Two-Level Classification in Determining the Age and Gender Group of a Speaker

Ergün Yücesoy
Vocational School of Technical Sciences, Ordu University, Turkey

Abstract: *In this study, the classification of the speakers according to age and gender was discussed. Age and gender classes were first examined separately, and then by combining these classes a classification with a total of 7 classes was made. Speech signals represented by Mel-Frequency Cepstral Coefficients (MFCC) and delta parameters were converted into Gaussian Mixture Model (GMM) mean supervectors and classified with a Support Vector Machine (SVM). While the GMM mean supervectors were formed according to the Maximum-A-Posteriori (MAP) adaptive GMM-Universal Background Model (UBM) configuration, the number of components was changed from 16 to 512, and the optimum number of components was decided. Gender classification accuracy of the system developed using aGender dataset was measured as 99.02% for two classes and 92.58% for three classes and age group classification accuracy was measured as 67.03% for female and 63.79% for male. In the classification of age and gender classes together in one step, an accuracy of 61.46% was obtained. In the study, a two-level approach was proposed for classifying age and gender classes together. According to this approach, the speakers were first divided into three classes as child, male and female, then males and females were classified according to their age groups and thus a 7-class classification was realized. This two-level approach was increased the accuracy of the classification in all other cases except when 32-component GMMs were used. While the highest improvement of 2.45% was achieved with 64 component GMMs, an improvement of 0.79 was achieved with 256 component GMMs.*

Keywords: *GMM, mean supervector, speaker age and gender classification, SVM, two level classification.*

Received November 20, 2019; accepted February 4, 2021
https://doi.org/10.34028/iajit/18/5/5

1. Introduction

In addition to the linguistic content of the vocalized phrase, the speech signal contains many important information such as the speaker's identity, age, gender, accent, social group, geographical area, health and psychological status. This information, called paralinguistic, is widely used in communication between people. For example, from a telephone conversation, we understand the speaker's identity, gender, age and/or psychological state, and determine how we address him/her. The automatic determination of this information contained in the speech signal by data processing methods has a wide range of applications, including commercial, forensic and medical applications [17]. For example, in the automatic dialogue system, the age of the user can be estimated from the voice and the speed of the speech synthesizer can be adjusted according to this information [8]. Similarly, in interactive voice response systems, customer satisfaction can be increased by offering music or advertising to customers waiting for the operator according to their ages and/or genders [25]. In forensic cases such as kidnapping and threats, this information can be used to determine the guilty by estimating the age of the speaker from the telephone records [24]. In addition, the speaker's gender can be used as a preliminary information in speaker and speech recognition systems. Thus, gender dependent models can be defined according to the gender of the speaker and the accuracy of the whole system can be increased.

Determination of age and gender from the speaker's short-term speech is a very difficult problem and interest in this issue has increased in recent years. In the studies, age and gender information is generally considered together [17, 18, 26]. However, there are studies in which this information is examined separately [10, 16]. Although adult speakers are generally used in gender recognition studies, there are also studies using children [21]. In studies involving children speeches, generally adult speakers were classified as male and female, while children were considered as a single class without gender discrimination. There are two different approaches in age recognition studies. In the first approach, it is aimed to classify the speakers according to age groups such as young, adult and senior, while in the other approach, age regression, the exact age of the speakers is tried to be estimated by years [9, 11]. Determining the exact age of the speaker is a more difficult problem than determining the age group. However, the age group rather than the exact age of the speaker is used in many areas, especially commercial applications. Therefore, there was more interest in determining the age group of the speaker. However, there is no standard for both the number of age groups and age ranges in studies on this subject. For example, there are

studies where speakers are divided into two groups as adult and senior, three groups as young, adult and senior, or four groups as child, young, adult and senior [15, 16]. There are also studies where both age and gender groups of the speaker are handled together [18, 26]. In these studies, the speakers are generally divided into 4 age groups, and the others except children are divided into two groups according to their gender, and a structure with a total of 7 classes is used.

