IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 6, December, 2023, pp. 2423-2432
Journal homepage: http://iieta.org/journals/ts

Optimizing Acoustic Feature Selection for Estimating Speaker Traits: A Novel Threshold-Based Approach

Umniah H. Jaid[1,2*], Alia K. AbdulHassan[2]

[1] Department of Computer Science, College of Science, University of Baghdad, Baghdad 10071, Iraq
[2] Department of Computer Science, University of Technology, Baghdad 10066, Iraq

Corresponding Author Email: umniah.h@sc.uobaghdad.edu.iq

https://doi.org/10.18280/ts.400608

Received: 23 March 2023
Revised: 25 September 2023
Accepted: 1 November 2023
Available online: 30 December 2023

Keywords:
acoustic features, age estimation, feature selection, gender detection, height estimation, speaker profiling, TIMIT dataset

ABSTRACT

Speech signals offer a rich array of information about a speaker, encompassing physical attributes and emotional or health states, with significant applications in forensics, security, surveillance, marketing, and customer service. This work aims to identify key acoustic features for estimating an unidentified speaker's height, age, and gender. A novel Forward Feature Selection with Threshold-Based Backward Elimination (FFS-TBE) algorithm is proposed, designed to optimize feature selection across various spectral, temporal, and prosodic dimensions of speech, including Mel-frequency cepstral coefficients (MFCCs), pitch, and formants. Tested against the TIMIT dataset, the FFS-TBE algorithm surpassed traditional feature selection methods like backward and forward sequential feature selection (BSFS/FSFS) and mutual information (MI) statistical methods. It achieved state-of-the-art results, with mean absolute errors (MAEs) of 4.87 cm for male and 4.5 cm for female speakers in height estimation, and MAEs of 4.82 years and 4.91 years for male and female speakers, respectively, in age estimation. Gender prediction accuracy reached 99%. Crucially, the study found that gender-specific feature selection enhances performance, highlighting the distinct acoustic differences between male and female speakers.

1. INTRODUCTION

Speech, as a primary mode of human communication, not only transmits thoughts, ideas, information, and emotions but also encodes a plethora of information about the speaker, including gender, age, and emotional and physical states. The extraction of such information from speech signals has far-reaching applications in surveillance [1], forensics [2, 3], commercial sectors [4, 5], and human-robot interaction [6-8]. Voice, as a biometric, offers the advantages of non-intrusiveness, cost-effectiveness, ease of deployment, and user acceptability [9].

Automated Speaker Profiling (ASP) is increasingly being recognized for its utility in processing vast quantities of voice recordings, a task that is impractical to perform manually [10]. In surveillance systems, ASP enhances data by providing comprehensive profiles, even in cases of obscured or blocked images [1]. In forensic applications, ASP aids in extracting information about suspects from various recordings [2, 3]. Additionally, in commercial settings, ASP finds utility in call routing, playing tailored music/messages, and enhancing customer service [4].

Research has established a correlation between physical build and voice characteristics. Early studies, such as that by Lass and Brown [11], identified strong correlations between voice and physical size. Subsequent research has enabled listeners to estimate speakers' relative size—height and weight—from their voices [12, 13]. A study employing magnetic resonance imaging (MRI) of 129 individuals further confirmed correlations among height, weight, and vocal tract length (VTL) [14]. Age-related characteristics, such as speech rate and fundamental frequency (F0), are indicative of a speaker's age, with younger speakers typically exhibiting a faster speech rate [4] and a decrease in F0 observed with increasing age [15], particularly among female speakers [16]. Furthermore, F0 is crucial for gender detection, given the lower-frequency voices typically associated with male speakers.

Although numerous acoustic features of speech have been identified, determining the most suitable characteristics for specific speaker-profiling applications remains challenging [17]. The estimation of height, age, and gender using a minimal feature set, particularly in the context of advanced machine-learning techniques, is complex due to the overlap of various factors such as sound production systems and the speaker's gender, health, and emotional state, all of which can influence speech characteristics [18].

The present study aims to identify optimal acoustic features for characterizing diverse physical speech traits of speakers. A baseline feature vector encompassing spectral, temporal, prosodic, and harmonic voice aspects is extracted. A novel wrapper feature selection algorithm is then applied and evaluated using the TIMIT dataset. Following quantile normalization of the extracted features, the algorithm selected representative features for specific profiling tasks. The proposed model demonstrated MAEs of 5.16 cm and 4.71 cm

for height estimation of male and female speakers, respectively, and 4.99 years and 5.3 years for age estimation. In gender detection, the algorithm achieved an accuracy of 98.5%.

The remainder of this paper is structured as follows: Section 2 reviews the related literature on speaker profiling, encompassing various features, models, and methods for feature selection and dimensionality reduction. Section 3 delineates the methodology, including feature extraction methods, the proposed algorithm, preprocessing methods, and regression techniques. Section 4 discusses the experimental setup, dataset, evaluation metrics, and results. Finally, Section 5 presents the conclusions drawn from this research.

2. RELATED WORK

Recent advancements in the extraction of paralinguistic content from speech signals have significantly contributed to the field of speaker profiling. This domain primarily involves extracting pivotal features from raw speech signals and employing machine-learning models to facilitate predictions. Historically, various features such as F0 [3, 19-21], MFCC [3, 19, 21], linear predictive coding (LPC) [9, 19, 22], and formants [9, 19, 22] have been proposed and utilized for these tasks. In a notable study, a phone-based approach was employed to estimate a speaker's height and vocal tract length, focusing on the correlation between phone-based short-term features, such as MFCC, Linear Predictive Coding (LPC), and formants [19]. It was observed that 57.15% of the variability in a speaker's height could be attributed to the combined influence of these features. Parallel research utilized vowel regions to predict height, applying formant track regression [9, 22]. This technique achieved MAE of 5.37 cm for male speakers and 5.49 cm for female speakers. Subsequent refinement of the feature set, with the inclusion of line spectral frequencies, led to reduced MAEs of 4.93 cm for male speakers and 4.76 cm for female speakers. However, such phone-based methods for height estimation present limitations, notably their reliance on specific vowels and potential need for speech transcription, which might render them impractical for certain applications.

