KLE TECH
KLE Technological University
Creating Value
Leveraging Knowledge

School

of

Electronics and Communication Engineering

SRP Project Report

on

Speaker Verification for Multi-lingual Scenarios : RawNet Model

By:

1. **Daivarth B N** USN: 01FE20BEC009

Semester: VIII, 2023-2024

Under the Guidance of

Dr. Nirmala S R

Prof.Satish Chikkamath

K.L.E SOCIETY'S
KLE Technological University,
HUBBALLI-580031
2023-24

SCHOOL OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that project entitled “ **Buiding a RawNet model and comparing with custom CNN for Multilingual Scenarios** ” is a bonafide work carried out by the student team of ” **Daivarath B N (01FE20BEC009)** ”. The project report has been approved as it satisfies the requirements with respect to the Minor project work prescribed by the university curriculum for BE (VIII semester) in School of Electronics and Communication Engineering of KLE Technological University for the academic year 2023-24.

Satish Chikkamath	**Dr.Suneeta V B**	**Prof. B. S. Anami**
Guide	**Head of School**	**Registrar**

External Viva:

Name of Examiners	**Signature with date**
1.	
2.	

ACKNOWLEDGMENT

Without mentioning the names of the people who kindly assisted us in the process of fulfilling the project's aims and with their persistent directives and support, brought about its fulfillment, these come with its effective completion that would have been incomplete. We express our sincere appreciation and gratitude to the School of Electronics and Communication an the Center for Artificial Intelligence Research[CAIR] lab. Dr. Suneeta V Budihal , the Head of the Department, and Prof. Nirmala and Assistant Prof. Satish C Chikkamath for their guidance and encouragement as we completed this project. We are grateful to our esteemed institution KLE Technological University, Hubballi which has given us this opportunity to realize the most cherished desire to achieve our goal. Finally, we would like to express our gratitude to everyone who assisted us, directly or indirectly, in fulfilling the project's goals and objectives.

-The project team

ABSTRACT

This paper showcases the efficiency of several algorithms in case of speaker verification using several features. Mel-frequency cepstral coefficients (MFCCs) are used with great care in the feature extraction process of our speaker recognition system development. MFCCs are well known for their effectiveness in encapsulating the inherent properties of speech, and they form the basis of our model's capacity to identify distinct speaker qualities. We build and train four different models: Convolutional Neural Network (CNN), Artificial Neural Network (ANN), SincNet, and VGG16. The architecture of each model is deliberately designed to navigate the complex complexities found in speech patterns in the English language. Training these models on a variety of multilingual data sets guarantees their flexibility and capacity to identify characteristics unique to individual speakers. Our dedication to accuracy and resilience is demonstrated throughout this procedure by the use of state-of-the-art techniques.

Contents

Chapter 1

Introduction

The project employs creative filter and model design using the primarily declared RawNet architecture, and Convolutional Neural Network (CNN) to advanced speaker recognition in multilingual settings in response to the demands of a globalised society. Driven by the need to build newer models, improve performance and adaptation in the face of linguistic diversity, the initiative strives to recognise speakers more accurately. The goals include building a non-existing newer model according to the architecture declared, evaluating multilingual proficiency, comparing filters, assessing robustness, and examining practical application. In order to achieve through feature extraction, effective neural network building, and classification, the project methodologically comprise RawNet, and Convolutional Neural Network (CNN). A literature review, an explaination of the technique, the specifies of the experimental setting, and the results with comparative insights are all included in the report structure. By offering flexibility in traversing linguistic intricacies on a large scale, this study hopes to advance speaker recognition technology.

This speaker recognition system is based on the careful selection of Mel-frequency Ceptstral Co-efficients (MFCCs) throughout the feature extraction process. MFCCs are well known for their ability to capture the fundamental aspects of speech, and they offer a solid basis for our models to identify minute differences in the acoustic characteristics of the spoken languages. The technique for feature extraction plays a crucial role in identifying the distinctive voice patterns that set each speaker apart. Meanwhile in the demonstration done with dedication to a multimodal strategy by choosing two different models: RawNet, and Convolotional Neural Network (CNN). Every model is customized to maximize the identification of patterns unique to a given speaker within the complex English language speech.

The architecture of each model is deliberately designed to navigate the complex complexities found in speech patterns in the English language. Training these models on a variety of datasets guarantees their flexibilities and capacity to identify characteristics unique to individual speakers. In addition to the more traditional CNN architecture, here this project complimented it with state-of-the-art techniques like RawNet, demonstrating the dedication of this report to accuracy and resilience throughout this process.

In testing, the assessment process is built to thoroughly evaluate the models' performance with speech samples in a language designated as the ”favourite language”. The onjective of this strategic approach is to evaluate the system's ability to generalise and adapt. By combining various models and concentrating on the nuances of the English language, this speaker identification system is positioned as a high-tech solution that can make a substancial contribution to the complex field of multilingual speaker recognition.

1.0.1 Overview and Scope:

The goal of improving speaker identification systems, this research has a broad scope that includes many important aspects. Ml-Frequency Cepstral Coefficients (MFCCs), which are known for their efficiency in capturing speech characteristics, are used meticulously in the feature extraction step. The project uses two different neural network architecture for the model development: Convolutional Neural Network (CNN), and RawNet. Each architecture is specifically designed to identify patterns unique to the speaker, especially in the subtitles of English language speech. By utilizing band-pass filters and hierarchical representations found in voice data, the project's scope is expanded into complex neural network designs through the use case of cutting-edge approaches like RawNet. Despite the fact that English laanguage patterns are the main emphasis, the project's design and techniques are prepared to contribute.

The context of this work, a thorough investigation of Mel-Frequency Cepstral Coefficients (MFCCs) is a fundamental component of feature extraction proceedure, with a focus on their function in encapsulating crucial speech attributes. By adding two neural network models- Convolutional Neural Network (CNN), and RawNet- the project broadens it's scope. Every model is painstakingly created to focus on identifying distinct patterns particularly to a certain speaker within the complex English language.

By incorporating the state-of-the-art method like RawNet, the project shows a forward-thinking attitude and digs into advanced methodology. Through the use of band-pass filters. RawNet introduces novel Neural Network design, improving the system's ability to collect important temporal characteristics. Concurrently, the system's capacity to recognise hierarchical representations in speech data is further improved by the custom CNN's deep and complex architecture.

