Received: January 29, 2021. Revised: February 18, 2021.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

CatBoost Machine Learning Based Feature Selection for Age and Gender Recognition in Short Speech Utterances

Ameer A. Badr[1,2]* **Alia K. Abdul-Hassan**[2]

[1]*Collage of Managerial and Financial Sciences, Imam Ja'afar Al-Sadiq University, Salahaddin, Iraq*
[2]*Department of Computer Science, University of Technology, Baghdad, Iraq*
* Corresponding author's Email: cs.19.53@grad.uotechnology.edu.iq

Abstract: Lately, with the rapid growth of various technologies, identifying the information of gender and age give short speech utterances has become a necessity for many applications in daily life like human-robot interaction, targeted marketing, identifying suspects in criminal cases, etc. Despite the comprehensive studies carried out to extract descriptive features, the recognition accuracy is still not satisfactory. In this study, an automatic system is proposed to classify age and gender in short speech utterances without depending on the text. Firstly, two groups of features are extracted from each utterance frame, followed by measuring 10 statistical functionals for each extracted feature dimension. After that, the extracted features dimensions are normalized using the Quantile technique. Then, the CatBoost machine is utilized as an important features detection to select the most discriminatory features for speaker age and gender recognition tasks. For classification purposes, the selected feature dimensions are fed into the Support Vector Machine (SVM). Experiments are conducted on the aGender data-set for measuring the suggested system's performance. The unweighted accuracies (UA) of the proposed system for gender, age, in addition to gender & age is 89.62%, 72.29%, and 71.96%, respectively. The achieved results outperform recent results on the same data-set.

Keywords: Speaker age and gender recognition, Statistical functionals, Quantile technique, CatBoost machine, aGender data-set.

1. Introduction

Individuals use nonverbal information sources (i.e., paralinguistic), not just verbal information sources (i.e., speech) in their daily voice-based communications. The former might include the characteristics of the speaker like identity, gender, emotional state, age, along with the existence of disease conditions. Nonverbal information can be used to render specific services on the internet without directly contacting the users [1]. In addition, an automatic system concerning the recognition of gender and age in short speech utterances was vital for the purposes of forensic medicine, for instance, to narrow down the list of suspects following committing crimes in the case when samples of speech are available. Besides, the system might be utilized for increasing the efficiency of targeted advertising and in health-care institutions [2]. Because of the wide-spreading as well as prevention attempts of coronavirus pandemic (COVID-19), such a system is particularly suitable for the last application by allowing the identification of seniors at risk. Concerning speech processing, the recognition of gender and age given short speech utterances was a difficult task. Challenges in this field include text-independent or text-dependent system design, accent variances resulting from various speakers, and background noise [3, 4].

In such a system, there are three major phases: feature extraction, feature selection, and classification. In terms of the stage of feature extraction, distinctive features are calculated from speech signals that uniquely represent the speaker's gender and age information. Generally, the feature sets might be evaluated via utilizing prosodic, spectral, or glottal properties related to the speech utterances [3]. The most used features in such a system include the Mel-frequency cepstral coefficients (MFCCs) [1–5], formants frequency [5], fundamental frequency (F0) [4, 5], jitter and shimmer [5], and i-vectors [6]. The feature selection stage aims at decreasing the number of feature dimensions to reduce the time as well as the storage space needed for experiments while enhancing the accuracy of recognition. Recently, ensemble methods are used to feature important detection [7]. Among them, the CatBoost machine is considered one of the most powerful machines [8]; it is used for feature selection in [7, 9, 10]. The stage of classification aims at designing a classifier that utilizes the chosen feature set to recognize the speaker's gender and age information. The most used classifiers in such system include deep neural networks (DNNs) [1, 2, 11], gaussian mixture model (GMM) [5, 11], Gaussian mixture model universal background model (GMM-UBM) [3], and Support Vector Machine (SVM) [4, 5, 12].

Schuller et al. (2010) [13] presented a paralinguistic three sub-challenges utilizing the aGender data-set [14]. In the gender sub-challenge, the speaker gender has to be determined in 3 groups: female, male, and child. In the age sub-challenge, the speaker's age has to be determined in 4 groups: senior, adult, youth, and, child. However, in the last sub-challenge, which is considered as a combination regarding such 2 sub-challenges, the speakers will be divided into 7 groups as senior female, adult female, youth female, senior male, adult male, youth male, and child.

