Age Regression via Vocal Characteristics

Maddalena Ghiotti
Politecnico di Torino
Student id: s332834
s332834@studenti.polito.it

Nunzio Licalzi
Politecnico di Torino
Student id: s344860
s344860@studenti.polito.it

***Abstract*—We present a solution to the regression task of estimating a speaker's age based on vocal characteristics and demographic features. Our approach combines features extracted from both the time and frequency domains, alongside key demographic attributes. For modeling, we selected the Histogram Gradient Boosting model, which demonstrated superior performance compared to several defined thresholds, baselines and other models. The chosen methodology achieved satisfactory results, with room for further improvement.**

I. Problem overview

We were tasked with participating in a contest titled ”Estimating the Age of a Speaker Based on Their Vocal Characteristics”. The objective was to develop a regression model capable of processing audio files, effectively handling the variability in vocal features, to infer speaker's age.

The initial dataset provided included the following information:

- File paths to the audio recordings;
- Demographic details about the speakers, such as gender and ethnicity;
- Previously extracted features, including Harmonics-to-Noise Ratio (HNR), jitter, shimmer and other audio-related attributes.

We assumed that the provided audio recordings were partially sourced from the Speech Accent Archive[1]. All audio samples from this archive feature speakers reciting the same phrase in English (the exact phrase is available on the Speech Accent Archive). In contrast, recordings from other sources include a broader variety of content and languages.

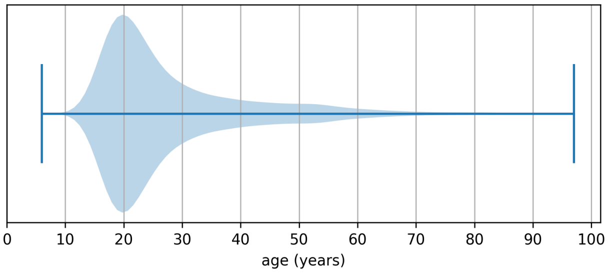
The given set of data was divided into:

- a **Development Set** containing around 2930 labeled samples.
- an **Evaluation Set** containing around 700 unlabeled samples, for which the model should correctly predict the speakers' ages.

All audio samples were recorded at a sampling rate of 22.05 kHz, which, according to the Nyquist-Shannon Theorem [1], implies that the range of captured frequencies is $[0, 11]$ kHz.

Moreover, the audios have different time lengths spanning from a few seconds to over a minute and a half, thus complicating the feature extraction process.

Fig. 1. Violin plot depicting the age distribution



The additional features are provided in two different *.csv files*, one for the Development set and the other for the Evaluation set. Using the Development set, we conducted an in-depth analysis to derive insights and observations.

First, we examined the distribution of the target variable. The age range spans from a minimum of 6 years to a maximum of 97 years with a mean of approximately 27.9 years. To better illustrate this distribution, we generated a violin plot, as shown in Figure 1.

As illustrated, the age distribution is right-skewed (the median is roughly 23 years), indicating that there are significantly more audio samples at the lower end of the age spectrum rather than the higher end. However, in spite of the minimum age being 6, only two speakers are actually younger than 15. This indicates that, overall, not all age groups are adequately represented and modeled with the available data.

Examining the provided *.csv files*, we observed that none of the given features contained missing values, ensuring a complete dataset to work with.

Moreover, we can see that the gender distribution is overall well-balanced in the Development set and less-balanced in the Evaluation set, having the distribution depicted in Table I. Gender is a particularly crucial feature, mainly because vocal characteristics evolve differently with age depending on the individual's sex [2]. In fact, if this factor is not adequately accounted for, it could lead to significant inaccuracies in age estimation.

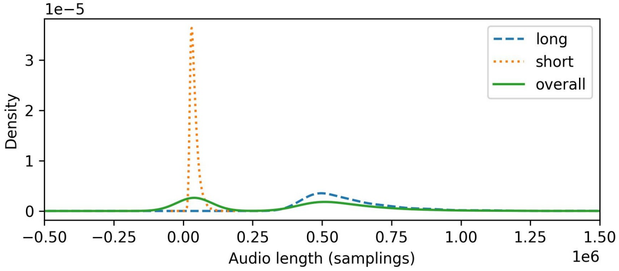
TABLE I
Gender distribution in the different data sets

	Gender		
Set	**Male**	**Female**	$\mathbf{N_M/N_F}$
Development	1465	1468	0.99796
Evaluation	393	298	1.31879

Looking at the distribution of audio lengths, shown in Figure 2

[1] Speech Accent Archive

Fig. 2. Distribution of audio length based on audio category



with a green continuous line, we can notice two distinct peaks. The right peak corresponds to recordings with exactly 69 words and 281 characters (blue dashed line), while all other shorter recordings contribute to the left peak (yellow dotted line). The first group, termed *long audios*, consists of 1,710 samples, which likely originate from the Speech Accent Archive. The second group is labeled *short audios* and it is of unknown origin.

