DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

Ensemble Feature Selection for Age Estimation from Speech

Umniah Hameed Jaid[1], Alia Karim Abdulhassan[2]
[1]Computer Science Department, College of Science, University of Baghdad, Baghdad, Iraq
[2]Computer Science Department, University of Technology, Baghdad, Iraq
[1]umniah.h@sc.uobaghdad.edu.iq, [2]alia.k.abdulhassan@uotechnology.edu.iq

Abstract*— The voice signal carries a wide range of data about the speaker, including their physical characteristics, feelings, and level of health. There are several uses for the estimate of these physical characteristics from the speech in forensics, security, surveillance, marketing, and customer service. The primary goal of this research is to identify the auditory characteristics that aid in estimating a speaker's age. To this end, an ensemble feature selection model is proposed that selects the best features from a baseline acoustic feature vector for age estimation from speech. Using a feature vector that covers various spectral, temporal, and prosodic aspects of speech, an ensemble-based automatic feature selection is performed by, first calculating the feature importance or ranks based on individual feature selection methods, then voting is applied to the resulting feature ranks to attain the top-ranked subset by all feature selection methods. The proposed method is evaluated on the TIMIT dataset and achieved a mean absolute error (MAE) of 5.58 years and 5.12 years for male and female age estimation.*

Index Terms*— Age Estimation, Feature Selection, Ensemble Selection, TIMIT dataset.*

I. INTRODUCTION

Speech is a key component of human interaction since it allows people to convey their thoughts, ideas, knowledge, and feelings. However, a voice signal can carry much more data than just uttered words. We can quickly determine a person's gender, age, and physical as well as emotional state by listening to their speech. There are several real-world uses for estimating a speaker's physical attributes from speech, including in the areas of surveillance [1], forensics [2, 3], commerce [4, 5], and human-robot interaction[6-8]. Voice has several benefits over other biometric procedures, including the fact that it is non-intrusive, affordable, simple to use, and well-accepted by users [9].

Large quantities of recordings that would normally be challenging to evaluate manually may be evaluated much more easily with the help of automated speaker profiling (ASP)[1]. It can also improve surveillance system data by giving a more thorough profile of a person, even when that person's picture is partially hidden or obscured [2]. It may be used in forensics to extract a suspect's description from phone conversations or other recordings [3, 5]. It may be utilized for phone routing, playing appropriate music or messaging, allocating customer assistance, and mobile purchasing in the business sector [5].

Numerous research has shown that voice and physical build are related. Because some vocal features, such as speech pace, might reveal a speaker's age [1], younger speakers tend to talk more quickly. Furthermore, the fundamental frequency tends to decline with age [2, 3], particularly for female speakers. The fundamental frequency (F0), which male speakers often have lower than female speakers, is crucial in determining gender.

Received 11/March/2023; Accepted 12/June/2023

ISSN (Print) 1811-9212 ISSN (Online) 2617-3352

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

Multiple acoustic aspects of a speaker's speech have been retrieved by numerous research, however, it is still unclear which acoustic elements are most appropriate for the various tasks of speaker profiling [10]. Furthermore, despite ongoing research, accurately estimating height, age, and gender with a minimal feature set using advanced machine learning techniques is still difficult [11]. This is because there are many sources of variability that overlap, including the speaker's gender, health, and emotional state, which can all affect speech as well as the design of the sound production system.

In this work, several feature selection methods are used to evaluate the importance of different features on the task of age estimation from speech, namely, mutual information, Correlation, Random Forests, Permutation, and Single Feature Performance (SFP). The features are then ranked based on the importance obtained with every feature selection method and voting is performed between the different feature selection methods to attain the final rank of features. Four different ensemble methods are applied to the acquired feature importance matrix to find the best features for the task. The ensemble methods used are average voting, majority voting, reciprocal voting, and Borda Counts. The resulting feature ranks are then used to estimate the age of speakers from their speech signals using Support Vector Regression (SVR). The proposed method is evaluated on the TIMIT dataset [12] achieving an MAE of 5.58 and 5.12 for female and male speakers respectively.

