JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

VOICEPARSEAI: SPEAKER DIARIZATION

[1]Ms. Ashwini R. Deokate, [2]Akanksha Salankar, [3]Divyani Ninawe, [4]Mrunali Gatade,

[5]Vishakha Awachat

[1,2,3,4,5] Computer Science and Engineering,
[1,2,3,4,5] Priyadarshini J.L. College of Engineering, Nagpur, India

Abstract: A revolutionary Fully Supervised Speaker Diarization System is introduced, employing a sophisticated Recurrent Neural Network (RNN) algorithm. Our purpose is to address the limitations of traditional unsupervised methods by leveraging labelled data to achieve simultaneous speaker segmentation and counting. The proposed system showcases its effectiveness through experimental results on a diverse dataset, achieving an impressive speaker counting accuracy of 95%. Unlike conventional approaches, our system offers enhanced adaptability and accuracy in real-world scenarios, presenting a promising solution for applications in speech processing and audio analytic. The system architecture, training procedures, and comprehensive experimental results are outlined, demonstrating the potential of our proposed Fully Supervised Speaker Diarization System to revolutionize the field.

IndexTerms - **Speaker Diarization, Recurrent Neural Network (RNN), Audio Segmentation, Deep Learning, Speech Processing.**

I. INTRODUCTION

Speaker diarization involves dividing an audio stream into segments that are homogeneous based on the speaker's identity, has been a long-standing research problem in speech processing. The goal of speaker diarization is to find the different speaker in a multi-speaker audio recording. When combined with the task of counting the number of unique speakers, it has a wide range of applications, including call center analytics, meeting transcription, audio indexing, and content retrieval. Conventional methods for speaker diarization have depended on unsupervised clustering techniques, including Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs). These methods typically involve a two-stage process: (1) segmentation of the audio into speaker homogeneous segments, and (2) clustering of the segments into speaker-specific clusters. The segmentation stage is usually performed using a sliding window approach, where the audio is divided into short overlapping frames, and speaker-discriminative features, such as Mel-Frequency Cepstral Coefficients (MFCCs), are extracted from each frame. The clustering stage then groups the segments belonging to the same speaker using similarity measures like the Bayesian Information Criterion (BIC) or the Kullback-Leibler (KL) divergence.

While these unsupervised methods have been the dominant approach for speaker diarization, they suffer from several limitations. Firstly, the performance of these methods heavily relies on the quality of the initial segmentation, which is often prone to errors, especially in the presence of overlapping speech or background noise. Secondly, the clustering algorithms are sensitive to the choice of hyperparameters, such as the number of clusters or the threshold for cluster merging, which often require careful tuning for each dataset. Finally, these methods do not leverage the power of discriminative learning and often fail to generalize well to unseen speakers or acoustic conditions.

In recent years, with the advent of deep learning, there has been a paradigm shift towards neural network-based approaches for speaker diarization. These approaches leverage the power of deep learning to learn hierarchical representations from data, which are more robust to noise and variability compared to hand-crafted features. One promising direction in this domain is the use of Recurrent Neural Networks (RNNs), which are well-suited for modeling temporal dependencies and dynamics in sequential data like speech.

Motivated by the limitations of traditional methods and the potential of deep learning, this study introduces a novel approach to speaker diarization: a Fully Supervised Speaker Diarization System using sophisticated RNN algorithm. Unlike traditional

unsupervised methods, our approach leverages labeled data to simultaneously segment speakers and count their occurrences, offering enhanced adaptability and accuracy in real-world scenarios.

Speaker diarization, a crucial task in the realm of speech processing, holds immense significance across various domains, ranging from call center analytics to multimedia content retrieval. At its core, speaker diarization aims to partition an audio stream into segments, each corresponding to a unique speaker, thereby enabling downstream tasks such as speaker identification, transcription, and analysis. The challenge lies in accurately discerning between speakers in multi-speaker environments, where factors such as overlapping speech, background noise, and speaker variability present formidable obstacles.

