Deep Learning
Gender Recognition By Voice

Max Marmer
Carmel Ron
Odelia Hochman
Efrat Cohen

January 20,2020

Abstract

Human voice is basically sound which is made by humans from their vocal tracts. Voice is made of different constituents and has various characteristics such as frequency, etc.

This paper reflects development of a system using these characteristics which are called Mel-frequency cepstral coefficients (MFCCs) to detect the gender of the speaker.
We have used three models to classify the genders:

- Logistic Regression
- Multilayer Perceptron
- Recurrent neural network

An RNN deep learning algorithm has been applied to recognize gender-specific traits. Our model achieves 97.58% 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. Human voice is easily differentiable by human ears [1]. The speaking mechanism can be divided in parts where the lung gives the air pressure which helps the vocal folds to vibrate, vocal folds use larynx muscle to adjust the pitch and tone. This combination of modulations and articulations is the trait which distinguishes human voice being a female voice or a male voice. An adult male usually has lower pitched voice and larger vocal folds whereas female tend to have high pitched voice and smaller vocal folds.

However, designing a computer program to do this may be more complicated. This article describes the design of a computer program to

model the analysis of short-term power spectrum of the sound and speech for determining gender. The model is constructed using 27,300 recorded samples of male and female voices that processed using Mel Frequency Cepstral Coefficients and then applied to an artificial intelligence algorithm to learn gender-specific traits.

Related work

Becker [2] 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.

According to used models, the results showed in "Table I"

TABLE I. 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

Background

Data set:

- The databases consist of around 1150 utterances [3].
- The databases include US English male (BDL) and female (SLT) speakers (both experienced voice talent) as well as other accented speakers.
- The distributions include 16KHz waveform and simultaneous EGG signals and include Festival CLUNITS based voices. Complete runnable Festival Voices are included with the database distributions, as examples though better voices can be made by improving labelling.
- The database contains 21,000 voice samples in train – 10,500 male 10,500 females, and 6300 voice samples in test - 3150 male 3150 female. The files are 100ms long.

MFCC:

The mel frequency cepstral coefficients (MFCCs) of a signal are a small set of features (usually about 10-20) which concisely describe the overall shape of a spectral envelope. MFCC is the well knowns timbre texture feature or spectrum features which is the highest performing individual feature used in speech recognition.
From theory of speech production, speech is assumed to be convolution of source (air expelled from lungs) and filter (our vocal tract).
The purpose here is to characterize the filter and remove the source part.
The first step in any automatic speech recognition system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc.

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. This page will provide a short tutorial on MFCCs.

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) (click here for a tutorial on cepstrum and LPCCs) and were the main feature type for automatic speech recognition (ASR), especially with HMM.

Previous attempts

We started with a simple model of Logistic regression. We first took wav files of women and men and cut them all out for about a second.
The features we used are features of the MFCC directory. We divided the data into two parts: train, test (70% - 30%) and we 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.
A logistic regression model from the above analysis gives us an accuracy of around 73.4% on the test set. Clearly, it's an improvement over the baseline algorithm, algorithm guessing at probability of 50%.

We move on to the next model which 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 add 4 hidden layers when the first hidden layer has 70 neurons, the second has 100 neurons the third has 60 neurons and the fifth has 30. Also, the model adds an input layer that has 13 neurons as the amount of the features and output layer that has 1 neuron as the amount of our possible output. Every hidden layer has 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. After applying it on our given dataset, it freely gives an accuracy about of 81.9% on the test set. This is again a jump in the improvement of the accuracy of our model.

Project Description

To resolve the problem of this article and recognize voice gender by using Mel-frequency cepstral coefficients, the recurrent neural network (RNN) deep learning model was chosen.

The RNN model implemented with Long short-term memory (LSTM). The model split the data to data_x (Matrix: examples X features) and data_y (Matrix: examples X labels).
All samples in data_x was divided to 15 examples 1400 batches and data_y was divided to 15 examples 420 batches.
21000 train-examples / 15 batch-size = 1400 batches
6300 test-examples / 15 batch-size = 420 batches

Event batch from data_x enters the LSTM. The output from LSTM is a matrix of size 15*30. In order to make the output matrix size 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 GradientDescentOptimizer 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 female.
The Softmax split up a vector that running on it the function:

$$J(W,b) = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(h(x_i))$$

The model uses this function minimize the error of the train.

Simulation result

In order to get the best result in model RNN, we made some changes to the model. In the RNN model we took all the second-length files and cut them to one-tenth of a second.

In RNN model the train step (optimizer) algorithm was changed from AdamOptimizer to GradientDescentOptimizer and also the model uses Long-short term memory to get a better result. The results of the model were loss 0.0008 and test accuracy 91.87% before we change to GradientDescentOptimizer. After change, we get loss 0.0006 and test accuracy 97.58%.

Model	Loss	Test accuracy
Logistic Regression	0.58	73.4
Multilayer Perceptron	0.61	81.9
Recurrent neural network	0.00068	97.58

Conclusions

The purpose of this project is to determine the gender of a person by his voice. In machine learning, there are many different models for solving this problem and they all give different accuracy of the answer. The main goal is to maximize the probability of a correct answer.

As a result of the project, in RNN the accuracy on the training set was 99% and about 97% accuracy on the test set. This is the highest accuracy among all used models that have been tested so far. In addition, it was found that when the average values of all the features of each sample were used, and the entire length of the sample features of the example was not used, the results were better. The model obtained in the article shows us that we can use features of the Mel-frequency cepstral coefficients of the sound and speech to determine the gender of a voice. Thus, RNN was used to obtain a model for classification from a dataset that has voice sample parameters.

In summary, RNN was used to obtain the model for classification from the dataset that has the parameters of voice samples. In this project, not a large number of samples of votes was used, but if it turns out to increase the number of samples of voices, this will help to create an almost perfect model, given the sufficient variety of votes and gender.

References

[1] Mel Frequency Cepstral Coefficient (MFCC) - http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/
[2] Kory Becker “Identifying the Gender of a Voice using Machine Learning”. 2016, unpublished
[3] http://festvox.org/cmu_arctic/