IJAMEC
e-ISSN: 2147-8228

INTERNATIONAL JOURNAL OF APPLIED MATHEMATICS ELECTRONICS AND COMPUTERS

www.dergipark.org.tr/ijamec

International Open Access

Volume 10
Issue 04

December, 2022

Research Article

Speech-to-Gender Recognition Based on Machine Learning Algorithms

Serhat Hızlısoy [a,*], ***Emel Çolakoğlu*** [b], ***Recep Sinan Arslan*** [a]

[a] *Kayseri University, Faculty of Engineering, Architecture and Design Faculty, 38280, Kayseri, Turkey*
[b] *Kayseri University, Graduate School of Education, 38280, Kayseri, Turkey*

ARTICLE INFO

Article history:
Received 17 November 2022
Accepted 28 December 2022

Keywords:
Gender Recognition, Machine Learning, MFCC, spectrogram, logistic regression, Turkish

ABSTRACT

Speech recognition has several application areas such as human machine interaction, classification of phone calls by gender, voice tagging, STT, etc. Predicting gender from audio signals is a problem that is easy for humans to solve, difficult to solve by a computer. In this study, a model based on MFCC and classification with machine learning is proposed for gender estimation from Turkish voice signals. Within the scope of the study, 58 different series and films were examined and a new original dataset was created with 894 audio recordings consisting of 5 sec sections taken from them. Mel-frequency cepstral coefficients (MFCC) and spectrogram, which are frequently used in the literature, were used for feature extraction from audio data. The results were first evaluated separately using two features in one way. A hybrid feature vector was then created using two feature vectors. Different machine learning algorithms (LR, DT, RF, XGB etc.) were tested in the classification process and it was seen that the best accuracy was achieved in the hybrid model and logistic regression with 89%. Recall, precision and f-score values were obtained as 86.8%, 92% and 89.3%, respectively. The obtained test results revealed that the proposed model, together with the hybrid feature vector used, the original dataset and the classifier based on machine learning, showed classification success in terms of accuracy and was a stable and robust model.

1. Introduction

Speaking is one of the most effective and natural methods of communication between people [1]. This causes them to want to use voice to exchange data and creates the expectation that voice will be used in the communication of people and computers. For this reason, studies in this field include complex problems from simple human-computer interaction applications to applications that convert from sound to text. In recent years, there has been a lot of interest from researchers in this field and they continue their research for the development of robust models with higher performance [2].

Speech is produced by vibrations of the vocal cords, the movement of articulators and the inhalation of air from the lungs. Sounds are created using lip, tongue and mouth positions [3].These audio signals are analog. In order for computers to operate on this data, the analog data must be converted into a digital signal. This conversion is provided by encoding the signal and sampling it many times in one second and recording the loudness of the sound [4].

The maximum rate for the hearing bandwidth of our ears is 22 kHz. Other important information about sound is frequency and amplitude. The number of vibrations that the sound creates in 1 second is called the frequency. As the frequency increases, sound becomes thinner and decreases sound thickens. Therefore, the frequency of women's voices is higher. While the tone of the sound is important in frequency, the intensity of the sound is important in terms of amplitude [4].

A speech signal contains not only the content of the conversation, but also information such as the speaker's age, gender, mood and identity. These signals also play an important role in computer and human interaction. Today, speech signals are used for automatic speech recognition [5], speaker recognition [6], emotion recognition from

* Corresponding author. E-mail address: *serhathizlisoy@kayseri.edu.tr*
DOI: 10.18100/ijamec.1221455

music [7], gender recognition, age estimation, and speech to emotion recognition systems [8] [9].

Speech recognition is one of the most important areas of human-computer interaction and allows the control of devices, services and systems along with the use of voice in communication with the computer. There are research projects that take advantage of this opportunity, as well as commercial personal support assistants such as Apple Siri, Baidu Speech, Google Speech, Azure Speech Service [10].

Speech recognition consists of real data. It is used with machine learning and deep learning algorithms. Machine learning is a very multidirectional field of artificial intelligence. Learning takes place based on the data in hand. It has many uses in many fields such as banking, education, finance, automotive, health. [11]

For humans, determining gender from sound is quite easy. However, with computer systems, it is not easy to guess whether the speaker is male or female. When we look at the literature, it is seen that many studies have been done on this subject. However, classification accuracy has still not reached the desired level [12].

There are many reasons why it is not easy to recognize gender without speaking. The first of these is that the speech characteristic of each individual is singular. Another problem is that the data contains noise. The presence of noise in the data is a problem that reduces the recognition rate considerably. Therefore, noise must be eliminated by using some pre-processing techniques to speech data. The fact that each speaker is different is due to the anatomy of the voice. In addition, at times the speech characteristics of a woman and a man can be very similar. For this reason, in order to achieve good results in gender recognition, it is necessary to train with a large number of different characteristics of data [13].