The remainder of this paper is organized as follows, the next section reviews related work on speaker age estimation and views different features and models, and feature selection methods used in the context of speaker age estimation.

Section III, presents the methodology followed in this work, the feature extraction methods, the proposed method, preprocessing methods, and the regression model used in this work. Section IV. Presents the experimental setup followed in this work, the dataset and evaluation metrics, and the results obtained from the proposed method are discussed. Finally, section V concludes the work.

II. RELATED WORK

In recent years, the study of paralinguistic content extraction has advanced quickly. In speaker profiling applications, important characteristics are frequently extracted from raw speech signals and used in prediction by a machine learning model, and this work focuses mainly on speaker age estimation from short utterances.

Several studies are conducted to estimate the age of speakers, one approach included adopting statistical approaches for speaker age estimation by capturing low-level representations from speech using short-term features like MFCC and Mel spectrogram as supervectors. Bahari [13] proposed a speaker’s age estimation approach when Hidden Markov Model (HMM) weight supervectors were modeled for each speaker. Then, the dimensionality of input space has been reduced by using Weighted Supervised Non-Negative Matrix Factorization (WSNMF). Finally, the speaker’s age has been estimated by using the Least Squares Support Vector Regressor (LS-SVR). Moreover, Fedorova. Et. Al. [14] employed i-vectors, along with a dimensionally reduced variant of supervectors (i-vectors), followed by Artificial Neural Networks (ANN) for age estimation. Similarly to this, Grzybowska et al. [15] investigate the application of i-vectors in conjunction with other auditory parameters for age estimation and age group detection. It was discovered

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

that employing i-vectors in combination with other acoustic features produced superior outcomes versus doing so alone.

Deep learning approaches are used for speaker representation and feature extraction in a distinct kind of speaker modeling. Sadjadi et al.[16] used i-vector modeling, a DNN-based alternative to the standard GMM method. This method generated i-vectors that are phonetically aware and fed them to an SVR model to estimate age. Zazo. et. al [17] proposed the use of Long short-term memory (LSTM) Recurrent Neural Networks (RNN), for age estimation using MFCC features and pitch information, producing an MAE of 6.97 and 7.79 for females and males respectively. A similar approach using LSTM-RNN for age and height estimation was employed in [18] using acoustic features extracted, resulting in a 6.08 and 5.62 MAE for female and male speakers respectively. Transfer learning with a DNN architecture called X-vectors has been also used in speaker profiling to estimate age and gender with encouraging results[19, 20]. A combination of convolutional neural networks and temporal neural networks are employed in [21] for age and gender classification using a short-time Fourier transform (STFT) and Mel-scale filterbanks as input achieving an error of 20% in age group prediction.

Kalluri et al. [10] employed various combinations of features, such as jitter, shimmer, and Harmonics-to-Noise Ratio (HNR), as well as Mel spectrograms, their first and second-order derivatives, formants, and fundamental frequency statistics. Their approach achieved an MAE of 5.2 years and 5.6 years for male and female speakers, respectively.

Badr. Et.al. [22, 23] used the accumulated statistic of MFCC and LPC with their first and second derivatives, Spectral Sub-band Coefficients (SSC), as well as the first 4 formants (f1-f4), and MAE of 10.3 and 9.25 years on the VoxCeleb dataset, and 7.73 and 4.96 for male and female age on the TIMIT dataset.

Finding the ideal feature set to represent various bodily features, nevertheless, is challenging. In the research on speaker profiling, several feature selection and dimensionality reduction techniques have been used for age feature selection including Catbthe oost optimization technique [23], Principle Component Analysis (PCA) [3], and Linear Discriminant Analysis (LDA) [22, 24, 25].

While these studies have demonstrated promising results, they often rely on specific feature sets or require extensive feature engineering, which can be computationally expensive. Additionally, many of the prior methods have not explored the potential benefits of using ensemble FS approaches or adequately addressed the challenges of finding the ideal feature set to represent various bodily features.

