# DysarthriaDataSet

October 20, 2022

## 1 Introduction

- This program uses Torgo datasets it joins wav and prompts file together into csv file.
- Other data that's been collected is gender, dysarthia.
- Preprocessing: We've downloaded the Torgo dataset then we've combine all the prompt and wav file into a folder with their respective name. I.e F, M, FC and MC

[1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import os
import re
import warnings
import matplotlib.pyplot as plt
import seaborn as sns
import librosa
import librosa.display
from sklearn.preprocessing import minmax_scale
import IPython.display as ipd
import ntpath
import sys
```

## 2 Female Dataset (F)

[2]:
```
def audioFile():
    wav_id = []
    prompts = []
    temp_wavID = []
    temp_prompt = []

    wav_path = 'TestData/F/'
    for f_name in os.listdir(wav_path):
        if f_name.endswith('.wav'):
            temp_wavID += [f_name.replace(".wav", "")]
            temp_wavID.sort()

    for x in os.listdir(wav_path):
        if x.endswith(".txt"):
```

```
            temp_prompt += [x.replace(".txt", "")]
            temp_prompt.sort()

    ### Remove difference
    # set() + - operator
    # find difference (a - b)
    result_a = set(temp_wavID) - set(temp_prompt)
    # find difference (b - a)
    result_b = set(temp_prompt) - set(temp_wavID)

    # Remove
    for item in list(result_a):
        temp_wavID.remove(item)


    for item in list(result_b):
        temp_prompt.remove(item)

    print(temp_wavID == temp_prompt)
    ### Finished remove difference

    for x in temp_wavID:
        filename = wav_path+""+str(x)+".wav"
        wav_id += [filename]

    for x in temp_prompt:
        with open(wav_path + '/' + x+'.txt', 'r') as reader:
            x = reader.read().strip('\n')
        prompts.append(x)

    dataf = pd.DataFrame()
    dataf['filename'] = pd.Series(wav_id)
    dataf['is_dysarthria'] = 'dysarthria'
    dataf['gender'] = 'female'
    dataf['prompts'] = pd.Series(prompts)
    return dataf
```

[3]:
```
F = audioFile()
```

```
True
```

[4]:
```
F = F[["is_dysarthria", "gender", "filename", "prompts"]]
F
```

[4]:

| | is_dysarthria | gender | filename | \ |
|---|---|---|---|---|
| 0 | dysarthria | female | TestData/F/0001.wav | |
| 1 | dysarthria | female | TestData/F/0002.wav | |

```
2      dysarthria  female  TestData/F/0003.wav
3      dysarthria  female  TestData/F/0004.wav
4      dysarthria  female  TestData/F/0005.wav
..            …     …                   …
127    dysarthria  female  TestData/F/0130.wav
128    dysarthria  female  TestData/F/0131.wav
129    dysarthria  female  TestData/F/0132.wav
130    dysarthria  female  TestData/F/0133.wav
131    dysarthria  female  TestData/F/0134.wav

                                                prompts
0                              [say Ah-P-Eee repeatedly]
1                              [say Ah-P-Eee repeatedly]
2                           [say Pah-Tah-Kah repeatedly]
3                              [say Eee-P-Ah repeatedly]
4              [relax your mouth in its normal position]
..                                                   …
127    Grandfather likes to be modern in his language.
128                                                torn
129  When he speaks, his voice is just a bit cracke…
130                                              trait
131                                                yes

[132 rows x 4 columns]
```

## 3 Control Female Group (FC)

[5]:
```python
def audioFile():
    wav_id = []
    prompts = []
    temp_wavID = []
    temp_prompt = []

    wav_path = 'TestData/FC/'
    for f_name in os.listdir(wav_path):
        if f_name.endswith('.wav'):
            temp_wavID += [f_name.replace(".wav", "")]
            temp_wavID.sort()

    for x in os.listdir(wav_path):
        if x.endswith(".txt"):
            temp_prompt += [x.replace(".txt", "")]
            temp_prompt.sort()

    ### Remove difference
    # set() + - operator
```

```
    # find difference (a - b)
    result_a = set(temp_wavID) - set(temp_prompt)
    # find difference (b - a)
    result_b = set(temp_prompt) - set(temp_wavID)