When previous studies are examined, it is seen that there are some differences between female and male speech. These are physiological (e.g. vocal tract length), phonetics, and sound quality differences [14].

The most important point in speech-to-gender recognition systems is the creation of feature vectors. The selection of appropriate features directly affects the recognition rate [4].

Extracting the values that reflect the speaker from his voice is called feature extraction [15]. There are number of features that are used to recognize gender from voice. The most common features used in gender recognition are mel-scale power spectrogram (Mel), mel frequency ceptral coefficients (MFCC), power spectrogram (Chroma), spectral contrast, and tone weight central features (Tonnetz). The most commonly used attribute in speech-to-gender recognition systems is MFCC. The reason is that the recognition performance is better than other attribute vectors [4].

After the extraction of the features, machine learning algorithms are used on the data set created to obtain high quality recognition rates [16]. In machine learning, classification is the distribution of data among the various classes defined in the dataset. The classification process begins with the separation of the dataset with a certain class label as training and test data. The model is then designed with training data and validated with test data. [17]

The success of the classification process is influenced by the fact that robust and distinctive features are given to the machine learning system [11]. The biggest problem encountered here is that the data distribution is not balanced. In this case, the group with little distribution is included in the training data little or not at all. As such, the risk of error is high in the group with low distribution after the training of the model [17].

When the studies in the literature are examined, it is seen that different algorithms are used in the process of gender recognition from speech. Some of these are SVM, ANN, logistic regression, linear regression, random forest, AdaBoost, Decision Logic, CNN, KNN, Genetic Algorithm and DNN.

In the evaluation of the success of the model created by using classification algorithms, a matrix that is the output of the model is used. This is called a confusion matrix. In this matrix, each row shows the actual values and each row shows the predicted values. Based on the data in this matrix, performance criteria values such as accuracy, precision and recall are obtained. From here, the success percentages of the results obtained are found.

2. Related Works

In this study, it is aimed to make gender recognition by voice. Many studies have been carried out in this field from past to present. Current ones from these studies are explained in detail in this section. When the literature is examined in general, there are basically two main approaches in speech gender recognition studies. The first approach uses the numerical properties and characteristics of the voice. For example, average frequency, mode and standard deviation. The second approach is to use the spectral properties of the sound, like MFCC's and Log-Mel features [18].

Alkhawaldeh (2019) [16] used a structure with 20 languages as a dataset and 16 sound samples in 8 files for each gender in each language. The features included in the study are MFCCs, Chroma, Mel, and Tonnetz. The best relationship is between MFCC and Chroma. The worst correlation is between Chroma and Contrast. For feature selection, Evolutionary Search, PSO search, and Wolf search were used. Best results were achieved with DL_norm (99.97%) and SVM (99.7%) without feature selection. After feature selection, the best result was obtained as SVM (100%).

Taspinar et al (2020) [19] used Acoustic Analysis of the

Dataset as a dataset in their study. The data set consists of 1584 female and 1584 male voices. A total of 22 features were extracted from the data set. In the classification process, ANN (Artificial Neural Network) was used. ANN consists of 20 input neurons, 100 hidden layer neurons, and 2 output neurons. Cross-validation was used. In this method, the data set is divided into k parts. In each training process, the k-1 section of the dataset is reserved for training. The remaining part was used in the testing process. This process continued until all k sections were used in the testing process. In this study, k value was selected as 10. The success rate achieved is 97.9%.

Levitan et al. (2016) [14] in their study investigated the effect of using spectral traits along with pitch features on determining gender. Praat was used to obtain the F0 feature. For each element of the data set, the minimum, maximum, median, mean and standard deviation values were calculated from the value of f0. In addition, 21 MFCC and energy features were obtained from each sample. Mean and standard deviation values were obtained within each element.

HMIHY ("How May I Help You") was used as the English data set. This dataset consists of 1654 speakers and 5002 data records. In this study, 4520 records of the data set were used. 80% of the data set is reserved for training, 10% for development and 10% for testing. In this study, 4 classifiers were used. These are logistic regression, linear regression, random forest and AdaBoost. The best result was obtained by logistic regression in all 3 data sets. In this study, it was observed that F0 feature alone were more successful in gender recognition than cepstral features. However, in the hybrid model, 95% accuracy has been achieved.

aGender was also used as the German data set. The data is labeled in 3 parts. (woman, man and child). In the German data set, the best results were obtained with Random Forest in the analysis including children's voices. If all features were included in the process, the success rate was 85%, and with only the F0 feature, this rate was 83.3%. With the combination of F0+Energy+Voice Quality, the rate was 84.1%.

For the hybrid training set, male and female labels were provided from the HMIHY dataset and child data were obtained from the German dataset. As a result, the training kit is built on British data, while the test data is built on German data. The best result was achieved in the Logistic Regression classifier in the hybrid model. (%92.1) (With F0 feature).

