UNIVERSITY OF NAIROBI

FACULTY OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND INFORMATION ENGINEERING

PROJECT: AUDIO-BASED AGE AND GENDER PREDICTOR

PROJECT INDEX: 081

NAME: BETT EMMANUEL KIPNGETICH
REGISTRATION NUMBER: F17/2052/2020

SUPERVISORS:

PROF. H.A OUMA

MR. K. WACHIRA

EXAMINER:

This project is submitted in partial fulfilment of the requirement for the award of the Degree of

Bachelor of Science in Electrical and Electronic Engineering at the University of Nairobi

Submitted on [date]

DECLARATION OF ORIGINALITY

NAME OF STUDENT:	Bett Emmanuel Kipngetich
REGISTRATION NUMBER:	F17/2052/2020
FACULTY/SCHOOL/INSTITUTE:	Faculty of Engineering
DEPARTMENT:	Department of Electrical and Information Engineering
COURSE NAME:	Bachelor of Science in Electrical & Electronic Engineering
TITLE OF WORK:	Audio-Based Age and Gender Predictor

1. I understand plagiarism and am aware of the university policy in this regard.
2. This final year project report is my original work and has not been submitted elsewhere for examination, award of a degree, or publication. Where other people's work or my work has been used, this has properly been acknowledged and referenced by the University of Nairobi's requirements.
3. I have not sought or used the services of any professional agencies to produce this work.
4. I have not allowed and shall not allow anyone to copy my work to pass it off as his/her work.
5. I understand that any false claim in respect of this work shall result in disciplinary action, in accordance with the university's anti-plagiarism policy.

Signature: …………………………...   Date: ………………………...

DEDICATION

[DEDICATION]

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABBREVIATIONS & ACRONYMS

Item	Meaning
AI	Artificial Intelligence
CNN	Convolutional Neural Network
DB	Database
dBFS	Decibels relative to Full Scale
FFT	Fast Fourier Transform
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SGD	Stochastic Gradient Descent
STFT	Short-Time Fourier Transform
UI	User Interface
VAD	Voice Activity Detection

ABSTRACT

CHAPTER 1: INTRODUCTION

1.1 Background

Voice**,** a simple yet complex aspect of human communication, has evolved significantly over the years, finding a wide range of applications. From basic voice commands with assistants like Siri and Alexa [1] to biometric security features such as Safaricom's "My Voice is My Password," [2] voice technology has become integral to many modern systems. In telemedicine, voice analysis is starting to play a role in identifying health conditions [3], though this remains an emerging area.

As the potential of voice technology continues to grow, there is increasing interest in pushing its capabilities beyond simple commands. One exciting direction is the use of voice data for demographic predictions, where the analysis of speech components can help identify characteristics such as age and gender [4]. This application is being explored in sectors like targeted marketing, personalized advertising, customer service [5], and healthcare [3], where voice analysis may improve the delivery of care.

The rise of machine learning and artificial intelligence [6] has fueled this progress, allowing for more accurate extraction of insights from voice data. Advanced techniques like deep neural networks and transfer learning have significantly improved the efficiency and effectiveness of these systems [6]. As a result, demographic analysis from voice data has gained momentum, enabling more sophisticated applications in voice recognition and opening up exciting new possibilities for voice technology.

However, while these advancements have shown the potential of voice recognition technology, the shift toward integrated systems capable of predicting both age and gender from audio data remains underexplored.

1.2 Problem Statement

Despite advancements in voice feature extraction techniques and machine learning, current models for predicting age and gender from audio data often face several critical limitations. Firstly, most systems focus on either age or gender prediction separately [7] , lacking an integrated approach that combines both predictions in a single model. Furthermore, these models are typically built without user-friendly interfaces, limiting their usability to technical experts rather than everyday users which is a potential issue [8].

Another significant challenge is the limited diversity in training datasets, which are often biased towards English-speaking populations [9]. This results in models that perform poorly across different languages and demographic groups, particularly in multilingual environments like Kenya, where Swahili and regional languages are common, reducing their applicability in such settings [10]. These datasets also tend to overlook cultural and dialectal variations that influence voice characteristics. Moreover, many systems are designed for high-resource settings, requiring powerful hardware or constant internet access, making them impractical in low-resource environments [11]. To address these issues, a simple, locally deployable interface optimized for basic infrastructure would greatly enhance accessibility and usability in under-resourced areas [12].

Addressing these limitations will ensure the model is inclusive, accessible, and capable of real-world impact in diverse settings.

1.3 Justification

This project is noteworthy as it offers an integrated approach to age and gender prediction from voice data, simultaneously addressing both attributes in a single model. Unlike existing systems that focus on only one prediction [7], this approach will enhance the accuracy and usability of demographic prediction in real-world applications.

Additionally, the project will extend the scope of demographic prediction beyond English-speaking populations by incorporating multiple languages and dialects available in open-source datasets, making the system more relevant to diverse populations. This ensures better inclusivity and performance in multilingual environments.

The system will be made user-friendly, enabling non-technical users to operate it easily. Moreover, it aims to be deployable in low-resource settings, such as rural hospitals, by minimizing the computational power required, making it accessible even in areas with limited infrastructure.

Finally, the project will prioritize the ethical handling of sensitive speech data by incorporating secure data handling methods, ensuring privacy and compliance with data protection standards.

1.4 Objectives

- ### Main Objective:

  To design and implement an audio-based system capable of accurately predicting the age and gender of individuals based on voice input through an intuitive user interface.

- **Specific Objectives:**
  - To investigate and use current machine learning models to achieve an accuracy rate of at least 80% in predicting gender and age from audio samples.
  - To utilize pre-trained machine learning models to minimize the computational burden and time required for model development while ensuring the system can process audio samples from diverse demographics.
  - To design and implement a database for efficient storage and retrieval of voice samples and corresponding demographic predictions, ensuring scalability and accessibility.
  - To design a user-friendly interface (UI**)** that enhances the user experience, making it intuitive for non-technical users to input audio and view demographic predictions.
  - To develop control systems that guarantee accurate processing and timely feedback of audio data to enhance the system's overall performance.
  - To complete the development and thorough testing of the system.

1.5 Scope

The primary focus of this project is to develop an audio-based system that predicts both age groups (child, teen, adult) and binary gender (male, female) from voice data, utilizing open-source datasets for the initial development phase.

The project will involve programming, developing machine learning algorithms for processing voice samples, and setting up a database to store both voice samples and demographic predictions. A user-friendly interface (UI) will be developed to allow easy interaction, where users can input voice samples and receive demographic predictions. Control mechanisms will ensure the smooth integration of the audio processing module with the UI, providing seamless feedback and user experience.

The following restrictions and assumptions are put in place to make the project feasible and doable within the allotted time:

- ✓ **Model Limitations:** This project will leverage pre-existing machine learning models for age and gender prediction instead of developing new models from scratch. This approach will help minimize development time and computational resources required for model training.