Interestingly, the *long audios* group includes recordings from 150 different ethnicities, whereas the short audio group contains only 18 ethnicities. Only three ethnicities are present in both groups. This prompted further analysis of ethnicity distribution within the audio files to uncover additional insights. Since the text provided in the Speech Accent Archive is spoken in English, native English speakers may naturally exhibit a faster speech rate. This could potentially result in these speakers being perceived as younger than their actual age.

However, even though the hypothesis seemed reasonable, the results obtained were inferior to those achieved without taking the hypothesis into consideration when testing it on the public leaderboard, provided to us by the organizer of the competition to test part of the Evaluation set. Consequently, it was ignored.

From the analysis of ethnicity and gender distributions, we can conclude that the training and evaluation sets are likely sampled from two different distributions.

To conclude, we have to recall that when it comes to working with audio files we can leverage both the time domain and the frequency domain, using, for example, the fast Fourier transform (fft) [3] and the wavelet transform [4].

II. Proposed Approach

We decided to utilize some of the pre-extracted features provided in the *.csv files* while also extracting additional features directly from the audio files.

A. Preprocessing

As previously discussed, both the time domain and the frequency domain can be leveraged to extract additional useful features beyond those already provided.

1) Time Domain Feature Extraction: In the time domain, several additional features were extracted from each audio sample:

- **Frames**: The total number of frames and the number of voiced frames.
- **Quantiles**: The following quantiles were calculated: 10%, 16%, 50%, 84%, and 90%.
- **Medians and ranges**: The following medians were computed: 84-median, 16-median, 50-median, as well as the $90\% - 10\%$ range.
- **Signal characteristics**: Statistical measures such as mean, variance, skewness and kurtosis were computed.
- **Times**: Specific time markers were calculated at 5%, 25%, 50%, 75%, and 95% of the signal duration.
- **Mean absolute slope**: via the use of `praat` we extracted the mean absolute slope that is the average absolute slope across all turning points in a pitch contour [5].

2) Frequency Domain Feature Extraction: In the frequency domain, additional features were extracted from each audio sample to capture spectral characteristics:

- **Spectral Energy**: the spectral energy was computed for the following frequency intervals: $[250, 650]$ Hz and $[1, 8]$ kHz [6].
- **Peak Frequency**: the peak frequency and its corresponding amplitude were calculated.

3) Mixed Domain Feature Extraction: Finally, we extracted features that establish a relationship between the time and frequency domains:

- **Hilbert Mean**: The Hilbert transform was applied, and the mean of the resulting signal was computed [7].
- **Entropy**: Temporal entropy, derived using the Hilbert transform [7], was calculated. This feature was selected due to its demonstrated effectiveness in analyzing audio signals [8].

4) Mel-frequency cepstral coefficients: The final extracted features were the Mel-frequency cepstral coefficients (MFCCs), which have been proven to be key in audio processing across various contexts (for example, [9] and [10]). The calculation of MFCCs involves a series of consecutive transformations [11], [12]. First, the signal is divided into temporal frames using overlapping windows. For each resulting frame, the Fourier transform is applied to generate a power spectrogram. The spectrogram coefficients are then mapped to a Mel scale through a bank of overlapping triangular filters, aligning them more closely with human sound perception. Subsequently, a discrete cosine transform (DCT) is applied. In some cases, the coefficients are further normalized.

We chose to use MFCCs instead of simple spectrograms converted to the decibel scale. This decision was based on an analysis of the correlation between the target variable and the coefficients extracted from spectrograms, which revealed that coefficients corresponding to lower frequencies had a stronger correlation. MFCCs inherently emphasize low frequencies, making them a more suitable choice for our purposes.