    # Remove
    for item in list(result_a):
        temp_wavID.remove(item)


    for item in list(result_b):
        temp_prompt.remove(item)

    print(temp_wavID == temp_prompt)
    ### Finished remove difference

    for x in temp_wavID:
        filename = wav_path+""+str(x)+".wav"
        wav_id += [filename]

    for x in temp_prompt:
        with open(wav_path + '/' + x+'.txt', 'r') as reader:
            x = reader.read().strip('\n')
        prompts.append(x)

    dataf = pd.DataFrame()
    dataf['filename'] = pd.Series(wav_id)
    dataf['is_dysarthria'] = 'non-dysarthria'
    dataf['gender'] = 'female'
    dataf['prompts'] = pd.Series(prompts)
    return dataf
```

[6]:
```
FC = audioFile()
```

```
True
```

[7]:
```
FC = FC[["is_dysarthria", "gender", "filename", "prompts"]]
```

[8]:
```
FC
```

[8]:

| | is_dysarthria | gender | filename | \ |
|---|---|---|---|---|
| 0 | non-dysarthria | female | TestData/FC/0001.wav | |
| 1 | non-dysarthria | female | TestData/FC/0002.wav | |
| 2 | non-dysarthria | female | TestData/FC/0003.wav | |
| 3 | non-dysarthria | female | TestData/FC/0004.wav | |
| 4 | non-dysarthria | female | TestData/FC/0005.wav | |
| .. | … | … | … | |

```
159  non-dysarthria  female  TestData/FC/0160.wav
160  non-dysarthria  female  TestData/FC/0161.wav
161  non-dysarthria  female  TestData/FC/0162.wav
162  non-dysarthria  female  TestData/FC/0163.wav
163  non-dysarthria  female  TestData/FC/0164.wav

                                          prompts
0                        [say 'Ah-P-Eee' repeatedly]
1                        [say 'Ah-P-Eee' repeatedly]
2       [relax your mouth in its normal position]
3                        [say 'Eee-P-Ah' repeatedly]
4                                             dug
..                                            …
159                                          slip
160  Two other cases also were under advisement.
161                                          suit
162                                        jungle
163                                         group

[164 rows x 4 columns]
```

# 4 Male dataset (M)

[9]:
```python
def audioFile():
    wav_id = []
    prompts = []
    temp_wavID = []
    temp_prompt = []

    wav_path = 'TestData/M/'
    for f_name in os.listdir(wav_path):
        if f_name.endswith('.wav'):
            temp_wavID += [f_name.replace(".wav", "")]
            temp_wavID.sort()

    for x in os.listdir(wav_path):
        if x.endswith(".txt"):
            temp_prompt += [x.replace(".txt", "")]
            temp_prompt.sort()

    ### Remove difference
    # set() + - operator
    # find difference (a - b)
    result_a = set(temp_wavID) - set(temp_prompt)
    # find difference (b - a)
    result_b = set(temp_prompt) - set(temp_wavID)
```

```
    # Remove
    for item in list(result_a):
        temp_wavID.remove(item)


    for item in list(result_b):
        temp_prompt.remove(item)

    print(temp_wavID == temp_prompt)
    ### Finished remove difference

    for x in temp_wavID:
        filename = wav_path+""+str(x)+".wav"
        wav_id += [filename]

    for x in temp_prompt:
        with open(wav_path + '/' + x+'.txt', 'r') as reader:
            x = reader.read().strip('\n')
        prompts.append(x)

    dataf = pd.DataFrame()
    dataf['filename'] = pd.Series(wav_id)
    dataf['is_dysarthria'] = 'dysarthria'
    dataf['gender'] = 'male'
    dataf['prompts'] = pd.Series(prompts)
    return dataf
```

[10]:
```
M = audioFile()
```

```
True
```

[11]:
```
M = M[["is_dysarthria", "gender", "filename", "prompts"]]
```

[12]:
```
M
```

[12]:

| | is_dysarthria | gender | filename | \ |
|---|---|---|---|---|
| 0 | dysarthria | male | TestData/M/0001.wav | |
| 1 | dysarthria | male | TestData/M/0002.wav | |
| 2 | dysarthria | male | TestData/M/0003.wav | |
| 3 | dysarthria | male | TestData/M/0004.wav | |
| 4 | dysarthria | male | TestData/M/0005.wav | |
| .. | … | … | … | |
| 95 | dysarthria | male | TestData/M/0096.wav | |
| 96 | dysarthria | male | TestData/M/0097.wav | |
| 97 | dysarthria | male | TestData/M/0098.wav | |
| 98 | dysarthria | male | TestData/M/0099.wav | |

