Gender Recognition By Voice

Yuval Mizrahi 209497445
Dvir Barzilay 318751971
Inon Peer 204115570

January 6, 2019

Abstract

In this article, a Recurrent neural network (RNN) deep learning model has been defined to recognize voice gender. The dataset has 3,168 recorded samples of male and female voices. The samples are given by using acoustic analysis. An RNN deep learning algorithm has been applied to recognize gender-specific traits. Our model achieves 95.9% accuracy on the test data set.

Introduction

Often, the human ear can easily determine a person's gender as male or female voice within the first few spoken words. However, designing a computer program to do this may be more complicated.

This article describes the design of a computer program to model the acoustic analysis of voices and speech for determining gender. The model is constructed using 3,168 recorded samples of male and female voices that processed using acoustic analysis and then applied to an artificial intelligence algorithm to learn gender-specific traits.

Related Work

Kory Becker [1] used a frequency-based baseline model, logistic regression model, classification and regression tree (CART) model, random forest model, boosted tree model, Support Vector Machine (SVM) model, XGBoost model, stacked model for recognition of voices data set. Mucahit Buyukyilmaz and Ali Osman Cibikdiken [2] used MLP model for recognition of voices data set.

Table 1: Accuracy of models for recognition voices

Accuracy (%)		
Model	**Train**	**Test**
Frequency-based baseline	61	59
Logistic regression	72	71
CART	81	78
Random forest	100	87
Boosted tree	91	84
SVM	96	85
XGBoost	100	87
Stacked	100	89
MLP	96	96

The Data-set

A training database was required in order to analyze gender by voice and speech. A database was built using 3,168 samples of male and female voices. Each of these samples was labeled by their gender - male or female.

Each voice sample is stored as a WAV file, which is then pre-processed for acoustic analysis using the special function from the WarbleR R package. Special measures 22 acoustic parameters on acoustic signals for which the start and end times are provided.

The output from the pre-processed WAV files was saved into a CSV file, containing 3168 rows and 21 columns (20 columns for each feature and one label column for the classification of male or female).

Acoustic Properties Measured

The following acoustic properties of each voice are measured and included within the CSV:

- **meanfreq:** mean frequency (in kHz)
- **sd:** standard deviation of frequency
- **median:** median frequency (in kHz)
- **Q25:** first quantile (in kHz)
- **Q75:** third quantile (in kHz)
- **IQR:** interquantile range (in kHz)
- **skew:** skewness
- **kurt:** kurtosis
- **sp.ent:** spectral entropy
- **sfm:** spectral flatness
- **mode:** mode frequency
- **centroid:** frequency centroid
- **peakf:** peak frequency (frequency with highest energy)

- **meanfun:** average of fundamental frequency measured across acoustic signal
- **minfun:** minimum fundamental frequency measured across acoustic signal
- **maxfun:** maximum fundamental frequency measured across acoustic signal
- **meandom:** average of dominant frequency measured across acoustic signal
- **mindom:** minimum of dominant frequency measured across acoustic signal
- **maxdom:** maximum of dominant frequency measured across acoustic signal
- **dfrange:** range of dominant frequency measured across acoustic signal
- **modindx:** modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- **label:** male or female

Previous Attempts

In order to determine whether a computer program is actually achieving better results than a non-artificial intelligence-based approach, a baseline model can be employed and used to measure initial accuracy.

In fact, in any model we ran, before the start of the learning process, the accuracy ratio was determined to be approximately 50%. Actually, we can consider it as a basic model with a simple algorithm to determine the gender of a voice. It simply always responds with “male” for a voice, regardless of the acoustic properties.

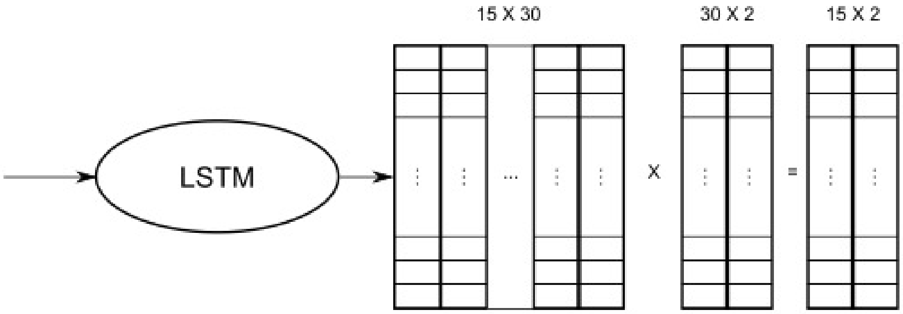
This algorithm results in an accuracy of 50% on both the training and test sets. This makes sense since the data-set is split evenly between male and female voice samples. This is the same accuracy as flipping a coin and guessing randomly. Smarter algorithms can certainly do better than this.