To account for the varying lengths of the audio samples, the MFCCs were averaged along the temporal axis, resulting in a row vector of m coefficients. An exception is made for the random crop (see subsection II-B). This compression does not result in a significant loss of information, as the pronunciation of characters and words (distinguishable along the temporal axis) is not relevant for age prediction. The number of MFCCs to extract was treated as a hyperparameter and, as such, a few possible values will be discussed further on.

To conclude, we also transformed some of the features provided in the *.csv file*. The most notable modifications include the encoding of the `gender` feature, where males were encoded as 1 and females as 0, and the conversion of the `HNR` feature into decibels (dB). Finally, we opted to retain only the following features from the provided set: `gender`, `HNR`, `jitter`, `shimmer`, `max_pitch`, `min_pitch` and `mean_pitch`. Some of these features were selected based on their demonstrated importance in similar voice-related tasks [13].

In addition to feature extraction, an audio cleaning procedure was applied. Noise was reduced in the recordings using the `noisereduce` Python package, and the `librosa` package was used to trim leading and trailing silence. The audio data were retained and tested in both their original and cleaned forms.

B. Model selection

The following models were tested:

- **Random Forest**: This regressor is the most interpretable among the tested models. It combines multiple decision trees, each trained on different subsets of the data and features, to mitigate overfitting [14]. Since the Random Forest algorithm leverages decision trees it does not require feature scaling.
- **Support Vector Machines (SVM)**: SVM aim to find the maximum-margin hyperplane that best separates the target feature. Given that the input features have different orders of magnitude, applying standardization or normalization is necessary for optimal performance.
- **Multi-layer Perceptron (MLP)**: MLPs are feedforward neural networks composed of fully connected artificial neurons. They tend to perform better when trained on datasets with a large number of samples.
- **Histogram Gradient Boosting**: This model builds an ensemble of weak learners (Decision Trees in this case), optimizing them sequentially to minimize the loss function. It is highly efficient for large datasets and can handle both continuous and categorical features. It does not require feature scaling.
  Note: from now on we will mostly refer to this model as ”Hist” for brevity.

Both Random Forest, SVM and MLP models showed great potential, as they have previously achieved strong results in audio-related problems [15] [16]. However, the performance metrics, computed as the average root mean squared error (RMSE) over a 15-fold cross-validation on the standardized Development set, indicated that the Hist model outperformed the other approaches.

The results achieved by these models with their default hyperparameters used by the Python package `scikit-learn` ([17], [18], [19], [20]) are summarized in Table II.

TABLE II
MEAN RMSE ACHIEVED BY DIFFERENT MODELS USING DEFAULT HYPERPARAMETERS WITH A 15-FOLD CROSS-VALIDATION

Regressor	Mean RMSE
Random Forest	10.1270
SVM	10.9371
MLP	13.1516
Hist	9.7871

Based on these results, we decided to pursue two approaches. Since the **Histogram Gradient Boosting (Hist)** model achieved the highest score in our tests and delivered outstanding results in similar audio-related tasks [21], we focused on fine-tuning its hyperparameters. Meanwhile, given the widespread use and strong performance of MLPs in the field [16], we aimed to improve its results using a data augmentation technique.

For both models, we evaluated predictions in their raw form and after rounding to the nearest half-integer. This rounding procedure was done based upon an analysis of the development target variable, which only takes integer or half-integer values.

Additionally, we trained the Hist model separately by gender, by audio length (*long* and *short*), and by combining both splits (four distinct models). For the MLP, our experiments were limited to data exclusively from *long audios*.

MLP model construction

The poor performance of the MLP model is likely due to the small size of the Development set. To address this, a random cropping strategy was implemented. During training, the model selects a random time window of the MFCCs for each sample at every iteration [22], [23]. The number of frequency bins for MFCC computation was set to either 15 or 35, while the time bin width and overlap lengths were 66,150 and 22,050 time samples. The random crop width, treated as a hyperparameter, varied between 3, 5, and 10 samples. The resulting coefficients were either flattened or averaged along the time axis, and *.csv file* features were appended. Given the already high feature count, no additional metrics were added.

C. Hyperparameters Tuning

We are going to discuss hyperparameters tuning exclusively for the Hist model as it was the better performing one.

There were primarily two sets of hyperparameters to optimize:

1) The number of MFCCs m to extract.
2) The hyperparameters of the Hist model.

1) Fine-tuning the Number of MFCCs (m): To determine the optimal number of MFCCs to extract, we adopted a trial-and-error approach. The tested values ranged from 20 to 80, incremented by 5 whilst keeping the default hyperparameters for the model and assuming their orthogonality with respect to m.

