Gender Detection through Speech Processing

AishwaryaMala GM[1], *ChristopherB*[2], *Dharanish NH*[1]

[1]Electrical Engineering, University of Twente
[2]Embedded Systems, University of Twente
Speech Processing Project - Group 7

Abstract

This paper presents models that can recognize the gender of an individual by analyzing their recorded speech signal. The vocal folds are an important aspect of the human voice. The length of the vocal cord is gender-dependent. Males tend to have longer vocal folds than females and in turn, males have a heavier voice with more intensity when compared to females. Based on this idea, a gender detection system is proposed in this paper. The approach presented is very simple. The process begins with the extraction of features from the speech signal. These features are then passed as input to the Machine Learning (ML) model which classifies the signal as either a voice that belongs to a male or a female. The database used during the work consists of both recorded as well as open-source audio clips. The acoustic features of the speech signal were extracted using the MFCC algorithm. We have used 12 MFCC coefficients as features to train our Machine Learning models to detect the gender of the speaker of the speech signal. In our work, we have also aimed at comparing the performances of different ML model on gender detection. Hence, we trained a Support Vector Machine (SVM), Neural Network (NN) and a Gaussian Mixture Model (GMM) on our training data and compared their performances on our test data. We were able to achieve a maximum accuracy of 95% using NN model on our test data. SVM and GMM models gave an accuracy of 90% on our testing data.

Index Terms: Gender recognition, SVM, GMM, Neural Networks, MFCC

1. Introduction

According to Oxford dictionary speech is “the expression of or the ability to express thoughts and feelings by articulate sounds”. Other than expressing the thoughts and feelings, speech also has the ability to express other paralinguistic characters of the individual. We believe that gender is one such feature that could be derived from the speech signal of a person.

Voice related bio-metric security systems are used for authentication, surveillance and security. Including a gender detection algorithm will help limit the search for a person’s identification by half. Human-Computer Interaction (HCI) is a technology that is focused on the interface between human and computers. Some HCI applications are developed for personal interaction as in medical care. In healthcare systems, gender detection can play a significant role. There are some medical complications that are biased to a particular gender. For example, vocal folds cyst can be seen particularly in female patients. Hence, analyzing the data from the speech signal and identifying the gender of the user can help make these kinds of HCI systems more personalized [7].

Common approaches for gender recognition are based on the analysis of the pitch of the speech. However, gender recognition using a single feature does not offer sufficient accuracy for a large variety of speakers. To capture differences in both time domain and frequency domain, a set of features known as Mel-frequency cepstrum coefficients (MFCC) are used [3]. In this paper, we aim to extract 12 MFCC coefficients from speech, and with the help of Machine learning classifiers SVM (Support Vector Machine), Neural Network and GMM (Gaussian Mixture Models) detect the gender of the person.

2. Related Works

The speech of the human voice consists of many essential paralinguistic features that are used in many voice recognition applications. For gender identification, certain features from the voice signal can be used to develop a model. Alkhawaldeh’s work [1] has demonstrated gender detection technique in three parts, (i) Studying a set of voice features and examining their effects on gender classification techniques, (ii) Use of several ML techniques to determine the gender from the extracted features. (iii) Evaluating and choosing the optimal feature of selection techniques. The efficiency of the model is around 99.7% for two models of deep learning and Support Vector Machine and with feature selection, the efficiency is 100% for SVM techniques.

In the work done by Jamil et.al.[3] the different classifiers like k nearest neighbor, naïve Bayes, multilayer perceptron, random forest and support vector machine are used for the determination of gender from telephonic speech. The MFCC feature is used to determine gender. The experiments were performed to determine the effects of training data size and length of speech streams and tuning of classification performance. The result shows the SVM classifier is the best among others.

Based on our study, we deduced a methodology to design a system that detects the gender of a person using his or her recorded speech signal.

