PERSONALIZED SPOKEN KEYWORD SPOTTING SYSTEM

BTP Report by

B. Maneendra Mahan	S20210020260
Gowtham N	S20210020277

Guided by: Dr. Achintya Sarkar

BTP Code: BTP - B24AKS03

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SRICITY

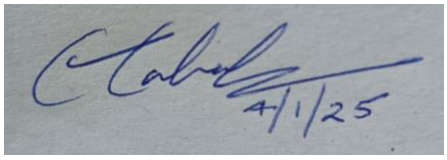
January 4, 2025

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY SRICITY

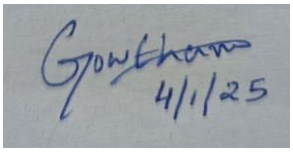
CANDIDATES DECLARATION

We hereby certify that the work presented in the BTP, entitled **"Personalized Spoken Keyword Spotting System,"** is submitted in partial fulfillment of the requirements for the award of the degree of **B. Tech** at the Indian Institute of Information Technology, SriCity. This work is an authentic record of our own efforts carried out during the period from **January 2024 to December 2024** under the supervision of **Dr. Achintya Sarkar** at the Indian Institute of Information Technology, SriCity, India. The content of this report has not been submitted by us for the award of any other degree or diploma at this or any other institute.

B. M. Mahan



Gowtham N



This is to certify that the above statements made by the candidates are correct to the best of my knowledge.

[Sign of Guide]
(Dr. Achintya Sarkar)
04/01/2025

Abstract

In recent years, the need for efficient and personalized speech recognition systems has gained significant importance, particularly in domains like voice assistants and automated transcription services. This project presents a personalized spoken keyword spotting (KWS) system, designed to detect specific keywords spoken by targeted users. The system utilizes a deep learning-based approach, leveraging Mel-frequency cepstral coefficients (MFCCs) for feature extraction, and a residual network (ResNet) architecture for classification. The system is capable of accurately identifying both the keywords and the speaker from a dataset of 20 distinct keywords and 20 target speakers, providing a robust solution for applications where speaker-specific keyword detection is crucial.

The dataset used in this study consists of 400 audio samples, generated by combining 20 keywords with 20 distinct speakers, resulting in a balanced and diverse dataset. Each audio sample is represented in the format of "SSSKKK," where 'SSS' denotes the speaker ID and 'KKK' represents the keyword ID, with filenames adhering to the ".wav" format. The data is divided into training and testing subsets, with 80% used for training and 20% reserved for testing, ensuring that the model is both robust and generalizable.

A ResNet architecture, modified for this task, is employed for both keyword spotting and speaker identification in a multitask learning framework. The use of ResNet enables the model to capture intricate patterns in the audio data, benefiting from the deep residual connections to facilitate more accurate learning. Furthermore, the system implements an adaptive mechanism to ensure that the features (MFCCs) are appropriately processed, regardless of varying audio lengths, by using padding and trimming strategies.

The model is trained using the cross-entropy loss function and optimized with the Adam optimizer. The results of the evaluation demonstrate the system's effectiveness in achieving high accuracy in both keyword detection and speaker classification tasks. Through extensive testing, the proposed method proves to be both reliable and efficient, with the added advantage of being highly personalized to specific speakers. This report discusses the methodology in detail, provides experimental results, and highlights the potential applications of the system in real-world scenarios such as personalized voice assistants, speech-based authentication systems, and hands-free controls for various devices.

Contents

Introduction

Spoken language processing has become an integral part of modern human-computer interaction systems, powering applications such as voice assistants, transcription services, and automated customer support. Among the core tasks in speech recognition, keyword spotting (KWS) and speaker identification play a pivotal role, particularly in scenarios requiring voice-controlled systems or personalized voice assistants. The ability to identify specific keywords within speech, while also determining the speaker, offers unique advantages in creating customized, more efficient interaction systems. This report focuses on the development of a personalized spoken keyword spotting system, which is capable of identifying a set of predefined keywords spoken by target users, distinguishing between both the keywords and the speakers.