Although English language speech is still the major focus, the project acknowledges the wider implications for multilingual applications. The system's versatility allows it to function in a variety of linguistic environments, perhaps opening up new application areas outside of English language. This flexibility is demonstrated by the use of several testing methedologies and speech samples in a selected ”favourite language” in the evaluation of the system's flexibility and generalisation capacities in a range of language circumstances. This study, taken as a whole, combines state-of-the-art approaches, diverse models, and excellent feature extraction with an eye towards the nuances of English speech patterns and the larger field of multilingual speaker recognition.

1.0.2 Applications of CNN:

For speaker recognition in the context of this project, Convolutional Neural Network (CNN) provide a flexible and effective tool. CNNs are well-known for their effectiveness in image and signal processing applications. They are also good at learning hierarchical representations, which makes them a good choice for identifying complex characteristics in speech data. In the field of feature extraction, CNNs are highly proficient in automatically recognising and extracting pertinent patterns from unprocessed input, which enables the identification of subtle acoustic properties present in spoken language. This project's primary feature extraction method relies on the network's convolutional layers, which functions as localised filters to efficiently capture temporal dependencies and spatial correlations within the MFCCs.

Additionally, CNN's greatly enhance the speaker identification system's flexibility. The model's ability to identify speaker-specific patterns even in the face of variances in pronounciation, accent, and other linguistic nuances is improved by the network's ability to generalise well to a variety of speaker characteristics thanks to the acquired hierarchical suitability for parallel processing, which facilitates effective computation and scalability. In conclusion, the use of CNNs in this project improves the process of extracting features, makes modelling intricate speech pattern easier, and increases the speaker recognition system's overall resiliance and adaptability.

1.0.3 Applications of RawNet:

The RawNet architecture offers a compelling alternative to traditional approaches in speaker recognition. Unlike the methods relying on pre-defined features like MFCCs, RawNet directly processes the Raw audio waveform. This eliminates the need for complex feature engineering and potentially captures richer information for speaker identification. Additionally, RawNet functions as an end-to-rnd system, directly converting raw audio into a speaker embedding, simplifying the overall process.

RawNet's ability to directly process raw audio waveforms presents several advantages over traditional methods based on handcrafted feaures like MFCCs. Firstly, it eliminates the need for expert knowledge and potentially leading to improved performance. Furthermore, this direct processing approach avoids potential information loss that can occur during feature extraction stages.

While primarily focus on speaker verification, the RawNet architecture's potential extends beyond the application itself[7]. It's ability to directly process the raw audio waveforms makes it suitable for various audio-related tasks, such as speech synthesis, audio anomaly detection, and even music genre classification. Researchers are actively exploring RawNet's potential for various applications in the audio domain, demonstrating it's versatility and promise for future advancements.

1.1 Motivation and Objectives:

The rationale behind the integration of different filtering mechanisms, namely Convolutional Neural Network (CNN), and RawNet is rooted in their distinct advantages and skills, each of which enhances speaker recognition systems in a unique manner.

Convolutional Neural Network (CNN) are good at capturing spatial hierarchies, they are especially useful for image and signal processing applications. CNNs are very good at automatically extracting hierarchical features from MFCCs in the context of speaker recognition, which makes it possible to identify subtle acoustic properties in the voice data. By capturing both temporal dependencies and spatial correlations, their localised filters help the network create a strong representation of speaker-specific patterns.

RawNet utilizes convolutional layers and residual blocks similar to the CNNs. However, these layers tend to operate directly on the raw audio data. This allows the network to automatically learn speaker-specific patterns from the complex temporal and spectral information embedded within the waveform. By capturing both the sequential nature (temporal dependencies) and the relationships between different frequency components(spatial correlations), RawNet aims to create a robust representation for speaker identification.

The objective of this project includes the thorough investigation and refinement of two different filtering mechanisms: RawNet, and Convolutional Neural Network (CNN) - in order to improve speaker recognition systems. RawNet is to be used on temporal dynamics, while CNNs are to capture hierarchical spatial features, and feature extraction techniques are to be improved[6]. Furthuremore, the study attempts to test ow well different filtering techniques compare in terms of being able to identify speaker-specific patterns in Mel-Frequency Cepstral Coeeficients (MFCCs). Through evaluation of the methods' effectiveness across a range of linguistic subleties and variations, the project aims to provide insights into the the complex field of speaker recognition and promote a better comprehension of the advantages and practicality of each filtering techniques.

1.2 Organization of Report

In section 2, the problem statement includes the problem definition with the parameters explained in brief.

In section 3, the system design includes a subsection dedicated to the methodology. This section provides a visual representation of the project flow, highlighting the various blocks involved in the system.

In section 4 of the document consists of three main sections: Specifications and system architecture, algorithm, and flowchart. These sections encompass the detailed information about the project, including the software utilized, the algorithms employed, and the project flowchart that illustrates the sequence of steps.

In section 5 of the document the focus is on the results and discussion, where there is presentation of conducted result analysis, and insights derived from the project are provided. The findings and observations obtained from the project are discussed, analyzed, and interpreted, shedding light on the outcomes and implications of the conducted work.

In section 6 of the document that is dedicated to the Conclusion and Future Scope, with the mentionings of the key findings and conclusions drawn from the project are summarized. The overall significance and implications of the work are discussed, highlighting it's contributions and potential impact. Additionally, the chapter explores avenues for future research and improvements, suggesting ares where the project can be further optimized or expanded upon to address any limitations or open research questions.

Chapter 2

Problem statement

Speaker verification, the process of confirming a person's identity based solely on their voice, holds immense potential across various applications. However, current systems often falter when faced with the diverse linguistic landscapes of this world. This project tackles this challenge by investigating the effectiveness of the RawNet model in comparision to the CNN model in verifying speaker identity impact of language diversity, and potentially refining the model for this purpose, to aim to contribute to the development of multilingual speaker verification systems that can bridge the language gap and enhance speaker identification in various real-world applications and algorithms.

Chapter 3

System design

This project's system design comprises of combining two filtering mechanisms - CNN, and RawNet - into a single, cohorent framework for speaker recognition. The fundamental technique for feature extraction is called Mel-Frequency Cepstral Coefficients, or MFCCs[5]. Each model in the architecture is trained on a variety of English language dataset in order to maximise it's flexibility and capacity to identify characteristics unique to individual speakers. CNNs exploit their natural ability to recognise complex patterns during training[10]. RawNet exploits it's ability to automatically learn features from the raw audio data and it's convolutional architecture to capture both temporal and spatial information.
The methodology for testing include assessing the models with speech samples in a specified "favourite language", guaranteeing a thorough evaluation of the system's flexibility[8]. By combining various filtering techniques, there is hope to develop a strong and adaptable speaker recognition system that can handle the challenges for multilingual scenario.