As the distribution among classes is not balanced, Unweighted Accuracies (UA) is the baseline performance measure. In addition, the baseline study [13] has been used SVM with the linear kernel; their best performance results for gender, age, in addition to gender & age recognition reached (80.42, 48.91, and 44.94) %. The best result among competitors in the paralinguistic sub-challenges was achieved by Kockmann et al. (2010) [12]. Their system used both prosodic and acoustic features, SVM and GMM classifiers. Their best performance results for gender, age, in addition to gender & age recognition reached (81.82, 52.88, and 53.86) %. After the paralinguistic sub-challenges [13], there was a growing interest in research in this field. Li et al. (2013) [15] combined seven different sub-systems to improve the baseline performance. Their system used prosodic and acoustic features, as well as SVM and GMM classifiers. They have demonstrated weighted summation-based fusion related to such 7 sub-systems at score level. Their best performance results for gender, age, in addition to gender & age recognition reached (81.70, 52.80, and 50.30) %. Barkana et al. (2015) [4] presented a pitch range-based feature set for gender and age recognition. Their system used k-nearest neighbors (KNN) and SVM classifiers for evaluating the efficiency of the suggested feature sets. Their best performance results for gender, age, in addition to gender& age reached (84.70, 66.20, and 63.70) %. Yücesoy et al. (2016) [5] presented an approach that includes score level fusion regarding 7 sub-systems. Their fused system used prosodic and spectral features, also SVM, GMM, and GMM supervector based SVM classifiers. Their best performance results for gender, age, in addition to gender & age reached (90.40, 54.10, and 53.50) %. Grzybowska et al. (2016) [6] examined using i-vectors for gender and age recognition. They used the cosine distance scoring for classification. Their best performance results reached 62.90% for age & gender recognition. Qawaqneh et al. (2017) [11] presented a system that applies a bottle neck feature (BNF) extractor in addition to DNN for gender and age recognition. Moreover, the BNF extractor was utilized for generating transformed MFCCs features, while DNN has been used to classify the transformed features. Their best results reached a performance of 58.98% for age & gender recognition. Abu-Mallouh et al. (2017) [3] proposed a model for generating bottleneck features from DNN in addition to GMM–UBM classifier in terms of speaker's gender and age recognition. The DNN with a bottleneck layer was trained in an unsupervised way to calculate the initial weights between layers. Then, it was tuned and trained in a supervised way for generating transformed MFCC features. GMM–UBM was utilized for building the model of GMM for each one of the classes, and the models were also utilized for classifying the speaker's gender and age. Their best performance results reached 57.63% for age & gender recognition. Markitantov et al. (2019) [1] presented different DNN topologies based on convolutional and fully-connected layers for the speaker's gender and age recognition. Their system uses MFCC and Mel-spectrogram (MEL) features, as well as fully-connected DNN and convolution neural networks (CNN) classifiers. Their best performance results for gender, age, in addition to gender & age recognition reached (88.80, 57.53, and 48.41) %. Markitantov (2020) [2] presented a transfer learning method with regard to gender and age recognition. They modified the pre-trained models, including VGG-16, AlexNet, ResNet34, ResNet18, ResNet50, along with recent EfficientNet-B4 from Google, also developed time-delay neural networks (TDNN) and 1D-CNN models for speaker's age and gender recognition. Their best performance results for gender, age, in addition to gender & age recognition reached (81.74, 51.71, and 48.96) %.

Although many feature groups were investigated to recognize the speaker's gender and age, researchers have not yet identified the best feature groups for this task. This is because most previous studies combine multiple feature groups to enhance the performance; however, recognition errors from these different feature groups may not be complementary. The present study aims at filling this research gap by exploring the strength of the CatBoost machine in features important detection. The main contributions of this study can be summarized in the following points:

- Combining two feature groups, which are MFCCs, Spectral Subband Centroids (SSCs). Then, measuring 10 statistical functionals for each extracted feature dimension to achieve the greatest possible gain from each feature vector.
- Exploring the strength of using the CatBoost machine as a supervised features selection approach.

The rest of this study is categorized in the following way: Section 2 provides the proposed methodology. The results of simulations and experiments are shown in Section 3. Finally, Section 4 sets out the study conclusions and ideas for future works.

2. The proposed methodology

As shown in Fig. 1, the methodology of this study consists of five main stages: features extracting, statistical functionals measuring, features standardizing, features selecting, and speaker age and gender recognizing. Initially, appropriate features are extracted from each speaker's utterance, followed by features scaling to fall within a smaller range using the standardization technique. Then, by using the features selection method, the high dimensional features will be transformed into more discriminative low dimensional features. Finally, the SVM classifier is used to recognize the speaker's age and gender.

2.1 Utterance based features extraction

Among all types of speech-based feature extraction domains, Cepstral domain features are the most successful ones in recognition of speaker's gender and age tasks, where a Cepstrum was acquired via taking an inverse Fourier transform related to signal spectrum. MFCC is the most important method to extract speech-based features in this domain [16, 17]. MFCCs magnificent role stems from the capability for exemplifying the speech amplitude's in a concise form. A speaker's voice is filtered by the articulator form of the vocal tract, such as the nasal cavity, teeth, and tongue. This shape affects the vibrational characteristics of the voice. If the shape is precisely controlled, this should give an accurate depiction of the phoneme being formed [4, 17]. The procedure for obtaining the MFCC features is explained in the following steps [18]:

1. Preemphasis: This relates to filtering that stresses the high frequencies. Therefore, a few glottal impacts were eliminated from the vocal tract parameters by preemphasis.
2. Frame blocking and windowing: The speech must be studied for a short-period (frames) for stable acoustic characteristics. A window was used on each one of the frames for tapering the signal towards the frame's limits. Hamming windows are usually used.
3. FFT spectrum: By applying the Fast Fourier Transform (FFT), each one of the windowed frames were converted into a magnitude spectrum.
4. Mel spectrum: it is evaluated via passing the FFT signal via a set of band pass filters referred to as Mel-filter bank. In addition, Mel is considered as a measurement unit based on the perceived frequency of the human ears. Approximately, the

