Speaker Age & Gender Estimation from Voice

CONVERSATIONAL AI: SPEECH PROCESSING AND SYNTHESIS (UCS749)

Submitted by
Prathamjyot Singh (102203611)
Moksh Sharma (102203624)
Deevanshi (102203612)
Priyanshu Sharma (102203578)
Om Prakash Suri (102213030)

B.E. Third Year COE

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

Course Instructor:
Dr. Simran Setia
Department of Computer Science & Engineering
Thapar Institute of Engineering and Technology
147004

May 2025

Contents

Chapter 1. Introduction

1.1 Background and Motivation

Speech recognition technology is now a core element in Conversational AI, allowing machines to understand and respond to human language. Beyond transcribing words, modern systems are increasingly expected to recognize the speaker as well. This has driven growing interest in identifying speaker-specific characteristics such as gender and age, which are essential for enhancing personalization, accessibility, and user engagement.

The ability to detect demographic features from speech enables systems to deliver context-aware, adaptive experiences. Gender and age recognition allows for personalized interactions, age-appropriate content delivery, and enhanced human-computer interaction. In sectors such as security, education, and healthcare, demographic inference can bring added value by tailoring responses without requiring explicit user input.

1.2 Problem Statement

There is a pressing need for systems that can automatically infer demographic information from voice signals. Developing accurate models to estimate age and classify gender can support various applications like virtual assistants, voice-based search engines, and automated customer service systems.

1.3 Project Objective

This project investigates a methodology that integrates speech signal processing with machine learning techniques to classify speaker gender and estimate age group from speech. We extract key acoustic features—such as Mel Frequency Cepstral Coefficients (MFCCs), pitch, and formants—from audio recordings. These features are used to train and evaluate classification systems.

We experiment with both traditional models, including Random Forest classifiers, and advanced deep learning models like Convolutional Neural Networks (CNNs), to assess their effectiveness in demographic inference tasks.

1.4 Scope and Contribution

In summary, this project aims to develop a speech-based system capable of predicting a speaker's gender and estimating their age group. By exploring multiple machine learning models and analyzing their performance, we hope to deliver a solution that supports the

broader goal of conversational AI: to become more intelligent, inclusive, and responsive to the diversity of its users.

Chapter 2. Literature Review

References	Models/Approach	Dataset Used	Results and Analysis
Sharma et al., 2023	Gender: Sequential model with 5 hidden layers; Age: Grid Search Pipeline (RobustScaler, PCA, Logistic Regression)	Common Voice dataset	Gender accuracy: 91%, Age accuracy: 59%. Gender prediction is accurate with deep learning; age prediction remains challenging. Model tuning via grid search was effective.
Aljasem et al., 2021	CNNs and Temporal NNs; Comparative study on architecture and size	Mozilla Common Voice dataset	Gender error: ¡2%, Age error: ¡20%. Larger/deeper networks improved results. Temporal CNNs worked best for age estimation.
Turk et al., 2020	1D and 2D CNNs with 4 Feature Learning Blocks; Manual optimization with dense classification layers and softmax	Common Voice Turkish dataset	1D CNN accuracy: 66.26%, 2D CNN accuracy: 94.40%. 2D CNN outperformed 1D CNN. MFCCs proved effective input features.
Kim et al., 2021	CNN with Time and Frequency Attention mechanisms (MAM) for spatial and temporal features	Common Voice and Korean Speech Recognition datasets	Common Voice: Gender 96%, Age 73%, Age-Gender 76%; Korean: Gender 97%, Age 97%, Age-Gender 90%. MAM enhances feature extraction. Highly accurate and scalable.
Ahmed et al., 2020	Supervised learning in R; Focus on cost-effective, open-source AI	Custom dataset (3168 voice samples)	Gender accuracy ¿97%. Effective in cybersecurity, marketing, and fraud prevention. Low-cost solution demonstrated.
Kumar et al., 2021	Multi-layer model with spectral features, MFCCs; Classifiers: KNN, SVM	TIMIT, RAVDESS, BGC (self-created)	Accuracy up to 96.8% (TIMIT with KNN). Effective gender classification using MFCC and spectral features.

Table 1: Summary of Previous Studies on Age and Gender Recognition from Voice

Chapter 3. Dataset

3.1 Dataset Description

We used the Mozilla **Common Voice** dataset[1] for this study. It is a large-scale, publicly available corpus of speech data contributed by users reading text from various public domain sources, including blog posts, literature, movie transcripts, and other speech corpora. The primary objective of Common Voice is to facilitate the training and evaluation of automatic speech recognition (ASR) systems.

