Malatya Turgut Özal University
Journal of Engineering and Natural Sciences

https://dergipark.org.tr/tr/pub/naturengs

Research Article

Gender Detection by Acoustic Characteristics of Sound with Machine Learning Algorithms

Hamit Mızrak[1], Serpil Aslan[2*]
[1]*Department of Informatics, Malatya Turgut Ozal University, Malatya, Turkey.*
[2]*Department of Software Engineering, Malatya Turgut Ozal University, Malatya, Turkey.*

ARTICLE INFO

Received: 2023-05-09

Accepted: 2023-06-02

DOI: 10.46572/naturengs.1294655

ABSTRACT

Sound has been studied in almost every field since the existence of humans, and sound science branches have emerged due to its increasing importance day by day. Sound has been studied in many fields from the past to the present, and it has become an essential factor in people's understanding of each other. It is possible to determine the voice of individuals with the help of computers in the digitalized world to determine the speech acoustics and the gender of that voice and to determine it with the prediction algorithms in machine learning by utilizing the characteristics of the voice's characteristic metric performance criteria. In this study, six different machine learning algorithms were used. By comparing these machine learning algorithms, prediction results were obtained to determine the best prediction result and the gender difference by using the voice characteristics of individuals. When the models' performances were evaluated, The Support Vector Machine (SVM) algorithm produced the gender prediction models with the lowest accuracy performance, 48.85%, and the Decision Tree (DT) algorithm produced the highest accuracy performance, 98.28%.

Keywords: *Sound Acoustics, Sound Characteristics, Gender Detection, Machine learning, Hertz*

Introduction

Sound, an indispensable part of people's lives, has passed through different stages from the past to the present. While expressing ourselves, we convey what we want to do and our thoughts to the other party through our voices. The sound wave is an acoustic pressure wave formed due to the voluntary movements of the anatomical structures that make up the sound production system. The main parts of this system are the lungs, trachea, larynx, throat, oral cavity, and nasal cavity. Technically, the throat and oral cavity are defined as the 'vocal tract' [1]. In short, sound is the type of energy produced as a result of vibrations of atomic molecules. We call the vibrational power in sound decibels (dB). In daily life, a person communicates in 45-60 decibels. Another factor that creates sound is frequency. In short, frequency is the number of vibrations per second, and its unit is Hertz. Examining the topographic structure of the sound, sound analysts and sound forensics (audio forensics) are considered electronic evidence, and this branch investigates the recorded data to reveal the crimes.

Speech is the most accessible and most natural form of communication between people. During speech, not only words are transmitted to the listener. At the same time, information about the speaker, such as identity, age, gender, and mood, is transmitted. Intensive studies are carried out to establish this communication between people with computers. The use of voice, especially in biometric systems, provides both cost and ease of use advantages. For gender estimation with speech, speech should be prepared for this process and included in the recognition process as computer-aided. For this purpose, speech needs to be converted into exemplary signals utilizing a microphone and labeled (for example, as sounds, phonemes, words, or phrases) and transformed into forms expressed with parametric structures or plain models with classification techniques that will form the basis of recognition processes [2,3]. In almost all of these studies, representation data to be obtained from raw audio data to represent high-quality audio is sought.

Gender detection is the techniques used to decide the gender category by processing the speaker's voice signals. A recorded speech's signals can be given acoustic characteristics like duration, intensity, frequency, and filtering [4]. Many application areas, including emotion recognition, human-machine interaction, gender-based call sequencing, automatic greetings, and audio/video categorization, use gender detection methods. The gender of the voice is of such importance. Additionally, facts and elements that will facilitate the differentiation between individuals such as ear, face region, fingerprint and eye iris pattern, are also used for gender determination [5]. It is essential to increase human impact and efficiency in computer interaction by using human voice and

*Corresponding author: serpil.aslan@ozal.edu.tr
ORCID : 0000-0001-8009-063X

describing human knowledge. There is a growing need to know not only the information a user is speaking but also how it is spoken and its meaning. Many parameters affect the speech feature. The most well-known are gender, age, health, language, dialect, accent, emotional state, and the speaker's attention while speaking [1].

In this article, machine learning algorithms are used for gender detection problems with voice signals, which are used in many different fields and become a popular topic. By using certain features of the voice, performance comparisons of different machine learning algorithms in detecting gender from voice signals were made.

Related Works

Estimating gender and age from sound with machine learning methods; Today, it has increasingly widespread usage areas and has been investigated and analyzed by many research groups with different methods before. While the effects of speakers on speech characteristics have been investigated since the 1950s [6], natural systems that attempt to estimate gender and age from the human voice have been the subject of severe studies since the early 2000s. The last ten years have significantly increased data mining and machine learning techniques for vocal gender recognition. These prediction models can identify a person's gender based on various characteristics, including the vocal cords' length and speech pattern.

