Audio Age Estimation: A Regression Approach

Enrico Chen
Politecnico di Torino
Student id: s337750
s337750@studenti.polito.it

Luca De Paola
Politecnico di Torino
Student id: s337556
s337556@studenti.polito.it

***Abstract*—This report introduces a possible approach to the *Audio Age Estimation* regression problem. More specifically, the proposed approach involves exploiting various spectral, rhythm, and acoustic features that can be extracted from speech signals and linguistic metadata. Various summary statistics are also used to process the features. The proposed approach outperforms a naive baseline defined for the problem and demonstrates promising performance.**

I. Problem Overview

The task at hand is a long-established task within the field of speech processing: its objective is to estimate the age of a person by analyzing the acoustic and linguistic properties extracted from a speech signal.

The given dataset is divided into two parts:

- a *development* set, containing $2,933$ recordings for which the target variable *age* is given
- an *evaluation* set, containing 691 recordings for which the label is not given.

The goal is to build a regression pipeline capable of predicting the speaker's age based on these recordings.

The following features are provided in the datasets:

- *Sampling rate*: The sampling rate of the audio signal.
- *Gender*
- *Ethnicity*
- *Pitch (mean, min, max)*
- *Jitter*: A measure of pitch variations, indicating voice stability.
- *Shimmer*: A measure of amplitude variations in the speech signal.
- *Energy*: The overall energy of the speech signal.
- *Zero-crossing rate (ZCR) (mean)*: The number of times the signal changes sign.
- *Spectral centroid (mean)*: The "center of mass" of the frequency spectrum.
- *Tempo*: The estimated speaking rate.
- *Harmonic-to-noise ratio (HNR)*: It indicates the ratio of harmonics to noise.
- *Number of words, characters, pauses*: The total number of words, characters and pauses in the speech signal.
- *Silence duration*: The duration of silence periods within the speech.

An analysis of the development set reveals that the problem is imbalanced, with most of the speech signals coming from individuals in the 20–30 age range, as shown in Figure 1.

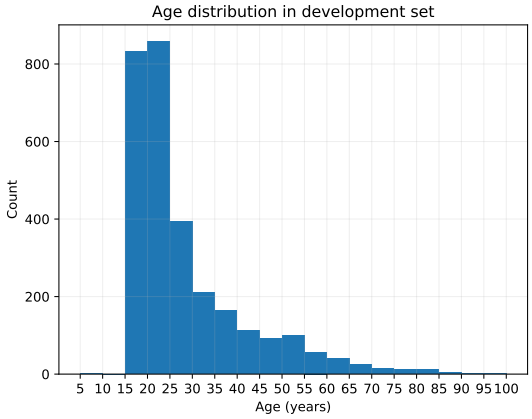

Fig. 1. Histogram of the target variable (*age*) distribution. The distribution is notably imbalanced, with certain age ranges being disproportionately represented, indicating potential challenges in ensuring uniform model performance across all age groups.

We observe that speech signals vary in sampling rate, as shown in Figure 2. During feature extraction, we use the native sampling rate for each recording instead of applying a default rate, which differs from the approach used for the provided features. Additionally, the duration of the samples also varies.

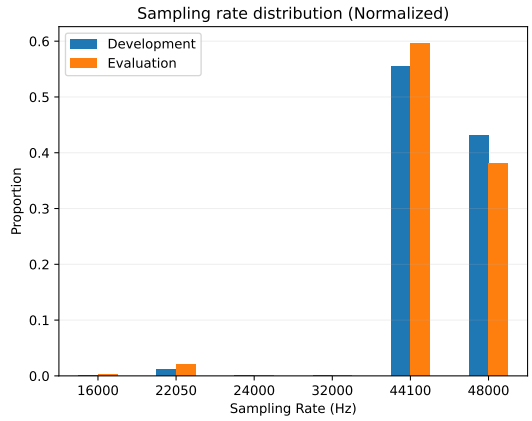

Fig. 2. Distribution of the sampling rate across both datasets. The figure highlights variations in the sampling rate, which are preserved during feature extraction by using the native rate for each recording.

