Make Train Valid Test Split

This file splits the 193 videos into Train/Validation/Test at a ratio of approx 60:20:20

This means approximately 117 videos in Train, 38 videos in Valid, 38 videos in Test. Also want to ensure that they are roughly proportional lengths

make train, valid, test at 60%, 20%, 20%

In [1]:

```
allVideosByPID = {
    111: [1, 3, 4],
    112: [1, 2, 3],
    113: [2, 3, 4, 5, 6],
    114: [1, 4, 5, 6],
    115: [2, 3, 5, 6],
    116: [1, 2, 3, 4, 5, 6],
    117: [1, 2, 3, 4, 5, 6],
    118: [1, 3, 4, 5, 6],
    119: [2, 4, 6],
    120: [1, 2, 3, 4],
    121: [1, 3, 5, 6],
    123: [1, 2, 3, 4, 5],
    124: [1, 2, 3, 6],
    127: [2, 3, 4, 5, 6],
    128: [2, 5, 6],
    129: [1, 2, 3, 4, 5, 6, 7],
    130: [1, 2, 4, 6],
    131: [1, 2, 3, 5, 6],
    134: [1, 3],
    135: [3],
    137: [1, 2, 4, 5, 6],
    141: [1, 6],
    142: [1, 2, 3],
    143: [3, 5],
    144: [1, 3],
    145: [1, 2, 3, 4],
    147: [1, 2, 3, 4, 5],
    149: [3, 6],
    151: [2, 6],
    153: [1, 2, 3, 4, 5, 6],
    154: [1, 4, 6],
    156: [1, 2, 3, 6],
    161: [1, 3, 4],
    162: [2, 3, 4, 5, 6],
    163: [1, 2, 5],
    164: [3, 4, 5],
    165: [1, 2, 3, 4, 5, 6, 7],
    167: [1, 2],
    168: [1],
    169: [2, 4],
    170: [2, 7],
    171: [1, 2, 3, 4, 5, 6],
    172: [1, 2, 3, 4, 5],
    173: [1, 3, 4, 6],
    174: [1, 2, 3, 5, 6, 7],
    178: [1, 2, 4, 5, 6],
    179: [1, 3, 4, 5, 6],
    180: [1, 2, 3, 4, 5, 6],
    181: [1, 2, 3, 4, 6]
}

## double check numVids = 193, yes.
# numVids = 0
# for thisKey in allVideosByPID:
#     #print str(thisKey) + ":" + str(allVideosByPID[thisKey])
#     numVids += len(allVideosByPID[thisKey])
# print numVids
```

```
vidLengths = {"111_1": 242, "111_3": 192, "111_4": 172,
              "112_1": 91, "112_2": 102, "112_3": 86,
              "113_2": 196, "113_3": 107, "113_4": 181, "113_5": 179, "113_6": 118,
              "114_1": 186, "114_4": 184, "114_5": 197, "114_6": 182,
              "115_2": 176, "115_3": 156, "115_5": 113, "115_6": 194,
              "116_1": 157, "116_2": 115, "116_3": 100, "116_4": 82, "116_5": 94, "1
              "117_1": 122, "117_2": 122, "117_3": 125, "117_4": 161, "117_5": 127,
              "118_1": 164, "118_3": 108, "118_4": 105, "118_5": 108, "118_6": 126,
              "119_2": 107, "119_4": 134, "119_6": 158,
              "120_1": 181, "120_2": 169, "120_3": 187, "120_4": 142,
              "121_1": 118, "121_3": 119, "121_5": 77, "121_6": 87,
              "123_1": 164, "123_2": 175, "123_3": 176, "123_4": 182, "123_5": 176,
              "124_1": 100, "124_2": 97, "124_3": 171, "124_6": 96,
              "127_2": 167, "127_3": 181, "127_4": 191, "127_5": 170, "127_6": 185,
              "128_2": 112, "128_5": 173, "128_6": 163,
              "129_1": 199, "129_2": 168, "129_3": 182, "129_4": 113, "129_5": 80, '
              "130_1": 152, "130_2": 115, "130_4": 175, "130_6": 112,
              "131_1": 133, "131_2": 153, "131_3": 139, "131_5": 105, "131_6": 100,
              "134_1": 176, "134_3": 178,
              "135_3": 153,
              "137_1": 187, "137_2": 182, "137_4": 174, "137_5": 171, "137_6": 163,
              "141_1": 60, "141_6": 68,
              "142_1": 69,  "142_2": 61, "142_3": 52,
              "143_3": 93, "143_5": 111,
              "144_1": 160, "144_3": 182,
              "145_1": 179, "145_2": 112, "145_3": 133, "145_4": 193,
              "147_1": 71, "147_2": 103, "147_3": 89, "147_4": 120, "147_5": 166,
              "149_3": 139, "149_6": 177,
              "151_2": 47, "151_6": 97,
              "153_1": 55, "153_2": 67, "153_3": 63, "153_4": 120, "153_5": 126, "15
              "154_1": 35, "154_4": 35,  "154_6": 48,
              "156_1": 166, "156_2": 194, "156_3": 140, "156_6": 189,
              "161_1": 181, "161_3": 134, "161_4": 177,
              "162_2": 137, "162_3": 171, "162_4": 154, "162_5": 147, "162_6": 145,
              "163_1": 159, "163_2": 131, "163_5": 178,
              "164_3": 176, "164_4": 92, "164_5": 106,
              "165_1": 111, "165_2": 131, "165_3": 104, "165_4": 115, "165_5": 83, '
              "167_1": 96, "167_2": 114,
              "168_1": 121,
              "169_2": 170, "169_4": 104,
              "170_2": 47, "170_7": 69,
              "171_1": 128, "171_2": 179, "171_3": 157, "171_4": 173, "171_5": 167,
              "172_1": 158, "172_2": 179, "172_3": 167, "172_4": 188, "172_5": 178,
              "173_1": 128, "173_3": 49, "173_4": 112, "173_6": 93,
              "174_1": 88, "174_2": 92, "174_3": 115, "174_5": 171, "174_6": 165,  '
              "178_1": 114, "178_2": 75, "178_4": 160, "178_5": 180, "178_6": 130,
              "179_1": 153, "179_3": 163, "179_4": 185, "179_5": 132, "179_6": 150,
              "180_1": 95, "180_2": 99, "180_3": 123, "180_4": 95, "180_5": 85, "180
              "181_1": 165, "181_2": 126, "181_3": 103, "181_4": 171, "181_6": 126 }

totalVidLengths = sum(vidLengths[thisVid] for thisVid in vidLengths)
```