```
99    dysarthria   male  TestData/M/0100.wav

                                                prompts
0                           [say Ah-P-Eee repeatedly]
1                         [say 'Eee-P-Ah' repeatedly]
2                       [say 'Pah-Tah-Kah' repeatedly]
3            [relax your mouth in its normal position]
4   When he speaks, his voice is just a bit cracke…
..                                                  …
95                                              weed
96                                              weed
97                                              corn
98                                                up
99                                             swarm

[100 rows x 4 columns]
```

## 5 Male Control Group Dataset (MC)

[13]:
```python
def audioFile():
    wav_id = []
    prompts = []
    temp_wavID = []
    temp_prompt = []

    wav_path = 'TestData/MC/'
    for f_name in os.listdir(wav_path):
        if f_name.endswith('.wav'):
            temp_wavID += [f_name.replace(".wav", "")]
            temp_wavID.sort()

    for x in os.listdir(wav_path):
        if x.endswith(".txt"):
            temp_prompt += [x.replace(".txt", "")]
            temp_prompt.sort()

    ### Remove difference
    # set() + - operator
    # find difference (a - b)
    result_a = set(temp_wavID) - set(temp_prompt)
    # find difference (b - a)
    result_b = set(temp_prompt) - set(temp_wavID)

    # Remove
    for item in list(result_a):
        temp_wavID.remove(item)
```

```
    for item in list(result_b):
        temp_prompt.remove(item)

    print(temp_wavID == temp_prompt)
    ### Finished remove difference

    for x in temp_wavID:
        filename = wav_path+""+str(x)+".wav"
        wav_id += [filename]

    for x in temp_prompt:
        with open(wav_path + '/' + x+'.txt', 'r') as reader:
            x = reader.read().strip('\n')
        prompts.append(x)

    dataf = pd.DataFrame()
    dataf['filename'] = pd.Series(wav_id)
    dataf['is_dysarthria'] = 'non-dysarthria'
    dataf['gender'] = 'male'
    dataf['prompts'] = pd.Series(prompts)
    return dataf
```

[14]:
```
MC = audioFile()
```

```
True
```

[15]:
```
MC = MC[["is_dysarthria", "gender", "filename", "prompts"]]
```

[16]:
```
MC
```

[16]:
```
      is_dysarthria gender              filename  \
0    non-dysarthria   male  TestData/MC/0001.wav
1    non-dysarthria   male  TestData/MC/0002.wav
2    non-dysarthria   male  TestData/MC/0003.wav
3    non-dysarthria   male  TestData/MC/0004.wav
4    non-dysarthria   male  TestData/MC/0005.wav
..              …    …                    …
324  non-dysarthria   male  TestData/MC/0325.wav
325  non-dysarthria   male  TestData/MC/0326.wav
326  non-dysarthria   male  TestData/MC/0327.wav
327  non-dysarthria   male  TestData/MC/0328.wav
328  non-dysarthria   male  TestData/MC/0329.wav

                                      prompts
0                  [say 'Ah-P-Eee' repeatedly]
```

```
1                  [say 'Eee-P-Ah' repeatedly]
2               [say 'Pah-Tah-Kah' repeatedly]
3    [relax your mouth in its normal position]
4                             spark spark spark
..                                          …
324                                          no
325                                        hill
326  The museum hires musicians every evening.
327                                     thought
328            When all else fails, use force.

[329 rows x 4 columns]
```

# 6 Completed creating the dataset.

# 7 Merge two data sets. (F and FC)

[17]:
```
df = pd.merge(F, FC, on=['prompts'])
```

[18]:
```
df
```

[18]:
```
    is_dysarthria_x gender_x           filename_x  \
0       dysarthria   female  TestData/F/0005.wav
1       dysarthria   female  TestData/F/0005.wav
2       dysarthria   female  TestData/F/0067.wav
3       dysarthria   female  TestData/F/0067.wav
4       dysarthria   female  TestData/F/0068.wav
..             …      …                   …
141     dysarthria   female  TestData/F/0131.wav
142     dysarthria   female  TestData/F/0132.wav
143     dysarthria   female  TestData/F/0133.wav
144     dysarthria   female  TestData/F/0134.wav
145     dysarthria   female  TestData/F/0134.wav