3.1 Methodology

Define the issue

Define the task of speaker recognition clearly and indicate that the model should be capable of recognising speakers in multiple languages.

Data Gathering

Collect a broad dataset of audio recordings of speaker from various languages. As certain that the dataset is well-balanced and indicative of the intended application.

Data Preparation

Convert audio into a format that can be fed into a CNN (for example, Mel-Frequency Cepstral Coefficients - MFCCs). To maintain consistency between recordings, normalise and standard is the data.

Optional Language Embeddings

Consider adding linguistic information to the model. This might be accomplished by intoducing a language embedding as a new input to the network.

Divide the data

Divide the dataset into three parts: training, validation, and testing. Make sure that each set includes a fair distribution of speakers and languages.

Creating the model architecture

Create a model of architecture capable of extracting speaker-related properties from audio data. To analyse temporal patterns in audio recordings, consider utilising ID convolutional layers. Experimenting with different layer counts, filter-sizes, and pooling algorithms. Include batch normalization and activation functions like ReLU.

Model Compilation

For both speaker and language recognition tasks, use appropriate loss functions. there might be a need to compile numerous loss functions because this is a multi-task situation. Choose an optimizer and metrics that are appropriate for the task.

Educating the model

Using the training set, train the model. Overfitting in the validation set should be monitored, and hyper-parameters should be adjusted accordingly. Consider using data augmentatoin strategies to boost the diversity of your training set.

Examine the model

Examine the extraction of language of language-independent features from audio data. This could imply employing embeddings or representations that are less reliant on language-specific properties.

Language-independent Features

Investigate the extraction of language-independent features from audio file. This could imply employing embedding or representations tat are less reliant on language specific properties.

Optimal fine-tuning and transfer learning

Experimenting with fine-tuning strategies, especially if you have access to huge audio datasets with pre-trained models. When labelled data is scarce, transfer leaning can be useful.

Optional post-processing

Implement post-processing techniques like as smoothing or voting processes across various time frames to improve final predictions.

Deployment

Deploy the model to the target environment, taking into account available computational resources and any limits imposed by real-time processing requirements.

Continuous monitoring and Upkeep

In the production environment, monitor the model's performance and update it as needed. Changes in speaker characteristics and language distributions must be accommodated. Implementation details When undertaking a project, it is to include specific parameters in the implementation details to ensure that the results can be easily comprehended at a glance[14]. These commonlyinclude specifications, architecture, flowchart, algorithm, and methodology. Here are some implementation details to consider.

3.1.1 Specifications and System Architecture

The programming language use is python in it's ipynb that is the jupyter notebook format, so that it is flexible to work either in Jupyter notebook, VS Code or even Google Colab. The framework includes the dataset used is from IIT Guwahati which is an enormous data of 100 speakers which includes the recording of them reading and conversing in English language, and another conversing in their favorite language which might be their mothertongue. The audio dataset is that are recorded from 5 different devices including, 1 Laptop (40K), 1 Table and Chair (10K), 1 Zoom Microphone (25K), 2 Mobiles (1 call IVR, 1 for Recording)/ 1 mobile and 1 medium quality microphone (1 Nokia 5130c and 1 Sony Errison W350i) (20K), and 1 multi-array setup (15K). Therefore, the data gathered is Multilingual using Multisensor (6 sensors) in Multi-environment (2 type) that is Clean (controlled environment) may be an Office and the another is a Noisy (Market/ Railway station) (may be outsource) environment in different styles that is reading and conversation.

3.1.2 Algorithms

Convolutional Neural Network

In order to interprit grid-like input, such as pictures or, in this example, Mel-Frequency Cepstral Coefficients (MFCCs) in speech recognition, Convolutional Neural Networks (CNNs) use a hierarchical and learnable architecture[2]. Convolutional, pooling, and fully linked layers are CNN's fundamental building blocks. Learnable filters are utilised by convolutional layers to capture local patterns and features by convolving over input data. Next, in order to reduce computational complexity while preserving crucial information, pooling layers downsample the spatial dimensions[9]. The network can learn intricate associations and generate predictions because to the hierarchical feature representations that are fed into fully connected layers. Non-linearity is introduced by using non-linear activation functions, like Rectified Linear Unit(ReLU), which improves the network's ability to recognise complex patterns[3]. Overfitting is prevented by regularisation strategies like dropout. The architecture is given with the with notation Fig. 3.1. CNN Architecture.

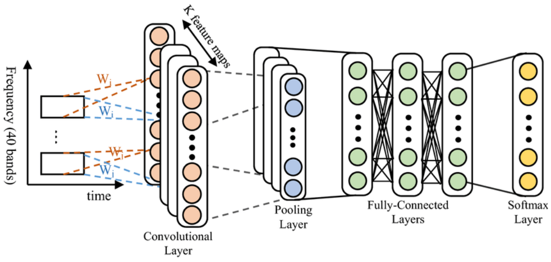
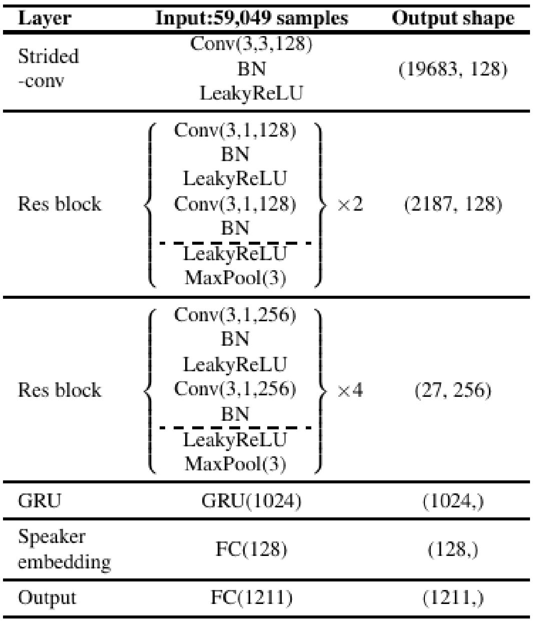


