Emotion detection from sound

Project proposal by Alena Churakova

Domain background

Much of the natural language understanding by a machine is based on study of text. This project focuses on recognition of emotions from the sound of human speech. Researchers recognize the importance of teaching machines to recognize human emotions, which can be used to adapt machine's action appropriately [1]. The rise of virtual personal assistants shows people willingness to interact with machines by voice. Analysis of tone in addition to the text information has a potential to make the interactions more natural, pleasant and effective.

Problem statement

A sentence with a neutral meaning, e.g. "This is a dog", carries factual content. Depending on the context and a person's attitude. The same sentence can be pronounced with happiness in the voice by a child who was secretly hoping their parents get a puppy a birthday present. A burglar, on the contrary, would likelier say it with fear.

The task is to recognise emotions from the audial signals of human speech, without involvement of text analysis. A model developed in this project can be useful in an application where a machine changes its action based on emotion detection, e.g. encourage the conversation partner when detecting sadness.

Datasets and inputs

Toronto emotional speech set (TESS)[2] was created by Kate Dupuis and M. Kathleen Pichora-Fuller in 2010. It consists of short audiofiles (.wav format) recorded by two actors of different ages. Short sentences of the form 'Say the word _____' were pronounced with different emotions (anger, happiness, fear, etc) and with a neutral tone. This project is concentrating on a few expressions - sad, happy, fear, angry and neutral. In the later work, the list can be extended. The total number of files in each category is 400.

[1] Maghilnan S and Rajesh Kumar M. Sentiment Analysis on Speaker Specific Speech Data https://arxiv.org/pdf/1802.06209.pdf

[2] Dupuis, Kate and Pichora-Fuller, M. Kathleen. Toronto emotional speech set (TESS) https://tspace.library.utoronto.ca/handle/1807/24487

Solution statement

The task of emotion recognition is a multi-class classification problem. Any classification algorithms that is natively supporting multiple classes or that can be tweeked via e.g. one-vs-all, could be appropriate.

The complete solution include among others:

- Interfaces for data input (via a microphone) and prediction display
- Preprocessing of audio data
- Feature generation
- Machine learning perdition model

Benchmark model

A no-model (random) prediction for a classification task could serve as a benchmark for all modelling approaches. With five categories (sad, happy, fear, angry, neutral) and a balanced data set, a baseline accuracy would be appx. 20%.

Evaluation metrics

In a described example scenario of changing machine's actions based on the detected emotion, an accurate prediction brings value of appropriate reaction by a machine. This motivates to use accuracy as an evaluation metric.

Project design

As mentioned in the solution statement, the final prediction pipeline encompasses steps from data ingest up to prediction. An ML development process suggested by Uber [3] carries credibility due to their market success that is at least in part attributable to deployed machine learning models.

This proposal outlines the first step of understanding business needs and defining minimal viable product. The prototype stage will include data preparation and training/evaluation of the models. MFCC (Mel-Frequency Cepstral Coefficients) is a well established feature for audio files that takes advantage of multiple ideas in sound preprocessing (overlapping windows, fast Fourier transforms, etc).[4] Definition of audio pre-processing steps to be used on every audio input

[3] Hermann, Jeremy and Del Balso, Mike. Scaling Machine Learning at Uber with Michelangelo https://eng.uber.com/scaling-michelangelo/

[4] Mendels, Gideon. How to apply machine learning and deep learning methods to audio analysis https://towardsdatascience.com/how-to-apply-machine-learning-and-deep-learning-methods-to-audio-analysis-615e286fcbbc

are one of the outputs of the prototype stage. Deep learning approach suits the multi-class classification in the speech domain and will be followed in this project. In particular, a Multilayer Perceptron (MLP) will be compared to the no-model benchmark performance in terms of accuracy. The productionizing of the solution will include model deployment and making prediction. The forth stage of measuring predictions in production can be realized via a feedback loop from the interface where the prediction is displayed. This last step is, however, out of scope of the current project.