- ✓ **Processing Capacity:** The system will be designed to perform well on standard hardware, meaning it should be able to function on devices with average processing power. Real-time audio processing will be optimized for lower computational loads to ensure it is deployable in areas with limited resources
- ✓ **Privacy and Data Handling:** The system will incorporate strict data handling protocols to ensure that sensitive speech data is managed securely, adhering to privacy standards. Access to the database will be restricted to ensure compliance with data protection regulations, particularly given the nature of the data.
- ✓ **Data Availability:** The availability of diverse, multilingual datasets for training and testing the model is assumed. This project will primarily rely on existing datasets that include a variety of demographics, but it is assumed that these datasets will be sufficiently representative to ensure the system works across different age groups, genders, and regional accents.
- ✓ **Deployment Environment:** The system will be designed to function in low-resource settings, such as rural healthcare environments, where internet access might be unreliable, and hardware may be less powerful. The design will prioritize simplicity, ensuring accessibility even with limited infrastructure.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The task of predicting age and gender from voice recordings has become increasingly significant, especially with advancements in machine learning (ML) and artificial intelligence (AI) [6]. These systems leverage the unique acoustic features of human speech to infer demographic attributes such as age and gender. In recent years, research has focused on improving the accuracy of such predictions by exploring various techniques in data collection, preprocessing, feature extraction, model training, and system integration.

This literature review aims to analyze existing methodologies, techniques, and technologies for age and gender prediction from audio data, providing a comprehensive understanding of the approaches that will guide the development of an accurate and efficient prediction system for this project. The insights gained will serve as the foundation for the methodology used in the design and implementation of the prediction system.

2.2 General Overview

The main objective of this project is to predict both age and gender from audio data by analyzing voice recordings. Human speech exhibits certain vocal characteristics that vary according to age and gender, making it possible to infer demographic information from voice recordings. The core task involves applying machine learning algorithms to process and analyze these vocal characteristics, generating accurate demographic predictions.

This project focuses on these key areas:

- **Data Collection and Preprocessing**
- **Feature Extraction**
- **Model Training and Prediction**
- **System Integration (User Interface and Database)**

Each area is critical to the success of the system, ensuring that it performs accurately, efficiently, and remains user-friendly. Below is an overview of these stages and their relevance to the project's objectives.

2.3 Overview of Key Areas

2.3.1 Age and Gender Prediction from Audio Data

Predicting age and gender from audio involves identifying distinct vocal characteristics that differ by age and gender. Various studies have identified pitch, speaking rate, and timbre as features that are particularly sensitive to these attributes [13]. The challenge lies in isolating these features from the broader spectrum of speech, as many factors—such as emotional state, speaking environment, and recording quality—can affect the raw audio signal [14]. Successful systems need to consider these variables to improve the robustness of predictions, especially in real-world environments.

Research into age and gender prediction has evolved with the integration of deep learning techniques, such as neural networks, which can automatically learn these vocal patterns without explicit human intervention [15]. However, the complexity of real-world speech signals [16], including multilingual and multi-accented variations, remains a challenge that requires ongoing innovation in model development.

2.3.2 Data Collection and Preprocessing

Data collection is critical in building robust models for age and gender prediction. Publicly available datasets such as VoxCeleb and TIMIT [17] offer diverse voice samples that include varying demographic characteristics like age and gender. CommonVoice [18], with its multilingual and global coverage, provides additional diversity in accents and languages. Figure 2.1 illustrates the distribution of speaker ages in these datasets, highlighting gaps and strengths in demographic coverage. To ensure the model generalizes well across linguistic groups, diversity in accents, age ranges, and gender balance is essential [19][20].

Preprocessing transforms raw audio data into a consistent and clean input format, addressing challenges such as noise, volume inconsistency, and irrelevant segments [21][22]. Key steps include noise reduction, volume normalization, segmentation into fixed-length windows, and silence removal. These steps enhance the quality of extracted features and ensure that the model focuses on relevant information, reducing overfitting and improving generalization to new datasets.

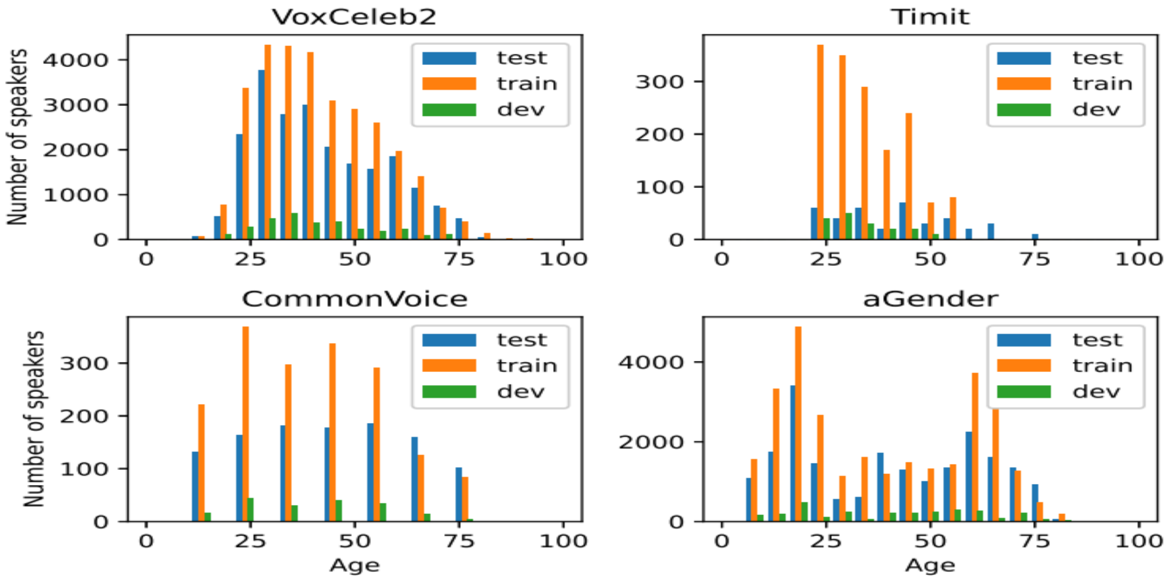


Figure 2.1: *Distribution of speaker age (#samples) in the datasets for the three splits its (CommonVoice age in mid-decade)*

2.3.3 Feature Extraction

Feature extraction transforms audio data into meaningful representations that facilitate accurate demographic predictions. Techniques like **Mel-Frequency Cepstral Coefficients (MFCCs)** are widely used for capturing vocal characteristics, offering noise resilience and computational efficiency [21]. Advanced methods such as **Wav2Vec** enrich the feature set by capturing contextual information from raw audio, improving model performance in noisy environments. Combining these techniques provides a robust feature representation for diverse use cases. **Table 2.1** shows a comparison of various feature extraction methods.

Variability in recordings, such as differences in speaker pitch, rate, or background noise, can affect feature extraction accuracy. Techniques like **Voice Activity Detection (VAD)** and speaker normalization reduce speaker-related inconsistencies, while methods like **band-pass filtering** and **spectral subtraction** mitigate acoustic variability. These processes ensure that extracted features are consistent and representative, leading to more reliable model predictions.