3. Methodology

3.1. Database

The speech database which was used for our experiment was recorded manually with the use of Alesis io4 USB Recording Interface. The iO4 is a compact audio-recording interface for the home, project and portable studio recording setups. By using this device and Praat software [6] which is used for speech analysis we recorded 24 recordings from 8 volunteers, 3 recordings per volunteer. We asked them to sign a consent form to use their voice for this research and provided volunteers with the complete information on the research process before proceeding with the recordings. Each individual was given with one common sentence and 2 sentences of their choice where all the recordings average up to 3.5 seconds of duration. There were equal numbers of male and female speakers. To increase the efficiency of the prediction algorithm we used an open dataset to increase the number of voice recordings which in turn used for training the classifier. As a whole, we had 50 voice recordings of male and female each.

Table 1: *Speech database*

	Male	**Female**
Training	30	30
Validation	10	10
Testing	10	10

3.2. MFCC

The shape of the vocal tract present in humans generates speech along with the help of teeth, tongue, and resonators present inside the mouth. The sound comes out of humans differs with respect to the shape of the vocal tract, so identifying the shape of vocal tract sound can be determined where the shape will be given out in the form of the power spectrum. MFCC is good at picking out those spectrums which will determine the shape of the vocal tract. MFCCs are used for extracting the features from a speech which in turn is used to train the classifier. MFCC has been used widely since its introduction and to calculate MFCC there are a series of steps involved [1].

- Frame the signal into short frames.
- For each frame calculate the power spectrum.
- Apply the mel filter bank to the power spectra, sum the energy in each filter.
- Take the logarithm of all filter bank energies.
- Take the DCT of the log filter bank energies.
- Keep DCT coefficients 13, discard the rest.

In this research, we used the software “Praat” to compute MFCC. A sound wave has been analyzed using the ‘Analyze Spectrum’ option under which ‘To MFCC’ is present which will compute the MFCC coefficients according to our requirement we insert like several coefficients, window length and time step. We used standard window length of 25ms and the number of coefficients to be calculated were taken as 12. MFCCs are computed with all these details inserted and we get 12 coefficients for each window length. For training purposes, we take mean for each coefficient respectively. We labeled the coefficients from male voices with 0 and female voices with 1.

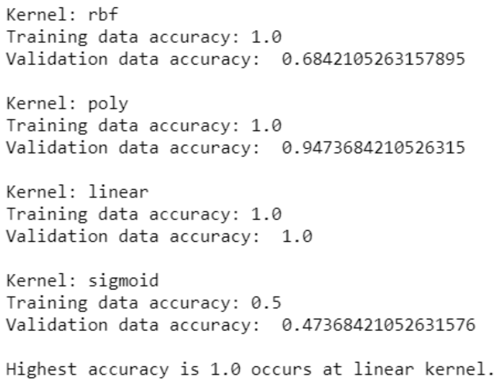
3.3. Training

Using the extracted features, we trained three different classifiers: Support Vector Machine (SVM), Neural Networks (NN) and Gaussian Mixture Model (GMM). We used Machine Learning models from the Scikit-learn Python library for training. The following are the details on how we trained each of the classifiers [2].

3.3.1 Support Vector Machine (SVM)

Support Vector Machine classification algorithm works by identifying the ‘support vectors’ of the data and based on these calculating the decision boundary that guarantees the widest margin between the two classes of data. Support vectors are defined as the subset of the training data that define the position of the decision boundary. These are the outermost data points of the classes. These data points are the most difficult to classify as they lie close to the boundary and to the data points that belong to the adjacent class. SVM algorithm finds the most optimum way to separate the classes by finding the boundary line or hyperplane that separates the support vectors maximally and this decision boundary will hence have the widest margin of separation between the two classes.

SVM algorithm works best on linearly separable data. In the case of complicated data that are non-linearly separable, we use a kernel. The kernel helps map the non-linearly separable data into a higher dimensional space where it becomes possible to find a hyperplane that can classify the data.
Scikit-learn library offers the following pre-defined kernels functions: linear, poly (polynomial), RBF (Radial Basis Function) and sigmoid [2][8]. We trained the SVM classifier with each of the mentioned kernels using the training dataset. The performance of each of the trained SVM model was then evaluated using the validation data. Figure 1 shows accuracies that were obtained on the training data and the validation data for the different kernels used [2].



```
Kernel: rbf
Training data accuracy: 1.0
Validation data accuracy:  0.6842105263157895

Kernel: poly
Training data accuracy: 1.0
Validation data accuracy:  0.9473684210526315

Kernel: linear
Training data accuracy: 1.0
Validation data accuracy:  1.0

Kernel: sigmoid
Training data accuracy: 0.5
Validation data accuracy:  0.47368421052631576

Highest accuracy is 1.0 occurs at linear kernel.
```

Figure 1: SVM Training Results

We obtained maximum accuracy on the linear kernel function. Hence it could be inferred that the data is linearly separable. We chose this SVM classifier that had the maximum accuracy on the validation data as our final SVM model and evaluated its performance on the test data. The results obtained are discussed in the results section of the report.

