Speaker Age & Gender Estimation from Voice

CONVERSATIONAL AI: SPEECH PROCESSING AND SYNTHESIS (UCS749)

Submitted by:

Dilpreet Singh (102203769)

Sayiam Handa (102203777)

Akshita Pathak (102203796)

Aarushi Bajaj (102203820)

B.E. Third Year COE

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Course Instructor:

Dr. Komal Bharti

Department of Computer Science & Engineering

Thapar Institute of Engineering and Technology , Patiala
(147004)

January - May 2025

Contents

1. Introduction

Speech is one of the most natural and efficient forms of human communication. Applications such as virtual assistants, customer profiling, personalised services, and voice analytics have made it more important than ever to analyse vocal characteristics to extract speaker information, such as age and gender, thanks to advancements in speech processing and conversational AI.

In contrast to simply classifying speakers into predefined categories, this project focusses on automatic gender and age recognition from voice, where a speaker's gender (e.g., male or female) and age are predicted. Mel Frequency Cepstral Coefficients (MFCCs), among other acoustic and prosodic features, are extracted from audio signals and fed into machine learning and deep learning models as part of this method.

We make use of the Mozilla Common Voice dataset, which provides a wealth of voice samples with speaker metadata such as age and gender. This allows us to train and test our models for accurate gender classification and age estimation.

Other noteworthy datasets commonly used in speaker attribute recognition research, aside from Common Voice, are:

- aGender – German dataset with age and gender annotations included.
- VoxCeleb 1 & 2 – YouTube interviews with labelled demographics are part of this extensive speaker recognition dataset.
- TIMIT – A traditional phoneme-rich dataset that is helpful for speech analysis
- LibriSpeech – Audiobook recordings that are helpful for voice tasks in general.

In order to determine which approaches are best for fine-grained speaker attribute prediction using voice, we plan to assess both deep learning models and conventional machine learning techniques.

2. Literature Review

References	Work	Models/Approach	Dataset Used	Results	Analysis
1	Age and Gender Estimation Using Deep Neural Networks for IVR Systems	Various Deep Neural Networks -CNNs and Temporal Neural Networks - Analysis based on network size and architecture	Mozilla's Common Voice dataset	-Gender classification error: < 2% -Age group classification error: < 20%	Accuracy increased with deeper and more intricate networks; CNNs in conjunction with temporal models produced the best age prediction results. Deep neural networks (DNNs) demonstrated remarkable efficacy in gender classification, rendering this method feasible for practical IVR applications.
2.	Machine Learning-Based Gender and Age Detection from Voice	- Gender: Sequential model with 5 hidden layers - Age: Grid Search Pipeline using RobustScaler, PCA, Logistic Regression	Common Voice dataset	-Gender accuracy: ~91% - Age accuracy: ~59%	By combining temporal and spatial attention mechanisms, the Multi-Attention Mechanism (MAM) improves the extraction of pertinent features. It is appropriate for multilingual age and gender recognition in human-computer interaction (HCI) systems due to its robustness, scalability, and high accuracy across multiple datasets.
3.	Age and Gender Recognition Using CNN with Multi-Attention Module (MAM)	- End-to-end CNN with Multi-Attention Module (MAM) - Time and Frequency Attention mechanisms for spatial and temporal feature extraction	Common Voice & Korean Speech Recognition datasets	Common Voice: - Gender: 96% - Age: 73% - Age-Gender: 76% Korean Dataset: - Gender: 97% - Age: 97% - Age-Gender: 90%	By utilising both spatial and temporal attention, the Multi-Attention Mechanism (MAM) improves feature extraction and improves classification accuracy on a variety of datasets. For multilingual age and gender recognition tasks in human-computer interaction frameworks, its scalability and resilience make it an excellent choice.
4.	Gender Detection Using AI	- Supervised learning algorithms implemented in R	Custom dataset (3168	Gender recognition accuracy: >97%	The study demonstrates the effectiveness of AI-driven gender detection using

