

Novel Approach towards Speaker Identification in Meetings

Anuj Apte¹, Vedang Atgur², Pranjay Beniwal³, Sanraaj Flora⁴ and Shilpa Sonawani⁵

¹⁻⁵Department of Computer Engineering and Technology, MIT World Peace University, Pune, India

Email: {anujapte07, vedang.mit, pranjayb12, sanraajflora2003}@gmail.com, shilpa.sonawani@mitwpu.edu.in

Abstract—Effective documentation and script recording of official proceedings, meetings and conferences is essential for decision-making. But manually recording who says what can be time-consuming and prone to mistakes. In order to automate the process of identifying speakers during meetings, we propose in this paper a speaker identification system that makes use of deep learning techniques. Our system extracts features from audio recordings using Mel-frequency cepstral coefficients (MFCCs) and Long Short-Term Memory (LSTM) networks. We validate and experiment a great deal to show that our method works well for reliably identifying speakers in real-time scenarios. This study helps to improve efficiency and accuracy in recording and analyzing meeting content by streamlining the formal meeting documentation process.

Index Terms— LSTM, MFCC, Recurrent Neural Networks and Deep Learning.

I. INTRODUCTION

Formal gatherings such as conferences and meetings are essential platforms for sharing knowledge, making decisions, and working together in a variety of fields. Accurate records must be kept, and further analysis will be made easier with the help of the meeting documentation, which includes speaker identities and contributions [15]. Whereas, manually recording meeting proceedings can be laborious, error-prone, and time-consuming [5]. There's a growing need for automated solutions to make this process more efficient as the number of meetings rises and the requirement for effective documentation grows [4]. Complex speaker identification systems have been made possible in recent years by developments in deep learning [7] [8]. These systems use methods like LSTM-based neural networks [15], [3] and MFCC extraction [4] [5] [6] [12] [13] to analyze audio recordings and recognize speakers based on their distinct vocal traits. Although speaker identification has been studied in great detail in a variety of settings, there are special opportunities and challenges when applying it in real-time, especially in official meetings. In this paper, we present a unique speaker identification method designed especially for conferences and official meetings. Our system seeks to precisely identify presenters in real-time so that meeting content can be seamlessly documented without the need for human intervention. We aim to overcome the drawbacks of manual documentation and improve the effectiveness and precision of meeting record-keeping by utilizing deep learning and signal processing techniques.

II. MOTIVATION AND OBJECTIVES

The need to automate and expedite the documentation process at formal conferences and meetings is what spurred

this research. Meeting records are frequently delayed in becoming available because manual documentation of the proceedings, including speaker identification and transcription, is time-consuming and prone to mistakes. Furthermore, mistakes or omissions in the documentation may have major repercussions in cases where accurate documentation is essential for making decisions or fulfilling legal requirements [15]. Our goal is to address these issues and enhance the effectiveness and precision of meeting documentation by creating an automated speaker identification system designed especially for official gatherings and conferences. In addition to alleviating the workload for human transcribers, automated speaker identification allows for real-time recording of meeting materials, guaranteeing that meeting minutes are accessible shortly after the meeting ends. Furthermore, real-time synchronized captions or transcripts made possible by automated speaker identification can improve the accessibility of meeting content for those with hearing impairments. The overall objective of this research is to create a reliable and accurate speaker identification system that satisfies the particular needs and difficulties of formal meetings and conferences by utilizing developments in deep learning [5] [8] [11] and audio processing techniques. By doing this, we hope to improve meeting documentation's effectiveness, precision, and accessibility, which will eventually lead to improved decision-making and teamwork in governmental and organizational contexts.