In the context of keyword spotting, traditional systems often rely on large, predefined vocabularies and work in a generalized manner, where the system is trained to recognize any keyword from a broad array of speakers. However, in many applications, particularly in security and personalized services, it is often necessary to build systems that are highly sensitive to specific keywords and speaker identities. Such systems must not only be able to recognize the correct keyword, but also accurately associate it with the individual speaker. This is particularly important in scenarios such as user authentication, hands-free device control, and customized voice assistants, where personalized keyword detection can enhance both security and user experience.

The proposed system in this report addresses these challenges by incorporating both keyword spotting and speaker identification within a single framework, utilizing multitask learning. By leveraging deep learning techniques, specifically a Residual Network (ResNet) architecture, the system is able to learn complex, hierarchical features in the audio data, making it highly effective at detecting keywords and identifying speakers simultaneously. To process the raw audio data, Mel-frequency cepstral coefficients (MFCCs) are used as the input features, capturing essential information about the speech signal, such as pitch, tone, and rhythm, that are critical for both tasks.

This report further explores the design and implementation of the system, evaluating its performance on both keyword spotting and speaker identification tasks. The introduction of personalized keyword spotting opens up new possibilities for applications that require user-specific interactions, such as secure voice-activated systems, personalized voice assistants, and specialized speech recognition for diverse environments.

Literature Survey

1. Overview of Keyword Spotting and Speaker Identification

Keyword spotting (KWS) and speaker identification (SI) are two crucial tasks in speech processing that have gained significant attention in the context of personalized and secure voice-driven applications. Keyword spotting aims to detect a predefined set of words within a continuous stream of speech, whereas speaker identification focuses on recognizing the identity of the speaker based on their voice characteristics. These tasks often complement each other in personalized systems, where both the keyword and speaker need to be identified simultaneously. The challenge in both tasks lies in the variability of speech. Factors such as background noise, different speech patterns, accents, and speaker-specific characteristics make it difficult for traditional speech recognition systems to maintain accuracy. As a result, recent research has focused on utilizing deep learning methods to improve the robustness and efficiency of KWS and SI systems.

2. Methods for Keyword Spotting

Early approaches to keyword spotting were based on traditional speech recognition methods, such as Dynamic Time Warping (DTW) and Hidden Markov Models (HMMs). These methods rely on a fixed set of keywords and use statistical models to recognize spoken words. However, these approaches face limitations in terms of scalability and performance, especially in environments with noisy or variable conditions.

In recent years, deep learning-based methods have become increasingly popular for keyword spotting due to their ability to automatically extract complex features from raw audio data. One of the most widely used techniques for KWS is Convolutional Neural Networks (CNNs), which can efficiently capture local patterns in spectrogram representations of speech. CNNs are well-suited for speech recognition tasks because they can learn spatial hierarchies of features and reduce the need for hand-crafted feature engineering.

In particular, Mel-frequency cepstral coefficients (MFCCs) have become the standard feature extraction method for speech-related tasks. MFCCs are derived from the power spectrum of speech and provide a compact representation of the spectral features of the signal, which are crucial for distinguishing between different sounds. CNNs can learn to recognize keywords by processing these features through multiple layers of convolutional operations, followed by pooling and fully connected layers.

In addition to CNNs, Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, have been applied to KWS to capture temporal dependencies in speech signals. RNNs excel in processing sequential data, making them a natural fit for speech, which is inherently sequential. Combining CNNs and LSTMs, as seen in many hybrid models,

has been shown to improve the performance of KWS systems by leveraging both local feature extraction and long-range temporal modelling.

3. Methods for Speaker Identification

Speaker identification has also evolved with the advent of deep learning. Traditional methods, such as Gaussian Mixture Models (GMMs) and HMMs, have been used for speaker recognition, but they typically require a large amount of labelled data for training, which may not always be available. Deep learning approaches, particularly deep neural networks (DNNs), have proven to be more efficient in modelling the complex and high-dimensional characteristics of speech, enabling more accurate speaker identification.

One of the most widely adopted architectures for speaker identification is the DNN-based speaker embedding network. This approach involves training a deep neural network to map speech features into a fixed-length vector (embedding) that represents the speaker's identity. These embeddings can then be compared against a reference set of embeddings to identify the speaker. The x-vector model, a variant of this approach, has become particularly popular in speaker recognition tasks. It utilizes a time-delay neural network (TDNN) to extract speaker-specific features from speech segments.