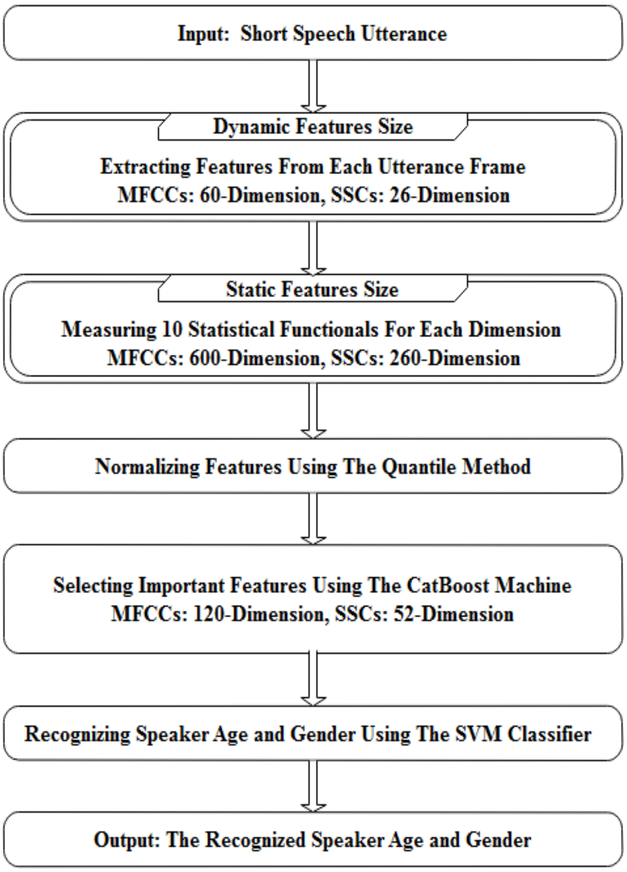
Input: Short Speech Utterance
Dynamic Features Size
Extracting Features From Each Utterance Frame
MFCCs: 60-Dimension, SSCs: 26-Dimension
Static Features Size
Measuring 10 Statistical Functionals For Each Dimension
MFCCs: 600-Dimension, SSCs: 260-Dimension
Normalizing Features Using The Quantile Method
Selecting Important Features Using The CatBoost Machine
MFCCs: 120-Dimension, SSCs: 52-Dimension
Recognizing Speaker Age and Gender Using The SVM Classifier
Output: The Recognized Speaker Age and Gender

Figure. 1 The general framework of the proposed system

Mel scale is less than 1 kHz linear frequency spacing and above 1 kHz logarithmic spacing. There is a possibility for expressing the approximation related to Mel from physical frequency, like in Eq. (1). In addition, the warped axis, on the basis of non-linear function provided in Eq. (1), was conducted for mimicking the human ears' perception. The most commonly used filter shape is triangular. When the spectrum magnitude is multiplied by each one of the triangular Mel weighting filters, the Mel related to magnitude spectrum X (k) is evaluated in Eq. (2).

$$f_{Mel} = 2595 \log_{10}(1 + \tfrac{f}{700}) \qquad (1)$$

$$s(m) = \sum_{k=0}^{N-1} [|X(k)|^2 \, H_m(k)]; \quad 0 \le m \le M - 1 \qquad (2)$$

Where f_{Mel} is the perceived frequency in Hz, while f represents the physical frequency. M represents the total number of triangular Mel weighting filters, $H_m(k)$ represents the k^{th} energy spectrum bin weight which contributes to m^{th} output band.

5. Discrete cosine transforms (DCT): The levels of energy in adjacent bands have a tendency for being correlated due to the fact that the vocal tract was smooth. A set of cepstral coefficients is produced by the DCT utilized to the transformed Mel frequency coefficients. Lastly, MFCC is calculated as expressed in Eq. (3).

$$c(n) = \sum_{m=0}^{M-1} log_{10}\left(s(m)\right) \cos\left(\frac{\pi n(m-0.5)}{M}\right); \quad n = 0,1,2,\ldots,C-1 \qquad (3)$$

In which *c(n)* represents the cepstral coefficients, while *C* represents the number of MFCCs.

6. Dynamic MFCC features: The additional information on the signal's time dynamics was acquired via evaluating the 1st and 2nd derivatives related to cepstral coefficients since they contain only information from the given frame. Eq. (4) shows the majorly utilized definition for dynamic parameter computation.

$$\Delta c_m(n) = \frac{\sum_{i=-T}^{T} k_i c_m(n+i)}{\sum_{i=-T}^{T} |i|} \qquad (4)$$

Where, $c_m(n)$ is the m^{th} feature for n^{th} time frame, k_i represents i^{th} weight and T represents the number of successive frames utilized for computations.