	Algorithms in Voice Data	- Emphasis on open-source, cost-effective AI solutions	voice samples)		voice data, showing great promise for use in targeted marketing, cybersecurity, and fraud prevention. It promotes the creation of affordable, open-source solutions to enable broad adoption in useful, real-world situations.
5.	Gender recognition using speech	Multi-layer architecture: -Layer1: Autocorrelation(fundamental frequency), spectral-entropy, spectral-flatness, mode frequency - Layer 2: Linear interpolation,MFCC - Classifiers: KNN, SVM	TIMIT, RAVDES, BGC (Self-created)	Highest accuracy: 96.8% (TIMIT with KNN)	By using spectral features in combination with Mel-Frequency Cepstral Coefficients (MFCCs), the suggested method achieves strong gender classification performance-while effectively-capturing speaker-specific vocal tract traits and acoustic patterns that are essential for discriminative modelling.

3. Dataset

3.1 Dataset Description

The dataset for this project is a carefully selected subset of the Mozilla Common Voice corpus, a vast, open-source collection of speech recordings donated by people all around the world. The chosen data has been pre-processed and filtered to allow efficient training of a gender and age prediction model from audio inputs.

The primary goal is multi-task learning for:

- Gender categorization (male vs female)
- Age regression (age range projection)

3.2 General Information

The dataset is organized into multiple subsets based on validation status: • Valid: Audio clips verified by at least two listeners to match the transcription. • Invalid: Clips marked by the majority of listeners as mismatched with the Transcription. • Other: Clips with fewer than two votes or equal votes for validity and invalidity. The valid and other subsets are further split into: • train: For training ASR models. • dev: For development and experimentation. • test: For evaluation using metrics like Word Error Rate (WER).

- Source: Mozilla Common Voice
- Language: English
- Format: .wav audio files with accompanying metadata (CSV format)
- Gender: encoded as 0 (male) or 1 (female)
- Age: converted from categories (e.g., "twenties") to numerical values (e.g., 20)
- Dropped Metadata: up_votes, down_votes, accent, duration, text

3.3 Data Structure and Conventions

Each audio sample in the dataset has:

- A filename pointing to the local path of the audio file
- Gender and age labels
- All entries are validated to ensure no missing values for gender or age
- Gender values are restricted to binary classes (samples labeled as "other" are removed)
- Age values are mapped to numerical midpoints of the original age range

To ensure gender balance, the dataset is down sampled to equalize male and female samples across age groups.

3.4 Preprocessing

The dataset was pre-processed before being used for training. Each MP3 clip was co verted into a log-mel spectrogram to serve as input to our deep learning model. Only samples from the cv-valid-train and cv-valid-dev sets were used for training and validation. The test set was kept aside for final evaluation.

- Sample Rate: Resampled to 22,050 Hz
- Clip Duration: 5 seconds (110,250 samples)
- Cropping:
  - Random cropping during training
  - Center cropping during testing
- Feature Extraction: MFCCs (Mel-Frequency Cepstral Coefficients) using librosa
- Standardization: MFCC features are normalized using mean and standard deviation
- TensorFlow Dataset API is used for pipeline creation:
  - Efficient batching (batch size = 256)
  - Shuffling and caching during training
  - Prefetching for performance

3.5 Acknowledgments

The Mozilla Common Voice project and its contributors are gratefully acknowledged for making this large-scale, high-quality dataset freely available to the research community. We gratefully acknowledge Michael Henretty, Tilman Kamp, Kelly Davis, and the Mozilla Common Voice team for compiling the dataset. We also thank the global contributor community whose donated voices made this possible, as well as organizations like Mycroft, SNIPS.ai, Mythic, Tatoeba.org, Bangor University, and SAP for their support.

4. Methodology

This chapter describes the structured method adopted for the prediction of age and gender from speech signals. The methodology includes six chief stages: data collection, data preprocessing, dataset splitting, model training, evaluation, and prediction.

