Data Science Lab: Estimation of speaker's age

Giacomo Scali
Nicola Biagioli
Politecnico di Torino
Student id: s346381, s344677

***Abstract*—This report investigates the task of estimating a speaker's age from vocal characteristics. Utilizing a dataset of 3,624 audio files, a pre-emphasis filter was applied to the signal prior to analysis, acoustic features such as Mel-Frequency Cepstral Coefficients (MFCCs) were extracted to complement those already available. Two models, Random Forest and LASSO regression, were evaluated. Model performance was assessed using Root Mean Square Error (RMSE), with results indicating that LASSO exhibited a slight advantage over Random Forest.**

I. Problem Overview

In this report, we address the task of estimating a speaker's age based on vocal characteristics, a regression problem in the field of speech processing. Using a dataset of 3,624 audio files collected under controlled conditions to ensure consistent feature extraction, the proposed solution combines acoustic feature extraction (such as MFCC) with machine learning models to minimize prediction errors measured by Root Mean Square Error (RMSE).
A preliminary analysis of the dataset reveals irregularities in the distribution of the recorded files. Indeed, as shown in

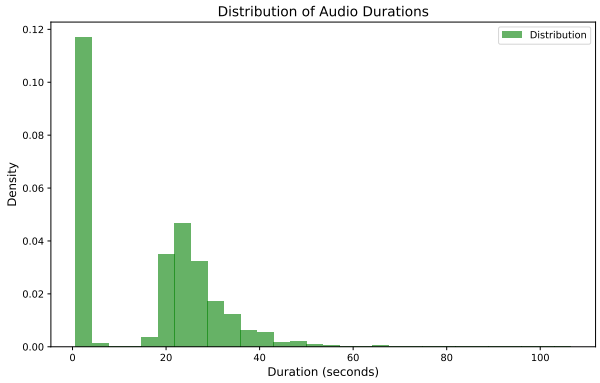

Fig. 1. Recordings duration distribution in the dataset,

Figure 1, some observations significantly differ in duration. Listening to a few samples from the dataset reveals that the longer tracks are recordings of the same text, whereas the shorter ones do not seem to follow any specific pattern.
Analyzing the age distribution in the dataset used for training the model, Figure 2 reveals a strongly skewed distribution: ages between 16 and 40 are overrepresented. This characteristic of the dataset will affect the model's ability to predict older ages, as they are underrepresented.

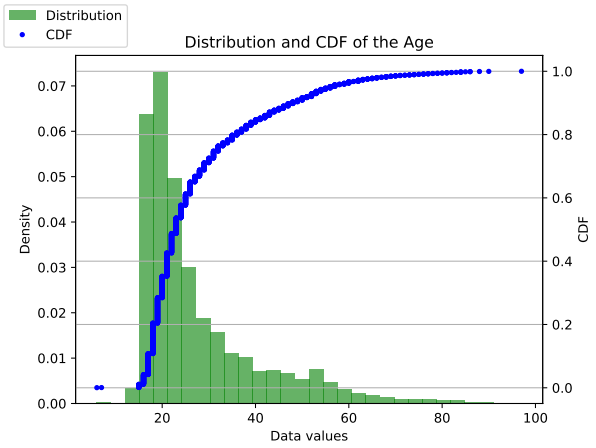

Fig. 2. Age distribution, density bin in green and CDF in blue.

For each sample, a variety of acoustic and linguistic features have already been extracted from the speech signal and are included in the dataset. Additionally, each sample includes information about the individual from whom the recording originates, specifically gender and ethnicity.
Analyzing the ethnicity variable further reveals that the dataset contains over 170 distinct ethnicities saved as a string.
We are also interested in extracting new features from the data. For this purpose, we can analyze a representation of a recording in the time and frequency domains and observe that high frequencies tend to be weaker than low frequencies.

II. Proposed approach

A. Preprocessing

The file path column has been removed. Both the gender and ethnicity variables have been refacotored using a one-hot encoding.
For the extraction of features from the signal, we must first consider what was previously observed: high frequencies in the signal are significantly weaker than low frequencies. To obtain better results, we aim to amplify these frequencies. Therefore, we can apply a simple high-pass filter (as shown in Fig.3, and Fig.4), known as a pre-emphasis filter [1].

$$y[n] = x[n] - \alpha \cdot x[n-1]$$

Where:

- $y[n]$: Sample of the pre-emphasized signal.
- $x[n]$: Current sample of the original signal.
- $x[n-1]$: Previous sample of the original signal.
- α: Pre-emphasis coefficient, usually between 0.9 and 0.99.