There are many studies in the literature about age and gender classification. In the study conducted by Fokoue and Ma [10] mean vectors calculated from MFCC features were applied as an input to Support Vector Machine (SVM) and speakers were classified with 7% prediction error according to their gender. In the study conducted by Müller *et al*. [15], age (non-elderly and elderly) and gender classification performances of five different classification methods Dynamic Time Warping (DTW), Artificial Neural Network (ANN), K-Nearest Neighbors (KNN), Naive Bayes and (NB) and SVM were compared. As a result of the tests performed using jitter and shimmer acoustic features, the highest gender and age classification accuracy was obtained by ANN method as 81.09% and 96.57% respectively. In the study of Porat *et al*. [16], a new age recognition system based on Gaussian Mixture Model (GMM) weight supervectors was proposed. The proposed system was tested with two datasets (aGender and local datasets) and as a result of these tests, the weighted average recall for the 4-class age classification was 53.75% and 56.18%, respectively. In the study of Qawaqneh *et al*. [18], a new age and gender classification method using Deep Neural Networks (DNN) as feature extraction and classifier was proposed. Mel-Frequency Cepstral Coefficients (MFCC) and Shifted Delta Cepstral (SDC) coefficients were used in the study. As a result of the tests, the proposed SDC speaker+class model was found to be more successful than other systems with 57.21% classification accuracy. In the study conducted by Safavi *et al*. [21], it was aimed to identify the gender and age group of the child speakers by three different classifiers GMM-Universal Background Model (GMM-UBM), GMM-SVM and Probabilistic Linear Discriminant Analysis (PLDA) based classification with i-vector). While the best full-band width gender recognition was obtained with age independent GMM-SVM system with an accuracy of 79.18%, the best performance in determining the age group was obtained with gender independent i-vector system as 83%. In the study conducted by Yücesoy [26], a new system based on GMM supervectors was proposed for age and gender classification. In the proposed system, after removing channel effects from GMM supervectors with the NAP method, an accuracy of 62.03% was achieved using the SVM classifier.

In this study, age and gender classes were first considered separately and then these classes were combined and a total of 7 classes were classified. In gender classification, firstly, adult speakers were divided into two classes as male and female, then children were included in the data set and 3-class gender classification was made. In the age classification, male and female speakers were handled separately and classified into three age groups as young, adult and senior. After examining the age and gender classes separately, these classes were combined and a total of 7 classes were classified. In the study, for combining age and gender classes a new two-level approach was proposed. This approach is based on the idea of first classifying the data according to their more specific features and then analyzing each class separately to reveal less specific features. In line with this idea, it was thought that an increase in the overall accuracy of the system could be achieved by first classifying the speakers according to their gender and then classifying adults by age group. In this approach, instead of using a single UBM model in creating age and gender models, a separate UBM model was used for gender in the first stage, and two separate UBM models for male and female are used in the second stage. Thus, a better representation of the speakers was achieved compared to models derived from a single UBM.

This article basically contributes to the use of a two-level classification approach in which gender is determined in the first stage and gender-based age models are used in the second stage in the determination of age and gender. It also contributes to the determination of the optimum GMM component number for age and gender models.

The remainder of the paper is organized as follows; Firstly, the database and class definitions are given, and then, feature extraction, Gussian Mixture Model and SVM methods are introduced. Finally, the study is completed with the experimental results and conclusion section.

2. Database and Class Definitions

In this study, aGender database [22] was used to develop the proposed system. The aGender consists of reading and semi-spontaneous speech recordings of 945 German speakers recorded over public telephone lines. In the database, 8 speech items are voiced in 6 different sessions by each speaker. However, both the date/time of the sessions and the number of sessions per speaker are not checked. The distribution of the speakers according to age and gender classes is approximately equal and the average length of the recordings is approximately 2.58 seconds. The aGender database, which includes a total of 47 hours of telephone conversations, consists of three parts: train, development and test. There are 32527 speeches of 471 speakers in the train part, 20549 speeches of 299 speakers in the development part and 17332 speeches

of 175 speakers in the test part. There is no difference between speakers and speeches in these parts, and each part consists of separate speaker groups. Therefore, a system that is trained with the train part and tested with the development/test part will be independent of the speaker. In the aGender database, the speakers are defined in four classes as child, young, adult and senior according to their age, and three classes according to their gender as child, male and female. The age and gender classes defined in the aGender and the number of speakers/utterances in these classes are given in Table 1.

Table 1. The classes in the aGender and the number of speakers/utterances in each class.