SSCs features intend to be a complement to the MFCCs features in speech recognition. The addition of noise to the speech signals affects the spectrum of speech power at all frequencies, but in the higher amplitude (i.e., formant) of the spectrum, the effect is less noticeable. Therefore, to ensure the robustness of the feature, some formant-like features have to be investigated; SSC features are similar to the formant frequencies and can be easily and reliably extracted [19]. The entire frequency band (0 to Fs/2) is divided into N number of sub-bands for computation of SSCs, where Fs is the speech signal sampling frequency. SSCs were found via using the filter banks to signal power spectrum and, after that, evaluating the first moment (centroid) regarding each one of the sub-bands. SSC of mth sub-band is calculated as seen in Eq. (5), where Fs represents the sampling frequency, P(f) represents the short time power spectrum, ωm(f) represents the frequency response related to mth bandpass filter, and γ represents the parameter which controls the dynamic range regarding power spectrum [20].

$$C_m = \frac{\int_0^{Fs/2} f \, \omega m(f) P\gamma(f) df}{\int_0^{Fs/2} f \, \omega m(f) P\gamma(f) df} \qquad (5)$$

Since SSCs features are proposed to be complementary to the MFCCs features, these two groups of features were incorporated in this study in which the recognition errors from these feature groups are complementary. In the beginning, each speaker's utterance is split into frames with a window size of 25 milliseconds and a frameshift of 10 milliseconds to ensure that each frame contains robust information. Then, two groups of features are extracted from each utterance frame, namely MFCCs (i.e., 20-dimensions with its first and second derivative), and SSCs (i.e., 26- dimensions).

2.2 Statistical features generation

To override the issue of varying features size between different speaker utterances as well as to achieve the greatest possible gain from each feature dimension, the features with dynamic size extracted from the previous stage are turned into features with static size by measuring 10 statistical functionals for each dimension. These statistical functionals include mean, min, max, median, stander deviation, skewness, kurtosis, first quantile, third quantile, and interquartile range (Iqr). The total output of features dimension in this stage is 600 for MFCC and 260 for SSC.

2.3 Features normalization using quantile technique

Quantile normalization is a global method of adjustment that assumes that each sample's statistical distribution is the same [21]. This method is supported by the idea that in the case where all N data vectors have the same distribution, the quantiles are plotted in N-dimensions and provide a straight line along the unit vector line. This indicates that if one projects the points of our N-dimensional quantile plot onto a diagonal, one might have the ability if creating a data-set with the same distribution. This implies that the same distribution can be given to each array via taking the mean quantile as well as replacing it with the data item value in the original data-set [22]. This motivates the following steps by giving them the same distribution to normalize a set of data vectors [21]:

1. given n arrays of length p, form X of dimension (p × n) in which each one of the arrays was a column;
2. sorting each one of the columns of X for giving X_{sort};
3. taking the means across rows of X_{sort} and assigning the mean for each one of the elements in the row for getting $\tilde{X}_{sort}$;
4. get $X_{\text{normalized}}$ via rearranging each one of the columns $\tilde{X}_{sort}$ for having the same ordering as the original X.

The expression of features in smaller units will result in a wider range of these features and thus will tend to give such features a greater effect. The normalization process involves transforming the data to fall within a smaller range. Therefore, due to the great usefulness of the normalization process in machine learning methods, the features extracted from the previous stage will be normalized using the quantile method.

2.4 Important features selection using catBoost machine

Gradient boosting can be defined as one of the significant approaches to machine learning, achieving excellent results in various practical tasks. Essentially, gradient boosting is one of the processes used to construct an ensemble predictor through conducting a gradient descent in the functional space [23]. Within the framework related to gradient boosting decision tree (GBDT), three main implements have been developed: Light Gradient Boosting Machine (LightGBM), eXtreme gradient Boosting (XGBoost), and CatBoost. LightGBM efficiently enhances the calculation efficiency, while XGBoost achieves massive parallelism. Yet, LightGBM and XGBoost have an inherent problem of prediction shift [24]. In addition, CatBoost is considered more significant compared to LightGBM and XGBoost in terms of accuracy as well as the generalization ability via solving such a problem. Particularly, CatBoost ensembles the symmetric decision trees with symmetry structures endowing its few parameters, faster testing and training, as well as high accuracy. CatBoost replaces the gradient estimation approach of the conventional gradient boosting algorithm with ordered boosting, thus decreasing the gradient estimation bias and enhancing the capability of generalization [24].

Assuming a data-set $D = \{(x_k, y_k)\}_{k=1}^{n}$, in which x_k represents a d-dimensional feature vector, while $y_k \in R$ represents the corresponding label. In addition, the symmetric decision trees have been designed when the whole feature space is recursively partitioned. Considering that the feature space R^d related to CatBoost was divided into J disjoint regions (tree nodes), each one of the regions (tree leaf) has a corresponding value b_j, which is an estimated value related to the predicted class label. The decision tree h might be written as a superposition regarding the estimated values of all regions [24]:

$$h(x) = \sum_{j=1}^{J} b_j \, I \tag{6}$$

where I is an indicator function [24]:

$$I_{\{x \in R_j\}} = \begin{cases} 1 & if\ x \in R_j \\ 0 & otherwise \end{cases} \tag{7}$$