These features have also been applied effectively in the domains of age estimation and gender detection as well. Zazo et al. [21] utilized MFCC features along with pitch information, employing Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) to estimate the ages of speakers. In a similar vein, Badr et al. [18, 23] harnessed cumulative MFCC and LPC statistics, including their first and second derivatives, along with spectral sub-band coefficients (SSC) and the first four formants (f1-f4). This approach yielded MAEs of 7.73 years for male and 4.96 years for female age estimations using the TIMIT dataset. When applied to the VoxCeleb dataset, the method resulted in MAEs of 10.3 years for males and 9.25 years for females, illustrating the versatility and effectiveness of these acoustic features in diverse datasets.

Several studies have integrated spectral features with temporal and prosodic elements, such as the Harmonic to Noise Ratio (HNR), shimmer, and jitter, to enhance speaker profiling accuracy. Kalluri et al. [17] experimented with various combinations of these features, incorporating jitter, shimmer, and HNR, along with mel spectrograms and their first- and second-order derivatives, formants, and fundamental frequency statistics. This multifaceted approach was applied to predict both the height and age of speakers. The results demonstrated MAEs of 5.2 years for male and 5.6 years for female speakers in age estimation, and 5.2 cm for males and 4.8 cm for females in height estimation, highlighting the efficacy of combining these diverse acoustic features for accurate speaker profiling.

Numerous researchers have adopted statistical methods for speaker profiling, focusing on capturing low-level speech representations. These methods predominantly use short-term features like MFCC and mel spectrograms, forming supervectors. Such supervectors are often based on Gaussian Mixture Models, Universal Background Models (GMM-UBMs) [24], or Hidden Markov Models (HMMs) [25], and are employed for the estimation of attributes such as age, gender, and height [26]. In addition, several studies have utilized i-vectors, which are dimensionally reduced versions of supervectors, coupled with various regression schemes like Support Vector Regression (SVR) and Artificial Neural Networks (ANN) to estimate the age and height of speakers [2, 27]. Further, Grzybowska and Kacprzak [28] investigated the integration of i-vectors with other acoustic features for age estimation, discovering that the combination of i-vectors with additional acoustic features yields more accurate results compared to using i-vectors in isolation.

In the realm of speaker modeling, deep learning methodologies have been increasingly utilized for feature extraction and speaker representation. A notable instance is the work of Sadjadi et al. [29], where Deep Neural Networks (DNN)-based i-vector modeling was implemented as a novel alternative to the conventional GMM approach. This technique yielded phonetically aware i-vectors, which were subsequently integrated into a SVR model for the purpose of age estimation. Moreover, x-vectors have gained prominence in the fields of age and gender estimation for speaker profiling, demonstrating their efficacy through a range of promising outcomes in various studies [30-32].

Identifying an optimal feature set to accurately represent multiple physical attributes in speaker profiling continues to be a formidable challenge. In addressing this, a variety of feature selection methods have been explored. Techniques such as CatBoost [33], relief-based algorithms [1, 34], and dimensionality reduction methods including Principal Component Analysis (PCA) [3, 35] and Linear Discriminant Analysis (LDA) [23, 29, 30] have been applied to diverse feature sets in the field. A significant study by Ganchev et al. [34] involved ranking 6,552 features using the OpenSMILE framework and a relief-based algorithm to identify the most relevant set for height estimation. This comprehensive analysis resulted in the identification of 200 pertinent features, with the top 50 features being utilized to achieve MAEs of 5.3 cm for male and 5.2 cm for female speakers. Additionally, the CatBoost optimization algorithm has been employed for age and gender classification [33], demonstrating notable accuracy rates of 89.62% for age group prediction and 72.29% for gender detection.

This paper's research builds upon prior studies, expanding the exploration into a broad array of spectral, prosodic, and temporal features. It introduces a novel feature selection algorithm designed to identify representative features for accurately estimating the height, age, and gender of both male and female speakers. While existing research provides valuable insights, it often concentrates on single physical traits [1, 23, 34] or limits speech utterances to constrained representations [2, 30]. This study sets itself apart by

examining an extensive set of features and employing a unique wrapper feature selection algorithm. This algorithm is distinct in its approach, implementing a threshold-based acceptance criterion during sequential forward selection and incorporating a hybrid method that amalgamates both forward selection and backward elimination techniques. Such an approach minimizes the inclusion of marginally beneficial features, thereby ensuring a more refined and effective feature set for speaker profiling.

3. METHODOLOGY

The methodology employed in this study encompasses three distinct stages. Initially, data were sourced from the TIMIT dataset, from which requisite features were extracted and subjected to pre-processing to ensure their relevance and applicability to the task of speaker trait estimation. Subsequently, a novel feature selection algorithm was applied to these extracted features to identify a subset optimal for modeling. The final stage involved the utilization of SVR for the estimation of height, age, and gender. The following subsections detail each of these stages.

3.1 Dataset

The TIMIT dataset, primarily designed for Automatic Speech Recognition (ASR), was utilized for speaker profiling in this research. This dataset encompasses metadata for each participant, including height, age, education level, ethnicity, and regional dialect. Contributions from each participant included 10 speech recordings, culminating in a total of 6,300 utterances. These utterances were segregated into distinct training and test sets, with the training set comprising 326 male and 136 female speakers (totaling 462 speakers) and the test set including 168 speakers (112 males and 56 females). The standard TIMIT train/test split was adopted for this study.

It is noteworthy that the original distribution of data in the TIMIT dataset introduced imbalances in the categories of height, age, and gender, primarily due to its design orientation towards speech recognition. Such imbalances have been reported to influence model performance and prediction accuracy [17, 26]. Nevertheless, to maintain comparability with other studies, the original train/test split was retained.

3.2 Feature extraction

In this study, an extensive set of spectral, prosodic, voice quality, and articulatory features was extracted employing the OpenSMILE toolkit. The (extended) Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [28] was chosen for its wide range of acoustic and prosodic features, offering a standardized baseline for evaluation and mitigating variability from different parameter sets. Two versions of eGeMAPS are noted: a minimalistic and an extended version. The minimalistic set comprises 18 low-level descriptors (LLDs), categorized into three groups:

- Frequency-related parameters, including F0; Jitter; formants 1, 2, and 3; frequency; and first-formant bandwidth.
- Energy/amplitude-related parameters, including shimmer, loudness, and HNR,
- Spectral parameters, such as alpha ratio; Hammarberg index; spectral slope; the relative energy of formants 1, 2, and 3; harmonic differences H1–H2; and harmonic differences H1–A3.

