國立台北大學資訊工程學系專題報告

**Identifying Speaker with Pitch and Timbre**

專題組員：范齊萱、鄧意頻、尤家歡、莊宜珊

專題編號：PRJ-NTPUCSIE-105-007

執行時間：105年09月至106年06月

1. **摘要**

聲紋辨識是一種無接觸，無干擾並容易使用的生物辨識。在本研究中，自相關函數及兩次傅利葉轉換出來的調變聲紋圖等技術會用來找出聲紋的特色及相應數據，通過計算不同音訊的相關係數並進行比較，便能根據不同的聲紋特色辨認出說話人。本研究的實驗所設定的聲音週期閥值為90.7毫秒，能100%正確的分辨出說話人性別；另外以實驗規模為七個人，比較相同及不同人的元音錄音與數字0-9錄音的相關係數皆能100%正確的辨認出說話人。

1. **簡介**

每個人都有獨特的聲紋、指紋、虹膜等等。在指紋辨識開始變得不安全後，虹膜辨識及語音辨識開始受到重視，然而獲取有聲紋的音訊之途徑比較方便及自然、成本低，語音辨識的演算法複雜度也較低。因此聲紋辨識雖然目前在市場上不普遍流行，但它具有很大的發展潛力。

聲音有三個基本元素，音色(Timbre)、音量(Volume)以及音高(Pitch)。當聲音通過時，嘴唇與舌頭相對位置和形狀的不同，就會產生不同的音色。音量為聲音的大小，當肺部壓縮力愈大，音量則會跟著愈大。聲音的高低取決於每個人聲帶的張力，如果聲帶張力愈大，則聲門愈小，導致聲帶振動愈快，使音高愈高。本研究會用到自相關函數去找出音訊聲音的週期及頻率，並使用到調變聲紋圖找出聲音的音色及音高。其中每個人聲音的特質反映於音高及音色中，為此本專題設計了一套系統來進行說話人的辨認。

1. **專題進行方式**

我們的專題分為兩個部分：

1. 藉由音高做男女性聲音之區分
2. 利用調變聲譜圖同時做音高與音色之區分

以下針對每個部份的說明實驗方法。

* 1. 藉由音高(pitch)做男女性聲音之區分

影響男女性之音高主要因素為男性與女性其聲帶的長短差異，男性的聲帶較長導致聲帶震動較慢(頻率低)，因此音高較低；女性的聲帶較短導致聲帶震動較快(頻率高)，因此音高較高。我們使用了以下四個步驟來完成這部分的區分。

圖表 1 專題第一部分流程圖

1. 計算自相關係數

自相關(Autocorrelation)，也叫序列相關，是一個訊號於其自身在不同時間點的互相關。而自相關函數(Auto-correlation Function, ACF) 是指一個訊號在兩次觀察之間的相似度對它們之間時間差的函數。

方程式 1 自相關函數(ACF)

當自相關係數(ACF值)愈高時，表示其相關性也愈高，並且如果訊號有週期性，其ACF值也具週期性。因此我們利用ACF值來找出音訊的基本週期，經過音訊之取樣率(sample rate)轉換後並可以得到聲音的音高(基本頻率)。

我們先以25毫秒(ms)作為音框大小，取樣率為每秒16,000個樣本，因此一個音框包含400個樣本(L=400)，並以10毫秒作為樣本音框與測試音框的時間差，經過取樣率換算後得出時間差為160個樣本數，再進行該音框之ACF值計算。