To address these limitations, the work reported in this paper builds on prior research to examine a broad range of spectral, prosodic, and temporal data. The main contribution of this study is the application of ensemble FS on these features to arrive at the optimal feature subset for the speaker's age estimation, specifically for male and female speakers. By doing so, the study aims to improve the age estimation accuracy and generalizability across different datasets and speaker groups and reduce the computational cost associated with feature extraction and selection.

Furthermore, this research highlights the importance of considering gender differences in age estimation, as previous studies have often reported varying Mean

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

Absolute Errors (MAEs) for male and female speakers. By focusing on gender-specific feature selection and optimization, this work strives to improve the robustness and reliability of age estimation models for both genders.

III. PROPOSED METHOD

In this work, the aim is to arrive at the best representative features for age estimation in male and female speakers. For this purpose, an ensemble feature selection method is proposed. The proposed method consists of three main stages, as can be seen in *Fig. 1*. After feature extraction and normalization, five different FS methods are used to evaluate the importance of each feature. The methods include filter methods such as Mutual Information (MI) and Correlation feature selection methods, wrapper methods such as Permutation and Single Feature Performance (SFC)), and an Embedded FS Random Forests feature (RF). An NxM feature importance matrix is obtained, where N is the number of features and M is the number of feature importance methods. For each feature importance method, the importance of each feature is obtained from FS methods. The feature importance matrix is normalized using the Min-Max normalization as given by (1):

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Where x is the original data to be transformed, x_{min} is the minimum value of x, x_{max} is the maximum value of x, and x_{scaled} is the new value rescaled to the range of (0-1) after transformation.

The features are then ranked based on their importance to arrive at an NxM matrix of ranked features. Subsequently, a voting algorithm is applied to the ranked data to find the final ranks of the features.

Four voting algorithms are employed, namely, Average voting, majority voting, reciprocal voting, and Borda count. The following subsections provide an overview of the different FS and voting methods employed.

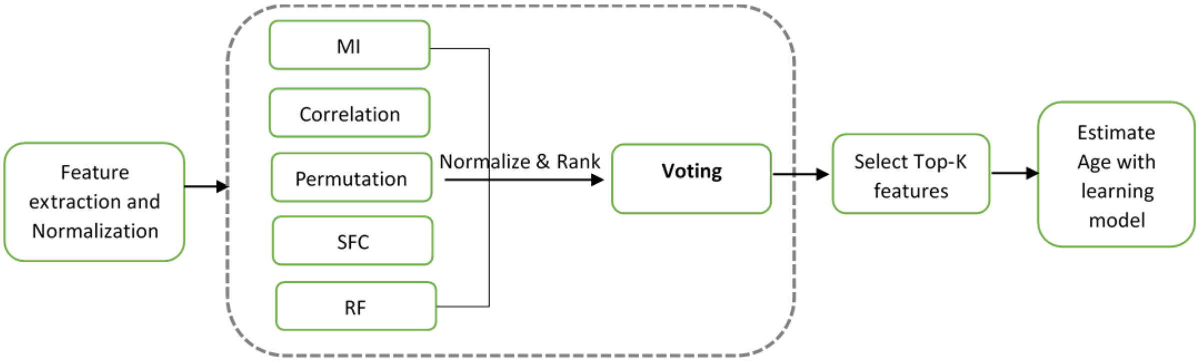


FIG. 1. PROPOSED FEATURE ENSEMBLE METHOD FOR SPEAKER AGE ESTIMATION.

The proposed methodology is divided into three stages. First, the data is collected from the TIMIT dataset. The required features are extracted from the dataset, and preprocessing is done to make the extracted features' data more meaningful and relevant for the estimation task. Next, the proposed method is applied to the extracted features to

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

identify the suboptimal feature set. Finally, the estimation of speaker age is achieved using SVR. The subsequent sections elaborate on these stages.

