Age Estimation from Speech: A Regression Problem

Ismail Aljosevic
Politecnico di Torino
s337769
ismail.aljosevic@studenti.polito.it

***Abstract*—This report presents a possible regression approach to estimate the age of the speaker. The proposed method consists of extracting widely used audio features from the signal, computing their summary statistics and combining them with the provided data. In addition, feature reduction techniques are applied. The proposed solution achieves result that exceeds the average of the published results and outperforms approximately 70% of them.**

I. Problem overview

The objective of the proposed assignment is to address the regression problem of estimating the speaker's age. The dataset consists of recordings where each one corresponds to a spoken sentence and it is divided into two subsets:

- a *development* set: contains 2,933 records
- an *evaluation* set: contains 691 records.

In addition to the dataset, acoustic and linguistic features extracted from the speech signals have been provided for each sample. These features will be utilized to construct a regression model to predict the age of the speakers in the evaluation set. Additional features may also be extracted directly from the audio samples to enhance the model's performance.

Based on the information obtained from the development set, certain observations can be made. The age target variable is not balanced, with most speakers falling within the younger age range, particularly between 15 and 30 years old. Figure 1 shows that the age distribution is right-skewed, meaning the number of speakers decreases as age increases. The dataset also includes two categorical features: gender and ethnicity of the speakers. The distribution of gender is balanced, while the distribution of ethnicity is imbalanced. There are 165 unique ethnicities in total, but the top 30 ethnicities account for over 85% of the data.

As shown in [1], pitch and speech rate are important for the age estimation of people. Speech rate has a greater impact on age prediction than pitch, as older individuals typically speak more slowly and take more pauses. This statement is useful for better understanding the data and its relationship to age perception.

Table I presents the 10 features with the highest absolute correlation values with age. Silence duration is intuitively linked to a greater number of pauses and slower speech rate. However, since the duration of recordings varies, silence duration is strongly influenced by that, making the relationships more complex. Unexpectedly, tempo is not among the ten most age-correlated features.

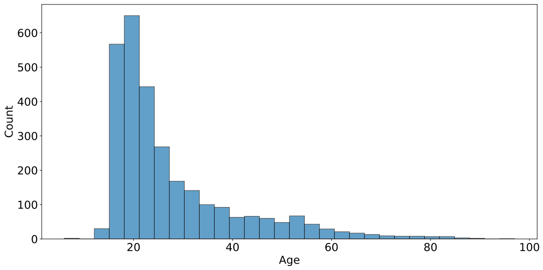

Fig. 1: Age distribution

Rank	Feature	Correlation
1	Silence Duration	0.5141
2	Number of Words	0.4735
3	Number of Characters	0.4734
4	Harmonics-to-Noise Ratio (HNR)	0.4469
5	Number of Pauses	0.4377
6	Mean Pitch	0.3156
7	Zero Crossing Rate (ZCR) Mean	0.2784
8	Jitter	0.2385
9	Maximum Pitch	0.2266
10	Minimum Pitch	0.2242

TABLE I: Absolute feature correlation with age

Besides features related to slow speech rate and pitch, spectral features such as jitter, zero-crossing rate (ZCR), and Harmonics-to-Noise Ratio (HNR), which are commonly used in speech processing, have also shown relevance. However, there are dependencies among all of the features, which highlight the complexity of the dataset, even before extracting additional features.

II. Proposed approach

A. Preprocessing

Due to the complexity of the problem, extracting additional features was necessary to enhance the efficiency and accuracy of age prediction. As part of this process, Librosa Python library was used to extract Mel Frequency Cepstral Coefficients (MFCC), which are one of the most commonly used features in speech processing application. The process of extracting MFCC is very straightforward and it involves the following steps [2]:

1) **Framing:** Dividing the audio signal into short, overlapping frames
2) **Fast Fourier Transform:** Transforming each frame into the frequency domain

3) **Mel Filter Bank:** Transforming the frequency spectrum onto the Mel scale, which is more relevant to human hearing
4) **Logarithm:** Applying a logarithmic operation on Mel-scaled spectrum
5) **Discrete Cosine Transform:** Reducing the number of coefficients by transforming the log-scaled spectrogram into a more compact form.

An example of a MFFC spectrogram of a signal is shown in Figure 2.

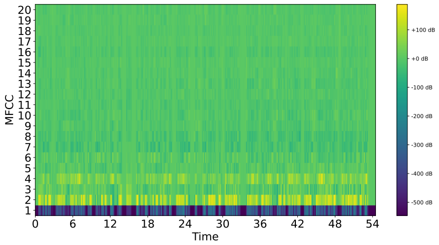

Fig. 2: Mel Frequency Cepstral Coefficients

In the spectrogram, 20 MFCC are represented for each frame of the input signal. Each coefficient captures different aspects of the speech signal. For example, the values for the first MFCC are much smaller than the values for the other coefficients. The reason for this is that the first MFCC primarily captures information from the lower frequencies of a signal, which are associated with the overall energy and loudness. In contrast, higher frequencies provide more information about the finer characteristics of the signal, and the other MFCC are more focused on these frequencies.

