Speaker's age prediction problem

Lanzillotta Andrea
Politecnico di Torino
343438
s343438@studenti.polito.it

Mallo Giuseppe
Politecnico di Torino
346884
s346884@studenti.polito.it

***Abstract*—This report presents a machine learning pipeline for the estimation of a speaker's age based on their spoken audio. The approach involves extracting key acoustic features such as pitch, jitter, shimmer, and spectral centroid, along with linguistic metadata like gender and ethnicity. These features are used as inputs to a regression model designed to predict the speaker's age as a continuous variable. Experimental results suggest that the model can reasonably map the extracted features to the target age variable, achieving promising performance when compared to a baseline. While the results are encouraging, further refinement is required to improve prediction accuracy and generalization.**

Problem Overview

This project aims to develop a machine learning pipeline to estimate a speaker's age based on audio files. The inputs consist of `.wav` files and metadata extracted from audio signals, while the model is designed to predict age as a continuous variable. This task belongs to the domain of spoken language processing, with a particular focus on the correlation between vocal features and biological parameters, such as age.

The provided dataset comprises two main sections:

- **Development:** 2,933 samples with age targets provided, used for training and validation.
- **Evaluation:** 691 samples without age targets, intended for model evaluation.

Each sample is described by 19 non-null acoustic and linguistic features extracted from the audio signals, along with the speaker's age as the target variable (only available in the development section). Key features include mean, minimum, and maximum signal frequency, word count, number of pauses, and the total duration of silences. During development, some features were deemed unnecessary and excluded, while others were added, as detailed below.

Proposed Approach

A. Preprocessing

As a first step, the dataset was loaded into suitable data structures, such as Pandas DataFrames, followed by an evaluation phase (not included in the code), which identified necessary preprocessing actions to render the data usable. The original dataset contained four nominal features:

- **Gender**
- **Ethnicity**
- **Tempo**
- **Path**

For the first two columns, *Gender* and *Ethnicity*, one-hot encoding was applied. This process resulted in a total of 223 new columns (2 from *Gender* and 221 from *Ethnicity*). For the *Tempo* column, which contained the speaking rate expressed in BPM, values were converted to floating-point format. Lastly, the *Path* column was retained after string manipulations to keep only the corresponding audio file names. This information was preserved for potential future use, although it does not influence regression models since the feature is dropped whenever necessary.

B. Model Selection

Following this preprocessing, an initial regression test was conducted using a generic `RandomForestRegressor` to gain a preliminary understanding of the dataset. Predictions were made on a pseudo-test set, which was derived from the development dataset using the `train_test_split` function. Similar procedures were followed for all subsequent intermediate regression tests, with the evaluation dataset updated alongside the development dataset but only provided to the model during the final evaluation phase. To improve clarity and conciseness, this process is not reiterated in later sections.

The results of the initial regression test were as follows:

- **R^2 Score:** 0.3688
- **Mean Absolute Error (MAE):** 6.8141

Convinced that there was a margin of improvement, a model evaluation function was implemented to test various regression models and select the one most suitable for this use case. Three models were tested:

- **MLP_Standard**
- **Support Vector Regressor (SVR)**
- **RandomForestRegressor**

Model	RMSE	R^2 score
MLP Standard	11.455701235598674	0.19546451227315387
SVR	13.101749336980067	-0.052350521092772606
RandomForest	10.007980742499175	0.3859626765189904

TABLE I
Results of evaluation model

As is clear from the table above, the statistics indicated that the `RandomForestRegressor` was the most suitable choice, as it achieved a significantly lower Root Mean Squared Error (RMSE) and higher R^2 score compared to the other models.

Feature Importance Analysis: Feature importance analysis was conducted, yielding the following top 10 features:

1) Silence Duration: 0.31885
2) Number of Pauses: 0.05545
3) Jitter: 0.05541
4) Harmonics-to-Noise Ratio (HNR): 0.05234
5) Spectral Centroid Mean: 0.05112
6) Energy: 0.04768
7) Mean Pitch: 0.04664
8) Minimum Pitch: 0.04652
9) Shimmer: 0.04608
10) Index: 0.04579

Based on these results, we reckoned that additional features were necessary. They have been extracted from the provided audio files. Specifically, the project focused on extracting spectrograms from the `.wav` files and computing their Mel-Frequency Cepstral Coefficients (MFCC).

