Age Classification from Acoustic Features

Rosario Interlandi
Politecnico di Torino
s334202
s334202studenti.polito.it

Ramadan Mehmetaj
Politecnico di Torino
s346213
s346213@studenti.polito.it

***Abstract*—This work explores the task of age regression from speech signals by extracting and analyzing relevant acoustic features. Specifically, Mel spectrograms and MFCCs were computed, along with additional features derived from these representations. Using this feature set, two regression models were evaluated: Support Vector Regressor (SVR) and Random Forest Regressor (RFR). The models were trained to predict the speaker's age based on the extracted acoustic features. Experimental results highlight the effectiveness of these approaches, providing insights into the relationship between speech characteristics and age estimation.**

I. Problem Overview

A. Dataset Description

The objective of the project is to develop a model that is capable of accurately classifying the age of each speaker. The audio dataset for development and evaluation consists of a collection of numerous audio records.

Each audio file is stored in WAV format and is associated with a specific path, which is recorded in the *development.csv* and *evaluation.csv* files.

The dataset consists of 3,624 samples: 2,933 samples for the development set and 691 samples for the evaluation set. The development set contains several attributes used to train the model. Each record in the data set contains the following attributes:

- *Id*: Identifier of the audio;
- *Sampling rate*: The sampling rate of the audio signal, in Hz;
- *Age*: The chronological age of the speaker (target label);
- *Gender*: The gender of the speaker;
- *Ethnicity*: The ethnicity of the speaker;
- *Mean pitch, Max pitch, Min pitch*: Mean, maximum, and minimum pitch of the speech signal, in Hz;
- *Jitter*: A measure of the variations in pitch, representing voice stability;
- *Shimmer*: A measure of amplitude variations in the speech signal;
- *Energy*: The overall energy of the speech signal;
- *ZCR mean*: The mean zero-crossing rate, indicating the number of times the signal changes sign;
- *Spectral centroid mean*: The mean spectral centroid, representing the "center of mass" of the frequency spectrum;
- *Tempo*: The estimated speaking rate, in beats per minute (BPM);
- *HNR*: The harmonic-to-noise ratio, indicating voice quality;
- *Num words, Num characters*: The number of words and characters in the spoken sentence;
- *Num pauses*: The number of pauses detected in the speech;
- *Silence duration*: The total duration of silence within the speech signal, in seconds;
- *Path*: The path of the audio recording file. The audios are stored in the .wav extension in the folders audios_development and audios_evaluation;

It is used the *development.csv* dataset to train the model, dividing it into a training set and a validation set, using 79% of the samples in the former and the remaining 21% in the latter.

The evaluation set contains the samples that must be classified for the test. Obviously, for these audios, the age label is not provided.

Fig. 1 illustrates the label distribution of the development.csv dataset. Younger individuals are clearly more prevalent, with a significant concentration in the 20–25 age range. As a result, one can expect the predictions to be more accurate for subjects within this age group.

Fig. 2 illustrates the distribution of audio durations across the dataset. Some records have longer durations than average, but were kept as they may contain valuable age-related information. However, a closer analysis of the audio files reveals notable variations in duration. These differences can be attributed not only to the speaker's natural speaking speed but also to the presence of silences at the beginning, middle, or end of certain recordings, which do not contribute meaningful information.