In the process of gradient boosting, a series of approximate functions $F^t: R^d \longrightarrow R$ were created for minimizing the expected loss $\mathcal{L}(F): EL(y, F(x))$ in a greedy manner [24]:

$$F^t = F^{t-1} + \alpha h^t \tag{8}$$

In which α represents the size of the step and h^t represents a tree that was chosen from the functions for minimizing $\mathcal{L}(F)$ in t^{th} iteration [24]:

$$h^t = argmin\, EL(y, F^{t-1}(x) + h(x)) \tag{9}$$

The least-squares function uses the loss function for the most part. The step of negative gradient was utilized for solving the minimization problem. Thus, Eq. (9) was transformed in the following [24]:

$$h^t = argmin\, E(-g^t(x,y) - h(x))^2 \qquad (10)$$

In which, $g^t(x,y) := \frac{\partial L(y, F^{t-1}(x))}{\partial F^{t-1}(x)}$, following N iterations, one might obtain a series of approximate functions $F^t(t = 0, 1, \dots)$ and summing them for getting the final model [24]:

$$F(x) = \sum_{t=1}^{N} h^t \qquad (11)$$

It must be indicated that the aim of the first tree h^1 was y, whereas the latter trees h^t regard the residuals r^t of targets y and the estimated results $h^{t-1}(x)$ regarding the previous model as their goals. Thus, the classification model F^t at each step of boosting depends on the target values related to all training samples that were utilized for building the previous model F^{t-1}. It indicates the target leakage, leading to a prediction shift of the learned model. In addition, the prediction shift is going to impact the model's generalization capability and the performance related to the driving style classification. Based on the ordered boosting principle, CatBoost enhances the standard gradient boosting process and implement no bias boosting [24]. Furthermore, the principle related to ordered boosting was as in Algorithm 1 [23]. In such a way, no target y_k was revealed in boosting's previous steps; also, CatBoost might be achieving high generalization performance.

In addition to the high ability of generalization, the CatBoost superiority over LightGBM and XGBoost motivates this study to choose the CatBoost machine as a supervised features selection technique.

Algorithm (1): The Ordered Boosting [23].

```
Input: ${\{(x_k, y_k)\}}_{k=1}^{n}, I$ ;
Step 1: $\sigma \leftarrow random\ permutation\ of\ [1, n]$ ;
Step 2: $M_i \leftarrow 0\ for\ i = 1 \dots n$ ;
Step 3: for $t \leftarrow 1$ to $I$ do
    for $i \leftarrow 1$ to $n$ do
        $r_i \leftarrow y_i - M_{\sigma(i)-1}(x_i)$ ;
    for $i \leftarrow 1$ to $n$ do
        $\Delta M \leftarrow LearnModel((x_j, r_j): \sigma(j) \le i)$ ;
        $M_i \leftarrow M_i + \Delta M$ ;
  return $M_n$
```

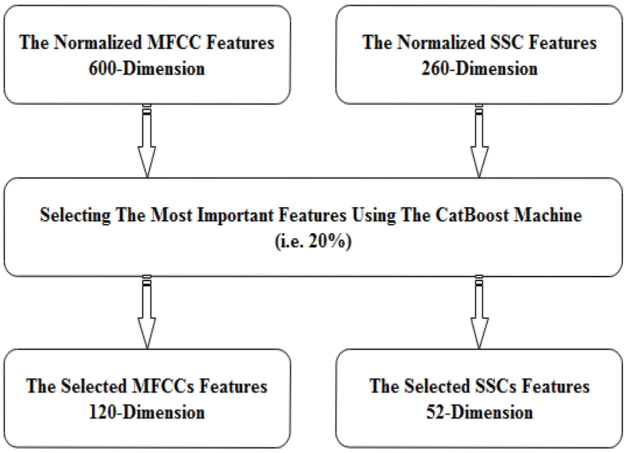
The Normalized MFCC Features
600-Dimension
The Normalized SSC Features
260-Dimension
Selecting The Most Important Features Using The CatBoost Machine
(i.e. 20%)
The Selected MFCCs Features
120-Dimension
The Selected SSCs Features
52-Dimension

Figure. 2 The proposed features selection stage.

At first, the CatBoost classifier with the learning rate of 0.08 and 500 iterations is fitted with training data features (i.e. MFCCs and SSCs) separately. Then, the CatBoost returns these features in order from most important to least important. Finally, the ratio of the most important features is chosen empirically. Fig. 2 demonstrates the main idea of the feature selection stage.

2.5. Speaker age and gender recognition using SVM classifier

SVM has become a prevalent classifier due to its promising performance in various studies. The major aim of SVM is to find a classifier that minimizes the expected error limits. SVM uses a two-step classification process. A kernel function performs a transformation of the feature from low to high dimensions in the first step. This transformation allows for linear separation at a higher dimension of non-linearly separable data. In the second step, it forms an optimum hyperplane to draw the boundary of the decision between classes [4]. SVM is considered one of the most accurate and robust methods among classification algorithms due to its use of optimum separation to prevent misclassification of outliers. SVM has been identified as one of the top 10 classification algorithms [25]. To guarantee that hyperplanes with the maximum margin are found, an SVM classifier tries to maximize the following function in terms of $\vec{w}$ and b as seen in Eq. (12) [25]:

$$L_P = \frac{1}{2}\|\vec{w}\| - \sum_{i=1}^{t} \alpha_i\, y_i(\vec{w}.\vec{x} + b) + \sum_{i=1}^{t} \alpha_i \qquad (12)$$

Where t is the training examples number, and α_i, $i = 1, \dots, t$ represents the non-negative numbers in a way that the L_P derivatives as regard to α_i is zero. L_P

is called the Lagrangian where α_i is Lagrange multipliers. In Eq. (12), the hyperplane is defined by the vectors $\vec{w}$ and constant b.