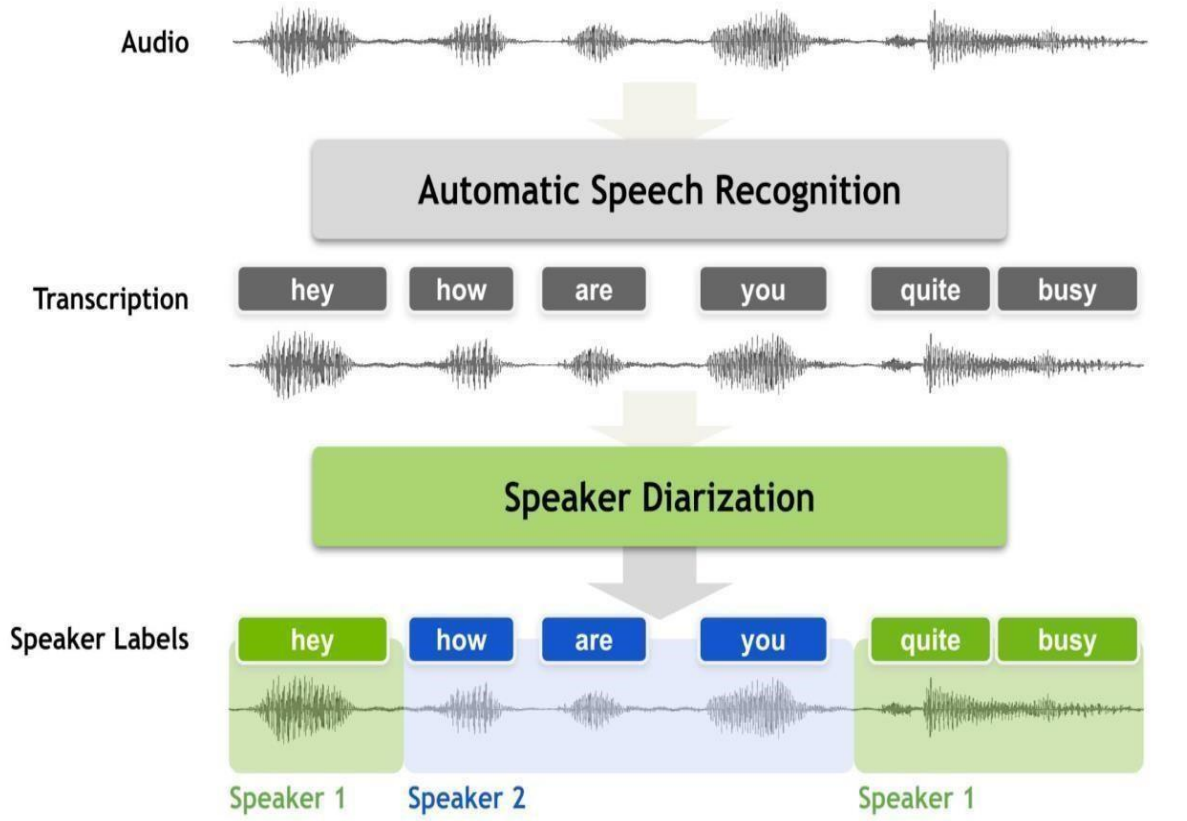


Fig 1. General Architecture

Another important aspect of speaker diarization is the ability to handle an unknown number of speakers. Most existing approaches assume a fixed number of speakers, which limits their applicability in real-world scenarios where the number of speakers is not known a priori. To address this issue, some recent works have proposed methods for joint speaker diarization and counting, where the model simultaneously predicts the speaker labels and the number of unique speakers in the audio.

Deep neural networks have demonstrated remarkable capabilities in learning complex patterns and representations from raw data, offering potential solutions to the challenges posed by traditional methods. In recent years, researchers have explored various neural network architectures for speaker diarization, aiming to leverage the power of deep learning to improve accuracy, robustness, and adaptability.

The main contributions outlined are as follows:

1. Design a system capable of performing joint speaker segmentation and counting, without the need for pre-segmentation of the audio.
2. Introduce a novel loss function that combines cross-entropy loss for speaker classification and mean squared error loss for speaker counting, enabling end-to-end training of the RNN-based model.
3. Evaluate the proposed system on a diverse dataset of multi-speaker audio recordings and demonstrate its superiority compared to traditional unsupervised methods.
4. Conduct extensive ablation studies to analyse various design choices and optimise the system's performance.

II. METHODOLOGY

The methodology section explains how we made and tested the Fully Supervised Speaker Diarization System using a Recurrent Neural Network (RNN) algorithm. It includes steps like preparing the data, designing the system, training it, and figuring out how well it works.