ID	Age group	Age	Gender	The number of speakers/utterances	
				Train	Develop
1	Child	7-14	X	68/4406	38/2396
2	Youth	15-24	Male	63/4638	36/2722
3	Youth	15-24	Female	55/4019	33/2170
4	Adult	25-54	Male	69/4573	44/3361
5	Adult	25-54	Female	66/4417	41/2512
6	Senior	55-80	Male	72/4924	51/3561
7	Senior	55-80	Female	78/5549	56/3826

3. Feature Extraction

Speech is a complex signal that occurs as a result of transformations at different levels and includes information such as the speaker's age, gender, and psychological state as well as the spoken text. However, all of this information is not always needed. The purpose of the feature extraction is to extract the relevant information from the speech signal and represent the speech signal with a limited number of parameters. There are many feature extraction methods used for this purpose [1]. Of these, MFCC is the most widely used and is also preferred in this study.

In the first stage of the MFCC, the larynx and lip effects are removed from the speech signal by pre-emphasis. For this purpose a one-order FIR filter is commonly used [7]. The signal is then divided into fixed length frames which are considered stationary. The frame length is usually selected between 20 and 40 ms. During this process, each frame is defined to contain the L samples ($L \le N$) of the preceding frame, thus preventing loss of information at the boundaries of the signal. Later, a window function is applied to each frame to increase continuity between adjacent frames and to reduce sudden changes in the end of the frame. The most commonly used window function in speech processing is the Hamming window [7]. Then, by calculating the Discrete Fourier Transform (DFT) of each frame, the amplitude spectrum of the signal is obtained and then by shifting it according to the Mel-scale given by Equation (1), the Mel-spectrum is calculated. For this transformation, a band-pass filter set evenly spaced along the Mel frequency is used. Each filter has a triangular band pass frequency response and these filters cover the entire frequency range from zero to Nyquist frequency.

$$f_{Mel} = 2595 \log_{10}\left(1 + \frac{f_{linear}}{700}\right) \quad (1)$$

In the last stage, the MFCC coefficients, C_j, are calculated by performing Discrete Cosine Transform (DCT) on the logarithm of the Mel spectrum with Equation (2).

$$C_j = \sqrt{\frac{2}{N_f}} \sum_{m=1}^{N_f} \log(\tilde{s}(m)) \cos\left[\frac{j\pi}{N_f}(m-0.5)\right] \quad (2)$$

Where $\tilde{s}(m)$ is the energy of the m-th Mel-filter output, J is the number of MFCC coefficients, N_f is the number of Mel-filters.

4. Gaussian Mixture Model (GMM)

The Gaussian mixture model is the weighted sum of M individual Gaussian densities and is represented by Equation (3) [19].

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i g(x \mid \lambda_i, \Sigma_i) \quad (3)$$

Where x represents the D dimensional continuous valued data vector, w_i, i=1,…,M are the mixture weights satisfying the $\sum_{i=1}^{M} w_i = 1$ requirement and $g(x \mid \lambda_i, \Sigma_i), i = 1, ..., M$ represents Gaussian density components. Each component is the D-variable Gaussian function represented by Equation (4).

$$g(x \mid \lambda_i, \Sigma_i) = \frac{1}{2\pi^{D/2} \mid \Sigma_i \mid^{1/2}} \exp\{-\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i)\} \quad (4)$$

Here, μ_i is the mean vector, and $\sum_i$ is the covariance matrix. The Gaussian mixture model is represented by mixture weights, covariance matrices and mean vectors of all components as in Equation (5). These parameters are estimated using the Expectation-Maximization (EM) algorithm [8].

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, ..., M \quad (5)$$

5. Support Vector Machines (SVM)

SVM are a two-class classifier consisting of sums of a kernel function $K(\cdot,\cdot)$, defined by Equation (6) [3].

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(x, x_i) + d \quad (6)$$

Where the t_i are the ideal outputs, $\sum_{i=1}^{L} \alpha_i t_i = 0$, and $\alpha_i > 0$. x_i support vectors are obtained from the training set at the end of an optimization process [6]. This process, which is based on the concept of maximum margin, results in a hyperplane that guarantees maximum separation between the two classes. At the classification stage, it is decided whether $f(x)$ is above

or below a limit. The kernel function $K(\cdot,\cdot)$ can be expressed as Equation (7) by limiting it to some properties called Mercer conditions.

$$K(x, y) = b(x)^t b(y) \tag{7}$$

Here $b(x)$ represents the function used to move the input vector into a high dimensional space and is selected according to the type of application. In this study, GMM supervector linear kernel, details of which are given in the next section, is used as kernel function.