Sadek et al. (2012) [20] created the dataset of their study themselves. PCM (Pulse Code Modulation) was used to digitize speech signals. FFT and Power Spectrum were preferred as features. Decision Logic was also used as a classifier. In this study, 10 speakers were located (5 men and 5 women) .All speakers were asked to voice the letter A and B and were recorded. For the letter A, the gender of all speakers has been guessed correctly. For B, 4 speakers were misestimated. As a general result, 80% success was achieved.

Zhong et al. (2019) [21] tried to recognize gender without speaking using a decision tree algorithm in their study. The model is trained with 2250 recordings of random sounds. All of these sounds are taken from the Festvox website. In the test process, the recognition rate of random sounds is 99.9%. The reason is that it is a ready-made dataset for this purpose. The test result from daily speech is 90%. These speeches were recorded by the academicians who conducted the study. Because the data type is different from the data in the training process, the recognition rate has decreased. Song data was obtained by downloading from popular music from the internet. Their duration is more than 3 minutes. The training set consisted of recordings between 3-5 seconds. Because of this difference, the recognition rate in the lyrics was also quite low.

Yücesoy et al. (2013) [22] used the TIMIT dataset in their studies. Here various parameters are proposed, focusing on different aspects of the laryngeal flow signal. These parameters are generally classified as time-based and frequency-based. Time-based parameters are calculated using the opening and closing moments of the laryngeal pulses or the maximum and minimum flow values. The most commonly used time-based parameters are OQ (Aperture rate), CIQ (Closing Rate) and SQ (Speed Ratio). In the study, the relationship between the sex of the speaker and the laryngeal flow signal that occurs in the larynx and is thought to be the source of the sound was investigated. For this purpose, a three-stage system has been proposed. In the first stage, from the sound recordings recorded by means of a microphone, the effect of the voice path and the lip is extracted by reverse filtering and the laryngeal flow signal is obtained. At the second stage, quantitative parameters are determined from the obtained laryngeal flow signal. Another parameter used in the study is dH1H2, which is obtained from the spectrum of the laryngeal flow signal and calculated as the difference of the first two harmonics in decibels. In the third and final stage, a threshold value is determined for each parameter and the gender of the speaker is decided according to this level. From experimental studies, it was seen that SQ, CIQ and dH1H2 parameters derived from the sound source were significantly related to the speaker gender. Among these parameters, 99% success was achieved with the laryngeal closure rate CIQ.

Uddin et al. (2021) [23] used a layered structure in feature extraction in their study. Fundamental frequency, spectral entropy, spectral flatness and mode frequency features are obtained in the first layer. In the second layer, the MFCC feature is obtained. In the third layer, linear predictive coding (LPC) is calculated.

The audio data set has noise and distortion. Therefore,

pre-processing was done before removing features from the data. During the pre-processing process, a high-pass filter, z-score normalization and Savitzky-Golay filter were applied. The data sets used in this study are TIMIT, RAVDESS and BGC. By combining 3 data sets, 1433 audio files were used for training and 615 audio files were used for testing. In the classification process, 1D CNN was preferred. When different optimization options were used and the results were compared, the best result was achieved with Adam. (93.01%)

Jena et al. (2020) [24] created 300 different audio recordings from high school students aged between 19-22 aged as a data set in their study. All the students said the same sentence. Different features were extracted using short-term, statistical and spectral analysis of speech signals. Short-term Mean Magnitude (STAM) and short-term mean energy (STE) features with short-term analysis; mean, variance and standard deviation features were obtained by statistical analysis; mean power density, average frequency and median frequency features were obtained by spectral analysis. KNN and SVM were also used as classifiers. A recognition rate of 87.5% was achieved with SVM and 80% with KNN. In addition, the highest recognition rate in KNN is achieved with the Euclidean and City Block distance functions.

Yücesoy et al (2016) [25] applied SVM based on GKM super vectors in their study, which combines the generalizing power of GKM (Gaussian Mixture Model) with the distinguishing feature of SVM (Support Vector Machines) approach. A total of 39 coefficients consisting of 13 MFCC coefficients from each frame and their first and second derivatives were used as features. In this study, classification was made in three breakdowns according to gender. (woman, man and child) The data set used is aGender. The highest classification successes in the tests were obtained as a result of modeling 16-second conversations with 64-component GCMs. These rates are 92.42% in the gender category. It is seen that most of the erroneous decisions are due to the classification of child speakers as female, while the transition between adult speakers is quite low.

Thangaiyan et al. (2017) [26] applied fuzzy logic and the neural network approach which yields the necessary results for gender classification due to the complexity of the educational network. Various evolutionary algorithms such as the Genetic Algorithm (GA) are applied in gender classification to overcome this problem. GA was also used in this study. The features used in the study are Short time energy (STE), Zero Crossing Rate (ZCR), and Energy Entropy (EE). 80 speech signals were used as inputs and then the data set was divided into 4 groups. In the classification process, Fuzzy Logic (FL), Neural Network (NN), Naive Bayes (NB) and Genetic Algorithm were used and their results were compared. Among these, the highest recognition rate was obtained with GA (76.75%).

