Speaker recognition with Speaker dataset

Harshit Choudhary
2021254
IIIT Delhi
harshit21254@iiitd.ac.in

Manya Tyagi
2021064
IIIT Delhi
manya21064@iiitd.ac.in

Viviana Longjam
2021115
IIIT Delhi
viviana21115@iiitd.ac.in

Ankur Tiwari
MT23017
IIIT Delhi
ankur23017@iiitd.ac.in

Shashank S. Pathak
MT23141
IIIT Delhi
shashank23141@iiitd.ac.in

Abstract

This project delves into Speaker Identification through a comprehensive approach. We commence with Exploratory Data Analysis (EDA) on a dataset featuring prominent figures like Nelson Mandela, Benjamin Netanyau, Jens Stoltenberg, Julia Gillard, and Margaret Thatcher. The EDA reveals nuanced audio characteristics, prompting the addition of noise to enhance model robustness. Subsequent steps involve Principal Component Analysis (PCA) and feature extraction using Mel Frequency Cepstral Coefficients (MFCC). The extracted features are then employed in a Gaussian Mixture Model (GMM) to develop a machine learning model for speaker identification. This holistic methodology aims to capture diverse speech attributes and achieve accurate speaker classification.

1 Introduction

Speaker Identification, a pivotal aspect of speech and audio processing, impacts diverse fields such as behavioral studies, linguistics, and sociolinguistics. It contributes to understanding linguistic diversity by unraveling regional accents, dialects, and speech variations.

In practical terms, speaker identification is employed by voice assistants and smart devices to deliver personalized user experiences. It proves valuable in forensic investigations for analyzing audio evidence and identifying speakers in recordings. Additionally, it serves as a secure means of authentication in various scenarios.

The project, Speaker recognition with Speaker datset aims to enhance the reliability and accuracy of speaker identification systems. Utilizing sophisticated machine learning algorithms, noise augmentation, and supervised learning, it seeks to develop a flexible model for precise speaker identification in diverse settings. This aligns with broader objectives, including strengthening security protocols, enabling customized communications, and advancing research in linguistics and voice analysis.

In this project, we employ a multifaceted approach for accurate speaker classification. Starting with Exploratory Data Analysis (EDA), we examine recordings to uncover distinct audio nuances. Noise augmentation is done in order to mimic real-world conditions, and Principal Component Analysis (PCA) to reduces= data dimensionality. Mel Frequency Cepstral Coefficients (MFCC) extraction was then performed to capture vital speech features. These features were fed into a Gaussian Mixture Model (GMM) for speaker classification and learning individual speech attributes.

Through a holistic methodology that includes EDA, noise augmentation, dimensionality reduction, and GMM classification, the project aims for precise and resilient speaker identification across diverse speech variations and environments.

2 Related Work

2.1 Speaker Identification on the SCOTUS Corpus by Jiahong Yuan and Mark Liberman

The paper employs a diverse strategy using Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) for speaker recognition in the SCOTUS corpus. It starts with robust feature extraction (MFCC, PLP), trains models with HTK and CMU Pronouncing Dictionary, and tests on various oral arguments, exploring sampling rate impacts. It delves into inter-speaker variability, pinpointing highly variable speech sounds like ’UH’ and ’NG’. The study evaluates alignment accuracy, revealing insights into speaker variability in courtroom speech data.

2.2 Gaussian Mixture Model-Based Speech Recognition System Using MATLAB by Manan Vyas

The study introduces a MATLAB®-based speaker-dependent speech recognition system. It outlines meticulous MFCC-based Feature Extraction and robust Voice Activity Detection algorithms for varied noise levels. Employing Gaussian Mixture Models (GMM) with MLE and EM algorithms refines parameter estimation. Performance Analysis exposes system strengths, limitations, and environmental influences. It highlights deployment complexities and envisions future integration in diverse fields like security, healthcare, automation, telephony, robotics, and assistive tech.