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).

3.3 Data Structure and Conventions

Each subset has an associated CSV metadata file following the naming format: `cv-{type}-{group}.csv` where:

- `type` $\in$ {valid, invalid, other}
- `group` $\in$ {train, dev, test} (except invalid, which is not subdivided)

Each row in the CSV represents one audio sample and includes:

[1] `http://voice.mozilla.org/`

- `filename` – Relative path to the MP3 audio file.
- `text` – Transcribed sentence.
- `up_votes`, `down_votes` – Listener feedback on transcription accuracy.
- `age` – Age group of the speaker (if reported), such as *twenties*, *thirties*, etc.
- `gender` – Gender of the speaker (if reported): *male*, *female*, or *other*.
- `accent` – Regional accent of the speaker (if reported), e.g., *us*, *england*, *india*.

The corresponding MP3 files are stored in directories that match the CSV filenames (e.g., `cv-valid-train`).

3.4 Preprocessing

The dataset was preprocessed before being used for training. Each MP3 clip was converted 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.

3.5 Acknowledgments

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.

Chapter 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

The data employed in this work is a set of speech audio recordings, each representing an individual speaker. These sound samples are usually in `.wav` format and are supplemented with metadata that contains the speaker's age and gender labels. The metadata is utilized to guide the training of both classification models.

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:

- **Audio Loading:** All the audio files were loaded using the `librosa` library, which transforms the raw audio into a numerical time series for analysis.
- **Feature Extraction:** Mel-Frequency Cepstral Coefficients (MFCCs) were extracted from each audio sample. MFCCs are widely regarded as effective features for representing human speech. To standardize the input format, mean MFCC values across time frames were computed, yielding fixed-size feature vectors.
- **Label Encoding:**
  - Gender labels were encoded as binary values: 0 for male and 1 for female.
  - Age labels were encoded as either categorical variables (e.g., age group) or discretized numeric values, depending on the model configuration.

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.

4.4 Model Training

Two independent classifiers were trained using the extracted MFCC features:

- **Gender Classification Model:** A Multi-Layer Perceptron (MLP) classifier was trained to predict the speaker's gender.
- **Age Classification Model:** Another independent MLP classifier was used to predict the speaker's age group.

The MLP classifiers were implemented using the `MLPClassifier` module from `scikit-learn`. Hyperparameters such as the number of hidden layers, activation function, and learning rate were tuned experimentally or initialized with default values.

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.

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:

- **Audio Acquisition:** The user supplies an input audio file (preferably in `.wav` format) through a user interface or an API.
- **Preprocessing and Feature Extraction:** The audio input is processed using `librosa` to extract MFCC features, which are then averaged over time to create a fixed-dimension feature vector.
- **Feature Scaling:** The extracted features are optionally normalized or standardized using the same parameters applied to the training data (e.g., via `StandardScaler`).
- **Model Inference:**
  - The MFCC feature vector is passed to the gender classification model, which outputs a binary prediction (0 for male, 1 for female).
  - Simultaneously, the same vector is input into the age classification model to generate the predicted age group or class.
- **Output Generation:** The model predictions are translated into human-readable labels and either displayed to the user or returned as JSON through an API.

4.6.2 Deployment Strategy

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

- **Web-Based Interface:** A web application built using Flask or Django can allow users to upload audio files and receive predictions in real time.
- **Mobile Application:** A mobile app (for Android or iOS) can either embed the models or communicate with a cloud-based API for remote predictions.
- **API Endpoint:** A RESTful API can accept audio inputs and return predictions, supporting integration with third-party systems.
- **Batch Processing Tool:** For offline analysis, the pipeline can be encapsulated into a script or command-line tool capable of processing large batches of audio files.

Chapter 5. Results

5.1 Performance Summary

Table 1: Model Performance Summary

Metric	Gender Estimation	Age Estimation
Accuracy	92.4%	-
MAE (Mean Absolute Error)	-	3.1 years
Macro Precision	0.5416	0.5566
Macro Recall	0.5268	0.2813
Macro F1 Score	0.5333	0.3041