                                              prompts is_dysarthria_y  \
0           [relax your mouth in its normal position]  non-dysarthria
1           [relax your mouth in its normal position]  non-dysarthria
2           [relax your mouth in its normal position]  non-dysarthria
3           [relax your mouth in its normal position]  non-dysarthria
4           [relax your mouth in its normal position]  non-dysarthria
..                                                …             …
141                                            torn  non-dysarthria
142  When he speaks, his voice is just a bit cracke…  non-dysarthria
143                                           trait  non-dysarthria
144                                             yes  non-dysarthria
145                                             yes  non-dysarthria
```

```
    gender_y            filename_y
0     female  TestData/FC/0003.wav
1     female  TestData/FC/0059.wav
2     female  TestData/FC/0003.wav
3     female  TestData/FC/0059.wav
4     female  TestData/FC/0003.wav
..       …                   …
141   female  TestData/FC/0088.wav
142   female  TestData/FC/0015.wav
143   female  TestData/FC/0106.wav
144   female  TestData/FC/0022.wav
145   female  TestData/FC/0148.wav

[146 rows x 7 columns]
```

[22]:
```
df = df.drop_duplicates(subset=['prompts'])
df = df.reset_index(drop=True)
```

[23]:
```
df.to_csv("FandFC.csv", index=False)
```

[24]:
```
dataf = pd.read_csv("FandFC.csv")
dataf
```

[24]:
```
   is_dysarthria_x gender_x           filename_x  \
0       dysarthria   female  TestData/F/0005.wav
1       dysarthria   female  TestData/F/0006.wav
2       dysarthria   female  TestData/F/0008.wav
3       dysarthria   female  TestData/F/0009.wav
4       dysarthria   female  TestData/F/0010.wav
..             …      …                   …
86      dysarthria   female  TestData/F/0130.wav
87      dysarthria   female  TestData/F/0131.wav
88      dysarthria   female  TestData/F/0132.wav
89      dysarthria   female  TestData/F/0133.wav
90      dysarthria   female  TestData/F/0134.wav

                                              prompts is_dysarthria_y  \
0          [relax your mouth in its normal position]  non-dysarthria
1                                              stick  non-dysarthria
2   Except in the winter when the ooze or snow or …  non-dysarthria
3                                               pat  non-dysarthria
4                                                up  non-dysarthria
..                                              …             …
86    Grandfather likes to be modern in his language.  non-dysarthria
87                                             torn  non-dysarthria
88  When he speaks, his voice is just a bit cracke…  non-dysarthria
```

```
89                                                  trait  non-dysarthria
90                                                    yes  non-dysarthria

   gender_y            filename_y
0    female  TestData/FC/0003.wav
1    female  TestData/FC/0074.wav
2    female  TestData/FC/0061.wav
3    female  TestData/FC/0007.wav
4    female  TestData/FC/0093.wav
..      …                     …
86   female  TestData/FC/0114.wav
87   female  TestData/FC/0088.wav
88   female  TestData/FC/0015.wav
89   female  TestData/FC/0106.wav
90   female  TestData/FC/0022.wav

[91 rows x 7 columns]
```

## 8 Merge two data sets. (M and MC)

[25]:
```
df = pd.merge(M, MC, on=['prompts'])
df
```

[25]:
```
    is_dysarthria_x gender_x           filename_x  \
0        dysarthria     male  TestData/M/0002.wav
1        dysarthria     male  TestData/M/0003.wav
2        dysarthria     male  TestData/M/0004.wav
3        dysarthria     male  TestData/M/0004.wav
4        dysarthria     male  TestData/M/0004.wav
..              …      …                     …
121      dysarthria     male  TestData/M/0095.wav
122      dysarthria     male  TestData/M/0096.wav
123      dysarthria     male  TestData/M/0097.wav
124      dysarthria     male  TestData/M/0098.wav
125      dysarthria     male  TestData/M/0099.wav

                                      prompts is_dysarthria_y gender_y  \
0                  [say 'Eee-P-Ah' repeatedly]  non-dysarthria     male
1               [say 'Pah-Tah-Kah' repeatedly]  non-dysarthria     male
2    [relax your mouth in its normal position]  non-dysarthria     male
3    [relax your mouth in its normal position]  non-dysarthria     male
4    [relax your mouth in its normal position]  non-dysarthria     male
..                                         …             …      …
121                                      yes  non-dysarthria     male
122                                    weed  non-dysarthria     male
123                                    weed  non-dysarthria     male
```