Maka et al.'s [7] experiments on the issue of gender inference through sound were carried out in various acoustic settings, including indoor and outdoor auditory scenes. 630 speakers made up the dataset used by 438 men and 192 women. Using the machine learning algorithms SVM, KNN (K-Nearest Neighbor), and GMM (Gaussian Mixture Model), Sedaaghi [8] targeted age and gender discrimination. The authors used two different sound databases in his studies. The first is the Danish emotion database DES, and the other is the English language voice database ELSDSR. Pahwa et al. [9] proposed a classification model using SVM and artificial neural network classifiers by extracting the Mel coefficient and first and second derivatives of speech characteristics to determine gender. They tested their proposed method using speech samples from 46 speakers.

In their study, Nguyen et al. [10] proposed a machine learning-based model for identifying age, gender, and accent. The proposed model used VQ (Vector Quantization), GMM, and SVM as classifiers. An Australian speech database with 108 speakers and 200 phrases per speaker was used to test the system. In their study, Pribil et al. [10] suggested a two-level GMM algorithm to identify age and gender. The proposed classifiers were first tested in the Czech and Slovak languages, identifying four different age groups—child, young adult, adult, and old—and the gender of all but children's voices. Then, using MFCC and LPCC (Linear Predictive Cepstrum Coefficients) speech features as well as SVM and DT (Decision Tree) classifiers, Lee et al. [11] created a system for gender and age group recognition. In the study, whose age categories were children and adults, 7 males and 7 female voices were used. Kabul et al. [12] proposed a method that uses CNN (Convolution Neural Network) followed by MLP (Multi-Layer Perceptron) in raw speech signals and uses a Softmax function for classification.

Materials and Methods

Data Collection

In this study, a dataset downloaded from Kaggle, an open-access site where datasets can be shared and downloaded, is used in a versatile manner by data scientists. This database was developed to classify voices as male or female based on speech and voice acoustics. 3,168 recorded voice samples from male and female speakers make up the dataset. In order to realize the experimental results, the training set was divided into 20% test and 80% training dataset.

Machine Learning Algorithms

- **SVM**

Support Vector Machines [13] is a high-performing learning algorithm built with a straightforward idea. Support Vector Machines, based on Statistical Learning theories, are theoretically based on support vectors that best linearly separate samples from two classes. Therefore, it aims to separate the data into two categorized classes appropriately and draws lines to separate the points on the plane. SVM uses the kernel function to convert input variables implicitly. Thanks to kernel functions, SVM can divide nonlinearly separable support vectors using a linear plane. To maximize performance for the majority of applications, the proper kernel function and kernel width must be chosen.

- **KNN**

It is used in supervised learning in classification and solving regression problems. It was proposed by Cover and Hart [14] in 1967. The calculation is made according to the existing data of the new data to be added to the data set, and close neighborhoods are checked in k units. Distance criterion functions such as Euclid, Manhattan, and Minkowski are used for distance. Finding the predetermined number of training examples that are most similar to the new point and inferring the label from them are the objectives of KNN. In the case of radius-based neighbor learning, the number of samples can either vary depending on the local density of points or be a user-defined constant (k-nearest neighbor learning). Despite being straightforward, nearest neighbors have been used to solve various classification and regression issues, including scenes from satellite images and handwritten numbers. Additionally, it is non-parametric and frequently works well in classification scenarios where the decision boundary is erratic.

- **DT**

A classification algorithm called Decision Tree builds a tree-like classification model from decision and leaf nodes. Using specific decision rules, it divides very large datasets into smaller subsets. It is a tree-based algorithm used to solve classification and regression problems. A decision tree consists of roots, nodes, and leaves. The first cell is called the root. Classification is made as No or Yes according to the status of each observation. Just below the stem cell are nodes, and just below the nodes are leaves.

- **LR**

A dataset with one or more independent variables that affect the outcome is examined using the LR statistical method [12]. The outcome is assessed using a binary variable. LR is a binary dependent variable, meaning it only contains information that can be expressed as 1 (true, etc.) or 0 (false, etc.). LR aims to identify the model that best captures the relationship between a set of related independent variables and a bidirectional characteristic.

- **NB**

A well-known statistical learning algorithm called NB [15] has been used as a fundamental classifier for comparison with other algorithms. Since NB presumes that the inputs for a particular class are independent, the class condition provides a fair estimate of the probabilities. Moreover, by finding the conditional marginal densities of classes—which indicate the likelihood that a given sample belongs to one of the potential target classes—NB simplifies the problem of class discrimination. As a result, NB outperforms competing algorithms unless the latter has correlated inputs.