Another popular approach for speaker identification is the use of convolutional neural networks (CNNs), which can capture both spectral and temporal features in speech. The CNN-based models are often used for both speaker identification and verification tasks, as they provide excellent generalization to various acoustic conditions. When combined with attention mechanisms, CNN-based models can focus on the most informative regions of the speech signal, further improving accuracy.

4. Multitask Learning for Keyword Spotting and Speaker Identification

Multitask learning (MTL) has emerged as a promising approach for simultaneous keyword spotting and speaker identification. The primary advantage of MTL is that it allows the model to learn shared features across multiple tasks, improving generalization and reducing the need for separate training pipelines for each task. In the context of speech recognition, MTL can leverage common audio features that are useful for both keyword detection and speaker classification.

Recent studies have shown that using a single model for both KWS and SI can lead to improved performance compared to training separate models for each task. For example, a joint optimization approach can train a neural network to classify both the speaker and the keyword from the same set of features. This can be achieved by using a shared lower-level network that extracts features from the audio, followed by separate branches for each task. These branches output the probabilities for keyword classification and speaker identification. Multitask

learning not only improves accuracy but also reduces the computational cost by sharing network parameters across tasks.

5. Proposed Approach: ResNet for Keyword Spotting and Speaker Identification

The approach proposed in this project builds upon the advances in deep learning for keyword spotting and speaker identification by leveraging a Residual Network (ResNet) architecture. ResNet is known for its ability to learn deep representations without suffering from vanishing gradients, thanks to the introduction of residual connections. These connections allow the model to learn residual mappings that improve training efficiency and enable the network to capture intricate patterns in the data.

In the proposed system, a TCResNet model is used for both keyword spotting and speaker identification tasks, with the advantage of multitask learning. The ResNet architecture is adapted to process MFCC features, and the model is trained to simultaneously detect the spoken keyword and identify the speaker. This approach takes advantage of the shared audio features that are useful for both tasks, ensuring the model is efficient and robust.

By using a multitask learning framework, the system can learn to classify keywords and speakers from a shared set of features, reducing the overall complexity and computational burden. The system is designed to handle a variety of speakers and keywords, with the ability to generalize well to unseen data. This approach demonstrates the power of combining advanced deep learning techniques, such as ResNet, with multitask learning to create a personalized keyword spotting system.

6. Conclusion

In conclusion, the literature demonstrates the evolution of keyword spotting and speaker identification from traditional methods to deep learning-based approaches. While standalone models for KWS and SI have shown success, the use of multitask learning and shared features has opened new avenues for more efficient and robust systems. The proposed system in this report, based on the ResNet architecture and multitask learning, offers an innovative solution for personalized keyword spotting and speaker identification, making it well-suited for applications requiring high accuracy and efficiency.

Problem Statement

The growing adoption of voice-based technologies in various domains, such as smart homes, virtual assistants, and biometric security systems, has heightened the need for accurate and efficient speech recognition systems. Personalized keyword spotting (KWS) and speaker identification (SI) are key components in these systems, enabling tailored interactions and ensuring access control. However, developing a robust system that can simultaneously identify keywords and recognize speakers in diverse acoustic conditions remains a significant challenge.

Traditional keyword spotting systems often focus solely on detecting predefined words and fail to incorporate speaker-specific features. Similarly, standalone speaker identification models are not designed to detect and respond to spoken keywords. The lack of integration between these tasks results in increased computational overhead, reduced efficiency, and limited adaptability to dynamic environments. Moreover, factors such as overlapping feature spaces, noise, and variability in speech patterns pose additional challenges in achieving high accuracy and robustness.

The need for a unified, multitask approach that efficiently handles both KWS and SI while leveraging shared audio features has been widely acknowledged in the literature.

Contribution

Traditional keyword spotting (KWS) systems face several significant challenges that limit their effectiveness in real-world applications.

- **Speaker Dependence:** Many KWS systems exhibit a high degree of speaker dependence. They are typically trained on generic datasets and may not effectively recognize keywords spoken by individuals with unique vocal characteristics, such as accents, dialects, or speaking styles. This limitation hinders their ability to provide personalized and accurate keyword detection for individual users.