2.3 Real-time speaker identification and verification by T. Kinnunen, E. Karpov and P. Franti

The text covers several methodologies in speaker recognition systems, emphasizing pre-quantization, speaker pruning, and cohort normalization for efficient verification Evaluations on TIMIT and NIST corpora showcase performance metrics like identification error rates, equal error rates, and speed-up factors for various techniques—highlighting the trade-offs between accuracy and computational efficiency The integration of Vector Quantization (VQ) and Gaussian Mixture Models (GMM) is explored, demonstrating adaptability across different modeling techniques. The text references various mathematical approaches - Bayesian likelihood ratio, Log-likelihood ratios LBG algorithm for pre-quantization

2.4 Speaker identification features extraction methods: A systematic review

The study employed the Kitchenham systematic review methodology, scrutinizing 160 publications on SI feature extraction from 2011 to 2016. It focused on identifying, comparing, and analyzing extraction approaches through exclusion criteria. The review answered three research questions and evaluated publications’ relevance based on citations and implementations. It used SLR guidelines to fill gaps in existing literature, categorizing SI applications and systematically assessing feature extraction methods’ significance and trends over the specified timeframe.

3 Exploratory Data Analysis

During Exploratory Data Analysis, an in-depth analysis was conducted on a Speaker Recognition dataset with the primary objective of comprehending the dataset characteristics, extracting essential audio features, and preparing the groundwork for robust speaker recognition models. The exploration encompassed various data preprocessing techniques, and visualization tools aimed at understanding the unique attributes of each speaker within the dataset. The dataset comprises five principal classes representing distinct speakers: Nelson Mandela, Benjamin Netanyau, Jens Stoltenberg, Julia Gillard, and Margaret Thatcher. Notably, the classes are balanced, ensuring equitable representation for model training. Additionally, two noise classes are included, augmenting the dataset’s robustness to diverse acoustic environments.

Then audio attribute analysis was performed which showed that all classes exhibit consistent 1.00-second duration clips at a sample rate of 22050 Hz, ensuring uniform audio quality across the dataset. Further, it was found that the audio characteristics vary among speakers, with Mandela and Thatcher exhibiting higher average power (0.06), indicative of consistent speech strength, while Netanyau and Stoltenberg have lower power (0.03 and 0.02, respectively). Gillard stands out with an average power of 0.07. Total energy levels further differentiated speakers, reflecting vocal dynamics.

The utilization of histograms and bar graphs elucidated the distribution patterns of essential audio features within the dataset, aiding in understanding the distinguishing attributes of each speaker class Leveraging Short-Time Fourier Transform (STFT) spectrograms, Mel spectrograms, and Mel-Frequency Cepstral Coefficients (MFCC) spectrograms, the analysis revealed distinct frequency distribution patterns unique to each speaker, providing crucial insights for subsequent speaker discrimination and model development.

4 Methodology

To train our model for prediction, we first preprocessed the data, performed feature extraction by selecting relevant features and then used these features to train a GMM model.

4.1 Preprocessing

4.1.1 Chunking Noise and Noise Augmentation

The audio samples of the speakers provided are clean. While these are good samples, to make the model robust, noise must be added during training in some of the samples. The noise available in the Speaker dataset included background noise as well as other noise such as pink noise. First, the noise audio was chunked to be the same size as the 1 second samples of the speakers. Then in 20% of the data, this chunked noise was augmented randomly, both additive and multiplicative.

4.1.2 Human Voice Layovers

Aside from the background noise, there is a possibility that an audio sample of a specific speaker may contain additional voices apart from theirs. To incorporate for this case while training the model, some audios in the above 20% of the data were also added with scaled-down audios of another speaker. The model should then be able to predict the speaker when another voice is also present.

4.1.3 Principal Component Analysis

Principal Component Analysis (PCA) is carried out after the following section of feature extraction. In order to reduce the dimensionality of extracted features while retaining the important information, PCA performs eigendecomposition on the computed covariance matrix of the dataset and selects the top principal components.