The SVM classifier is utilized in the suggested system of the speaker's gender and age recognition due to its strong classification ability. In the presented work, the radial basis function (RBF) kernel is used as recommended by [4] since it yielded a higher accuracy. The classification stage involves two steps, which are the testing step and the training step. Regarding the latter, the SVM classifier will be trained with the training part of the database to find the optimal SVM model. This model is fed then to the testing step for recognizing the speaker's age and gender.

3. Experimental results and analysis

The data-set used in this study is described, and the experiments that were performed are explained and discussed in detail in this section.

3.1 The agender data-set

The proposed system is assessed using aGender database that has been created via "InterSpeech 2010 Paralinguistic Challenge" [13, 14]. Approximately, it consists of phone conversations of 954 German speakers for 47 hrs. Also, it consists of 7 categories related to gender and age, as can be seen in Table 1. In addition, there are 65,364 utterances in the database, while the average length of utterance was 2.58 s; therefore, the utterances are specified to be short. Furthermore, the database was divided into 2 categories; the test-set includes 17,332 utterances (175 speakers), and the training-set includes 53,076 utterances (770 speakers). Since the age and gender information for the test set is not shared, the 53,076 utterances are divided randomly for the training and testing process. To avoid overfitting during the training process, 80% of utterances within each class are used for system training, and 20 % for testing purposes.

3.2 Results

Various experiments have been carried out for finding the optimal configuration of the proposed speaker's gender and age recognition system parameters. At first, the performance evaluation regarding the suggested system is carried out for the speaker's gender, age, in addition to age & gender recognition in terms of UA. Table 2 shows the results of this experiment. Tables 3, 4, and 5 show the Confusion matrix for the best result achieved in Table 2 for speaker's gender, age, in addition to age & gender recognition, respectively. Finally, a comparison of the suggested system with related works utilizing the same data-set (i.e., aGender data-set) in terms of UA is presented. Table 6 shows the results of this experiment.

Table 1. The age and gender class of the aGender dataset

Class	Category	Age range	Gender	Abb.
1	Children	7-14	Male + Female	C
2	Youth	15-24	Female	YF
3	Youth	15-24	Male	YM
4	Adult	25-54	Female	AF
5	Adult	25-54	Male	AM
6	Senior	55-80	Female	SF
7	Senior	55-80	Male	SM

As seen in Table 2, the effectiveness of the proposed system is evaluated using two measures: time complexity and UA. The table shows results that are obtained using MFCC and SSC as a source for feature extraction tasks individually—taking into consideration the trade-off between time complexity and UA, the 3 ID-numbered system leading to the highest success rate when using MFCC. This is when the CatBoost machine selection ratio is 20%. On the other hand, the 7 ID-numbered system resulted in the most acceptable success rate when using SSC. This is also when the CatBoost machine selection ratio is 20%. The table also demonstrates the high efficiency of the CatBoost machine as features selection method, where the time complexity is decreased from 2735.33 sec, 1012.66 sec (i.e., without selection) to 436.66 sec, 249.33 sec (i.e., 20% selection) for the MFCC and SSC respectively. In order to ensure the balance between the time complexity and UA, this study proposed a combination of 3 and 7 ID-numbered subsystems, which gives the highest UA for speaker gender, speaker age, and speaker age & gender recognition. This indicates the fact that the CatBoost machine selected features hold most of the discriminative information about the speaker's age and gender.

As seen in Table 3, the confusion matrix is presented in terms of gender category in which speakers were specified as male (M), child (C), and female (F). M is the most recognized gender with an accuracy of 93.87%, while C is the least recognized with an accuracy of 86.23%. Most of the overlap occurs between F and C genders, indicating the similarity in the traits of the proposed system of these two genders.

As seen in Table 4, the confusion matrix is presented in terms of the age category in which

Table 2. Performance evaluation of the proposed system for the three categories in terms of UA

ID	Feature	CatBoost Selection Ratio	No. Dimension	Time (sec)	UA (%)		
					Gender	Age	Age & Gender
1	MFCC	Without	600	2735.33	86.86	60.63	59.42
2	MFCC	10%	60	275.00	86.39	66.99	63.80
3	MFCC	20%	120	436.66	87.98	69.43	68.51
4	MFCC	50%	300	1076.33	88.25	66.00	65.08
5	SSC	Without	260	1012.66	87.84	66.86	66.22
6	SSC	10%	26	169.66	81.41	56.57	52.39
7	SSC	20%	52	249.33	84.76	61.78	60.55
8	SSC	50%	130	489.66	87.42	67.37	66.54
9	Combine 3, 7	-	172	502.66	**89.62**	**72.29**	**71.96**

