Speech Regression for Age Estimation

Mattia Molinari
Politecnico di Torino
s337194
mattia.molinari@studenti.polito.it

Giacomo Maino
Politecnico di Torino
s338682
s338682@studenti.polito.it

***Abstract*—In this report we introduce a regression pipeline for age estimation of speakers based on audio signals. The proposed approach extracts acoustic and linguistic features from a dataset of controlled speech samples, focusing particularly on the Mel-Frequency Cepstral Coefficients, and uses different models to perform predictions. The final algorithm outperforms the prefixed baseline and achieves overall satisfactory results despite the skewness of the dataset.**

I. Problem overview

The automatic extraction of useful information from speech signals has a wide range of commercial and forensic applications, ranging from human-machine interaction to security. In this context, the identification of the speaker's age is a key task, opening the door to a variety of opportunities and challenges. This project develops a regression model to estimate the age of speakers starting from controlled speech samples, where speakers of different age, gender and ethnicity are asked to read a sentence. The dataset presents 3624 samples, divided into:

- *development* set, including 2933 samples with the corresponding ages;
- *evaluation* set, including 691 samples.

In addition to the gender, ethnicity and raw audio, both partition includes a variety of precomputed features, both acoustic and linguistic, such as pitch values, jitter, shimmer, number of words and characters, etc. We will need to use the development set to train and validate our model, and then evaluate its performances on the evaluation set.

We can make some considerations based on the development set. First, all the audios are sampled at 22.05 kHz, which is more than enough to satisfy the Nyquist-Shannon sampling theorem. Second, the age distribution is not uniform and it is skewed, with a higher concentration of samples in the interval $[18, 30]$ years, as can be notices in Figure 1: this could lead to a bias in the model, since it could be more accurate in predicting ages in this range and less effective in the others. On the other hand, the gender distribution is well balanced. Finally, audio samples differ in duration, depending on the speaker's pace and the length of the spoken sentence: this variability needs to be taken into account when extracting features from the audio signals, since most regression algorithms require a fixed number of dimensions. Moreover, we notice that speakers read from a small set of distinct sentences, differencing principally in length: this means that the number of words and characters is a good proxy for the specific sentence.

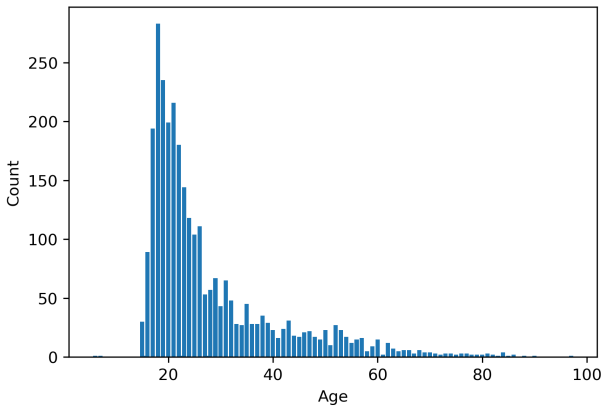

Fig. 1. Age distribution in the development set.

II. Proposed approach

A. Preprocessing

The dataset includes two categorical features that need to be addresses: *gender* and *ethnicity*.

The gender is expected to be a fundamental piece of information, since it influences voice's acoustic features [1]: for this reason, it is encoded using One-Hot Encoding.

The ethnicity is more problematic, since it has a large and very asymmetric distribution: it is possible to attenuate this problem by creating a default category (i.e. "other") including rare ethnicity, each representing only a small portion of speakers; this would lower the number of added dimensions by the encoding technique. However, we observe that ethnicity values coverage vastly differ between development and evaluation dataset: this implies that, during evaluation, all the ethnicity absent in the development phase would be assigned to the default category, determining its prevail over the others and a loss of information. This aspect is shown in Figure 2. In addition, even though sentences are spoken in a specific language, after a manual inspection we observed that the ethnicity of the speaker is not necessarily correlated with its proficiency and, consequentially, its pace: as a matter of fact, these last two characteristics are more likely to be represented by the overall duration, silence duration and number of pauses. For this reasons, we chose to remove the *ethnicity* feature from the dataset.