4.1 Data Collection

Data was obtained from the Mozilla Common Voice platform. It includes audio clips with metadata describing speaker gender and age. The dataset was downloaded and locally prepared in **.wav** format.

4.2 Data Preprocessing

Preprocessing of the audio data is essential to ensure efficient feature extraction and model performance. The following operations were carried out:

- Non-binary genders and missing age/gender samples were removed.
- Audio was converted to mono channel and resampled to 22,050 Hz.
- Each clip was cropped or padded to exactly 5 seconds.
- MFCC features were extracted and normalized.
- Final features were reshaped to serve as input for the model.

4.3 Split of Dataset

To evaluate the models' generalization performance, the dataset was divided into training and test subsets. This was accomplished using the train test split function from the scikit-learn library, ensuring a balanced distribution of age and gender labels in both subsets.

- Dataset was split into training and testing sets.
- A validation set was implicitly derived during training.
- Gender distribution was kept balanced in both sets.

4.4 Model Training

- A hybrid CNN-RNN model was used:
    - CNN layers (Conv1D) extract spatial features from MFCCs
    - A GRU layer captures temporal dependencies
    - Output branches for:

        - Gender (sigmoid activation, binary cross-entropy loss)
        - Age (linear activation, MSLE loss)
- Loss Weights: 0.5 each for gender and age
- Optimizer: Adam with learning rate = 0.0001
- Metrics:
    - Gender: Accuracy
    - Age: Mean Absolute Error (in years)
- Training run for 30 epochs with batch size of 256

4.5 Model Evaluation

The trained classifiers were assessed using standard classification metrics. The primary evaluation metric was accuracy, calculated on the test set. In addition, confusion matrices and classification reports were used to evaluate the performance of each model across various classes.

Model performance was validated using the test set:

- Gender prediction accuracy
- Mean Absolute Error for age prediction

Test preprocessing used center cropping for consistency.

4.6 Prediction and Deployment

The final stage of the methodology involves applying the trained models to predict the age and gender of new speakers and preparing the system for real-world deployment.

4.6.1 Prediction Pipeline

To forecast the age and gender of an unknown audio input, the system performs the same preprocessing and feature extraction steps used during training. The steps are as follows:

- Custom .wav files can be uploaded for inference.
- Files are processed to match training format:
    - Resampled, cropped, MFCC extracted, normalized
- Pre-trained model predicts gender and age.
- Audio visualizations include:
    - Waveform
    - MFCC spectrogram

  - Mel-spectrogram
  - RMS energy

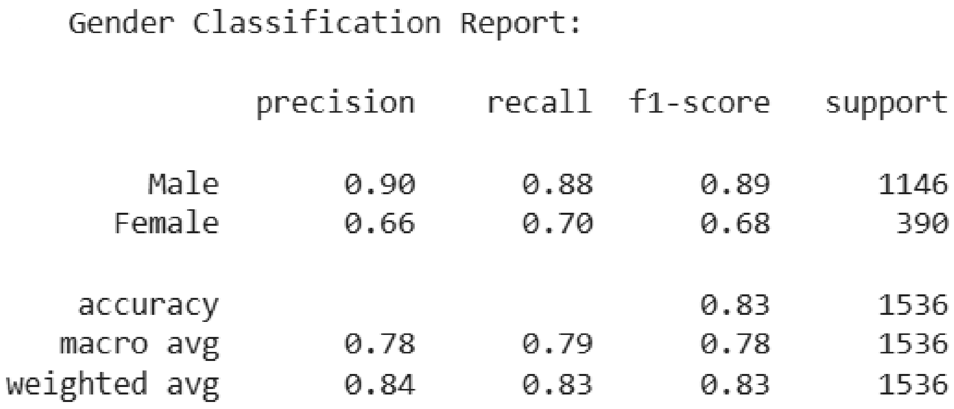
4.6.2 Deployment Strategy

To enable real-world deployment, the trained models and prediction pipeline can be deployed using any of the following strategy:

- The model is stored in .keras format and can be reloaded for inference.
- A user-friendly pipeline is implemented in the notebook to handle real-time predictions with audio upload.