Figure 3.1: CNN Architecture



Layer	Input:59,049 samples	Output shape
Strided -conv	Conv(3,3,128) BN LeakyReLU	(19683, 128)
Res block	{ Conv(3,1,128) BN LeakyReLU Conv(3,1,128) BN LeakyReLU MaxPool(3) } ×2	(2187, 128)
Res block	{ Conv(3,1,256) BN LeakyReLU Conv(3,1,256) BN LeakyReLU MaxPool(3) } ×4	(27, 256)
GRU	GRU(1024)	(1024,)
Speaker embedding	FC(128)	(128,)
Output	FC(1211)	(1211,)

Figure 3.2: Standard RawNet Architecture

RawNet

The DNN used in the standard study comprised residual blocks, agated recurrent unit (GRU) layer[15.16], a fully-connected layer (used for extraction of speaker embedding), and an output layer[1]. In the architecture, input features are first processed using the residual blocks[10] to extract the frame-level embeddings. Where as the model built for multiclass classification tasks, potentially suited for processing 1D audio data, the key component incorporated like the convolutional block for feature extraction with batch normalization and dropout for regularization[13]. A single GRU layer captures temporal information within the audio sequence. The flattened output is further processed by a fully-connected layer with ReLU activation and dropout[4]. Finally, the output layer with softmax activation generates class optimizations like reduced GRU units and dropout rates suggest a more comprehensive analysis and potential for further refinement. The standard RawNet Architecture is displayed as Fig. 3.2. Standard RawNet Architecture.

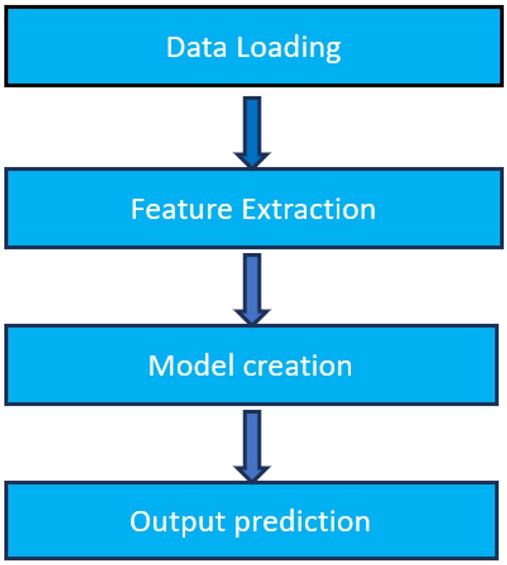


Figure 3.3: Flow Chart

3.1.3 Flowchart

Fig. 3.3. Flow Chart represents the project's flowchart that is the block diagram of the model pipeline and workflow, illustrating the sequential execution of the Machine Learning (ML) algorithm, where each block is individually represented with it's own set of operations[11]. Primarily, the data is loaded using the Librosa library which helps to load and pre-process the data, followed by the feature extrction where around 40 features are extracted, removing unwanted columns and having the desired csv file with only required extracted feature[12]. Using the sklearn library the data is then divided into test and train by splitting the English spoken audio data as the input for training, and later use the favourite language as the testing dataset to verify the autenticity and efficiency of the models.
Followed by creating the model with the chosen algorithm where the layer creation and all takes place, like CNN, and RawNet.
Finally, to know the efficiency by testing with favourite audio dataset the efficiency, loss function are all plotted, and the prediction is displayed showing off the eficiency of the respective model.

Chapter 4

Results and discussions

The performance of two filtering mechanisms - CNN and RawNet - in speaker recognition is highlighted in the findings and discussion. By comparing them, it is possible to see how flwxible CNNs are in identifying subtle speaker-specific patterns in MFCCs. The RawNet exhibits efficacy in maintaining important temporal properties thanks to it's raw audio processing specialisation in both temporal and spatial information. A detailed examination of accuracy, precision, and recall in realtion to linguistic subtleties highlights the advantages of each mechanism, with RawNet and CNN performing well in a variety of temporal subtleties. RawNet is useful for raw audio processing and CNN is better for analyzing. In general, the combination of these filtration methods offers a complex output.

4.1 Result Analysis

4.1.1 Result Analysis of Device-D01

- The speaker verification obtained an impressive 87% accuracy by using 15 epochs to train the Convolutional Neural Network (CNN) (D01) as shown in Fig. 4.1 and Fig. 4.2 is digital voice recoder of 16 kHz/16 bits.

- With this modification, the CNN model can further hone it's capacity to identify complex speaker-specific patterns within Mel-Frequency Cepstral Coefficients (MFCCs), reflecting the effects of extended training.

- The choice to increase the number of training epochs on Device D01 emphasises the importance of the iterative learning process in improving accuracy for real-world speaker recognition applications, while also strengthening the system's resilience and ability to capture complex speech patterns.

- After 300 training epochs, this speaker recognition system's RawNet component attained the accuracy of 53.5%as shown in Fig. 4.3 and Fig. 4.4 is digital voice recoder of 16 kHz/16 bits.. While this might seem a bit low at first glance, it's crucial to consider the context. Comparing this accuracy to a simple baseline, such as Random guessing, provides a clearer picture of RawNet's performance.

- It suggests that RawNet captures some meaningful patterns in the data, if the baseline accuracy is significantly lower, even if it hasn't yet reached the performance of more sophisticated models.