- **RF**

An algorithm improves the classification rate by creating multiple decision trees during the classification process. The decision forest comprises a collection of randomly chosen decision trees. For many datasets, it generates more accurate results than Support Vector Machines. For example, it performs well in datasets with missing data, an uneven distribution, or categorical variables with many variables and class labels.

Experimental Results

In this study, gender prediction was made using voice data using six machine learning algorithms. The dataset we use consists of male and female voice data in 2 different classes taken from the open-access site Kaggle, which also contains datasets. Experiments were conducted on the Google Colab platform, widely used for training deep learning and machine learning models. All experiments are tested on Intel Core i7 Windows 10 operating system 8 GB RAM computer.

The experimental results were examined in light of the evaluation criteria to determine the effectiveness of the machine learning algorithms. Evaluation metrics such as the "Accuracy, Precision, Recall, and F1-score" were frequently used to compare the compared machine learning models' performances. The complexity matrix, a NxN matrix, displays the proportion of accurate and inaccurate predictions of the classification model concerning the data's actual outcomes. N is the number of classes. The most popular metric for assessing the performance of models is the confusion matrix.

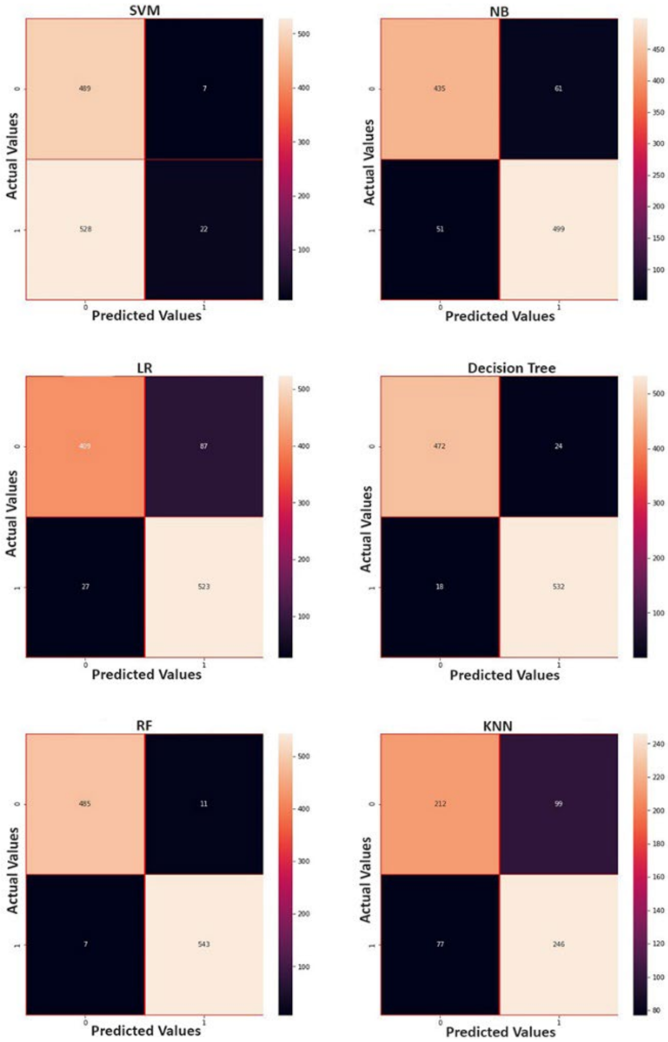

Figure 1. Comparison of Machine Learning Algorithms' Complexity Matrix

Figure 1 represents the complexity matrix comparison of machine learning algorithms. As can be seen in the confusion matrix analysis, the closest estimates to the actual values were obtained by the RF machine learning algorithm according to the results of the independent sampling method.

Table 1. Performance comparison of machine learning algorithms.

	Machine Learning Algorithms					
	SVM	**NB**	**LR**	**DT**	**RF**	**KNN**
Accuracy	0,4885	0,8929	0,8910	0,9637	0,9828	0,7224
Precision	0,9859	0,8770	0,8246	0,9536	0,9778	0,6817
F1 Score	0,6464	0,8859	%8777	0,9614	0,9818	0,7067
Recall	0,9858	0,8770	%8245	0,9536	0,9778	0,6816

From the sound samples, 0hz-280hz (which is the range of human vocal voice). There is more than one performance criterion from the sound analysis criteria. More than one metric criterion is used in the gender classification of acoustic data. Each sound sample is stored as a .wav file in non-volatile memory (hard disk). Afterward, see wave and tuneR are pre-functioned for acoustic analysis using the particular function in the R package, and the resulting data is saved as a CSV file. More than one machine learning from Machine Learning algorithm was used from this dataset. Support Vector Machine(SVM) success rate is 48.85%), Naive Bayes success rate is 89.29%, Logistic Regression success rate is 89.10%, Decision Tree success rate is 96.37%, Random Forrest success rate is 98.28%, KNN success rate is 72.24%. Found as a result of machine learning. Predictions are made on each machine learning dataset.