Table 3. Confusion matrix (%) of gender recognition for the best result achieved in Table 2 (i.e., ID-9 system)

	C	F	M
C	**86.23**	10.08	1.73
F	12.92	**86.38**	4.39
M	0.84	3.52	**93.87**

Table 4. Confusion matrix (%) of age recognition for the best result achieved in Table 2 (i.e., ID-9 system)

	C	Y	A	S
C	**79.81**	7.41	3.04	2.99
Y	12.22	**69.00**	11.58	9.13
A	4.96	13.85	**71.81**	15.44
S	2.99	9.72	13.55	**72.42**

speakers were specified as a youth (Y), child (C), senior (S), and adult (A). C is the most recognized age with an accuracy equals to 79.81%, while Y is the least recognized one with an accuracy of 69.00%. Mostly, faults are originating from neighbour age groups, which demonstrated the effectiveness of the proposed selected features.

As seen in Table 5, the confusion matrix for age & gender category where speakers are defined as youth female (YF), child (C), adult female (AF), youth male (YM), senior female (SF), adult male (AM) and senior male (SM) is presented. C is the most recognized age & gender with an accuracy equals to 78.71%, while YF is the least recognized one with an accuracy of 66.82%. Faults regarding such a category mostly happened in discrimination between same-gender classes. This indicates why the age category achieved the highest success rate than the category of age & gender.

Table 6 shows a comparison made between the proposed system and other works utilized in the same database (i.e. aGender). The table shows the achieved UA by each work for the three categories. The suggested system is considered the first in age and age & gender categories, and it is considered the second in the gender category of all systems. As it is

Table 5. Confusion matrix (%) of age & gender recognition for the best result achieved in Table 2 (i.e., ID-9 system)

	C	YF	YM	AF	AM	SF	SM
C	**78.71**	11.67	1.98	4.36	0.99	4.87	0.88
YF	11.79	**66.82**	0.33	8.73	0.83	5.88	0.44
YM	0.68	0.79	**72.65**	0.81	10.82	1.68	7.81
AF	5.20	12.40	1.90	**69.44**	1.67	15.30	1.27
AM	0.38	0.59	12.09	1.77	**73.09**	2.52	11.50
SF	2.75	7.63	1.98	13.91	1.82	**67.54**	2.06
SM	0.45	0.06	9.03	0.95	10.74	2.18	**76.00**

Table 6. A comparison of the suggested system with related works utilizing the same data-set (i.e., aGender) for the three categories in terms of UA

Authors	**UA(%)**		
	Gender	**Age**	**Age & Gender**
Schuller et al. (2010) [13]	80.42	48.91	44.94
Kockmann et al. (2010) [12]	81.82	52.88	53.86
Li et al. (2013) [15]	81.70	52.80	50.30
Barkana et al. (2015) [4]	84.70	66.20	63.70
Yücesoy et al. (2016) [5]	**90.40**	54.10	53.50
Grzybowska et al. (2016) [6]	-	-	62.90
Qawaqneh et al. (2017) [11]	-	-	58.98
Abu-Mallouh et al. (2017) [3]	-	-	57.63
Markitantov et al. (2019) [1]	88.80	57.53	48.41
Markitantov (2020) [2]	81.74	51.71	48.96
The proposed system	89.62	**72.29**	**71.96**

clearly shown in the table, the UA achieved through the proposed system outperforms those achieved by other systems with an improvement equals to 6.09 % and 8.26% for age and age & gender categories, respectively.

4. Conclusions and future works

In this study, an automatic system is proposed to identify gender and age in short speech utterances without depending on the text. Firstly, two groups of features are combined to further improve system performance. After that, the dynamic size features are turned into static size features by measuring 10 statistical functionals for each dimension. Then, the use of the CatBoost machine as a features selection method has a vital effect on the system efficiency via generating a 172-dimensional informative feature vector. Finally, the SVM classifier gives the proposed system the classification power while taking advantage of the RBF kernel strength. The experimental results show the efficiency of the suggested system with UA of 89.62%, 72.29%, and 71.96% for gender, age, and age & gender categories, respectively, using the aGender data-set. In future works, some improvements can be considered at the feature extraction stage, for instance, implementing the voice activity detection (VAD) process before extracting the features, using the CatBoost machine as a multi-level features selection, and adding other groups of features such as jitter and shimmer.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Author Contributions

Conceptualization, methodology, and implementation writing—original draft preparation, Ameer A. Badr; writing—review, editing, supervision, and funding acquisition, Alia K. Abdul-Hassan.

References

[1] M. Markitantov and O. Verkholyak, “Automatic Recognition of Speaker Age and Gender Based on Deep Neural Networks”, In: *Proc. of International Conf. on Speech and Computer*, Istanbul, Turkey, pp. 327–336, 2019.

[2] M. Markitantov, “Transfer Learning in Speaker’s Age and Gender Recognition”, In: *Proc. of International Conf. on Speech and Computer*, St. Petersburg, Russia, pp. 326–335, 2020.

