Regression Pipeline for Age Estimation from Speech Signals

Stefan Alexandru Patru
Politecnico di Torino
s338342
s338342@studenti.polito.it

Andrea van der Putten
Politecnico di Torino
s284202
s284202@studenti.polito.it

***Abstract*—In this report, we introduce a possible approach to the speaker age estimation problem using the provided speech dataset. Specifically, the proposed approach involves utilizing both acoustic features extracted from the audio signals and additional features provided with the dataset to build a regression pipeline. By combining these feature sets, we aim to enhance the model's ability to capture relevant patterns and improve the accuracy of age predictions. We implemented and tested three regression models to assess their performance on the provided evaluation dataset, aiming to identify the most effective model. Eventually, two baseline models have been constructed without performing any preprocessing and feature extraction operations in order to highlight the importance of these techniques in improving the accuracy of age predictions.**

I. Problem Overview

The dataset consists of:

- 2933 samples designated for the *development* set
- 691 samples allocated for the *evaluation* set.

Each sample represents a spoken sentence, and the corresponding speaker's age serves as the target variable.

The development set will be employed to construct a classification model capable of accurately categorizing the data points in the evaluation set.

Upon analyzing the development set, we observed that:

- All the samples have the same sampling rate of 22050 Hz, ensuring consistency in the temporal resolution of the audio data.
- The duration of the recordings varies across samples. As illustrated in Figure 1, there is a significant spike at the shortest durations, followed by a moderate increase peaking around 20-30 seconds. Frequencies gradually decline as the durations increase beyond 30 seconds, showing fewer longer recordings. A few outliers are observed near the upper limit of the duration range, specifically between 80 and 100 seconds.
- The dataset exhibits an imbalance, with a significantly higher number of younger individuals compared to older ones, as illustrated in Figure 2. This imbalance could bias the regression model towards predicting ages closer to the majority class, potentially leading to less accurate predictions for older individuals.

To gain a better understanding of the data at hand, we chose to visually inspect some signals in both the time and frequency domains. As illustrated in Figure 3, we can observe some noise in the waveform as random, irregular fluctuations or spikes that do not seem to correspond to the main signal patterns.

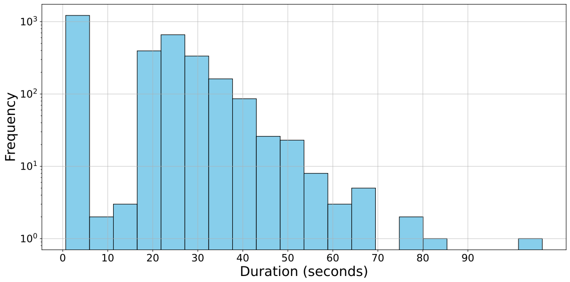

Fig. 1. Distribution of the duration of the recordings

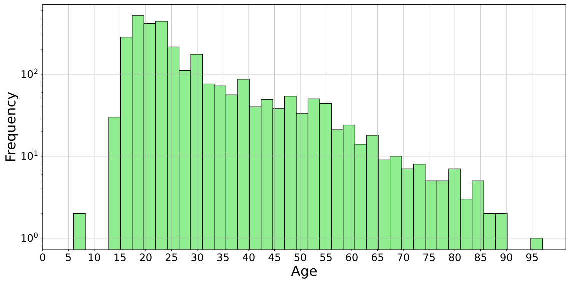

Fig. 2. Distribution of the age of the people

Similarly, in the Mel spectrogram [1] in Figure 4, noise can be observed as sporadic energy spread across the frequency spectrum, especially in regions where there is no clear structure or consistent frequency bands. The noise may degrade the model's performance by reducing the precision of feature extraction and increasing the variability in the training data.

We observed that certain values in the dataset appeared to be erroneous. For instance, the 'num_words' feature was often equal to 0, and the silence duration was disproportionately high compared to the total duration of the recording. Upon analyzing these samples, we discovered that the language spoken in the recordings was not English.

