Age Estimation Through Speech: A Feature-Driven Random Forest Approach

Fabio Depetro
Politecnico di Torino
Student id: s347102
fabio.depetro@studenti.polito.it

Duccio Lalli
Politecnico di Torino
Student id: s345062
duccio.lalli@studenti.polito.it

***Abstract*—This report presents a regression model to estimate a speaker's age based on vocal characteristics. The model leverages the extraction of acoustic and linguistic features from audio recordings, including Mel Frequency Cepstral Coefficients (MFCCs), and uses a random forest regressor as a driver for the regression. The model achieves good accuracy, exceeding the reference threshold set by the challenge.**

I. Problem overview

This report focuses on the development of a regression model designed to estimate the age of a speaker from audio tracks. This problem falls under the speech processing subfield known as speaker profiling, which involves characterizing demographic, and personal information from vocal signals.

For the development of the model, a dataset was provided in two parts:

- A *development set* containing 2,933 recordings and a table with additional information related to the audio tracks, including the age of the speaker (target).
- An *evaluation set* consisting of 691 recordings and its corresponding table.

In both sets, the additional table has a number of rows equal to the number of recordings and a total of 19 columns containing:

- metadata (e.g., sample rate, gender, ethnicity, speaker ID, etc.)
- pre-computed audio features (e.g., pitch, jitter, shimmer, etc.)
- linguistic features (e.g., num_words, num_characters, etc.) related to the audio tracks.

The purpose of the challenge was to predict the age (target) of the subjects in the evaluation set, starting with the provided features and adding others.

Upon examining the provided dataset, it becomes clear that certain features in the tables require careful handling, particularly:

- The categorical feature 'ethnicity' must be preprocessed taking into account the cardinality of ethnic groups, since there is an imbalance in their representation in the original dataset (Figure 1).
- The 'tempo' feature must be converted into a usable format.
- Particular attention should be given to the correlation between similar features (e.g., 'num_words' and 'num_characters') and distinguish those that may be informative for estimating age from those that only correlate with the sentence spoken by the subjects, enabling an effective initial pruning of the features.

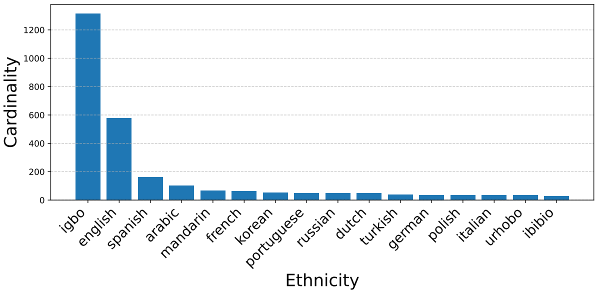

Fig. 1. Ethnic group cardinalities in the given dataset

Regarding the provided set of audio recordings, they were sampled at 22 kHz, which, according to the Nyquist-Shannon sampling theorem, ensures that the signal is adequately captured without loss of information. The frequency spectrum generally considered for speech processing is between 50 Hz and 8 kHz. In particular, it is known that the fundamental frequency of men, women and children is between 80 and 300 Hz [1], with formants developing up to 4 kHz, therefore within this frequency we expect to have most of the information content for optimal feature extraction.

In order to determine which features to extract from these recordings, considerations on the nature of the problem were made, taking into account various scientific studies that investigate the vocal changes associated with aging. Key aspects of these changes include variation in the fundamental frequency and overall frequency properties [1], as well as an increase in irregularities in voice stability and a decrease in voice amplitude due to the weakening of the vocal cords, a reduced lung capacity and the loss of elasticity in the larynx [2].

Starting from these considerations on the aging of the voice, it was possible to decide which features to choose and select those that were faithfully representative of such variations, both in the time and frequency domains in order to consider complementary information on each signal.

II. Proposed approach

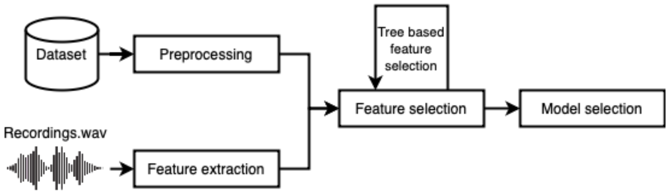

Fig. 2. Pipeline overview

A. Preprocessing

The preprocessing of the provided tables was carried out following some basic steps to prepare the dataset in a suitable format for analysis and modeling.

- The dataset did not contain any NaN values; therefore, no data was removed.
- One-hot encoding was applied to the ‘gender’ variable to make the feature usable for modeling.
- We grouped ethnicities with fewer than 50 samples into an ”other” category (Figure 3), to mitigate sparsity issues in both the development and evaluation sets. The threshold was chosen to ensure a good characterization of the dataset and a manageable number of distinct ethnic groups. This also prevented having too many low-significance categories, making it feasible to apply one-hot encoding to these categorical variables.
- In line with the considerations outlined in the Introduction regarding the correlation between similar features, we performed a detailed analysis to identify and address multicollinearity. Specifically, we calculated the correlation matrix for ’num_words’, ’num_characters’, ’silence_duration’ and ’time_len’ -a feature that accounts for the duration of the recording- to quantify their interdependence and their correlation with the target variable ’age’ (Figure 4). The three features with the weakest correlation to age were pruned, retaining only ’time_len’.
- Other labels such as ’sample ID’, ’sampling_rate’ and ’file_path’ were removed.

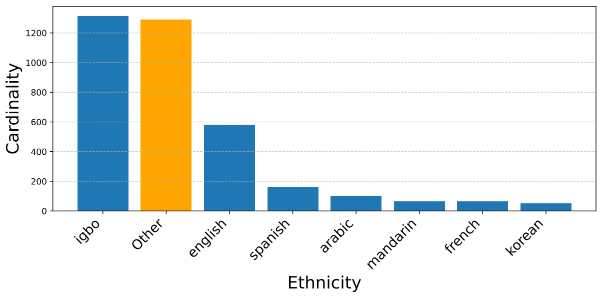

Fig. 3. Grouped ethnic groups

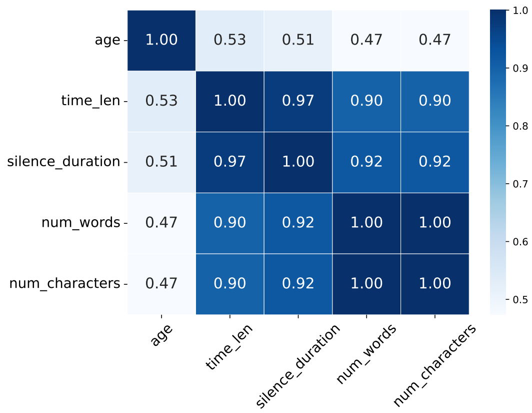

Fig. 4. Correlation matrix

As part of the audio preprocessing, since some audio recordings contained high-frequency electronic background noise, we applied a 4th-order Butterworth low-pass filter with a cutoff frequency of 4000 Hz to prepare them for feature extraction, keeping only the information in the frequency bands relevant to us.

To reduce computational complexity, only the first 5 seconds of each recording were extracted. This duration was selected because it captures enough meaningful information while keeping the segment size manageable. Additionally, by focusing on the initial part of the audio, we minimize the influence of the specific sentence spoken by the subjects and its temporal variations, ensuring that the extracted frequency features are more consistent and generalizable. To ensure that all subjects spoke for a sufficient duration within the initial 5 seconds, enabling meaningful analysis, we first examined the onset of speech for each subject. As shown in Figure 5, all subjects started speaking within the first 3.6 seconds.

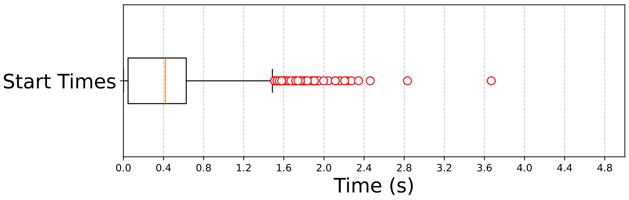

Fig. 5. Distribution of speech start times

After completing the audio preprocessing stage, we extracted features from the recordings using the open-source Python library Librosa.

Given the diversity of sentences and languages in the dataset, including tonal languages, we selected a wide range of features to capture the varied nuances in speech characteristics:

- ***Temporal features***: These were used to capture variations in amplitude (voice intensity), root mean square energy and voice stability over time. Normalization of the signal was avoided to preserve the natural dynamics of these features and maintain their inherent variations.
- ***Spectral features***: These features provide insights into the energy distribution across frequencies, which is crucial

for capturing distinct voice timbres and tonal variations. We computed features such as spectral centroid, roll-off, flatness and bandwidth.

- ***Mel-Frequency Cepstral Coefficients***: MFCCs are widely used in speech processing [3] to capture the power spectrum's envelope on the perceptual mel scale (Figure 6). They effectively represent vocal characteristics like pitch, tone, and timbre.
- ***Chroma features***: These features represent the harmonic and tonal content of the audio by analyzing energy distribution across the 12 chromatic pitch classes [4].

All these features were extracted from overlapping signal frames and then aggregated over time using classical aggregation functions, including mean, median, and standard deviation. A total of 108 features were extracted from each audio sample.

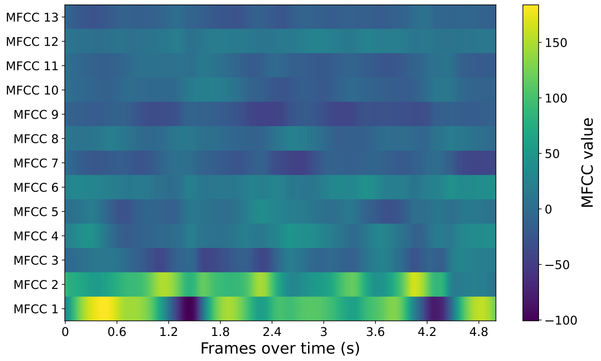

Fig. 6. Visualization of MFCCs Computed on Signal Frames

B. Model selection

The following algorithms have been tested:

- *Random Forest Regressor*: it is well-suited for predicting a subject's age, as it can model complex, nonlinear relationships. It is robust to noise and overfitting, requiring minimal data preprocessing (e.g., no normalization). Furthermore, it can automatically identify the most important features, which is particularly useful when relationships between features are not easily discernible, as in this case.
- *Support Vector Regression(SVR)*: it is also a good model for our prediction task [5]. However, it requires some pre-processing, such as feature scaling, for optimal performance. It should be noted that SVR performance is highly dependent on the choice of kernel, making it more effective when the data has a well-defined structure that can be captured in a higher dimensional space.

By combining the extracted features with those from the original tables, a total of 131 features are obtained. At this stage, feature selection is performed to reduce dimensionality, remove redundant features, and prevent overfitting. To perform feature selection, we used a tree-based feature selection approach: we applied a default random forest regressor to identify an initial set of relevant features by selecting all those with a 'feature_importance' greater than 0.009, i.e. a total of 12 *'best features'*.

During this process, we noticed that the one-hot encoding of the 'English' ethnicity had a very high feature importance. Further analysis of ethnicity distributions revealed that this specific group was exclusively present in the development set, suggesting potential overfitting. To address this issue, we reintroduced the one-hot encoding for all ethnicities.

To further expand the *'best features'* set, we introduced additional variables based on a well-defined criterion: we selected only features with a high correlation with the 'age' while maintaining a low correlation with the features already present in the list of *'best features'*. This approach aimed to generalize the model, avoiding the inclusion of highly correlated features that could add redundancy.

To validate this process, we performed 5-fold cross-validation on the 80/20 train/test split for each new feature. A decrease in the mean RMSE after adding a feature (or a subset of features) indicated its positive contribution to the model's performance. Features that did not improve the RMSE were excluded from the model.

With this approach, we obtained a selected set consisting of 23 *'best features'* : {*time_len*, *ethnicity_english*, *Chroma7_median*, *MFCC10_std*, *jitter*, *Chroma6_median*, *Chroma12_std*, *spectral_centroid_mean*, *MFCC7_median*, *Chroma6_std*, *MFCC6_mean*, *shimmer*, *ethnicity_igbo*, *ethnicity_spanish*, *ethnicity_arabic*, *ethnicity_korean*, *ethnicity_mandarin*, *ethnicity_french*, *ethnicity_Other*, *time_skew*, *mean_pitch*, *MFCC2_mean*, *time_max*}

C. Hyperparameters tuning

After the feature selection stage, we performed a grid search on the development set, split in an 80/20 ratio as before, to identify the optimal hyperparameter configuration for both the SVR and Random Forest model, based on the parameters defined in Table I.

Model	**Parameter**	**Values**
Random Forest Regressor	`n_estimators`	{500}
	`max_depth`	{20, 25, 30}
	`min_samples_split`	{2, 5, 6, 7}
	`min_samples_leaf`	{2, 3, 4, 5}
	`max_features`	{'sqrt', 'log2'}
	`random_state`	42
SVR	`kernel`	{'linear', 'rbf', 'poly'}
	`C`	{0.1, 1, 5, 50, 100}
	`epsilon`	{0.01, 0.1, 1, 5, 10}
	`gamma`	{'scale', 'auto', 0.1, 0.01}

TABLE I
HYPERPARAMETERS FOR RANDOM FOREST REGRESSOR AND SVR MODELS.

III. RESULTS

By performing hyperparameter tuning, the following optimal hyperparameter configurations were obtained from grid search:

Model	Parameter	Value
Random Forest Regressor	`n_estimators`	500
	`max_depth`	20
	`min_samples_split`	5
	`min_samples_leaf`	2
	`max_features`	'sqrt'
	`random_state`	42
SVR	`kernel`	'rbf'
	`C`	100
	`epsilon`	5
	`gamma`	0.01

TABLE II
BEST HYPERPARAMETER CONFIGURATIONS FOR THE RANDOM FOREST REGRESSOR AND SVR MODELS.

As for the RMSE scores, the results are shown in Figure 7: for the Random Forest Regressor, the average RMSE across the 5 folds was ≈ *9.651*, and an RMSE of ≈ *9.719* was achieved on the Public Leaderboard. In contrast, for the SVR model, the average RMSE across the 5 folds was ≈ *9.787*, with an RMSE of ≈ *9.549* on the Public Leaderboard.

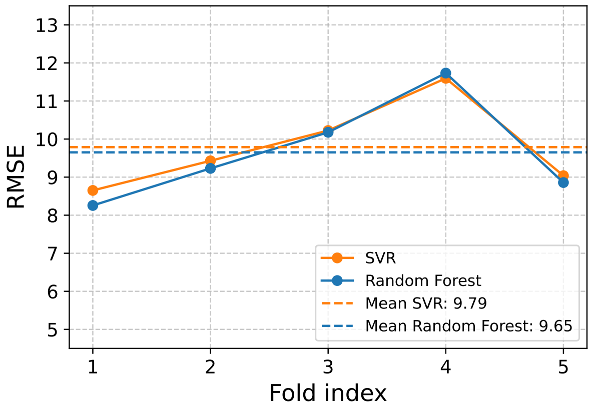

Fig. 7. RMSE scores obtained across the 5 folds

Figure 7 illustrates that both models produce satisfactory and comparable results. The difference in RMSE for the SVR could be attributed to variability between the cross-validation splits and the evaluation set.

To provide context for the obtained scores, we referred to the minimum threshold set on the public leaderboard (RMSE = **11.179**) and a second reference threshold. The second threshold was defined by using only the provided table, without any feature extraction or selection, and training a default Random Forest Regressor. This approach achieved an RMSE of **10.192** on the leaderboard, which we adopted as the second benchmark.

Compared to the public leaderboard, our approach ranks slightly above the middle, which indicates that our solution is solid and demonstrates the effectiveness of our approach.

IV. DISCUSSION

The proposed approach achieves results that surpass both the baseline set by the public leaderboard and the naive solution obtained using the previously mentioned simple approach.

Several aspects can be considered to further improve the results.

- Formant extraction could be performed, as formants are known to vary with age [1] and could potentially enhance the model's performance.
- Feature selection could also be extended by exploring non-linear correlations using Spearman and Kendall coefficients to identify additional feature combinations. This may lead to the discovery of a better performing subset of *best features*, since only a limited selection of potential features has been used in the current approach.
- A promising alternative lies in utilizing automated feature extraction techniques from signal spectrograms, such as Convolutional Neural Networks (CNNs).
- A more extensive grid search could be undertaken, as only a limited set of hyperparameters has been explored so far.

The results, however, are already solid, both in absolute terms and when compared to those on the public leaderboard and the baselines, demonstrating the robustness of the proposed approach.

This regression task proved challenging, as identifying features derived from audio signals that correlate with age is not easy. Furthermore, the distribution of ethnicities, languages, sentences, and ages in the dataset was not uniform, further complicating the task.

REFERENCES

[1] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," vol. 2, 01 1995.

[2] A. Dehqan, R. Scherer, G. Dashti, A. Ansari-Moghaddam, and S. Fanaie, "The effects of aging on acoustic parameters of voice," *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, vol. 64, pp. 265–270, 01 2013.

[3] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122136–122158, 2022.

[4] V. Chernykh and P. Prikhodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[5] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, pp. IV–1085–IV–1088, 2007.