4.2 Feature Extraction

For feature extraction we used MFCC, Delta, Double delta, LPC and LBD.

4.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs are widely used for capturing the spectral characteristics of an audio signal. They mimic the human ear's sensitivity to different frequencies. MFCCs are computed by taking the discrete cosine transform of the logarithm of the short-term power spectrum of the audio signal. They effectively represent the power spectrum, emphasizing the frequency bands that are more relevant for speech perception.

4.2.2 Delta and Double Delta

Delta coefficients represent the rate of change of MFCCs over time. They provide information about the dynamic aspects of speech signals. Delta coefficients are calculated by taking the first-order difference of consecutive MFCCs. They capture the temporal dynamics in speech, contributing to the characterization of speaker-specific patterns.

Double delta coefficients (order = 2) capture the acceleration or curvature of the MFCCs, providing further information on the speech signal's dynamics. Double delta coefficients are computed by taking the first-order difference of consecutive delta coefficients. They enhance the representation of temporal variations, contributing to the modeling of speaker-specific speech patterns.

4.2.3 Linear Prediction Coefficients (LPC)

LPC represents the coefficients of a linear predictive model applied to the speech signal. It is used to model the vocal tract system. LPC is computed by estimating a predictive model that minimizes the prediction error of the speech signal. LPC is particularly useful for modeling the resonant characteristics of a speaker's vocal tract, providing information about speech production.

4.2.4 Logarithmic Bandwidth Detector (LBD)

LBD characterizes the spectral bandwidth of the speech signal in a logarithmic scale. It is used to capture information about the distribution of energy in different frequency bands. LBD is computed by dividing the frequency spectrum into logarithmically spaced bands and calculating the energy in each band. LBD can be effective in capturing the unique spectral features of a speaker's voice, particularly in the context of distinguishing speakers with different vocal characteristics.

4.3 Model Training

After performing feature extraction and PCA, and after splitting the dataset for training and testing, we used GMM for model training. As there are 5 speakers, 1 GMM was trained for each speaker, i.e. 5 GMM models. During prediction, scores were calculated to see which of the GMMs gave the closest match to the test audio, and this was the predicted speaker.

Using the extracted features, we also tried to use LinearSVC as a model for prediction instead. While we found similar accuracy, GMMs were slightly better.

5 Types of Model Used

A **Gaussian Mixture Model** was used in our project, which is basically a probabilistic model that represents a mixture of multiple Gaussian distributions. Each Gaussian component in the mixture model corresponds to a cluster or mode in the data.

As found in the literature review, GMM's find a notable application in speaker Identification, due to their ability to effectively model the distribution of acoustic features (such as MFCCs - Mel-Frequency Cepstral Coefficients) associated with different speakers. Further, they provide several advantages starting from Statistical Representation, which helps in capturing the variability in speech features associated with different individuals to being adaptable to specific speakers over time, allowing the model to better match individual variation in speech patterns. Further, their probabilistic decision-making framework, allowing for more robust speaker identification by considering the uncertainty associated with each decision and being used as a mixture of acoustic models, where each Gaussian component represents a different phoneme or speaker makes them suitable for modeling complex relationships in speech data.

6 Analysis

In the pursuit of enhancing the robustness and performance of our speaker identification model, several preprocessing and feature extraction techniques were employed post-selection of the model between GMM and Linear SVC, where GMM gave us better accuracy in every case. These strategies aimed to improve the model's adaptability to real-world scenarios and diverse speech variations.

6.1 Noise Augmentation

We initiated the preprocessing phase with noise augmentation, a critical step in mitigating the impact of environmental factors on the model. The noise augmentation involved a multi-faceted approach, including noise multiplication and addition. To further enhance the model's resilience, voice layovers were introduced, incorporating a 25

6.2 Delta and Double Delta in MFCC

Building upon the noise augmentation, we incorporated delta and double delta features in the Mel Frequency Cepstral Coefficients (MFCC). The inclusion of delta and double delta features adds valuable temporal information to the static MFCC, capturing variations in speech dynamics over time. This proves particularly beneficial in our speaker identification project, as it allows the model to discern nuanced changes in speech patterns and improves the system's ability to identify speakers across diverse utterances.

