Introduction to Digital Speech Processing – Homework 1

資工三 B05902023 李澤諺

1. Environment

(1) Environment：Ubuntu 16.04

(2) Language：C

(3) Compiler：gcc

※已確認過可以在 CSIE workstation 上編譯並執行

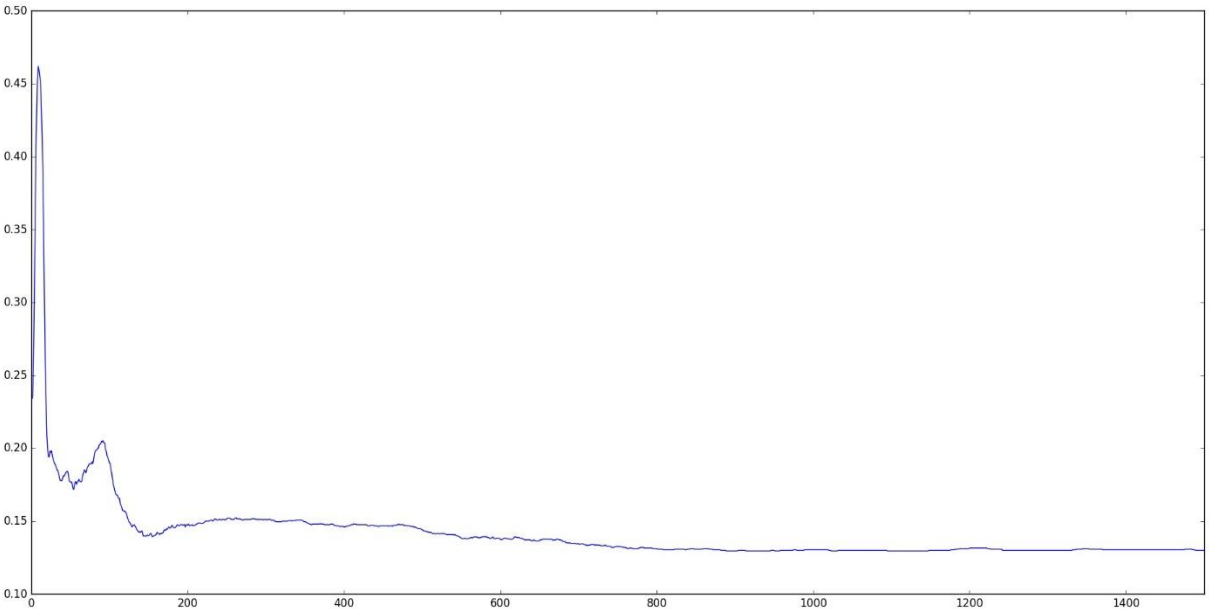
2. Usage

```
#  用 make 指令將 train.c 和 test.c 編譯為執行檔 train 和 test
>  make

#  執行檔 train 的命令列參數分別為：iteration 次數、用來初始化該 HMM 的
   檔案所在的路徑、用來訓練該 HMM 的 training data 所在的路徑、用來儲
   存該 HMM 參數的檔案所在的路徑
>  ./train  900  model_init.txt  seq_model_01.txt  model_01.txt
>  ./train  900  model_init.txt  seq_model_02.txt  model_02.txt
>  ./train  900  model_init.txt  seq_model_03.txt  model_03.txt
>  ./train  900  model_init.txt  seq_model_04.txt  model_04.txt
>  ./train  900  model_init.txt  seq_model_05.txt  model_05.txt

#  執行檔 test 的命令列參數分別為：記錄所有 HMM 參數所在位置的檔案之
   路徑、testing data 所在的路徑、用來儲存測試結果的檔案所在的路徑
>  ./test  modellist.txt  testing_data1.txt  result1.txt
>  ./test  modellist.txt  testing_data2.txt  result2.txt

#  使用 make clean 指令將執行檔 train 和 test 刪除
>  make clean
```

3. Summary

其實本次作業就是一個 machine learning 的問題：想辦法讓電腦在看到 training data 中的 observation sequence 後，訓練出各個 HMM，用以分辨不同的 observation sequence 是由哪一個 model 所產生。既然如此，可以預料到當 iteration 次數越多時，各個 HMM 在其 training data 上的表現就會越好，但與此同時，若 iteration 次數過多，便很有可能會發生 overfitting。若將本次作業中提供的 testing_data1.txt 視為 validation data，則利用其所計算出來的 accuracy，便能用來反映我們訓練出來的 HMM 在其它 testing data 上大致可得的 accuracy。

以下圖表為每次 iteration 過後，便計算當前訓練出來的 HMM 在 testing_data1.txt 上所得到的 error rate(橫軸為 iteration 次數，最多 1500 次，縱軸則為 error rate)：



由此可以看出，當 iteration 達到 800 次以上時，訓練出的 HMM 在 testing_data1.txt 上的 error rate 已趨於平緩(約為 0.13，即 accuracy 約為 0.87)，並且再沒有升高，因此推斷，iteration 約為 800 次時，HMM 在其餘 testing data 上的表現大致上已經不錯，而當 iteration 極大(接近 1500 次)時，並未觀察到 error rate 上升，即 overfitting 的現象，我推測是因為 testing_data1.txt 和 seq_model_01～05.txt 中的 observation sequence 皆為自極為相近的機率分布所產生，因此有著相同的 bias，故當 iteration 極大，訓練出的 HMM 在 seq_model_01～05.txt 上表現很好時，其在 testing_data1.txt 上的表現也便會很好，另一種可能的解釋是因為 iteration 的次數仍不夠多，所以還未觀察到 overfitting，但不論如何，當 iteration 約為 800 次時，訓練出的 HMM 表現已經不錯，故我在本次作業要上傳的檔案中，model_01～05.txt 儲存的即是使用 iteration 為 900 次所訓練出來的 HMM 的參數。