For each of these 18 LLDs, arithmetic means and standard deviations (stddevs) were computed, resulting in 36 parameters. Additionally, 8 functionals were applied to loudness and pitch across various percentiles, alongside means and stddevs of the slopes of rising/falling signal parts, culminating in 52 parameters. The arithmetic means of alpha ratios, Hammarberg indices, and spectral slopes further extended this to 56 parameters. Including 6 temporal features such as the rate of loudness peaks, the mean lengths and stddevs of continuously voiced regions where F0 was non-zero, the mean lengths and stddevs of unvoiced regions, and the number of continuous voiced regions per second, brought the total to 88 acoustic features.

Additional feature extraction was performed using OpenSMILE, supplementing the features obtained from eGeMAPS. For this, nineteen functionals were computed for the first 12 MFCCs, along with voicing probabilities and F0, generating a total of 255 features. These functionals comprised various statistical measures, including minimum and maximum values, the range between these extremes, and their temporal positions. Additionally, the arithmetic means and the slope coefficient derived from a linear regression applied to the first coefficient of the smoothed MFCCs were calculated. The mean absolute error of this regression, along with measures of kurtosis, skewness, and the first three quartiles, including their interquartile range, were also included.

Furthermore, three additional sets of features were extracted utilizing Praat software [36], focusing on prosodic aspects of speech. This included the measurement of phonation time, which reflects the duration of vowel sustenance. Additionally, the number of pauses in speech was recorded, providing insights into speech patterns. Speech rate was also assessed, gauging the speed at which a speaker communicates. Similarly, articulation rate was evaluated, which measures the velocity of syllable pronunciation. Lastly, articulation was analyzed, offering a metric for the rapidity of syllable enunciation by the speaker.

Additionally, the study incorporated the statistical functions of the first four formants, encompassing their means, stddevs, and medians. A range of amplitude-related features of the speech signal was also analyzed to gauge amplitude variations. This analysis included local jitter, local absolute jitter, RAP Jitter, PPQ5 jitter, DDP jitter, local shimmer, APQ3 shimmer, APQ5 shimmer, APQ11 shimmer, and DDA shimmer. These features were instrumental in capturing the nuances of voice quality and fluctuation. The inclusion of these comprehensive measures resulted in a final feature set comprising a total of 408 distinct features, offering a broad and detailed perspective for accurate speaker profiling.

3.3 Data pre-processing

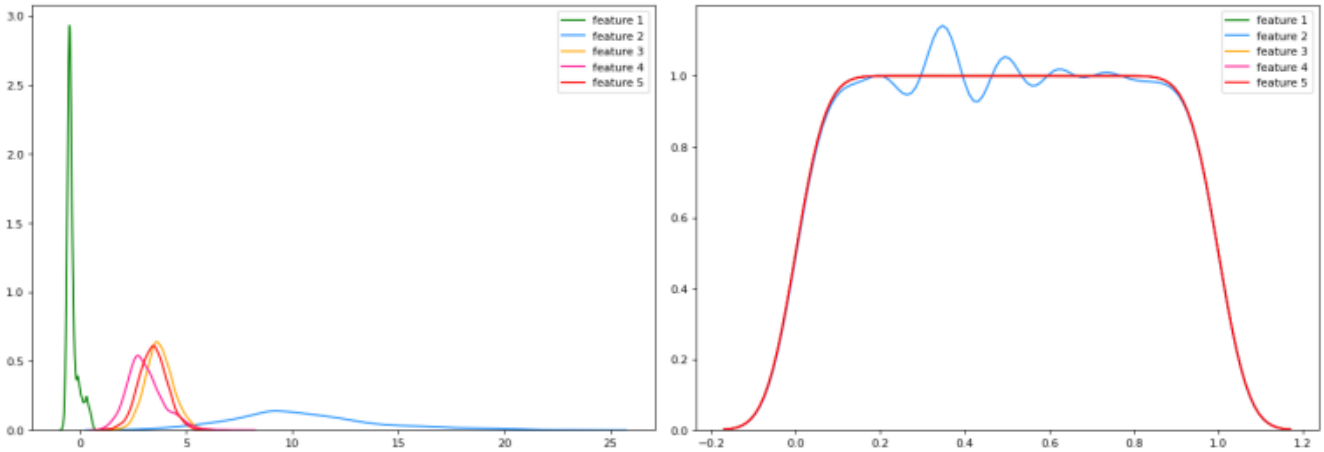
In this phase, quantile normalization was employed to standardize the feature expressions, thereby reducing variance and mitigating bias in the learning model towards specific values. This technique transforms the features, aligning various samples with differing statistical distributions to a uniform target distribution.

Although quantile normalization may potentially obscure certain distinct differences between features, it offers the advantage of being less susceptible to extreme values compared to methods such as Min-Max scaling. Crucially, it preserves the inherent relationships between samples. By

ensuring that features adhere to a common distribution, this method facilitates more effective subsequent analysis.

The quantile normalization process involves initially sorting all feature values in ascending order. Subsequently, these values are assigned ranks, followed by the calculation of the mean for each rank. This procedure ensures that the resultant dataset exhibits a uniform distribution of values across all features, while maintaining the original inter-sample relationships.

Figure 1 illustrates the impact of three different transformations on a selection of five features randomly chosen from the dataset. Panel 1a displays the feature distributions prior to transformation, while Panel 1b depicts the distributions post quantile normalization, clearly demonstrating the effect of this normalization technique on the feature set.



(a) No transformation (b) Quantile normalisation

Figure 1. Effects of different feature transformation methods on the data

3.4 Proposed feature selection

In this study, a novel wrapper feature selection algorithm, named FFS-TBE, was developed and implemented. Wrapper feature selection is a method in machine learning that employs a predictive model to identify the most pertinent features. This approach relies on the model's performance to evaluate and validate the effectiveness of selected features. The FFS-TBE algorithm operates by sequentially incorporating features into a selected pool and then recursively reassessing each feature's contribution based on performance metrics and a predefined threshold.

The algorithm considers two critical factors: the performance of the selected features and their interactions. As depicted in Figure 2, the FFS-TBE process encompasses two principal phases: sequential feature addition and re-evaluation.

Given a feature set $X = \{x_1, x_2, x_3, .., x_d\}$ in a d-dimensional space, where d represents the total number of features, the objective is to extract a subset of features $S_k = \{x_j | j = 1, 2, 3, \ldots, k, x_j \in X\}$, where k is the count of selected features. This subset S_k aims to maximize the criterion function, which is indicative of optimal model performance.