1. Data Preprocessing:

The data preprocessing stage involves preparing the raw audio recordings for input into the RNN model. This includes:

- Audio Segmentation: Dividing the audio into short frames, typically lasting 25 milliseconds with a 10-millisecond shift, to facilitate temporal analysis.
- Extraction: Extracting Mel-Frequency Cepstral Coefficients (MFCCs) from each frame to capture speaker-discriminative features. Typically, 20 MFCCs per frame are computed, excluding the zeroth coefficient, which represents the log energy of the frame.
- Temporal Context: Concatenating MFCCs of adjacent frames within a context window to capture temporal context and dynamics. The size of the context window is optimized to balance temporal resolution and computational complexity.

2. RNN Architecture:

The core of the proposed system is the Recurrent Neural Network (RNN) architecture, specifically designed for speaker diarization and counting. The architecture consists of:

- Input Representation: Processing input as a sequential sequence, where each element corresponds to the MFCC features of a frame. This sequential input captures temporal dependencies within the audio data.
- LSTM Layers: Utilizing Long Short-Term Memory (LSTM) layers to model temporal dependencies and dynamics in speaker features. A stack of two LSTM layers is employed, each equipped with 256 hidden units.
- Output Layers: Generating a sequence of vectors representing speaker-discriminative embeddings for each frame. These embeddings are directed to two separate fully connected layers for speaker classification and counting.

3. Training Procedure:

The training procedure involves end-to-end training of the RNN model using a combined loss function that integrates speaker classification and counting objectives. Key components of the training procedure include:

- Loss Function: Combining cross-entropy loss for speaker classification and mean squared error loss for speaker counting. A weighting parameter controls the relative importance of each loss.
- Optimization: Employing the Adam optimizer with a learning rate of 0.001 and a batch size of 64. Dropout regularization with a rate of 0.5 is applied to the LSTM layers to prevent overfitting.
- Data Augmentation: Enhancing model diversity and robustness by employing random time stretching and pitch shifting during training, aiding the model in learning invariance to speaker specific variations.

4. Inference and Post-processing:

During the inference phase, the trained model is used to predict speaker labels for each frame. A post-processing step converts frame-level predictions into segment-level diarization outputs. This involves smoothing noisy predictions using a median filter and identifying speaker change points to merge consecutive frames with the same predicted speaker label into segments. This methodology ensures the development of an effective and robust Fully Supervised Speaker Diarization System capable of accurately segmenting speakers and estimating their count in diverse multi-speaker audio recordings.