In [2]:

```
TrainVideosNum = 117
ValidVideosNum = 38
TestVideosNum = 38

# First, want to choose some targets as held out in Validation and held out in Test
# We want some unique targets that are in Valid/Test that are not in Train.
# These are the following targets that have 1 or 2 videos. Split them equally into
#     134: [1, 3], /
#     135: [3], /
#     141: [1, 6], /
#     143: [3, 5], /
#     144: [1, 3], /
#     149: [3, 6], /
#     151: [2, 6], /
#     167: [1, 2], /
#     168: [1], /
#     169: [2, 4], /
#     170: [2, 7], /

TrainSet = []
ValidSet = ["134_1", "134_3", "143_3", "143_5", "149_3", "149_6", "167_1", "167_2",
TestSet = ["141_1", "141_6", "144_1", "144_3", "151_2", "151_6", "169_2", "169_4", "

allVideosByPID_Step2 = {
    111: [1, 3, 4],
    112: [1, 2, 3],
    113: [2, 3, 4, 5, 6],
    114: [1, 4, 5, 6],
    115: [2, 3, 5, 6],
    116: [1, 2, 3, 4, 5, 6],
    117: [1, 2, 3, 4, 5, 6],
    118: [1, 3, 4, 5, 6],
    119: [2, 4, 6],
    120: [1, 2, 3, 4],
    121: [1, 3, 5, 6],
    123: [1, 2, 3, 4, 5],
    124: [1, 2, 3, 6],
    127: [2, 3, 4, 5, 6],
    128: [2, 5, 6],
    129: [1, 2, 3, 4, 5, 6, 7],
    130: [1, 2, 4, 6],
    131: [1, 2, 3, 5, 6],
    137: [1, 2, 4, 5, 6],
    142: [1, 2, 3],
    145: [1, 2, 3, 4],
    147: [1, 2, 3, 4, 5],
    153: [1, 2, 3, 4, 5, 6],
    154: [1, 4, 6],
    156: [1, 2, 3, 6],
    161: [1, 3, 4],
    162: [2, 3, 4, 5, 6],
    163: [1, 2, 5],
    164: [3, 4, 5],
    165: [1, 2, 3, 4, 5, 6, 7],
    171: [1, 2, 3, 4, 5, 6],
    172: [1, 2, 3, 4, 5],
    173: [1, 3, 4, 6],
    174: [1, 2, 3, 5, 6, 7],
    178: [1, 2, 4, 5, 6],
```

```
    179: [1, 3, 4, 5, 6],
    180: [1, 2, 3, 4, 5, 6],
    181: [1, 2, 3, 4, 6]
}

allVidsSet = [str(thisKey) + "_" + str(thisValue)
              for thisKey in allVideosByPID_Step2 for thisValue in allVideosByPID_St
assignmentsLeft = ["Train" for _ in xrange(117)] + ["Valid" for _ in xrange(28)] + [
```

In [3]:

```
# Rest, split equally
import random
random.seed(113)
random.shuffle(assignmentsLeft)
random.shuffle(allVidsSet)

while (len(allVidsSet) > 0):
    whichAssignment = assignmentsLeft.pop()
    if whichAssignment == "Train":
        TrainSet.append(allVidsSet.pop())
    elif whichAssignment == "Valid":
        ValidSet.append(allVidsSet.pop())
    else:
        TestSet.append(allVidsSet.pop())

TrainSet.sort()
ValidSet.sort()
TestSet.sort()

TrainSetDuration = sum([vidLengths[thisVid] for thisVid in TrainSet])
ValidSetDuration = sum([vidLengths[thisVid] for thisVid in ValidSet])
TestSetDuration  = sum([vidLengths[thisVid] for thisVid in TestSet])

print "Train Length: " + str(TrainSetDuration) + ",  expected: " + str(int(0.60 * to
print "Valid Length: " + str(ValidSetDuration) + ",    expected: " + str(int(0.20 *
print "Test  Length: " + str(TestSetDuration) + ",    expected: " + str(int(0.20 * t
```

```
Train Length: 15973,  expected: 15662 diff: 311
Valid Length: 4981,    expected: 5220 diff: -239
Test  Length: 5150,    expected: 5220 diff: -70
```