The FFS-TBE algorithm initiates with an empty set S ($S = \emptyset$) for selected features. Each feature x_i from X is sequentially appended to S for evaluation. If the inclusion of x_i fails to enhance performance, it is discarded; otherwise, if x_i improves performance within a specific threshold, it is retained in S, and the set undergoes re-evaluation. The threshold value plays a pivotal role in this process, allowing a feature with potential performance improvement to be compared with existing features in the set. A higher threshold enables a more extensive evaluation of feature interactions but risks overfitting and reduced generalization to new data. In contrast, a lower threshold streamlines the algorithm but may limit thorough re-evaluation.

During the re-evaluation phase, the significance of each feature in the current set S is assessed by individually removing features and observing the resultant impact on model performance. Features that contribute the least to performance are eliminated, refining the feature set to achieve the best performance. This recursive procedure continues until the most effective subset of features is identified. The detailed methodology of the process is outlined in Algorithm 1.

	Function: Re-evaluate
	Input: *Dataset (X, Y) // X is the feature vector and Y is the target vector* *Best performance (best_performance), Selected features (selected_features)* ***Output:*** *Updated Best performance (best_performance)* *Updated Selected features (selected_features)*
1	***Initialise:*** *empty set (performance_drops)*
2	***For (i = 1 to number of features in X)***
3	*Remove feature i from selected_features*
4	*Evaluate performance of model without feature i // Evaluation Metric such as: MAE, accuracy, etc.*
5	*record performance in performance_drop set*
6	*Add feature i back to selected_features*
7	***End***
8	*Find minimum value in performance drop*
9	Compare the minimum drop with the current best performance
10	If the minimum drop is better than the current best performance, then:
11	remove corresponding feature from selected features
12	update best performance with the minimum drop
13	***End***

3.5 SVR

In the implementation of the proposed wrapper feature selection algorithm, Epsilon-Support Vector Regression (ε-SVR) was utilized as the predictive model for the physical parameters of height, age, and gender. Throughout the feature selection process, SVR functioned as the primary model to fit the selected features and facilitate the evaluation of their performance.

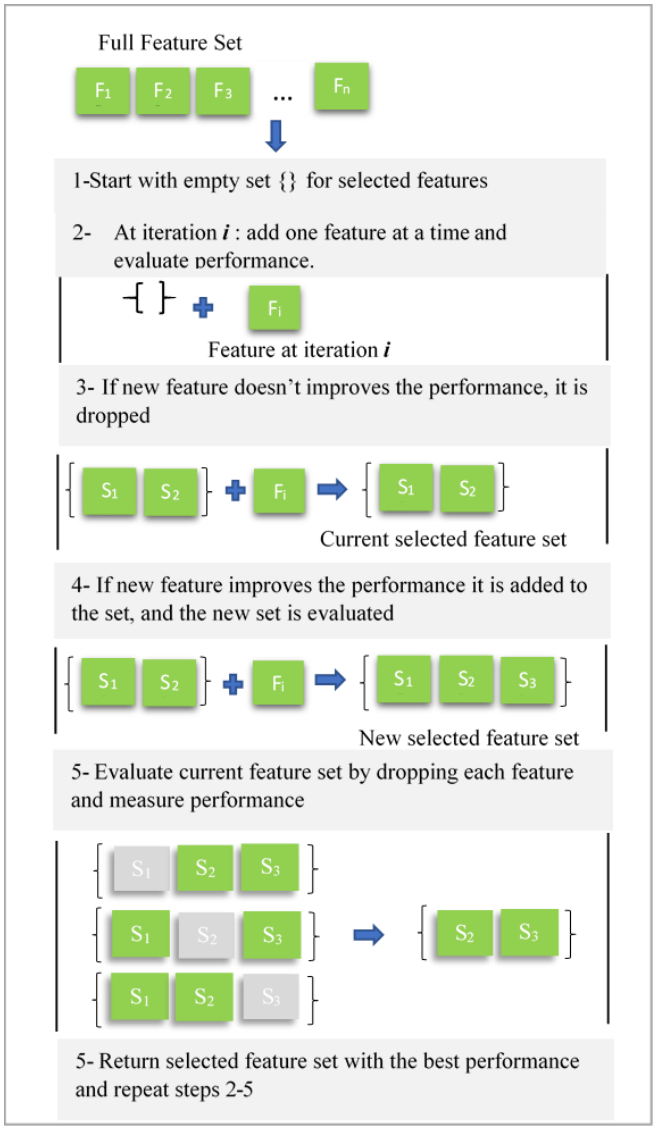


Figure 2. An example of the workflow of the proposed FFS-TBE feature selection algorithm

SVR, an extension of Support Vector Machines (SVM) to regression tasks, is well-suited for such applications. In this research, ε-SVR was specifically chosen for its proficiency in predicting the aforementioned physical parameters. The selection of SVR was based on its ability to efficiently handle high-dimensional data, along with its capability to effectively model non-linear relationships that are often not readily apparent. This characteristic of SVR makes it an ideal tool for dealing with complex data patterns, thereby enhancing the accuracy and reliability of the physical parameter predictions in speaker profiling.

Given a set of training data $\{(x_1,y_1), (x_2,y_2) \ldots (x_n, y_n)\}$, where x_i is the d-dimensional feature vector and y_i is the corresponding target, to predict a target output, the relationship between the target and the features is given by Eq. (1):

$$f(x) = w^T x + b \tag{1}$$

where, w is the weight vector, and b is the bias term. The objective is to minimize Eq. (2):

$$MIN\ \frac{1}{2}\ \|w\|^2 \tag{2}$$

Within the constraints of Eq. (3):

$$|w^T x_i + b - y_i| < \varepsilon \tag{3}$$

where, ε is the margin of accepted error (i.e., deviation from the target output).

Although a notable limitation of SVR is its sensitivity to hyperparameters, this study utilized the default parameters of the scikit-learn SVR implementation to ensure consistency and reproducibility in the experimental results.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section details the experiments conducted for estimating height, age, and gender. The impact of various feature set combinations on each of these tasks is evaluated using specific metrics that are appropriately tailored to each estimation problem. For height and age estimation, two key metrics are employed: the MAE and Pearson's correlation coefficient. Gender prediction, on the other hand, is assessed using accuracy as the primary metric.

