A logo with a yellow and purple design

Description automatically generated with medium confidenceSpeaker identification

Mahmoud Ayman abdelhamid  
faculty of computers and artificial intelligence ,cairo universityGIZA,Egypt m.aymanabdelhamid@gmail.com

Shreen osama abdelrahman  
faculty of computers and artificial intelligence,cairo universityGIZA,Egypt  
shreenosama3311@gmail.com

Ahmed yasser ahmed meshref  
faculty of computers and artificial intelligence,cairo universityGIZA,Egypt  
ahmrdyasser38@gmail.com

*Abstract*—

Introduction

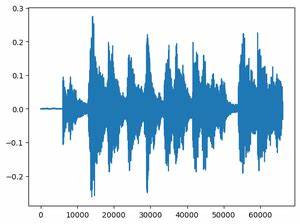
The first thing I want to mention that CNN become one of the most Applied technologies in AI section and make The AI field became famous People wonders in how that a machine can detect the voice gender or voice language detection this things make the regular people ask about what is AI? What is this thing which make this machine can detect our voices!! In our research we implement a speak identification detection model. Voice identification, also known as speaker recognition or voice authentication, is a fascinating and critical field in the domain of biometric authentication and security. It involves the process of verifying or identifying individuals based on their unique vocal characteristics, such as the pitch, tone, cadence, and spectral features of their voice. This technology has a wide range of applications, including access control, fraud prevention, forensic analysis, and improving user experiences in various voice-operated systems, such as virtual assistants and voice-controlled devices. One type of biometric authentication is voice recognition, which refers to dynamic biometric techniques and allows an individual to be identified using a Combination of distinctive voice traits. Technology known as "speaker Recognition" uses the speech waveform to automatically identify the speaker. Voice identification has evolved significantly over the years, benefiting from advances in machine learning and deep learning techniques, particularly Convolutional Neural Networks (CNNs). CNNs, originally developed for image Processing tasks, have found application in the analysis of acoustic features extracted from audio signals for voice identification. They have shown remarkable potential in accurately distinguishing different speakers and enhancing the robustness and security of voice-based authentication systems. This research focuses on the application of CNNs in the field of voice identification. CNNs are a class of deep learning models that have demonstrated exceptional performance in various image-related tasks due to them.

ability to capture hierarchical and discriminative features from complex data. Similarly, the acoustic features extracted from voice signals can be analyzed effectively using CNNs for the purpose of speaker recognition. The objectives of this research are: Investigate the potential of CNNs for voice identification: This study aims to explore the capabilities of CNNs in analyzing voice data and their effectiveness in speaker recognition tasks. We will assess the advantages of using CNNs over traditional methods and other deep learning architectures. Feature extraction and representation: One of the key aspects of voice identification with CNNs is the extraction and representation of relevant acoustic features from the raw audio data. This research will delve into different feature extraction techniques and their impact on the performance of CNN-based voice identification systems. Model architecture and training: We will design and implement a CNN-based model for voice identification and investigate the optimal architecture for this task. Training processes, data augmentation strategies, and hyper parameter tuning will be explored to maximize the model's accuracy and robustness. Evaluation and benchmarking: To assess the effectiveness of CNN-based voice identification, we will conduct comprehensive evaluations and benchmark comparisons with existing methods. Performance metrics such as accuracy, precision, recall, and F1-score will be used to quantify the model's capabilities. Real-world applications: Finally, we will discuss potential real-world applications of CNN-based voice identification, including security and access control systems, voiceactivated devices, and forensic investigations. Modeling a speaker identification algorithm very exciting from a to z. Beginning of collecting datasets (.wav), Encoding it from analog voice signal to digital signal, features extraction using Mel frequency cebstral coefficients (MFCC), Model training using CNN and finally Evaluating the model In our first phase which is feature extraction Following MFCCS Approach, MFCCs are a set of coefficients derived from the power spectrum of an audio signal, focusing on the representation of the spectral envelope of the signal. They are obtained through a multi-step process, involving taking the logarithm of the power spectrum, applying Mel filter banks to the spectrum, computing the Discrete Cosine Transform (DCT) of the logarithm of the filter bank energies, and extracting a subset of these DCT coefficients.

# DataSet

[Speaker Recognition Dataset (kaggle.com)](https://www.kaggle.com/datasets/kongaevans/speaker-recognition-dataset)

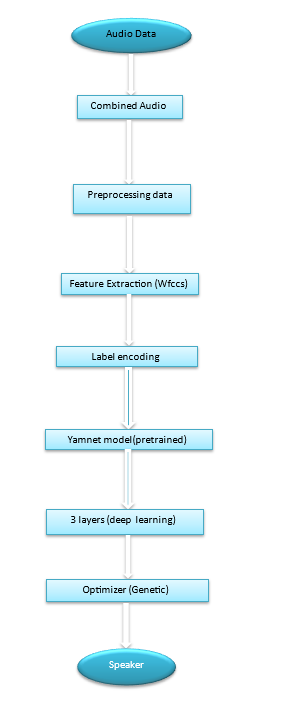
This dataset contains speeches of five prominent leaders namely; Benjamin Netanyahu, Jens Stoltenberg, Julia Gillard, Margaret. Each speaker's speech was originally one long audio file, so I divided them up into manageable chunks of one second each. The chunked audio files from 0.wav to 1500.wav can be combined to create the speaker's entire speech. PCM stands for Pulse Code Modulation, which is a standard method used to digitally represent analog signals, particularly in audio recording and transmission. In PCM, an analog signal, such as sound waves in the case of audio, is sampled at regular intervals. These samples are then quantized to a specific bit depth (e.g., 8-bit, 16-bit, 24- bit), which determines the resolution of the digital representation. The more bits used for quantization, the more accurately the original analog signal can be represented digitally Which are used in our dataset to convert analog voice signal to digital signal.