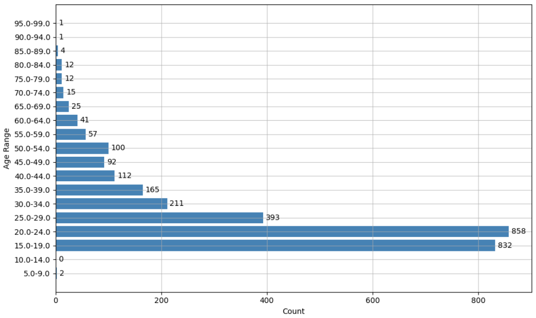

Fig. 1. Labels distribution

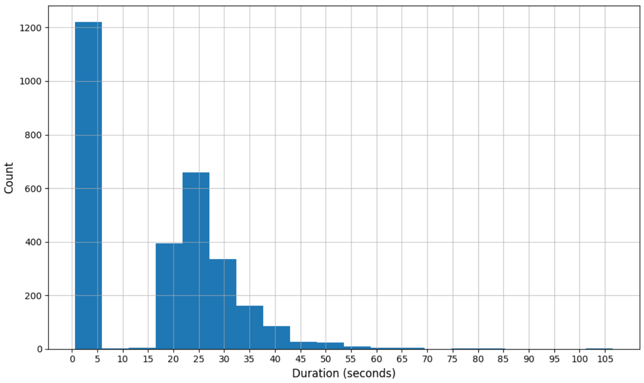

Fig. 2. Audio duration distribution

II. Proposed Approach

A. Data Preprocessing

The preprocessing of the audio recordings involved encoding the categorical *gender* label using a label encoder, mapping each gender (e.g., 'male', 'female') to an integer. The *tempo* feature is handled to ensure usability for the model and no missing values were found, so all samples are retained. Observing the dataframe values and listening to the dataset's audio, *ethnicity*, *sampling_rate*, *num_pauses*, *num_words* and *num_characters* were removed due to low correlation for the first two and inconsistencies with the audio for the latter.

Eliminating silences throughout the duration of the audio filtering frequencies with a difference of less than 25 dB relative to the maximum peak results in a compact audio representation with a meaningful mean duration. The recordings are loaded at a constant sample rate of 22.05 kHz and are modified to 44.10 kHz to improve audio quality to extract detailed audio features.

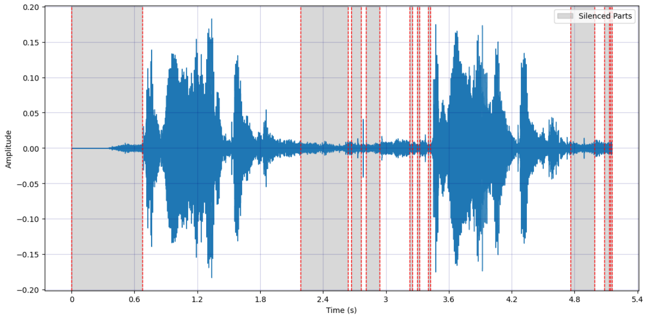

Fig. 3. Mel spectrogram from audio 807

Feature extraction is performed using Mel-Spectrogram and MFCC. The Mel-Spectrogram represents a signal's frequency content over time, using the Mel scale, which approximates the human auditory system better than traditional spectrograms [4] . The relationship between Mels (m) and Hertz (f) is given by the following formula

$$m = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right)$$

All the features already provided in the *development.csv* are scalar values. To maintain consistency with these features, we opted to use a Mel-Spectrogram(*Fig. 4*) with *mels = 40*, which divides the frequency spectrum into 40 equally spaced Mel bands. For each of these Mel bands, we computed the mean and standard deviation across time to obtain scalar values that summarize the spectral content of each band.

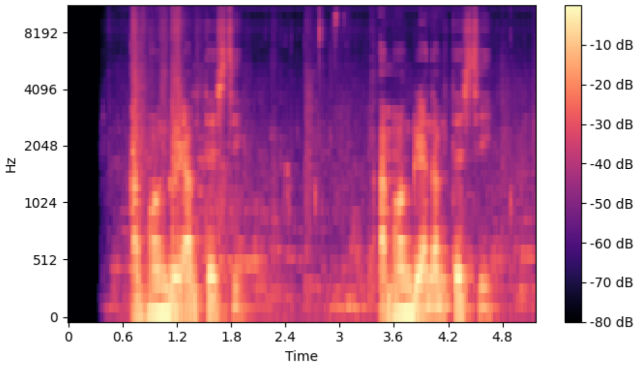

Fig. 4. Mel spectrogram from audio 807

In order to obtain a comprehensive view of the data and enhance the feature set, the second part of the process involves computing the MFCC features. MFCC is an audio feature extraction technique that extracts parameters from speech, similar to how humans perceive speech, while deemphasizing irrelevant information. The librosa API is utilized to extract Mel-frequency cepstral coefficients (MFCCs) from each audio file. These features are then processed by calculating the mean and variance of the entire MFCC matrix (*Fig. 5*) to summarize the data. A major limitation of MFCCs and Mel spectrograms is their static representation of the spectral content of each audio frame, ignoring the temporal relationships between frames. To overcome this limitation, we used deltas and delta-deltas for both MFCCs and Mel spectrograms: deltas measure the rate of change of coefficients between consecutive frames, while delta-deltas calculate their acceleration. These descriptors allow us to capture the temporal dynamics of the signal, making the representation richer and more capable of modeling complex patterns and variations over time.