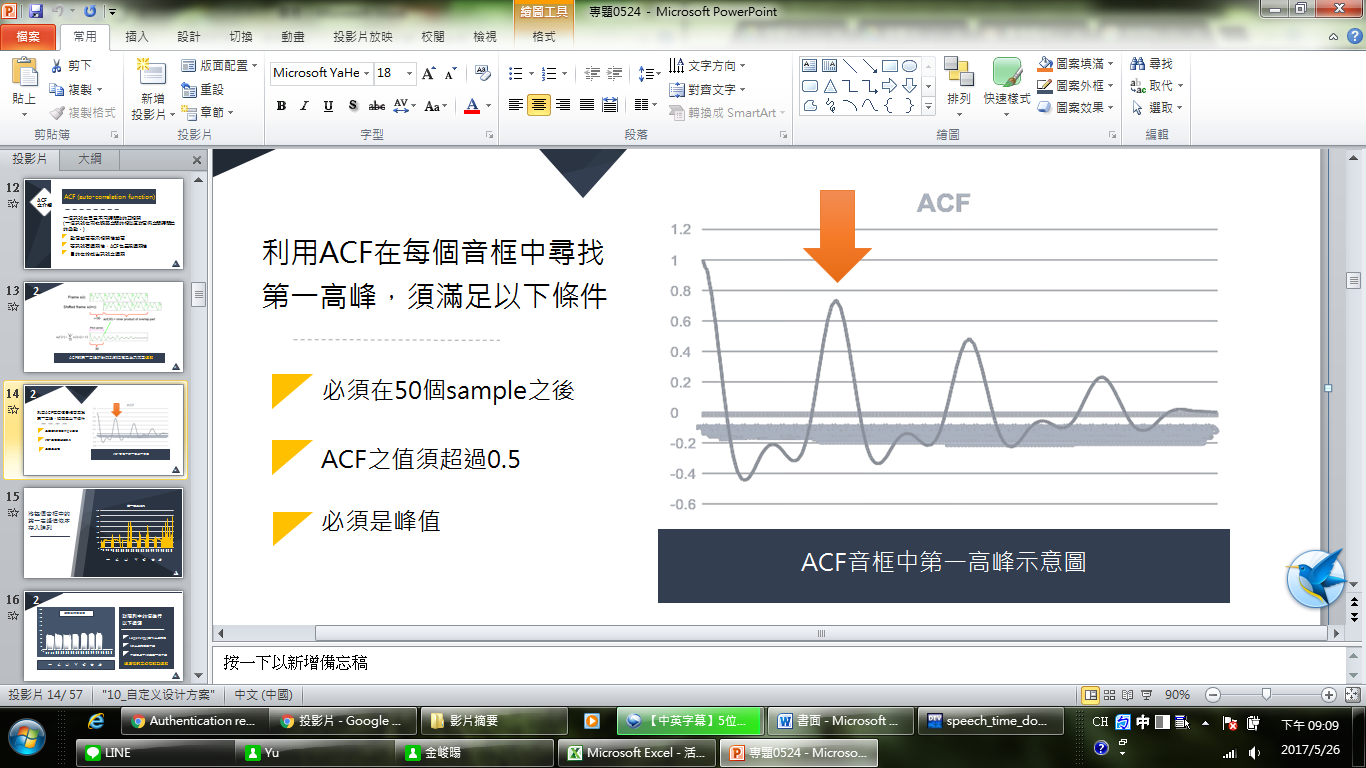
1. 找出自相關的第一高峰值

當訊號有週期性，其自相關係數也具週期性，因此自相關係數的第一高峰所對應到的時間為此音訊的週期，如圖表2。而搜尋第一高峰，我們利用了以下三項條件。

首先我們預期人聲的基本頻率會在300赫茲(Hz)以下，換算成週期約為33毫秒以上，再利用音檔的取樣率計算得出第一高峰位置應約在50個樣本之後，因此我們在每個音框中的50個樣本之後開始尋找。

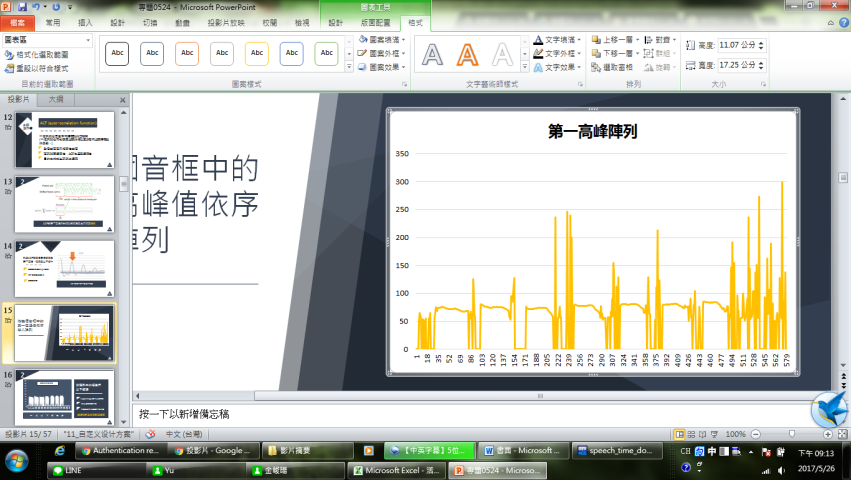
第二，自相關係數必須超過0.5，若低於0.5，則相關性太低不採用。

最後一項條件為此數值必須為峰值。



圖表 2 ACF第一高峰值示意圖

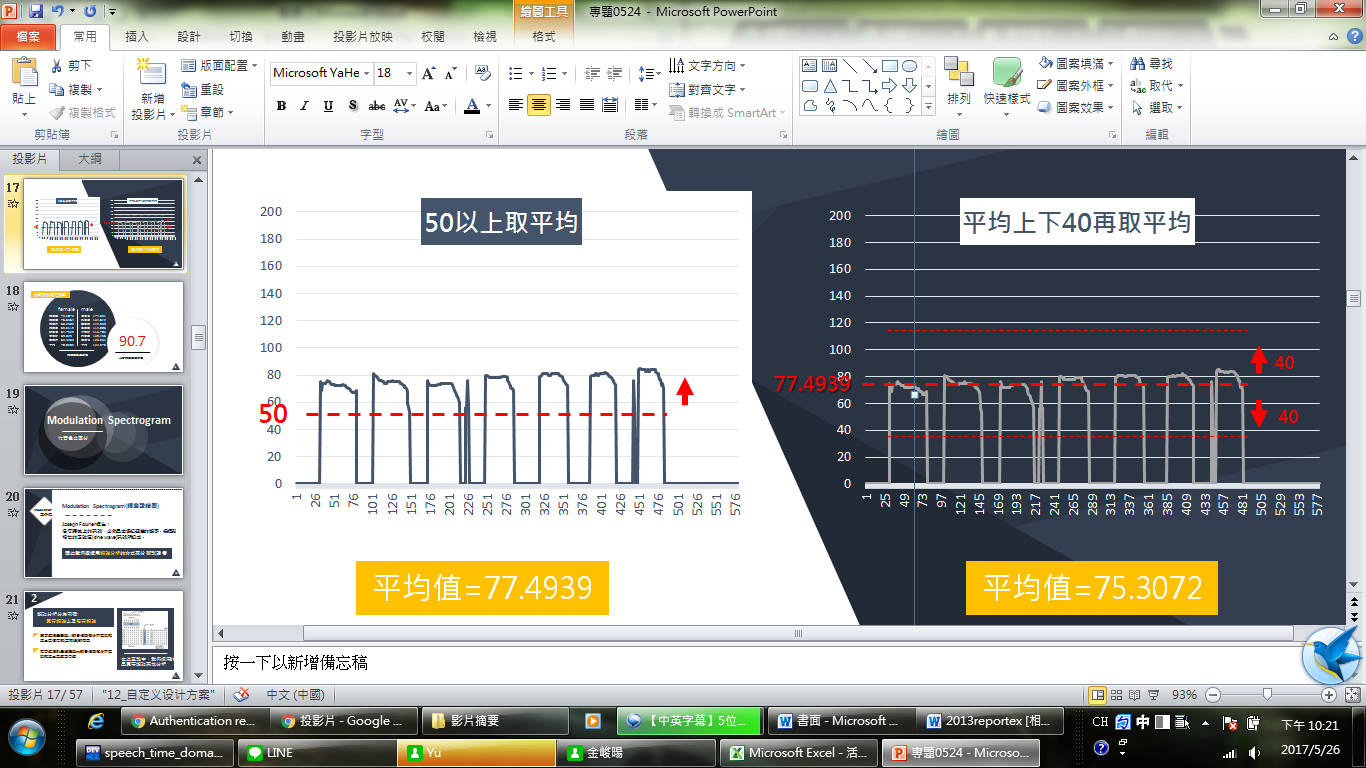
由以上條件得出每個音框的第一高峰值，並將這些數值所對應的時間依序存入陣列中，如圖表3所示。