A. Feature Extraction

In this study, various features are extracted from speech using the openSMILE toolbox. The extracted features are intended to cover a broad range of characteristics, including spectral, prosodic, voice quality, and articulatory features. The first feature set used in the study is the Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which was chosen because of its range of acoustic and prosodic data, the consistency as a baseline assessment, and lack of variation caused by different parameter choices. The eGeMAPS feature set consists of 36 parameters derived from 18 Low-level descriptors (LLD) of frequency, energy/amplitude, and spectral characteristics. The LLDs include F0 (fundamental frequency), Jitter (variation in F0), formant frequencies, shimmer (variation in amplitude), loudness, and other spectral parameters.

To supplement the features provided by eGeMAPS, the study also uses openSmile to extract 255 additional features, including Mel Frequency Cepstral Coefficients and various funclo998jtionals of these coefficients. The study also uses Praat software to extract prosodic features, such as the number of pauses, phonation time, articulation rate, and speech rate.

The final feature set includes a total of 408 features, including amplitude-related features of the speech signal, such as various measures of jitter and shimmer. These features provide a comprehensive description of various elements of speech and music signals, including spectral, prosodic, voice quality, and articulatory characteristics.

B. Feature Selection Methods

Before commencing the ensemble FS process, feature importance is derived and the features are sorted based on the obtained importance. Five different FS methods are employed in this work:

Mutual Information (MI): A statistical measure that quantifies the dependence between two random variables. In the context of feature importance, MI can be used to assess the relationship between each feature in a dataset and the target variable. Features with high MI with the target are considered more informative and important in making predictions. However, MI can be biased towards features with a large number of unique values.

Correlation: A statistical measure that quantifies the linear relationship between two variables. Pearson's correlation is used in this work to assess the relationship between each feature and the target variable. The correlation coefficient ranges from -1 to 1, with a high positive correlation indicating a more important feature for prediction. Features with a high positive correlation with the target are considered to be more informative.

Single Feature Performance (SFP): In this method, the importance of each feature is evaluated individually for each task. The model is trained with each feature, and its importance is determined by the model's performance score. The higher the model performance on a specific feature, the greater the discriminative power of that feature. However, this method does not account for the effect of feature interaction with each other.

Permutation: This method measures feature importance based on the decrease in a model's performance score when the values of a single feature are randomly shuffled. Shuffling the order of the feature values alters the original relationship between the feature

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

and the target, so the drop in the model's performance score is indicative of how much the model depends on that feature.

Random Forests: A popular machine learning algorithm for both classification and regression problems. In the context of feature selection, random forests can be used to identify the most important features for making predictions. The algorithm uses a measure of feature importance based on the reduction in impurity achieved by splitting on a given feature, averaged over all trees in the forest. Features with a high average reduction in impurity are considered more important, as they lead to more homogeneous and informative leaf nodes in the trees. The feature importances can be computed for each tree in the forest and averaged over all trees to obtain a more stable estimate of feature importance.

C. Ensemble Methods

The Borda Count is a voting-based ensemble feature selection method where each feature is assigned a score based on its ranking by individual feature selectors. These selectors can be any feature selection method, such as wrapper, filter, or embedded methods. The Borda Count method combines results from multiple FS methods to produce a final ranking of features that is more robust and less sensitive to the choice of FS method. The ranking of a feature in Borda count is given by (2) as described in [26]:

$$r(f) = \sum r_i(f) \tag{2}$$

Where i = 1, 2, …, N are the FS methods, and ri(f) is the rank of the feature f according to the ith method.

The Reciprocal Rank (RR) is a measure of the performance of a feature selection algorithm that can be used to combine the results of multiple feature selectors. It is defined as the reciprocal of the rank of the first relevant feature in the ranked list of features produced by the algorithm. According to [26], the Reciprocal Rank is based on the calculation of the final rank r(f) of a feature f using (3):

$$r(f) = \frac{1}{\sum \frac{1}{r_i(f)}} \tag{3}$$

Where i = 1, 2, …, N are the FS methods, and $r_{i(f)}$ is the rank of the feature f according to the ith method. In ensemble feature selection, the RR is calculated for each feature selector, and the scores are combined to obtain the final ranking of features. The feature with the highest average RR across all feature selectors is considered to be the most relevant.