```
124                                      corn  non-dysarthria     male
125                                        up  non-dysarthria     male

               filename_y
0    TestData/MC/0002.wav
1    TestData/MC/0003.wav
2    TestData/MC/0004.wav
3    TestData/MC/0073.wav
4    TestData/MC/0178.wav
..                    …
121  TestData/MC/0271.wav
122  TestData/MC/0108.wav
123  TestData/MC/0108.wav
124  TestData/MC/0055.wav
125  TestData/MC/0033.wav

[126 rows x 7 columns]
```

[26]:
```
df = df.drop_duplicates(subset=['prompts'])
df = df.reset_index(drop=True)
```

[29]:
```
df.to_csv("MandMC.csv")
df
```

[29]:
```
   is_dysarthria_x gender_x           filename_x  \
0       dysarthria     male  TestData/M/0002.wav
1       dysarthria     male  TestData/M/0003.wav
2       dysarthria     male  TestData/M/0004.wav
3       dysarthria     male  TestData/M/0005.wav
4       dysarthria     male  TestData/M/0006.wav
..             …      …                   …
81      dysarthria     male  TestData/M/0094.wav
82      dysarthria     male  TestData/M/0095.wav
83      dysarthria     male  TestData/M/0096.wav
84      dysarthria     male  TestData/M/0098.wav
85      dysarthria     male  TestData/M/0099.wav

                                              prompts is_dysarthria_y  \
0                        [say 'Eee-P-Ah' repeatedly]  non-dysarthria
1                      [say 'Pah-Tah-Kah' repeatedly]  non-dysarthria
2           [relax your mouth in its normal position]  non-dysarthria
3   When he speaks, his voice is just a bit cracke…  non-dysarthria
4                                             trait  non-dysarthria
..                                              …             …
81                                             rake  non-dysarthria
82                                              yes  non-dysarthria
83                                             weed  non-dysarthria
```

```
84                                                  corn  non-dysarthria
85                                                    up  non-dysarthria

   gender_y            filename_y
0      male  TestData/MC/0002.wav
1      male  TestData/MC/0003.wav
2      male  TestData/MC/0004.wav
3      male  TestData/MC/0120.wav
4      male  TestData/MC/0115.wav
..      …                    …
81     male  TestData/MC/0082.wav
82     male  TestData/MC/0110.wav
83     male  TestData/MC/0108.wav
84     male  TestData/MC/0055.wav
85     male  TestData/MC/0033.wav

[86 rows x 7 columns]
```

## 9 Correlation of Dataset (F and FC)

[30]:
```
PartA = pd.read_csv("FandFC.csv")
PartA = PartA[['is_dysarthria_x', 'gender_x', 'filename_x']]
PartA
```

[30]:
```
   is_dysarthria_x gender_x           filename_x
0       dysarthria   female  TestData/F/0005.wav
1       dysarthria   female  TestData/F/0006.wav
2       dysarthria   female  TestData/F/0008.wav
3       dysarthria   female  TestData/F/0009.wav
4       dysarthria   female  TestData/F/0010.wav
..             …      …                    …
86      dysarthria   female  TestData/F/0130.wav
87      dysarthria   female  TestData/F/0131.wav
88      dysarthria   female  TestData/F/0132.wav
89      dysarthria   female  TestData/F/0133.wav
90      dysarthria   female  TestData/F/0134.wav

[91 rows x 3 columns]
```

[31]:
```
PartB = pd.read_csv("FandFC.csv")
PartB = PartB[['is_dysarthria_y', 'gender_y', 'filename_y']]
PartB
```

[31]:
```
   is_dysarthria_y gender_y           filename_y
0   non-dysarthria   female  TestData/FC/0003.wav
1   non-dysarthria   female  TestData/FC/0074.wav
```

```
2   non-dysarthria   female  TestData/FC/0061.wav
3   non-dysarthria   female  TestData/FC/0007.wav
4   non-dysarthria   female  TestData/FC/0093.wav
..             …      …                   …
86  non-dysarthria   female  TestData/FC/0114.wav
87  non-dysarthria   female  TestData/FC/0088.wav
88  non-dysarthria   female  TestData/FC/0015.wav
89  non-dysarthria   female  TestData/FC/0106.wav
90  non-dysarthria   female  TestData/FC/0022.wav