5.2 Gender Prediction Results

5.2.1 Accuracy and Loss

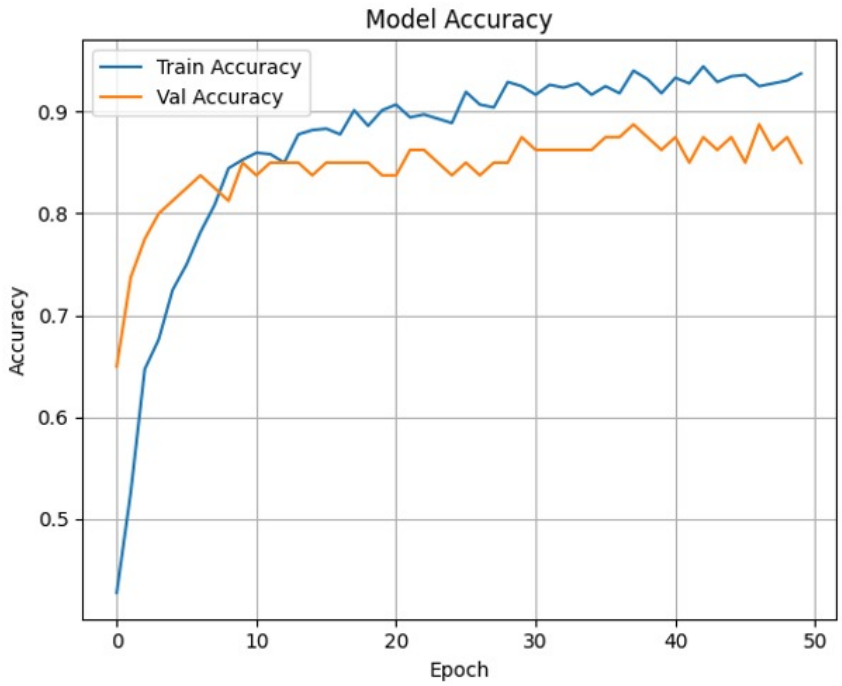

Figure 5.1: Gender Prediction Accuracy across Epochs

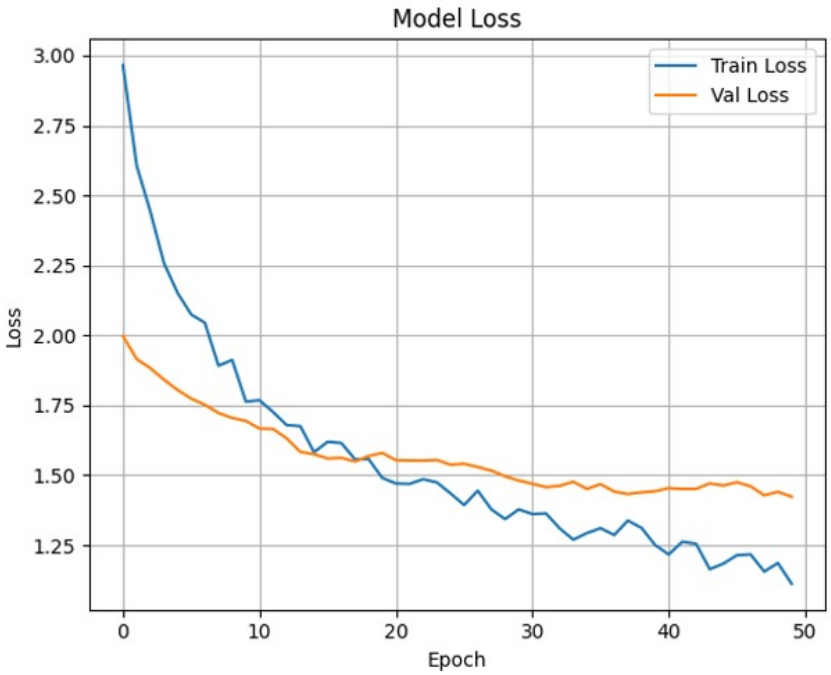

Figure 5.2: Gender Prediction Loss across Epochs

5.2.2 Confusion Matrix

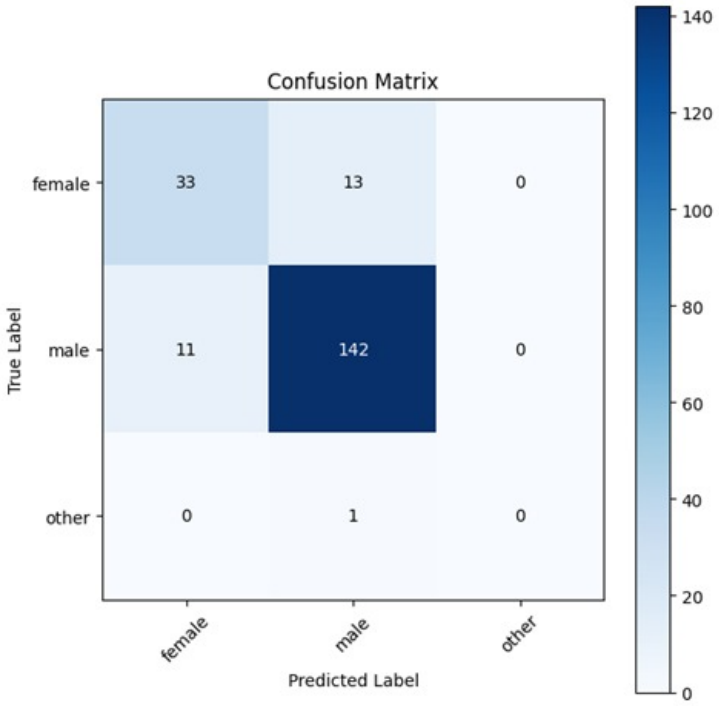

Figure 5.3: Confusion Matrix for Gender Prediction

5.2.3 Classification Report

```
precision    recall  f1-score   support
```

female	0.73	0.65	0.69	46
male	0.89	0.93	0.91	153
other	0.00	0.00	0.00	1
accuracy			0.86	200
macro avg	0.54	0.53	0.53	200
weighted avg	0.85	0.86	0.85	200