5.1. GMM Supervectors

GMM supervectors are typically generated using MAP adaptive GMM-UBM configuration. GMM-UBM has been successfully applied in many areas, especially speaker, language and dialect recognition [5]. According to this configuration, a GMM model called UBM that represents all speaker characteristics is first trained using a large dataset. The selection of the database used in the training of UBM should ensure that the distribution of all speaker groups is equal. Otherwise, the resulting background model will tend to a particular group. There are different approaches used in UBM training. The most common of these approaches is combining the features obtained from all training data and training a single GMM model with EM algorithm and using this model as UBM. The trained UBM model is considered as the initial model, and the class dependent models corresponding to each conversation are adapted from this model. There are different methods for adapting speaker models. One of these methods, Maximum-A-Posteriori (MAP) is similar to EM algorithm and is used in this study. In practice, since it gives better results in terms of both speed and performance, instead of adapting all parameters, only the mean vectors are adapted and other parameters are taken directly from UBM [20].

After the GMM model corresponding to each speech is generated by adaptation from UBM, the mean vectors of the mixture components are concatenated to obtain a fixed length vector. This vector, GMM mean supervector, is a MDx1 dimensional (M: component number, D: feature size) vector and is considered as the conversion between speech and high dimensional vector. The natural distance between the two GMMs indicated by g_a and g_b, corresponding to utt_a and utt_b conversations, is given by Equation (8).

$$D(g_a \| g_b) = \int_{R^N} g_a(x) \log\left(\frac{g_a(x)}{g_b(x)}\right) dx \tag{8}$$

However, this expression does not satisfy the Mercer requirement and is difficult to use with SVM. Therefore, instead of the direct use of natural distance (KL-divergence), the expression given by Equation (9) is used as its prediction.

$$K(utt_a, utt_b) = \sum_{i=1}^{M} (\sqrt{w_i}\,\Sigma_i^{-\frac{1}{2}}\mu_i^a)^t (\sqrt{w_i}\,\Sigma_i^{-\frac{1}{2}}\mu_i^b) \tag{9}$$

This expression is linear and can be used as a kernel function because it satisfies the Mercer condition. The linear GMM kernel based on the KL-divergence normalizes all component means with the expression $\sqrt{w_i}\,\Sigma_i^{-\frac{1}{2}}$ and thus variance and weight parameters are included in the supervectors.

5.2. GMM Supervector SVM

The GMM supervector SVM classification is a hybrid approach that combines the generative power of the GMM with the discriminative features of the SVM. It was first used by Campbell *et al.* [4] in the speaker verification system and then it was used in different studies such as age, gender, psychological status and language recognition [21, 23]. In the first stage of approach, a GMM model called UBM is trained using a large data set. The class-dependent models that correspond to each utterances in the training database are then created using the UBM-MAP configuration. In the final stage, the class-dependent GMM models are converted into supervectors and applied to the SVM classifier to train the SVM. After the training phase, the system is tested using a different dataset. In the testing phase, the supervectors corresponding to each speech are first created as in the training phase. These supervectors are then applied as input to the trained SVM, and a class prediction is made for each entry and the classification process is completed. In this study, this approach was used to classify speakers according to age and gender groups.

5.3. Two-level Age and Gender Classification

In the GMM supervector SVM classification approach, a single UBM is defined and all class dependent models are adapted from this model. Considering the age and gender classification problem discussed in this study, there are 7 different classes and a single UBM is used in the training of all class dependent models. In fact, in the training of class-dependent models, the use of only UBMs created with speakers from their gender group will provide better modelling of the classes. Based on this idea, a two-level classification approach was proposed in this study. According to this approach shown in Figure 1, the speakers are first divided into three groups as child, male and female. At this stage, a background model is used in which all gender groups (child, male, female) are represented equally. Then, a second classification is made on male and female groups and the speakers are divided into three age groups: young, adult and senior. Thus, at the end of the two stages, the speakers are divided into a total of 7 age and gender classes. In the second stage, in which the speakers are classified according to age groups,

two different UBMs are defined for male and female gender groups and each age group model is formed by adapting from the relevant gender dependent UBM. As in the first stage, in the training of the UBMs defined in the second stage, homogeneous datasets in which each subgroup (age) is represented approximately equally, should be used.

