|  |
| --- |
| Isolated Digit Recognition Using Dynamic Time Warping Algorithm |
|  |
| **Sowmya Voona** |
| **6/7/2016** |

|  |
| --- |
|  |

# Summary

This report presents how a speech recognition system can use DTW algorithm to recognize isolated digits uttered by a speaker.

# Introduction

Speech recognition system basically does extraction of features of an utterance and classify it using any algorithm (here, we use DTW). In this case, MFCC was used to extract feature vectors of speech signals. Since same word/digit uttered by speakers may have different duration (i.e., speaking speeds), we use DTW to find out the similarity between different temporal sequences. The DTW measures a distance like quantity between two sequences. Using this score between an unknown data and known data, we classify the unknown data to the class with which it has least score.

# DTW Algorithm For Isolated Digit Recognition

**Input:** Sequence of feature vectors for each digit uttered by a speaker are stored in a file in the following format: (mfcc file)

38 98

-1.129163e+01 8.201429e+00 -4.876589e+00 ....

-1.112325e+01 8.435195e+00 -4.208691e+00 ....

-1.074037e+01 7.301194e+00 -5.066467e+00 ....

First Line represents the dimension of each feature vector and number of feature vectors.

Here, 38 is dimension of feature vector and 98 is the number of feature vectors extracted for a digit uttered.

**Train Data**: File containing list of mfcc files of recognized digit spoken by different speakers.

**Test Data**: File containing list of mfcc files of unrecognized digit spoken by different speakers.

**Algorithm:** Using the input we find out distance matrix and cost matrix as follows:

For a test file of digit to be recognized, we first find out the score/similarity between the test file and each train file.

**Distance Matrix**: This matrix contains Euclidian distance between each feature vector of test file and each feature vector of train file. So it will be M\*N matrix where M is number of feature vectors of test file while N is the number of feature vectors of train file.

Lesser distance between two feature vectors indicates that these two feature vectors can be matched together. Diagonal entries are likely to have low distances.

Let X = {x1, x2, x3 ..xm } be the test feature vector, Y = {y1,y2,y3..yn } be the train feature vector.

Distance (i, j) = |xi-yj|

**Cost Matrix:** Obtaining cost matrix implicitly means warping the path. We try to find out the path of minimum distance from (1, 1) to (M, N).

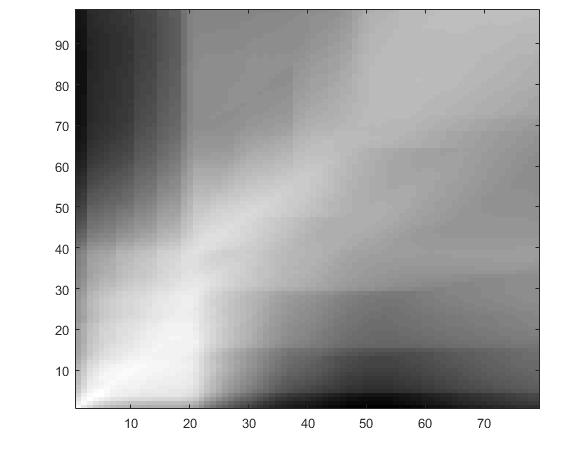
Since the sequence is temporal, we cannot go back and hence the path only moves in forward direction. Either (i+1, j) or (i, j+1) or (i+1, j+1).

Cost (1, 1) = distance (1, 1)

Cost (i, 1) = distance (i, 1) + distance(i-1, 1)

Cost (1, j) = distance (1, j) + distance(1, j-1)

Cost (i, j) = distance (i, j) + minimum{cost(i-1,j),cost(i,j-1), cost(i-1,j-1)}



Plot of cost matrix between test file: ft\_1a.mfcc and train file: st\_1b.mfcc

It is clear from the picture that minimum path lies along the diagonal.

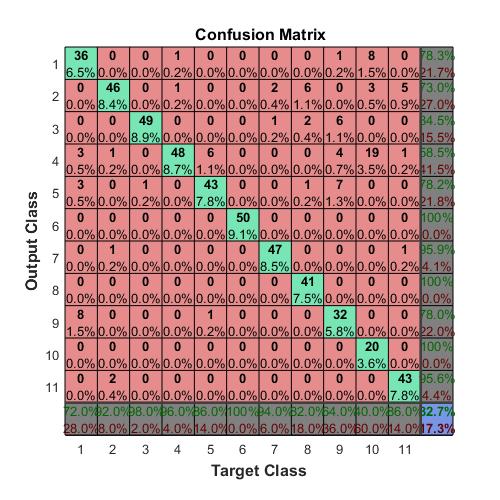
**Score**: Score between test file and train file is taken as: cost(M,N)/(M+N)

Find out the score between test file and each train file of digit uttered by a speaker. The test file is classified into a digit for which score obtained was minimum. If we have train files of many speakers, we conduct voting between each digit recognized against each speaker.

# Results

Scores were calculated between several test files and train files of different speakers. Using these scores test files were recognized as digits. Confusion matrix has been plotted and the results were as follows:

Number of test files: 550 Number of Digits: 11 (includes letter ‘O’) There are 50 test cases for each digit

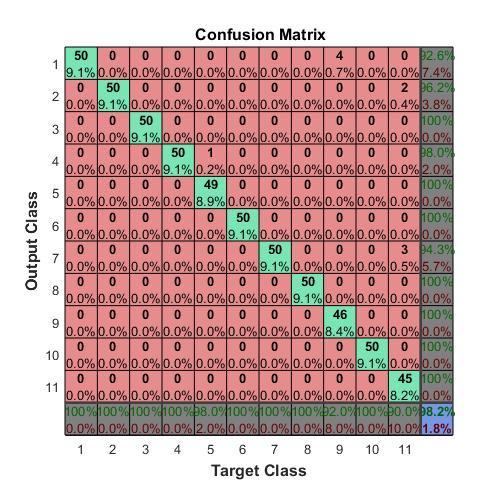


Plot of confusion matrix where Number of train files: 11 and Number of speakers: 1

Element (i,j) represents number of test cases that were recognized as ‘i’ which are actually ‘j’. Diagonal entries indicate how many test cases were recognized correctly.

(1,1):36 indicates that out of 50,36 were recognized correctly as digit-1. Here digit 10 was confused mostly with 4 (19) and 1(8). **Accuracy in this case is 82.7%**

On increasing the number of train files and speakers, digit recognition becomes more accurate.



Plot of confusion matrix where Number of train files: 550 and Number of speakers: 500

In this case, most of the diagonal entries are 50 which indicates almost all the test files have been recognized correctly.

Digit 9 has been confused with digit 1 and Digit 11 has been confused with 7.

This happens because the phoneme sequence for digits 9 and 1 are similar which reduces the distance between their feature vectors, thereby minimizing the cost.

Here **accuracy is 98.2%**

Thus, more the number of train files , more is the accuracy.