Kiani et al. (2017) [27] preferred Hamming windowing w (n) to reduce the distortions of sound during frequency conversion and for spectral analysis of speech in their proposed method. In the training process, voice data from 10 female and 10 male speakers were used. MFCC is used as the feature and DNN Tool is used for classification. In the test process, the data of 5 female and 5 male speakers were advanced. The success rate of gender recognition is 90%.

Yücesoy (2020) [28] addressed the issue of automatic determination of the age and gender group of the speaker from short-term telephone conversations in his study. A 39-element vector formed by the first and second derivatives of coefficients with 13 MFCC coefficients (including zero) is used as the feature. Gaussian Mixture Models (GKM) were used to model the age and gender class. The generated GKM models were converted into super vectors and classified with SVM. There are several alternatives that have been proposed for the core, which is the most important component of SVM training. In this study, performance comparison of linear core, polynomial nucleus, radial-based (RBF) core and GKM KL divergence core was made. The system developed using the aGender database was tested with four different SVM cores and five different component numbers. In the tests conducted with the aGender database, the highest classification rate was obtained as 60.95% as a result of the classification of 256-component GKMs with GKM-KL cores. This rate is both the age and gender estimate ratio.

3. Methodology

In this section, the architectural structure of the study is mentioned. In the study, audio signals were obtained over the speaker and text independent data set. From these audio signals, MFCC and spectrogram were extracted as features. The data set is divided into training and testing. Classification was made with 8 different machine learning algorithms and the results were compared. As another analysis, the results of this study were compared with recent studies. Accuracy, precision, recall and f-score values were calculated on the 2-class result output.

3.1. System Architecture

The architectural structure of the study is shown in Figure 1. Within the scope of the study, 58 single films and series were examined. As a result of this examination, 894 audio recordings of 5 seconds were prepared. Camtasia Recorder application is used for this process. In addition, all cutting and correction operations on audio files were done with Camtasia Studio. The file format of audio recordings is wav. All of the recordings are single-channel and the sampling frequency is 22050.

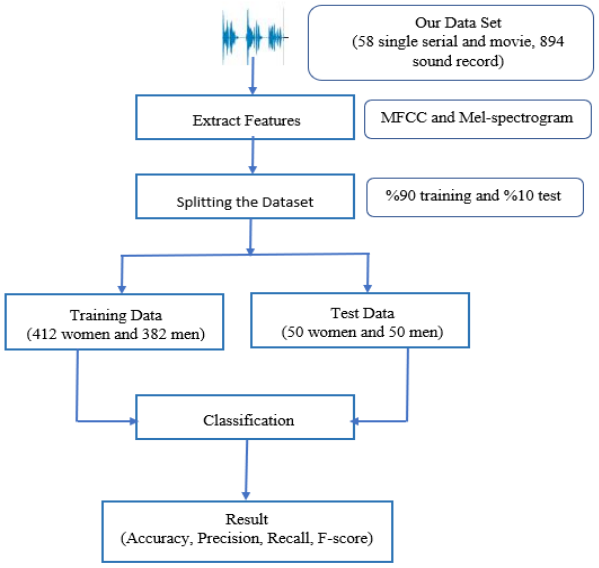

Figure 1. Architectural Structure

3.2. Dataset

The distribution of the data set by gender is listed in Table 1.In this study, the data set we prepared ourselves was used. Each recording is 5sec long. As a result, a data set was created in the distribution shown in the table below. Then the data set is divided into training and test data as 90% - 10%.

Women	**Men**
462	432

Table 1. Gender Distribution of Data Set

3.3. Feature Extraction

In this way, after the data set was created, the process of extracting the features was continued. Here, MFCC and spectrogram, which are among the most preferred features in gender recognition among the features of speech, were used as features. The reason for using MFCC is that it can mimic the frequency selectivity of the human ear and provides values that can distinguish speakers in a good way [15]. Figure 2 shows the MFCC feature extraction steps.

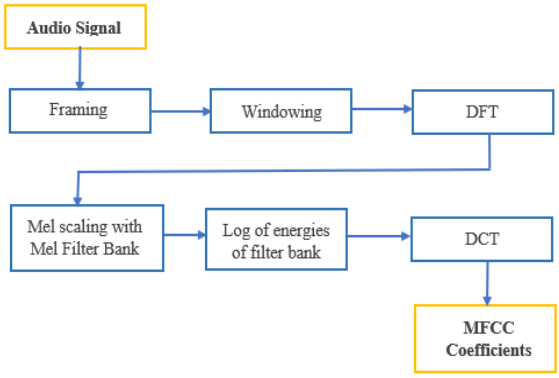

Figure 2. MFCC feature inference steps

Framing: Speech marks should be broken down into small pieces. Because if FFT is calculated along the entire mark, it can create losses in the retention of spectral information of different phenomena. Here the FFT of the respective framework is calculated [15]. Frame length must be between 20-30 ms.