Majority Voting is a method used to aggregate the scores or rankings of different feature selection algorithms. After obtaining feature ranks, it selects features that have received the most votes, or that have been ranked the highest by the majority of the algorithms. The advantage of using majority voting is that it can help overcome the limitations of individual feature selection algorithms, as well as increase the stability and reliability of the final feature selection results.

Average Voting is a method used to aggregate the scores or rankings of different feature selection algorithms. Multiple feature selection algorithms rank the features and calculate the average score of each feature across all algorithms. The features with the

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

highest average scores are then selected as the most relevant or important features. Average Voting is given as (4):

$$r(f) = \frac{\sum r_i(f)}{M} \tag{4}$$

D. SVR

The goal of SVR is to create a model that can predict continuous values based on a set of input variables. It does this by finding the hyperplane in a high-dimensional space that has the maximum distance to the nearest data points. This hyperplane is called the "support vector," and the data points closest to it are called "support vectors."

The SVR algorithm can handle non-linear data by using a kernel function to transform the input variables into a higher dimensional space, where the data becomes separable by a hyperplane. Given a set of training data [(x1,y1),(x2,y2) … (xn, yn)] where a is the d-dimensional feature vector and yi is the corresponding target, to predict a target output, we model the relationship between that target and the features as given in [27] by using (5):

$$f(x) = w^T x + b \tag{5}$$

Where w is the weights vector, and b is the bias term. The objective is to minimize (6):

$$MIN \ \frac{1}{2} \ ||w||^2 \tag{6}$$

With the constraints of $|w^T x_i + b - y_i| < \varepsilon$, where ε is the margin of accepted error (i.e. deviation from the target output).

IV. EXPERIMENTAL RESULTS

In this section, we perform experiments to estimate an unknown speaker's age from speech using ensemble feature selection. As such we evaluate the effect of different FS methods and ensemble FS on age estimation accuracy. For Evaluation, MAE is used to measure the error in age estimation as given in (7):

$$MAE = \frac{\sum |y_i - x_i|}{n} \tag{7}$$

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

The remaining section describes the dataset used in the experiments and discusses the results obtained using individual FS and ensemble feature selection.

A. Dataset

This work uses the TIMIT dataset for multitask speaker profiling. TIMIT is an automatic speech recognition (ASR) dataset that contains metadata about the participants such as height, age, education level, ethnicity, and regional dialect. Each participant contributed 10 recordings of speech transcripts, resulting in 6300 utterances that are divided into non-overlapping Train and Test sets. The Train set consists of 326 male speakers and 136 female speakers resulting in a total of 462 speakers, and the Test set

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

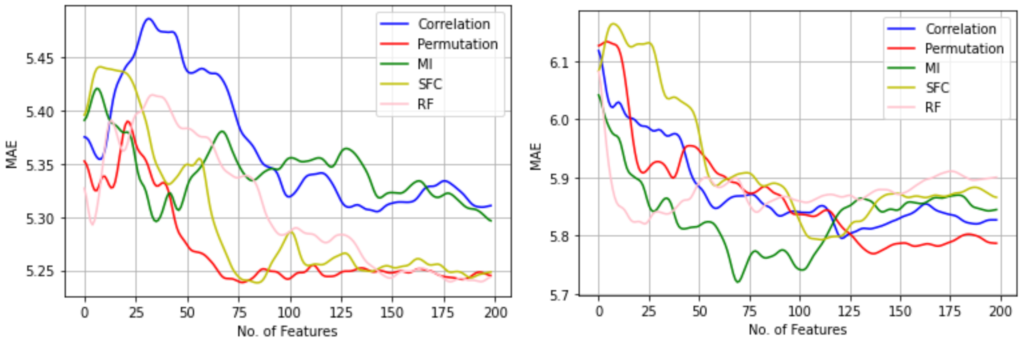
consists of 168 speakers, with 112 male and 56 female speakers. The standard train and test split of TIMIT is used in this work. Table I shows the statistics of the dataset.

