2 Data Acquisition

Data Science - Capstone Project Submission

- Student Name: **James Toop**
- Student Pace: **Self Paced**
- Scheduled project review date/time: **29th October 2021 @ 21:30 BST**
- Instructor name: **Jeff Herman / James Irving**
- Blog URL: **https://toopster.github.io/ (https://toopster.github.io/)**

IMPORTANT NOTE:

This section presents code and instructions for downloading each dataset.

The datasets and transformed JSON files have not been included in the GitHub repository with this notebook and will need to be downloaded and stored in the local repository for the code to run correctly.

The code below will however download and store the datasets. The notebook (3_preprocessing.ipynb) entitled `3_preprocessing.ipynb` contains code for transforming the datasets as required for the models to run.

To ensure ease of use, however, it is also possible to download the raw and transformed datasets using this link (https://drive.google.com/file/d/11lKYIZiwEQJ-pp0G1bJPHXLJLj8uKPqW/view?usp=sharing).

In [1]:

```
# Import required libraries and modules for data acquisition
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import tensorflow as tf

import os
import pathlib
from pathlib import Path
import shutil
```

2.1 Download Speech Commands dataset

A dataset for limited-vocabulary speech recognition by Pete Warden, TensorFlow team at Google.

https://arxiv.org/abs/1804.03209 (https://arxiv.org/abs/1804.03209)

The Speech Commands dataset is an attempt to build a standard training and evaluation dataset for a class of simple speech recognition tasks. Its primary goal is to provide a way to build and test small models that detect when a single word is spoken, from a set of ten or fewer target words, with as few false background noise or unrelated speech.

In [2]:

```
def download_speech_commands():
    '''
    Code adapted from Simple audio recognition: Recognizing keywords
    https://www.tensorflow.org/tutorials/audio/simple_audio

    Downloads and unpacks the speech commands dataset, removing any
    unnecessary files
    '''
    data_dir = pathlib.Path('data/speech_commands_v0.02')
    origin = 'http://download.tensorflow.org/data/speech_commands_v0.02.tar.gz'

    # Check to see if data directory already exists, download if not
    if not data_dir.exists():
        tf.keras.utils.get_file(
            'speech_commands_v0.02.zip',
            origin=origin,
            extract=True,
            cache_dir='.',
            cache_subdir='data/speech_commands_v0.02')
    else:
        print('Speech Commands dataset already exists')

    # Remove the _background_noise_ samples as these are not required
    try:
        shutil.rmtree(str(data_dir) + '/_background_noise_')
    except OSError as e:
        print('Error: %s - %s.' % (e.filename, e.strerror),
              'Check if directory has already been removed.')


    # Remove the extracted zip file for politeness as this is not required
    zip_file = str(data_dir) + '/speech_commands_v0.02.zip'
    if os.path.exists(zip_file):
        os.remove(zip_file)
```

In [3]:

```
# Call the function to download the Speech Commands v0.02 dataset
download_speech_commands()
```

```
Downloading data from http://download.tensorflow.org/data/speech_comma
nds_v0.02.tar.gz (http://download.tensorflow.org/data/speech_commands_
v0.02.tar.gz)
2428928000/2428923189 [==============================] - 160s 0us/step
2428936192/2428923189 [==============================] - 160s 0us/step
```

2.2 Download the Ultrasuite dataset

A collection of ultrasound and acoustic speech data from child speech therapy sessions – University of Edinburgh, School of Infomatics

https://ultrasuite.github.io/ (https://ultrasuite.github.io/)

Ultrasuite is a collection of ultrasound and acoustic speech data from child speech therapy sessions. The current release includes three datasets, one from typically developing children and two from speech disordered children:

- [Ultrax Typically Developing (UXTD) (https://ultrasuite.github.io/data/uxtd/)](https://ultrasuite.github.io/data/uxtd/) - A dataset of 58 typically developing children.
- [Ultrax Speech Sound Disorders (UXSSD) (https://ultrasuite.github.io/data/uxssd/)](https://ultrasuite.github.io/data/uxssd/) - A dataset of 8 children with speech sound disorders.
- [UltraPhonix (UPX) (https://ultrasuite.github.io/data/upx/)](https://ultrasuite.github.io/data/upx/) - A second dataset of children with speech sound disorders, collected from 20 children.

Source:

Eshky, A., Ribeiro, M. S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J., & Wrench, A. (2018) Ultrasuite: A repository of ultrasound and acoustic data from child speech therapy sessions. Proceedings of INTERSPEECH. Hyderabad, India. [[paper (https://ultrasuite.github.io/papers/ultrasuite_IS18.pdf)](https://ultrasuite.github.io/papers/ultrasuite_IS18.pdf)]

In [4]:

```python
# Function for downloading the Ultrasuite datasets
def download_ultrasuite(datasets):
    '''
    Sets up a remote sync for the Ultrasuite datasets and labels

        Params:
            datasets (list): Specific ultrasuite dataset to sync can be
                             'upx', 'uxtd' or 'uxssd'

    '''
    orig_loc = Path.cwd()
    data_dir = pathlib.Path('data/ultrasuite')

    # Check to see if data directory already exists, download if not
    if not os.path.isdir(data_dir):
        os.makedirs(data_dir)

        # Change working directory
        os.chdir(data_dir)

        for dataset in datasets:
            rsync_data = 'rsync -av --include="*/" --include="*.wav" \
            --exclude="*" ultrasuite-rsync.inf.ed.ac.uk::ultrasuite/core-'
            os.system(rsync_data + dataset + ' .')
            print(dataset, 'dataset has been downloaded.')

        rsync_labels = 'rsync -av \
        ultrasuite-rsync.inf.ed.ac.uk::ultrasuite/labels-uxtd-uxssd-upx .'
        os.system(rsync_labels)
        print('The ultrasuite labels have been downloaded.')

        # Change working directory back
        os.chdir(orig_loc)
```

In [5]:

```python
# Download the Ultrasuite datasets
download_ultrasuite(['upx', 'uxtd', 'uxssd'])
```

```
upx dataset has been downloaded.
uxtd dataset has been downloaded.
uxssd dataset has been downloaded.
The ultrasuite labels have been downloaded.
```

Sources / Code adapted from:

* Simple audio recognition: Recognizing keywords - Tensorflow (https://www.tensorflow.org/tutorials/audio/simple_audio)