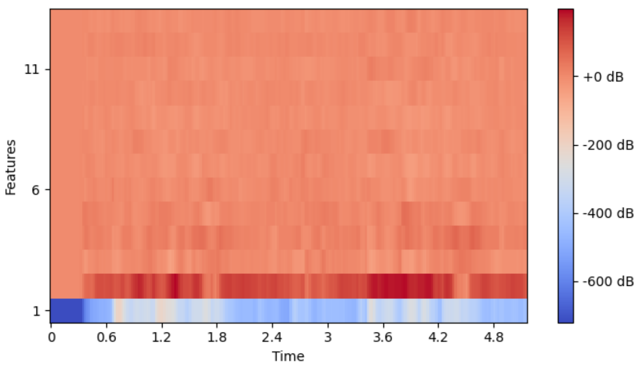

Fig. 5. MFCC spectrogram from audio 807

We also included features derived from the autocorrelation function (ACF), performing the mean and the standard deviation, which help capture repeating patterns or rhythmic structures within the audio.

After completing the feature extraction process, we implemented a strategy to prevent the model from overfitting on the most represented samples in the training dataset. Specifically, we grouped the samples into age intervals and established a maximum threshold for the number of samples allowed in each group. By capping the number of samples per age group, we ensured a more balanced distribution of the data across different intervals. This adjustment not only reduced the bias towards overrepresented groups but also contributed to improving the generalization ability of the model during training.

B. Model Selection

In this study, two popular machine learning algorithms, Support Vector Regressor (SVR) and Random Forest Regression (RF), were tested as the regression models:

- *Random Forest Regressor*: RF is one of the most well-known ensemble algorithms that use decision trees as a base classifier. It selects a random sample with replacement from the training set and trains the trees. Each tree is learned on a random set of features, typically formed by $\sqrt{p}$ features. Finally, the prediction is made by averaging the predictions from all the trees.

- *Support Vector Regressor*: SVR is a supervised learning algorithm commonly used for regression tasks. It extends the principles of Support Vector Machines to predict continuous values rather than discrete categories. Instead of finding a hyperplane that separates data points, SVR identifies a function that best approximates the relationship between the input features and the target variable. By maximizing the margin within a specified error threshold, the algorithm ensures robust predictions while minimizing the impact of outliers. During training, SVR learns the optimal mapping between audio features and the target values, enabling precise estimation of continuous outputs.

C. Hyperparameters tuning

The tuning process focuses on adjusting three sets of hyperparameters:

- Age group thereshold parameter: n is number of samples allowed in each age group, which is set to control the distribution of samples

- Classification Model parameters: Each of the two models examined in this study has its own distinct set of hyperparameters.

To identify the optimal parameters for our data, a grid search was conducted on the development set for the thereshold and Random Forest model whie to finetune SVR hyperparameters, the Optuna library was used. Unlike Grid Search, which evaluates a predefined set of values, Optuna efficiently explores the hyperparameter space using Bayesian optimization and Tree-structured Parzen Estimators (TPE). This approach is particularly useful for SVR, as its hyperparameters are continuous, making Grid Search less effective. By dynamically adjusting the search process, Optuna identifies optimal configurations with fewer evaluations, improving efficiency and performance.The hyperparameters and suggestion (for Optuna) used are outlined in Table 1.

Hyperparameters	**Values**
Thereshold max	n 140 $\rightarrow$ 180, step 10
RF	max estimators: [100, 300] max depth: [5, 10, None]
SVR	C: [10,11] ϵ: [0.01, 0.02]

TABLE I
HYPERPARAMETER VALUES

III. RESULTS

The RMSE of the Random Forest and Support Vector Regression models as n-sample changes is shown in *Fig. 6*. As seen from the graph, the Support Vector Regressor performs better than the Random Forest Regressor.

When looking at the RMSE of the model, an increase is observed. Using *grid search* and *optuna*, it was found that the best parameter configuration for the task is setting *n = 150* and using a Support Vector Regressor with *C = 10.7693* and *epsilon = 0.01931*. It was also noticed that combining MFCC features with those extracted from the spectrogram, instead of using MFCC features alone, slightly improves the overall accuracy, as shown in *Fig. 7*. Therefore, both methods can be considered.