圖表 3 ACF第一高峰值陣列

1. 過濾雜音

由於背景雜音有時也具週期性，因此我們先利用聲音的能量大小來做刪去的動作。我們將能量取對數後，經過實驗來找出雜訊能量大約的數值，並將閥值定在-5。若音訊之能量取對數小於-5，則此位置對應到第一高峰陣列中的值設為0，並於第一高峰陣列中數值為50以上的做平均。



圖表 4 利用ACD第一高峰陣列做第一次過濾

得到平均值後於此平均值以上40及40以下範圍內的值再取一次平均。



圖表 5 利用ACD第一高峰陣列做第二次過濾

得到的平均值即為此音檔的基本頻率之週期。再利用音檔之取樣率求出其頻率。

1. 比較

我們先收集不同性別的音檔資料並將這些聲音設定為訓練集 (train set)，從訓練集中找出男性與女性聲音基本頻率之週期，我們發現男性的聲音週期皆大於100；而女性的皆小於80。我們分別計算出男性與女性聲音週期的平均，並將兩個平均的中間值90.7設定為判別男性與女性的閥值。



圖表 6 判斷男女之週期

* 1. 利用調變聲譜圖(Modulation Spectrogram)同時做音高與音色之區分

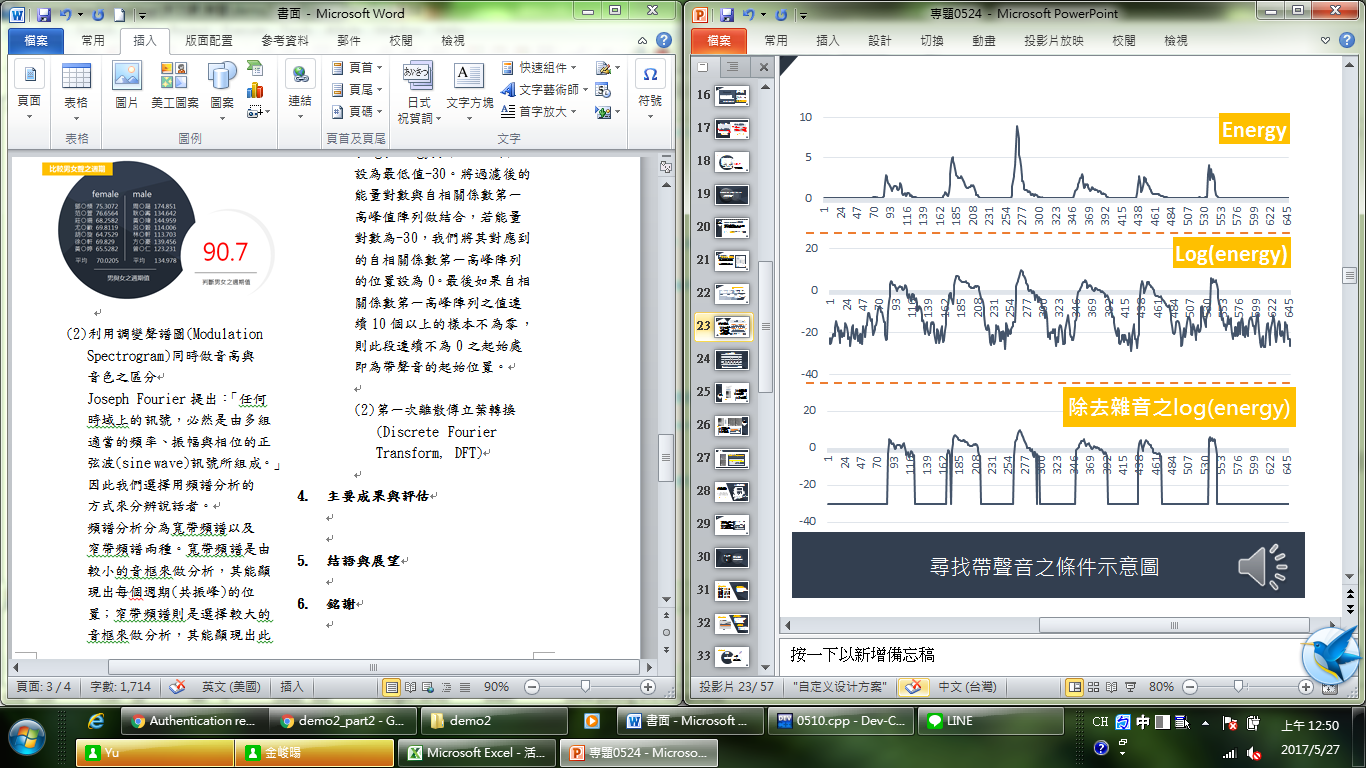
Joseph Fourier提出：「任何時域上的訊號，必然是由多組適當的頻率、振幅與相位的正弦波(sine wave)訊號所組成。」因此我們選擇用頻譜分析的方式來分辨說話者。

