Gender detection from raw audio signals

Knowledge discovery lab

Faculty of informatics

Objectives

Goal is to find a way to process raw audio data and then feed that data to various ML models in order to determine a more complex feature, for the purposes of this experiment gender detection is the goal. The main tools are:

- Mathematical analysis of audio signals using Fourier analysis
- Extraction of features from the Fourier image which are useful for gender detection
- If necessary remove any unneeded features and if possible create additional ones
- Use the data to train ML models which are then manually tuned and measure their performance by certain metric.

Introduction

The time-domain analysis of audio signals allows only the analysis of amplitude strength and its change in time, but for the purposes of **extracting more complex information**(energy distribution and similar metrics) **frequency-domain** analysis is much more appropriate. The purpose of this experiment is to **identify** at least a part of the **important features** necessary for solving the mentioned problem.

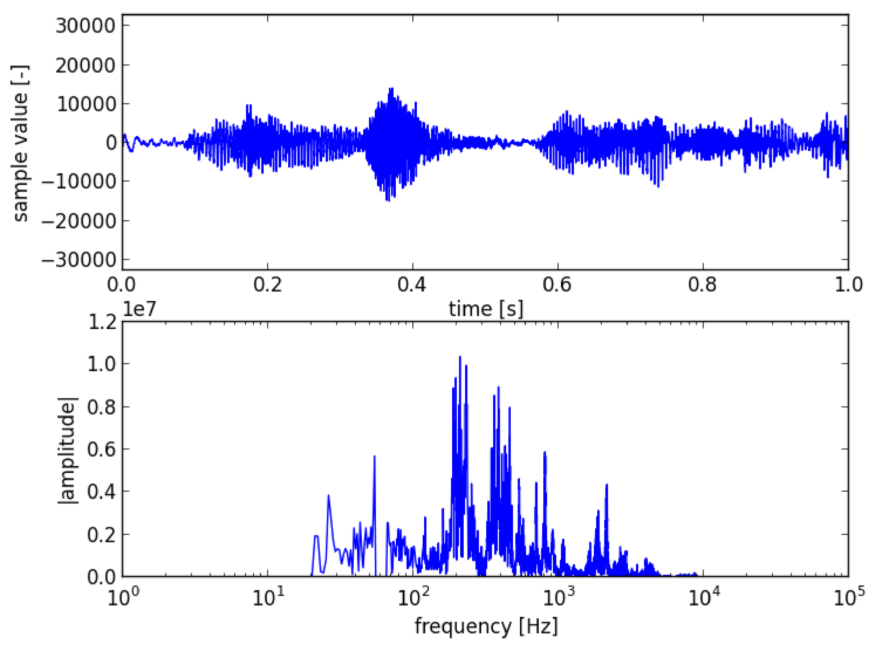

Figure 1:Signal represented in time and frequency domains[1]

Feature creation

The most important tool used for this purpose is the Fourier transformation implemented in numpy library. The resulting image is shown in figure 1. The resulting image offers easier signal analysis since the data is not defined in a sequential manner, from this the following properties can be extracted:

- Mean, median and standard deviation of frequency
- Mode frequency and centroid of frequency
- Spectral entropy and flatness
- Mean, median and standard deviation of amplitude
- Lower, upper quartiles and interquartile range
- Skewness and kurtosis

Using the extracted features, data was **clustered** in 6 groups representing different types of voice as an additional feature.

Using the obtained features on a deep learning model

Deep learning model using keras library has been created and tested on the dataset. The model used 10% of the data as test set, 13.5% for validation and the rest was used for training.

First layer is the 16 features extracted from the data, which are then transformed through the $64(tanh) \rightarrow 128(ReLU) \rightarrow 128(tanh) \rightarrow 128(ReLU) \rightarrow 128(tanh) \rightarrow 64(ReLU) \rightarrow 2(softmax)$ network architecture with dropout(20%) layers after each hidden layer.

Some combinations of optimizers, loss functions and learning rate schedules have been tested out and for this particular dataset, model and purpose the combination, optimizer: RMSprop with learning rate of 0.001, loss: binary crossentropy has achieved the best results of slightly above 80% which has been achieved with random forest classifier.

Importance of different features

The importance of each feature depends on the purpose of the used data, in this case, the most defining properties are the mode frequency, which represents the frequency which carry the largest amount of energy of the signal, and skewness of the signal and spectral flatness.

Performance of machine learning models

The dataset consists of labeled audio signals split evenly between the two target labels(male and female). All of the generated features were normalized. First model to be used was **decision tree** from sklearn library, which is particularly useful since it shows the amount of information contained in each property. This model was able to achieve around 70% accuracy using ten-fold validation score. Another model which has significantly outperformed the baseline models is **random forest classifier**, with 300 estimators used it was able to achieve around 80% accuracy.

Results

The extracted features which have been used here are not fully describing the signal in frequency-domain, but have been sufficient to achieve some results with the classical ML and deep learning methods.

The performance of all three models is comparable of around 80% which is around 30% better than the baseline models. The interesting note is that every model better recall for one label(around 5% difference), while for that same label it has a lower precision(around 3-5%).

Conclusion

Most of the defining features of speakers voice is contained in frequency-domain, which is much more appropriate for machine learning models and feed-forward neural networks, than the same signal represented in time-domain.

The human auditory system solves this problem with very high accuracy in real-time, which shows the capabilities of the biological system used to analyse the audio signal and it is interesting that those capabilities are somewhat imitated using already known mathematical transformations.

Additional Information

The extraction of features in frequency-domain is done somewhat differently than in time-domain, but the principles are mainly the same. For example average frequency is defined as sum of the product of the spectrum and appropriate frequencies, which is then divided by the total sum of the spectrum.

Other common machine learning models have been used(support vector machines, logistic regression), but have not been able to achieve significantly better results than baseline models and have been omitted.

References

[1] Roland Smith.
Fourier transformation example.
https://stackoverflow.com/a/36259069, Mar 2016.
Accessed on 2021-10.

[2] H. Hu P. Phukpattaranont A.Phinyomark, S. Thongpanja and C. Limsakul.
Computational intelligence in electromyography analysis.
IntechOpen, 2012.

[3] Mücahit Büyükyılmaz and Ali Çıbıkdiken.
Voice gender recognition using deep learning.
12 2016.

Prepared by Faruk Herenda