The best performance was achieved with $m = 35$.

TABLE III
HYPERPARAMETERS CHOSEN FOR HIST

Parameter	Tested	Chosen
loss	{quantile, absolute_error, poisson, squared_error}	squared_error
max_iter	{100, 200, 300, 400, 500, 1000}	100
max_depth	{None, 15, 20, 25, 35, 40, 50}	None
learning_rate	{0.3, 0.1, 0.01, 0.001}	0.1
min_samples_leaf	{10, 20, 30, 40, 50}	20
quantile	{from 0.1 to 0.9 with step 0.1, None}	None

TABLE IV
THRESHOLDS SET AND FINAL SCORE

Threshold name	Public score
Public	11.179
Naive	12.177
Decision Tree	11.660
csv Feature	10.367
Ensemble	9.470
Our model (Hist)	**9.059**

2) Fine-tuning the Hist model: To determine the optimal configuration of the hyperparameters, we performed a grid search. Each combination of parameters was evaluated using the RMSE score, computed through 15-fold cross-validation on the Development set. The tested hyperparameter values, along with the best-performing configuration, are summarized in Table III. Finally, we also set `warm_start` to `True`. When this option is on, more estimators are added to the ensemble, starting from the estimators obtained from the previous call to `fit`, as such, it is needed that the Hist model is rigorously trained on the same exact data of the previous `fit` call. While not a standard hyperparameter, enabling this option can improve performance when applicable. Notably, this change alone resulted in a gain of 0.040 on the public leaderboard.

III. RESULTS

The MLP model and its variants did not produce significant results, consistently failing to outperform the Hist model with default parameters trained exclusively on *long audios*. Shallow networks performed poorly on both the training and validation sets, while deeper networks achieved excellent training results but demonstrated clear overfitting.

Separation of data by gender for the Hist training did not lead to improvements in performances, as well as dividing the dataset by audio length or cleaning input recordings. Using the best configuration of hyperparameters on the whole original Development set, we achieved a result that was significantly above average, as shown in Table III. Before presenting the final public score obtained, we would like to discuss a few thresholds or baselines that were established during the development of the solution. These baselines are as follows:

- **Public Baseline:** This baseline was the only one not set by us, as it was provided as part of the competition.
- **Naive Baseline:** This threshold was obtained by predicting, for every audio in the evaluation set, the age as the mean of the mean and median ages in the training set for every sample.
- **Decision Tree Baseline:** Since the Hist model is fundamentally based on decision trees, we compared its performance to that of a single fine-tuned decision tree. This serves as a less naive baseline.
- **CSV Feature Baseline:** Another baseline used to evaluate feature extraction involves training the default Hist model solely on the data provided in the *.csv files*, without incorporating any additional feature extracted from the audio files.
- **Ensemble Baseline:** A more sophisticated baseline was established by creating an ensemble of the SVM and Random Forest models.

The scores for these baselines achieved on the public leaderboard, along with the final score gained by our model, are summarized in Table IV. Finally, as shown in Table IV, our model achieved a remarkable public score of 9.059, surpassing the established baselines and demonstrating its effectiveness. The corresponding model with rounded predictions, performed as specified in subsection II-B, reached a score of 9.060 on the public leaderboard.

IV. DISCUSSION

As shown in Table IV, our model consistently outperforms all the established baselines, thanks in part to the effective feature extraction process detailed in Section II-A and the efficiency of the chosen model.

Looking ahead, several avenues could be explored to further improve the achieved performance:

- **Exploring advanced models:** Investigating more complex models, such as convolutional neural networks (CNNs), which have demonstrated strong performance with spectrogram features and audio data in general [24].
- **Using pre-trained models:** Leveraging pre-trained models specifically designed for audio-related tasks, as they have been shown to be highly effective [25], [26].
- **Feature engineering:** Collaborating with medical experts and audio specialists to identify and extract additional features that may better capture age-related vocal characteristics.
- **Demographic feature extraction:** Incorporating additional demographic features to gain further insights. For instance, understanding an individual's level of English proficiency could help account for speaking speed variations. Similarly, considering health conditions, such as smoking, which can alter the fundamental frequency of speech [27], may help address potential outliers and improve the overall performance of the model.

In conclusion, we achieved quite good results, especially considering the inherent difficulty of the task, the scarcity of the provided data, and the significant variability within the dataset. These challenges underscore the robustness and reliability of our proposed approach.