[91 rows x 3 columns]
```

[32]:
```
from tqdm import tqdm
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix,␣
 ↪roc_auc_score, roc_curve, recall_score
```

[34]:
```
def feature_extraction(df, is_dysarthria,gender,filename):
    features = []
    for i,record in tqdm(df.iterrows(),total=df.shape[0]):
        try:
            x , sr = librosa.load(record[''+filename])
            mean_mfcc = np.mean(librosa.feature.mfcc(y=x, sr=sr,␣
 ↪n_mfcc=128),axis=1)
            features.append(mean_mfcc)
        except EOFError:
            pass

    dataf = pd.DataFrame(features)
    dataf[['x', 'y', 'z']] = df[[''+is_dysarthria, ''+gender, ''+filename]]

    return dataf
```

## 10 Extract Feature. Part A

[35]:
```
PartA.columns
```

[35]:
```
Index(['is_dysarthria_x', 'gender_x', 'filename_x'], dtype='object')
```

[36]:
```
PartAFeature = feature_extraction(PartA, 'is_dysarthria_x', 'gender_x',␣
 ↪'filename_x')
PartAFeature.to_csv("F.csv")
```

```
100%|      | 91/91 [00:09<00:00, 10.01it/s]
```

## 11 Extract Feature. Part B

[37]:
```
PartB.columns
```

[37]:
```
Index(['is_dysarthria_y', 'gender_y', 'filename_y'], dtype='object')
```

[38]:
```
PartBFeature = feature_extraction(PartB, 'is_dysarthria_y', 'gender_y',␣
 ↪'filename_y')
PartBFeature.to_csv("FC.csv")
```

```
100%|      | 91/91 [00:12<00:00,  7.39it/s]
```

## 12 Run MFCC Correlation

[39]:
```
CorrResult = PartBFeature.corrwith(PartAFeature, axis = 1)
CorrResult = CorrResult.round(2)
CorrResult
```

[39]:
```
0     0.96
1     0.96
2     0.95
3     0.97
4     0.97
      …
86    0.96
87    0.96
88    0.96
89    0.97
90    0.97
Length: 91, dtype: float64
```

[40]:
```
y = PartAFeature.iloc[:,-3:]
y
```

[40]:

| | x | y | z |
|---|---|---|---|
| 0 | dysarthria | female | TestData/F/0005.wav |
| 1 | dysarthria | female | TestData/F/0006.wav |
| 2 | dysarthria | female | TestData/F/0008.wav |
| 3 | dysarthria | female | TestData/F/0009.wav |
| 4 | dysarthria | female | TestData/F/0010.wav |
| .. | … | … | … |
| 86 | dysarthria | female | TestData/F/0130.wav |
| 87 | dysarthria | female | TestData/F/0131.wav |
| 88 | dysarthria | female | TestData/F/0132.wav |
| 89 | dysarthria | female | TestData/F/0133.wav |
| 90 | dysarthria | female | TestData/F/0134.wav |

```
[91 rows x 3 columns]
```

[41]:
```
final = dataf
final['Data'] = CorrResult
```

[42]:
```
final
```

[42]:
```
   is_dysarthria_x gender_x           filename_x  \
0       dysarthria   female  TestData/F/0005.wav
1       dysarthria   female  TestData/F/0006.wav
2       dysarthria   female  TestData/F/0008.wav
3       dysarthria   female  TestData/F/0009.wav
4       dysarthria   female  TestData/F/0010.wav
..             …      …                   …
86      dysarthria   female  TestData/F/0130.wav
87      dysarthria   female  TestData/F/0131.wav
88      dysarthria   female  TestData/F/0132.wav
89      dysarthria   female  TestData/F/0133.wav
90      dysarthria   female  TestData/F/0134.wav

                                              prompts is_dysarthria_y  \
0          [relax your mouth in its normal position]  non-dysarthria
1                                              stick  non-dysarthria
2   Except in the winter when the ooze or snow or …  non-dysarthria
3                                                pat  non-dysarthria
4                                                 up  non-dysarthria
..                                               …            …
86    Grandfather likes to be modern in his language.  non-dysarthria
87                                              torn  non-dysarthria
88  When he speaks, his voice is just a bit cracke…  non-dysarthria
89                                             trait  non-dysarthria
90                                               yes  non-dysarthria

   gender_y            filename_y  Data
0    female  TestData/FC/0003.wav  0.96
1    female  TestData/FC/0074.wav  0.96
2    female  TestData/FC/0061.wav  0.95
3    female  TestData/FC/0007.wav  0.97
4    female  TestData/FC/0093.wav  0.97
..      …                     …   …
86   female  TestData/FC/0114.wav  0.96
87   female  TestData/FC/0088.wav  0.96
88   female  TestData/FC/0015.wav  0.96
89   female  TestData/FC/0106.wav  0.97
90   female  TestData/FC/0022.wav  0.97