The ethnicity feature shows significant variation between

the two datasets; only 17 ethnicities are shared between them, as summarized in Table I. Figure 3 illustrates the distribution of the top 10 ethnicities in the development and evaluation sets. Nearly 20 distinct ethnicities are identified, with most being unique to one dataset. The Igbo ethnicity is the only one represented in a significant proportion in both datasets, highlighting the substantial differences in ethnic composition.

TABLE I
NUMBER OF DISTINCT ETHNICITIES IN THE DEVELOPMENT AND EVALUATION SETS. ONLY 17 ETHNICITIES ARE COMMON TO BOTH DATASETS, HIGHLIGHTING SIGNIFICANT DIFFERENCES IN THEIR COMPOSITION.

	Number of distinct ethnicities
Development set	165
Evaluation set	73
Common to both sets	17

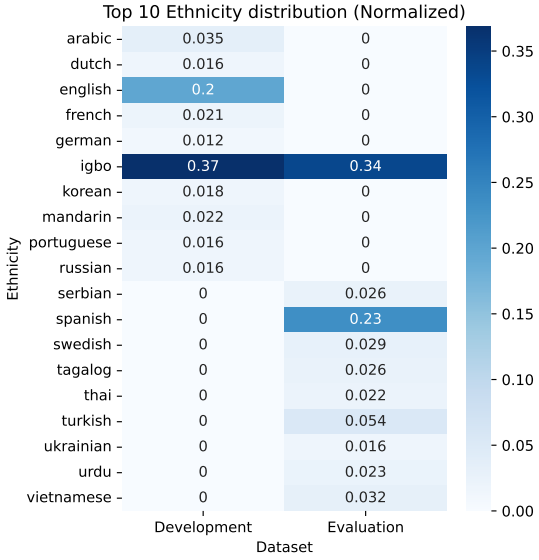

Fig. 3. Distribution of the top 10 ethnicities in the development and evaluation sets. The figure highlights that, except for the Igbo ethnicity, most prominent ethnicities are unique to either dataset.

The *tempo* feature is stored as an array, but it always contains a single scalar value. Additionally, in the evaluation set, the *gender* feature has three possible values: *male*, *female*, and *famale*. The distribution of *gender* is shown in Figure 4.

The distributions of the number of words and characters are similar in both datasets. A large portion of the speakers (approximately 60%) produce the same long phrase, while the remaining speakers utter very short phrases or minimal vocalizations. These patterns are depicted in Figure 5.

II. PROPOSED APPROACH

A. Preprocessing

Based on the *Data Exploration* step, we begin by dropping the sampling rate feature, as it is constant across all audio files and therefore unnecessary. Since the tempo is stored as a list, we transform it into a scalar. Due to inconsistencies between

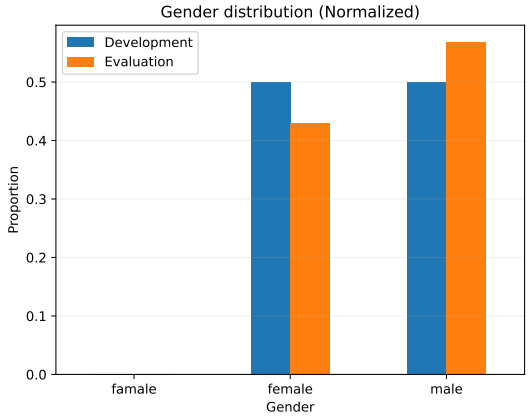

Fig. 4. Normalized distribution of the gender in both sets, we can observe that the gender in the development set is well balanced, while in the evaluation set slightly less. We can observe also the presence of the outlier *famale*.