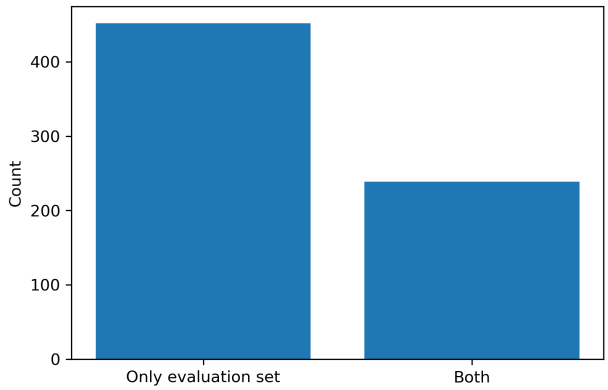

Fig. 2. Number of samples with ethnicity present only in the evaluation set (left) and in both evaluation and development sets (right).

We observed that the numbers of words and characters are highly correlated, presenting a Pearson coefficient of almost 1, so we decided to keep only first.

In addition to the precomputed features, we extended the dataset by extracting additional information directly from the audio signals, more precisely the sample duration and the Mel-Frequency Cepstral Coefficients (MFCCs). MFCCs are one of the most common features used in speech processing: they are derived in the frequency domain and capture the spectral characteristics of an audio signal, specifically the short-term power spectrum of a sound, aligning with human auditory perception. [2] MFCCs are computed dividing the audio signal into overlapping frames of fixed length, meaning that the number of values per coefficient is not fixed and depends on the sample duration: to handle this and maintain consistency, we calculated the mean, standard deviation, maximum and minimum of each coefficient across all frames. The total number of MFCC features extracted, denoted as n_{mfcc}, is a hyperparameter to tune.

As last step, numerical features are standardized using the Z-score normalization.

B. Model selection

The following algorithms were considered for the regression problem:

- *Random Forest Regressor*: it fits a number of decision tree regressors on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
- *Support Vector Regressor*: it is type of SVM, which is a method applying a transformation to the data points and identifying the maximum-margin hyperplane that separates the classes.
- *Voting Regressor*: this algorithm fits multiple models on the same dataset and averages the predictions. It can be used to combine conceptually different regressors and can be a good choice when the individual models perform well on different parts of the dataset. In our case, we will use a combination of the two previous approaches.

Model	Parameter	Values
Random Forest	`n_estimators`	{50, 100, 200, 300}
	`max_features`	{`sqrt`, `log2`, None}
	`max_depth`	{3, 6, 9, None}
	`max_leaf_nodes`	{3, 6, 9, None}
SVR	`C`	{1, 5, 10, 50, 100, 500}
	`kernel`	{`rbf`, `linear`, `sigmoid`}

TABLE I
HYPERPARAMETERS CONSIDERED

This models have been chosen because they are well-known and widely used in regression tasks, in particular they have been show to perform well in the context of speech processing. [3]

C. Hyperparameters tuning

There are two main sets of hyperparameters to tune:

- number of MFCC n_{mfcc};
- parameters of the Random Forest Regressor and Support Vector Regressor;
- weights of the Voting Regressor.

In order to determine the optimal value of n_{mfcc}, it is necessary to evaluate the impact of the quantity of MFCCs on the performance of the models: in consequence, the approach consists in extending the development set with an increasing number of MFCC, then split it into a training and validation set with proportion 80% and 20% respectively; this partitions are used to train the Random Forest Regressor and SVR with default hyperparameters, and assess performances through $RMSE$ score. Then, we can tune the two regressors using a grid search with 3-fold cross-validation: the tested hyperparameters are shown in Table I. Finally, the Voting Regressor tuning is performed using a grid search with 5-fold cross-validation and weights in the range $[0, 4]$.