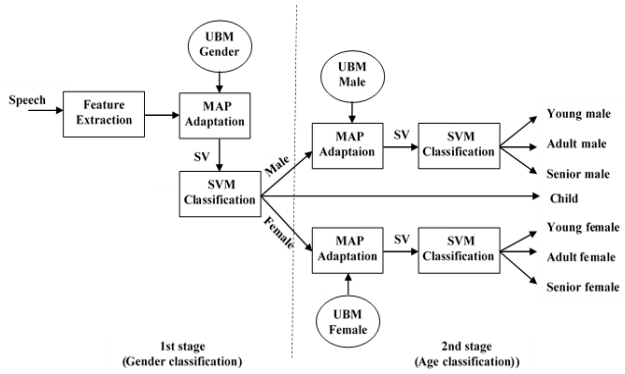

Figure 1. Two-level age and gender classification system.

6. Experimental Results

In the experiments the age and gender characteristics of the speakers were first considered separately, then these classes were combined and a 7-class classification was made. In the study, a two-level approach is proposed to classify age and gender classes together. This approach, details of which are given in Section 5.3, has been applied with the aGender data set and the results are presented in the following section together with the results of the single-level approach.

6.1. Gender Classification

The gender classification system have been tested in two cases. In the first of these tests, only adult speakers were classified as female and male, then the number of classes was increased to three by including children in the data set. In the case where the speakers were classified as male and female, a background model trained only with adult speakers was used, while in the three-class case an UBM trained with all speaker groups was used. The data used in the training of UBM were randomly selected from the training section of the aGender database, 20 speakers of each age and gender group. The background model was trained with the EM algorithm and a vector with 39 elements, created with 13 MFCC coefficients and delta coefficients, was used as features. In addition, the number of GMM components were changed from 16 to 512 and the optimum component number was determined. In the study, the speakers except the speakers selected from the training section of the aGender for background were used in the training of SVM. GMM supervectors corresponding to speeches in each session were created according to the UBM-MAP configuration, and an SVM was trained using these vectors. In this study, 1355 speeches of 283 speakers were used for the training of SVM with two classes, and 1571 speeches of 331 speakers were used in the training of SVM with 3 classes. In the final stage, the trained SVM classifier was tested with the speeches in the development section of the aGender database. 1226 speeches of 261 adult speakers were used in two-class tests, and 1388 speeches of 299 speakers, 38 of whom were children, were used in three-class tests. In these tests given in Table 2, the highest classification accuracy was obtained as 99.02% for two classes and 92.58% for three classes by using 64-component GMM.

Table 2. Gender classification accuracy (%).

Number of components	Two-class accuracy	Three-class accuracy
16	97.15%	89.91%
32	98.12%	91.14%
64	**99.02%**	**92.58%**
128	98.86%	92.44%
256	**99.02%**	92.36%
512	**99.02%**	91.86%

6.2. Age Classification

In the gender classification system, details of which are given in the previous section, only data sets and class definitions were changed and the same system was used as an age classifier. In the study, two different age classification systems were defined for male and female speakers and the speakers were divided into three groups as young, adult and senior according to age groups. 665 speeches of 139 speakers were used in the training of the male age classification system and 690 speeches of 144 speakers were used in the training of the female age classification system. In the test phase, the male age classification system was tested with 580 speeches of 130 speakers, and the female age classification system was tested with 646 speeches of 131 speakers. In these tests given in Table 3, the highest classification accuracy was obtained as 67.03% for female and 63.79% for male as a result of the use of 256 component GMMs.

Table 3. Age classification accuracy (%).

Number of GMM components	Accuracy for females	Accuracy for males
16	57.74%	57.59%
32	62.85%	57.76%
64	64.55%	61.55%
128	64.86%	61.55%
256	**67.03%**	**63.79%**
512	65.63%	63.79%

6.3. Age and Gender Classification

In this section, age and gender classes were combined to make a classification of 7 classes. Speakers defined as three classes (child, male and female) in gender classification were divided into 7 classes and the same system was used as age and gender classifier by changing the class labels of the speakers. In the training and test stages of the age and gender

classification system, the same data sets used in the 3-class gender classification system were used. In the tests given in Table 4, the highest age and gender classification accuracy was obtained as a result of the use of 256 component GMMs with 61.46%.