III. LITERATURE REVIEW

The papers we surveyed used various techniques for recognising speaker and found comparable results. Bharti et al. [1] provided a comprehensive overview of various deep learning techniques involved in speaker recognition and processing. The review covers publications from 1991 to 2021 and is from various sources. Shome et al. [2] showcased the importance of dataset size and noise levels for speaker recognition tasks. They prepared a dataset of 80 individuals with five voice samples each. After feature extraction and comparing various models such as Dense Neural Network, CNN-Dense hybrid and LSTMs, the results proved that Dense Neural Network outperformed the rest due to the relatively small dataset. Naveen et al. [3] worked on audio biometrics, focusing on advancements, challenges, and applications in speaker recognition systems. They compared SVM with LSTM on VoxCeleb dataset and proved that the SVM showed better results, showcasing its ability to fit the training data well and generalize effectively. Most of these papers implemented Mel-frequency cepstral coefficients (MFCC) for feature extraction of the audio as a pre-processing technique. For instance, Roumiassa et al. [4] in their paper used MFCC, along with Linear Predictive Coding (LPC) combined with Principal Component Analysis (PCA) for feature reduction, and Dynamic Time Warping (DTW) for speaker identification and verification in Arabic and Berber languages. The performance of the system varies between 30% to 100%, which is influenced by the speaker's emotional and physical state, speech speed, recording conditions and noise levels, and the results demonstrate the effectiveness of the proposed methodology (using MFCC and LPC with PCA). Nehra et al. [5] in their approach used a similar system as that of ours, firstly with the dataset - comprising of speeches of prominent leaders such as Nelson Mandela, Julia Gillard, Benjamin Netanyahu, Jens Stoltenberg, and Margaret Thatcher. Instead of using traditional method of MFCC for feature extraction, they utilized CNN for both feature extraction and the classifier, and managed to score 96.54% accuracy on the test data. On performing a comparative analysis, they found out that this accuracy outperformed existing techniques such as MFCC with VQ, MFCC with SVM, and MFCC with ANN, showcasing the effectiveness of the CNN-based methodology for speaker identification. Pentapati et al. [6] performed experiments on a simplified CNN. They utilized Log-MelSpectrum to represent speech signals for speaker identification. They experimented on Libri Speech corpus with a limited dataset. The CNN had dilated convolution layers, Leaky ReLu activation, Batch Normalization layers, max pooling, and a simple neural network as a fully connected layer, while they aimed to train the network effectively even with small data samples per speaker. The accuracy obtained was 80.6% showcasing the effectiveness of the proposed approach and it also outperformed traditional CNN models without dilation factors. Sarangi et al. [7] in their paper introduced the Speech signal-based frequency cepstral coefficient (SFCC) as a promising alternative to traditional MFCC. By comparing the inverse of SFCC (ISFCC) with the inverse of MFCC (IMFCC), the study aimed to enhance the performance of speaker identification systems. Experimenting on the POLYCOST database which contains recording sessions of 134 speaker from 14 different countries, the Gaussian mixture model- maximum likelihood (GMM-ML) classifier is employed. The results demonstrate that in standalone conditions, SFCC and ISFCC outperform MFCC and IMFCC, with ISFCC showing better performance than IMFCC. Furthermore, the fusion of SFCC with ISFCC yields significant improvements in accuracy, with relative improvements of up to 4% and 7% over MFCC and IMFCC,