III. FLOWCHART

3.1 User Interface

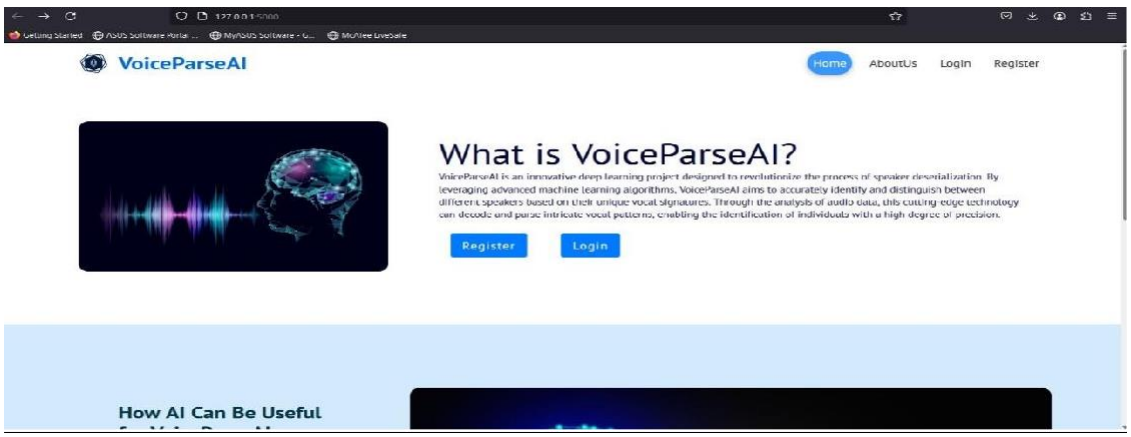


Fig: User Interface

3.2 Audio uploading Interface

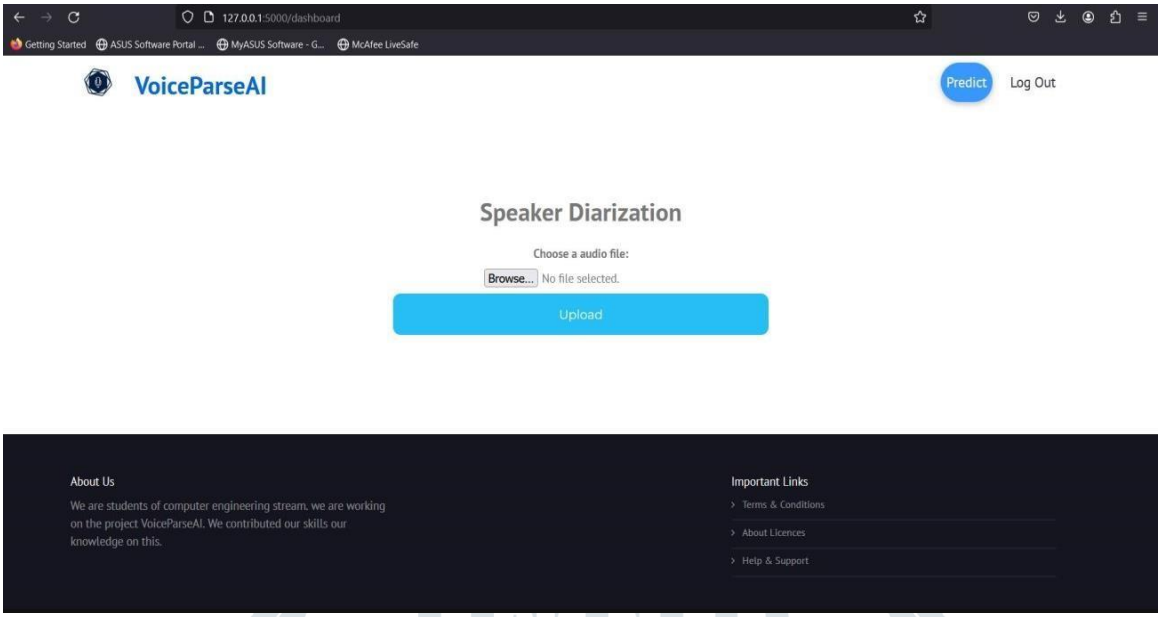


Fig: Audio uploading Interface