In addition to MFCC, other spectral and rhythmic features offered by the Librosa library were extracted, including chroma features, contrast, bandwidth, roll-off, tempogram and tonnetz [3]. For all of these features, the minimum, maximum, mean values, along with their standard deviations, were computed.

For ethnicity and gender features, one-hot encoding must be performed to transform them into a format suitable for the regression models. This process has created separate binary columns for each category.

Intuitively, the features that were irrelevant or clearly redundant were removed. For instance, the sampling rate and ID column were excluded, as they do not contribute to predicting age. Additionally, the number of characters was dropped since it has shown a strong correlation with the number of words, making it redundant for the task.

The addition of new features provided more information about each sample, but it also increased the dataset's dimensionality, especially with the inclusion of one-hot encoded ethnicity values. As a result, the task of feature selection has become more demanding. To address this issue, the Recursive Feature Elimination with Cross-Validation (RFECV) was implemented. RFECV is a technique that iteratively removes the least important features based on model performance, using cross-validation to evaluate performance at each step. It selects the most relevant features by combining results from different folds, improving model performance and generalization.

Based on the [4] where the combination of Principal Component Analysis (PCA) with MFCC, as well as other spectral and rhythmic features, achieved good results, this approach was also applied. PCA is a widely used dimensionality reduction technique. It simplifies high-dimensional datasets by transforming their features into a smaller number of components, known as principal components. These components are linear combinations of the original features and they are ordered by the amount of variance they cover from the data. A component that explains more variance contains more information and captures greater variability in the dataset.

B. Model selection

The following models have been evaluated:

- *Linear Regression:* this model is straightforward to understand and implement. It works by finding the best-fitting line that minimizes the sum of squared differences between the true values and the predicted values, assuming a linear relationship between the input features and the target variable. Its computational efficiency ensures quick training, even when working with a relatively large number of features
- *Ridge Regression:* this model is a regularized version of Linear Regression, particularly effective for high-dimensional datasets where most features are relevant, such as the given dataset. Unlike methods that eliminate features, Ridge regression retains all coefficients but reduces the impact of less important ones by shrinking their values, thereby maintaining feature information while improving generalization and reducing overfitting
- *Random Forest Regressor (RFR):* this model offers several advantages that are suitable for the characteristics of the given dataset. It utilizes multiple decision trees, each trained on a different subset of the data. By averaging the results of these trees, the model improves prediction accuracy and generalization. Additionally, tree-based algorithms handle skewed distributions of the target variable effectively. As the model estimates feature importances, combining it with RFECV is an intuitive step. In related work [4], it has demonstrated competitive performance compared to more complex models.

C. Hyperparameters tuning

Since techniques such as RFECV and PCA were utilized, it was first necessary to determine the appropriate number of selected features and principal components. RFECV with a RFR was applied to identify the optimal number of features. The optimal number of principal components was determined by evaluating results of all models across different thresholds. After identifying the optimal number of features from RFECV and the principal components, Grid Search Cross-Validation was performed for the models. The parameters and their values

Method	Parameters	Values
RFECV	step min_features_to_select	10 145
PCA	n_components	{0.95, 0.96, 0.97, 0.98, 0.99}

(a) Preprocessing parameters

Model	Parameters	Values
Linear Regression	None	None
Ridge Regression	alpha	{0.1:100:0.5}

RFR	n_estimators
	max_depth
	min_samples_split
	min_samples_leaf
	max_features
		{100, 200, 300}
		{None, 10, 20, 30}
		{2, 5, 10}
		{1, 2, 4}
		{log2, sqrt}

(b) Hyperparameter Search Space

TABLE II: Preprocessing and Hyperparameter Tuning

for both preprocessing and model hyperparameters are shown in Table II (a) and Table II (b), respectively.

The Root Mean Squared Error (RMSE) metric was used to evaluate the model's performance, including preprocessing methods, hyperparameter tuning, and results on the evaluation set. RMSE is calculated for n samples with targets $y_1, y_2, \ldots, y_n$ and predictions $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ as follows:

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

III. Results

The optimal number of features selected using RFECV is 171, which reduced the dimensionality of the dataset by 120 features. In particular, all the removed features were some of the one-hot encoded ethnicity features. This outcome was anticipated, as the top 30 ethnicities accounted for more than 85% of the dataset. A comparison of the results with all the features included and with the reduced set of features is shown in Figure 3, where the results remained consistent.