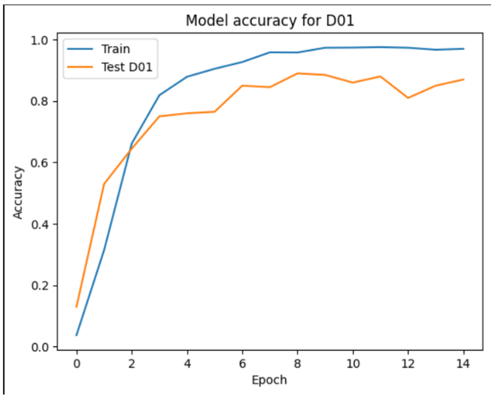


Figure 4.1: CNN Accuracy D01

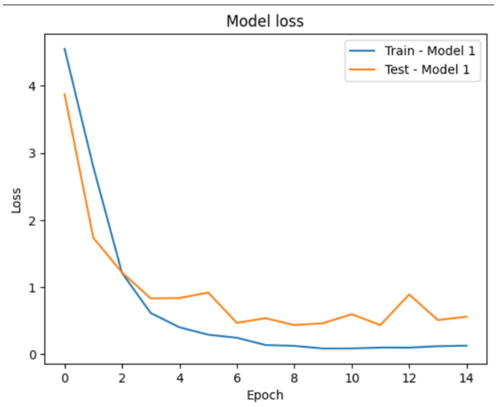


Figure 4.2: CNN Loss D01

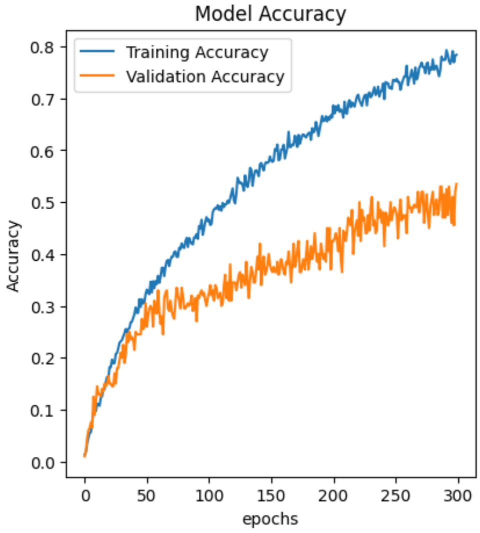


Figure 4.3: RawNet Accuracy D01

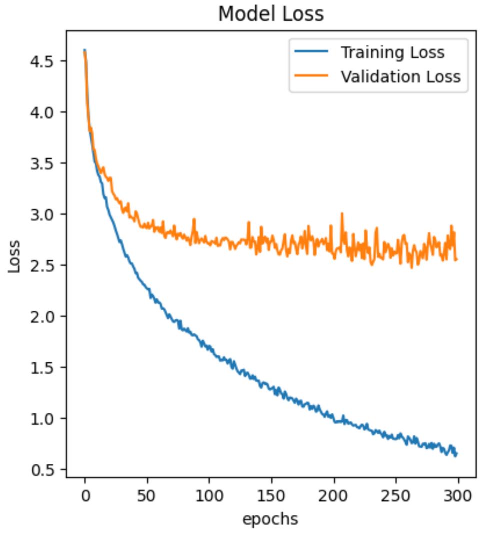


Figure 4.4: RawNet Loss D01

- Further analysis is necessary to understand the limitations of RawNet. Additionally, potential explanations for the current performance, such as model limitations, data issues, or training factors, should be explored to guide future improvements.

4.1.2 Result Analysis of Device-H01

- The speaker recognition system for the Convolutional Neural network (CNN) implementation on Device Headset (H01) with a digital sound featuring a sampling rate of 16 kH/ 16 bits attained an accuracy of 89.99% following the training proceedure as shown in the Fig. 4.5 AND fig. 4.6. This result highlights how flexible and effective the CNN architecture is in capturing complex speaker-specific patterns in Mel-Frequency Cepstral Coefficients (MFCCs) under particular conditions of the headgear.

- The achieved accuracy highlights the CNN model's effectiveness for speaker recognition within the speaker recognition within the specific headset device and shows

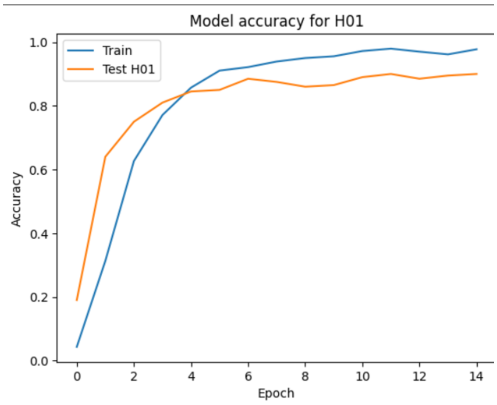


Figure 4.5: CNN Accuracy H01

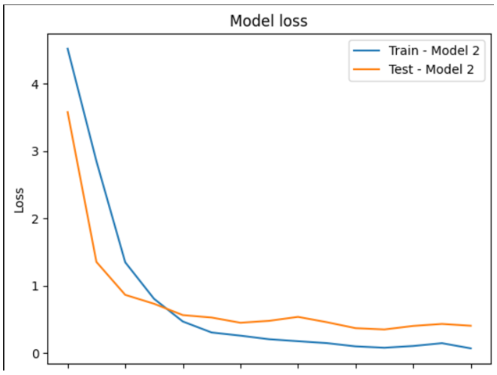


Figure 4.6: CNN Loss H01

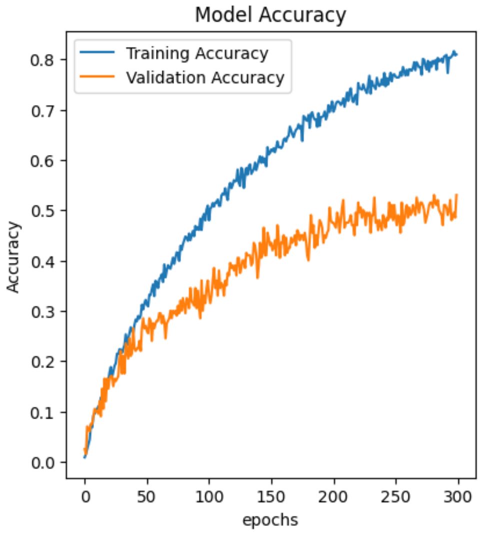


Figure 4.7: RawNet Accuracy H01

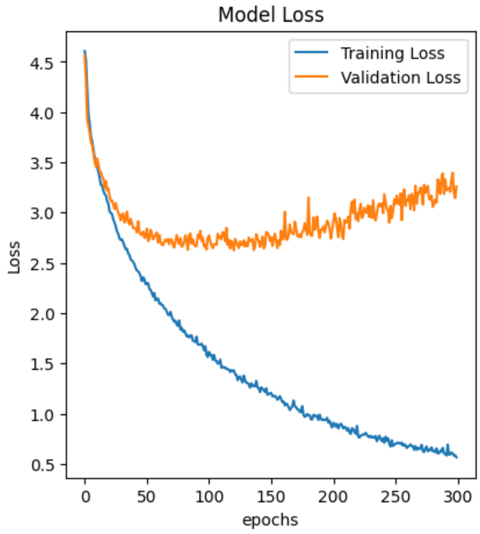


Figure 4.8: RawNet Loss H01

how well it has converged in identifying subtle speech elements.

- While the RawNet model has an accuracy of 52.99% as shown in the Fig. 4.7 AND Fig. 4.8. It is crucial to evaluate it's performance. Compared to a basic benchmark like random guessing, a significantly higher accuracy implies RawNet learns some valuable patterns from the data.

- Additionally, investigating potential reasons for the current performance, such as model complexity, data quality, or training settings, will be crucial for guiding future enhancements.