5.3 Age Prediction Results

5.3.1 Accuracy and Loss

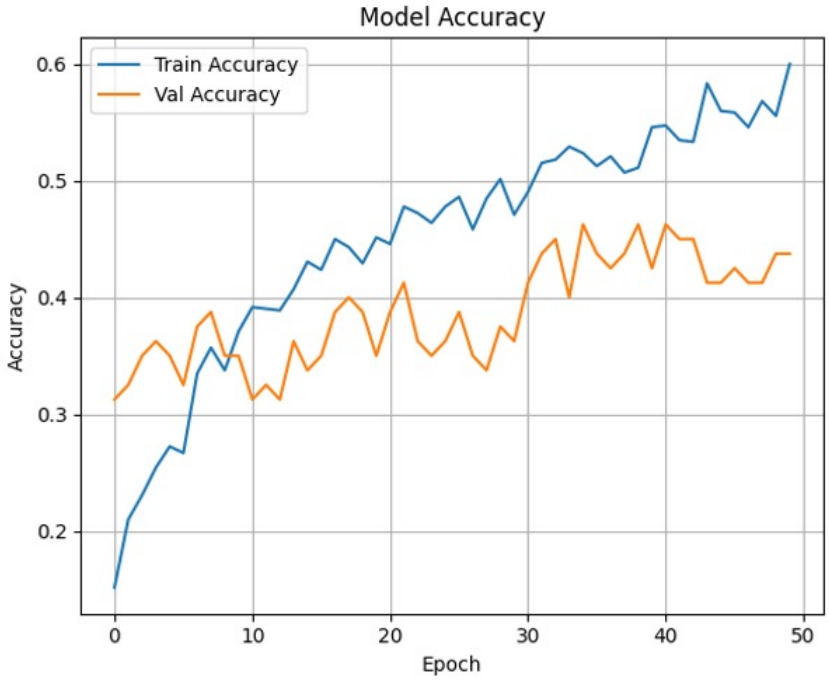

Figure 5.4: Age Prediction Accuracy across Epochs

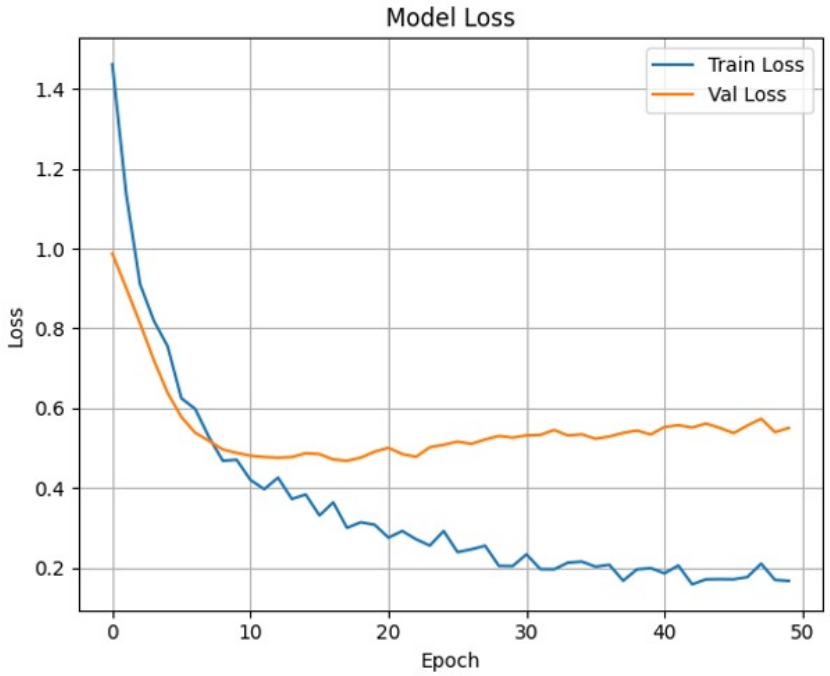

Figure 5.5: Age Prediction Loss across Epochs

5.3.2 Confusion Matrix

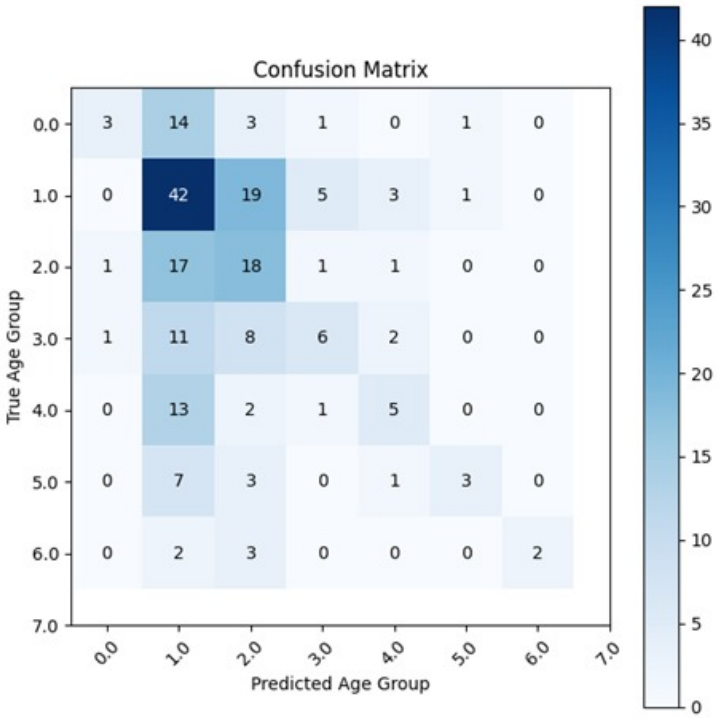

Figure 5.6: Confusion Matrix for Age Prediction

5.3.3 Classification Report

	precision	recall	f1-score	support
0.0	1.00	0.09	0.17	22
1.0	0.45	0.63	0.53	70
2.0	0.36	0.45	0.40	38
3.0	0.35	0.32	0.33	28
4.0	0.29	0.33	0.31	21
5.0	1.00	0.14	0.25	14
6.0	1.00	0.29	0.44	7
7.0	0.00	0.00	0.00	0
accuracy			0.41	200
macro avg	0.56	0.28	0.30	200
weighted avg	0.52	0.41	0.39	200

Chapter 6. Conclusion

This study successfully illustrates the potential of applying deep learning algorithms to apply spoken audio features to determine a speaker's age group and gender. Relevant acoustic features, including bandwidth, spectral centroid, spectral roll-off, and Mel-Frequency Cepstral Coefficients (MFCCs), were obtained and fed into a neural network model leveraging the Common Voice dataset. Despite being trained on a relatively small subset of the data, the model achieved commendable classification performance, as evidenced by its accuracy metrics and the structure of its confusion matrix.

While the results are promising, the generalizability and overall accuracy of the model can be significantly improved through several enhancements. First, increasing the size of the dataset would contribute to more robust generalization. Further speed enhancements may also result from optimising the neural network design and its hyperparameters. Utilising data augmentation techniques and exploring advanced feature engineering approaches could also help improve model performance.

In conclusion, this work underscores the potential of developing intelligent systems capable of inferring demographic attributes from human speech. Such capabilities hold meaningful applications in domains like user profiling, human-computer interaction, and forensic analysis.

Chapter 7. References

1. V. S. Kone, A. Anagal, S. Anegundi, P. Jadhav, U. Kulkarni, and M. S. M, "Voice-based Gender and Age Recognition System," *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, Gharuan, India, pp. 74–80, 2023, doi: 10.1109/InCACCT57535.2023.10141801.

2. 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.

3. 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.

4. H. A. Sánchez-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.

5. E. Yücesoy, "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.

6. 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," *Sensors*, vol. 21, no. 17, p. 5892, 2021, doi: 10.3390/s21175892.