頻譜分析分為寬帶頻譜以及窄帶頻譜兩種。寬帶頻譜是由較小的音框來做分析，其能顯現出每個週期(共振峰)的位置；窄帶頻譜則是選擇較大的音框來做分析，其能顯現出此音框基頻之倍頻。而在此實驗中，我們採用寬帶頻譜做分析。在這部分的研究我們分為以下四個步驟。

圖表 7 專題第二部分流程圖

1. 尋找帶聲音(voiced sound)位置

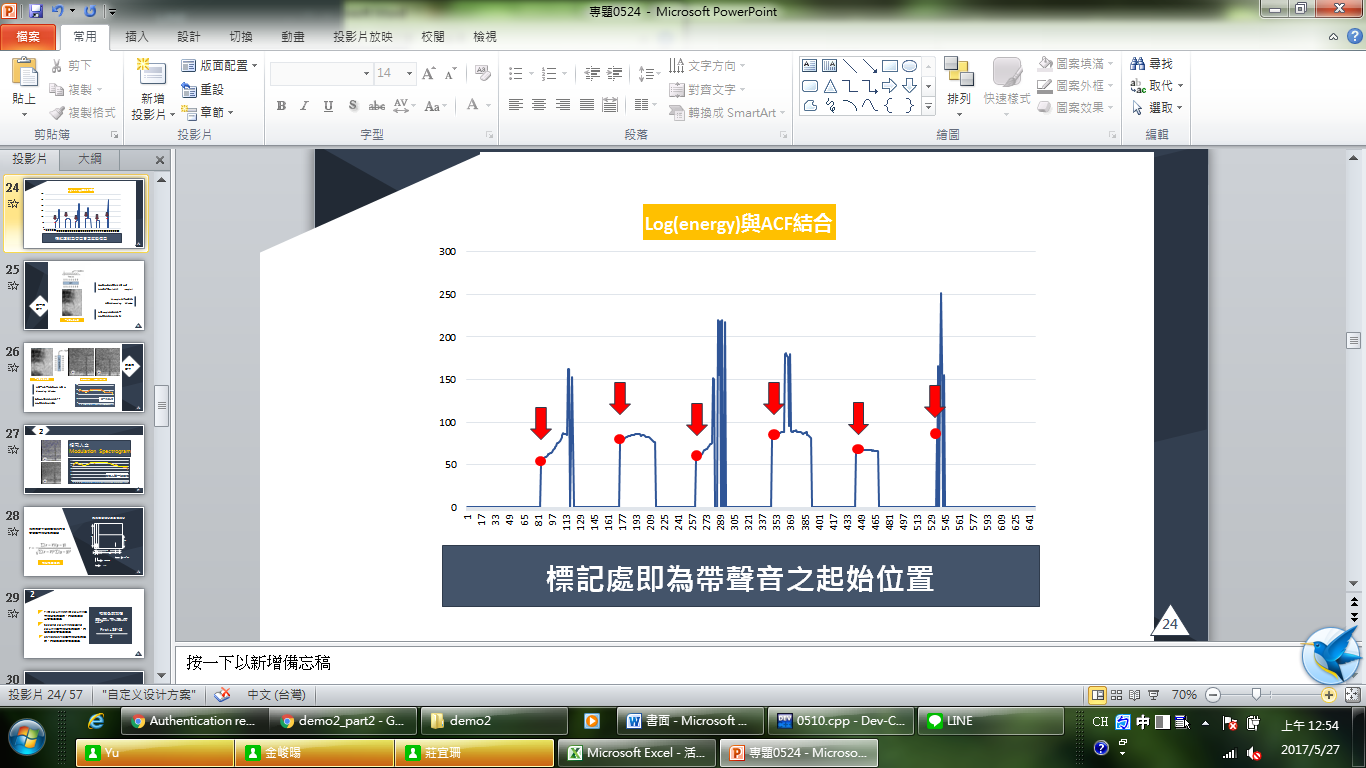
我們利用以下條件從錄音檔中找出帶聲音之起始位置。將音訊之能量對數 (log(energy))小於-5的值設為我們觀察到音訊的最低值-30。



圖表 8 利用能量尋找帶聲音

將過濾後的能量對數與自相關係數第一高峰值陣列做結合，若能量對數為-30，我們將其對應到的自相關係數第一高峰陣列的位置設為0。

最後如果自相關係數第一高峰陣列之值連續10個以上的音框不為零，則此段連續不為0之起始處即為帶聲音的起始位置。



圖表 9 能量對數與ACF第一高峰做結合

1. 第一次離散傅立葉轉換(Discrete Fourier Transform, DFT)

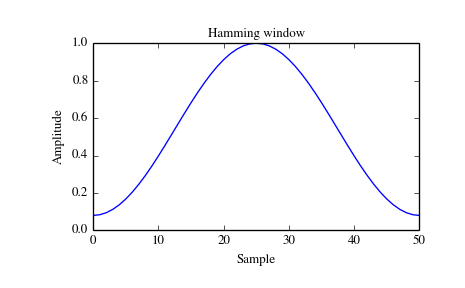
我們將在第一步驟中找到的帶聲音起始位置取100個大小為7毫秒的音框，每個音框有112個樣本，且音框與音框之間的時間差為1毫秒。

由於語音中有低音較強、高音較弱的情形，為了便於分析較高頻域中的變化，我們在做DFT之前先對每個音框進行高頻增強(high frequency emphasis)的動作，其做法為將音框內的每一個樣本減去其前一個樣本的若干比例。