TABLE I. STATISTICS OF AGE IN TIMIT DATASET

	Male Speakers	Female Speakers
No. of Speakers	438	192
Minimum Age	20	21
Maximum Age	75	67
Mean	30.5	30
Standard Deviation	7.6	8.7

B. FS Performance Analysis

This study compares the mentioned FS methods to assess their performance. Table II shows the minimum MAE obtained with each method for male and female ages. As can be seen in *Fig. 2* (A) and (B), all methods improve their performance as the number of features increases until it reaches a point where the performance degrades gradually. For female age estimation, MI FS shows the best performance, while permutation consistently outperforms other methods, both methods show the best performance around 75 features.



(A)

(B)

FIG. 2. AGE MAE OF FIVE FEATURE SELECTION METHODS WITH VARYING NUMBERS OF FEATURES FOR (A) MALE, AND (B) FEMALE.

These differences in performance between male and female age estimation using different methods of FS suggests that different feature sets should be selected for male and female tasks, and any speaker profiling application should be preceded by gender detection.

TABLE II. MINIMUM MAE IN AGE ESTIMATION USING DIFFERENT FEATURE SELECTION METHODS

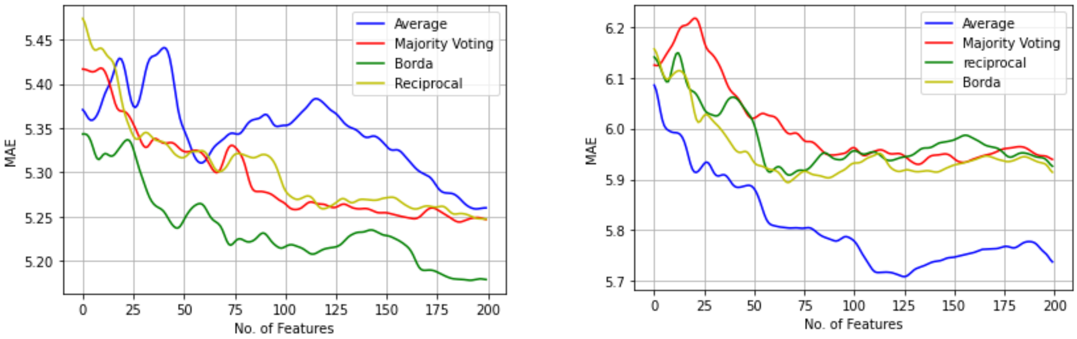
Method		AGE MAE	
	Female	Male	All
Correlation	5.78	5.3	5.47
Permutation	5.76	5.23	5.42
MI	5.7	5.28	5.46
SFC	5.78	5.23	5.43
RF	5.8	5.23	5.45

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

C. Ensemble FS Performance

In this section, several ensemble methods are compared to find the minimal feature set for the task that achieves the best performance. *Fig. 3* (A) and (B) show obtained results using majority voting, average voting, reciprocal voting, and Borda count. The comparison results show that average voting performs best for female age prediction while Borda counts significantly outperform other methods in male age estimation. These two methods achieved the best estimation performance with a small number of features as shown in Table III. On the other hand, majority voting and reciprocal voting perform comparably on both sets. The results also confirm that using ensemble methods for FS outperforms the results obtained with individual FS algorithms. However, the choice of the ensemble method can greatly influence the outcome of the results.



(A) (B)

FIG. 3. AGE MAE OF FOUR ENSEMBLE METHODS FOR FEATURE SELECTION WITH VARYING NUMBERS OF FEATURES FOR (A) MALE, AND (B) FEMALE.