5. Results

5.1 Performance Summary



```
Gender Classification Report:

              precision    recall  f1-score   support

        Male       0.90      0.88      0.89      1146
      Female       0.66      0.70      0.68       390

    accuracy                           0.83      1536
   macro avg       0.78      0.79      0.78      1536
weighted avg       0.84      0.83      0.83      1536
```

5.2 Prediction Results

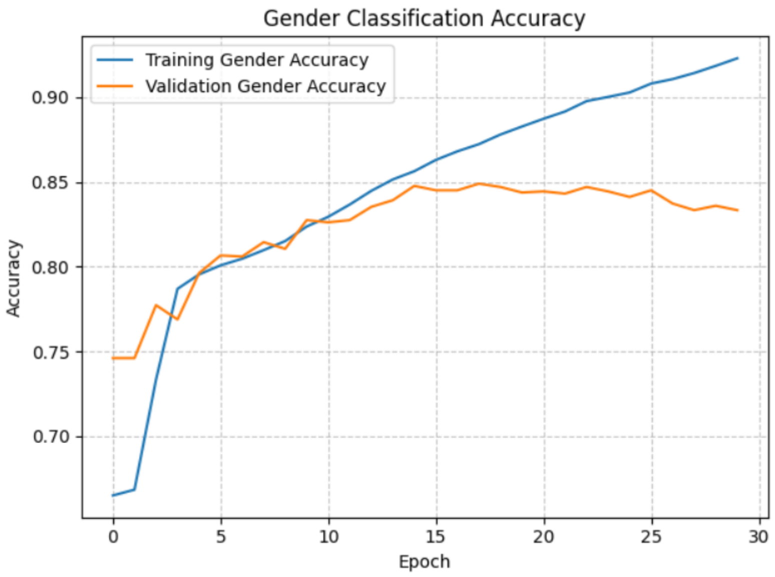


Figure 5.1: Gender Prediction Accuracy across Epochs

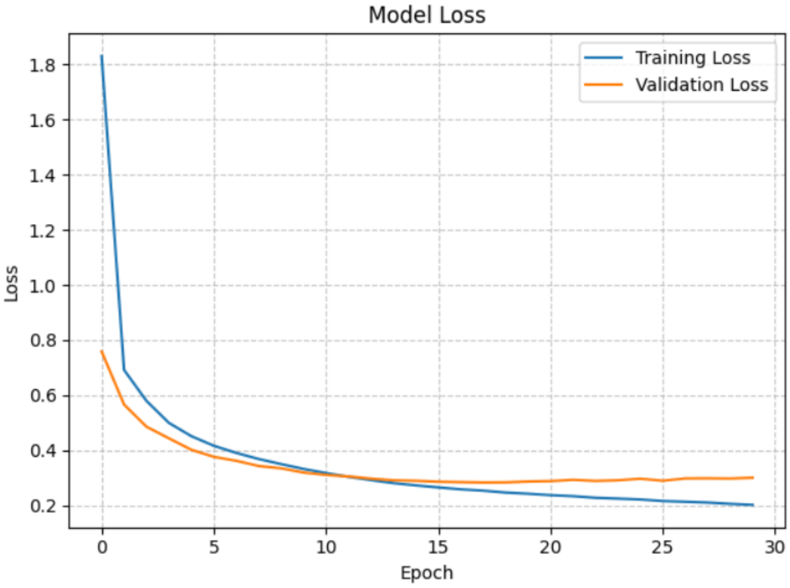


Figure 5.2: Model Loss across Epochs

5.2.2 Confusion Matrix

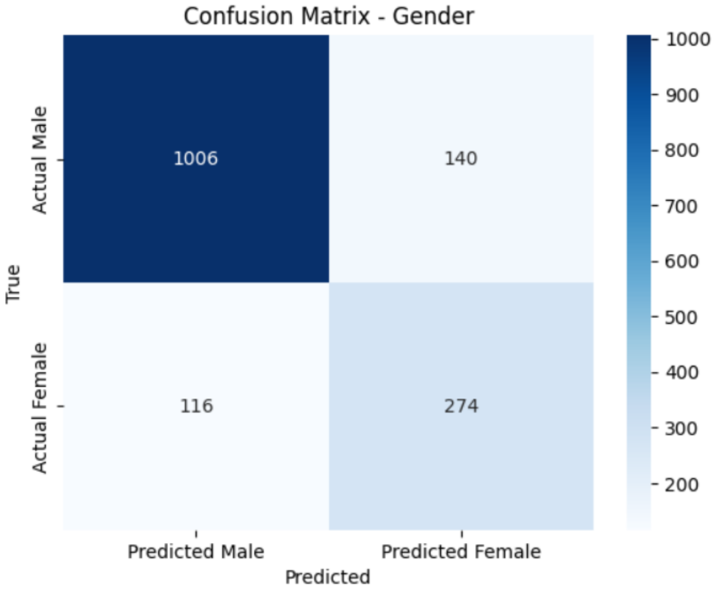


Figure 5.3: Confusion Matrix for Gender Prediction

5.3 Age Prediction Results

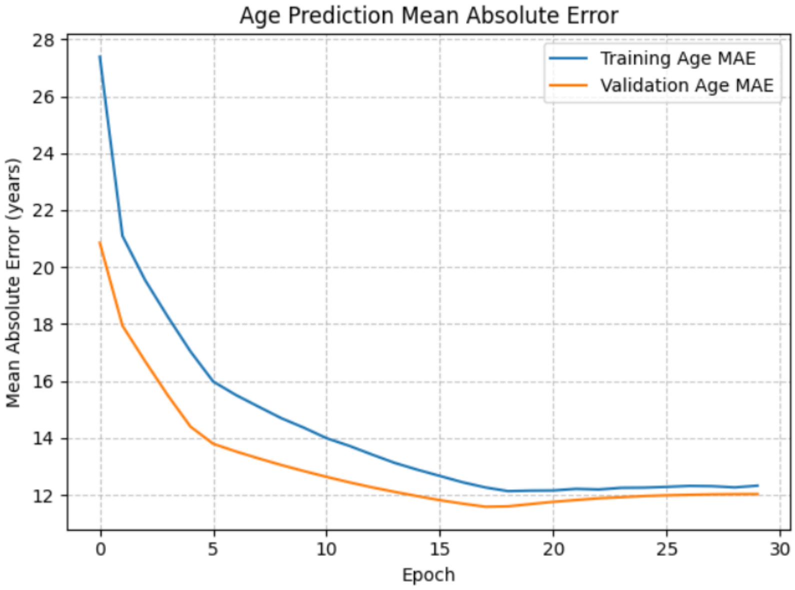


Figure 5.4: Age Prediction Mean Absolute Error across Epochs

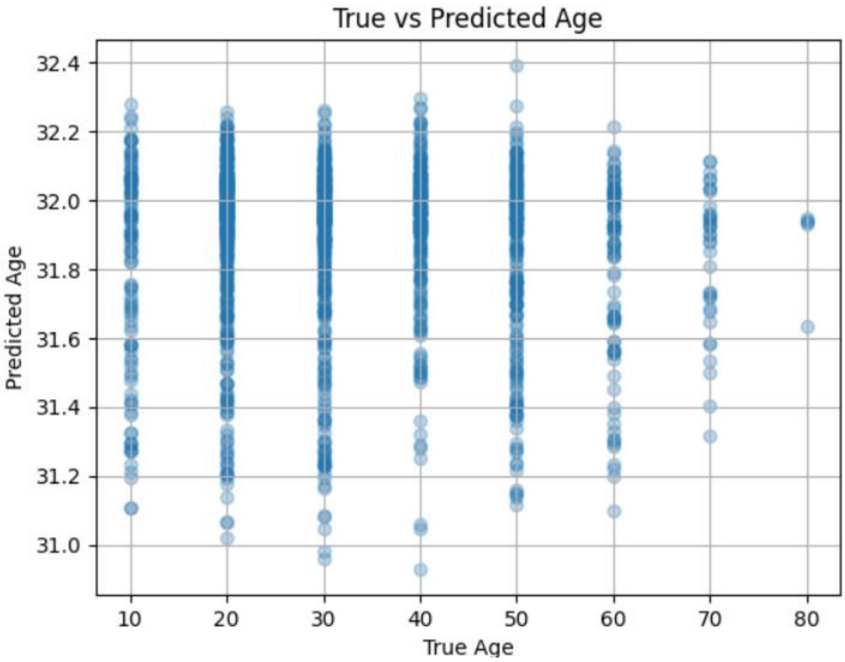


Figure 5.5: True vs Predicted Age

6. Conclusion

The increasing use of voice-based interaction systems in contemporary applications emphasises the necessity of speech-based intelligent demographic recognition. In this project, we used a machine learning-based approach to address the problem of automatically identifying age and gender from voice data. We trained supervised models that can learn speaker characteristics embedded in audio signals by using the Mozilla Common Voice dataset to extract important acoustic features, especially Mel-Frequency Cepstral Coefficients (MFCCs).