4.1.3 Result Analysis of Device M01

- The accuracy result of the convolutional Neural Network (CNN) implementation in this speeach Recognition system is 75.49% as shown in Fig. 4.9 and Fig. 4.10 for the Mobile phone (Nokia 5130c) (M01) with a sample rate of 8 kHz/ 16 bits.

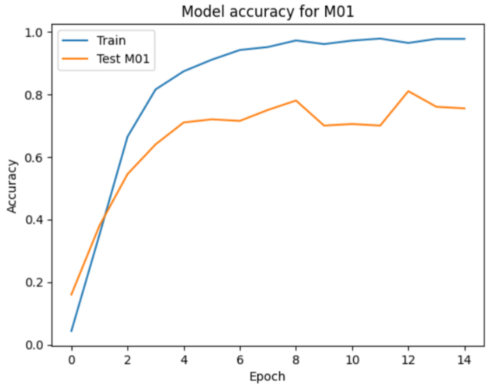


Figure 4.9: CNN Accuracy M01

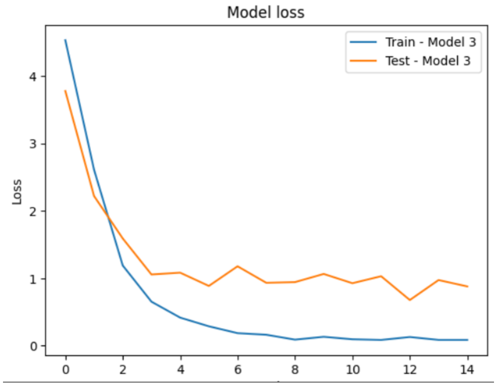


Figure 4.10: CNN Loss M01

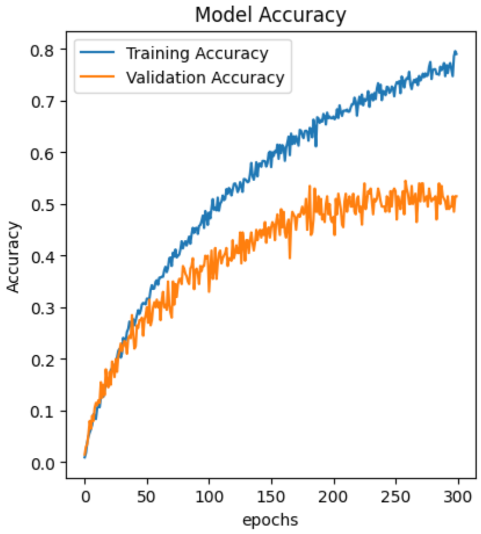


Figure 4.11: RawNet Accuracy M01

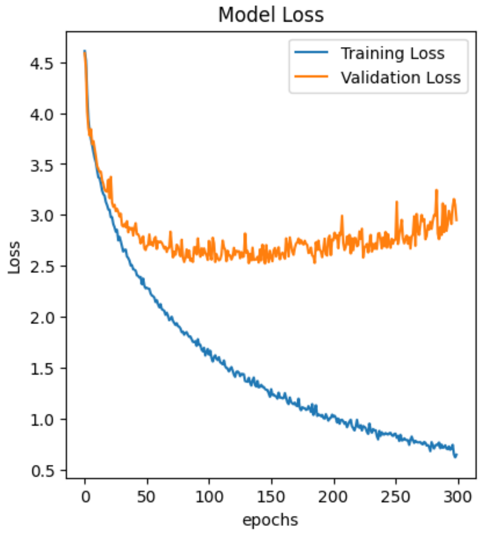


Figure 4.12: RawNet Loss M01

- The results highlight how flexible and effective the CNN model is at identifying speaker-specific patterns in mel-Frequency Cepstral Coefficients (MFCCs) on a mobile device with distinct audio properties. A more thorough assessment of the model's performance of the Nokia 5130c would require more information regarding the accuracy result.

- The RawNet model has achieved 51.49% as shown in the Fig. 4.11 and Fig. 4.12 for Nokia 5130c (M01).

- While this might not be as high as hoped, it signifies the model's ability to perform above chance baseline. For tasks with inheritance difficulty or limited data, achieving an accuracy exceeding random guessing can be a significant first step.

4.1.4 Result Analysis of Device M02

- The application of Convolutional Neural Network (CNN) in this speaker recognition system obtained an excellent accuracy of roughly 76.50% with a loss of 14% as shown in the Fig. 4.13 and Fig. 4.14 for the Mobile phone (Sony Errison W350i) (M02) with a sampling rate of 8 kHz/ 16 bits.

- This result Highlights how flexible and effective the CNN model is in identifying speaker-specific patterns in Mel-Frequency Cepstral Coefficients (MFCCs) on the Sony Errison W350i. The high accuracy emphasizes CNN's strong performance on this particular mobile deviice and further supports it's usefulness for real-world speaker detection applications.

- The application of RawNet in this speaker recognition system obtained accuracy of roughly 48.5% as shown in Fig. 4.13 abd Fig. 4.16 for Sony Erisson W350i (M02). While this accuracy falls short of ideal performance, it's crucial to consider the context.

- Further investigation is crucial to understand the RawNet's limitations. Additionally, exploring potential reasons for the current performance, such as model complexity, data quality, or training settings, will be vital for guiding future improvements and potentially boosting performance.

4.1.5 Result Analysis of Device T01

- The Convolutional Neural Network (CNN) implementation in this speaker recognition system achieved an excellent accuracy of roughly 88.50%, with a low loss of 12% as shown in Fig. 4.17, and Fig. 4.18 for the Tablet (T01) with a sample rate of 16 kHz/ 16 bits.

- This result highlights the flexibility of CNN model in identifying speaker-specific patterns in Mel-Frequency Cepstral Coefficients (MFCCs) on the given tablet. While high accuracy demonstrates the robust performance of the CNN on the tablet (T01) and it's potential for accurate speaker recognition in Tablet environments, the minimal loss suggests effective convergence during training.

- The RawNet implementation in this speaker recognition system achieved roughly 55.5% as shown in the Fig. 4.19 and Fig. 4.20 for the device Tablet (T01). This performance falls short of the desired level and necessitates further exploration.

- To gain deeper understanding, further analysis is recommended. Additionally, investigating potential explanations, such as model limitations, data quality issues, or training regime inefficiencies, is crucial for guiding future improvements and potentially bridging the gap between current perfomance and desired accuracy.