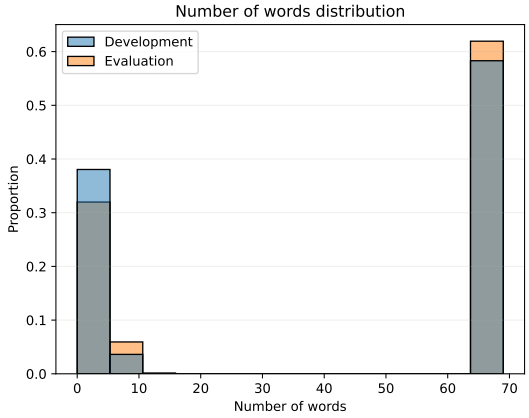

Fig. 5. Distribution of the number of words in the datasets. The figure shows a bimodal distribution, with most speakers producing either a long phrase ($\tilde{6}0\%$) or very short phrases/minimal vocalizations

the two datasets, we decide to drop the ethnicity feature. We also encode the categorical attribute gender using *dummy encoding*, with the outlier value *famale* manually mapped to *female*.

We extract the following additional features from the audio signals:

- *Mel-frequency Cepstral Coefficients (MFCC)*: These coefficients represent the power spectrum of the audio (Figure 6). Based on the literature [1], we select the first 13 MFCC features, which are widely recognized as effective for this task. To ensure robustness to outliers, we compute the 5th, 50th, and 95th percentiles for each coefficient, representing the minimum, median, and maximum values. The utility of MFCC features for age estimation is supported by previous studies [2], [3].
- *Delta features*: These features estimate the first derivative (Δ) of the MFCCs (Figure 6). Similarly to MFCCs, we calculate the 5th, 50th, and 95th percentiles of the MFCC-

Δ.

- *Spectral bandwidth*: This feature represents the range of frequencies where the magnitude of the spectral components is significant. For this feature, we compute the mean value over time to provide a concise summary.
- *Spectral rolloff*: This feature represents the frequency below which a specified percentage of the spectrum's energy is contained. Similarly to spectral bandwidth, we calculate the mean value over time as a representative measure.

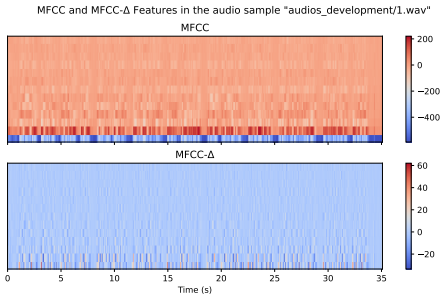

Fig. 6. *MFCC* and *MFCC-*Δ features extracted from the audio sample `audios_development/1.wav`.

We also introduce some additional features by exploiting the existing ones, such as:

- *mean_silence*: the average duration of silent pauses
- *silence_ratio*: the ratio of the silence duration over the entire audio
- *wps*: the rate of words per second

This allows us to drop *duration, num_characters, num_words, num_pauses* as, otherwise, they would provide redundant information. Moreover, these features are dependent on the duration of the audio clip, which is not relevant for age estimation.

To prevent features with large values from dominating the behavior of our models, we apply z-score normalization to all attributes. This step also enables meaningful comparison of the coefficients of linear models, which would otherwise be difficult to interpret.

B. Model selection

We evaluate several models using their default hyperparameters, with performance assessed through 5-fold *cross-validation* based on the *root mean squared error (RMSE)*. The models considered are the following: *Linear regression*, *Lasso regression*, *Ridge regression*, *Random forest regressor*, *Decision tree regressor*, *Support vector regression (SVR)*, *k-Nearest Neighbors (KNN) regression*, *Multi-layer perceptron regression*.

The RMSE scores of these models are compared against two baseline approaches: one that predicts the mean age and another that predicts the median age. Table II summarizes the RMSE scores obtained from 5-fold cross-validation alongside those from the public leaderboard, which considers performance on the evaluation dataset.

