PROJECT REPORT-SPEAKER IDENTIFICATION

Team number-19

Team members

Name	Email ID
Leya kurian	leyahoney2003@gmail.com
Hemal shaji	hemalshaji77@gmail.com
Jubelin elizabeth joji	jeliza.jo11@gmail.com

Title: Speaker Identification System Development

Problem Statement:

This project aims to develop a robust speaker identification system capable of accurately recognizing and verifying speakers from audio recordings.

Dataset Details:

https://www.kaggle.com/kongaevans/speaker-recognition-dataset

The "Speaker Recognition Dataset" available on Kaggle, provided by user Kongaevans, is a collection of audio recordings intended for use in speaker recognition research and development. The dataset contains audio recordings of multiple speakers, each uttering a specific phrase. The audio files are in WAV format, which is a common format for uncompressed audio data. The dataset is designed to capture variability among speakers, including differences in voice characteristics, accents, intonations, and other speech-related attributes. This dataset is suitable for various tasks related to speaker recognition, including speaker verification (determining if two samples come from the same speaker) and speaker identification (assigning an identity to a speaker from a pool of known speakers).

Challenges: The primary hurdle we encountered was the sheer scale of the dataset, which led to extensive training times averaging between 4 to 5 hours per iteration. To navigate this challenge, we devised a simplification tactic. Specifically, we opted to strategically reduce the dataset size by randomly selecting a set number of audio samples per speaker. This adjustment not only alleviated the strain on computational resources but also expedited the training process significantly. Consequently, we achieved a more streamlined workflow, enabling us to optimize resource allocation and enhance overall efficiency in our research endeavors.

Method or Experimental Setup:

In our project, we meticulously designed and implemented a sophisticated speaker identification system. Leveraging cutting-edge techniques in deep learning, we constructed a Convolutional Neural Network (CNN) architecture with residual blocks, ensuring robust feature extraction and classification capabilities from audio recordings.

Model Architecture:

Our model architecture is meticulously crafted to leverage the power of CNNs with residual blocks. Each residual block is intricately designed, incorporating convolutional layers with batch normalization and ReLU activation, coupled with a skip connection. These residual blocks play a pivotal role in enabling the model to learn intricate patterns in audio data effectively, while the skip connections mitigate the notorious vanishing gradient problem. Following the residual blocks, pooling layers are employed to reduce spatial dimensions, enhancing computational efficiency. The flattened features are then fed into dense layers for final classification, with the output layer utilizing softmax activation to provide probabilities for each speaker class.

Training Configuration:

During model training, we meticulously configured various parameters to ensure optimal performance. We employed the Adam optimizer along with sparse categorical cross-entropy loss function, specifically tailored for multi-class classification tasks. Training proceeded for a predetermined number of epochs, with early stopping mechanisms in place to prevent overfitting by monitoring validation loss. To track the best-performing model, we implemented model checkpoints. Furthermore, we employed advanced data augmentation techniques, such as injecting noise into input audio samples, to bolster the model's resilience to environmental variations and noise types. Additionally, to mitigate model bias and enhance training efficiency, we shuffled and batched the training and validation datasets.

Observations and Analysis:

The performance of our implemented model surpassed expectations, boasting an impressive accuracy rate of approximately 94% on the validation dataset. This stellar performance can be attributed to various factors, including the adept utilization of residual blocks within the CNN architecture, effectively addressing gradient vanishing issues and enabling the model to discern complex patterns in audio data. Furthermore, the augmentation of training data with noise and the utilization of Fast Fourier Transform (FFT)-

based feature extraction techniques significantly contributed to the model's prowess in speaker identification tasks.

Results:

Metric	Value
Accuracy	0.9480
Loss	0.1756

Hypotheses and Further Analysis:

Moving forward, we envision several avenues for further exploration and refinement of our model. Hyperparameter tuning, including experimenting with different configurations of CNN architectures and fine-tuning parameters such as learning rate and batch size, holds promise for optimizing model performance further. Exploring additional data augmentation techniques, such as time-domain transformations or spectrogram augmentation, could bolster the model's robustness to various types of noise and environmental variability. Additionally, in-depth analysis of misclassifications and leveraging transfer learning techniques from related tasks such as speech recognition present exciting opportunities for enhancing model accuracy and real-world applicability. Through continued experimentation and refinement, we aim to push the boundaries of speaker identification technology and deliver even more accurate and reliable solutions.

Conclusion:

The implemented CNN-based model, augmented with noise and utilizing FFT-based feature extraction, showcases promising results in speaker recognition. Through careful experimentation and analysis, including hyperparameter tuning, exploration of data augmentation techniques, and in-depth examination of misclassifications, further enhancements can be achieved. Continual refinement of the model and methodologies will contribute to the development of robust and accurate speaker recognition systems for various real-world applications.

Acknowledgment:

We would like to extend our sincere appreciation to IIIT Hyderabad for providing us with the invaluable opportunity to participate in the Research Teaser Program. This program has been instrumental in expanding our knowledge and skills in the field of research, and we are grateful for the platform it has provided for our professional development.

We would also like to express our heartfelt gratitude to the team at IIIT Hyderabad for their guidance and support throughout the program. Their expertise and mentorship have been invaluable in helping us navigate the complexities of research and learn new concepts and methodologies.

We would also like to thank the management and faculty of Saintgits College of Engineering for facilitating our participation in the IIIT Hyderabad Research Teaser Program. Their support and encouragement have been instrumental in shaping our academic journey and fostering our passion for research.

Participating in the IIIT Hyderabad Research Teaser Program has been a transformative experience, and we are confident that the skills and knowledge gained during this program will greatly benefit our future endeavors in research and academia. We are honored to have been a part of this prestigious program and look forward to applying what we have learned to make meaningful contributions to the field of research.

Code:

https://colab.research.google.com/drive/1b8tvs8Shi6sg9erG0DHH6InEPEiXuMKa?usp=sharing