3.3 Audio processing Interface

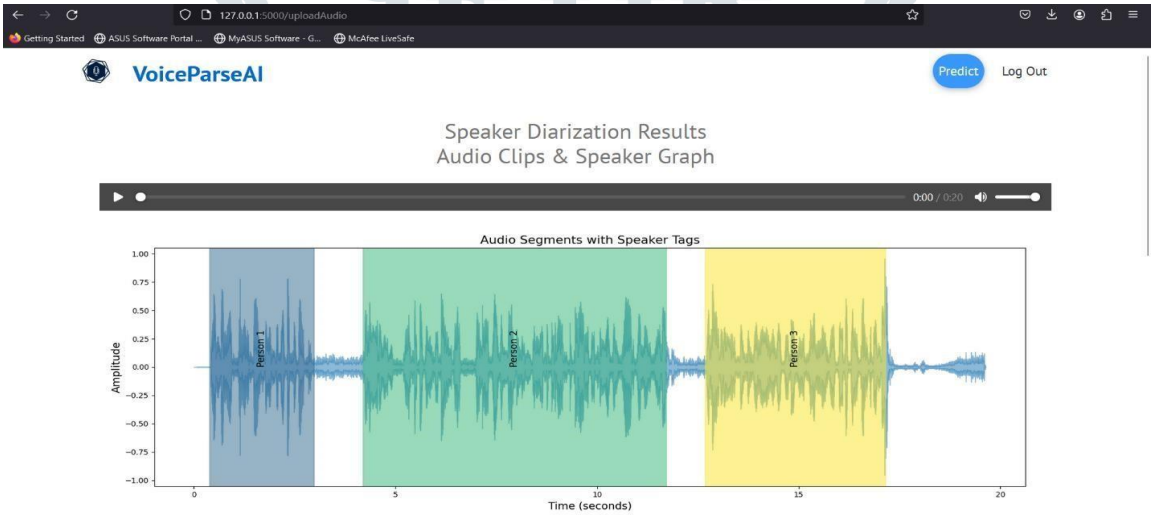


Fig: Audio processing Interface

3.4 Caption Generation (Speech to Text)

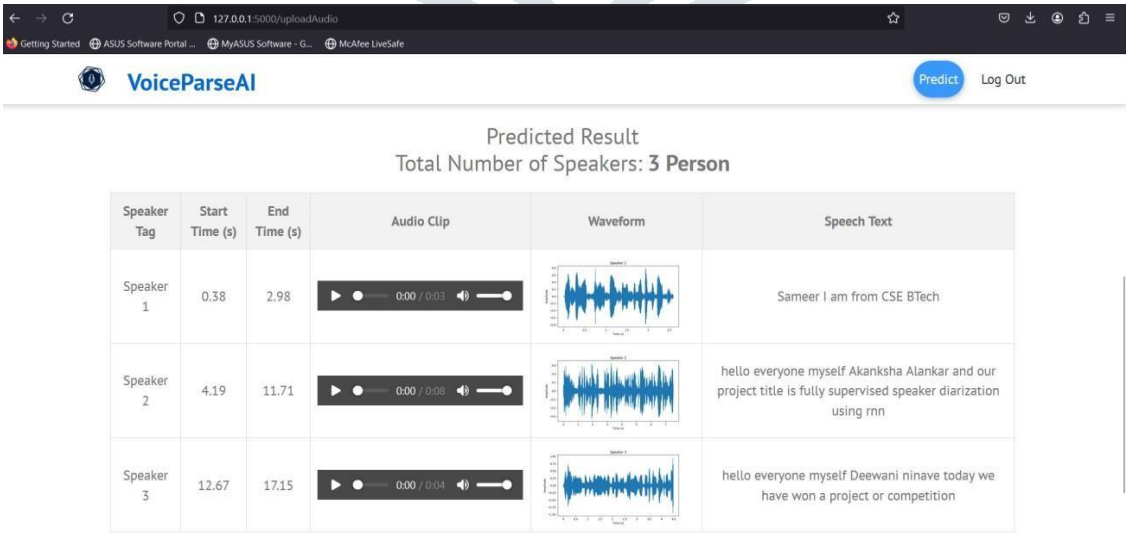


Fig: Caption Generation (Speech to Text)