Table 2.1 Comparison of Feature Extraction Methods

Feature Extraction Method	Description	Advantages	Limitations
MFCC	Mimics human auditory perception, capturing speech-relevant frequency features.	Efficient, widely used, effective for gender classification	May not capture complex speech patterns for age prediction
Wav2Vec	Utilizes deep learning to process raw audio waveforms, producing embeddings directly.	Captures deeper contextual features, no need for labeled data	Computationally expensive, requires large datasets
Spectral Features	Extracts pitch, timbre, and temporal variations like centroid and roll off.	Captures dynamic speech characteristics	May require additional features for reliable age and gender prediction

2.3.4 Model Training and Prediction

Model training and prediction are crucial stages in building a machine learning system for age and gender classification from audio. These processes involve transforming preprocessed audio data into meaningful demographic predictions.

Model Training begins with feeding labeled data into a machine learning model, where the aim is to optimize its parameters for accurate predictions. Various architectures, including Convolutional Neural Networks (CNNs) for spatial feature extraction, Recurrent Neural Networks (RNNs) for sequential data, and hybrid CNN-LSTM models, are commonly used due to their ability to combine spatial and temporal analysis. The training process incorporates optimization techniques (e.g., SGD, Adam) and loss functions tailored to classification or regression tasks. Evaluation metrics such as accuracy (for gender) and RMSE (for age) are used to measure model performance.

Prediction involves applying the trained model to new, unseen data. The pipeline consists of preprocessing the input, extracting features, and feeding them into the trained model to generate predictions, such as gender labels or approximate age groups. Generalization is critical, and maintaining consistency in preprocessing and feature extraction across training and prediction phases is key to achieving accurate results.

Additional details on these processes can be found in Appendix 2, which includes a detailed breakdown of algorithms, evaluation metrics, and training strategies. Table 2.2 provides a summary of commonly used models, describing their applications and limitations, while Figure 2.2 visually illustrates the entire workflow from raw audio to model prediction and output generation.

Table 2.2: Model Architectures

Model	Description	Applications	Limitations
CNN	Captures spatial features in audio representations (e.g., spectrograms or MFCCs).	Gender classification.	Limited in capturing temporal data dependencies.
RNN with LSTM	Analyzes sequential data, ideal for capturing temporal variations in speech.	Age range predictions.	Computationally expensive.
Pre-Trained Models	Leverages architectures like ResNet or Wav2Vec, fine-tuned for specific tasks.	High-performance systems.	Requires advanced understanding of transfer learning.

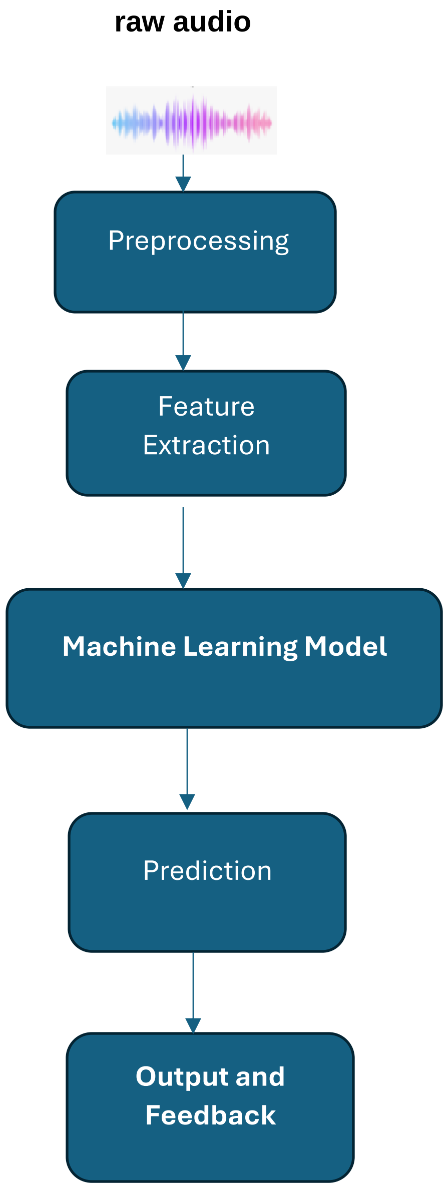


Figure 2.2: *Flow of Audio-Based Prediction System*

2.4 Review of Related Work

2.4.1 Data Collection and Preprocessing

Data diversity and quality are fundamental for training models capable of accurate age and gender predictions. The reviewed studies employed a range of datasets and preprocessing techniques to standardize audio data and minimize demographic bias.

- **Kwasny and Hemmerling [28]** utilized VoxCeleb1, Common Voice, and TIMIT datasets, leveraging their varied demographic coverage. VoxCeleb1 was used for pre-training, while Common Voice and TIMIT were fine-tuned for age and gender tasks. Preprocessing steps included **Voice Activity Detection (VAD)** to isolate relevant speech, random cropping to standardize audio length at 5 seconds, and **volume normalization** to enhance feature consistency. **Figure 2.4** visualizes the age and gender distribution in these datasets, showcasing their wide coverage of speaker demographics.

- **Shameem et al.,[29]** relied on the Common Voice dataset, focusing on English-language recordings. They applied basic noise reduction and resampled audio to 16 kHz to improve consistency. While useful for general voice analysis, its limited diversity in accents and age groups posed challenges for broader generalization.

- **Kandasamy and Bera [30]** adopted a comprehensive approach, combining TIMIT, VCTK, NISP, and GMU datasets for a diverse range of voices, accents, and demographics. Their preprocessing pipeline included **VAD**, padding, and cropping to a uniform 5-second duration, optimizing computational efficiency while maintaining feature richness. As illustrated in **Figure 2.3**, this process transitions audio from raw form (A) through speech-focused VAD (B), cropped segments (C), and final normalized output (D), ensuring consistent quality for effective feature extraction.