Description	Test Accuracy	Train Accuracy
Using LPC	0.92	0.9
Using LPCC	0.89	0.93
Using LDB	0.92	0.88

Table 1: Test and Train Accuracy using various Feature Extraction Methods

6.3 Linear Prediction (LPC), Linear Predictive Cepstral Coefficients (LPCC), and Logarithmic Bandwidth Detector (LDB)

For feature extraction, we explored three distinct techniques: LPC, LPCC, and LDB. While LPCC demonstrated test accuracy exceeding train accuracy, suggesting potential overfitting, LPC and LDB were implemented due to their promising performance. Linear Prediction (LPC) involves predicting future samples based on a linear combination of past samples, offering insights into the underlying speech signal characteristics. LDB, or Logarithmic Bandwidth Detector, leverages logarithmic frequency scales, providing an alternative perspective on the spectral content of the speech signal. Both LPC and LDB contribute significantly to the discriminative power of our model, enhancing its ability to accurately identify speakers.

Therefore, to establish a robust model, we opted to incorporate Noise Augmentation, Delta, and Double Delta features in MFCC, along with LPC and LDB in our system thus achieving a final accuracy of 94% in Test.

7 Results

We were able to achieve an accuracy of 94% while testing in the task of Speaker Identification using classical machine learning techniques.

GMM, the model we used, is therefore an effective model for speaker identification tasks.

We concluded that the use of MFCCs is highly useful for feature extraction, with other methods such as Delta, Double Delta, LPC and LBD helpful in improving the accuracy slightly if taken as features as well.

Further, an integral finding during the model training was that addition of background noise improves the model greatly, as well as addition of other voices into the audio of a speaker. This makes the model much more robust, supporting previous findings.

References

[1] Kinnunen, Tomi, Evgeny Karpov, and Pasi Franti. "Real-time speaker identification and verification." IEEE Transactions on Audio, Speech, and Language Processing 14, no. 1 (2005): 277-288.

[2] R. Togneri and D. Pullella, "An Overview of Speaker Identification: Accuracy and Robustness Issues," in IEEE Circuits and Systems Magazine, vol. 11, no. 2, pp. 23-61, Secondquarter 2011, doi: 10.1109/MCAS.2011.941079.

[3] Tirumala, Sreenivas Sremath, Seyed Reza Shahamiri, Abhimanyu Singh Garhwal, and Ruili Wang. "Speaker identification features extraction methods: A systematic review." Expert Systems with Applications 90 (2017): 250-271.

[4] Yuan, Jiahong, and Mark Liberman. "Speaker identification on the SCOTUS corpus." Journal of the Acoustical Society of America 123, no. 5 (2008): 3878.

[5] Auishikpyne. (2022, July 24). Speaker identification. Kaggle. https://www.kaggle.com/code/auishikpyne/speaker-identification

[6] Masoudmzb. (2021, February 24). Gradient Tape Tutorial (Audio Proccesing example). Kaggle. https://www.kaggle.com/code/masoudmzb/gradient-tape-tutorial-audio-proccesing-example

[7] Voice Processing, Speech Recognition . GitHub. (n.d.). https://github.com/A4Ayub/Voice-Processing/blob/master/speech-recognition/speech-recognition-exploratory-data-analysis.ipynb

[8] W. L. L. Phyu and W. P. Pa, "Building Speaker Identification Dataset for Noisy Conditions," 2020 IEEE Conference on Computer Applications(ICCA), Yangon, Myanmar, 2020, pp. 1-6, doi: 10.1109/ICCA49400.2020.9022844.