III. RESULTS

The tuning of the number of MFCCs is summarized in Figure 3: we observe that the $RMSE$ score decreases as the number of MFCCs increases until a certain threshold, after which it stabilize or even worsen. Considering also that each MFCC adds four features to the dataset and increase computational time, an adeguate value for both models is $n_{\text{mfcc}} = 50$.

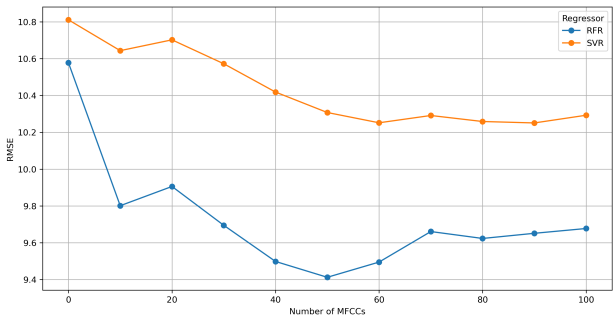

Fig. 3. Impact of the number of MFCCs on the $RMSE$ score.

Model	Parameter	Optimal Values
Random Forest	`n_estimators`	300
	`max_features`	`sqrt`
	`max_depth`	9
	`max_leaf_nodes`	9
SVR	`C`	50
	`kernel`	`rbf`

TABLE II
OPTIMAL HYPERPARAMETERS

The best configurations for the Random Forest Regressor and the SVR can be observed in Table II: more precisely, the two algorithms achieve respectively a $RMSE$ score of 10.718 and 9.835 on the validation set. Even though the SVR vastly outperforms the Random Forest Regressor, the Voting Regressor combining the two achieves a lower result of 9.616 with optimal weights 1 for the Random Forest Regressor and 4 for the SVR.

With all the hyperparameters defined, the three models are trained on the entire development set and evaluated: the results for the public score are shown in Table III. Since public and private partitions have the same statistical distribution, we can assume that the performances are analogous and the models do not overfit. For comparison purposes, a naive solution that predicts the age applying only a Random Forest Regressor on the initial features, without standardization, extraction of MFCCs and tuning of hyperparameters, did not achieve a public score lower than 11.187.

Model	Score
Random Forest Regressor	10.274
SVR	9.516
Voting Regressor	9.395

TABLE III
PUBLIC $RMSE$ SCORES

IV. DISCUSSION

The Voting Regressor clearly outperforms the naive model and achieves a far greater score than the prefixed baseline of 11.179. Nonetheless, some considerations for improvement can be made:

- Balance the dataset skewness towards younger speakers in order to reduce the bias of the model and improve the prediction of older speakers. A possible solution could be to augment the dataset with synthetic samples of older speakers, however this approach needs to be carefully evaluated to avoid overfitting or adding a new incorrect bias. [4]
- Additional features extractions. Extending the dataset with other information, such as delta-MFCC or spectral bandwidth and flatness, could improve the representation of the audio signals and the performance of the models. [5]
- MFCC extraction optimization. MFCCs clearly provide a useful tool to represent the audio signals, but they also introduce a large number of dimensions. This aspect could be attenuated by considering the correlation between the coefficients and reducing the number of features, for example by applying Principal Component Analysis.
- New model selection. The Voting Regressor combines two models that are already well-known and widely used in regression tasks, but it could be interesting to explore other algorithms, such as a Neural Network or a Gradient Boosting Regressor.

These promising opportunities open the door to a variety of possible improvements, however the proposed solution result is already satisfying, since it combines the strengths of the two models and achieves a better performance than the individual ones.

REFERENCES

[1] Y. Samuelsson, "Gender effects on phonetic variation and speaking styles," 2006.

[2] 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.

[3] H. Phan, M. Maaß, R. Mazur, and A. Mertins, "Random regression forests for acoustic event detection and classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 20–31, 2015.

[4] J. Sivakumar, K. Ramamurthy, M. Radhakrishnan, and D. Won, "Synthetic sampling from small datasets: A modified mega-trend diffusion approach using k-nearest neighbors," *Knowledge-Based Systems*, vol. 236, p. 107687, 2022.

[5] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," 2004.