Homework #4

EECS 498 – Intelligent Interactive Systems

# Overview

This homework will step through the basics underlying speech recognition. The problems in this assignment are more open-ended than the previous two assignments. This is intentional (and is factored into grading). The goal is to get you thinking about (and implementing!) speech modeling and the limitations of simpler (dynamic time warping) approaches.

# Problems

## Problem 1: Conceptual [2 pts]

Explain why the word “think” is more similar to the word “thing” than to the word “thin”.

## Problem 2: Dynamic Time Warping [8 pts]

As we saw in class, Dynamic Time Warping (DTW) can be used in the context of small vocabulary, isolated word recognition.

1. Why is accurate segmentation crucial?

We defined templates describing how speech features varied in time and w e hypothesized that we could understand the “similarity” between an input speech signal (Xin) and a reference speech signal (Xref) by calculating the total distortion between the two:

latex-image-1.pdf (1)

1. Why is this technique called time warping?
2. Why do we use the boundary conditions requiring that *i1* = 1, *ik* = n and *j1* = 1 and *jk* = m?

Consider two sequences: A = [1 2 4 9] and B = [1 2 2 4 1]. Use the symmetric DTW approach and the following distance measure: the absolute value of the difference between frames *i* and *j*. The distortion calculation can be calculated as slide 53 in lecture 8. Turn in: (1) the matrix of frame-level distances, (2) the matrix used to accumulate distortion, (3) a matrix of pointers (visualize with arrows as in the Lecture 8, Slide 54) that allows you to reconstruct the warping path.

1. What is the total distortion between A and B?
2. What is the final warping path?

## Problem 3: Dynamic Time Warping for Speech [30 pts]

Now, we want to use DTW to classify the content of a set of audio samples. You are given a series of mfcc values (1 coefficient, extracted at 1000 Hz) for a series of digits. You have a series of audio clips. Each digit is recorded six times. The first five recordings are training data (e.g., one1, one2, one3, one4, one5), the final recording is testing data (one6). The raw audio files are located in the folder Audio. I have already extracted features from the audio data. The features are located in the folder Features. The Features folder contains two sub-folders folders: training and testing. The features include: harmonic-to-noise ratio (hnr), intensity, pitch, mfcc, mfb.

In all cases:

* The distance between frame i of Xin and frame j of Xref is: d(i, j) = [ Xin(i) – Xref(j) ]2
* The distortion can be calculated as in slide 53 from lecture 8 (symmetric!).

Code your implementation in the language of your choice (some Matlab infrastructure has been provided). The code should run on a CAEN computer without additional dependencies. **Note: it is not necessary to provide the warping path! Just provide the accumulated distortion (much simpler, and in this case, much more useful!).**

1. The following table discusses the results of models **trained and tested on the training data only**. The first column is the training row number (if you are using data.mat). The second column is the file name (if you are not). For each highlighted point, find the three closest training templates using DTW operating on the **intensity** features of the data. Notes: (1) Do not compare the template to itself! (2) Closest “1” is the closest template to the training template, etc. What can you conclude?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training ID** | **Audio File** | **Distortion (D)** | **Closest Labels** | | |
| **1** | **2** | **3** |
| **1** | one1 | one5: 0.755541  one4: 0.889214  one3: 0.902575 | one5 | one4 | one3 |
| **7** | two2 | two5: 1.654490  two3: 3.004929  two4: 3.352084 | two5 | two3 | two4 |
| **13** | three3 | three2: 0.568741  three5: 0.907649  three4: 1.899511 | three2 | three5 | three4 |
| **16** | four1 | four3: 1.505547  four2: 1.606117  four5: 2.854807 | four3 | four2 | four5 |
| **22** | five2 |  | five3 | five5 | five4 |
| **28** | six3 |  |  |  |  |
| **34** | seven3 |  |  |  |  |
| **36** | eight1 |  |  |  |  |
| **41** | nine1 |  |  |  |  |
| **50** | ten5 |  |  |  |  |

1. Repeat using an amalgam of **all features**. Let the total distortion be the sum of each feature-specific distortion (weighted using the symmetric DTW). What can you conclude?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Training ID** | **Audio File** | **Distortion (D)** | **Closest Labels** | | |
| **1** | **2** | **3** |
| **1** | one1 |  |  |  |  |
| **7** | two2 |  |  |  |  |
| **13** | three3 |  |  |  |  |
| **16** | four1 |  |  |  |  |
| **22** | five2 |  |  |  |  |
| **28** | six3 |  |  |  |  |
| **34** | seven3 |  |  |  |  |
| **36** | eight1 |  |  |  |  |
| **41** | nine1 |  |  |  |  |
| **50** | ten5 |  |  |  |  |

1. The following table discusses the results of models **trained on the training data and tested on the testing data**. The first column is the testing row number (if you are using data.mat). The second column is the file name (if you are not). Find the closest three points (where “1” is the closest point). What can you conclude? What could be done to further improve the results?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Testing ID** | **Audio File** | **Distortion (D)** | **Closest Labels** | | |
| **1** | **2** | **3** |
| **1** | one6 |  |  |  |  |
| **2** | two6 |  |  |  |  |
| **3** | three6 |  |  |  |  |
| **4** | four6 |  |  |  |  |
| **5** | five6 |  |  |  |  |
| **6** | six6 |  |  |  |  |
| **7** | seven6 |  |  |  |  |
| **8** | eight6 |  |  |  |  |
| **9** | nine6 |  |  |  |  |
| **10** | ten6 |  |  |  |  |

Note: the class labels are given here – in general this is not the case for classification problems.