Table 4. Age and gender classification accuracy (%).

Number of GMM components	Accuracy
16	53.24%
32	59.15%
64	58.14%
128	59.58%
256	**61.46%**
512	61.02%

The confusion matrix of the system for the case with the highest accuracy is given in Table 5. The confusion matrix, also known as the error matrix, is a commonly used representation to represent the performance of a classification system. In this representation, together with the individual classification rate of each class, the distribution of erroneous decisions between classes is also shown. From the given confusion matrix, it is seen that the most confusion is between the age groups of same gender speakers (such as AM-YM, SF-AF, etc.). The confusion between the age groups of same gender speakers is followed by the child and female age groups, and the child and male age groups, respectively. While the age and gender group classified with the highest accuracy was the young female group with 72.3%, the group with the lowest accuracy was the adult male group with 43.8%.

Table 5. Confusion matrix of single level age and gender classification system.

	C	YF	YM	AF	AM	SF	SM
C	**65,4**	12,3	6,8	6,2	0,6	8,6	0,0
YF	11,3	**72,3**	0,0	13,6	0,0	2,8	0,0
YM	1,4	2,0	**63,9**	2,0	17,7	1,4	11,6
AF	1,8	22,8	0,4	**54,5**	0,4	20,1	0,0
AM	0,0	0,0	28,4	2,4	**43,8**	0,0	25,4
SF	0,8	9,8	0,0	28,2	0,0	**60,8**	0,4
SM	0,0	0,0	6,1	0,4	23,9	1,5	**68,2**

*C: Child, YM: Young Male, AM: Adult Male, SM: Senior Male, YF: Young Female, AF: Adult Female, SF: Senior Female.

6.4. Two-Level Age and Gender Classification

In the age and gender classification system, details of which are given in Section 6.3, all speaker models are derived from the same background model and the age and gender class of the speakers is decided at a single stage. In this study, instead of this one-stage approach, a new two-level approach in which age and gender groups are handled separately is proposed. According to this approach, details of which are given in Section 5.3, the speakers were first classified as child, male and female according to gender groups, then male and female speakers were classified according to age groups and a total of 7 classes were classified. In the training and testing stages of the proposed two-level classification system, the same data sets used in the one-level approach were used. In the tests given in Table 6, the highest age and gender classification accuracy was obtained as a result of the use of 256 component GMMs with 62.25%.

Table 6. Two-Level age and gender classification accuracy.

Number of GMM components	Accuracy
16	54.68%
32	57.56%
64	60.59%
128	60.45%
256	**62.25%**
512	61.31%

In the case where 256 components are used, the number of samples and accuracy rates at each stage of the system are as follows. In the first stage, a total of 1388 speech, 580 male, 646 female and 162 children, were applied to the gender classifier and the speeches were successfully classified according to gender by 92.36% (1282/1388). At this stage, the classification accuracies of male, female and child utterances were measured as 97.75% (567/580), 95.20% (615/646) and 61.72% (100/162), respectively. Then, a second classification was made on male and female utterances classified in the first stage and 63.84% (362/567) of male utterances and 65.36% (402/615) of female utterances were correctly classified according to age group. Thus, the overall classification accuracy of the system was determined to be 62.25% (864/1388). The individual classification accuracy of each age and gender group was measured as 67.85% for young male, 46.70% for adult male, 72.69% for senior male, 61.72% for children, 74.65% for young female, 61.60% for adult female and 63.26% for senior female.

When the results of the proposed two-level approach and the one-level approach are compared, it is seen that the accuracy of the proposed two-level approach is higher in all other cases, except for the case where 32-component GMMs are used. While the highest increase in accuracy was achieved with the use of 64-component GMMs with 2.45%, the overall accuracy of the system increased from 61.46% to 62.25% in the case of using 256 component GMMs.

Results of recent studies on age and gender classification are tabulated in Table 7. When the results in this table are examined, it is seen that the two-level approach in this study is higher than the success of the other study. In addition, considering the extra processing requirement of the Nuisance Attribute Projection (NAP) method in [26], which has the closest accuracy to the accuracy obtained in this study, the low processing requirement of the proposed approach can be considered as another advantage.