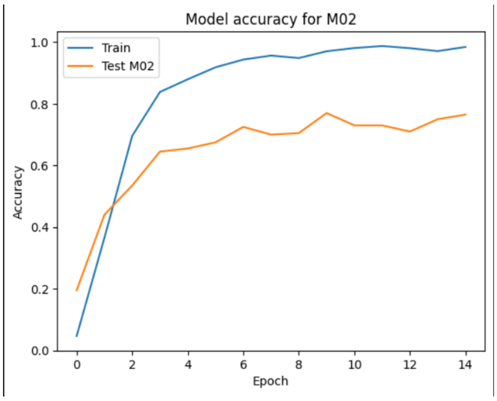


Figure 4.13: CNN Accuracy M02

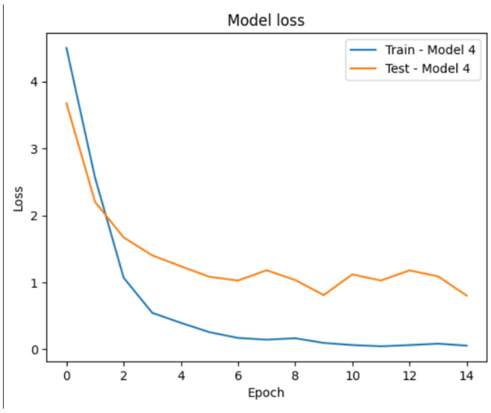


Figure 4.14: CNN Loss M02

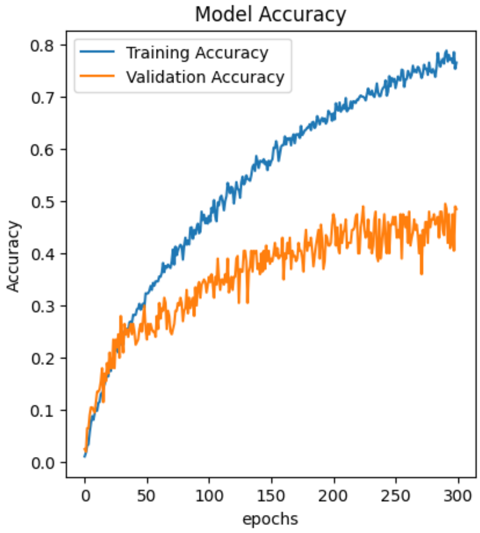


Figure 4.15: RawNet Accuracy M02

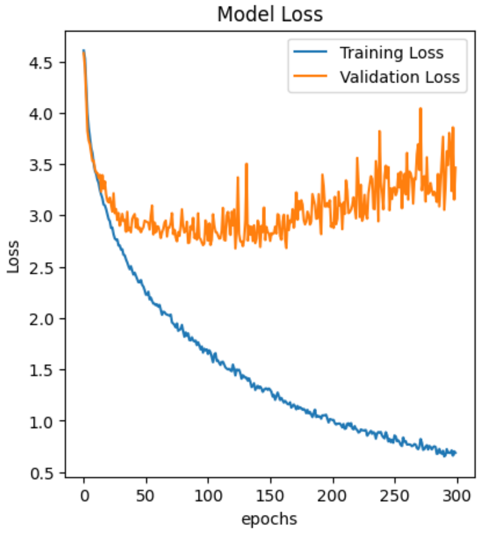


Figure 4.16: RawNet Loss M02

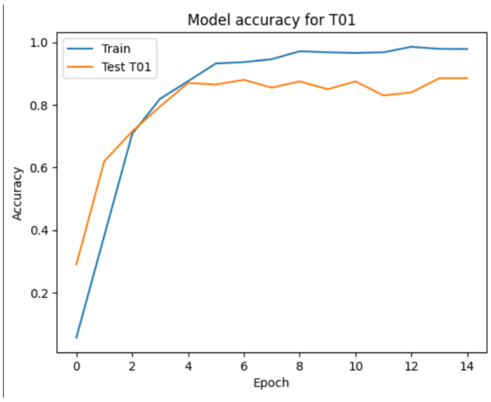


Figure 4.17: CNN Accuracy T01

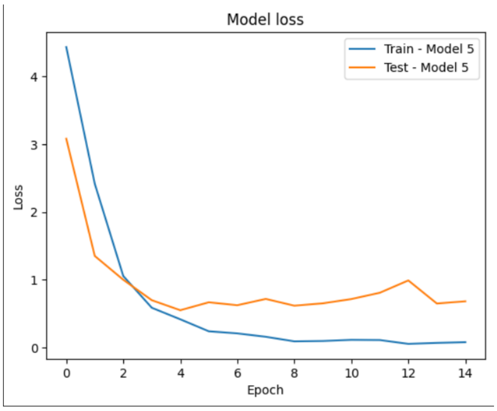


Figure 4.18: CNN Loss T01

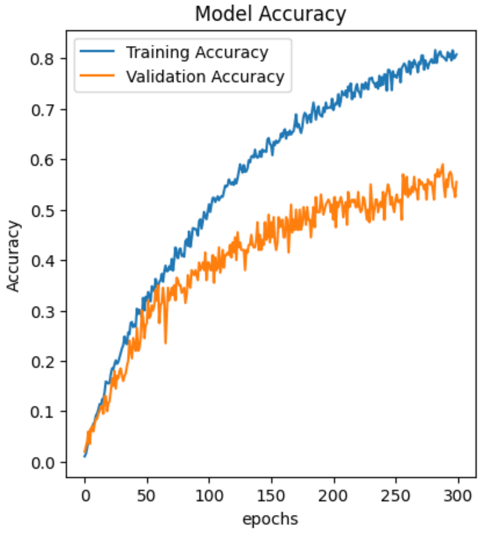


Figure 4.19: RawNet Accuracy T01

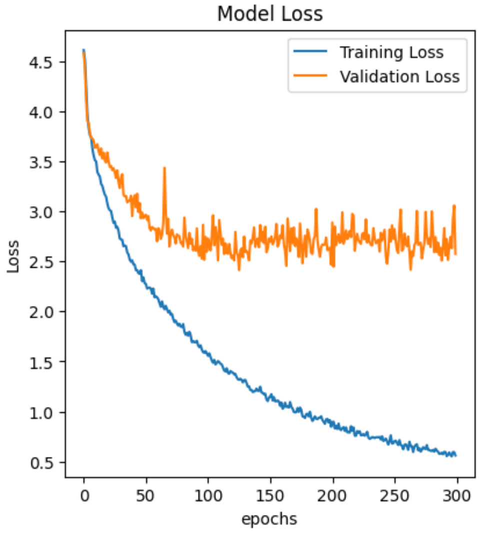


Figure 4.20: RawNet Loss T01

Chapter 5

Conclusions and future scope

The recognition of speakers in many different languages utilizing a combination of Convolutional Neural Networks (CNNs), and RawNet has been an important topic of research in the field of signal processing and machine learning.