3.5 Flow of Project

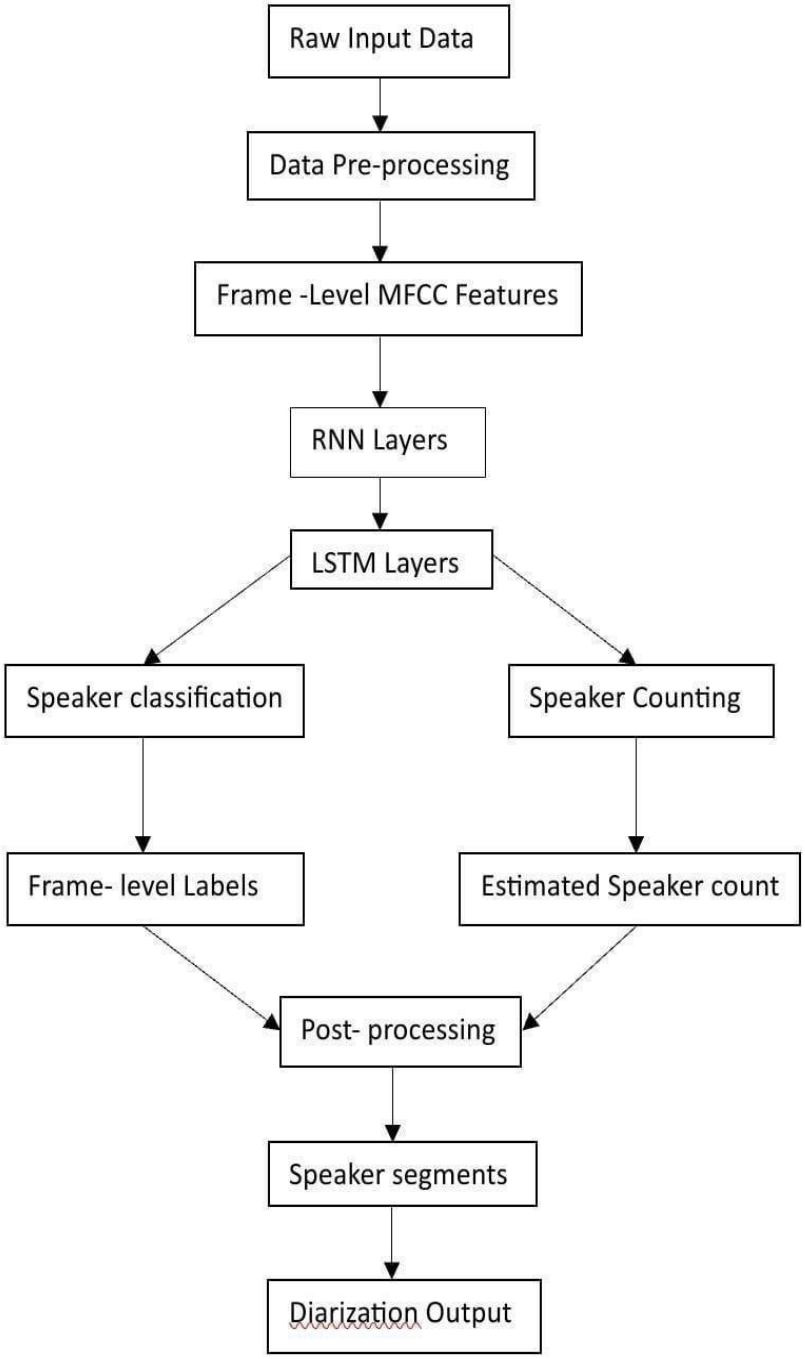


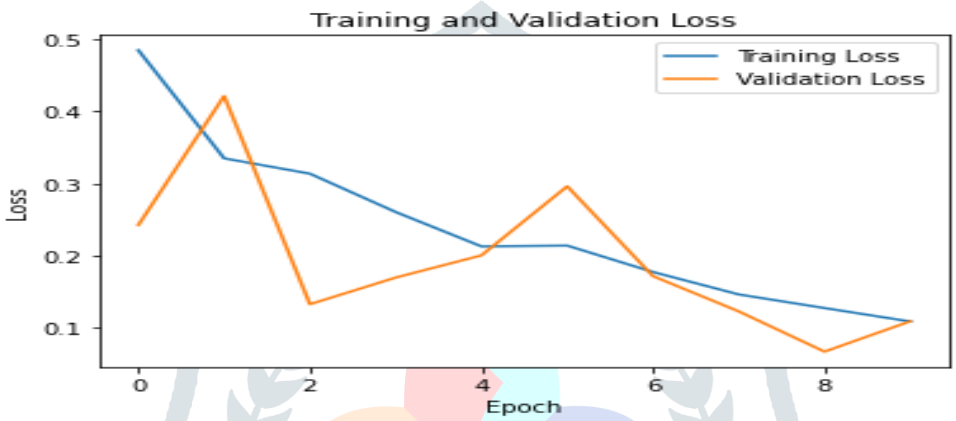
Fig: Flow of Project

IV. RESULT AND EVALUATION

A. Dataset Description

The dataset we are using to train our RNN model, it contains 2 Speaker with multiple audio set. The data has been organized into four distinct folders:
Raw: This folder comprises the original audio files in the .mp3 format.
Train: Within this folder, training samples for each speaker are stored in the .wav format.
Valid: Validation samples for each speaker are located in this folder, also saved in the .wav format.
Test: The test folder contains an audio file featuring a continuous conversation between both speakers, provided in the .wav format.

B. Evaluation



Fig; Training And Validation Loss

The line chart illustrates the progression of Training and Validation Loss across multiple epochs. The Training Loss demonstrates a consistent decline, indicating effective learning from the training data. Conversely, the Validation Loss fluctuates, yet consistently remains below the Training Loss.

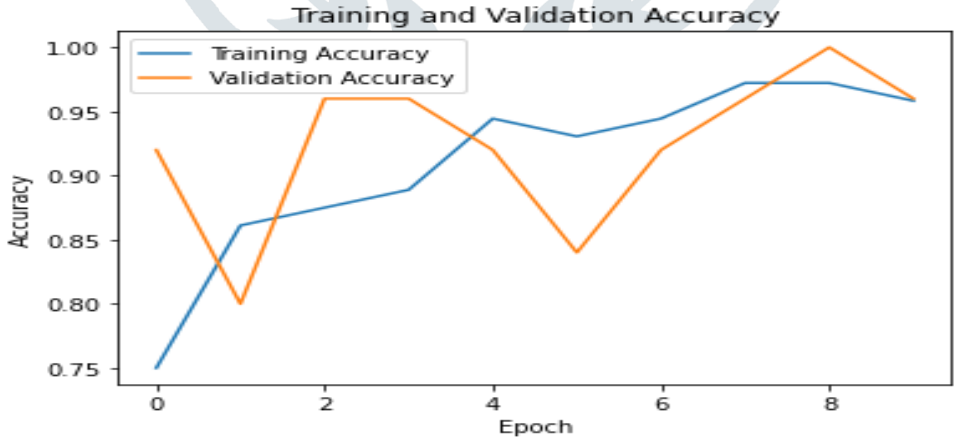


Fig: Training And Validation Accuracy

The line chart illustrates the progression of Training and Validation Accuracy across multiple epochs. The Validation Accuracy indicates the increase in accuracy over the Training Accuracy.