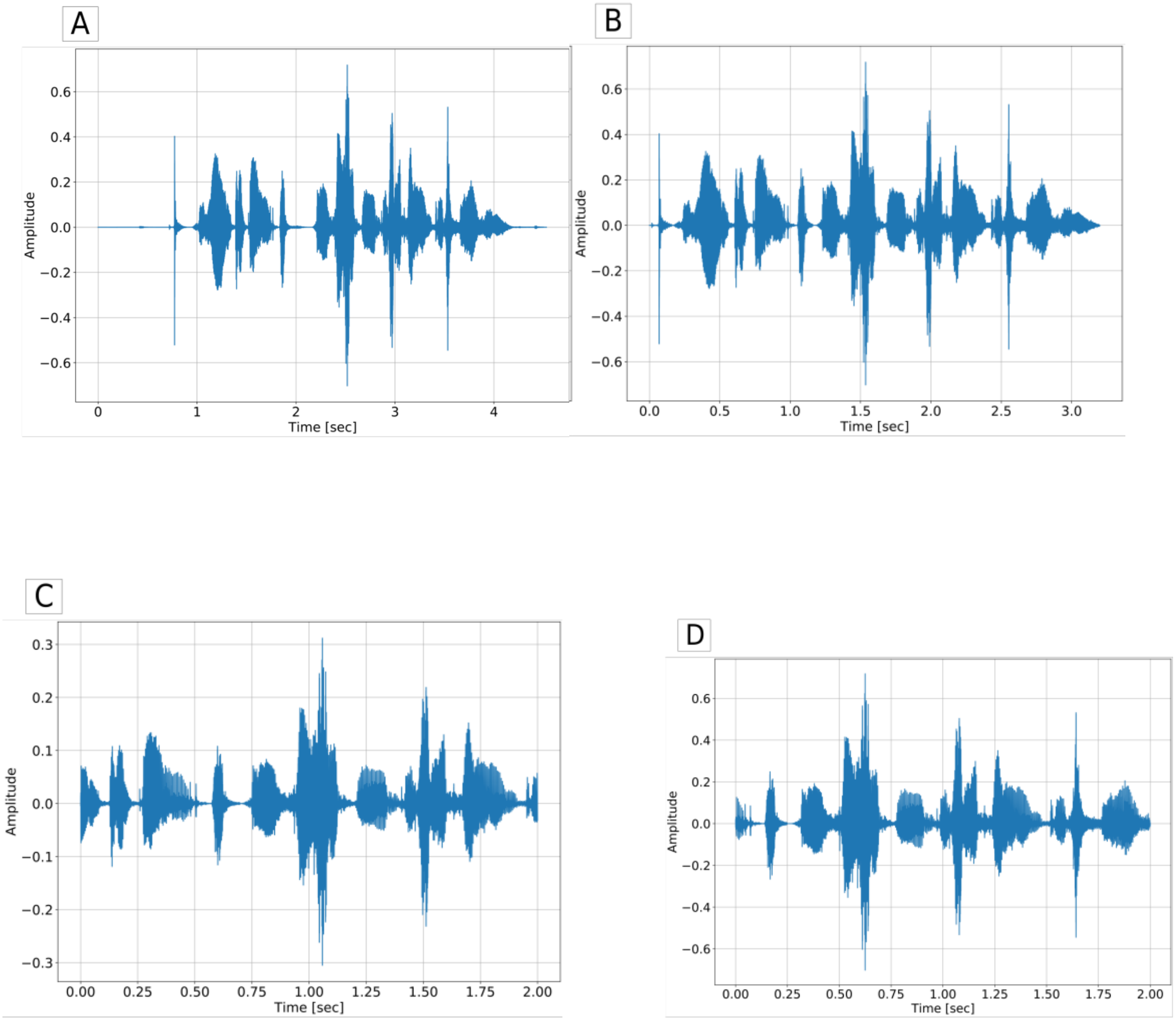


Figure 2.3. *Different waveform preprocessor stages illustrated. (A) unprocessed waveform, (B) waveform after VAD, (C) waveform after random cropping 2 s of the VAD output, (D) applying dBFS normalization to the level of −30 dB to the output of stage C.*

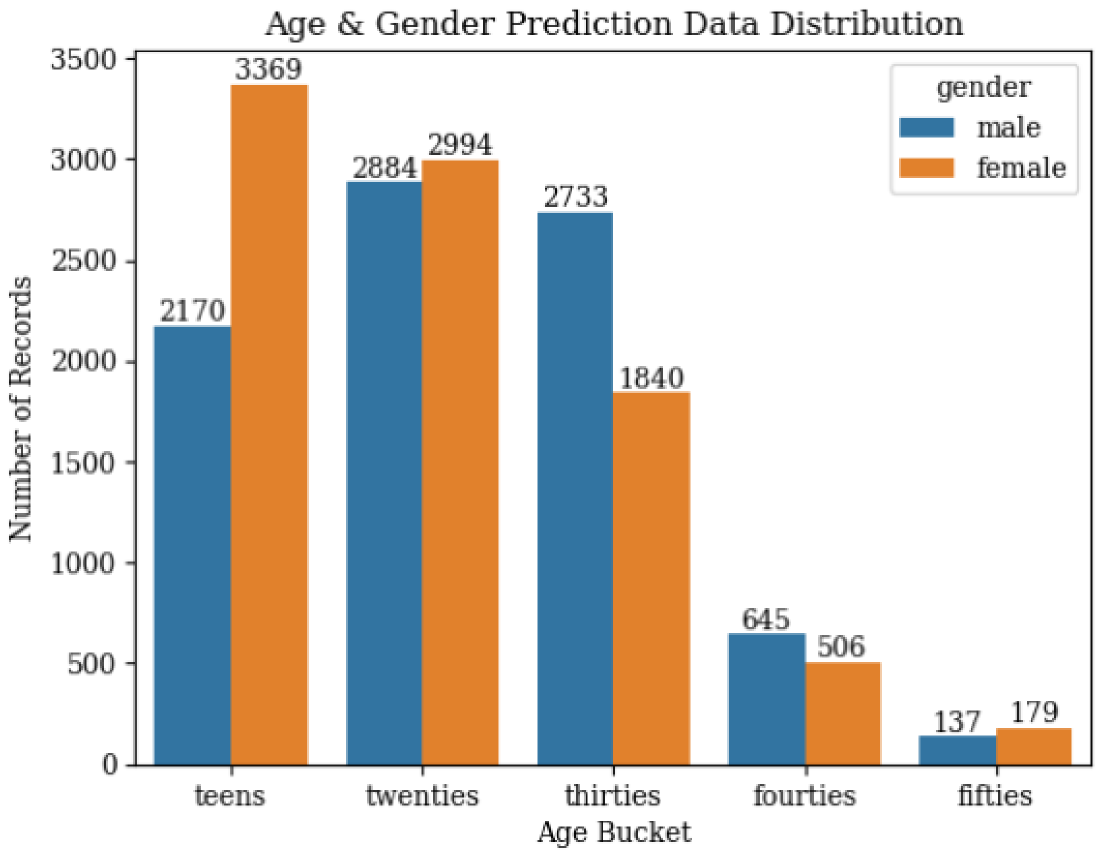


Figure 2.4. *Age & Gender Dataset Distribution*

2.4.2 Feature Extraction Techniques

Feature extraction is crucial in transforming raw audio signals into informative features that facilitate accurate demographic predictions. Each of the reviewed studies employed unique feature extraction techniques to capture essential characteristics of voice data.

- **Kwasny and Hemmerling [28]** explored **MFCCs** with an extended 30-dimensional setup for noise resilience and **Wav2Vec** for deeper contextual insights. Wav2Vec captured nuanced speech subtleties, improving prediction accuracy in variable noise conditions. The combination of MFCC and Wav2Vec enriched their feature set, enhancing model robustness in challenging environments.

- **Shameem et al., [29]** relied primarily on **MFCCs**, complemented by **spectral features** like centroid and bandwidth, offering a balanced feature set for moderate accuracy. However, the simpler feature extraction limited adaptability to real-world complexities.

- **Kandasamy and Bera [30]** integrated **MFCCs and Wav2Vec**, leveraging the latter's deep-learning-based contextual representation for robustness in noisy settings. Their dual approach enriched inputs for their CNN+LSTM model, yielding improved accuracy and generalizability in diverse scenarios.

2.4.3 Model Architectures and Designs

The architectural design of a model plays a significant role in determining its prediction accuracy and computational efficiency. Each study adopted distinct model architectures tailored to their specific goals in age and gender prediction.

- **Kwasny and Hemmerling [28]** evaluated x-vector, QuartzNet, and d-vector architectures. QuartzNet, with residual connections, excelled in handling short-duration speech, achieving 99.6% accuracy in gender classification and low RMSE values for age prediction. Its integration of CNNs captured spatial voice features, while its temporal capabilities enhanced demographic insights.