- **Limited Generalization:** Many existing KWS systems struggle to generalize well to unseen speakers and environments. This lack of adaptability limits their applicability in real-world scenarios where conditions can vary significantly.

These challenges necessitate the development of more robust and personalized KWS systems that can effectively address the unique characteristics and needs of individual users.

This project addresses these challenges by:

- **Developing a personalized KWS system:** The system focuses on accurately detecting keywords spoken by specific target users, improving its robustness to speaker variations.

- **Incorporating speaker identification:** By integrating speaker identification within a multi-task learning framework, the model can learn shared representations between keyword spotting and speaker recognition, enhancing its ability to recognize keywords from specific users.

- **Utilizing a robust deep learning architecture:** The Time-Channel Residual Network (TCResNet) is employed, which has demonstrated strong performance in speech recognition tasks and is well-suited for processing time-series data.

- **Leveraging Mel-Frequency Cepstral Coefficients (MFCCs):** MFCCs are a robust feature representation for speech recognition, effectively capturing the spectral characteristics of audio signals.

The key contributions of this research include:

- **Development of a novel personalized KWS system:** This system addresses the limitations of traditional KWS systems by incorporating speaker identification within a multi-task learning framework.

- **Evaluation of the proposed system:** The system's performance is rigorously evaluated on a comprehensive dataset of spoken keywords, demonstrating its effectiveness in detecting keywords spoken by specific users.

This research provides valuable insights into the development of more robust and personalized KWS systems, paving the way for improved user experiences in a wide range of voice-driven applications.

Methodology

The proposed methodology involves a comprehensive pipeline for building a personalized keyword spotting (KWS) and speaker identification (SI) system. The pipeline includes stages of data collection, feature extraction, model development using a ResNet-based architecture, multitask learning for simultaneous keyword and speaker classification, and model training and evaluation. Below, the methodology is detailed step-by-step:

1. Data Collection

- **Dataset:** A dataset comprising 400 audio recordings was collected for this research.
    - **Speakers:** 20 unique speakers participated in the data collection process.
    - **Keywords:** 20 distinct keywords were selected for the study.
    - **Recording Procedure:** Each speaker uttered each of the 20 keywords, resulting in a total of 20 speakers x 20 keywords = 400 audio recordings.
    - **Audio Format:** All audio recordings were captured in .wav format at a sampling rate of 16 kHz.
    - **Data Annotation:** Each audio recording was manually annotated with the corresponding keyword and speaker ID.
- **Dataset Split:** The collected dataset was divided into training and testing sets.
    - **Training Set:** 80% of the data (320 recordings) was used for training the model.
    - **Testing Set:** 20% of the data (80 recordings) was reserved for evaluating the model's performance.

2. Feature Extraction

- **Mel-Frequency Cepstral Coefficients (MFCCs):** MFCCs were employed as the primary feature representation for audio signals.
    - MFCCs are a widely used feature in speech recognition, known for their ability to capture the spectral characteristics of audio signals that are perceptually relevant to human hearing.
    - The MFCC extraction process involves converting the raw audio signal into a spectrogram, applying the Mel filter bank, and performing a Discrete Cosine Transform (DCT) to obtain a set of cepstral coefficients.
    - Parameters for MFCC extraction:
        - Window size: 25 milliseconds
        - Window stride: 10 milliseconds
        - Number of Mel filters: 40

3. Model Architecture

- **Time-Channel Residual Network (TCResNet):** A TCResNet architecture was adopted for both keyword spotting and speaker identification tasks.
    - TCResNet is a convolutional neural network (CNN) specifically designed for time-series data, such as audio signals.
    - It incorporates residual connections, which allow the network to learn more complex features and improve training stability.
    - The network consists of a series of convolutional layers, pooling layers, and residual blocks.