方程式 2 高頻增強

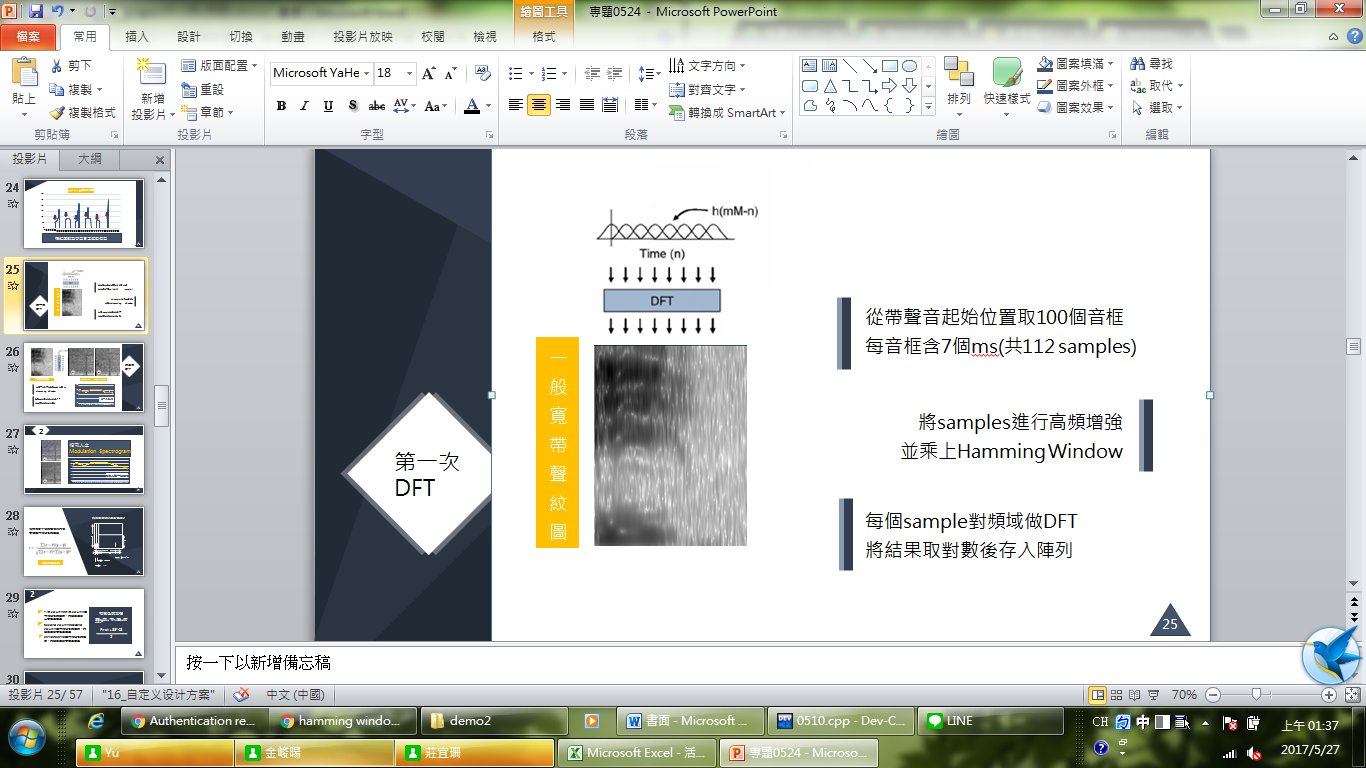
漢明窗(Hamming Window)擁有將主瓣(main lobe)加寬的效果，並有助於產生較佳的寬帶頻譜，因此我們選擇將每個音框套用漢明窗來幫助我們做頻譜的分析。

方程式 3 漢明窗



圖表 10 漢明窗示意圖

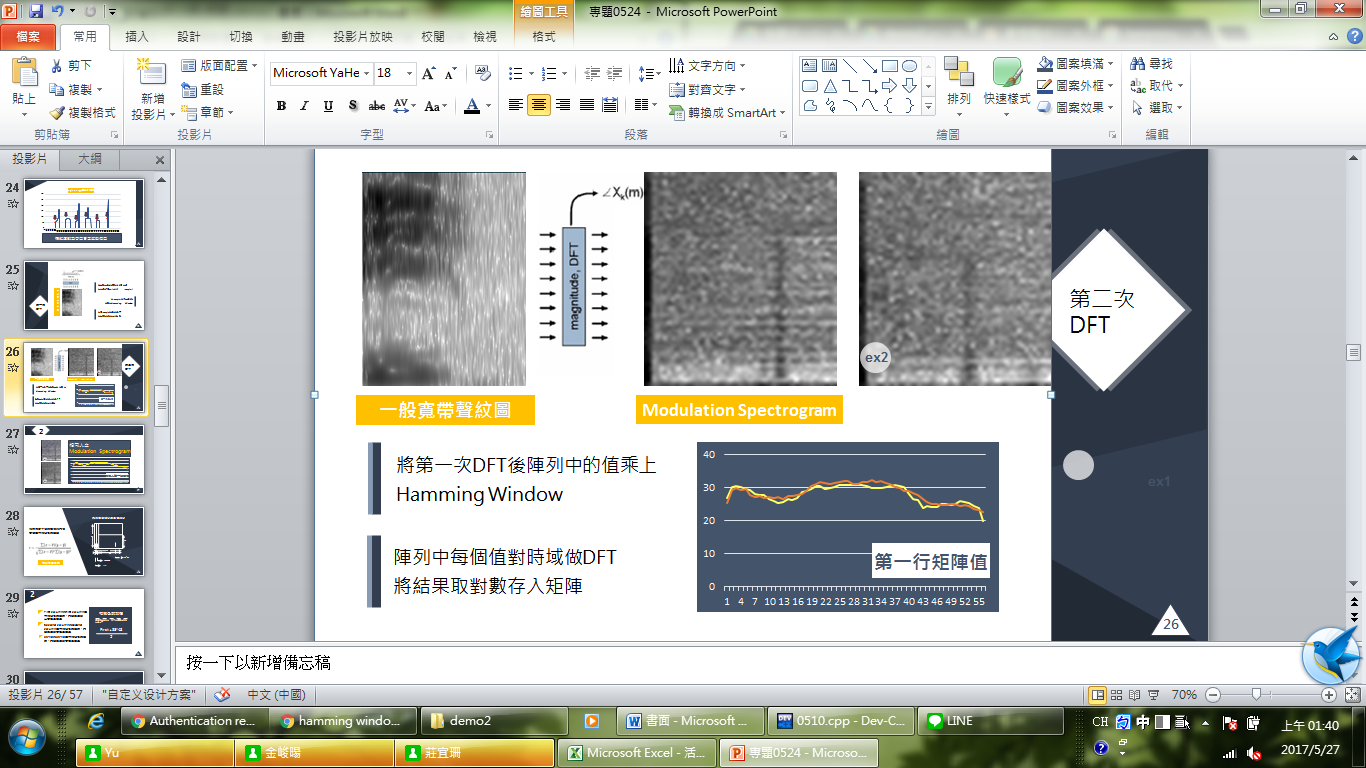
最後對每個音框做離散傅立葉轉換，為了減少數值差異的大小，我們將其結果取對數，再將數值存入陣列中，得到56\*100的二維矩陣。而將矩陣以灰階圖的形式表現，可以得到一般的寬帶聲紋圖。