TABLE II
THE TABLE PRESENTS THE *RMSE* SCORES OBTAINED WITH 5-FOLD *cross-validation*, COMPARING THE PERFORMANCE OF OUR MODELS WITH TWO CONSTANT MODELS THAT ALWAYS PREDICT THE MEAN AND MEDIAN AGE. WE OBSERVE THAT THE WORST-PERFORMING MODEL IS THE *Decision Tree Regressor*, WHILE THE TWO CONSTANT MODELS (MEAN AND MEDIAN PREDICTORS) OUTPERFORM IT. AS NOTED, ALL MODELS ARE TRAINED USING DEFAULT HYPERPARAMETERS

Regressor	cross-validation RMSE	leaderboard RMSE
LinearRegression	10.481	9.728
Lasso	10.832	10.053
Ridge	10.480	9.880
RandomForestRegressor	10.425	9.820
DecisionTreeRegressor	14.854	14.647
SVR	11.165	11.931
KNeighborsRegressor	11.179	12.125
MLPRegressor	10.330	10.819
Mean	13.091	11.960
Median	13.978	12.866

During the hyperparameter tuning phase, we narrow down our selection to four models: *Random forest regressor*, *SVR*, *Ridge regression*, and *Linear regression*.

- *SVR* is chosen based on its demonstrated effectiveness in previous studies for age estimation tasks utilizing MFCC features [2], [3].
- *Linear regression* is included as it is the best performer on the public leaderboard and serves as a strong baseline model.
- *Ridge regression* is included as a regularized variant of *linear regression*, expected to deliver comparable or improved performance compared to *linear Regression*.
- *Random forest regressor* is selected due to its support in the literature for modeling age estimation tasks [3].

For *Ridge* and *Linear regression*, we apply two techniques to improve model performance: *Recursive Feature Elimination (RFE)* and *Polynomial Features Extraction*.

- *RFE* is used to systematically remove irrelevant or redundant features, ensuring that only the most influential variables are used in the model. It is a technique that recursively removes the least important feature (or features) according to the given model. This process continues until only the specified number of most relevant features remain.
- *Polynomial Features Extraction* allows the models to capture higher-order relationships between features. This is particularly useful when data involve complex interactions that are not purely linear.

C. Hyperparameters tuning

To compute the optimal configuration of the hyperparameters, a 5-fold *grid search* is performed. The hyperparameter sets considered for each model are summarized in Table III. Z-score normalization is applied to all models. Furthermore, for the *Ridge regression* and *Linear regression* models, the hyperparameter sets for the preprocessing steps (i.e., *RFE* and *Polynomial Features Extraction*) are also included in the table.

TABLE III
PARAMETER GRID USED FOR HYPERPARAMETER TUNING OF THE CHOSEN MODELS AND THE EVENTUAL CORRESPONDING PREPROCESSING STEPS

Model/Pipeline	Parameter	Values
SVR	*epsilon*	$\{0.1, 0.5, 1, 2, 5, 10\}$
	C	$\{5, 10, 20, 50, 100, 200, 1000\}$
	gamma	{scale, auto}
RFE	*n_features_to_select*	$\{0.1, 0.28, 0.46, 0.64, 0.82, 1.0\}$
Polynomial Features E.	*degree*	$\{2, 3\}$
Ridge regression	*alpha*	$\{1, 10, 100, 1000\}$
RFE	*n_features_to_select*	$\{0.100, 0.182, 0.264, 0.345, 0.427, 0.509, 0.591, 0.673, 0.755, 0.836, 0.918, 1.000\}$
Polynomial Features E.	*degree*	$\{2, 3\}$
Linear regression		
Random forest r.	*max_depth*	{15, 30, 50, None}
	min_samples_split	$\{2, 5\}$
	min_samples_leaf	$\{1, 4, 16\}$
	max_features	{sqrt, log2}

III. RESULTS

The best configurations found for each model are shown in Table IV alongside the scores obtained at the end of the grid search and on the public leaderboard. In general, the models outperform the baseline models shown in Table II both in the local score and on the public leaderboard.