References

[1] H. J. Landau, "Sampling, data transmission, and the nyquist rate," *Proceedings of the IEEE*, vol. 55, pp. 1701–1706, Oct 1967.

[2] S. H. Chen, "Sex differences in frequency and intensity in reading and voice range profiles for taiwanese adult speakers," *Folia Phoniatr Logop / PubMed*, 2007.

[3] Nussbaumer, *The Fast Fourier Transform. In: Fast Fourier Transform and Convolution Algorithms*. Springer-Verlag, 1982.

[4] Zhang, *Wavelet Transform. In: Fundamentals of Image Data Mining*. Springer, Cham, 2019.

[5] T. Wang and Y. C. Lee, "Does restriction of pitch variation affect the perception of vocal emotions in mandarin chinese?," *J. Acoust. Soc. Am.*, Jan 2015.

[6] J. Deng, W. Wan, X. Yu, and W. Yang, "Audio fingerprinting based on spectral energy structure and nmf," in *2011 IEEE 13th International Conference on Communication Technology*, pp. 1103–1106, 2011.

[7] F. W. King, *Hilbert Transforms: Volume 2*. Cambridge University Press, Apr 2009.

[8] W. Jin and X. Fan, "Audio classification algorithm for hearing aids based on robust band entropy information," *Information*, vol. 13, no. 2, 2022.

[9] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using mfcc," in *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 2257–2260, 2017.

[10] S. H. Chen and Y. R. Luo, "Speaker verification using mfcc and support vector machine," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009*, vol. I, March 2009.

[11] 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.

[12] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (cat. No. 01CH37221)*, vol. 1, pp. 73–76, IEEE, 2001.

[13] J. P. Teixeira and P. O. Fernandes, "Jitter, shimmer and hnr classification within gender, tones and vowels in healthy voices," *Procedia Technology*, vol. 16, pp. 1228–1237, 2014.

[14] H. Erblin and B. Eliot, "Overfitting in machine learning: A comparative analysis of decision trees and random forests," *Intelligent Automation & Soft Computing*, 2024.

[15] A. M. Rifat and et al., "Comparison between support vector machine and random forest for audio classification," in *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)*, pp. 1–4, 2021.

[16] B. T. Atmaja and M. Akagi, "Deep multilayer perceptrons for dimensional speech emotion recognition," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 325–331, IEEE, 2020.

[17] "Scikit-learn: Random forest documentation. accessed: 2025-01-26." https://scikit-learn.org/1.6/modules/generated/sklearn.ensemble.RandomForestRegressor.html. Accessed: 2025-01-26.

[18] "Scikit-learn: Support vector machine documentation. accessed: 2025-01-26." https://scikit-learn.org/1.5/modules/generated/sklearn.svm.SVR.html. Accessed: 2025-01-26.

[19] "Scikit-learn: Multi-layer perceptron documentation. accessed: 2025-01-26." https://scikit-learn.org/1.6/modules/generated/sklearn.neural_network.MLPRegressor.html.

[20] "Scikit-learn: Histogram gradient boosting documentation. accessed: 2025-01-26." https://scikit-learn.org/1.5/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html.

[21] N. T. Ira and M. O. Rahman, "An efficient speech emotion recognition using ensemble method of supervised classifiers," in *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pp. 1–5, 2020.

[22] Q. Zheng, P. Zhao, Y. Li, H. Wang, and Y. Yang, "Spectrum interference-based two-level data augmentation method in deep learning for automatic modulation classification," *Neural Computing and Applications*, vol. 33, no. 13, pp. 7723–7745, 2021.

[23] O. O. Abayomi-Alli, R. Damaševičius, A. Qazi, M. Adedoyin-Olowe, and S. Misra, "Data augmentation and deep learning methods in sound classification: A systematic review," *Electronics*, vol. 11, no. 22, p. 3795, 2022.

[24] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An ensemble of convolutional neural networks for audio classification," *Applied Sciences*, vol. 11, no. 13, 2021.

[25] M. Tran and M. Soleymani, "A pre-trained audio-visual transformer for emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4698–4702, 2022.

[26] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[27] M. R. Ayoub, P. Larrouy-Maestri, and D. Morsomme, "The effect of smoking on the fundamental frequency of the speaking voice," *Journal of Voice*, vol. 33, no. 5, pp. 802.e11–802.e16, 2019.