TABLE III. MINIMUM MAE IN AGE ESTIMATION USING DIFFERENT ENSEMBLE METHODS

Method		**AGE MAE**	
	Female	**Male**	**All**
Average	5.58	5.2	5.4
Majority	5.92	5.24	5.47
Borda	5.85	5.12	5.42
Reciprocal	5.88	5.24	5.47

In order to demonstrate the effectiveness of the proposed ensemble feature selection methods for age estimation, we compare our results with previous studies in Table IV. The comparison highlights the advantages of our approach in terms of age estimation accuracy, particularly when considering gender-specific performance.

TABLE IV. COMPARISON OF AGE ESTIMATION MAE WITH PREVIOUS STUDIES

Study	**AGE MAE**	
	Female	**Male**
Badr Et. al.[22]	**4.96**	7.73
Kaushik Et. al [18]	6.08	5.62
Kalluri Et. al.[10]	5.6	5.2
Kwasny. Et. al [19]	5.29	**5.12**
Proposed Method	5.58	**5.12**

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

As shown in Table IV, our proposed ensemble feature selection methods achieve competitive performance compared to prior studies. In particular, our approach outperforms several previous studies in terms of gender-specific MAE, illustrating the importance of considering gender differences when selecting features for age estimation. Moreover, the overall MAE achieved by our method is lower than most of the prior studies, indicating that the ensemble FS approach is effective in improving the accuracy of speaker age estimation.

The ensemble feature selection approach proposed in this work demonstrates promising results for age estimation, effectively addressing some of the limitations of previous studies. The improved performance and the focus on gender-specific optimization suggest that the proposed methods can be valuable for various speaker profiling applications.

V. CONCLUSIONS

In this work, an automatic text-independent system is proposed for estimating age from speech utterances. Four feature selection (FS) algorithms are employed to select the most representative features from a 400-feature vector. Ensemble feature selection is performed using these algorithms to find features nominated by all FS methods. Four types of ensemble voting methods are evaluated in this study: average voting, majority voting, reciprocal voting, and Borda count. Support Vector Regression (SVR) is utilized to accurately predict the speaker's age. Experimental results demonstrate the effectiveness of the ensemble methods over individual FS methods, as evidenced by the performance on the TIMIT dataset with Mean Absolute Errors (MAE) of 5.58 and 5.12 for female and male speakers, respectively. The results also show that average voting outperforms other ensemble methods, indicating its potential as a robust approach for age estimation in speaker profiling applications.

Future work in this area could explore the applicability of the proposed methods to other speaker profiling tasks, such as height estimation, emotion recognition, or speaker verification, to assess the generalizability of the ensemble feature selection approach. Moreover, evaluate the performance of the proposed methods on larger and more diverse datasets, including speakers from various language backgrounds, dialects, and age ranges, to ensure the robustness of the age estimation system in real-world scenarios.

REFERENCES

[1] I. Mporas and T. Ganchev, "Estimation of unknown speaker’s height from speech," International Journal of Speech Technology, vol. 12, no. 4, pp. 149-160, 2010, doi: 10.1007/s10772-010-9064-2.

[2] A. H. Poorjam and M. H. Bahari, "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), 2014: IEEE, pp. 7-12.

[3] A. Beke, "Forensic speaker profiling in a Hungarian speech corpus," in 2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 2018: IEEE, pp. 000379-000384.

[4] C. Müller, "Automatic recognition of speakers' age and gender on the basis of empirical studies," in Ninth International Conference on Spoken Language Processing, 2006: Interspeech, pp. 2118–2121

[5] C. Müller and F. Burkhardt, "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, 2007: INTERSPEECH pp. 2277–2280.

[6] H.-J. Kim, K. Bae, and H.-S. Yoon, "Age and gender classification for a home-robot service," in RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication, 2007: IEEE, pp. 122-126.

Received 11/March/2023; Accepted 12/June/2023

DOI: https://doi.org/10.33103/uot.ijccce.23.4.2

[7] M.-W. Lee and K.-C. Kwak, "Performance comparison of gender and age group recognition for human-robot interaction," IJACSA) International Journal of Advanced Computer Science and Applications, vol. 3, no. 12, 2012.