[3] A. Mallouh, Z. Qawaqneh, and B. Barkana, “New transformed features generated by deep bottleneck extractor and a GMm_UBM classifier for speaker age and gender classification”, *Neural Comput. Appl.*, Vol. 30, pp. 2581–2593, 2017.

[4] B. Barkana and J. Zhou, “A new pitch-range based feature set for a speaker’s age and gender classification”, *Appl. Acoust.*, Vol. 98, pp. 52–61, 2015.

[5] E. Yücesoy and V. Nabiyev, “A new approach with score-level fusion for the classification of a speaker age and gender”, *Comput. Electr. Eng.*, Vol. 53, pp. 29–39, 2016.

[6] J. Grzybowska and S. Kacprzak, “Speaker age classification and regression using i-vectors”, In: *Proc. of Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, San Francisco, pp. 1402–1406, 2016.

[7] M. Al-Sarem, F. Saeed, W. Boulila, A. Emara, M. Al-Mohaimeed, and M. Errais, “Feature Selection and Classification Using CatBoost Method for Improving the Performance of Predicting Parkinson’s Disease”, In: *Proc. of First International Conf. of Advanced Computing and Informatics ICACIn*, Casablanca, Morocco, pp. 189–199, 2020.

[8] J. Hancock and T. Khoshgoftaar, “CatBoost for big data: an interdisciplinary review”, *J. big data*, Vol. 7, No. 1, p. 94, 2020.

[9] K. Ghori, A. Ayaz, M. Awais, M. Imran, A. Ullah, and L. Szathmary, “Impact of Feature Selection on Non-technical Loss Detection”, In: *Proc. of 6th Conf. on Data Science and Machine Learning Applications (CDMA)*, Riyadh, KSA, pp. 19–24, 2020.

[10] F. Zhang and H. Fleyeh, “Short Term Electricity Spot Price Forecasting Using CatBoost and Bidirectional Long Short Term Memory Neural Network”, In: *Proc. of 16th International Conf. on the European Energy Market (EEM)*, Ljubljana, Slovenia, pp. 1–6, 2019.

[11] Z. Qawaqneh, A. Mallouh, and B. Barkana, “Deep neural network framework and transformed MFCCs for speaker's age and gender classification”, *Knowledge-Based Syst.*, Vol. 115, pp. 5–14, 2017.

[12] M. Kockmann, L. Burget, and J. Cernocký, “Brno University of Technology system for Interspeech 2010 Paralinguistic Challenge”, In: *Proc. of 11th Annual Conf. of the International Speech Communication Association*, Makuhari, Chiba, Japan, pp. 2822-2825, 2010.

[13] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, “The INTERSPEECH 2010 paralinguistic challenge”, In: *Proc. of 11th Annual Conf. of the International Speech Communication Association*, Makuhari, Chiba, Japan, pp. 2794-2797, 2010.

[14] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann, “A Database of Age and Gender Annotated Telephone Speech”, In: *Proc. of the Seventh International Conf. on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 1562-1565, 2010.

[15] M. Li, K. J. Han, and S. Narayanan, “Automatic speaker age and gender recognition using acoustic and prosodic level information fusion”, *Comput. Speech Lang.*, Vol. 27, No. 1, pp. 151–167, 2013.

[16] G. Sharma, K. Umapathy, and S. Krishnan, “Trends in audio signal feature extraction methods”, *Appl. Acoust.*, Vol. 158, p. 107020, 2020.

[17] 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. 91–102, 2020.

[18] K. Rao and K. Manjunath, *Speech recognition using Articulatory and Excitation Source Features*, Cham, Switzerland, Springer International Publishing AG, 2017.

[19] K. Paliwal, “Spectral subband centroid features for speech recognition”, In: *Proc. of the 1998 IEEE International Conf. on Acoustics, Speech and Signal Processing*, Seattle, Washington, USA, Vol. 2, pp. 617–620, 1998.

[20] S. Chougule and M. Chavan, “Speaker Recognition in Mismatch Conditions: A Feature Level Approach”, *Int. J. Image, Graph. Signal Process.*, Vol. 9, No. 4, pp. 37–43, 2017.

[21] B. Bolstad, R. Irizarry, M. Åstrand, and T. Speed, “A comparison of normalization methods for high density oligonucleotide array data based on variance and bias”, *Bioinformatics*, Vol. 19, No. 2, pp. 185–193, 2003.

[22] M. Pan and J. Zhang, “Quantile normalization for combining gene-expression datasets”, *Biotechnol. Biotechnol. Equip.*, Vol. 32, No. 3, pp. 751–758, 2018.

[23] A. Dorogush, V. Ershov, and A. Gulin, “CatBoost: Gradient boosting with categorical features support”, *arXiv*, No. Section 4, pp. 1–11, 2018.

[24] W. Liu, K. Deng, X. Zhang, Y. Cheng, Z. Zhang, F. Jiang, and J. Peng, “A Semi-Supervised Tri-CatBoost Method for Driving Style Recognition”, *Symmetry* , Vol. 12, No. 3. 2020.

[25] X. Wu, V. Kumar, J. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, and D. Steinberg, “Top 10 algorithms in data mining”, *Knowl. Inf. Syst.*, Vol. 14, No. 1, pp. 1–37, 2008.