

Data Science Lab: Process and methods

Politecnico di Torino

Project Assignment

Winter Call, A.Y. 2024/2025

Last update: January 6, 2025

1 Project dates

Start date: January 6, 2025 at 23:59 (CET)
Due date: January 28, 2025 at 23:59 (CET)

Due date is a **strict deadline**.

2 Problem description

Within the field of speech processing, a long-standing task of interest is estimating the age of a speaker based on their vocal characteristics. Some systems are capable of analyzing spoken sentences to extract features that correlate with the speaker's age. One such approach involves leveraging acoustic and linguistic features derived from the speech signal.

When a speaker produces a spoken sentence, various feature extraction methods are applied to analyze the acoustic properties of the speech signal. In the specific system of interest, these features include, among others, pitch, formants, energy levels, and phonetic patterns. These extracted features serve as inputs to the problem, with their properties being correlated with the speaker's age.

The process of estimating a speaker's age from their speech is referred to as *age estimation*. The target (output) of the system is a single value representing the estimated age of the speaker.

The goal of this project is to build a data science pipeline that predicts, for each spoken sentence, the target age of the speaker, using the available metadata as inputs, along with any other feature that can be potentially extracted from the input speech.

2.1 Dataset

The dataset is comprised of 3,624 samples: 2,933 samples for the development set and 691 samples for the evaluation set. Each sample corresponds to a spoken sentence, and the associated speaker's age is provided as the target label. The speech samples have been collected in controlled conditions to ensure consistent feature extraction.

For each sample, a variety of acoustic and linguistic features have been extracted from the speech signal. These features form the dataset and include:

- **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 file path to the audio recording

These features comprehensively capture the acoustic properties of the speech signal along with some linguistic metadata, but you are allowed to extract additional features from the speech signal.

For each sample, the target variable is the speaker’s age, represented as a continuous numerical value in the `age` column of the dataset.



Warning: For this project, you are not allowed to use external datasets other than the one provided. Adoption of external resources, including pre-trained models, will result in failure of the exam.

The dataset is located at [this URL](#).

- **`development.csv`** (development set): a comma-separated values file containing the 2,933 samples for the development set. This portion includes the target age for each sample, which is used to train and validate your models.
- **`evaluation.csv`** (evaluation set): a comma-separated values file containing the 691 samples corresponding to the evaluation set. This portion does not contain the age target.
- **`sample_submission.csv`**: a sample submission file.
- **`audio_development/`**: a folder containing the audio files corresponding to the development set.
- **`audio_evaluation/`**: a folder containing the audio files corresponding to the evaluation set.

2.2 Task

You are required to build a regression pipeline to predict the age of the speaker.

2.3 Evaluation metric

Your submissions will be evaluated in terms of the Root Mean Square Error (RMSE) between your predictions and the target values. RMSE is a standard metric for regression tasks that measures the average magnitude of the error. For a single prediction with target y_1 and prediction \hat{y}_1 , the squared error is $(y_1 - \hat{y}_1)^2$.

For all n samples with targets y_1, y_2, \dots, y_n and predictions $\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$, the RMSE is computed as:

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

This metric provides an aggregate measure of prediction accuracy, penalizing larger errors more significantly than smaller ones.

3 Submit your result

Submission file To get your results evaluated, you have to upload a result file on our submission competition platform, Data Science Lab Environment (DSLE). The submission file must be a CSV file formatted as follows:

```
Id,Predicted
1,33.0
2,41.0
3,24.0
4,31.0
...
```

The submission file must contain a header line and a row for each record in the Evaluation collection. Each row must have two fields:

- the Id of the corresponding record in the Evaluation set. It corresponds to the column Id in the evaluation CSV file.
- the Predicted age (e.g. 21.0).

You can find a sample submission file in the project material (see 2.1).

Submission platform The submission platform is the same one you used during the course laboratories. Therefore, you have to use your personal key. If you do not have the key or have problems using it, please send an email to lorenzo.vaiani@polito.it. Please refer to [the guide](#) on the course website to go through the submission procedure.

You can find the DSLE platform at <http://trinidad.polito.it:8888>

4 Upload the report and the software

The report and the software have to be submitted by the due date reported in Section 1. This is a strict deadline.

Submission All the required files (i.e., for the report and the software) must be included in a **single ZIP archive**. The archive must be uploaded to the “[Portale della Didattica](#)”, under the *Homework* section. Please use as description: **report_exam_winter_2025**.



Info: A ZIP archive is a ZIP archive, not a RAR, a 7z, or a tarball archive, nor any of those renamed with a trailing .zip extension.

Formatting rules The formatting rules for both the report and the software are described in the document with exam rules. You can find it on the course website.

5 Fill in the LLM usage form

As discussed in the exam rules, adoption of Large Language Models (e.g. ChatGPT) is allowed for the production of the report (**not** for the implementation of the solution). Each team **must** provide information about whether they used, and to which extent they did, LLM-based tools.

To do so, please fill in [this form](#) by the due date of this project. Failure to do so will result in a void project.



Warning: Make sure you remember to fill in the form by the due date, or your project will not be considered valid!