Table 7. Comparing of recent studies.

Studies	Task	Data Set	Accuracy (%)
Kockmann *et al.*[12]	Gender Age Age and Gender	aGender	81.82 56.03 53.86
Porat *et al.* [16]	Age	Local aGender	56.18 53.75
Li *et al.* [13]	Gender Age	aGender	85.0 52.0
Qawaqneh *et al.* [18]	Age and Gender	aGender	57.21
Safavi *et al.* [21]	Gender Age	OGI Kids	79.18 83.0
Büyük and Arslan [2]	Age	Multi-language	59.9 (for female) 49.7 (for male)
Markitantov and Verkholyak [14]	Gender Age	aGender	88.80 57.53
Yücesoy [26]	Age and Gender Age Gender	aGender	62.03 61.82 92.30
This Study	Age and Gender	aGender	62.25
	Age		63.79 (for male) 67.03 (for female)
	Gender		92.58 (3-class) 99.02 (2-class)

7. Conclusions

In this study, a new two-level approach is proposed for classifying speakers by age and gender groups. In the first phase of the system, which was developed based on the GMM supervector SVM method, the speakers were divided into three classes according to gender: child, male and female. Then, male and female speakers were divided into three classes (young, adult and senior) according to their age groups and a total of 7 classes were classified. In this study, 39 element vectors consisting of 13 MFCC coefficients and first and second derivatives of these coefficients were used as features. GMMs created according to GMM-UBM configuration were converted to supervectors and applied to SVM to determine the age and gender group of the speakers. Firstly, adult speakers were classified as male and female. In the tests conducted with a total of 1226 speeches, 646 of which were female and 580 were male, 100% of female speeches and 97.93% of male speeches were correctly classified and 99.02% accuracy was achieved. Later, 162 children's speeches were included in the data set and they were classified into three gender classes. In the tests, the accuracy of the three-class classification was measured as 92.58%. In the study, in classifying the speakers according to age groups, the speeches of male and female were examined separately. In the tests, age classification accuracy was measured as 67.03% for 646 female speeches, and 63.79% for 580 male speeches. After the age and gender classes were handled separately, these classes were combined to investigate the 7-class situation in which the age and gender groups of the speakers were identified together. For this purpose, only the data set and class labels of the developed system were changed and the same system was used for age and gender classification with 7 classes. In tests conducted with 1388 speeches of 299 speakers, 38 of whom were children, the accuracy of classification in 7 classes was measured as 61.46%. Finally, the performance evaluation of the two-level age and gender classification system proposed for the 7-class situation was performed. In tests performed using the same data sets, it was observed that the overall accuracy of the system was increased from 61.46% to 62.25% by using the proposed approach.

In this study, the number of GMM components was changed between 16 and 512 and the optimum number of components was decided for each case. As a result of the tests, the optimum number of components was found to be 256 for age classification, 64 for gender classification and 256 for classification of both together. The accuracy of the age group classification system developed in the study is quite low compared to the accuracy of the gender classification system. In addition, there was a 7% decrease in the accuracy of gender classification when children were included in the data set. Considering these results, it is seen that there is a need for new features that better represent age group differences and children's speech. In this context, it is thought that it would be beneficial to investigate prosodic features such as jitter, shimmer, spectral slope and HNR and to use different feature vectors at every stage of the proposed system.

References

[1] Alim S. and Rashid N., *From Natural to Artificial Intelligence-Algorithms and Applications*, IntechOpen, 2018.

[2] Büyük O. and Arslan L., "An Investigation of Multi-Language Age Classification from Voice," *in Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies*, Prague, pp. 85-92, 2019.

[3] Burges C., "A Tutorial on Support Vector Machines for Pattern Recognition," *Data mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.

[4] Campbell W., Sturim D., and Reynolds D., "Support Vector Machines Using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308-311, 2006.

[5] Chellali S., Al-Maadeed S., Kenai O., Ahfir M., and Hidouci W., "Middle Eastern and North African English Speech Corpus (MENAESC): Automatic Identification of MENA English Accents," *The International Arab Journal of Information Technology*, vol. 18, no. 1, pp. 67-76, 2021.

[6] Collobert R. and Bengio S., "Svmtorch: Support Vector Machines for Large-Scale Regression Problems," *Journal of Machine Learning*

Research, vol. 1, no. 2, pp. 143-160, 2001.