Then, we used the emphasized signal to compute the Mel-Frequency Cepstral Coefficients (MFCCs) [1].
To obtain features usable in the model, we computed the mean and standard deviation of the 13 MFCC coefficients, (it makes sense to compute them, as we obtain a value for each coefficient for every 30 ms window of the sample, as shown in Fig.5,), in this way we obtain an even number of features across all samples.
Also other features have been extracted following the approach described in the paper "Talker age estimation using machine learning," [2]:

- **Mean Fundamental Frequency (F0 mean):** The average pitch of the voice, representing the central tendency of vocal frequency.
- **Standard Deviation of Fundamental Frequency (F0 std):** The variability in pitch, indicating how much the frequency deviates from the mean.
- **Pitch Strength:** A measure of the prominence of the perceived pitch, which reflects the clarity and robustness of the fundamental frequency.
- **Standard Deviation of Voicing Intensity:** The variability in the loudness of the voice, representing fluctuations in vocal energy.
- **Alpha Ratio:** The ratio of spectral energy between the 1-5 kHz and 0-1 kHz frequency ranges, used to quantify spectral balance.
- **Spectral Slope:** The difference in spectral energy between the 0-1 kHz and 1-10 kHz ranges, providing insight into the high-frequency content of the speech signal.
- **Spectral Tilt:** The slope of the trendline of the long-time average spectrum, indicating the overall distribution of energy across frequencies.
- **Human Factor Cepstral Coefficients (HFCCs):** Acoustic features derived from the cepstrum, capturing important characteristics of the speech signal.

All the mentioned features have been extracted using the librosa Python library [3].

B. Model selection

Two models have beeen tested:

1) Random Forest: This algorithm combines multiple decision trees, each trained on different subsets of data and features, to make predictions. By aggregating the results, it reduces the risk of overfitting commonly associated with individual decision trees while retaining some level of interpretability. Like decision trees, random forests process one feature at a time, so normalization is not required.
2) Lasso (Least Absolute Shrinkage and Selection Operator): This regression algorithm enhances prediction accuracy and interpretability by adding an L1 regularization term to the linear regression cost function. This penalty shrinks some coefficients to zero, effectively performing feature selection by eliminating less important variables, in this way it prevent overfitting and reduces model complexity. Since it operates on standardized coefficients, normalization of the input data is required before applying the algorithm, in this case we used the Z-standardization,

$$Z = \frac{X - \mu}{\sigma}.$$

For both the best-working configuration of hyperparameters has been identified through a grid search.

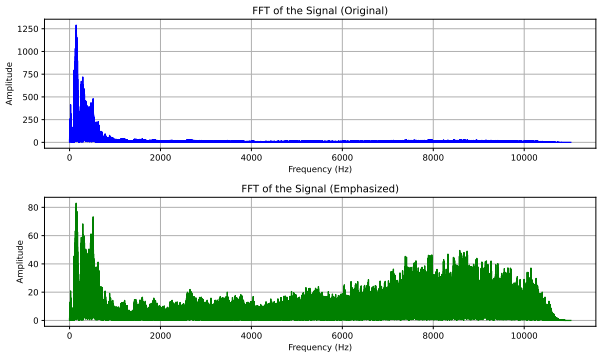

Fig. 3. Original signal (in blue) and enphasized signal (in green), comparison in the frequency domain.

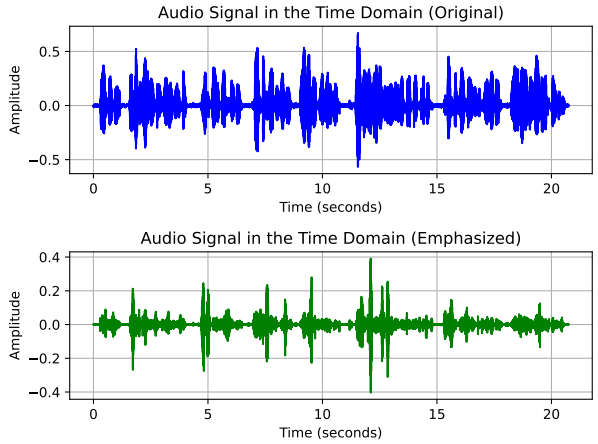

Fig. 4. Original signal (in blue) and enphasized signal (in green), comparison in the time domain.

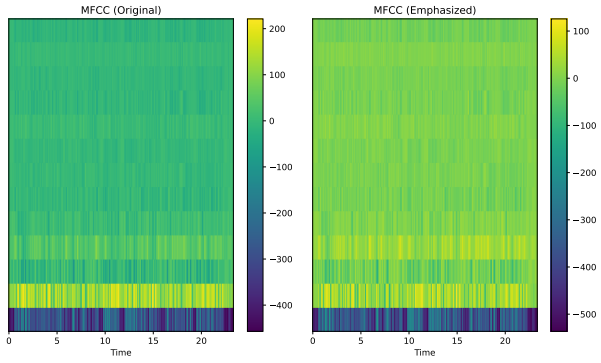

Fig. 5. MFCC coefficients of a sample, arranged in ascending order along the y-axis, with the right showing the values after and the left showing the values before applying the emphasis filter.