The MAE is a critical metric for these evaluations and is calculated as follows:

$$MAE = \frac{\sum |y_i - x_i|}{n} \tag{4}$$

where, y_i is the predicted value, x_i is the target value, and n is the number of observations.

The Pearson's correlation coefficient, which assesses the linear relationship between the actual and estimated vectors, is calculated using the following formula:

$$p = \frac{1}{N-1} \sum_{n=1}^{N} \left(\frac{\hat{y}_n - \overline{\hat{y}}}{\sigma_{\hat{y}}}\right)\left(\frac{y_n - \overline{y}}{\sigma_y}\right) \tag{5}$$

where, $\overline{y}$ and σ_y represent the mean and standard deviation of the true values, respectively, while $\overline{\hat{y}}$ and $\sigma_{\hat{y}}$ are the same metrics for the estimated values.

Accuracy is defined as:

$$Accuracy = \frac{No.of\ Correct\ Predictions}{Total\ No.of\ Predictions} \tag{6}$$

For the model evaluation, two distinct approaches were utilized in feature selection using the FFS-TBE algorithm. The first approach strictly confined feature selection to the training data, ensuring an unbiased methodology in line with conventional model evaluation practices. In contrast, the second approach adopted a more inclusive strategy, incorporating both training and testing data during feature selection. This method initially identified relevant features from the training set, followed by further refinement and

validation using the test set. Although this approach may introduce an optimistic bias due to the inclusion of future (testing) data in the feature selection process, it offers valuable insights into the model's potential performance under comprehensive data consideration.

In this study, each speaker contributed 10 recordings, thus the MAE was calculated at the speaker level as an average across these recordings. This approach ensures a more consistent and representative error measurement for each speaker, accounting for variations across multiple sessions.

Additionally, gender-dependent experiments were conducted separately for male and female speakers. The rationale behind this segmentation was to isolate the analysis to a specific gender, thereby eliminating the influence of gender-specific variances. SVR was employed as the predictive model, with MAE serving as the evaluation metric for height and age estimations, and accuracy as the metric for gender detection. All experiments were rigorously evaluated using the test data set.

4.1 Effect of feature transformation

This study examined the efficacy of three distinct feature transformation techniques within the feature set: min-max normalization, z-score normalization, and quantile normalization, alongside an analysis with no transformation applied. An initial experiment was conducted using the entire dataset to identify the transformation technique that yielded the best performance. As indicated in Table 1, quantile normalization demonstrated notable improvements: a 10% enhancement in age estimation, a 3% increment in height estimation, and a 6.5% increase in gender detection accuracy, compared to the untransformed case. Furthermore, quantile normalization surpassed the other transformation methods across all tasks. Consequently, this transformation method was adopted for all subsequent experiments to ensure optimal performance.

Table 1. Effect of different feature transformation methods on the estimation of height, age, and gender for both male and female speakers

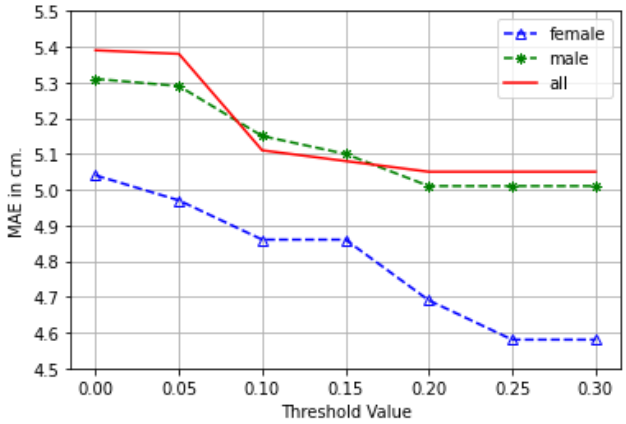
Normalization Method	Age		Height		Gender
	Female	Male	Female	Male	
No Normalization	6.4	5.85	6.11	5.39	93%
Min-Max	5.68	5.32	6.01	5.35	98%
z-score	5.4	5.42	5.95	5.3	99%
Quantile	5.29	5.42	5.92	5.22	99%

4.2 Choosing a threshold value

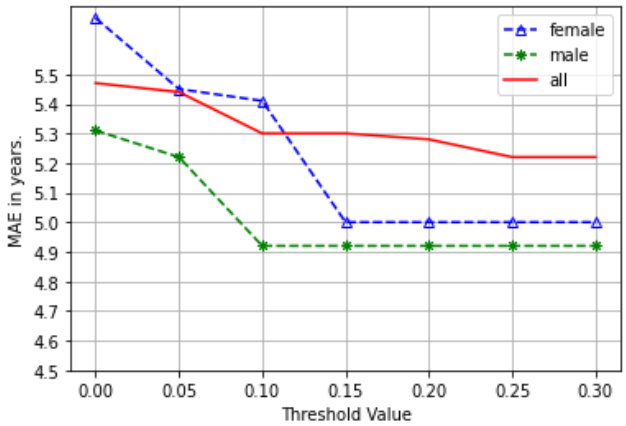
The threshold value plays a pivotal role in the proposed algorithm, as it governs the re-evaluation process. Figures 3(a) and 3(b) illustrate the performance associated with varying threshold values. These visual representations indicate a trend where an escalation in the threshold value initially leads to an enhancement in performance, up to a certain point where it plateaus. Notably, the experimental results revealed that the data for female speakers necessitated higher threshold values for achieving optimal performance compared to male data. This observation is reflective of the sample size disparity, with female speaker samples being less abundant. Consequently, the threshold value demonstrates a proportional relationship to the sample count in the dataset, underscoring its significance in the algorithm's efficacy.

4.3 Individual feature contributions

In order to discern the influence of various feature types on the estimation of height, age, and gender, the feature set was categorized into ten groups based on their characteristics. These groups included features based on fundamental frequency (F0), formants, spectral, temporal, energy, MFCC, voice quality, voicing probability, and prosodic elements, as well as harmonic differences.



(a) Height



(b) Age

Figure 3. Effect of threshold value on the performance of the FFS-TBE feature selection algorithm

Each feature group was individually evaluated for its effectiveness in each task. Figures 4, 5, and 6 showcase the performance of each feature category in estimating age, height, and gender, respectively. Overall, F0 and voice quality features emerged as the most effective, particularly notable in the lower MAEs for female height compared to male height. However, MFCC-related features exhibited a balanced performance, with similar MAEs of 5.38 for females and 5.33 for males, indicating an equitable effectiveness for both genders. Additionally, MFCC features generally outperformed other groups in most tasks.