The results of the experiment show that deep neural network architectures greatly improve gender prediction accuracy. Due to the subtle and variable nature of vocal traits influenced by variables like emotional state, health, and speaking style, it proved more difficult to predict the precise age rather than broad age groups. Nonetheless, the system continued to demonstrate encouraging patterns in accurately determining speaker age.

This study demonstrates how well signal processing and machine learning work together in the field of speech-based user profiling. The suggested approach is appropriate for real-world applications like smart assistants, interactive voice response (IVR) systems, healthcare diagnostics, and user behaviour analytics due to its scalability and adaptability. Additionally, using an open-source dataset guarantees that the solution will continue to be reproducible and accessible.

Future research could include adding more speakers from a wider range of languages and geographical areas, improving the models for real-time, on-device processing, and integrating more resilient temporal models like RNNs or transformers. In the end, this project successfully decodes speaker demographics through speech, laying the groundwork for creating more intelligent, secure, and personalised voice-driven systems.

7. References

1. H. A. S´anchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, et al., “Age group classification and gender recognition from speech with temporal convolutional neural networks,” *Multimedia Tools and Applications*, vol. 81, pp. 3535–3552, 2022, doi: 10.1007/s11042-021-11614-4.

2. A. Tursunov, Mustaqeem, J. Y. Choeh, and S. Kwon, “Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms,” vol. 21, no. 17, p. 5892, 2021, doi: 10.3390/s21175892.

3. M. A. Uddin, M. S. Hossain, R. K. Pathan, and M. Biswas, “Gender Recognition from Human Voice using Multi-Layer Architecture,” in *2020 International Conference on Innovations in Intelligent Systems and Applications (INISTA)*, Novi Sad, Serbia, 2020, doi: 10.1109/INISTA49547.2020.9194654.

4. L. Jasuja, A. Rasool, and G. Hajela, “Voice Gender Recognizer: Recognition of Gender from Voice using Deep Neural Networks,” *2020 International Conference on Smart Electronics and Communication (ICOSEC)*, Trichy, India, pp. 319–324, 2020, doi: 10.1109/ICOSEC49089.2020.9215254.

5. E. Yu¨cesoy, “Speaker age and gender recognition using 1D and 2D convolutional neural networks,” *Neural Computing and Applications*, vol. 36, pp. 3065–3075, 2024, doi: 10.1007/s00521-023-09153-0.