A smarter model can be the Logistic regression model. At first, we take the data-set and changes the parameter *label* from male or female to 0 or 1 respectively. After this, the model mixes the samples. In the second part, the model split the data-set to train and set with a ratio of 70-30. Every part also split to x that contain all the features except *label* and to y that contain only the *label*. After this the model start to train himself with the data that contain in the train data, when he use the logistic function $h(x) = \frac{1}{1+e^{-(xW+b)}}$. The model repeats on the train part 10000 and after this, the model starts the test part when he uses the data that contained in the test data.

Another model is the Multilayer Perceptron model. The model in his first two parts does the same as the Logistic Regression model. After these two parts the model define 4 hidden layers when the first hidden layer has 40 neurons, the second has 40 neurons also and the third has 4 neurons. Also, the model defines an input layer that has 20 neurons as the amount of the features and output layer that has 1 neuron as the amount of our possible output. Every hidden layers have *ReLU* activation function. The output layer uses the logistic function. After this, the model continues as the logistic regression model to train himself with the data that contain in the train data and check the accuracy with the data that contain in the test data.

Project Description

To resolve the problem of this article, to recognize voice gender by using acoustic analysis, chosen to use Recurrent neural network model (RNN) deep learning model. The RNN model implemented with Long short-term memory (LSTM). The model split the data to $data_x$ and $data_y$. Every sample in $data_x$ split to batch in size of 211*15*20*1 and $data_y$ split to batch in size of 211*15*2. Evert batch from $data_x$ enters the LSTM. The output from LSTM is a matrix of size 15*30. In order to make the size of the output matrix and $data_y$ size be equal, the output gets multiplied by a matrix in size of 30*2. After this, the model finds the weights by using *AdamOptimizer* and get from this a matrix in size of 15*2 that be added by the *Bias*. This matrix enters to $Softmax$ that check for each by the $Softmax$ function if the row represents a male or a woman. The $Softmax$ split up a vector that running on it the function $J(W,b) = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(h(x_i))$ (Loss function). The model use this function minimize the error of the train.



Simulation Results

In order to get the best result in all the models, be made changes in every model and model.

In Logistic regression model the optimizer change from GradientDescentOptimizer to AdamOptimizer. The results of the model were *loss* 0.59, *train accuracy* 80.5% and *test accuracy* 80.4%.

In Multilayer Perceptron model the optimizer change from GradientDescentOptimizer to AdamOptimizer and also change the amount of the hidden layers and amount of the neurons to find the amount that gives the best result. The results of the model were *loss* 0.36, *train accuracy* 96.7% and *test accuracy* 94.4%.

In Recurrent neural network model, the optimizer change from GradientDescentOptimizer to AdamOptimizer and also the model use Long-short term memory to get a better result. The results of the model were *loss* 0.28, *train accuracy* 97.4% and *test accuracy* 95.9%.

Table 2: Results

Model	**Loss**	**Train Accuracy**	**Test Accuracy**
Logistic regression	0.59	80.5%	80.4%
Multilayer Perceptron	0.36	96.7%	94.4%
Recurrent neural network	0.28	97.4%	95.9%

Conclusion

The model obtained in the paper shows us that we can use the acoustic properties of the voices and speech to detect the voice gender. RNN has been used to obtain the model for classification from the dataset which has the parameters of voice samples. A larger data set of voice samples can be minimized incorrect classifications from intonation.

References

[1] Kory Becker *"Identifying the Gender of a Voice using Machine Learning"*. 2016, unpublished

[2] Mucahit Buyukyilmaz and Ali Osman Cibikdiken *"Voice Gender Recognition Using Deep Learning"*. 2016, MSOTA