Knowledge discovery laboratory

Detection of speakers gender using various models

November 14, 2021

Introduction

Inputs: audio signals in form of .wav files, usually up to 5 seconds long.
Tools and methods:

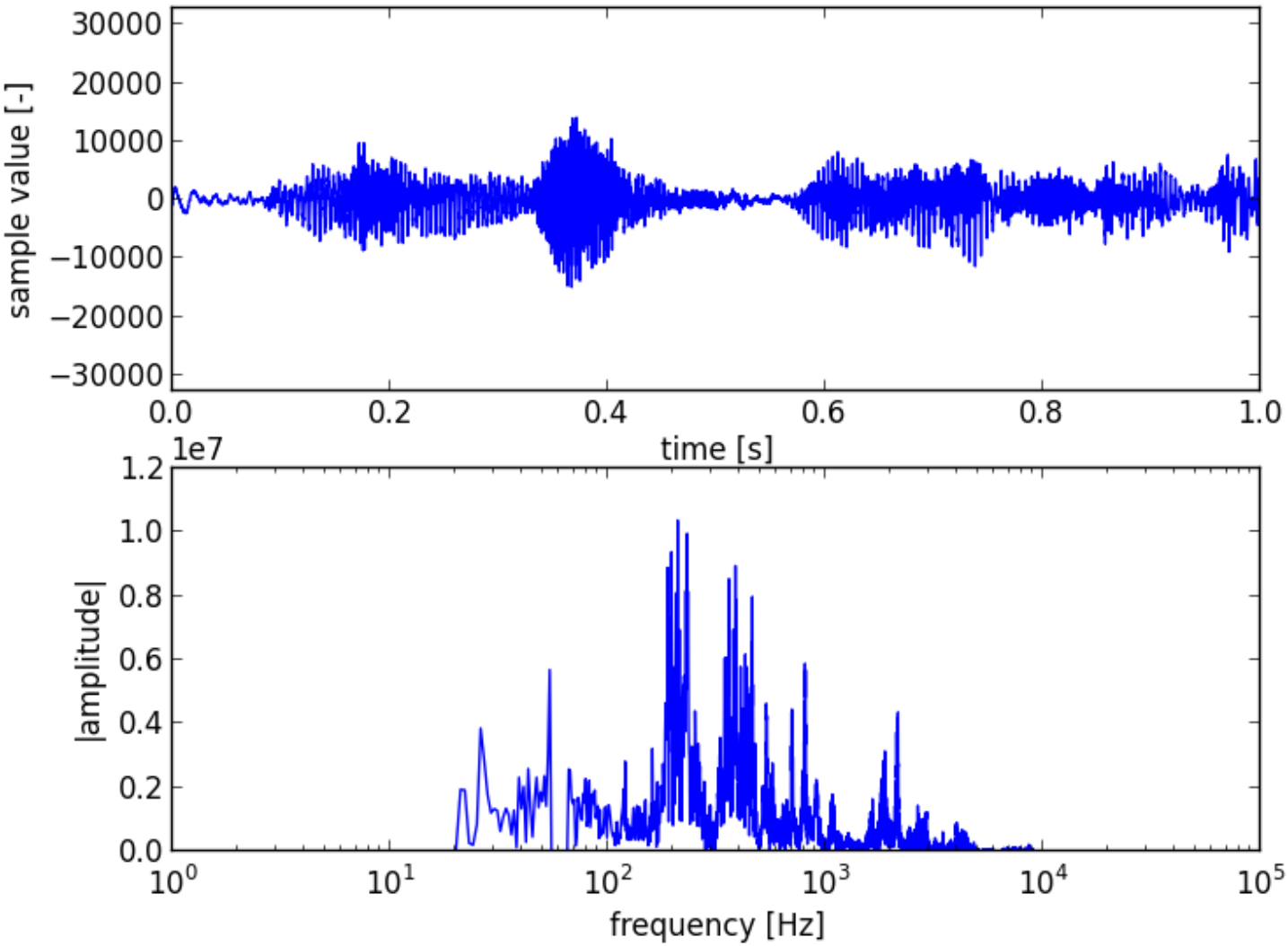
- signal analysis for feature creation
- basic preprocessing
- previous experience using ML and deep learning

Outputs: models whose inputs are the extracted numerical features and output is the predicted gender of the speaker

Signal analysis and feature creation

Fourier analysis

Representation of time-domain signal with frequency-domain one.



Signal analysis and feature creation

Feature extraction

In the frequency domain concepts as average, median and deviation are calculated a bit differently.

- Average frequency - sum of the product of the spectrum and the frequency, divided by total sum of the spectrum
- Median - is the frequency at which the spectrum is divided into two regions with equal amplitude
- Standard deviation - is calculated over the frequency multiplied by amplitude
- Mode - frequency at which the amplitude is the strongest
- Centroid - weighted mean of the frequencies multiplied by amplitudes

Feature creation

- Spectral entropy - measure of signals spectral power distribution
- Spectral flatness - another measure to characterize audio spectrum
- Mean, median and standard deviation of amplitude value
- Lower, upper quartile and interquartile range
- Skewness and kurtosis - two measures describing amplitude distribution in the frequency-domain

Using the features extracted from the frequency-domain, data can be clustered in voice types, trying to imitate the types used in music.

Machine learning models

Dataset consists of around 37 000 voice samples which has roughly equal parts for both genders, 25% was used as testing data.
Baseline models achieve around 50% accuracy. Other models were graded by cross-validation(ten fold) score.

- Decision tree: 0.71 (+/- 0.03), gave insight that mode frequency is one of the most defining aspects of the voice.
- MLP classifier(sklearn library): 0.74 (+/- 0.04), architecture:(45, 180, 180, 45) with adaptive learning rate.
- Random forest classifiers: 0.79 (+/- 0.03)

Other machine learning models have achieved significantly lower accuracy and have been omitted.
All models were manually tuned.

Keras model

For keras model data was also standardized using standard scaler from sklearn. Split into training(76.5%), validation(13.5%) and test data(10%).

- The structure of the used network is: (64, 128, 128, 128, 128, 64, 2) with a dropout layer(20%) after each hidden layer.
- The activation functions were tanh for every odd layer and relu otherwise. The output layer used softmax activation function.
- Optimizers: Adadelta, Nadam, RMSprop and SGD in combination with lr schedules(when appropriate): polynomial or exponential.
- Loss functions: binary crossentropy, KLDivergence, MeanAbsoluteError and MeanSquaredError used with the same metrics.

Keras model

Due to the large dataset and dropout layers it was decided to use large number of epochs along with a larger batch size. Many configurations have been tested and the best outcome was given by 400 epochs and batch size of 64.
The model which achieved best results has the following parameters:

- Optimizer Nadam with learning rate of 0.001
- Loss function is binary crossentropy with measured accuracy as a metric

Conclusion

- There is a lot of data hidden behind Fourier image of a signal.
- Additional features can be extracted from the image which would help increase the accuracy.
- A trivial task for humans, but not so much for ML models, shows the complexity of the human auditory system.

s Angkoon Phinyomark, Sirinee Thongpanja, Huosheng Hu, Pornchai Phukpattaranont and Chusak Limsakul
The usefulness of mean and median frequencies in electromyography analysis
DOI: 10.5772/50639.

Mucahit Buyukyilmaz and Ali Osman Cibikdiken
Voice Gender Recognition Using Deep Learning
DOI:10.2991/msota-16.2016.90.