Windowing: It is called the process to prevent discontinuity at the beginning and end of each frame [29]. The commonly used window structure is Hamming. Its function is as in formula 1.

$$w(n) = 0.54 - 0.46\,cos\left[\frac{2\pi n}{(N-1)}\right] \qquad N-1 \geq n \geq 0 \quad (1)$$

Fast Fourier Transform (FFT): FFT is applied to translate each frame with N samples from time zone to frequency region. The function is described in formula 2 [29].

$$X_n = \sum_{k=0}^{N-1} x_k e^{\frac{-2\pi jkn}{N}} \qquad ,n = 0,1,2,\ldots N-1 \qquad (2)$$

Pre-highlighting: In famous sounds, a total of -6 dB/octave attenuation occurs in the vocal path. In famous sounds, a first-order high-pass filter is usually used to eliminate this weakening. In the case of consonant sounds, there is no need for pre-highlighting. Because their spectrum is smooth [15].

Mel-scale Filter Sequences: In this step, the logarithm of the resulting sign is taken by passing through triangular filter arrays arranged according to one of the signal frequency scales (Bark, Mel, Linear, ERB) whose spectrum was previously taken [30].

Discrete Cosine Transform: Here, the last sign is applied to the discrete cosine transform to obtain attribute vectors known as cepstrum coefficients [30].

Spectrogram is one of the basic methods used in the creation of the time-frequency structure of the signal. One of the main uses of spectrograms is sound analysis. The [18] display of signals in the time-frequency field provides many benefits in terms of sound classification. First, the time-frequency conversion is reversible. This

representation covers all the features of the voice data. What is more important is that the time-frequency representation of sound includes many different characteristic features of the sound signal. In this way, distinctive data about sound can be obtained from Spectro-temporal shapes and classification results with a high success rate can be obtained. Then, during the classification process, MFCC and spectrogram were first tested by logistic regression as a single. Then, the results with logistic regression, Linear Discriminant Analysis, Ada Boost, Gradient Boosting, Extra Trees, XGB, Random Forest and SVC classifiers in the hybrid model including MFCC and spectrogram were obtained and compared.

3.4. Evaluation Metrics

Criteria's such as accuracy, precision and recall, which are frequently used in the literature, were used to compare the performance of classification metrics. Figure 3 shows the Confusion Matrix.

		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 3. Confusion Matrix

True Positive (TP): It is an outcome where the model correctly estimates the positive class. [32]
True Negative (TN): It is an outcome where the model correctly estimates the negative class.
False Positive (FP): It is an outcome where the model incorrectly estimates the positive class.
False Negative (FN): It is an outcome where the model incorrectly estimates the negative class.
Formulas used to calculate accuracy, precision and recall values are shown in Equation (3), (4) and (5).

$$Accuracy = \frac{TN+TP}{TN+TP+FP+FN} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

We trained our models on a machine with 8 cores Intel i7 processor with 32 GB of RAM installed.

4. Experimental Results

In this study, a system that determines the gender of the speaker independent of the speech text is proposed. Our self-prepared data set consists of 462 female and 432 male speakers. Turkish TV series and movies were used as sources. Structure is independent of both speaker and text. With the MFCC and spectrogram features extracted from the speaker data of certain gender, the model was first tested as a single and then as a hybrid in the classification process. The data set is divided into 90%-10% training and testing. In the confusion matrix, 1 corresponds to the female gender and 0 to the male gender.

The Confusion Matrix obtained after classification by Logistic regression using MFCC is shown in Figure 4 and Figure 5.

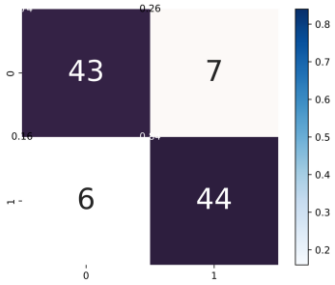

Figure 4. Confusion Matrix (MFCC)

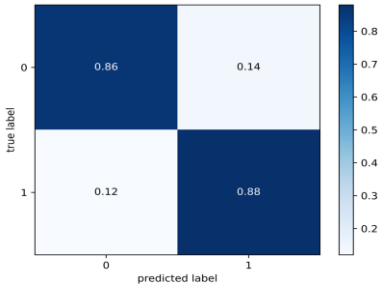

Figure 5. Confusion Matrix MFCC (Percentage Value)

When the results are examined according to Figures 4 and 5, while the male voice recognition rate is 86%, the female voice recognition rate is 88%. Looking at the overall result, 87% recognition rate was achieved by using the MFCC feature and logistic regression classifier.
The Confusion Matrix obtained after classification by Logistic regression using spectrogram is shown in Figure 6 and Figure 7.