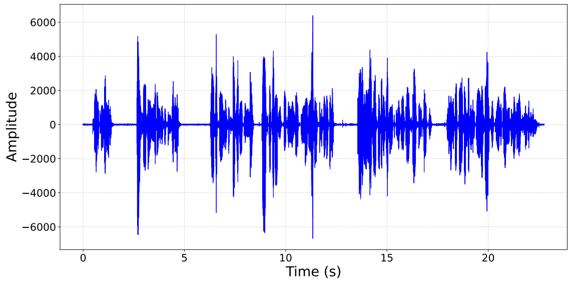

Fig. 3. Waveform of a recording

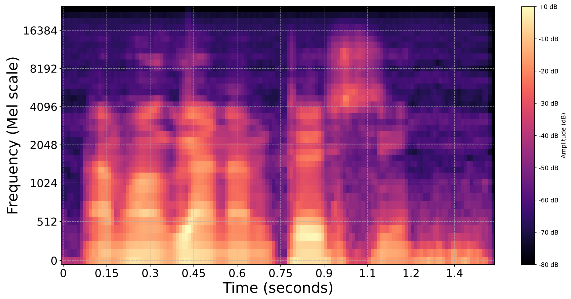

Fig. 4. Mel spectrogram of a recording

II. Proposed Approach

A. Data Preprocessing

1) Feature Selection: Based on the features provided in the dataset, we deemed it appropriate to conduct an analysis to select the most significant features for our model. To this end, the correlation matrix of all the features was plotted. For visual clarity, Figure 5 displays the correlation matrix limited to the most relevant features.

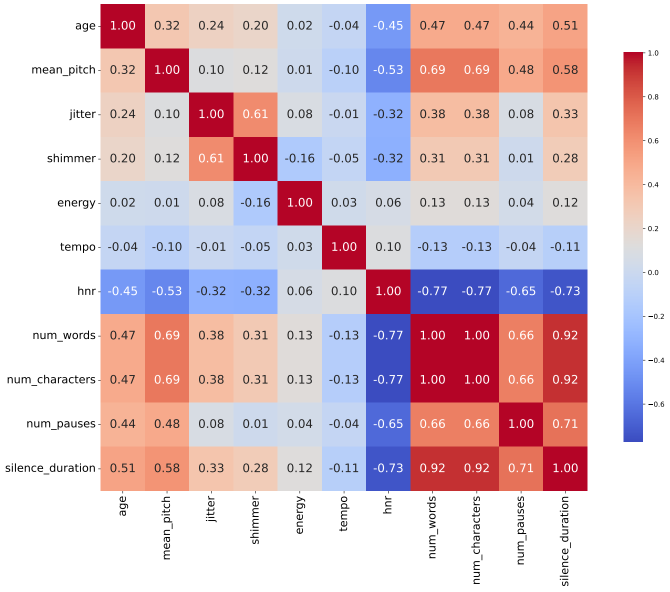

Fig. 5. Correlation matrix

As observed in the lower-right corner of the matrix, several features exhibit high correlation. Consequently, we decided to exclude the number of words and the number of characters, as they were strongly correlated with the silence duration of the recording.

Additionally, the matrix provides valuable insights into which features are most strongly associated with the speaker's age in the recording. For instance, silence duration appears to be very useful in predicting ages. Other notable features include the harmonic-to-noise ratio (HNR), which shows a negative correlation with age, and the mean pitch, which also demonstrates relevance in age determination.

Due to their low correlation with age, with values close to zero, we have also decided to exclude the tempo and energy features from our analysis, as well as the id, the path, and the sampling rate of the samples.

2) Data Cleaning: As outlined in the problem overview, some portions of the data were identified as inaccurate. During this phase of preprocessing, we carried out the following steps to address these issues:

- We examined the distribution of each feature using boxplots to identify potential outliers. However, after conducting a meticulous analysis of the audio samples exhibiting unusual values, we found that most of these values were consistent with the actual recordings and not indicative of errors. Consequently, we decided to retain these samples in the dataset.
- We assessed the dataset for missing values and duplicate entries but did not identify any occurrences of either.
- Some invalid values were found in the gender feature and were replaced with the correct ones.

3) Feature Extraction: To gain a deeper understanding of the recording characteristics, we have decided to extract additional features directly from the recordings using the `librosa` library [2]. The features that have proven to be particularly valuable for our regression task include:

- **Mel Spectrogram:** To further analyze the spectral properties of the audio, we extracted some features from the Mel spectrogram, which represents the signal's energy distribution across 10 frequency bands on the Mel scale. Specifically, for each frequency band, the mean and standard deviation of the log-transformed Mel spectrogram values were computed. These statistics provide a summary of the energy distribution and variability within each band, capturing key patterns while mitigating the influence of sporadic noise.
- **MFCCs:** Mel-Frequency Cepstral Coefficients (MFCCs) are a collection of features commonly used to represent speech signals. Derived from the signal's frequency spectrum, MFCCs are widely utilized in applications like automatic speech recognition. They capture key characteristics of the voice signal, such as the shape of the vocal tract [3]. In this study, 13 MFCCs were extracted for each audio recording. To summarize the spectral characteristics, the mean and standard deviation

of each coefficient were calculated across all time frames, representing both the average behavior and variability of the spectral features throughout the recording.