respectively. Xie et al. [8] in their research work, integrated multi-channel information, transfer learning and Harmonic-Percussive Source Separation (HPSS) techniques which helped in large-scale speaker identification tasks, using a modified LeViT transformer- based architecture to process the spectrograms of three channels. They used VoxCeleb1 dataset to conduct experiments and demonstrated the effectiveness. Saritha et al. [9] introduced DNN-based system for Speaker Identification which has a CNN-based architecture utilising on IITG Multi-variability speaker recognition database. The study shows a 5% improvement in identification accuracy for sensor mismatches, a 3% improvement for environment mismatches, and a 4% improvement for speaking style discrepancies. Boubakeur et al. [10] presents a study on Prosodic modelling, basically pitch and intensity, for speaker identification in noisy environments. It involves the use of Two Automatic Speaker Identification (ASI) systems: one based on Hidden Markov Models (HMM) while the other is using Support Vector Machines (SVM). After performing feature extraction, model training and testing, the results show that the SVM outperforms HMM-based system with better accuracy and robustness. Overall, the addition of prosodic features improves the process, especially in high-noise environments. Shetty et al. [11] in their paper implemented Siamese Neural Network (SNN) along with a Gaussian Mixture Model (GMM) for speaker identification. They implemented MFCC prior, for preprocessing audio data, followed by the removal of silent segments. Results show that SNN achieved an Equal Error Rate (ERR) of 4.88% while GMM achieved 6.55%, indicating that SNN is superior as compared to GMM here. Oo et al. [12] used a similar approach as that of ours, but in addition, they introduced Voice Active Detection techniques through their research. They have prepared a custom dataset containing 10 Myanmar speakers – 5 males and 5 females, with each having 60 repetitions of 3 utterances. They focused on utilizing a combination of zero-crossing rate (ZCR) and short-term energy (STE) as active voice detection features, which aim to remove unwanted components and reduce background noise. After applying MFCC and comparing them, the model using ZCR, STE, and MFCC features achieved a significantly higher accuracy of 89.25% compared to the model using only MFCC features, which scored 84.44%, indicating that integration of active voice detection techniques enhances the performance of speaker identification systems. Loina et al. [13] explored the feasibility of using small artificial neural networks with a limited data for speaker recognition. Along with MFCC, various preprocessing techniques were used such as Fast Fourier Transformation (FFT) and Short-term Fourier Transformation (STFT) for assessment. Results showcase that the preprocessing done with MFCC proved to be superior as compared to that done with FFT and STFT. The research identified a residual neural network with 387,461 parameters as the optimal model for achieving high accuracy in speaker recognition tasks. Also, the study compared the training times of different neural network architectures with X-vectors based network, which although required more time to train, demonstrated the best performance in terms of accuracy. Nandhitha et al. [14] in their research perform speaker recognition using autocorrelation features, Approximation statistical features, and detailed statistical features with the application of Recurrent neural network (RNN) and Long-Short term memory (LSTM) neural networks. They involve methods for extracting various statistical parameters such as Pitch, Period, NOP, Mean, Skewness, Kurtosis, Energy, ZCR, Second Order Moment, and Third Order Moment for the same. They compare the performance of RNN and LSTM using different class of inputs and show that LSTM has a higher accuracy and sensitivity than RNN. Wang et al. [15] worked on speaker diarization using d-vector based systems in combination with LSTM-based audio embeddings and non-parametric clustering algorithms, aiming to improve the efficiency and accuracy of speaker diarization which is crucial in speaker identification and speech recognition. They employed two main models here, namely i-vector which is trained using PLP coefficients with GMM-UBM and LDA dimensionality reduction, and d-vector model which consists of a 3-layer LSTM network with specific node configurations. Additionally, a Voice Activity Detection (VAD) model is utilized for segmenting speech and non-speech regions. The performance metrics used were Diarization Error Rates (DER) on standard public datasets such as CALLHOME American English, NIST Rich Transcription, and NIST Speaker Recognition Evaluation. The results showed that the d-vector based diarization systems which leveraged LSTM outperformed the i-vector based systems significantly, achieving lower DER rates compared to traditional i-vector systems.

IV. METHODOLOGY

To get a base-level understanding of the flow of the entire system, refer to Fig. 1

A. Dataset Preprocessing

The dataset we used was the Speaker Recognition Dataset taken from Kaggle which contained speeches of 5 prominent leaders [5] [11], namely Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret Thatcher and Nelson Mandela. It consisted of 7500 files, split into 5 directories, named after the particular speaker, with a lengthy WAV file for each speaker which was divided into 1500 equal sized WAV files, each being 1 second long. We resampled these files to 16000 Hz [6] to bring about uniformity and consistency [5]. Then using the os library in Python, we identified the dataset directory and all the audio files in it, by creating labels for each directory.

B. Feature Extraction

To extract relevant features from the dataset, we used the librosa Library [9] to extract features using the Mel-Frequency Cepstral Coefficient (MFCC) [4] [5] [6] [7] [10] [11], taking 13 coefficients [8] to extract features, after which we used the Standard Scaler to normalize these features. We then used the Label Encoder to encode speaker labels to numerical format.