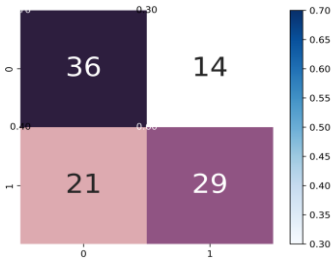

Figure 6. Confusion Matrix (Spectrogram)

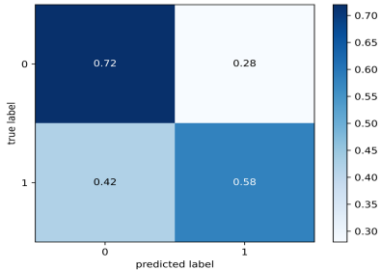

Figure 7. Conf. Matrix Spectrogram (Percentage Value)

When the results are examined according to Figures 6 and 7, while the male voice recognition rate is 72%, the female voice recognition rate is 58%. Looking at the overall result, 65% recognition rate was achieved by using the spectrogram feature and logistic regression classifier.

Finally, both features were incorporated into the process together and tested with different machine learning algorithms. The results obtained are shown in Table 2.

Algorithm Type	Accuracy	Precision	Recall	F1 Measure
Logistic Regression	89%	86.8%	92%	89.3%
Linear Discriminant Analysis	82%	79.6%	86%	82.7%
Ada Boost Classifier	80%	81.2%	78%	79.6%
Gradient Boosting Classifier	82%	79.6%	86%	82.7%
Extra Trees Classifier	80%	75.9%	88%	81.5%
XGB Classifier	82%	77.6%	90%	83.3%
Random Forest Classifier	84%	80.4%	90%	84.9%
SVC	86%	82.1%	92%	86.8%

Table 2. Results and Comparison of All Algorithms

According to Table 2, the best result was obtained by logistic regression. The recognition rate is 89%.

In addition, the Confusion Matrix obtained after the classification of the hybrid model by logistic regression is shown in Figure 8 and Figure 9.

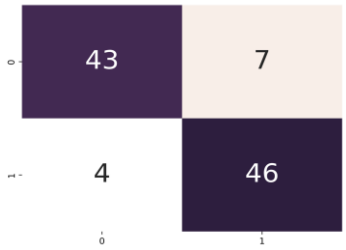

Figure 8. Confusion Matrix (Hybrid Model)

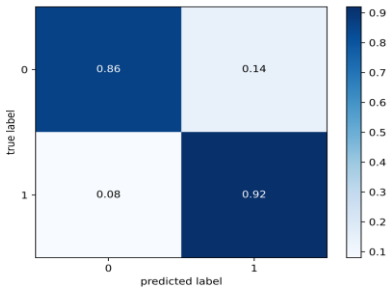

Figure 9. Conf. Matrix Hybrid Model (Percentage Value)

Figure 10 shows the distributions of the cross-validation values. It is seen that the accuracy does not change much when the proposed method is performed with k-fold cross validation.

In classification problems, the data set is usually divided into two as training and testing. However, in the separation of training and test data, if there is no random distribution, it may cause over fitting problem. In order to solve this problem, cross validation is used. [33] This allows it to be understood whether the success achieved in the model is random or not. In this study, the fact that the values do not change much with cross validation shows that the randomness of the data set and the model created are successful.

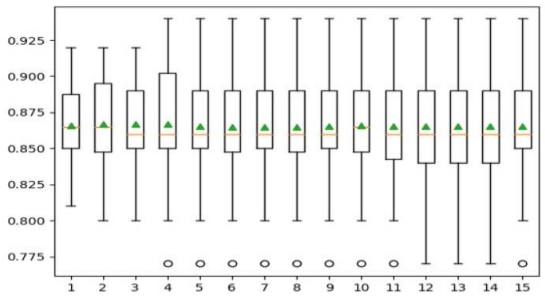

Figure 10. Box Plot of Means of Proposed Method over k-Fold Cross-Validation

5. Discussion

The list of results obtained when our proposed model is compared with similar studies in this area is given in Table 3.

Work	Dataset	Best Accuracy
Alkhawaldeh (2019) [16]	Custom	99.7% with SMO (recall)
Taspinar et al (2020) [19]	Acoustic Analysis of the Dataset	97.9%
Levitan et al (2016) [14]	HMIHY aGender	95%
Sadek et al (2012) [20]	Custom	80%
Zhong et al (2019) [21]	Custom	99.9%
Yücesoy et al (2013) [22]	TIMIT	99%
Uddin et al (2021) [23]	TIMIT, BGC RAVDESS	93.01%
Jena et al (2020) [24]	Custom	87.5%
Yücesoy (2016) et al [25]	aGender	92.42%
Thangaiyan et al (2017) [26]	Custom	76.75%
Kiani et al (2017) [27]	Custom	90%
Yücesoy (2020) [28]	aGender	60.95%
Proposed Method	**Original own dataset**	**89%**