3.3.2 Neural Networks

Neural networks are machine learning algorithms that are based on the biological concept of a neuron [4].
Figure 2 represents a simple neural network consisting of artificial neurons. It consists of an input layer (i), a hidden layer (h) and the output layer (o). Each layer is assigned an activation function. In the below equation for the output of a neuron, W and X are the weights and the input vector. W_i is the weight that is associated with the input X_i and N is the number of inputs that are

present for the neuron under consideration. σ(y) is the activation function of the layer in which the neuron is present.

The output of a neuron = $\sigma(\sum_{i=0}^{N} w_i x_i)$

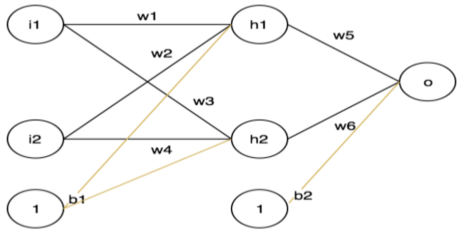


Figure 2: Neural Networks

In each network, the input is passed through the hidden layer to arrive at the output. The output will be a value that is in the range (0,1). This output is classified as class 0 (in our case as male) if the output value of the output neuron is close to zero or it is classified as class 1 (in our case as female) if the output is close to 1. This is process is called thresholding. During training, the weights of the neural network are adjusted to fit the training data in such a way that on thresholding the output the labels that correspond to the data are obtained.

For our problem statement, we have considered a neural network with 12 input neurons to take the 12 MFCC coefficients as input. This is passed through a hidden layer to obtain the output that could be thresholded to classify the speech signal as either belonging to a male or a female. We know that the input layer contains 12 neurons and the output layer contains 1 output neuron but the optimum number of neurons that need to be present in the hidden layer needs to be determined. For this purpose, we trained networks with a varying number of neurons, from 1 to 10 in the hidden layer and evaluated their performances on the validation data. Figure 3 shows a plot of the error rate versus the size of the hidden layer. We obtained the minimum error rate when the size of the hidden layer was 8 and so this network was finalized and it was used to evaluate the performance on the test data.

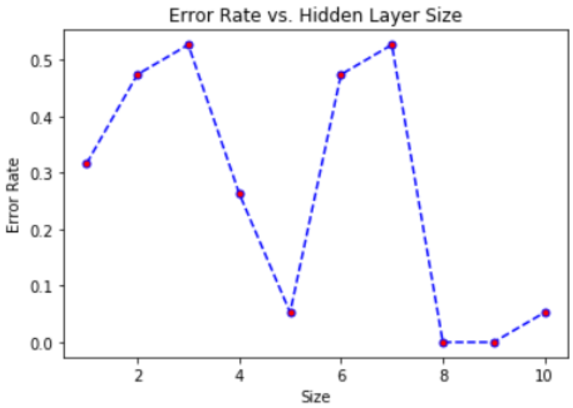


Figure 3: Neural Networks Training Results

3.3.3 Gaussian Mixture Models (GMM)

A Gaussian mixture model is a probabilistic model that assumes the data points are generated from a mixture of a finite number of Gaussian distributions. The parameters of the Gaussian distribution mean (μ), covariance (Σ) and the mixing probability of the distribution (π) are calculated based on the training data using the Expectation-Maximization (EM) algorithm.