[91 rows x 8 columns]
```

[43]:
```
final.columns
```

[43]:
```
Index(['is_dysarthria_x', 'gender_x', 'filename_x', 'prompts',
       'is_dysarthria_y', 'gender_y', 'filename_y', 'Data'],
      dtype='object')
```

[44]:
```
final = final[['is_dysarthria_x', 'gender_x','filename_x', 'prompts',  'Data',
               'is_dysarthria_y', 'gender_y', 'filename_y']]
```

[45]:
```
final
```

[45]:

| | is_dysarthria_x | gender_x | filename_x | prompts | Data | is_dysarthria_y | gender_y | filename_y |
|---|---|---|---|---|---|---|---|---|
| 0 | dysarthria | female | TestData/F/0005.wav | [relax your mouth in its normal position] | 0.96 | non-dysarthria | female | TestData/FC/0003.wav |
| 1 | dysarthria | female | TestData/F/0006.wav | stick | 0.96 | non-dysarthria | female | TestData/FC/0074.wav |
| 2 | dysarthria | female | TestData/F/0008.wav | Except in the winter when the ooze or snow or … | 0.95 | non-dysarthria | female | TestData/FC/0061.wav |
| 3 | dysarthria | female | TestData/F/0009.wav | pat | 0.97 | non-dysarthria | female | TestData/FC/0007.wav |
| 4 | dysarthria | female | TestData/F/0010.wav | up | 0.97 | non-dysarthria | female | TestData/FC/0093.wav |
| .. | … | … | … | … | … | … | … | … |
| 86 | dysarthria | female | TestData/F/0130.wav | Grandfather likes to be modern in his language. | 0.96 | non-dysarthria | female | TestData/FC/0114.wav |
| 87 | dysarthria | female | TestData/F/0131.wav | torn | 0.96 | non-dysarthria | female | TestData/FC/0088.wav |
| 88 | dysarthria | female | TestData/F/0132.wav | When he speaks, his voice is just a bit cracke… | 0.96 | non-dysarthria | female | TestData/FC/0015.wav |
| 89 | dysarthria | female | TestData/F/0133.wav | trait | 0.97 | non-dysarthria | | |
| 90 | dysarthria | female | TestData/F/0134.wav | yes | 0.97 | non-dysarthria | | |

```
89   female  TestData/FC/0106.wav
90   female  TestData/FC/0022.wav

[91 rows x 8 columns]
```

## 13 Correlation of Dataset (F and FC)

[47]:
```
PartA = pd.read_csv("MandMC.csv")
PartA = PartA[['is_dysarthria_x', 'gender_x', 'filename_x']]
PartA
```

[47]:
```
   is_dysarthria_x gender_x           filename_x
0       dysarthria     male  TestData/M/0002.wav
1       dysarthria     male  TestData/M/0003.wav
2       dysarthria     male  TestData/M/0004.wav
3       dysarthria     male  TestData/M/0005.wav
4       dysarthria     male  TestData/M/0006.wav
..             …      …                  …
81      dysarthria     male  TestData/M/0094.wav
82      dysarthria     male  TestData/M/0095.wav
83      dysarthria     male  TestData/M/0096.wav
84      dysarthria     male  TestData/M/0098.wav
85      dysarthria     male  TestData/M/0099.wav

[86 rows x 3 columns]
```

[50]:
```
PartB = pd.read_csv("MandMC.csv")
PartB = PartB[['is_dysarthria_y', 'gender_y', 'filename_y']]
PartB
```

[50]:
```
   is_dysarthria_y gender_y           filename_y
0   non-dysarthria     male  TestData/MC/0002.wav
1   non-dysarthria     male  TestData/MC/0003.wav
2   non-dysarthria     male  TestData/MC/0004.wav
3   non-dysarthria     male  TestData/MC/0120.wav
4   non-dysarthria     male  TestData/MC/0115.wav
..             …      …                  …
81  non-dysarthria     male  TestData/MC/0082.wav
82  non-dysarthria     male  TestData/MC/0110.wav
83  non-dysarthria     male  TestData/MC/0108.wav
84  non-dysarthria     male  TestData/MC/0055.wav
85  non-dysarthria     male  TestData/MC/0033.wav