圖表 11聲紋圖

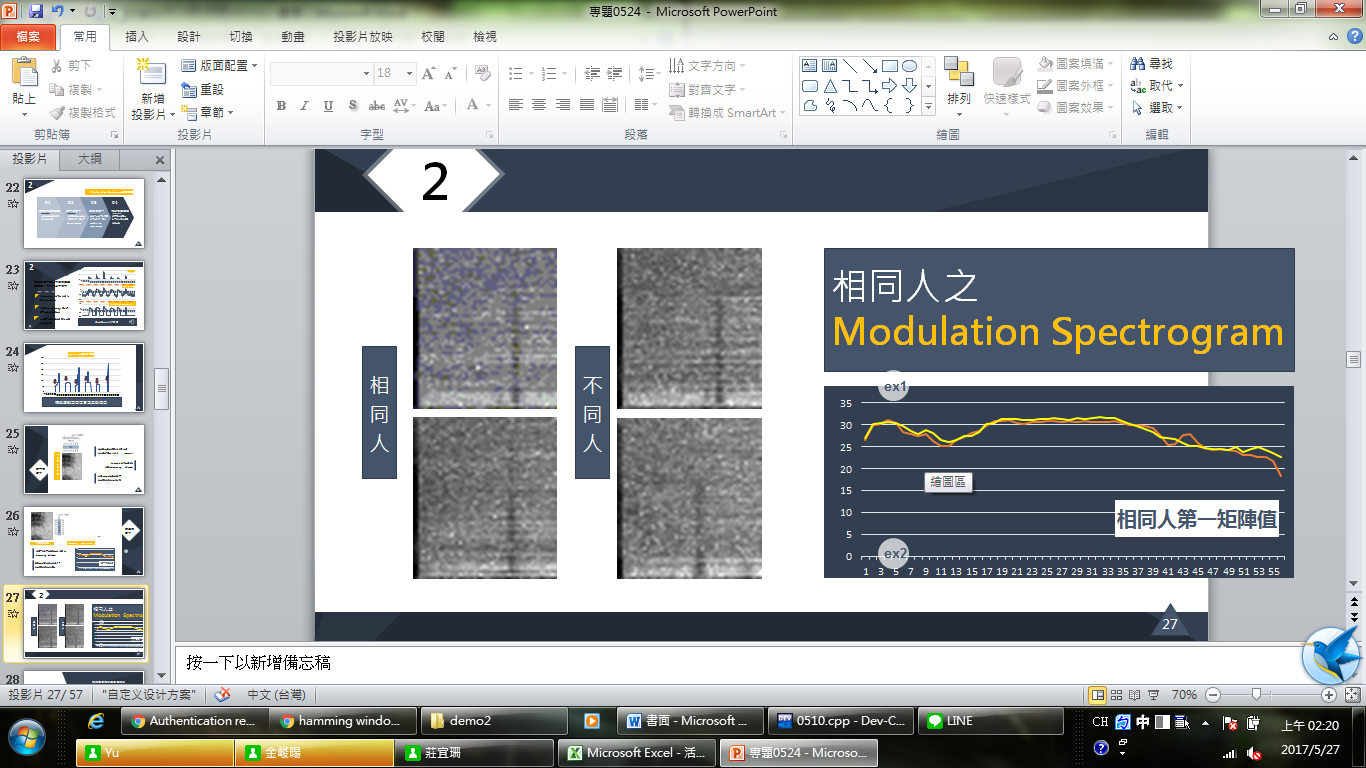
1. 第二次離散傅立葉轉換

我們利用步驟二所產生之聲紋圖，以頻率為單位，乘上漢明窗後，再對時域做一次離散傅立葉轉換。與第二步驟相同，為減少數值間差距，我們將轉換出來的數值取對數，並存入陣列，得到56\*50的二維矩陣。而一般聲紋圖經過離散傅立葉轉換後得到調變聲紋圖。圖表12右邊的灰階圖(BMP)中，黑線位置部分為該聲音之基本頻率大小。