After applying PCA with different thresholds to all models, the best results were achieved by selecting 99% of the principal components for both Linear and Ridge Regression, resulting in a reduction in RMSE for both models. This reduced the dataset's dimensionality by 29 features. Due to the increased RMSE for RFR, PCA was not combined with it, as it significantly impacted the important features for RFR. The results of models at different percentages of explained variance are presented in Figure 4.

The best configuration for RFR was found for { *max_depth*=None, *max_features*='sqrt', *min_samples_leaf*=4, *min_samples_split*=2, *n_estimators*=100 } (RMSE $\approx$ 10.43). For the Ridge Regression model, the best *alpha* values were found to be 99.7 (RMSE $\approx$ 10.49) when only RFECV was applied, and 7.6 (RMSE $\approx$ 10.30) when PCA was additionally applied.

The results achieved on the evaluation set are presented in Table III. Models labeled with "PCA" indicate that PCA was applied, while those without it used only RFECV-selected features. Linear Regression with RFECV-selected features achieved the lowest RMSE on the evaluation set. While PCA improved the results during the analysis, this improvement was not observed on the evaluation set. Unexpectedly, the RFR, despite being a more complex model, obtained the highest RMSE.

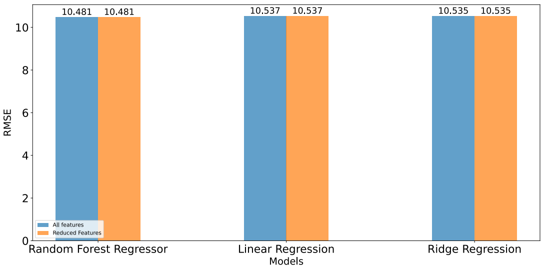

Fig. 3: Comparison of results using all features and the reduced set of features

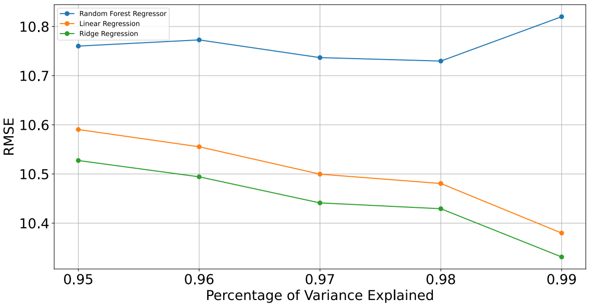

Fig. 4: Results for different perecentages of variance explained

Rank	Model	RMSE
1	Linear Regression	9.590
2	Ridge Regression	9.677
3	Ridge Regression PCA	9.823
4	Linear Regression PCA	9.935
5	RFR	10.062

TABLE III: Evaluation results

All results on the evaluation set outperformed those observed during cross-validation in the analysis phase.

Compared to other results on the leaderboard, the result achieved with Linear Regression and RFECV selection exceeds the average of the published results and outperforms approximately 60% of them.

IV. Discussion

An interesting observation is that the results obtained from Linear and Ridge Regression models on the hold-out split are very similar to those achieved on the evaluation set. This similarity could indicate some degree of overfitting, especially when considering the differences observed in cross-validation results.

Although the RFR showed the poorest results, its application was crucial in reducing the dimensionality of the dataset. To improve the performance of RFR, it would be beneficial to experiment with different feature sets and hyperparameter tuning.

For further improvement of the Linear Regression model, additional and alternative approaches could be implemented. These include:

- **Residual analysis**: analyzing the residuals in more detail can help identify any patterns or biases in the model's predictions, guiding improvements for better generalization
- **Applying PCA on different subsets of features**: instead of applying PCA to all acoustic features, experimenting with different subsets of features could reveal more meaningful components, thus capture better patterns in the data.

REFERENCES

[1] R. Winkler, "Influences of pitch and speech rate on the perception of age from voice," *Institute of Language and Communication, Technical University Berlin*, 2007.

[2] E. Deruty, "Intuitive understanding of mel-frequency cepstral coefficients," 2022. Accessed: January 23, 2025.

[3] M. P. Shameem, M. Faheem, M. V. V. Sahad, N. Ashraf, and S. Shaharyar, "Age and gender prediction from human voice for customized ads in e-commerce," *Computer Science and Engineering, MEA Engineering College, Perinthalmanna, India*, 2023.

[4] S. Ravishankar, S. Tiwari, P. M. K. Kumar, V. V. Patage, and S. Goyal, "Prediction of age from speech features using a multi-layer perceptron model," *Voice Intelligence Group, Samsung R&D Institute*, 2023.