MFCC Analysis: MFCCs represent numerical characteristics of audio signals and are widely used in speech and language processing applications. They capture the spectral shape of human voice signals, emphasizing timbral features relevant to auditory perception. MFCC computation involves the following steps:

1) **Windowing:** The audio signal is divided into time windows to capture temporal evolution.
2) **Spectrogram Calculation:** Each window undergoes a Fourier Transform to convert the signal from the time domain to the frequency domain.
3) **Mel Filter Bank Application:** A nonlinear frequency mapping, based on the Mel scale, is applied to emphasize low-to-mid frequencies.
4) **Cepstral Coefficients Calculation:** An inverse transform is performed to obtain compact representations of the spectral structure.

The resulting MFCCs reflect the evolution of acoustic features over time and can differentiate changes in pitch and timbre.

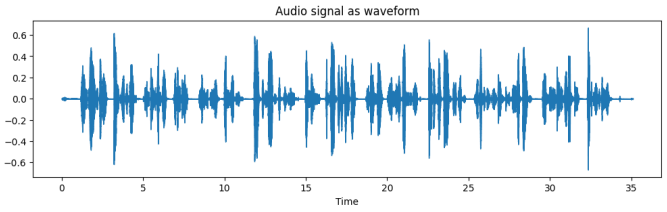

Fig. 1. Waveform of audio 0.wav from developing dataset

However, we believed that better results could be achieved by performing the calculation not on the entire spectrogram, but on smaller windows of it. This is because we considered that focusing on limited portions of the spectrogram could ensure the extraction of more precise and distinctive features for determining the speaker's age. For instance, although the frequency range of the human voice remains unchanged, the voice of a child and that of an adult develop within different frequency bands. Therefore, processing individual spectrograms in greater detail could lead to a model better capable of distinguishing the speaker's age. For each window we have considered the mean, the standard deviation and the maximum. The following images explain the approach better:

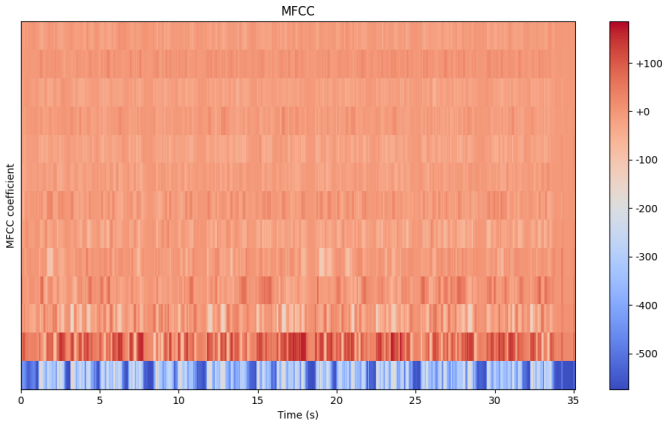

Fig. 2. MFCC of the audio 0.wav from developing dataset

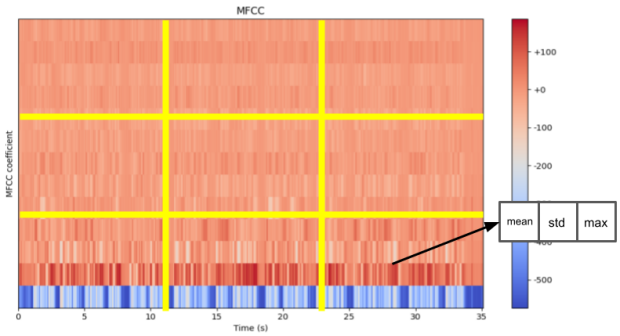

Fig. 3. Graphical example of how features are extracted

Testing various spectrogram segmentations (ranging from 1 to 100 windows), the optimal configuration was determined to be 9 rows and 2 columns. Thanks to this segmentation, the `RandomForestRegressor` has given us the following metrics:

- **R^2:** 0.4625
- **MAE:** 6.4511
- **RMSE:** 9.4353

Additional Features

Words Per Second: This feature, hypothesized to correlate with age, was calculated as the number of words pronounced per second in each audio sample. Younger individuals typically have faster speech rates compared to older speakers. Incorporating this feature improved performance metrics further:

- **R^2:** 0.4586
- **MAE:** 6.4258
- **RMSE:** 9.4696

Now, we report below the 10 features ranked by feature importance as declared by the regressor:

1) Duration: 0.2876
2) Silence Duration: 0.0548
3) Ethnicity English: 0.0490
4) Words Per Second: 0.0336
5) Mfcc max (1,0): 0.0286
6) Jitter: 0.0278
7) Mfcc std (4,0): 0.0239
8) Spectral centroid mean: 0.0205
9) Mfcc mean (4,0): 0.01997
10) Mfcc mean (5,0): 0.01946

As can be observed, the added features rank among the top positions in terms of importance, from the number of words per second, which becomes the fourth most important feature, to the various MFCCs of the individual spectrogram windows, which start appearing from the fifth feature in order of importance. However the gender features are ranked in low positions, and it is not very helpful since a voice can change radically based on the genetic sex of the speaker, which can significantly impact age recognition.

Gender-Specific Models: To account for significant differences in vocal frequencies between genders, separate `RandomForestRegressor` models were trained for male and female speakers. This specialization reduced potential confusion caused by overlapping frequency ranges between genders and improved accuracy.

C. Hyperparameters tuning

Finally, a *GridSearch* (which has RMSE as the *scoring* parameter) was performed to identify the best hyperparameters for each of the two regressors, testing the *number of estimators* and the *max_depth* from a set of different values for each parameter. The implemented *GridSearch* report as results the following *best regressor*:

Regressor	n_estimators	max_depth
RFR_Male	100	10
RFR_Female	100	10

TABLE II
BEST REGRESSORS FROM GRID SEARCH

RESULTS

The final model achieved an **RMSE of approximately 9.8**, positioning it in the upper-middle range of the leaderboard. We repeat that the model used as the base for performing regression is the Random Forest Regressor, as it performs better according to the chosen metrics, which are RMSE and R2 score. In fact, if we reference the comparison between the three regressors from the code provided, the `RandomForestRegressor` consistently outperformed the alternatives. Note: We discarded models such as the Linear Regressor a priori, as the relationship between age and the considered features is certainly not linear. The final submission leveraged two gender-specific models optimized via Grid Search, further enhancing the results.

DISCUSSION

In general, the final result is considered satisfactory.

However, when considering alternative methodologies, an analysis of the development dataframe reveals a significant imbalance in the distribution of audio samples across the different age ranges. Specifically, there is a notable overrepresentation of individuals in the younger age ranges, which is detrimental to the representation of individuals in the older age groups. A regressor trained on such a dataset would exhibit a bias toward predicting lower ages when applied to a new dataset. This is evident from the output file, where almost all of the approximately 600 evaluation records are assigned an age below 50 years.

range of age	
0-10	2.0
11-20	1031.0
21-30	1095.0
31-40	356.0
41-50	204.0
51-60	149.0
61-70	55.0
71-80	25.0
81-90	15.0
91-100	1.0

Fig. 4. Data development distribution among age ranges

Consequently, an attempt was made to address this imbalance by rebalancing the development dataframe to ensure an equal number of samples for each age range. However, this approach presents a trade-off: if an optimal balance between the cardinalities of the individual age ranges cannot be achieved, the number of development data points could be significantly reduced, thereby depriving the regressor of a substantial amount of training samples. In conclusion, this rebalancing approach did not result in significant improvements in the RMSE, and as such, it was decided not to implement it.