III.Related work

A sizable body of research on the recognition of voice commands has already been published. These studies cover topics such as machine learning algorithms for textto-voice conversion techniques for teaching voice control and voice conversion techniques. The works provide descriptions of current speech recognition systems. EX: Improved Speaker Identification System Based on MFCC and DMFCC Feature Extraction Technique Publisher: IEEE Speaker Identification (SI) is the process of comprehending an individual's voice through the application of machine learning algorithms. Accurately classifying the speakers in the speaker identification process requires extracting feature information from their utterances. Because of their ability to capture the repetitive nature of speech signals, Mel Frequency cebstral Coefficients (MFCC) and Dynamic Mel Frequency cebstral Coefficients (DMFCC) are utilized as features in many speaker identification systems. The goal of this work is to examine the MFCC and DMFCC coefficients, which are used to enhance speaker system recognition accuracy. The extracted MFCC and DMFCC features were used as input for a Gaussian Mixture Model (GMM) and Bayesian Classifier in order to build the speaker identity model. The GMM and Bayesian .Performance Analysis is done on the GMM and Bayesian classifier performance for varying numbers of mixtures. For MFCC and DMFCC, the GNN Model's maximum accuracy is 82.7% and 80.12%, respectively. The Bayesian classifier performance, on the other hand, is 79.43 for MFCC and 83.92% for DMFCC. The extraction of features and the classification model can be

# IV.purpose model



In our model we use dataset Audio that contains four speakers, for each one 16000 second then we combine audio in 2 minutes That mean that we used 120 second only for each one.

After that we make preprocessing on our data using spectrogram, MFCCS.

Spectrogram: representation of the spectrum of frequencies of a signal as it varies with time. It displays how the frequency content of a signal changes over time.

MFCCS: Feature Extraction: MFCCs are a feature extraction technique widely used in speech and audio processing. They aim to represent the short-term power spectrum of a sound.

After that we make label encoding for classes, after that we spilt data on train, validation and test then build in model yamnet pretrained and make 3 layers and fully connected layer with SoftMax, then we used optimizer genetic with population size=10, number of generations=5

And the final output is speaker who the owner of the voice.

##### V. Experiment result

After applying the genetic om our model, the best accuracy 99.11% for classification on four speakers(dataset)

We split the dataset: 4200 samples for train, 901 samples for test, validation 900.

A screenshot of a computer

Description automatically generated

##### figure for results of confucion matrix

VI. conclusion

##### References

1\_ [https://www.sciencedirect.com/science/article/abs/pii/S0950705118304 8](https://www.sciencedirect.com/science/article/abs/pii/S0950705118304%208)

2\_ [https://www.sciencedirect.com/science/article/abs/pii/S0957417421002 4](https://www.sciencedirect.com/science/article/abs/pii/S0957417421002%204)

3\_ [(PDF) Analysis of Methods and Techniques Used for Speaker Identification, Recognition, and Verification: A Study on Quarter-Century Research Outcomes (researchgate.net)](https://www.researchgate.net/publication/354974533_Analysis_of_Methods_and_Techniques_Used_for_Speaker_Identification_Recognition_and_Verification_A_Study_on_Quarter-Century_Research_Outcomes)

4\_ Herbig , T., & Gerl, W. Minker. (2012). Self-learning speaker identification for enhanced speech recognition. Computer Speech & Language, 26(3), 210–227

5\_\_ <https://www.sciencedirect.com/science/article/abs/pii/S0885230811000635?via%3Dihub>

6\_ <https://www.researchgate.net/publichttps://www.researchgate.net/publication/352046546_Recognition_of_Voice_Commands_Based_on_Neural_Network>

7\_ <https://ieeexplore.ieee.org/document/9616805>

8\_M. McLaren, N. Scheffer, M. Graciarena, L. Ferrer and Y Lei, "Improving speaker identification robustness to highly channel- degraded speech through multiple system fusion", 2013 IEEE international conference on acoustics speech and signal processing, vol. 6, pp. 6773-6777.

9\_ <https://ieeexplore.ieee.org/document/9616805>

10\_

<https://www.bing.com/ck/a?!&&p=b7d3caedb6804db9JmltdHM9MTY5ODc5NjgwMCZpZ3VpZD0xNGE4ZDY0Mi1lNDI5LTYxYmEtMThiZS1jNmEwZTU1NTYwMzQmaW5zaWQ9NTQ2MQ&ptn=3&ver=2&hsh=3&fclid=14a8d642-e429-61ba-18be-c6a0e5556034&psq=application+of+speaker+identification&u=a1aHR0cHM6Ly9hcnhpdi5vcmcvcGRmLzIyMDUuMTQ2NDkjOn46dGV4dD1TcGVha2VyJTIwcmVjb2duaXRpb24lMjBoYXMlMjBtYW55JTIwcmVhbC13b3JsZCUyMGFwcGxpY2F0aW9ucyUyMGluY2x1ZGluZyUyMGJpb21ldHJpYyxhJTIwY3JpbWUlMkMlMjBvciUyMHN1cnZlaWxsYW5jZSUyMGFuZCUyMGF1dG9tYXRpYyUyMGlkZW50aXR5JTIwdGFnZ2luZy4&ntb=1>

11\_ <https://blog.tensorflow.org/2021/03/transfer-learning-for-audio-data-with-yamnet.html>

12\_ <https://www.kaggle.com/models/google/yamnet>

.