which addresses challenges faced by traditional methods in real-world environments. The addition of a transformation module enhances DF-ResNet's resilience to background noise and speaker fluctuations, allowing it to adapt to changing acoustic conditions. Comprehensive evaluations demonstrate that DF-ResNet achieves higher accuracy and greater efficiency compared to conventional approaches, positioning it as a superior choice for real-time processing tasks.By optimizing feature extraction and using a depth-first search method, DF-ResNet maintains strong performance with minimal computational cost. This study advances scalable and efficient speaker verification systems. Future work will explore incorporating additional modalities, domain adaptation, lightweight model optimization, and self-supervised learning to enhance robustness and applicability across diverse settings.

ACKNOWLEDGEMENTS

The authors would like to thank Veaux, C., Yamagishi, J., and MacDonald, K. for providing the CSTR VCTK Corpus [1], which was used in this research. The dataset is available at the Centre for Speech Technology Research (CSTR), University of Edinburgh (https://datashare.ed.ac.uk/handle/10283/3443).

REFERENCES

[1] Abdrakhmanova M., Kuzdeuov A., Jarju S., Khassanov Y., and Varol H. (2021), ‘Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams’, Sensors, Vol. 21, No. 10, pp. 3465.

[2] Wang, Q., and Lee, K. A. (2022), "Cosine Scoring with Uncertainty for Neural SpeakeEmbedding," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

[3] Lin, W., and Mak, M. W. (2022), "Robust Speaker Verification Using Deep Weight Space Ensemble," Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)

[4] Han, B., et al. (2023), "Self-Supervised Learning With Cluster-AwareDINO for High-Performance Robust Speaker Verification," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

[5] Wang, H., Lin, X., and Zhang, J. (2023), "A Lightweight CNNConformer Model for Automatic Speaker Verification," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

[6] Zhu Y., and Mak B. (2023), 'Bayesian Self-Attentive Speaker Embeddings for Text-Independent Speaker Verification', IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 31, pp. 1000-1005.

[7] Liu, B., et al. (2023), "Towards Lightweight Speaker Verification via Adaptive Neural Network Quantization," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)

[8] Wang Q., and Lee K. A. (2024), 'Cosine Scoring with Uncertainty for Neural Speaker Embedding', IEEE Signal Processing Letters, Vol. 31

[9] Wang H., Lin X., and Zhang J. (2024), 'A Lightweight CNN52 Conformer Model for Automatic Speaker Verification', IEEE SignalProcessing Letters, Vol. 31, pp. 56-60.

[10] Veaux, C., Yamagishi, J., & MacDonald, K. (2017), "CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit," Centre for Speech Technology Research (CSTR), University of Edinburgh. Available at: https://datashare.ed.ac.uk/handle/10283/3443

AUTHORS

Vinoodhini D is presently enrolled in the Master of Engineering (M.E.) program in Computer Science and Engineering at the College of Engineering, Guindy, Anna University, India. Her research interests include biometric authentication, deep learning

Ajai Ram is a Research Scholar at the Department of Computer Science and Engineering, College of Engineering Guindy, and also Assistant Professor at the Department of Computer Science and Engineering, College of Engineering Trivandrum. His research interests include Cybersecurity, Network Security, Machine Learning for Intrusion Detection Systems (IDS), and Deep Learning using Generative Adversarial Networks (GANs). He holds an M.Tech in Computer and Information Science from Cochin University of Science and Technology. For correspondence, he can be reached at ajairam@cet.ac.in. His research contributions are available on GoogleScholar: https://scholar.google.com/citations?user=paadfikAAAAJ&hl=en

Dr. Arockia Xavier Annie R, is currently an Associate Professor in the Department of Computer Science and Engineering at Anna University, Chennai. She pursued her doctoral research under the guidance of Dr. P. Yogesh, Professor at the Department of Information Science and Technology. She earned her doctorate in Information and Communication Engineering from CEG, Anna University, Chennai. She is well-motivated and is appreciated by her peers. Her willingness to work even in a confined environment is her major achievement. She has several publications both National and International Journals to her sleeve. Her research interests include video processing, machine learning, security, computing methods in biomedical synthesis, and processing of medical forums through neural networks. Currently, she is the Principal Investigator of Information Security Education and Awareness (ISEA) Project -Phase III, funded by Ministry of Electronics Information and Technology (MeitY), Government of India. Her research contributions are available on GoogleScholar: https://scholar.google.com/citations?user=Yd3f0mAAAAAJ&hl=en