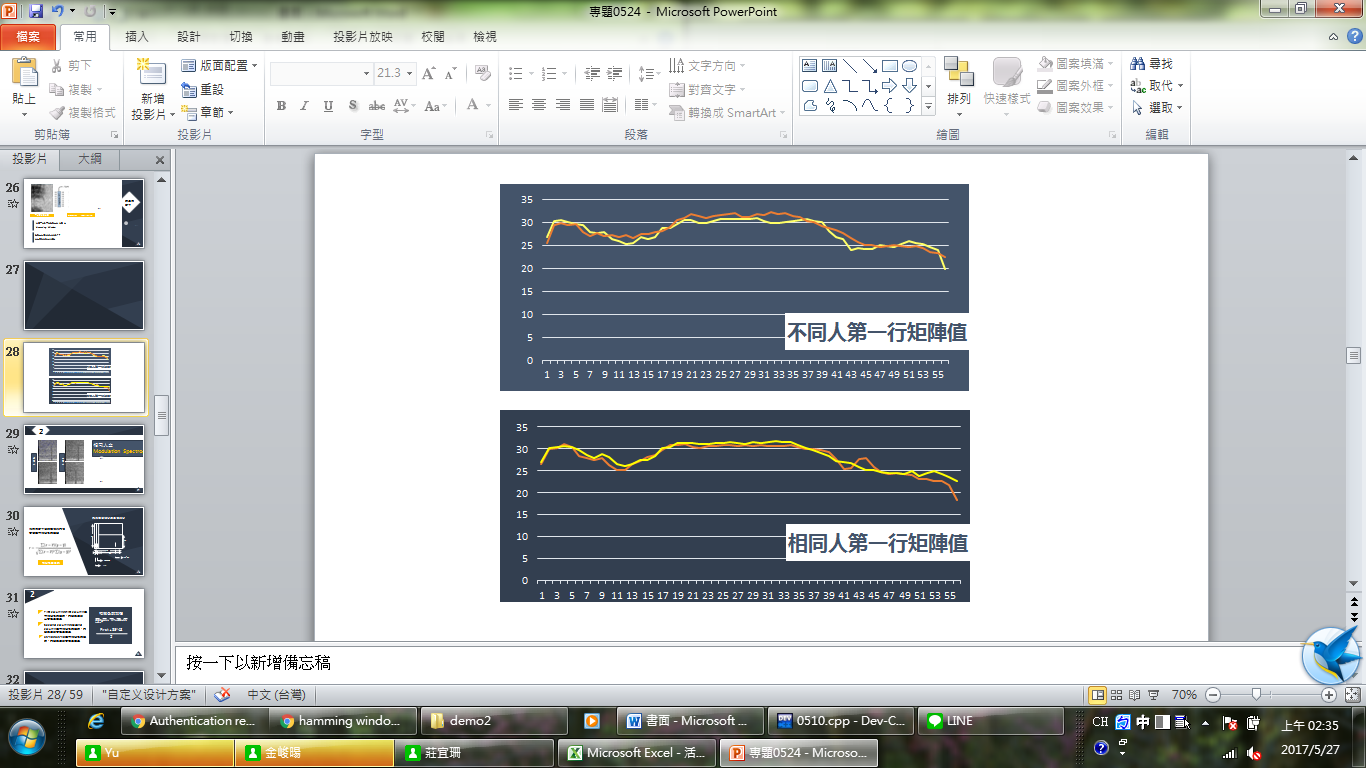
圖表 12 調變聲紋圖

由此陣列產生之灰階圖，我們可以明顯看出，相同人聲音基本頻率是一致的，而不同人的則不重合，如圖表13所示。



圖表 13 相同人與不同人之調變聲紋圖比較

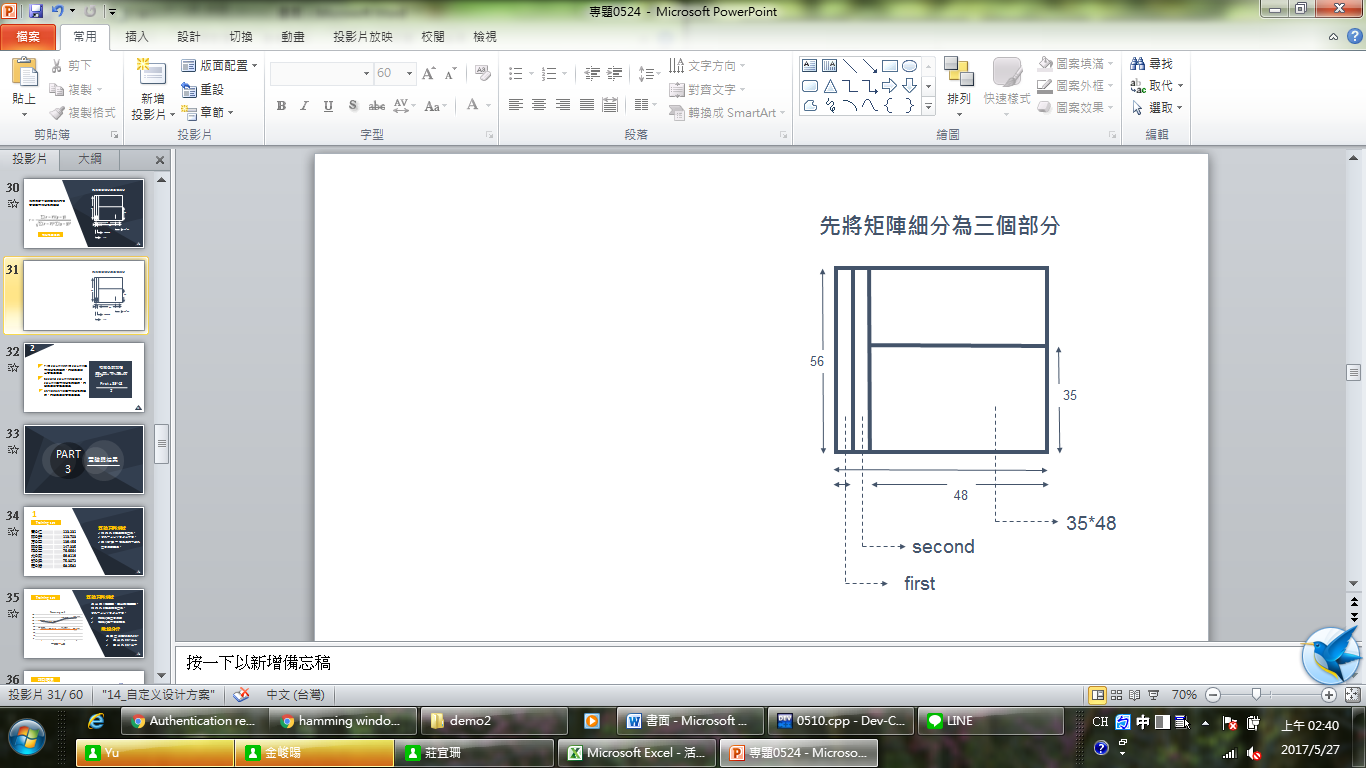
而在圖表14中，可以看出相同人與不同人之第一行矩陣值重疊的狀況也有所不同。不同人線條位置明顯有差異，相同人線條位置則較為相似。