Table 3. Comparison with Previous Studies

In this study, unlike many previous studies, a data set independent of speaker and content was used. Although the speaker and content are independent, the high recognition rate is a point that makes the study successful compared to previous studies. In addition, the data set consists of Turkish speech sentences. In general, data sets are prepared in a laboratory environment and contain foreign languages (English, German). The fact that it is based on Turkish is another difference point. However, recognition rates are higher in studies with ready-made data sets than in this study. Although this study was not applied, it was observed that the recognition rate increased in the studies where attribute selection was applied. MFCC and deltas are most preferred in the attribute. In previous studies, SVM, ANN, logistic regression, linear regression, random forest, AdaBoost, Decision Logic, CNN, KNN, Genetic Algorithm and DNN were preferred in classification processes. In this study, the results of many classifiers were compared and the best result was obtained by logistic regression.

6. Conclusion

Within the scope of this study, gender estimation was made through the dataset we prepared independently of the speaker and text. 3 different models were tested. In the first model, only MFCC was used as a feature and 87% success was achieved with the logistic regression classifier. In this model, the female voice recognition rate is higher than the male voice recognition rate. In the second model, spectrogram was used as a feature and logistic regression was used as a classifier. In this model, the recognition rate remained lower with 65%. In this model, where spectrogram was used as a feature, the female voice recognition rate was very low and reduced the overall average (female voice 58% and male voice 72%). The third model, in namely the hybrid structure where MFCC and spectrogram are used together, has achieved the best result.

Different machine learning algorithms were compared in this model and the most successful result was obtained in logistic regression. The main reason why the best result is obtained by logistic regression in all 3 models is that, unlike other ML algorithms, logistic regression is based on a statistical approach, has different decision limits with different weights close to the optimal point instead of the best value and was able to achieve more successful results in data sets consisting of smaller data can be said.

In the future, further studies will be carried out on expanding the dataset and evaluating the age status of individuals along with gender. In addition, the level of success in the results will be increased by using deep learning models.

Author's Note

Abstract version of this paper was presented at International Conference on Engineering Technologies (ICENTE'22), 17-19 November 2022, Konya, Turkey.

References

[1] R. S. Arslan and N. Barışçı, "Development of output correction methodology for long short term memory-based speech recognition," *Sustainability, ,* cilt 11(15), 2019.

[2] R. S. Arslan and N. Barışçı, "A detailed survey of Turkish automatic speech recognition," *Turkish journal of electrical engineering and computer science*, pp. 3253-3269, 2020.

[3] H. Erokyar, "Age and Gender Recognition for Speech Applications based on Support Vector Machines," Florida, 2014.

[4] A. Oğuz, "Ses Sinyallerinden Yaş Grubu ve Cinsiyet Bilgisinin Tahmin Edilmesi," Siirt, 2018.

[5] S. Hızlısoy and Z. Tüfekçi, "Noise robust speech recogniton using parallel model compensation and voice activity detection methods," *2015 5th international conference on electronics, devices, systems, and applications(ICEDSA)*, pp. 1-4, 2016.

[6] S. Hızlısoy and R. S. Arslan, "Text independent speaker recognition based on MFCC and machine learning," *Selcuk University Journal of Engineering Sciences*, no. 20(3), pp. 73-78, 2021.

[7] S. Hızlısoy, S. Yıldırım and Z. Tüfekçi, "Music emotional recognition using convolutional long short term memory deep neural networks," *Engineering science and technology, an international journal*, no. 24(3), pp. 760-767, 2021.

[8] A. Tursunov, Mustaqeem, J. Y. Choeh and S. Kwon, "Age and Gender Recognition Using a Convolutional Neural Network with a Specially Designed Multi-Attention Module through Speech Spectrograms," *Sensors*, 09 2021.

[9] E. Çolakoğlu, S. Hızlısoy ve A. Recep Sinan, "Konuşmadan duygu tanıma üzerine detaylı bir inceleme: özellikler ve sınıflandırma metodları," *Avrupa bilim ve teknoloji dergisi*, pp. 471-483, 2021.

[10] A. Recep Sinan ve N. Barışçı, "Farklı optimizasyon tekniklerinin bağlantıcı zamansal sınıflandırma kullanılan uçtan uca Türkçe konuşma tanıma sistemlerine etkisi," *2nd International Symposium on Multidisciplinary Studies and Innovative Technologies(ISMSIT)*, pp. 19-21, 10 2018.

[11] A. Pahwa and G. Aggarwal, "Speech Feature Extraction for Gender Recognition," *I.J. Image, Graphics and Signal Processing,*, pp. 17-25, 9 2016.

[12] F. Ertam, "An effective gender recognition approach using voice data via deeper LSTM networks," *Applied Acoustics*, pp. 351-358, 08 2019.

[13] A. Oğuz, "Ses sinyallerinden yaş grubu ve cinsiyet bilgisinin tahmin edilmesi" Siirt Üniversitei Fen Bilimleri Enstitüsü, Siirt, 2018.