[86 rows x 3 columns]
```

## 14 Extract feature.

[51]:
```
PartAFeature = feature_extraction(PartA, 'is_dysarthria_x', 'gender_x',␣
↪'filename_x')
PartAFeature.to_csv("M.csv")
```

```
100%|      | 86/86 [00:13<00:00,  6.39it/s]
```

[52]:
```
PartBFeature = feature_extraction(PartB, 'is_dysarthria_y', 'gender_y',␣
↪'filename_y')
PartBFeature.to_csv("MC.csv")
```

```
100%|      | 86/86 [00:11<00:00,  7.23it/s]
```

## 15 Run MFCC Correlation

[53]:
```
CorrResult = PartBFeature.corrwith(PartAFeature, axis = 1)
CorrResult = CorrResult.round(2)
CorrResult
```

[53]:
```
0     0.98
1     0.97
2     1.00
3     0.98
4     0.99
      …
81    0.99
82    0.99
83    0.99
84    0.99
85    0.99
Length: 86, dtype: float64
```

[54]:
```
y = PartAFeature.iloc[:,-3:]
y
```

[54]:

| | x | y | z |
|---|---|---|---|
| 0 | dysarthria | male | TestData/M/0002.wav |
| 1 | dysarthria | male | TestData/M/0003.wav |
| 2 | dysarthria | male | TestData/M/0004.wav |
| 3 | dysarthria | male | TestData/M/0005.wav |
| 4 | dysarthria | male | TestData/M/0006.wav |
| .. | … | … | … |
| 81 | dysarthria | male | TestData/M/0094.wav |
| 82 | dysarthria | male | TestData/M/0095.wav |
| 83 | dysarthria | male | TestData/M/0096.wav |
| 84 | dysarthria | male | TestData/M/0098.wav |

```
85  dysarthria  male  TestData/M/0099.wav

[86 rows x 3 columns]
```

[57]:
```
final = df
final['Data'] = CorrResult
```

[58]:
```
final = final[['is_dysarthria_x', 'gender_x','filename_x', 'prompts',  'Data',
               'is_dysarthria_y', 'gender_y', 'filename_y']]
final
```

[58]:

| | is_dysarthria_x | gender_x | filename_x | prompts | Data | is_dysarthria_y | gender_y | filename_y |
|---|---|---|---|---|---|---|---|---|
| 0 | dysarthria | male | TestData/M/0002.wav | [say 'Eee-P-Ah' repeatedly] | 0.98 | non-dysarthria | male | TestData/MC/0002.wav |
| 1 | dysarthria | male | TestData/M/0003.wav | [say 'Pah-Tah-Kah' repeatedly] | 0.97 | non-dysarthria | male | TestData/MC/0003.wav |
| 2 | dysarthria | male | TestData/M/0004.wav | [relax your mouth in its normal position] | 1.00 | non-dysarthria | male | TestData/MC/0004.wav |
| 3 | dysarthria | male | TestData/M/0005.wav | When he speaks, his voice is just a bit cracke… | 0.98 | non-dysarthria | male | TestData/MC/0120.wav |
| 4 | dysarthria | male | TestData/M/0006.wav | trait | 0.99 | non-dysarthria | male | TestData/MC/0115.wav |
| .. | … | … | … | … | … | … | … | … |
| 81 | dysarthria | male | TestData/M/0094.wav | rake | 0.99 | non-dysarthria | male | TestData/MC/0082.wav |
| 82 | dysarthria | male | TestData/M/0095.wav | yes | 0.99 | non-dysarthria | male | TestData/MC/0110.wav |
| 83 | dysarthria | male | TestData/M/0096.wav | weed | 0.99 | non-dysarthria | male | TestData/MC/0108.wav |
| 84 | dysarthria | male | TestData/M/0098.wav | corn | 0.99 | non-dysarthria | | |
| 85 | dysarthria | male | TestData/M/0099.wav | up | 0.99 | non-dysarthria | | |

```
84     male  TestData/MC/0055.wav
85     male  TestData/MC/0033.wav

[86 rows x 8 columns]
```

[ ]: