Digital Signal Processing Laboratory
Lab 12 Baby Sounds challenge

---

1. Introduction.

現今的醫療照護主要都是以人力為主，不過因為護士人數不夠、工時長導致精神不濟等因素，常常會讓嬰兒的照顧不周；在家裡也會遇到一樣的問題，長輩可能在廚房煮飯、在陽台曬衣服，而不知道房間裡嬰兒的狀況。因此若在嬰兒身邊有一些能偵測嬰兒狀態的工具，將會大幅降低醫療體系、家庭照顧者的負擔。

2. Lab objectives.

此次實驗要我們對不同種類的嬰兒聲音做分類辨識。透過取 MFCC，將各種不同的聲音如 canonical、non-canonical、laughing、crying、junk 等等進行特徵萃取，接著使用 SVM(支援向量機)做分類，調整其不同的參數做 cross validation，取 accuracy 最好的那個來當作最終的模型，最後用這個模型對 testing data 做預測，並將結果上傳 kaggle 進行辨識率的競賽。

3. Method.

一開始我先 load 所有檔案位置並對它們做 sorting、labeling 以及使用 librosa 套件將每個檔案的音訊資料取出，接著透過前幾次 lab 學到的知識，實作 MFCC 流程的每一步驟，包括 pre-emphasis、STFT、mel banks filtering、取 log、以及使用 dct 函數加強頻率表現，最後得出每個音訊檔的 MFCC。

由於輸入進 SVM 的資料需要是一維的向量，但 MFCC 是二維的矩陣，因此我們需要對 MFCC 陣列取特徵值以降低維度，我對每個 MFCC 特徵取平均與標準差，因此最終的 feature 數會有 40 個。

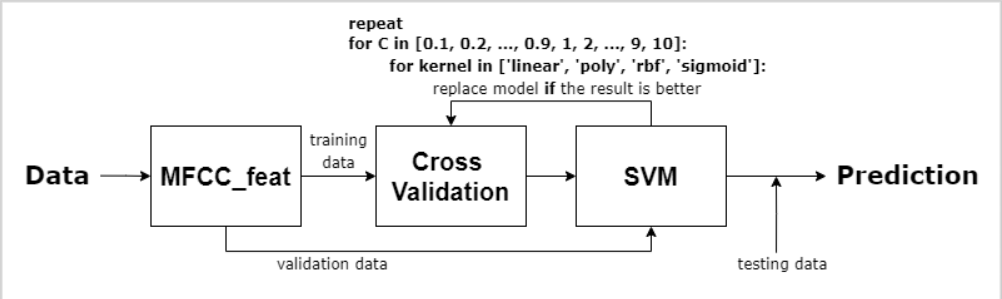
接著將資料輸入至 SVC 函數，SVC 函數是用來做 classification 的 SVM，有 kernel 和 C (Regularization parameter)可以調整，kernel 有'linear'、'poly'、'rbf' 和'sigmoid'這四種可以換，須根據資料的特性做最佳化，而 C 參數為懲罰係數；另外若 kernel 設定為'poly'時，還能調整參數 degree，代表多項式函數的最高次數。經過好幾次 trial and error，我發現參數設定在[kernel='rbf', C=2]時，validation 的 accuracy 會最好，因此選定此設定。

另外由於模型會受到 training set 與 validation set 的分法影響，因此在訓練 SVM 模型之前，我利用 KFold 函數做五次 cross validation，以得到不同的 training 和 validation data，並且對兩者做 normalize 以降低資料中極端值對模型訓練的影響，接著訓練完模型後預測五次 validation data 結果，取平均當作這個模型的準確率。

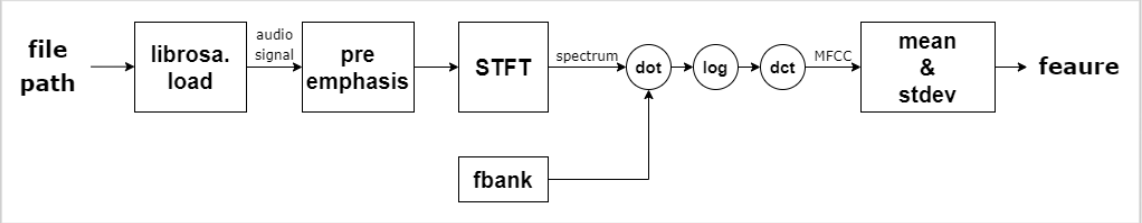
最後使用 validation 結果最好的模型預測 testing data，將得到的結果儲存成 csv 檔，上傳 kaggle 後得到 67.486%的準確度。

4. Flow chart.

主程式之流程圖：



副程式 MFCC_feat 之流程圖：



5. Results and discussion of report questions.

(1) 關於 MFCC：

在用 MFCC 函數算出其特徵時，我設定 MFCC 特徵的個數為 20。MFCC 的特徵數增加會讓辨識率增加，不過加到了一定程度後就沒甚麼影響了，而且反而會增加計算量與時間，因此選在中間值 20。

(2) 關於 SVM：

SVM 有 kernel、C、degree 等參數可以調整。kernel 有 linear、poly、rbf 和 sigmoid 這四種可以選，這需要根據資料的特性做最佳化；而 C 參數 (Regularization parameter)為懲罰係數，愈大代表錯誤的容忍程度愈低，在 training dataset 樣本中會區分的愈精細，因此設定太大容易造成 overfitting 的問題。反之若設定太小則會造成 underfitting。參數 degree 只有在 kernel 為 poly 時才有效，它用來調整 polynomial 的最高次數，也是需要針對資料的分布去做最佳化。

而我透過 for 迴圈，將 kernel 的所有可能、C 從 0.1 到 10、degree 從 2 到 5 這些可能全部都做了一次 cross validation，並找出結果最好的一組參數，也就是[kernel='rbf', C=2]，預測正確率達到 72%。然而，對於測試資料這樣的模型只達到 67%的辨識率，我認為主要原因是對於訓練資料 overfitting 了，因此可能需要降低 C 的值、或是增加訓練資料的變異性。

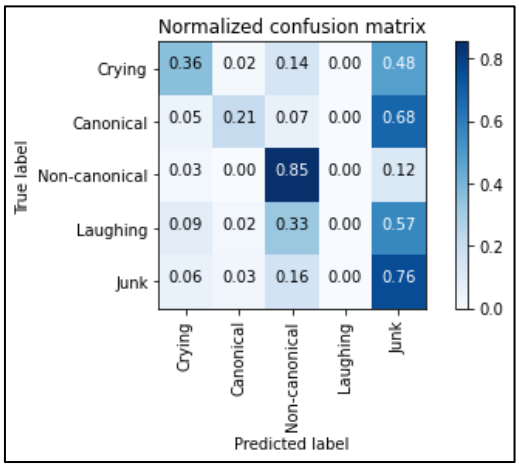
(3) 關於 dataset：

下表是訓練資料集中每個類別佔總訓練資料數的比例，可以看出五種類別的佔比差異極大，這樣會造成在訓練模型時不夠完整，例如 Crying、Laughing 和 Canonical 的資料量過少，導致這個類別沒有足夠多的資料變異，因此訓練不起來這個類別的辨識，這也是造成最終 SVM 辨識率不佳的原因之一。

Crying	Canonical	Non- canonical	Laughing	Junk
6.08 %	11.11 %	35.96 %	1.15 %	45.70 %

(4) 關於 result：

訓練資料的 validation result 顯示於下圖，由 confusion matrix 可以看出 Non-canonical 類別的辨識結果最好，可以達到 85%、而 Junk 類別也有 76%會正確判成 Junk；然而剩下 Crying、Canonical、Laughing 這三個類別就辨識的非常糟，都會有一大部分判成 Junk，剛好這三類在前面有提到資料數量都過少，也因此造成這樣的結果。



6. Conclusion.

這次實驗利用 MFCC 取特徵、再透過調整參數對 SVM 模型做最佳化，讓我們實作出嬰兒各種聲音的辨識，也讓我對於音訊處理的流程有了更深的認識。另外，我也了解到模型在訓練時達到的準確率不一定能出現在測試資料上，因為在訓練模型時為了使結果最好，常常會不小心讓模型對於訓練資料 overfitting，像我訓練玩模型發現準確率有達到 72%的時候以為我在 kaggle 上可以排第一了，但沒想到最後結果只有 67%而已。而我認為其他人可以做到 70%應該是因為取對了 feature，而不識相我只取了 mean 和 stdev，但我也實在想不到還有什麼統計的參數可以當作特徵了。可見在整個音訊辨識的處理流程中，每個步驟都會對預測結果造成不小的影響。