A set of Gaussians is fitted on the training features extracted from the female speech signals and another set of Gaussians is fitted on the training data extracted from male speech signals. The log-likelihoods of a given input set of features that are calculated using the Gaussian set fitted on the male and female training data are compared. If the log-likelihood calculated from the set of Gaussians fitted on the female training data is greater than the input is classified as a voice belonging to a female otherwise it is classified as a voice belonging to a male. The number of Gaussians are determined by evaluating the performance of each of the models with 1 to 5 Gaussians on the validation data. Figure 4 shows the plot of the accuracy versus the number of Gaussian components in the network. We obtained maximum accuracy when just a single Gaussian was used to fit the training data. Considering that our data is simple and linearly separable this seems like a sensible outcome [5].

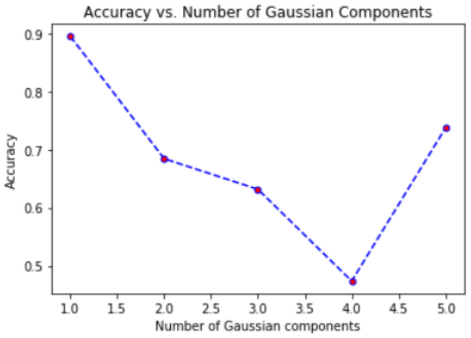


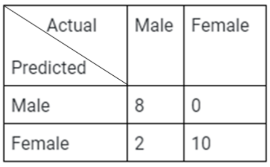
Figure 4: Gaussian Mixture Models Training Results

4. Results

As comparison metrics, we have calculated the accuracy, precision, and recall of each of the classification models on the test data in Table 2. We have also created the confusion matrices in Figure 5 and Figure 6.

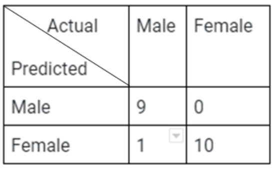
Table 2: *Performance of the ML models on the testing data*

	NN	SVM and GMM
Accuracy	0.95	0.9
Precision	Male: 1.0 Female: 0.91	Male: 1.0 Female: 0.83
Recall	Male: 0.9 Female: 1.0	Male: 0.8 Female: 1.0



Actual Predicted	Male	Female
Male	8	0
Female	2	10

Figure 5: Confusion Matrix SVM and GMM



Actual Predicted	Male	Female
Male	9	0
Female	1	10

Figure 6: Confusion Matrix NN

5. Discussion & Conclusions

SVM and GMM networks yielded almost similar results. Neural networks performed the best on the test data with an accuracy of 95 percent. As per our dataset, we found neural networks to have an edge over SVM and GMM networks when it comes to recognizing the gender to which a given voice belongs to. Although we feel that this conclusion drawn from our results could not be generalized mainly due to the restricted dataset that we worked with. A wider dataset with voices collected from more number of individuals could have helped make the dataset more diverse and the problem-solution more generalizable. In the future, we could explore modeling solutions with lesser computation complexity. For example, using features that are simpler than MFCC to model a solution could be explored. Adding additional criteria like detecting the age group of the individual along with the gender could add complexity to the problem statement but at the same time will also require adequate data that are rightly labeled.

6. References

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

[2] Jmlr.csail.mit.edu. (2020). *Scikit-learn: Machine Learning in Python*. Available at: http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

[3] Ahmad, J., Fiaz, M., Kwon, S.I., Sodanil, M., Vo, B. and Baik, S.W., 2016. Gender identification using mfcc for telephone applications-a comparative study. *arXiv preprint arXiv:1601.01577*.

[4] En.wikipedia.org. (2020). *Artificial neural network*. Available at: https://en.wikipedia.org/wiki/Artificial_neural_network.

[5] Medium. (2020). *Gaussian Mixture Models Explained*. [online] Available at: https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95.

[6] Boersma, Paul & Weenink, David (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.09, retrieved 26 January 2020 from http://www.praat.org/

[7] Alhussein M, Ali Z, Imran M, Abdul W. Automatic gender detection based on characteristics of vocal folds for mobile healthcare system. Mobile Information Systems. 2016;2016.

[8] Medium. (2020). *Understanding Support Vector Machine: Part 2: Kernel Trick; Mercer's Theorem*. [online] Available at: https://towardsdatascience.com/understanding-support-vector-machine-part-2-kernel-trick-mercers-theorem-e1e6848c6c4d