C. Splitting

We split the dataset into training (50%), testing (30%), and validation (20%) sets using the train_test_split method from Scikitlearn with random state set to 42.

D. Model Definition and Compilation

A Keras Recurrent Neural Network (RNN) Model [2] of 3 layers was defined for decoding. This model was compared with 2 other models that used 7 CNN [1] [2] [3] layers and a hybrid model that used 7 layers of both CNN and RNN- LSTM. We chose the RNN-LSTM model among the 3 models as it was the best performer in terms of Accuracy, Precision, Recall and F1-score [1] [2], as shown in Table I and Fig. 2. First, we added an RNN layer with 128 LSTM (Long Short Term Memory) [14] [15] [1] neurons. Next a dense layer [3] was added, with 64 neurons and a Rectified Linear Unit (ReLU) activation function [5] [6][9] [3] to add non-linearity to the model. We added another dense layer which had 5 neurons and the softmax activation function [6] [3] which converts the output into probability scores that should add up to 1. The model was then compiled using Adam optimizer which is an adaptive learning rate optimizer algorithm, and sparse_categorical_crossentropy was used to calculate loss between true labels and the predicted probabilities generated by the softmax function.

E. Training

Next, we trained the model using 20 epochs and batch size of 32. The model was then saved as an HDF5 file and the encoded labels of the 5 speakers [1] are saved as a pkl file for later use to decode predictions using the Pickle library.

F. Implementation

To implement this code, we first used the HDF5 and pkl files to load the model weights and encoded labels. We then used sounddevice library to record live audio as input data for prediction. The recorded audio undergoes feature extraction by the MFCC method. The resulting features are then transposed and vectorized. For prediction of the speaker, the features' dimensions are expanded and matched with those in the dataset [11], and the probabilities of each speaker are obtained, after which the name of the speaker is displayed on the output console.

V. RESULTS

As shown in Fig. 3 and Fig. 4, the training process for this model had of 20 epochs, with its performance evaluated at each epoch on both the training and validation datasets. Throughout the training process, the model's accuracy on the training dataset increased gradually, the first epoch yielding a training accuracy and loss of 67% and 79% respectively, whereas the last epoch yielded a training accuracy and loss of 99% and 2% respectively. Over the course of the epochs, the training loss decreased significantly, demonstrating the model's capacity to learn and reduce errors. The validation dataset performance of the model also demonstrated an increasing trend, yielding a training accuracy and loss of 92% and 22% respectively, whereas the last epoch yielded a training accuracy and loss of 98% and 8% respectively. This suggests that the model was not over-fitting to the training set and that it performed well when applied to new data. The validation loss also steadily dropped over the course of the epochs, indicating the model's capacity to generalize and produce precise predictions on fresh data. with reference to Table I, the model's overall metrics were evaluated by calculating the accuracy, precision, recall