Contrastingly, other feature categories such as prosodic, temporal, and energy demonstrated higher MAEs in female height estimation but similar results for male and female height. Harmonic, spectral, and formant features showed a consistent performance across all tasks for both genders. These findings suggest distinct strengths and weaknesses in different feature groups for male and female speakers.

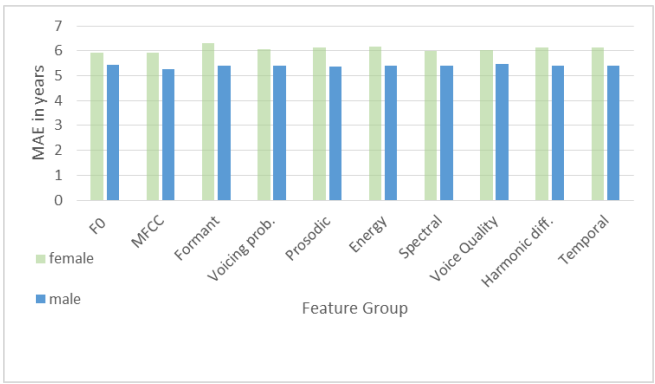


Figure 4. Performance of different feature groups on age estimation

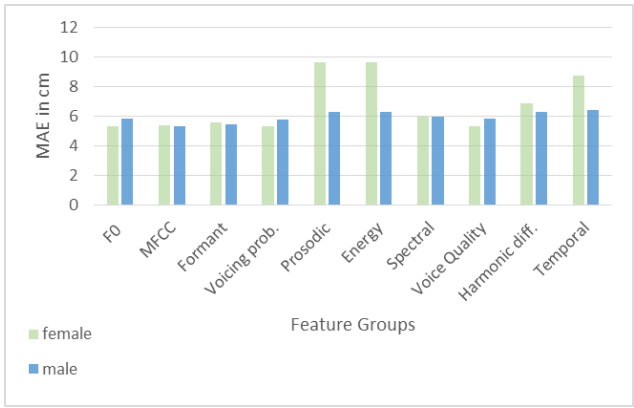


Figure 5. Performance of different feature groups on height estimation

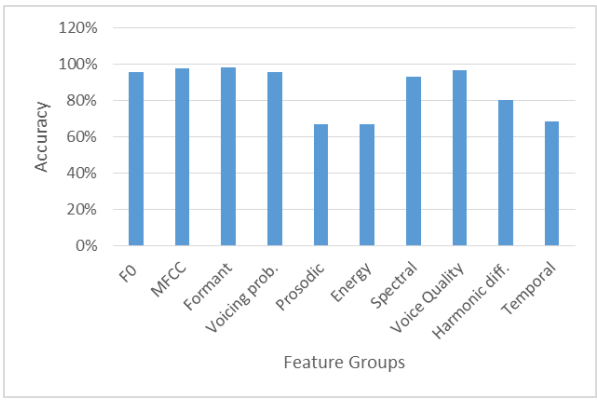


Figure 6. Performance of different feature groups on gender detection

In terms of age estimation, no single feature group significantly outperformed others in discerning ages of both females and males. The MAE values were relatively uniform across genders, with some groups showing slightly higher MAEs for females and others for males. Nevertheless, MFCC features consistently delivered the best performance in age estimation for both genders. For gender detection, formant features showed impressive accuracy (97.9%), closely rivaling the accuracy achieved with MFCC features (97.6%). Additionally, F0 features and voice quality were effective in gender discrimination, achieving accuracies of 96% and 96.8%, respectively, with the mean F0 alone yielding a 96.5% accuracy in gender detection. However, features such as temporal, prosodic, and energy demonstrated lesser efficacy in gender detection tasks.

Optimal estimation results across all tasks were consistently achieved when utilizing the complete set of features. Notably, specific feature groups such as prosodic, energy, and temporal features exhibited lower performance levels in height and gender detection tasks. A recurrent observation was that the accuracy of estimations for female speakers tended to be inferior compared to male speakers across all feature categories. This discrepancy is primarily attributable to the relatively limited data available for female speakers. Among the various groups of features, MFCC features distinctly outperformed others in all tasks, demonstrating their robustness and effectiveness in speaker profiling.

4.4 Height estimation

In the female height estimation experiment utilizing the FFS-TBE method with only training data, an MAE of 5.06 and a Pearson's correlation coefficient (p) of 0.30 were achieved, with 29 features selected. This p-value of 0.30 indicates a weak-to-moderate positive linear correlation between estimated and actual heights. However, when the test set was incorporated into the feature selection process, a notable improvement was observed, with the MAE decreasing to 4.5 and the p-value rising to 0.51, achieved with just 14 features. This higher p-value signifies a more pronounced moderate positive correlation, suggesting a more linearly predictable relationship between estimated and actual heights, despite a reduction in the number of features.

For male height estimation, the algorithm, relying solely on training data, selected 47 features and resulted in an MAE of 5.27 and a p value of 0.172, indicating a weak positive linear correlation between the estimated and actual heights. Conversely, integrating both training and testing data in the feature selection process led to a lower MAE of 4.87 and an improved p value of 0.346, achieved with only 16 features. The enhanced p value points to a stronger linear relationship between estimated and real height values, even with fewer features involved.

These variations in Pearson's correlation coefficient values underscore the significance of feature selection in bolstering the linear predictability of the models. An increased p-value, coupled with a lower MAE, suggests not only more accurate height predictions but also a more linear correlation between predicted and actual values, thereby reinforcing confidence in the model's efficacy.

Table 2 presents the detailed experimental outcomes for height estimation, comparing the results obtained using only the training set for feature selection with those achieved when both training and testing sets were utilized.

Table 2. MAE and Pearson's correlation coefficient (p) results of the proposed feature selection algorithm (FFS-TBE) for height estimation of male and female speakers

Mode of Evaluation	Female (MAE / p)	Male (MAE / p)	No. of Features
Baseline Features	5.31/0.185	5.44/0.118	408
Training Data Only	5.06/0.30	5.27/0.172	29/47
Training + Testing Data	4.5/0.51	4.87/ 0.346	14/16

4.5 Age estimation

In the gender-specific age estimation experiments, the

selection of features resulted in 15 features for female age estimation and 23 features for male age estimation. The MAEs achieved were 5.0 years for females and 4.9 years for males. In contrast, the gender-agnostic approach in age estimation yielded MAEs of 5.76 years for female speakers and 5.04 years for male speakers. The outcomes derived from the selected features are detailed in Table 3.