[14] S. Levitan, T. Mishra and S. Bangalore, "Automatic identification of gender from speech," *Speech Prosody 2016*, Boston, USA, 2016.

[15] Ö. Eskidere ve F. Ertaş, "Mel Frekansı Kepstrum Katsayılarındaki Değişimlerin Konuşmacı Tanımaya Etkisi," *Uludağ Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi Cilt 14, Sayı 2,* 2009.

[16] R. S. Alkhawaldeh, "DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network," *Scientific Programming*, pp. 1-12, 12 2019.

[17] A. Alan ve M. Karabatak, "Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin

Değerlendirilmesi," *Fırat Üniversitesi Müh. Bil. Dergisi* , cilt 32(2), no. 531-540, pp. 531-540, 8 2020.

[18] M. M. Nasef, A. M. Sauber and . M. M. Nabil, "Voice gender recognition under unconstrained environments using self-attention," *Applied Acoustics,* 11 2020.

[19] Y. S. Taspinar, M. M. Saritas, İ. Cinar and M. Koklu, "Gender Determination Using Voice Data," *International Journal of Applied Mathematics, Electronics and Computers,* 11 2020.

[20] A. Sadek, I. Shariful and H. Alamgir , "Gender Recognition System Using Speech Signal," *International Journal of Computer Science, Engineering and Information Technology,* pp. Vol.2, No.1, 02 2012.

[21] B. Zhong, Y. Liang, J. Wu, B. Quan, C. Li, W. Wang, J. Zhang and Z. Li, "Gender Recognition of Speech based on Decision Tree Model," %1 içinde *Proceedings of the 3rd International Conference on Computer Engineering, Information Science & Application Technology*, Chongqing, China, 2019.

[22] E. Yücesoy and V. V. Nabiyev, "Gender Identification Of A Speaker From Voice Source," %1 içinde *21st Signal Processing and Communications Applications Conference*, Haspolat, Turkey, 2013.

[23] M. A. Uddin, R. K. Pathan, H. Sayem and M. Biswas, "Gender and region detection from human voice using the three-layer feature extraction method with 1D CNN," *Journal of Information and Telecommunication,* pp. 27-42, 08 2021.

[24] B. Jena, A. Mohanty and S. K. Mohanty, "Gender Recognition of Speech Signal using KNN and SVM," %1 içinde *International Conference on IoT based Control Networks and Intelligent Systems*, Kottayam, Kerala,India, 2020.

[25] E. Yücesoy ve V. V. Nabiyev, "Konuşmacı yaş ve cinsiyetinin GKM süpervektörlerine dayalı bir DVM sınıflandırıcısı ile belirlenmesi," *Journal of the Faculty of Engineering and Architecture of Gazi University*, 09 2016.

[26] J. Thangaiyan, K. Vinothkumar and A. Vijayaselvi, "Automatic Gender Identification in Speech Recognition by Genetic Algorithm," *Applied Mathematics & Information Sciences,* pp. 907-913, 05 2017.

[27] F. Kiani, M. A. Kutlugün ve M. Y. Çakır, "Derin Sinir Ağları ile Konuşma Tespiti ve Cinsiyet Tahmini," %1 içinde *22. Türkiye'de Internet Konferansı*, İstanbul, 2017.

[28] E. Yücesoy, "Konuşmacının Yaş ve Cinsiyetine Göre Sınıflandırılmasında DVM Çekirdeğinin Etkisi," *El-Cezerî Fen ve Mühendislik Dergisi,* pp. 970-982, 05 2020.

[29] S. KARASARTOVA, "Metinden Bağımsız Konuşmacı Tanıma Sistemlerinin İncelenmesi ve Gerçekleştirilmesi," Ankara, 2011.

[30] Ö. Eskidere and F. Ertaş, "The Effects of Filter Frequency Scale Variability On Speaker Identification Performance," *Journal of Engineering and Natural Sciences,* pp. 197-207, 09 2009.

[31] İ. TÜRKER, "Ses Sinyallerinin Graf Tabanlı Temsillerinin Yapay Zekâ Yöntemleri İle Sınıflandırılması," Karabük, 2022.

[32] "wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Confusion_matrix.

[33] G. Öğündür, "Model Seçimi-K Fold Cross Validation," 13 01 2020. [Online]. Available: https://medium.com/@gulcanogundur/model-se%C3%A7imi-k-fold-cross-validation-4635b61f143c. [Access: 12 2022].

[34] Ö. Eskidere and F. Ertaş, "The effects of filter frequency scale variability on speaker identification performance" *Journal of Engineering and Natural Sciences Mühendislik ve Fen Bilimleri Dergisi,* 9 2009.

[35] S. Aksu, "Ses sinyallerinin graf tabanlı temsillerinin yapay zeka yöntemleri ile sınıflandırılması " Karabük Üniversitesi, Karabük, 2022.