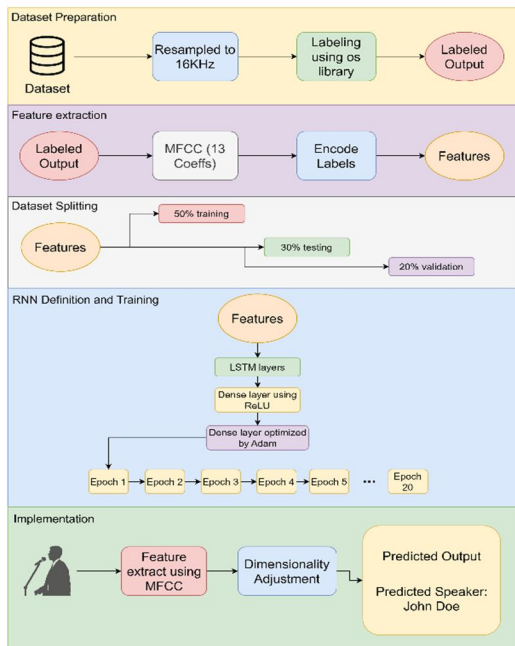


Figure 1. System Architecture of the Model

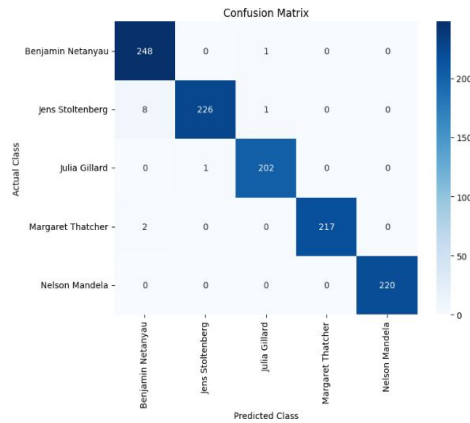


Figure 2. Confusion Matrix of the results obtained

and F1-score [1] [2], being 98.8%, 98.9%, 98.8% and 98.8% respectively. The confusion matrix, as shown in Fig. 2, represents all the correctly and incorrectly classified instances, with 1 misclassified instance of Benjamin Netanyahu being misclassified as Julia Gillard, Jens Stoltenberg as Julia Gillard and Julia Gillard as Jens Stoltenberg and 2 mis-classifications of Margaret Thatcher with Benjamin Netanyahu.

TABLE I. METRICS OF ALL THE MODELS USED

Model	Layers Used	Accuracy	Precision	Recall	F1-Score
CNN	7	97.6%	97.68%	97.64%	97.62%
CNN with RNN-LSTM	7	98.66%	98.66%	98.65%	98.64%
RNN-LSTM	3	98.84%	98.94%	98.87%	98.89%

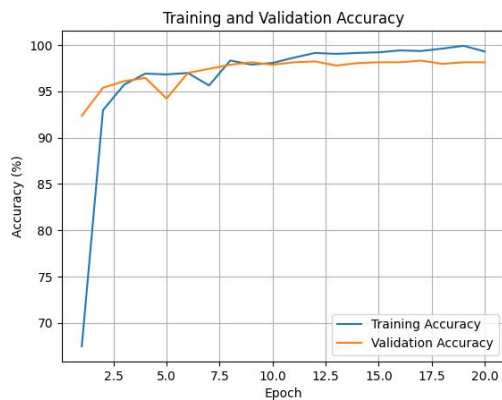


Figure 3. Training and Validation Accuracy of RNN-LSTM model

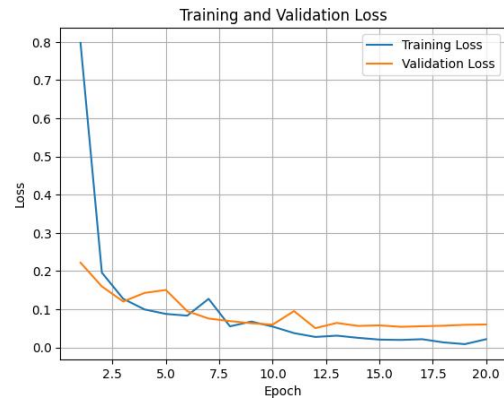


Figure 4. Training and Validation Loss of RNN-LSTM model

VI. CONCLUSIONS

Through extensive experimentation, our system demonstrated high accuracy in real-time speaker identification tasks. By automating the labor-intensive process of identifying speakers, our system streamlines meeting

documentation, enhancing efficiency, accuracy, and accessibility in official proceedings. Future research could focus on further optimization, scalability, and integration with existing meeting transcription tools to enhance functionality and usability. Overall, our study highlights the potential of deep learning in automating speaker identification tasks, thereby facilitating better decision-making and collaboration in organizational and governmental settings.