5.1 Conclusion

- Implementing speaker recognition across many languages presents a number of obstacles, including differences in pronounciation, accent, and linguistic traits. Combining CNNs, and specialised architecture such as RawNet indicates an effort to capture and comprehend the varied properties associated with many languages.
- RawNet is well-known for it's raw audio data processing of spatial and temporal patterns with audio recordings feature, notably for speech-related applications. The incorparation of RawNet into a speaker recognition system aids in capturing essential auditory information, making the model more resiliant across multiple languages.
- CNN allows for a thorough analysis of both spatial and temporal patterns with audio recordings. This Neural Network topology improves the model's ability to generalise across a wide range of linguistic traits and settings, and providing more accuracy.
- The success of such a system would most likely be measured using measures such as accuracy, precision, recall, and F1 score. It is crucial to evaluate the model's performance not just in terms of individual languages, but also in terms of it's capacity to handle the complexity imposed by the multilingual component.

5.2 Future Scope

RawNet's strength in processing raw audio waveforms presents a unique path for speaker recognition. Future research aims to enhance it's robustness against real-world noise and explore it's potential in speaker diarization, which involves identifying and seperating speech from multiple speakers within a recording. Additionally, combining RawNet with existing techniques could leverage their complementary strengths, paving the way for it's potential as a valuable tool in complex speaker recognition in standard English or even in multi-lingual scenarios.

Though well established in speaker recognition, CNN continues to eveolve. Future advancements involve exploring novel architectures for improving performance and efficiency. Integrating attention mechanisms allows the model to prioritize specific audio regions, leading to more robust feature extraction. Additionally, leveraging pre-trained models on large datasets promises faster training and improved resource efficiency, particularly in resource-limited scenarios. These advancements position CNNs to remain a vital force in speaker recognition, continuously adapting to overcome challenges and push the boundaries of accuracy and efficiency.

Acknowledgment

Without mentioning the names of the people who kindly assisted in the process of fulfilling the project's aims and with their persistent directives and support, brought about it's fulfillment, these come with it's effective completion that would have been incomplete. We express our sincere appreciation and gratitude to the School of Electronics and Communications and the Center for Artificial Intelligence Research (CAIR) Lab. Dr. Suneeta V Budihal, the head of the Department, and Prof. Nirmala and Assistant Prof. Satish Chikkamath for their guidance and encouragement for the completion of this project. We am grateful to the esteemed institution KLE Technological University, Hubballi which has given this opprtunity to realize the most cherished desire to achieve this goal. Finally, would like to express gratitude to everyonr who assisted, directly or indirectly, in fulfilling the project's goals and objectives.

Bibliography

[1] Junjee Kim, Chunghyun Park, and Junho Song. RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. arXiv preprint arXiv:1904.08104, 2019.

[2] M. Jakubec, E. Lieskovska and R. Jarina, "Speaker Recognition with ResNet and VGG Networks," 2021 31st International Conference Radioelektronika (RADIOELEKTRONIKA), Brno, Czech Republic, 2021, pp. 1-5, doi: 10.1109/RADIOELEKTRONIKA52220.2021.9420202.

[3] R. Jagiasi, S. Ghosalkar, P. Kulal and A. Bharambe, "CNN based speaker recognition in language and text-independent small scale system," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 2019, pp. 176-179, doi: 10.1109/I-SMAC47947.2019.9032667.

[4] R. Auckenthaler and J. S. Mason, "Analytical and iterative approaches to the equalisation of sub-band errors in speech and speaker recognition," 9th European Signal Processing Conference (EUSIPCO 1998), Rhodes, Greece, 1998, pp. 1-4.

[5] S. Shon, H. Tang and J. Glass, "Frame-Level Speaker Embeddings for Text-Independent Speaker Recognition and Analysis of End-to-End Model," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 1007-1013, doi: 10.1109/SLT.2018.8639622.

[6] R. Chakroun, L. Beltaïfa Zouari, M. Frikha and A. Ben Hamida, "Improving text-independent speaker recognition with GMM," 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, 2016, pp. 693-696, doi: 10.1109/ATSIP.2016.7523169.

[7] M. Hamidi, H. Satori, N. Laaidi and K. Satori, "Conception of Speaker Recognition Methods: A Review," 2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Meknes, Morocco, 2020, pp. 1-6, doi: 10.1109/IRASET48871.2020.9092118.

[8] S. H. Koya, I. Shahin, Y. Iraqi, E. Damiani and N. Werghi, "EA-VGG: A new approach for emotional speech classification," 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 2022, pp. 1-5, doi: 10.1109/ICECCME55909.2022.9988334.

[9] Z. Gan, Z. Qu, J. Li and Y. Yu, "A method of noisy Tibetan speakers verification based on SKA-TDNN," 2023 2nd Asia Conference on Electrical, Power and Computer Engineering (EPCE), Xiamen, China, 2023, pp. 1-6, doi: 10.1109/EPCE58798.2023.00009.

[10] N. N. An, N. Q. Thanh and Y. Liu, ”Deep CNNs With Self-Attention for Speaker Identification,” in IEEE Access, vol. 7, pp. 85327-85337, 2019, doi: 10.1109/ACCESS.2019.2917470.

[11] L. Jiahong, B. Jie, C. Yingshuang and L. Chun, ”An Adaptive ResNet Based Speaker Recognition in Radio Communication,” 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021, pp. 161-164, doi: 10.1109/ICESIT53460.2021.9696720.

[12] Z. Gan, J. Li, Z. Qu and Y. Yu, ”An improved Tibetan speaker recognition method based on ResNet,” 2023 2nd Asia Conference on Electrical, Power and Computer Engineering (EPCE), Xiamen, China, 2023, pp. 35-39, doi: 10.1109/EPCE58798.2023.00015.

[13] Z. Shao, ”Combining I-vector and ResNet by Knowledge Distillation for Text-Independent Speaker Verification,” 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 2021, pp. 802-806, doi: 10.1109/ICBAIE52039.2021.9390059.

[14] M. Ravanelli and Y. Bengio, ”Speaker Recognition from Raw Waveform with SincNet,” 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 1021-1028, doi: 10.1109/SLT.2018.8639585.