[8] A. Badr and A. Abdul-Hassan, "A Review on Voice-based Interface for Human-Robot Interaction," Iraqi Journal for Electrical and Electronic Engineering, vol. 16, no. 2, pp. 1-12, 2020, doi: 10.37917/ijeee.16.2.10.

[9] J. H. Hansen, K. Williams, and H. Boril, "Speaker height estimation from speech: Fusing spectral regression and statistical acoustic models," J Acoust Soc Am, vol. 138, no. 2, pp. 1052-67, Aug 2015, doi: 10.1121/1.4927554.

[10] S. B. Kalluri, D. Vijayasenan, and S. Ganapathy, "Automatic speaker profiling from short duration speech data," Speech Communication, vol. 121, pp. 16-28, 2020, doi: 10.1016/j.specom.2020.03.008.

[11] A. A. Badr and A. K. Abdul-Hassan, "Age Estimation in Short Speech Utterances Based on Bidirectional Gated-Recurrent Neural Networks," Engineering and Technology Journal, vol. 39, no. 1B, pp. 129-140, 2021, doi: 10.30684/etj.v39i1B.1905.

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," NASA STI/Recon technical report n, vol. 93, p. 27403, 1993.

[13] M. H. Bahari, "Speaker age estimation using Hidden Markov Model weight supervectors," in 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012: IEEE, pp. 517-521.

[14] A. Fedorova, O. Glembek, T. Kinnunen, and P. Matějka, "Exploring ANN back-ends for i-vector based speaker age estimation," in Sixteenth Annual Conference of the International Speech Communication Association, 2015, pp. 3036–3040.

[15] J. Grzybowska and S. Kacprzak, "Speaker Age Classification and Regression Using i-Vectors," presented at the Interspeech 2016, 2016.

[16] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "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), 2016: IEEE, pp. 5040-5044.

[17] R. Zazo, P. Sankar Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, "Age Estimation in Short Speech Utterances Based on LSTM Recurrent Neural Networks," IEEE Access, vol. 6, pp. 22524-22530, 2018, doi: 10.1109/access.2018.2816163.

[18] M. Kaushik, V. T. Pham, and E. S. Chng, "End-to-end speaker height and age estimation using attention mechanism with LSTM-RNN," arXiv preprint arXiv:2101.05056, 2021.

[19] D. Kwasny and D. Hemmerling, "Gender and Age Estimation Methods Based on Speech Using Deep Neural Networks," Sensors (Basel), vol. 21, no. 14, Jul 13 2021, doi: 10.3390/s21144785.

[20] D. Kwasny and D. Hemmerling, "Joint gender and age estimation based on speech signals using x-vectors and transfer learning," arXiv preprint arXiv:2012.01551, 2020.

[21] H. A. Sánchez-Hevia, R. Gil-Pita, M. Utrilla-Manso, and M. Rosa-Zurera, "Age group classification and gender recognition from speech with temporal convolutional neural networks," Multimedia Tools and Applications, vol. 81, no. 3, pp. 3535-3552, 2022.

[22] A. A. Badr and A. K. Abdul-Hassan, "Estimating Age in Short Utterances Based on Multi-Class Classification Approach," Computers, Materials & Continua, vol. 68, no. 2, pp. 1713-1729, 2021, doi: 10.32604/cmc.2021.016732.

[23] A. Badr and A. Abdul-Hassan, "CatBoost Machine Learning Based Feature Selection for Age and Gender Recognition in Short Speech Utterances," International Journal of Intelligent Engineering and Systems, vol. 14, no. 3, pp. 150-159, 2021, doi: 10.22266/ijies2021.0630.14.

[24] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," Expert Systems with Applications, vol. 136, pp. 252-263, 2019.

[25] P. Ghahremani et al., "End-to-end Deep Neural Network Age Estimation," presented at the Interspeech 2018, 2018.

[26] D. Effrosynidis and A. Arampatzis, "An evaluation of feature selection methods for environmental data," Ecological Informatics, vol. 61, p. 101224, 2021.

[27] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189-215, 2020.

Received 11/March/2023; Accepted 12/June/2023