[7] Deller J., Hanse J., and Proakis J., *Discrete Time Processing of Speech Signals*, IEEE Press, 2000.

[8] Dempster A., Laird N., and Rubin D., "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1-22, 1977.

[9] Dobry G., Hecht R., Avigal M., and Zigel Y., "Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 1975-1985, 2011.

[10] Fokoue E. and Ma Z., "Speaker Gender Recognition via Mfccs and Svms," Technical Report, Digital Media Library Running, 2013.

[11] Grzybowska J. and Kacprzak S., "Speaker Age Classification and Regression Using i-Vectors," *in Proceedings of 17th Annual Conference of the International Speech Communication Association*, San Francisco, pp. 1402-1406, 2016.

[12] Kockmann, M., Burget, L., and Černocký, J., "Brno University of Technology System for Interspeech 2010 Paralinguistic Challenge," *in Proceedings of 11th Annual Conference of the International Speech Communication Association*, Makuhari, pp. 2822-2825, 2010.

[13] Li M., Han K., and Narayanan S., "Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level Information Fusion," *Computer Speech and Language*, vol. 27, no. 1, pp. 151-167, 2013.

[14] Markitantov M. and Verkholyak O., "Automatic Recognition of Speaker Age and Gender Based on Deep Neural Networks," *in Proceedings of International Conference on Speech and Computer*, Istanbul, pp. 327-336, 2019.

[15] Muller C., Wittig F., and Baus J., "Exploiting Speech for Recognizing Elderly Users to Respond to Their Special Needs," *in Proceedings of 8th European Conference on Speech Communication and Technology*, Geneva, pp. 1305-1308, 2003.

[16] Porat R., Lange D., and Zigel Y., "Age Recognition Based on Speech Signals Using Weights Supervector," *in Proceedings of 11th Annual Conference of the International Speech Communication Association*, Makuhari, pp. 2814-2817, 2010.

[17] Přibil J., Přibilová A., and Matoušek J., "GMM-Based Speaker Age and Gender Classification in Czech and Slovak," *Journal of Electrical Engineering*, vol. 68, no. 1, pp. 3-12, 2017.

[18] Qawaqneh Z., Mallouh A., and Barkana B., "Deep Neural Network Framework and Transformed Mfccs for Speaker's Age and Gender Classification," *Knowledge-Based Systems*, vol. 115, no. 1, pp. 5-14, 2017.

[19] Reynolds D. and Rose R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp.72-83, 1995.

[20] Reynolds D., Quatieri T., and Dunn R., "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.

[21] Safavi S., Russell M., and Jančovič P., "Automatic Speaker, Age-Group and Gender Identification from Children's Speech," *Computer Speech and Language*, vol. 50, no.1, pp. 141-156, 2018.

[22] Schuller B., Steidl S., Batliner A., and Burkhard F., "The Interspeech 2010 Paralinguistic Challenge," *in Proceedings of 11th Annual Conference of the International Speech Communication Association*, Makuhari, pp. 2795-2798, 2010.

[23] Schwenker F., Scherer S., Magdi Y., and Palm G., "The GMM-SVM Supervector Approach for the Recognition of the Emotional Status from Speech," *Lecture Notes in Computer Science*, Limassol, pp. 894-903, 2009.

[24] Tanner D. and Tanner M., *Forensic Aspects of Speech Patterns: Voice Prints, Speaker Profiling, Lie And Intoxication Detection*, Lawyers and Judges Publishing Company, 2004.

[25] Van-Heerden C., Barnard E., Davel M., Walt C., Van-Dyk E., Feld M., and Müller C., "Combining Regression and Classification Methods for Improving Automatic Speaker Age Recognition," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, pp. 5174-5177, 2010.

[26] Yücesoy E., "Speaker Age and Gender Classification Using GMM Supervector and NAP Channel Compensation Method," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-10, 2020.

Ergün Yücesoy received his BSc., MSc. and Ph.D. Degrees from Department of Computer Engineering, Karadeniz Technical University, Trabzon, Turkey in 1999, 2004 and 2017 respectively. Currently, He is an Assistant Professor in Ordu Vocational School, Ordu University, Ordu, Turkey. His research interests include biometric security and machine learning.