- **Shameem et al., [29]** implemented a straightforward CNN, emphasizing ease of real-time implementation and computational efficiency. While achieving ~70% accuracy for age group and gender predictions, the model lacked the sophistication for precise age predictions.

- **Kandasamy and Bera [30]** employed a **hybrid CNN+LSTM** model, combining CNNs for spatial feature extraction and LSTMs for sequential processing. This design achieved 99.5% gender classification accuracy and an RMSE of 4.1 years for age prediction, demonstrating strong performance in varied demographic scenarios.

2.4.4 Performance Metrics

Each study used performance metrics to assess the performance of their models, providing a quantitative basis for comparing effectiveness across age and gender prediction tasks.

- **Gender Classification Accuracy** was a primary metric in all studies, with Kwasny and Hemmerling [28] achieving the highest accuracy at 99.6% using their hybrid model architectures.
- **Age Prediction Metrics** varied by study focus. Kwasny and Hemmerling [28] utilized Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for accurate age prediction, achieving an MAE of 5.12 years for males and 5.29 years for females. Kandasamy and Bera [30], using similar metrics, achieved an RMSE of 4.1, reflecting their model's precise age prediction capability. Shameem et al., [29] in contrast, concentrated on age group accuracy, which limited the precision of their age-related insights.

Table 2.3: Comparison of Feature Extraction Techniques, Model Architectures, and Performance Metrics in Age and Gender Prediction from Audio

Study	Feature Extraction Techniques	Model Architecture	Performance Metrics	Strengths	Limitations
Kwasny & Hemmerling [28]	MFCC (30-dim), Wav2Vec	QuartzNet x-vector, d-vector (LSTM)	Gender Accuracy: 99.6% Age MAE: 5.12 (male), 5.29 (female)	Robust noise resilience multi-stage learning	Complex model; higher computational needs
Shameem et al., [29]	MFCC, Spectral Features (Centroid, Bandwidth)	CNN	Age Group Accuracy: ~70%	Lightweight and real-time application	Limited precision in age prediction
Kandasamy & Bera [30]	MFCC, Wav2Vec	Hybrid CNN+LSTM	Gender Accuracy: 99.5% Age RMSE: 4.1	Effective in diverse scenarios	Higher resource requirements

2.4.5 Conclusion

The literature review has addressed the key areas of this project, providing a solid basis for the proposed approach in the next chapter. It covered data collection and preprocessing, highlighting techniques for sourcing diverse datasets, noise reduction, and feature scaling to ensure robust input quality. Feature extraction methods such as MFCCs and Wav2Vec were analyzed for their ability to capture critical demographic traits like pitch and timbre. Various model architectures, including CNNs, RNNs, and hybrid CNN-LSTM models, were reviewed, focusing on their strengths in feature learning and their performance metrics like accuracy and RMSE. Existing methodologies were examined to identify gaps, including limited dataset diversity and challenges with noise resilience and real-time optimization. This review lays the groundwork for a proposed system that will leverage the most effective techniques, aligning with the specific requirements of this project to enhance performance.

CHAPTER 3: DESIGN

3.1 Introduction

This chapter outlines the design approach for the audio-based age and gender prediction system. The design is informed by the literature review, aiming to meet the project objectives while optimizing performance, accuracy, and usability. This chapter will cover the system architecture, core functionalities, preprocessing steps, model prediction workflow, user interface design, backend integration, and the database structure. It will also include details on the required hardware and software specifications to implement the system, as well as potential challenges that may arise during development.

3.2 System Overview

The proposed system is designed to predict the age and gender of an individual based on audio input. The process involves the following key steps:

a) **Audio Input**: The system receives an audio file as input from the user.

b) **Preprocessing**: The audio input is processed to standardize its length and enhance its quality by reducing noise.

c) **Feature Extraction**: Relevant features such as Mel-frequency cepstral coefficients (MFCCs), pitch, and zero-crossing rate are extracted from the audio signal.

d) **Model Prediction**: These features are passed to pre-trained machine learning models to predict the age and gender of the speaker.

e) **Evaluation**: The system computes the prediction's accuracy, errors, and confidence metrics.

f) **Output**: The system displays the predicted age, gender, and confidence levels to the user via a graphical user interface (GUI).

Below is a flowchart that outlines this overall process, illustrating the major components of the system.

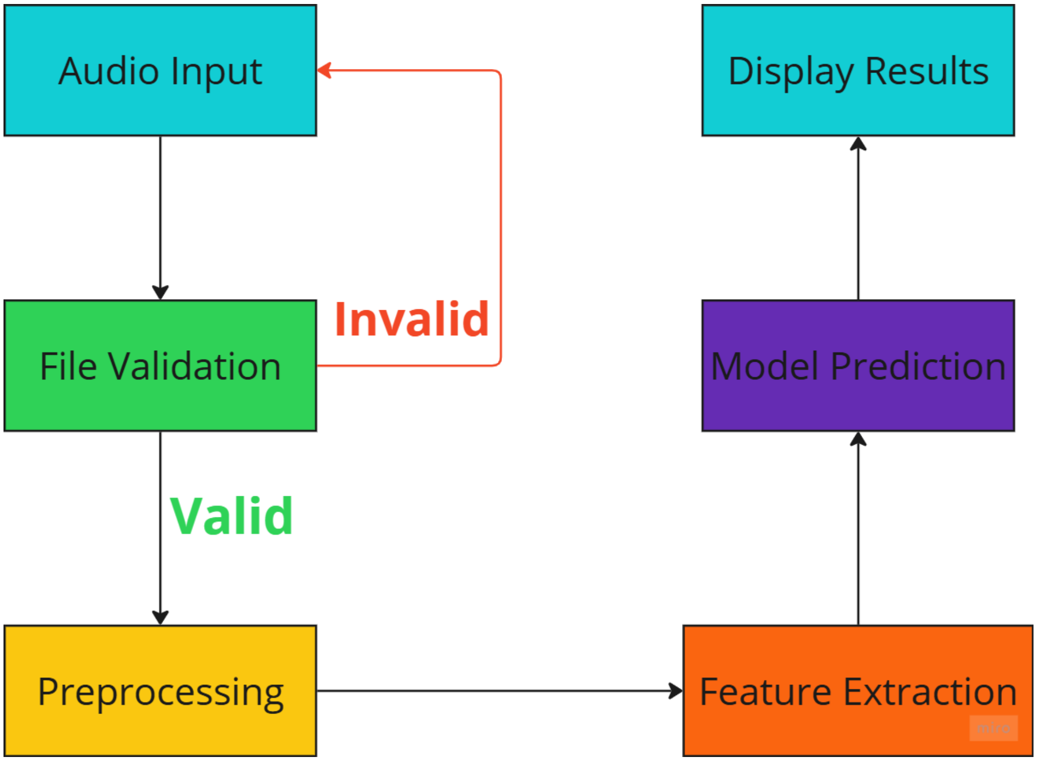


Figure 3.1. *System Architecture Block Diagram*

3.3 Core Functionalities

3.3.1 Preprocessing Audio Input

Preprocessing is a critical step that prepares the raw audio data for feature extraction. The following processes are carried out during preprocessing:

- ✓ **Resampling**: The audio is resampled to a consistent sample rate to ensure uniformity.
- ✓ **Trimming or Padding**: To ensure that all audio files have a consistent duration, audio files are trimmed or padded to 5 seconds.
- ✓ **Noise Reduction**: Various noise reduction techniques are applied to clean the audio, improving the quality of the features that are extracted.