These results demonstrate a significant improvement in estimation accuracy, with a 12% enhancement in female age estimation and a 6.3% improvement in male age estimation, compared to the outcomes obtained using the complete set of features.

Table 3. MAE and Pearson's correlation coefficient (p) results of the proposed feature selection algorithm (FFS-TBE) for age estimation of male and female speakers

Mode of Evaluation	Female (MAE / p)	Male (MAE / p)	No. of Features
Baseline Features	6.03/0.397	5.33/0.226	408
Training Data Only	5.3 /0.645	5.32/0.321	29/56
Training + Testing Data	4.91/0.71	4.82/0.476	14/27

4.6 Comparative analysis of the proposed method

This study also includes a comparative analysis to evaluate the effectiveness of the proposed feature selection scheme. This was achieved by contrasting its performance with other established feature selection algorithms, namely, the wrapper BSFS, FSFS, and the MI statistical method. The comparative results, as detailed in Table 4, reveal that the proposed FFS-TBE algorithm surpassed these conventional algorithms in performance when applied to the same dataset. This comparison underscores the enhanced efficacy and robustness of the proposed feature selection approach in speaker profiling tasks.

Table 4. The performance of the proposed feature selection method compared to other methods

Feature Selection Method	Height		Age		Gender
	female	male	female	Male	
MI	5.35	5.55	5.98	5.41	97%
BSFS	5.14	5.46	5.8	5.3	97.5%
FSFS	5.22	5.45	5.75	5.37	98%
FFS-TBE	5.06	5.27	5.3	5.32	99%

Table 5. Comparison of the performance of the proposed features with state-of-the-art results using the TIMIT dataset

Study	Height MAE		Age MAE	
	Male	Female	Male	Female
Kaushik et al. [20]	5.24	5.09	5.62	6.08
Badr et al. [23]	-	-	7.73	4.96
Williams and Hansen [22]	5.37	5.49	-	-
Ganchev et al. [34]	5.3	5.2	-	-
Kalluri et al. [17]	5.2	4.8	5.2	5.6
Proposed FFS-TBE	4.87	4.5	4.82	4.91

Furthermore, the results achieved with the proposed algorithm surpassed those documented in existing literature, setting a new benchmark for height and age estimation using the TIMIT dataset. This advancement is clearly demonstrated in Table 5, which compares the performance of the proposed method against previous state-of-the-art achievements.

5. CONCLUSIONS

This research delved into the speech characteristics indicative of various physical traits of speakers. A novel wrapper feature selection algorithm was proposed, focusing on a broad array of spectral, temporal, energy, and prosodic features to predict a speaker's height, age, and gender solely from speech signals. The algorithm was rigorously trained and tested using the TIMIT dataset. Results from the experiments highlighted inherent gender-specific variations in speech and audio features, impacting sound production. While some overlap was observed in features delineating male and female characteristics, distinct features were identified that capture the unique spectral properties of each gender, underscoring the necessity of gender prediction prior to broader speaker profiling tasks.

The study also revealed that the choice of feature normalization substantially affects the outcome quality, with quantile normalization markedly enhancing model performance compared to other methods. The implemented feature selection algorithm demonstrated superior performance over other wrapper methods applied to the same dataset, attaining exemplary results in height and age estimation across genders, along with a 99% accuracy rate in gender prediction.

However, it is imperative to acknowledge the imbalance in the dataset utilized, which could influence the generalizability of these findings to other datasets or real-world scenarios. While the efficiency of the feature selection algorithm is noteworthy, its impact on overall results and its scalability and adaptability in more complex scenarios warrant further investigation.

Future research directions include the exploration of multitask profiling using a unified set of features. Additionally, there is potential in examining a wider spectrum of speaker characteristics such as emotions, health status, and regional accents or dialects, which could be invaluable in various applications ranging from health monitoring to forensic analysis. Furthermore, the adaptation of these algorithms to datasets recorded in diverse environments is a crucial step toward applicability in real-world scenarios.

The conclusions drawn from this research provide significant insights into the relationship between speech characteristics and physical traits, while also highlighting areas for further exploration to enhance the robustness and applicability of speaker profiling methodologies.

REFERENCES

[1] Mporas, I., Ganchev, T. (2009). Estimation of unknown speaker's height from speech. International Journal of Speech Technology, 12: 149-160. https://doi.org/10.1007/s10772-010-9064-2

[2] Poorjam, A.H., Bahari, M.H. (2014). Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals. In 2014 4th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, pp. 7-