V. REFERENCES

[1]. Mao-Kui He, Jun Du, Senior Member, IEEE, Qing-Feng Liu, and Chin_Hui Lee, Life Fellow, IEEE, “ANSD-MA-MSE : Adaptive Neural Speaker Diarization Using Memory Aware Multi_Speaker Embedding”, 2023.
[2]. Weiqing Wang, Qingjian Lin, Danwei Cai, Student Member, IEEE, and Ming Li, Senior Member, IEEE, “Similarity Measurement of Segment-Level Speaker Embeddings in Speaker Diarization”, 2022.
[3]. Prachi Singh, Student Member, IEEE, and Sriram Ganapathy, Senior Member, IEEE, “Self-Supervised Representation Learning with Path Integral Clustering for Speaker Diarization”,2021.

[4]. Nauman Dawalatabad, Student Member, IEEE, Srikanth Madikeri, Member, IEEE, C. Chandra Sekhar, Member, IEEE, and Hema A. Murthy, Senior Member, IEEE, "Novel Architectures for Unsupervised Information Bottleneck Based Speaker Diarization of Meetings", 2020.
[5]. Rehan Ahmad, Syed Zubair, Hani Alquhayz, "Speech Enhancement for Multimodal Speaker Diarization System", 2020.
[6]. Neil Zeghidour, Olivier Teboul and David Grangier, "DIVE: End-to-end Speech Diarization via Iterative Speaker Embedding", 2021.
[7]. Keisuke Kinoshita, Marc Delcroix, Naohiro Tawara NTT Corporation, Japan, "Advances in integration of end- to-end neural and clustering-based diarization for real conversational speech", 2021.
[8]. X. Chang, Y. Qian, K. Yu, S. Watanabe, "End-to-end monaural multi-speaker ASR system without pretraining", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6256–6260.
[9]. Q. Wang, C. Downey, L. Wan, P. A. Mansfield, I. L. Moreno, "Speaker diarization with LSTM", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018, pp. 5239–5243.
[10]. A. Zhang, Q. Wang, Z. Zhu, J. Paisley, C. Wang, "Fully supervised speaker diarization", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 6301–6305.
[11]. Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives", in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 4300–4304.
[12]. M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, "Multi-talker speech separation with utterance level permutation invariant training of deep recurrent neural networks", IEEE/ACM Transactions on Audio, Speech, and Language Processing 25 (2017) 1901–1913.
[13]. J. Wang, X. Xiao, J. Wu, R. Ramamurthy, F. Rudzicz, M. Brudno, "Speaker diarization with session-level speaker embedding refinement using graph neural networks", in: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2020, pp. 7109–7113.
[14]. I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Ko-renevskaya, I. Sorokin, T. Timofeeva, A. Mitrofanov, A. Andrusenko, I. Podluzhny, A. Laptev, A. Romanenko, "Target speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario", in: Proceedings of the Annual Conference of the International Speech Communication Association, 2020, pp. 274–278.
[15]. M. Diez, L. Burget, S. Wang, J. Rohdin, J. Cernocky, "Bayesian HMM based x-vector clustering for speaker diarization", in: Proceedings of the Annual Conference of the International Speech Communication Association, 2019, pp. 346–350.
[16]. Nauman Dawalatabad, Student Member, IEEE, Srikanth Madikeri, Member, IEEE, C. Chandra Sekhar, Member, IEEE, and Hema A. Murthy, Senior Member, IEEE, "Novel Architectures for Unsupervised Information Bottleneck Based Speaker Diarization of Meetings", 2020.
[17]. Shaojin Ding2, Quan Wang, Shuo-yiin Chang, Li Wan, Ignacio Lopez Moreno, "Personal VAD: "Speaker Conditioned Voice Activity Detection",2020.