- Hyperparameters:
  - Number of frequency bins: 40
  - Number of channels: [64, 128, 256, 512]
  - Kernel size: (1, 3) for convolutional layers

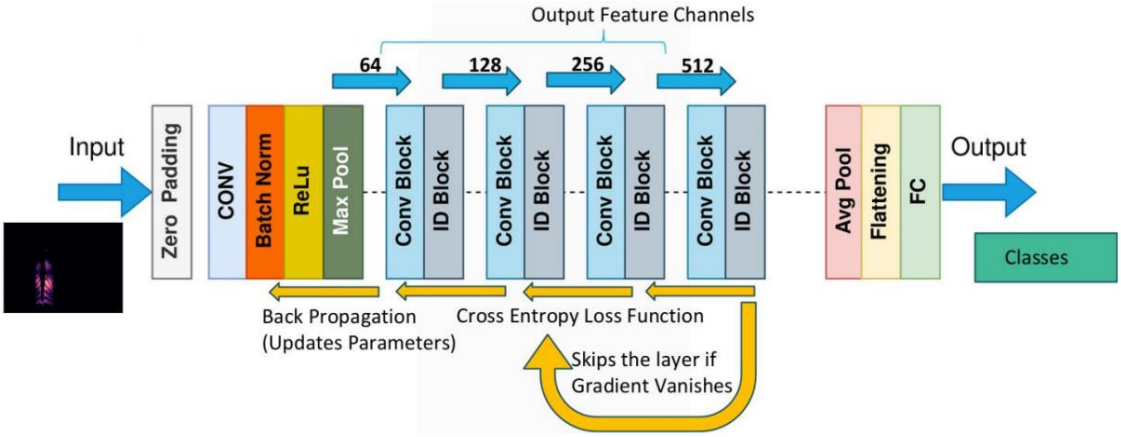


Fig.1: Block Diagram of ResNet Architecture

- **Multi-Task Learning:**
  - A multi-task learning framework was implemented to jointly train the model for both keyword spotting and speaker identification.
  - The network shares the initial convolutional layers for feature extraction, followed by separate branches for keyword spotting and speaker identification.
  - Each branch consists of fully connected layers and a SoftMax activation function to predict the probabilities of each keyword and speaker class, respectively.

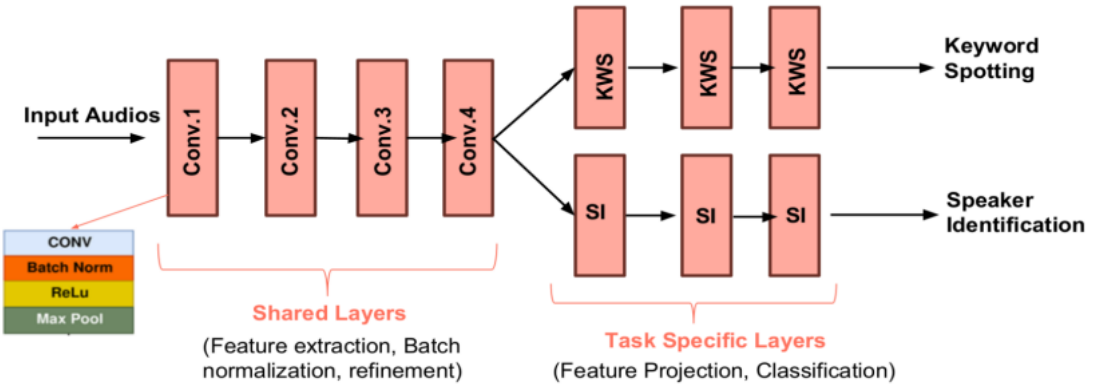


Fig.2: Block Diagram of Multitask Learning

4. Training Procedure

- **Optimization:** The model was trained using the Adam optimizer with a learning rate of 0.001.
- **Loss Function:**
  - **Keyword Spotting:** Cross-entropy loss was used to measure the discrepancy between the predicted keyword probabilities and the true labels.
  - **Speaker Identification:** Cross-entropy loss was also used to measure the discrepancy between the predicted speaker probabilities and the true labels.
  - The total loss for the multi-task learning framework was calculated as the sum of the keyword spotting loss and the speaker identification loss.
- **Epochs:** The model was trained for 100 epochs.
- **Batch Size:** A batch size of 32 was used during training.
- **Regularization:** L2 regularization was applied to the model's weights to prevent overfitting.