- **Chroma:** Chroma features are a representation of the energy distribution across the 12 pitch classes of the chromatic scale, making them a powerful tool for analyzing harmonic content in audio signals [4] .To understand the harmonic content of the audio, we determined the average energy and its fluctuation for each musical note (pitch class) across the entire sound. This analysis provides a concise summary of the harmonic structure of the audio by revealing the typical energy level and its variability for each note.
- **Spectral Flatness:** It is a measure that quantifies how ”flat” or ”tonal” a spectrum is. It is calculated as the ratio of the geometric mean to the arithmetic mean of the power spectrum. [5]. We summarized it by computing the mean and skewness of the spectral flatness values across time frames.
- **Spectral Bandwidth:** Spectral bandwidth quantifies the spread of energy across the frequency spectrum of a signal. It is calculated as the weighted standard deviation of the signal's frequencies relative to its center frequency. Signals with focused harmonic content have narrow spectral bandwidth, while those with broader energy distribution, such as noisy or breathy signals, exhibit wider bandwidth [6]. We computed the mean, standard deviation, and maximum values, which capture the overall spread, variability, and peak energy distribution of the signal.

4) Feature Transformation: Following feature selection, we retained some categorical features in the dataset, such as ethnicity and gender. Since most models perform better with numerical data, we applied one-hot encoding to these features.

However, due to the large number of possible ethnicities, which added unnecessary complexity to the problem, we chose to retain only the English ethnicity as a binary feature, as it was the most strongly correlated with the age of the person.

Since the features in the dataset had significantly different scales, some of the models we plan to use require the data to be scaled for optimal performance. To address this, we applied standard scaling and min-max scaling in order to normalize the features.

B. Model Selection

The following regressors have been tested:

- **Random Forest Regressor:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and averages their predictions to provide a final continuous output, making it ideal for tasks with continuous target variables like age [7]. It provides an intrinsic measure of feature importance, enabling us to identify which audio features contribute most significantly to age prediction. Random forests, like decision trees, work on one feature at a time, so no standardization of the data is necessary.
- **Ridge:** The Ridge model is a type of linear regression that incorporates L2 regularization to improve generalization and mitigate overfitting, especially when working with datasets that have multicollinearity or high-dimensional feature spaces. It minimizes the sum of squared residuals while adding a penalty term proportional to the square of the coefficients.
  To enhance the model's performance, we built a pipeline that includes the application of polynomial features, allowing the model to capture higher-order relationships between the input variables. This step helps in modeling non-linear patterns in the data by adding polynomial terms. Additionally, we applied min-max scaling to the features to ensure they are on the same scale, which is important for regularization methods like Ridge, as they are sensitive to the magnitudes of the features.
- **Lasso:** The Lasso model is another form of linear regression that incorporates regularization, but instead of L2 regularization like Ridge, it uses L1 regularization, which applies a penalty proportional to the absolute value of the coefficients.
  We constructed the same pipeline for the Lasso model as we did for the Ridge model, but using the standard scaling instead of the min-max one.

C. Hyperparameters Tuning

To tune the hyperparameters, we divided the development set into training and validation subsets using an 80/20 split. This procedure was carried out to perform the hyperparameter tuning on the training set, allowing us to optimize the model based on its performance during training. Subsequently, the tuned model was tested on the remaining validation set, which consisted entirely of unseen data, to evaluate how well the chosen hyperparameters generalized to new, unseen samples.

The tuning process was conducted for both the hyperparameters of the regression models and the degree of the polynomial feature transformations. For Ridge and Lasso regression and polynomial features, the tuning was performed within a pipeline, allowing us to jointly select the optimal value of α in both regression models corresponding to the degree used in the polynomial feature transformation. This approach ensured that the interaction between these two hyperparameters was taken into account, leading to a more effective model configuration.

The hyperparameters (presented in Table I) were optimized through a grid search to identify the best combination that maximizes the model's performance, with a 5-fold cross-validation.

III. Results

After running the grid search, we found the optimal hyperparameters for the three models. This procedure helps prevent overfitting, as the models were consistently optimized using the validation set. To ensure unbiased evaluation, the test set was never used to inform decisions during the model-building process.

Model	Parameter	Values
Polynomial Features + Ridge	*degree*	{1, 2, 3}
	alpha	{0.5, 1, 2, 3, 4, 5}
	max_iter	{500, 1000, 2500}
Polynomial Features + Lasso	*degree*	{1, 2, 3}
	alpha	{0.1, 0.15, 0.2, 0.5, 1}
	max_iter	{500, 1000, 2500}
Random Forest Regressor	*n_estimators*	{100, 500, 800}
	max_depth	{50, 100, 250, *None*}
	min_samples_leaf	{3, 5}
	min_samples_split	{2, 4, 6}
	max_features	{*sqrt*, log_2}

TABLE I
HYPERPARAMETERS CONSIDERED

The best configuration for Random Forest was found for {*n_estimators=800, max_depth=50, min_samples_leaf=3, min_samples_split=2, max_features=sqrt*}, which leads to a value of RMSE ≈ 9.70476 on the validation set.