12. https://doi.org/10.1109/ICCKE.2014.6993339
[3] Beke, A. (2018). Forensic speaker profiling in a Hungarian speech corpus. In 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Budapest, Hungary, pp. 000379-000384. https://doi.org/10.1109/CogInfoCom.2018.8639932
[4] Müller, C. (2006). Automatic recognition of speakers' age and gender on the basis of empirical studies. In Ninth International Conference on Spoken Language Processing, pp. 2118-2121.
[5] Müller, C., Burkhardt, F. (2007). Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age. In Eighth Annual Conference of the International Speech Communication Association, pp. 2277-2280.
[6] Kim, H.J., Bae, K., Yoon, H.S. (2007). Age and gender classification for a home-robot service. In RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication, Jeju, Korea (South), pp. 122-126. https://doi.org/10.1109/ROMAN.2007.4415065
[7] Lee, M.W., Kwak, K.C. (2012). Performance comparison of gender and age group recognition for human-robot interaction. International Journal of Advanced Computer Science and Applications, 3(12): 207-211.
[8] Badr, A.A., Abdul-Hassan, A.K. (2020). A review on voice-based interface for human-robot interaction. Iraqi Journal for Electrical and Electronic Engineering, 16(2): 91-102. https://doi.org/10.37917/ijeee.16.2.10
[9] Hansen, J.H., Williams, K., Bořil, H. (2015). Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models. The Journal of the Acoustical Society of America, 138(2): 1052-1067. https://doi.org/10.1121/1.4927554
[10] Kelly, F., Forth, O., Atreya, A., Kent, S., Alexander, A. (2017). What your voice says about you: Automatic Speaker Profiling using i-vectors. IAFPA. Proceedings of IAFPA2017. Split, Croatia: IAFPA, 72-75.
[11] Lass, N.J., Brown, W.S. (1978). Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. The Journal of the Acoustical Society of America, 63(4): 1218-1220. https://doi.org/10.1121/1.381808
[12] Smith, D.R., Patterson, R.D., Turner, R., Kawahara, H., Irino, T. (2005). The processing and perception of size information in speech sounds. The Journal of the Acoustical Society of America, 117(1): 305-318. https://doi.org/10.1121/1.1828637
[13] Van Dommelen, W.A., Moxness, B.H. (1995). Acoustic parameters in speaker height and weight identification: Sex-specific behaviour. Language and Speech, 38(3): 267-287. https://doi.org/10.1177/002383099503800304
[14] Fitch, W.T., Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. The Journal of the Acoustical Society of America, 106(3): 1511-1522. https://doi.org/10.1121/1.427148
[15] Nishio, M., Niimi, S. (2008). Changes in speaking fundamental frequency characteristics with aging. Folia Phoniatrica et Logopaedica, 60(3): 120-127. https://doi.org/10.1159/000118510
[16] Eichhorn, J.T., Kent, R.D., Austin, D., Vorperian, H.K. (2018). Effects of aging on vocal fundamental frequency and vowel formants in men and women. Journal of Voice, 32(5): 644-e1-644.e9. https://doi.org/10.1016/j.jvoice.2017.08.003
[17] Kalluri, S.B., Vijayasenan, D., Ganapathy, S. (2020). Automatic speaker profiling from short duration speech data. Speech Communication, 121: 16-28. https://doi.org/10.1016/j.specom.2020.03.008
[18] Badr, A.A., Abdul-Hassan, A.K. (2021). Age estimation in short speech utterances based on bidirectional gated-recurrent neural networks. Engineering and Technology Journal, 39(1B): 129-140. https://doi.org/10.30684/etj.v39i1B.1905
[19] Dusan, S. (2005). Estimation of speaker's height and vocal tract length from speech signal. In Ninth European Conference on Speech Communication and Technology, pp. 1989-1992.
[20] Kaushik, M., Pham, V.T., Chng, E.S. (2021). End-to-end speaker height and age estimation using attention mechanism with LSTM-RNN. arXiv preprint arXiv:2101.05056. https://doi.org/10.48550/arXiv.2101.05056
[21] Zazo, R., Nidadavolu, P.S., Chen, N., Gonzalez-Rodriguez, J., Dehak, N. (2018). Age estimation in short speech utterances based on LSTM recurrent neural networks. IEEE Access, 6: 22524-22530. https://doi.org/10.1109/ACCESS.2018.2816163
[22] Williams, K.A., Hansen, J.H. (2013). Speaker height estimation combining GMM and linear regression subsystems. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, pp. 7552-7556. https://doi.org/10.1109/ICASSP.2013.6639131
[23] Badr, A.A., Abdul-Hassan, A.K. (2021). Estimating age in short utterances based on multi-class classification approach. Computers, Materials & Continua, 68: 1713-1729. https://doi.org/10.32604/cmc.2021.016732
[24] Bahari, M.H., Van Hamme, H. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. In 2011 IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), Milan, Italy, pp. 1-6. https://doi.org/10.1109/BIOMS.2011.6052385
[25] Bahari, M.H. (2012). Speaker age estimation using Hidden Markov Model weight supervectors. In 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), Montreal, QC, Canada, pp. 517-521. https://doi.org/10.1109/ISSPA.2012.6310606
[26] Kalluri, S.B., Vijayasenan, D., Ganapathy, S. (2019). A deep neural network based end to end model for joint height and age estimation from short duration speech. In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, pp. 6580-6584. https://doi.org/10.1109/ICASSP.2019.8683397
[27] Poorjam, A.H., Bahari, M.H., Vasilakakis, V. (2015). Height estimation from speech signals using i-vectors and least-squares support vector regression. In 2015 38th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, pp. 1-5. https://doi.org/10.1109/TSP.2015.7296469
[28] Grzybowska, J., Kacprzak, S. (2016). Speaker age classification and regression using i-vectors. In

INTERSPEECH, pp. 1402-1406. http://dx.doi.org/10.21437/Interspeech.2016-1118
[29] Sadjadi, S.O., Ganapathy, S., Pelecanos, J.W. (2016). Speaker age estimation on conversational telephone speech using senone posterior based i-vectors. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, pp. 5040-5044. https://doi.org/10.1109/ICASSP.2016.7472637
[30] Ghahremani, P., Nidadavolu, P.S., Chen, N., Villalba, J., Povey, D., Khudanpur, S., Dehak, N. (2018). End-to-end deep neural network age estimation. In Interspeech, 2018: 277-281.
[31] Kwasny, D., Hemmerling, D. (2021). Gender and age estimation methods based on speech using deep neural networks. Sensors, 21(14): 4785. https://doi.org/10.3390/s21144785
[32] Kwasny, D., Hemmerling, D. (2020). Joint gender and age estimation based on speech signals using x-vectors and transfer learning. arXiv preprint arXiv:2012.01551. https://doi.org/10.48550/arXiv.2012.01551
[33] Badr, A.A., Abdul-Hassan, A.K. (2021). CatBoost machine learning based feature selection for age and gender recognition in short speech utterances. International Journal of Intelligent Engineering & Systems, 14(3): 150-159. https://doi.org/10.22266/ijies2021.0630.14
[34] Ganchev, T., Mporas, I., Fakotakis, N. (2010). Audio features selection for automatic height estimation from speech. In: Konstantopoulos, S., Perantonis, S., Karkaletsis, V., Spyropoulos, C.D., Vouros, G. (eds) Artificial Intelligence: Theories, Models and Applications. SETN 2010. Lecture Notes in Computer Science(), vol 6040. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-12842-4_12
[35] Babu, K.S., Vijayasenan, D. (2017). Robust features for automatic estimation of physical parameters from speech. In TENCON 2017-2017 IEEE Region 10 Conference, Penang, Malaysia, pp. 1515-1519. https://doi.org/10.1109/TENCON.2017.8228097
[36] Boersma, P. (2006). Praat: Doing phonetics by computer (version 4.4. 24). http://www. praat.org/.