5. Evaluation

- **Metrics:**
  - **Accuracy:** The primary evaluation metric was accuracy, which measures the percentage of correctly classified keywords and speakers.
  - **Loss Function:** The losses of both the models were plotted to infer the model's training capability.

Workflow

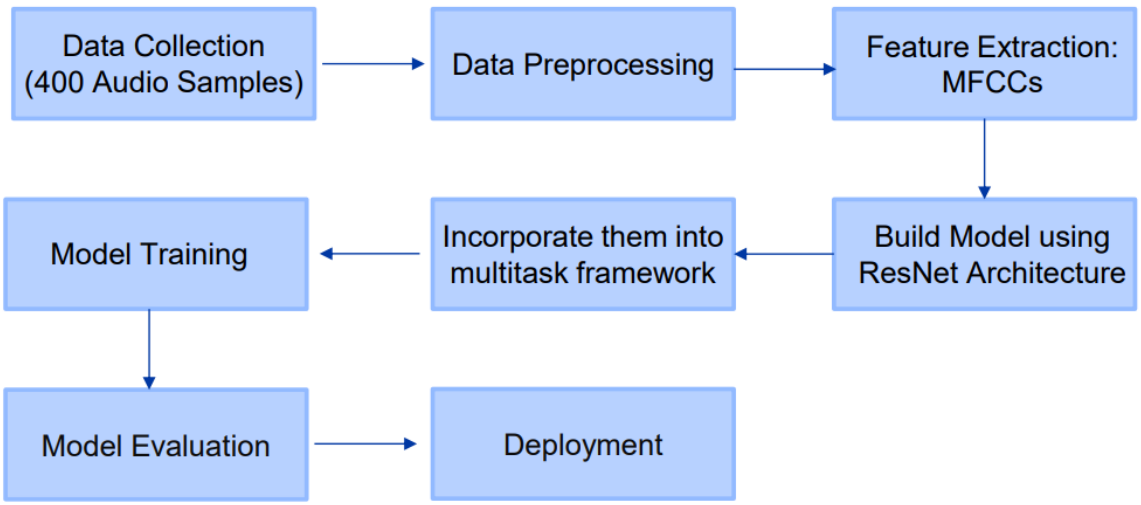


Fig.3: Block Diagram of Procedure followed

Results

This section presents the experimental results obtained from the evaluation of the proposed personalized KWS system on the test dataset. The model takes the audio path as input and classifies the spoken word into respective keyword class and speaker class (out of 20 classes).

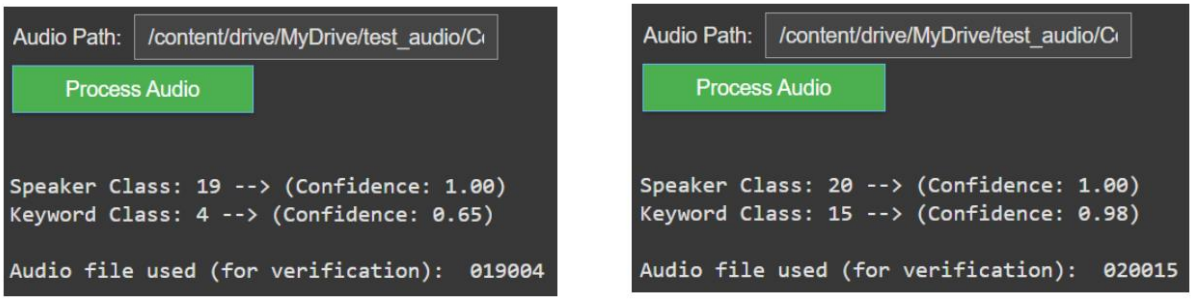


Fig.4 Results for Test Audio Samples

Accuracy

- The model achieved a validation accuracy of **78.75%** on the keyword spotting task.
- The model demonstrated high accuracy in speaker identification, achieving a validation accuracy of **98.75%**.