In [4]:

```
print TrainSet
```

```
['111_4', '112_1', '112_3', '113_2', '113_3', '113_4', '113_6', '114_
4', '114_5', '115_2', '115_5', '115_6', '116_1', '116_3', '116_5', '11
6_6', '117_1', '117_2', '117_4', '117_5', '117_6', '118_1', '118_3',
'118_4', '118_5', '118_6', '119_2', '119_4', '119_6', '120_1', '120_
2', '120_3', '120_4', '121_5', '121_6', '123_1', '123_2', '123_5', '12
4_1', '124_2', '124_3', '124_6', '127_2', '127_3', '127_4', '127_5',
'128_5', '129_2', '129_3', '130_1', '130_2', '130_4', '130_6', '131_
1', '131_3', '131_5', '137_1', '137_2', '137_4', '137_6', '142_1', '14
2_3', '145_1', '145_3', '147_1', '147_3', '147_5', '153_2', '153_3',
'153_4', '153_5', '153_6', '154_1', '154_6', '156_3', '156_6', '161_
1', '161_3', '161_4', '162_3', '162_4', '162_5', '163_1', '163_2', '16
3_5', '164_3', '164_5', '165_1', '165_2', '165_3', '165_4', '165_5',
'165_6', '165_7', '171_2', '171_3', '171_6', '172_2', '173_1', '173_
4', '173_6', '174_3', '174_7', '178_1', '178_5', '178_6', '179_1', '17
9_3', '179_4', '179_6', '180_2', '180_3', '180_5', '180_6', '181_2',
'181_3', '181_6']
```

In [5]:

```
print ValidSet
```

```
['111_3', '113_5', '114_1', '117_3', '121_3', '123_4', '127_6', '128_
2', '128_6', '129_1', '129_4', '129_5', '129_6', '129_7', '131_6', '13
4_1', '134_3', '143_3', '143_5', '145_2', '147_4', '149_3', '149_6',
'154_4', '156_1', '162_2', '162_6', '164_4', '167_1', '167_2', '170_
2', '170_7', '171_4', '172_5', '173_3', '174_2', '178_2', '181_1']
```

In [6]:

```
print TestSet
```

```
['111_1', '112_2', '114_6', '115_3', '116_2', '116_4', '121_1', '123_
3', '131_2', '135_3', '137_5', '141_1', '141_6', '142_2', '144_1', '14
4_3', '145_4', '147_2', '151_2', '151_6', '153_1', '156_2', '168_1',
'169_2', '169_4', '171_1', '171_5', '172_1', '172_3', '172_4', '174_
1', '174_5', '174_6', '178_4', '179_5', '180_1', '180_4', '181_4']
```

In [7]:

```
parsed = open("TrainSetAssignments/TrainSet.csv", 'w')
parsed.write('\n'.join(TrainSet))
parsed.close()
parsed = open("TrainSetAssignments/ValidSet.csv", 'w')
parsed.write('\n'.join(ValidSet))
parsed.close()
parsed = open("TrainSetAssignments/TestSet.csv", 'w')
parsed.write('\n'.join(TestSet))
parsed.close()
```

In [8]:

```
# This just helps to sort the files from "UNSORTED" into the appropriate folders

import os
import shutil

#

# UNSORTED_DIRECTORY = "features/Unsorted/acoustic/"
# TRAIN_DIRECTORY = "features/Train/acoustic/"
# VALID_DIRECTORY = "features/Valid/acoustic/"
# TEST_DIRECTORY = "features/Test/acoustic/"

UNSORTED_DIRECTORY = "features/Unsorted/emotient/"
TRAIN_DIRECTORY = "features/Train/emotient/"
VALID_DIRECTORY = "features/Valid/emotient/"
TEST_DIRECTORY = "features/Test/emotient/"


for thisFilename in os.listdir(UNSORTED_DIRECTORY):
    if(thisFilename == ".DS_Store"):
        os.remove(UNSORTED_DIRECTORY + ".DS_Store")
    else:
        vidID = thisFilename[2:6] + thisFilename[9]
        if vidID in TrainSet:
            shutil.move(UNSORTED_DIRECTORY + thisFilename, TRAIN_DIRECTORY + thisFil
        elif vidID in ValidSet:
            shutil.move(UNSORTED_DIRECTORY + thisFilename, VALID_DIRECTORY + thisFil
        elif vidID in TestSet:
            shutil.move(UNSORTED_DIRECTORY + thisFilename, TEST_DIRECTORY + thisFile
        else:
            print "Not found! : " + thisFilename
```

In [ ]:

```

```

In [ ]:

```

```