On the other hand, the optimal hyperparameters for the pipeline consisting of polynomial features and Lasso regression were identified as {*degree=2, alpha=0.15, max_iter=500* }, while the pipeline with Ridge was optimized for {*degree=2, alpha=5, max_iter=500* }.

Evaluating these models on the validation data gives a value of RMSE ≈ 9.1978 for the Lasso and ≈ 9.1626 for the Ridge.

After training the three models on the entire development dataset, we employed them to generate predictions for the evaluation set. The Random Forest achieved a public score of RMSE ≈ 9.682, while the Lasso and the Ridge obtained respectively an RMSE ≈ 9.280 and ≈ 9.404. These two submissions will be evaluated on the private set.

In order to make a comparison with our results, we implemented two baseline models without applying any preprocessing or incorporating additional features. Moreover, we have not performed any tuning for the hyperparameters. The first baseline model was a random forest regressor, achieving an RMSE ≈ 10.521 on the public leaderboard. The second baseline was a simple linear regressor, which performed slightly better, with an RMSE of approximately 10.248.

IV. Discussion

As we can observe from the results, our customized models achieved a lower RMSE compared to the baseline models.

The enhancements achieved through preprocessing and hyperparameter tuning highlight the importance of fine-tuning and careful data preparation in achieving more accurate and reliable predictions. In particular, the extracted features played a crucial role in improving the model's performance by providing more informative inputs, enabling the model to better capture underlying patterns in the data.

One of the major challenges we faced during model optimization was the uneven age distribution in the dataset. This imbalance led to the model performing more accurately when predicting the ages of younger individuals, particularly those between 20 and 50 years old, while struggling with older age groups. This issue is evident in the Figure 6, which illustrates the distribution of absolute prediction errors on the validation set based on the age of the individuals being predicted.

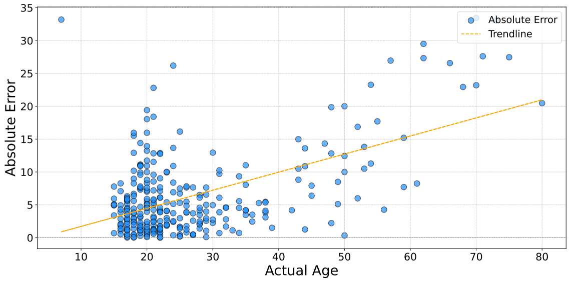

Fig. 6. Distribution of the errors using Random Forest Regressor

The orange trendline emphasizes the general tendency of the model's errors, showing how the prediction errors increase as the age of the individuals moves beyond the 50-year mark.

Here are some aspects that could be considered to further enhance the results obtained:

- As highlighted in the problem overview, several recordings in the dataset are not in English. This affects various features, such as the number of words, characters, and silence duration. To address this issue, it would be beneficial to utilize specialized libraries capable of handling multiple languages, enabling the accurate processing and updating of these features for non-English recordings.
- We may consider extracting additional features that could improve the model's ability to better recognize older individuals.
- Instead of predicting the exact age, grouping individuals into broader age categories (e.g., 20-40, 41-60, 61+) might help the model focus on learning the distinctions between these categories, improving accuracy for older individuals.
- A potential direction for improving the model's performance is the exploration of Neural Networks, which could leverage their capacity to learn hierarchical and complex representations from the data. Unlike traditional models, Neural Networks can automatically discover patterns in features that might not be immediately apparent.

References

[1] W. Endres, W. Bambach and G. Flosser, ”Voice Spectrograms as a Function of Age Voice Disguise and Voice Imitation”, The Journal of the Acoustical Society of America, vol. 49, no. 6, pp. 1842-1848, Jun. 1971

[2] B. McFee et al., ”librosa: Audio and Music Signal Analysis in Python”, Proceedings of the 14th Python in Science Conference, 2015.

[3] E. Joliveau, J. Smith and J. Wolfe, ”Vocal tract resonances in singing: The soprano voice”, The Journal of the Acoustical Society of America, vol. 116, no. 4, pp. 2434-2439, Oct. 2004

[4] K.Shah, M.Kattel, D. Shrestha Department of Computer Science and Engineering, School of Engineering Kathmandu University, Nepal, 2019

[5] Madhu, Nilesh. (2009). Note on measures for spectral flatness. Electronics Letters. 45. 1195 - 1196. 10.1049/el.2009.1977.

[6] S. Mlot, E. Buss, and J.W. Hall III, Ear Hear. Spectral integration and bandwidth effects on speech recognition in school-aged children and adults. 2010 February

[7] Sriram Ravishankar; Prasanna Kumar M.K. - Prediction of Age from Speech Features Using a Multi-Layer Perceptron Model