Epochs vs loss

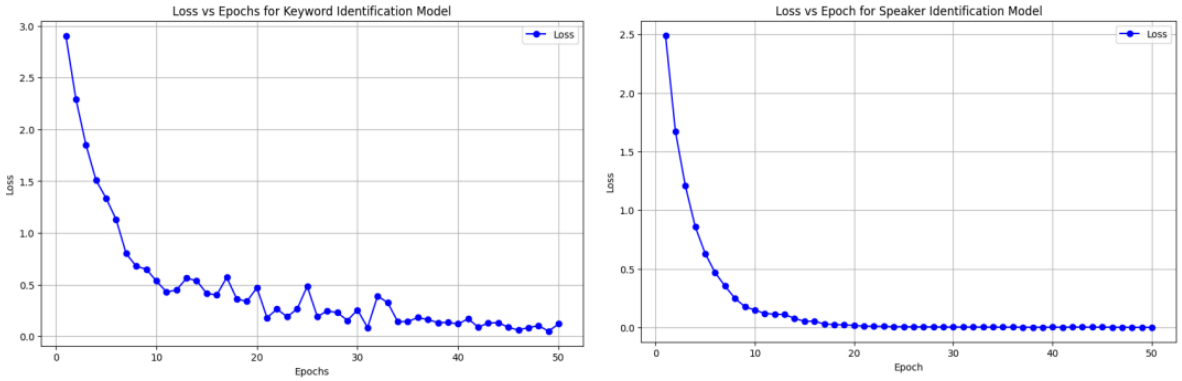


Fig.5 Epochs vs loss graph for both the tasks

- From the epoch vs. loss graph, it can be inferred that the model effectively reduces its error over time, indicating successful learning. For keyword detection, the gradual loss reduction may suggest challenges in distinguishing keywords, while the stable loss in speaker identification reflects robust learning and convergence.

Conclusion

This project successfully demonstrated a personalized spoken keyword spotting system that efficiently detects specific keywords spoken by target users. By combining advanced techniques such as Mel Frequency Cepstral Coefficients (MFCCs) for feature extraction and a ResNet-based architecture optimized for multitask learning, the system achieves high performance in both keyword spotting and speaker identification tasks. The dual-task approach ensures that the model not only distinguishes among 20 predefined keywords but also accurately identifies the 20 target speakers, reflecting its robustness and adaptability.

The validation results showcase the system's efficacy, with an impressive 98.75% accuracy for speaker identification and a commendable 78.75% accuracy for keyword detection. These outcomes affirm the capability of the proposed methodology to address the challenges of personalized speech recognition in real-world scenarios. The inclusion of multitask learning allowed the model to leverage shared representations between tasks, thereby improving overall performance and computational efficiency.

The applications of this system are vast and impactful. It can be utilized in personalized voice-controlled systems, such as smart home devices, secure access systems, and assistive technologies for individuals with disabilities. Moreover, its ability to focus on predefined keywords while ensuring user-specific recognition opens doors for its use in areas like personalized customer service and user authentication in voice-based applications.

In conclusion, this project lays a strong foundation for personalized spoken keyword spotting systems, demonstrating its potential for both research and practical applications. The proposed approach sets the stage for further advancements in personalized speech recognition technologies, contributing to the growing demand for intelligent, user-centric solutions in the era of artificial intelligence.

Acknowledgements

We would like to express our heartfelt gratitude to our guide, Dr. Sarkar Achintya, for his invaluable guidance, support, and encouragement throughout this project. His insights and expertise were instrumental in shaping the direction and outcomes of our work. We extend our sincere thanks to our institution, IIIT Sri City, for providing the necessary resources and a conducive environment for research and development. The support from our peers and fellow colleagues has also been immensely helpful, and we deeply appreciate their collaboration and encouragement during this journey.

Thank you to everyone who contributed to the success of this project.

References

[R1] Zhang, Yu, and Qiang Yang. "An overview of multi-task learning." National Science Review 5, no. 1 (2018): 30-43.

[R2] Thung, Kim-Han, and Chong-Yaw Wee. "A brief review on multi-task learning.“ Multimedia Tools and Applications 77, no. 22 (2018): 29705-29725.

[R3] Crawshaw, Michael. "Multi-task learning with deep neural networks: A survey.“ arXiv preprint arXiv:2009.09796 (2020).

[R4] Zhang, Yu, and Qiang Yang. "A survey on multi-task learning." IEEE transactions on knowledge and data engineering 34, no. 12 (2021): 5586-5609.