This step is crucial to ensure that the features extracted from the audio are of high quality, minimizing distortions that could affect the accuracy of the predictions.

3.3.2 Feature Extraction

Feature extraction plays a significant role in machine learning for audio analysis. The following features are extracted from the preprocessed audio to serve as inputs for the prediction models:

i. **Mel-frequency cepstral coefficients (MFCCs)**: These coefficients represent the power spectrum of the audio signal and are commonly used in speech and audio processing.

ii. **Chroma Features**: These features describe the harmonic content of the signal and are useful for music and speech analysis.

iii. **Spectral Contrast**: This feature captures the difference in amplitude between peaks and valleys in a sound spectrum, providing information about the texture of the audio signal.

iv. **Zero-crossing rate**: This feature reflects the number of times the signal crosses zero, giving insights into the noisiness of the audio.

v. **Energy and Entropy of Energy**: These features describe the loudness and variation in loudness of the signal, offering valuable information for speech recognition tasks.

vi. **Harmonic-to-noise ratio (HNR)**: HNR measures the ratio of harmonic sound to noise in the signal, indicating the clarity of the voice.

vii. **Pitch and Fundamental Frequency (F0)**: These features capture the pitch and tone of the speaker, which are essential for differentiating age and gender.

Additional features may be considered during model training to optimize performance further, depending on the results from initial experiments.

3.3.3 Model Prediction

The system will use two separate machine learning models to predict age and gender based on the extracted features:

1. **Age Prediction Model**: This model predicts the age of the speaker based on the audio features.
2. **Gender Prediction Model**: This model predicts the gender of the speaker using the same set of features.

Both models will be trained using a common dataset, and the predictions will be combined into a single output, which will be displayed on the user interface. The prediction results will include the predicted age, gender, and associated confidence levels, helping the user gauge the reliability of the predictions.

(Insert flowchart: This should depict the flow of data through preprocessing, feature extraction, and model prediction stages.)

3.4 Supporting Features

3.4.1 Graphical User Interface (GUI)

The graphical user interface (GUI) is the front-facing component of the system, allowing users to interact with the system. The key features of the GUI will include:

- **Audio Upload Button**: A button that allows users to upload audio files for processing.
- **Prediction Display**: The predicted age and gender will be displayed in real-time, along with confidence levels (e.g., "Male, Age: 25, Confidence: 95%").
- **Error Warnings**: If the audio file does not meet the required format or length, the system will display appropriate warnings (e.g., "Audio file too short").
- **Clear and Intuitive Layout**: The layout will be designed for ease of use, ensuring a smooth user experience.

(Insert GUI Mockup: A diagram showing the layout of the upload button, prediction output area, and error messages.)

3.4.2 Backend Workflow

The backend system will handle the core logic of the system, integrating the features from the user interface with the prediction models and the database. The backend will perform the following tasks:

- **Audio Input Processing**: Receive the audio file from the GUI, validate its format, and ensure it meets the necessary criteria.
- **Preprocessing**: Apply the resampling, trimming, padding, and noise reduction processes to prepare the audio for feature extraction.
- **Feature Extraction**: Extract the relevant audio features from the preprocessed audio.

- ❖ **Prediction**: Pass the extracted features to the trained models for age and gender prediction.
- ❖ **Results Display**: The predictions, along with the confidence levels, will be sent to the GUI for user display.

The backend will be built using Python and frameworks such as Flask or FastAPI, which will handle API requests and ensure smooth communication between the components.

(Insert Backend Interaction Diagram: Show the flow between the GUI, preprocessing, feature extraction, models, and output.)

3.4.3 Database

A database will be used to store essential system data, including:

- ✓ **User Inputs and Predictions**: Information about the uploaded audio files, extracted features, and the corresponding predictions will be stored.
- ✓ **System Logs**: Logs related to system performance, including errors and prediction metrics, will be recorded for debugging and analysis.
- ✓ **Training Data**: A copy of the training data, including the labeled audio files, will be stored for future use in model retraining or fine-tuning.

The database will be implemented using a lightweight system like SQLite, ensuring minimal overhead for system operations.

3.5 Data Collection

Data collection is a crucial step in building and training a machine learning model. For the audio-based age and gender prediction system, a diverse and representative dataset is essential to ensure that the model performs well across various age groups, genders, and audio qualities. The following steps outline the data collection process:

Source Selection: The dataset for training the models will be sourced from publicly available audio datasets, such as the **VoxCeleb** or **CommonVoice**, which provide labeled audio samples that include both age and gender information.

Data Augmentation: To enrich the dataset, audio samples may undergo augmentation techniques such as pitch shifting, time-stretching, and adding noise, which will help in building a more robust model by simulating a variety of real-world scenarios.

Data Labeling: The audio data must be labeled with accurate age and gender information. If the source dataset does not provide this, manual labeling will be performed by experts to ensure the accuracy of the training data.

Data Preprocessing: Once the data is collected, it will be cleaned and preprocessed to ensure that it is suitable for feature extraction. This may involve:

Removing unwanted noise or irrelevant sections.

Standardizing the format (e.g., sample rate, mono/stereo).

Trimming or padding the audio clips to ensure uniform length for consistency in model training.

Data Storage: The processed and labeled audio data will be stored in a structured format in the system database or file storage, ready to be used for training the model.

Once the data collection phase is complete, the system will be able to proceed with feature extraction and model training.

3.6 System Requirements and Specifications

3.6.1 Hardware Requirements

- **RAM**: At least 8 GB (16 GB recommended) to handle large datasets and run machine learning models.
- **Storage**: A minimum of 100 GB for datasets, models, and system logs.
- **Processor**: Multi-core processor (e.g., Intel i7 or equivalent) to support fast processing and model inference.
- **GPU**: A dedicated GPU with at least 4 GB VRAM (e.g., NVIDIA GTX 1650 or better) is recommended for model training and inference acceleration.

3.6.2 Software Requirements

a) **Programming Language**: Python is selected for its wide support for machine learning and audio processing libraries.

b) **Libraries**:

   i. **Librosa**: For audio preprocessing and feature extraction.

   ii. **TensorFlow/Keras**: For model training and inference.

   iii. **Flask/FastAPI**: For backend development, handling API requests between the frontend and models.

   iv. **Tkinter/PyQt**: For GUI development.

c) **Operating System**: Windows (preferred for compatibility with various development tools).

CHAPTER 4: IMPLEMENTATION

CHAPTER 5: RESULTS AND ANALYSIS

CHAPTER 6:CONCLUSIONS & RECOMMENDATIONS

REFERENCES

[1] (PDF) Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants, ResearchGate. https://www.researchgate.net/publication/322456429_Alexa_Siri_Cortana_and_More_An_Introduction_to_Voice_Assistants

[2] W. Ngumi, “Mobile-based voice biometric identity: an emerging technology that could assist vulnerable populations,” Mobile for Development, Oct. 2019. https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-for-development/programme/digital-identity/mobile-based-voice-biometric-identity-an-emerging-technology-that-could-assist-vulnerable-populations/ (accessed Nov. 07, 2024).