The best result in the estimations is the algorithm that predicts whether individuals are male or female from voice metric measurements with a random Forrest success rate of 98.28% machine learning.

Conclusions

This study aims to predict the characteristic criteria of the voice with the help of general features of the voice, such as Hertz, dB, spectral entropy, spectral flatness, etc., and to predict whether the voice is male or female with the help of machine learning algorithms. In the past, we used it in every field, from sound theater architecture in the Ancient Age to today. Training and development of the voice is a phenomenon that takes time and cost. Most newscasters, voice artists, teachers, orators, etc., must use their voices accurately and in moderation. Using these structural components of sound and today's technologies can be used in many interdisciplinary fields that will shed light on our future with metric value measurements of the human voice.

In increasing the recognition accuracy of speech signals and emotion systems, it is sometimes possible that people's voices may not be understood with absolute truth over time, in our emotional times, or the echoing moments of the sound resonance of the individual's environment, or when people who have been vocal artists for a long time force their voices too much, or sometimes in the sound distortions brought by old age. We can provide identification from the voice analysis of the individual with machine learning prediction algorithms to determine the voice gender in the voice character of a person from the voice analysis. Gender identification is today's most significant problem: acoustic data, namely pitch, median, frequency, etc. Gender monitoring can be done based on quality. Machine learning shows promising results in solving the abovementioned problem in all research areas. In this study, machine learning was made by using individuals' voices in speech acoustics, understanding whether that voice is male or female. It consists of 3168 recorded sound samples from the data set. More than one machine learning from Machine Learning algorithm was used from this dataset. It is aimed at machine learning that gives the best results from the algorithms. Predictions are made on each machine learning dataset. Success prediction results in a machine learning DT success rate 96.37%, SVM, KNN success rate 72.24%, LR success rate 89.10%, NB success rate 89.29% and RF success rate % As seen in the 98.28 results, RF was obtained among the best performance machine learning methods. RF accuracy gave the best result in estimations, 0.9828 estimation results.

Acknowledgments

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Competing Interest

There is no conflict of interest.

Author Contribution

All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

References

[1] Chaudhari, S. J., & Kagalkar, R. M. (2015). Automatic speaker age estimation and gender dependent emotion recognition. International Journal of Computer Applications, 117(17), 5-10.

[2] Eckert, M. A., Matthews, L. J., & Dubno, J. R. (2017). Self-assessed hearing handicap in older adults with poorer-than-predicted speech recognition in noise. Journal of Speech, Language, and Hearing Research, 60(1), 251-262.

[3] Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017, May). An all-in-one convolutional neural network for face analysis. In 2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017) (pp. 17-24). IEEE.

[4] Gamit, M. R., Dhameliya, K., & Bhatt, N. S. (2015). Classification techniques for speech recognition: a review. International Journal of Emerging Technology and Advanced Engineering, 5(2), 58-63.

[5] Karasulu, B., Yücalar, F., & Borandağ, E. (2022). A hybrid approach based on deep learning for gender recognition using human ear images. Journal of the Faculty of Engineering and Architecture of Gazi University, 37(3), 1579-1594.

[6] Mysak, E. D. (1959). Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research*, *2*(1), 46-54.

[7] Maka, T., & Dziurzanski, P. (2014, April). An analysis of the influence of acoustical adverse conditions on speaker gender identification. In *XXII Annual Pacific Voice Conference (PVC)* (pp. 1-4). IEEE.

[8] Sedaghi, M. (2009). A comparative study of gender and age classification in speech signals. Iranian Journal of Electrical & Electronic Engineering, 5(1), 1-12.

[9] Pahwa, A., & Aggarwal, G. (2016). Speech feature extraction for gender recognition. *International Journal of Image, Graphics and Signal Processing*, *8*(9), 17.

[10] La Mura, M., & Lamberti, P. (2020, June). Human-machine interaction personalization: a review on gender and emotion recognition through speech analysis. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT* (pp. 319-323). IEEE.

[11] Lee, M. W., & Kwak, K. C. (2012). Performance comparison of gender and age group recognition for human-robot interaction. *International Journal of Advanced Computer Science and Applications*, *3*(12), 207-211.

[12] Kabil, S. H., Muckenhirn, H., & Magimai-Doss, M. (2018, September). On Learning to Identify Genders from Raw Speech Signal Using CNNs. In *Interspeech,* 287, 291.

[13] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.

[14] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27.

[15] Michie, D., Spiegelhalter, D. J., Taylor, C. C., & Campbell, J. (Eds.). (1995). *Machine learning, neural and statistical classification*. Ellis Horwood.