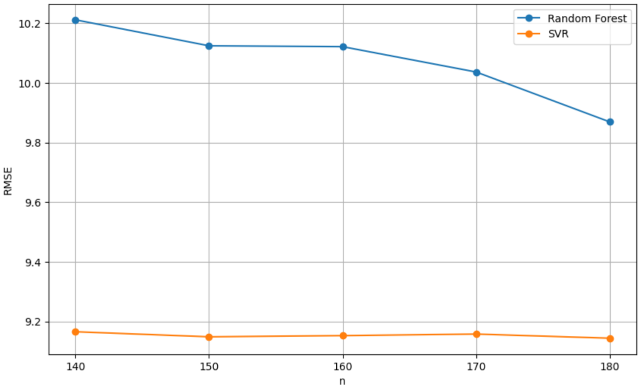

Fig. 6. Compared classification model RMSE

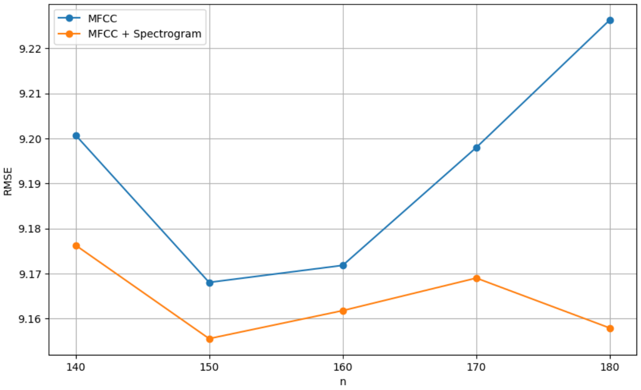

Fig. 7. Compared SVR and adding Spectrogram features

It was decided to choose the version incorporating both feature extraction methods, as the best performance for this model is achieved with a combination of MFCC and spectrogram features.

IV. Discussion

The models are trained with the optimal hyperparameters on the entire development set and evaluated using the evaluation dataset. The public score obtained with this approach is RMSE = 9.114, which, based on the leaderboard, can be considered a satisfactory result.

However, there are several opportunities for improvement to refine the approach and achieve better outcomes:

- *Use of Pre-trained Models*: Incorporating pre-trained models, such as Google Speech Recognition, for speech-to-text conversion can significantly improve the classification process for predicting age. These models are trained on extensive datasets and offer better accuracy in transcribing audio into text. This transcription makes it easier to correlate the audio features with the predicted age.
- *Convolutional Neural Networks (CNNs)*: CNNs can serve as a feature extraction model for audio signals, extracting relevant features that can then be used as input to other regressors, such as recurrent neural networks (RNNs) or support vector regressors (SVRs), for predicting age. Alternatively, CNNs can be trained as a standalone classifier to directly predict age. This approach has demonstrated good results for the task [8].

References

[1] V. Tiwari, ‘MFCC and its applications in speaker recognition’ International Journal on Emerging Technologies 1(1): 19-22(2010).

[2] MFCC features extraction [Online]. Available: https://librosa.org/doc/main/generated/librosa.feature.mfcc.html

[3] MFCC delta [Online]. https://librosa.org/doc/main/generated/librosa.feature.delta.html

[4] M. Xu, L. Duan, J. Cai, L. Chia, C. Xu, Q. Tian, ’HMM-Based Audio Keyword Generation’, Advances in Multimedia Information Processing – PCM 2004: 5th Pacific Rim Conference on Multimedia. Springer. ISBN 978-3-540-23985-7.

[5] B. Zhang, J. Leitner, S. Thornton, ’Audio Recognition using Mel Spectrograms and Convolution Neural Networks’, conference paper.

[6] L. Grama, C. Rusu, ’Audio Signal Classification Using Linear Predictive Coding and Random Forests’, Available:https://ieeexplore.ieee.org/abstract/document/7990431

[7] L. Grama, L. Tuns, C. Rusu, ’On the Optimization of SVM Kernel Parameters for Improving Audio Classification Accuracy’, 14th International Conference on Engineering of Modern Electric Systems (EMES) 2017.

[8] D. Palaz, M. Magimai, R. Collobert, ’Analysis of CNN-based Speech Recognition System using Raw Speech as Input’, Idiap-RR-23-2015, June 2015.