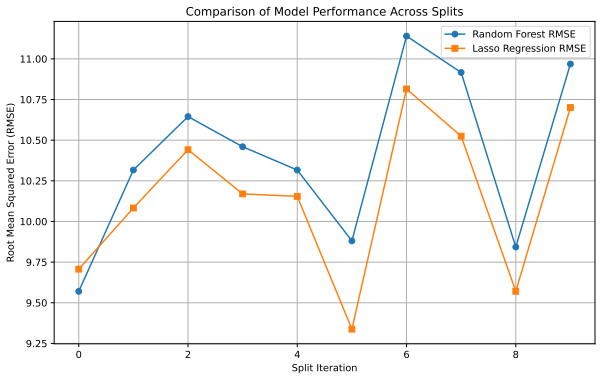

Fig. 6. Comparison between RMSE obtained from LASSO and random forest, considering different training/test split of the dataset using the best model obtained from the parameter grid search.

C. Hyperparameters tuning

The only hyperparameters to set are those related to Random Forest and LASSO. To choose these, we used an 80/20 train/test split on the development set and performed a grid search for the following hyperparameters:

TABLE I
HYPERPARAMETERS CONSIDERED

Model	Parameter	Values
Random forest	max_depth	{None, 10, 20, 30}
	$n\_estimators$	{100, 200, 300, 400, 500 }
	$min\_samples\_split$	{2, 5, 10 }
	$min\_samples\_leaf$	{1, 2, 4, 6, 8 }
LASSO	α	{0.001, 0.01, 0.1, 1.0, 10.0}

After this process we selected the best model for both.

III. RESULTS

Using the best models obtained from the grid search, we proceed by comparing the two models, performing several 80/20 splits of the development set and verifying the performance of the two models for each split by calculating the RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

where $\hat{y}_i$ denotes the predicted value, and y_i is the true value of the i-th observation. The results obtained for each split are shown in Figure 6, from witch we can observe that, generally, better results are obtained using LASSO compared to Random Forest. To conclude, we trained the best-performing Random Forest classifier and LASSO on all available development data. The scores achieved for the public part of the evaluation test are: RMSE for random forest $= 9.97$,
RMSE for LASSO $= 9.67$.
Again LASSO performed better than random forest but not by a big margin.

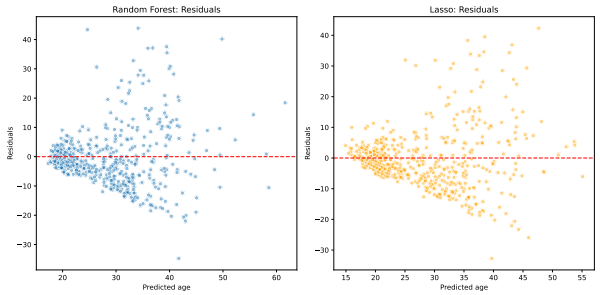

Fig. 7. Residual graph for the best models calculated over an 80/20 split on the development dataset.

IV. DISCUSSION

Analyzing the residuals obtained from the models on the test set (Fig.7), we can observe that these values for both models tend to increase as the estimated age grows. This is a consequence of what was observed regarding the distribution of ages in the development dataset (Fig.2), in section 1.

To further enhance the results, the following aspects could be explored:

- **Higher-Order Polynomial Features with LASSO**: While we employed polynomial features of degree 2 due to computational constraints, it would be valuable to investigate higher-degree polynomial features. This could reveal whether LASSO can identify more complex patterns.
- **Advanced Classification Algorithms**: Employing more complex models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) could provide insights into how advanced architectures might address this problem. These methods are particularly promising for automated feature extraction and could potentially yield better performance.
- **Exploring Windowing Techniques for Enhanced Feature Extraction**: One aspect that could be explored is the use of windowing techniques to extract features at a higher level of detail by segmenting the audio signal.

This involves experimenting with different configurations, such as overlapping and non-overlapping windows, as well as fixed-size and adaptive windows, to assess whether newly identified, potentially correlated features could enhance the model's performance.

REFERENCES

[1] A. A. Abdulsatar, V. Davydov, V. Yushkova, A. Glinushkin, and V. Y. Rud, "Age and gender recognition from speech signals," in *Journal of Physics: Conference Series*, vol. 1410, p. 012073, IOP Publishing, 2019.

[2] M. L. Berardi, E. J. Hunter, and S. H. Ferguson, "Talker age estimation using machine learning," in *Proceedings of Meetings on Acoustics*, vol. 30, AIP Publishing, 2017.

[3] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python.," in *SciPy*, pp. 18–24, 2015.