[3] J. Rusz, “Detecting speech disorders in early Parkinson’s disease by acoustic analysis,” Cvut.cz, 2018, doi: http://hdl.handle.net/10467/78544.

[4] A. A. Abdulsatar, V. V. Davydov, V. V. Yushkova, A. P. Glinushkin, and V. Y. Rud, “Age and gender recognition from speech signals,” Journal of Physics: Conference Series, vol. 1410, p. 012073, Dec. 2019, doi: https://doi.org/10.1088/1742-6596/1410/1/012073.

[5] S. Tao, L. Han, H. Cao, J. Wang, J. Li, and Y. Wang, “The Development of a Speech Recognition and Semantic Understanding System for Marketing Customer Service,” The Development of a Speech Recognition and Semantic Understanding System for Marketing Customer Service, pp. 457–463, Aug. 2024, doi: https://doi.org/10.1109/icsece61636.2024.10729529.

[6] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, “Age group classification and gender recognition from speech with temporal convolutional neural networks,” Multimedia Tools and Applications, vol. 81, Jan. 2022, doi: https://doi.org/10.1007/s11042-021-11614-4.

[7] L. M. David, “Age Prediction by Voice Using Deep Learning,” Upc.edu, Jan. 2023, doi: http://hdl.handle.net/2117/386585.

[8] F. Javier, “The Barriers for Implementing AI,” Apress eBooks, pp. 85–110, Jan. 2023, doi: https://doi.org/10.1007/978-1-4842-9669-1_4.

[9] D. Galvez et al., “The People’s Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage,” arXiv.org, 2021. https://arxiv.org/abs/2111.09344 (accessed Nov. 15, 2024).

[10] A. Torralba and A. A. Efros, “Unbiased Look at Dataset Bias,” IEEE Xplore, Jun. 01, 2011. https://ieeexplore.ieee.org/document/5995347

[11] I. Stoica et al., “A Berkeley View of Systems Challenges for AI,” arXiv:1712.05855 [cs], Dec. 2017, Available: https://arxiv.org/abs/1712.05855

[12] C. Peñaranda, C. Reaño, and F. Silla, “Exploring the Use of Data Compression for Accelerating Machine Learning in the Edge with Remote Virtual Graphics Processing Units,” Concurrency and Computation: Practice and Experience, Oct. 2022, doi: https://doi.org/10.1002/cpe.7328.

[13] T. F. Cleveland, “Acoustic Properties of Voice Timbre Types and Their Influence on Voice Classification,” The Journal of the Acoustical Society of America, vol. 61, no. 6, pp. 1622–1629, Jun. 1977, doi: https://doi.org/10.1121/1.381438.

[14] A. Al-Talabani, H. Sellahewa, and S. A. Jassim, “Emotion recognition from speech: tools and challenges,” SPIE Proceedings, May 2015, doi: https://doi.org/10.1117/12.2191623.

[15] G. Levi and T. Hassner, “Age and Gender Classification Using Convolutional Neural Networks,” www.cv-foundation.org, 2015. https://www.cv-foundation.org/openaccess/content_cvpr_workshops_2015/W08/html/Levi_Age_and_Gender_2015_CVPR_paper.html

[16] S. Basak et al., “Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems,” Computer Modeling in Engineering & Sciences, vol. 135, no. 2, pp. 1053–1089, 2023, doi: https://doi.org/10.32604/cmes.2022.021755.

[17] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, “VoxCeleb: Large-scale Speaker Verification in the Wild,” Computer Speech & Language, p. 101027, Oct. 2019, doi: https://doi.org/10.1016/j.csl.2019.101027.

[18] R. Ardila et al., “Common Voice: A Massively Multilingual Speech Corpus,” arXiv.org, Mar. 05, 2020. https://arxiv.org/abs/1912.06670

[19] Y. Bensoussan, J. Pinto, M. Crowson, P. R. Walden, F. Rudzicz, and M. Johns, “Deep Learning for Voice Gender Identification: Proof-of-concept for Gender-Affirming Voice Care,” The Laryngoscope, Nov. 2020, doi: https://doi.org/10.1002/lary.29281.

[20] N. Hossain and M. Naznin, “Finding Emotion from Multi-lingual Voice Data,” 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 408–417, Jul. 2020, doi: https://doi.org/10.1109/compsac48688.2020.0-214.

[21] J. Meyer, L. Dentel, and F. Meunier, “Speech Recognition in Natural Background Noise,” PLoS ONE, vol. 8, no. 11, Nov. 2013, doi: https://doi.org/10.1371/journal.pone.0079279.

[22] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, “An Overview of Noise-Robust Automatic Speech Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 745–777, Apr. 2014, doi: https://doi.org/10.1109/taslp.2014.2304637.

[23] B. Milner and J. Darch, “Robust Acoustic Speech Feature Prediction from Noisy Mel-Frequency Cepstral Coefficients,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 2, pp. 338–347, Feb. 2011, doi: https://doi.org/10.1109/tasl.2010.2047811.

[24] S. Sadhu et al., “Wav2vec-C: A Self-supervised Model for Speech Representation Learning,” arXiv.org, 2021. https://arxiv.org/abs/2103.08393 (accessed Nov. 20, 2024).

[25] C. Verma, Zoltán Illés, and V. Stoffová, “Age Group Predictive Models for the Real Time Prediction of the University Students Using Machine Learning: Preliminary Results,” Feb. 2019, doi: https://doi.org/10.1109/icecct.2019.8869136.

[26] T. O. Hodson, “Root-mean-square Error (RMSE) or Mean Absolute Error (MAE): When to Use Them or Not,” Geoscientific Model Development, vol. 15, no. 14, pp. 5481–5487, Jul. 2022, Available: https://gmd.copernicus.org/articles/15/5481/2022/

[27] P. V. L. Narasimha Rao and S. Meher, “ORG-RGRU: an Automated Diagnosed Model for Multiple Diseases by Heuristically Based Optimized Deep Learning Using speech/voice Signal,” Biomedical Signal Processing and Control, vol. 88, p. 105493, Oct. 2023, doi: https://doi.org/10.1016/j.bspc.2023.105493.