TABLE IV
BEST HYPERPARAMETER CONFIGURATIONS AND RELATIVE RMSE SCORES

Model/Pipeline	Parameter	Value	Local[1] RMSE	Test[2] RMSE
SVR	*epsilon*	5	10.103	9.311
	C	20		
	gamma	scale		
RFE	*n_features_to_select*	0.28		
Polynomial F. E.	*degree*	2		
Ridge r.	*alpha*	1000	10.261	9.578
RFE	*n_features_to_select*	0.1		
Polynomial F. E.	*degree*	2		
Linear r.			10.386	9.689
Random forest r.	*max_depth*	50	10.339	9.647
	min_samples_split	2		
	min_samples_leaf	4		
	max_features	sqrt		

Among the models tested, the two that are ultimately submitted for performance evaluation are *Ridge regression* and *SVR*. The configuration and performance of these models are summarized below:

- *SVR*: The optimal configuration is {`C: 20, epsilon: 5, gamma: scale`}. The model achieves a cross-validation RMSE of 10.103 and a leaderboard score of 9.311.
- *Ridge regression*: For the Ridge regression pipeline, the best parameters are a polynomial degree of 2, a regularization parameter $\alpha = 1000$, and 28% of features selected through recursive feature elimination. The model obtains a cross-validation RMSE of 10.261 and a leaderboard score of 9.578.

The SVR model outperforms the Ridge regression model in both the grid-search cross-validation RMSE on the development set and the leaderboard score. It is reasonable to expect that the models' performance on the private leaderboard will be similar to that on the public leaderboard.

At the time of writing, the SVR model places in the 22nd position out of 162 participants. The top-performing model achieves a score of 9.056, while the lowest score is 14.769. Additionally, the organizers provide a baseline model with a score of 11.179. These results demonstrate that our model not only surpasses the baseline, but is also competitive, performing close to the best solution and maintaining a significant margin above the least effective models.

IV. DISCUSSION

The proposed solution to the audio estimation problem demonstrates a competitive performance, as shown by the score and placement on the leaderboard.

During the evaluation phase on the test dataset, we observe that the RMSE scores on the public leaderboard are consistently lower than those obtained during cross-validation on the development set. This discrepancy, combined with the differing ethnic representation across datasets, suggests that the label distribution in the test dataset may differ significantly. It is plausible that the test dataset contains an even greater prevalence of the most represented age ranges, where both models demonstrate superior performances during local evaluation.

Future improvements could explore deep learning models such as *Multi-Layer Perceptrons (MLPs)* and *Convolutional Neural Networks (CNNs)* [3], which excel in capturing non-linear patterns and processing audio signals. Additionally, *i-vectors* [4] could be utilized to represent audio features effectively, and recurrent networks like LSTMs or GRUs [5] could model temporal dependencies. These techniques could enhance generalization and handle more diverse datasets.

[1] RMSE score obtained during cross-validation grid search

[2] RMSE score obtained on the public leaderboard, which is considered to be the test set of the regression task

REFERENCES

[1] S. Gupta, J. Jaafar, W. W. Ahmad, and A. Bansal, "Feature extraction using mfcc," *Signal & Image Processing: An International Journal*, vol. 4, no. 4, pp. 101–108, 2013.

[2] A. Sadeghi Naini and M. M. Homayounpour, "Speaker age interval and sex identification based on jitters, shimmers and mean mfcc using supervised and unsupervised discriminative classification methods," in *2006 8th international Conference on Signal Processing*, vol. 1, 2006.

[3] R. Djeradi and A. Djeradi, "Gender and age extraction from audio signal using convolutional neural network, mfcc," in *Proceedings of the Third International Conference on Innovations in Computing Research (ICR'24)*, vol. 1058, p. 163, Springer Nature, 2024.

[4] J. Grzybowska and S. Kacprzak, "Speaker age classification and regression using i-vectors.," in *Interspeech*, pp. 1402–1406, 2016.

[5] A. Derdour, F. Henni, and L. Boubchir, "On the use of recurrent neural network based on lstm model for voice-based age estimation," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 6378–6383, IEEE, 2024.