FUTURE SCOPE

In the future, there is potential to deploy the speaker identification system in real-time environments, where integration with speaker diarization techniques could significantly enhance accuracy by segmenting audio recordings into individual speaker segments. Moreover, the implementation of adaptive learning mechanisms would enable the continuous refinement of the model based on new data and feedback, ensuring its adaptability to evolving speaker characteristics and environmental conditions. Additionally, exploring multimodal integration with video or text-based information could offer complementary cues for more robust speaker identification, especially in complex audio scenarios. Addressing privacy and security concerns, ensuring scalability across diverse meeting environments, and seamless integration with existing tools are critical considerations for maximizing the system's effectiveness and widespread adoption in organizational and governmental settings.

REFERENCES

- [1] Bharti, S., Kumari, A., Verma, D. & Kumar, A. Speaker identification using various machine learning algorithms. 2024 14th International Conference On Cloud Computing, Data Science & Engineering (Confluence). (2024,1)
- [2] Shome, N., Sarkar, A., Ghosh, A., Laskar, R. & Kashyap, R. Speaker recognition through deep learning techniques. *Period. Polytech. Electr. Eng. Comput. Sci.* 67, 300-336 (2023,7)
- [3] Naveen, Reddy, J., Tanguturu, R. & Kumar, A. Speaker Identification and Verification using Deep Learning. 2022 International Conference On Signal And Information Processing (IConSIP). (2022,8)
- [4] Roumiassa, F. & Chelali, F. Speaker identification and verification system for Arabic and Berber language. 020 1st International Conference On Communications, Control Systems And Signal Processing (CCSSP). (2020,5)
- [5] Nehra, N., Sangwan, P. & Kumar, D. Speaker identification system using CNN approach. 2021 International Conference On Industrial Electronics Research And Applications (ICIERA). (2021,12)
- [6] Pentapati, H. & Sri Devi Dilated convolution and MelSpectrum for speaker identification using simple deep network. 2022 8th International Conference On Advanced Computing And Communication Systems (ICACCS). (2022,3)
- [7] Sarangi, S., Panda, M. & Sahu, B. Use of complementary feature to enhance the performance of speaker identification system. 2021 1st Odisha International Conference On Electrical Power Engineering, Communication And Computing Technology (ODICON). (2021,1)
- [8] Xie, Y., Tang, J., Yang, N. & Pun, M. Large-scale multi-channel transformer- based speaker identification with knowledge transfer using harmonic-percussive source separation. 2022 31st Wireless And Optical Communications Conference (WOCC). (2022,8)
- [9] Saritha, B., Sharma, S., Thiru Venkata Naga Manoj, T., Monsley, A., Hussain Laskar, R. & Choudhury, M. DNN based speaker identification system under multi-variability speech conditions. 2022 IEEE International Women In Engineering (WIE) Conference On Electrical And Computer Engineering (WIECON-ECE). (2022,12)
- [10] Boubakeur, K., Debyeche, M., Amrouche, A. & Bentrchia, Y. Prosodic Modelling based Speaker Identification. 2022 2nd International Conference On New Technologies Of Information And Communication (NTIC). (2022,12)
- [11] Shetty, P., Rodricks, R., Malgundkar, S., Pamnani, H. & Katke, S. SPEECH BIO- METRICS: A comprehensive deep learning-based speaker identification system. 2023 International Mobile, Intelligent, And Ubiquitous Computing Conference (MI- UCC). (2023,9)
- [12] Oo, K., Thu, L. & Aung, Z. Enhancement of speaker identification system based on voice active detection techniques using machine learning. 2024 Conference Of Young Researchers In Electrical And Electronic Engineering (ElCon). (2024,1)
- [13] Loina, L. Speaker identification using small artificial neural network on small dataset. 2022 International Conference On Smart Systems And Technologies (SST). (2022,10)
- [14] Nandhitha, N., Roslin, S. & Rajasekhar, B. LSTM based speaker recognition system with optimized features. 2023 Fifth International Conference On Electrical, Computer And Communication Technologies (ICECCT). (2023,2)
- [15] Wang, Q., Downey, C., Wan, L., Mansfield, P. & Moreno, I. Speaker Diarization with LSTM. 2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP). (2018,4)