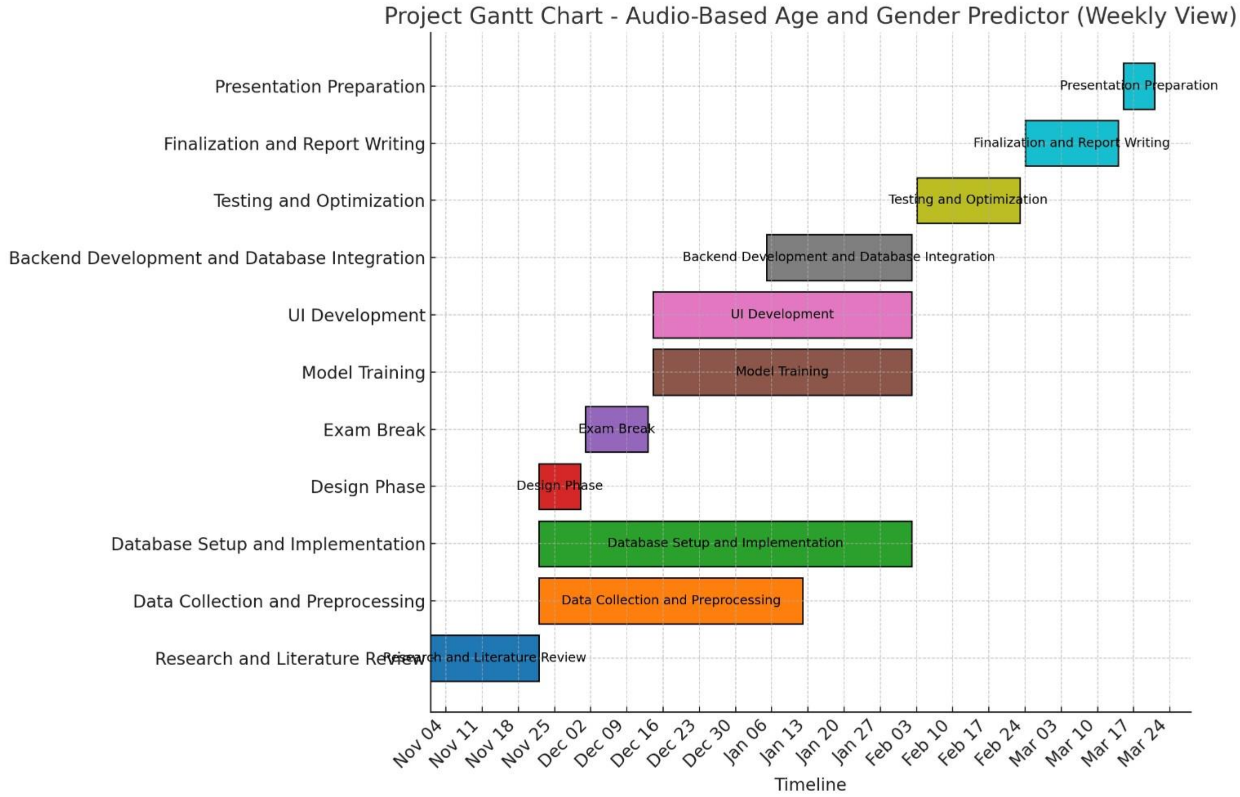
[28] D. Kwasny and D. Hemmerling, “Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks,” Sensors, vol. 21, no. 14, p. 4785, Jul. 2021, doi: https://doi.org/10.3390/s21144785.

[29] V. V. Kandasamy and A. Bera, “Improving Robustness of Age and Gender Prediction Based on Custom Speech Data,” 8th International Conference on Signal, Image Processing and Embedded Systems (SIGEM 2022), vol. 12, no. 20, pp. 69–83, Nov. 2022, doi: https://doi.org/10.5121/csit.2022.122005.

[30] Muhammed Shameem P, M. Faheem, Muhammed, N. Ashraf, and S. Shaharyar, “Age And Gender Prediction From Human Voice For Customized Ads In E-Commerce,” International Journal of Engineering Research & Technology, vol. 11, no. 1, Jun. 2023, doi: https://doi.org/10.17577/ICCIDT2K23-103.

APPENDICES

Appendix A: Gantt Chart



Appendix 2: Model Training and Prediction

Model Training

Model training involves feeding preprocessed and feature-extracted audio data into machine learning algorithms to learn patterns and relationships between audio features and demographic attributes like age and gender. The primary objective is to optimize the model for accurate predictions on unseen data.

Key Machine Learning Algorithms

Several machine learning algorithms [25] have been employed in age and gender prediction tasks, each offering specific advantages based on the nature of the data.

- **Convolutional Neural Networks (CNNs)**: CNNs are highly effective for spatial data processing, and their ability to extract local patterns from data makes them suitable for feature extraction tasks in image processing. In speech-based tasks, they are used to learn patterns from spectrograms or other 2D feature maps derived from audio signals. CNNs can automatically learn hierarchical features from raw data, which makes them suitable for tasks involving high-dimensional feature spaces, such as audio classification.

- **Recurrent Neural Networks (RNNs)**: RNNs excel at handling sequential data, as they can retain information about previous inputs in their internal state. In the context of audio, which is inherently sequential, RNNs can capture time-dependent relationships in the audio features. However, RNNs can suffer from issues like vanishing gradients when processing long sequences.

- **Long Short-Term Memory Networks (LSTMs)**: LSTMs are a specialized type of RNN designed to overcome the limitations of vanilla RNNs, particularly in handling long-term dependencies. By using memory cells, LSTMs retain relevant information for extended periods, which is particularly useful in processing audio sequences where context from earlier parts of the signal is important for understanding the speech.

- **Hybrid Models (CNN-LSTM)**: Some of the most effective models for age and gender prediction combine CNNs with LSTMs. This hybrid approach leverages CNNs to extract features from the raw audio signal (such as spectral features or Mel-spectrograms) and LSTMs to model the temporal dependencies in the data. The combination of spatial feature extraction (from CNNs) and temporal sequence learning (from LSTMs) allows these models to outperform individual CNN or RNN architectures.

Model Training Process

Once the model architecture is chosen, the training process involves feeding the labeled data (i.e., audio features with known age and gender) into the model and adjusting the model’s parameters to minimize the prediction error. This process involves:

- **Loss Function**: The choice of loss function depends on the type of prediction. For gender prediction, a **binary cross-entropy loss** function is commonly used, while for age prediction, a **mean squared error (MSE)** or **root mean square error (RMSE)** function is preferred. [26]

- **Optimization**: During training, optimization algorithms such as **Stochastic Gradient Descent (SGD)**, **Adam**, or **RMSprop** are used to adjust the model parameters and minimize the loss function. These optimizers update the weights of the model iteratively based on the gradient of the loss function.[27]

- **Evaluation Metrics**: The model’s performance is evaluated using different metrics, depending on the type of prediction task:

    - **Classification Accuracy**: For gender prediction, accuracy is a common metric used to assess how many predictions match the true gender labels.

    - **Root Mean Square Error (RMSE)**: For age prediction, RMSE is used to measure the model's prediction error, with a lower RMSE indicating better performance.

Prediction

Once the model is trained, it can be used to make predictions on new, unseen data. During prediction, the model processes the extracted features from the input audio and outputs the predicted age and/or gender.

Prediction Workflow

- **Feature Extraction:** Incoming audio undergoes the same preprocessing and feature extraction as used during training, ensuring consistency and avoiding data mismatches [27].

- **Model Inference**: Extracted features are passed into the trained model, which uses its learned parameters to generate predictions. For gender prediction, the output is a class label (e.g., male or female), while age prediction produces a numerical value for the estimated age [27].
- **Handling New Data**: Model generalization relies on diverse training data and consistent preprocessing pipelines. Variations in feature extraction during prediction can degrade accuracy, underscoring the importance of uniformity across all stages [26].

Model Tuning

After initial predictions, the model may be tuned to improve its performance further. Hyperparameter optimization techniques such as **grid search** or **random search** can be employed to find the best combination of hyperparameters (e.g., learning rate, number of layers, hidden units, etc.) to improve prediction accuracy.

Additional Details

Refer to **Figure 2.2** for a visual representation of the workflow from raw audio preprocessing to prediction and feedback, and **Table 2.3** for a comparative summary of models, their descriptions, applications, and limitations.