圖表 14 相同與不同人之第一行矩陣值比較

1. 相關係數比較

方程式 4 相關係數



圖表 15 調變聲紋圖之切割

在比較相關係數前，我們先將第三步驟產生之矩陣細分為三個部分。第一部分為矩陣的第一行(first)，第二部分為矩陣的第二行(second)，第三部分為二維矩陣，從第一列至第三十五列、第三行開始到第五十行 (35\*48)。由於剩下部分的資料中，數值皆沒有太大的差異，數據特徵非常不明顯，因此我們將其捨去不取。

利用上述的切割方式，每一個帶聲音都有first、second以及35\*48三個相關係數。

其中first與second代表此音訊的音色，35\*48代表此音訊的音高(基頻)。由於first與second皆代表著一樣的特徵，我們透過實驗比較出first之表現優於second，因此為了實驗的精確度，我們在做相關係數的計算時捨棄了second這項元素。又因為每個人聲音特質反映於音高及音色中，所以我們將first與35\*48的相關係數做了結合，得到下列方程式。

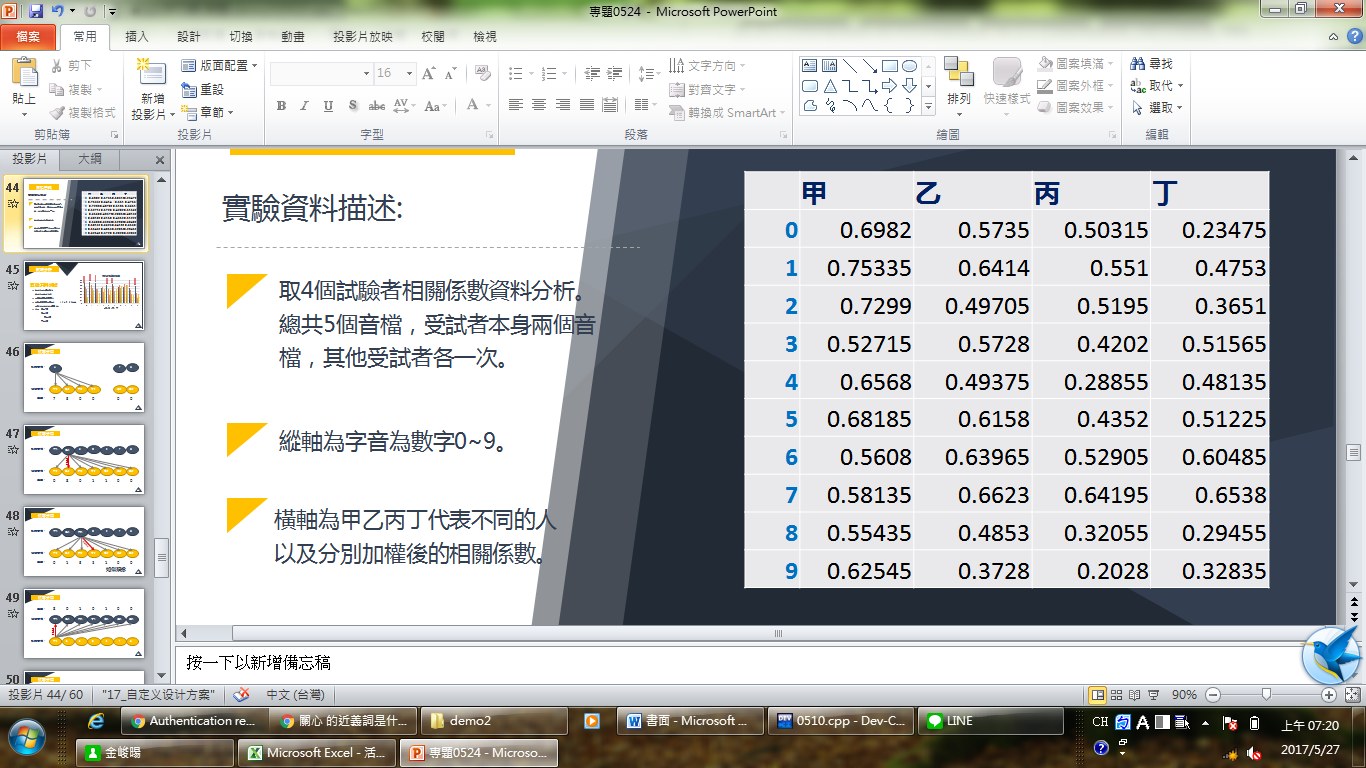
方程式 5 相關係數之加權

最後數據分析的部份我們利用鄰近分類以及投票的方式來做判別。

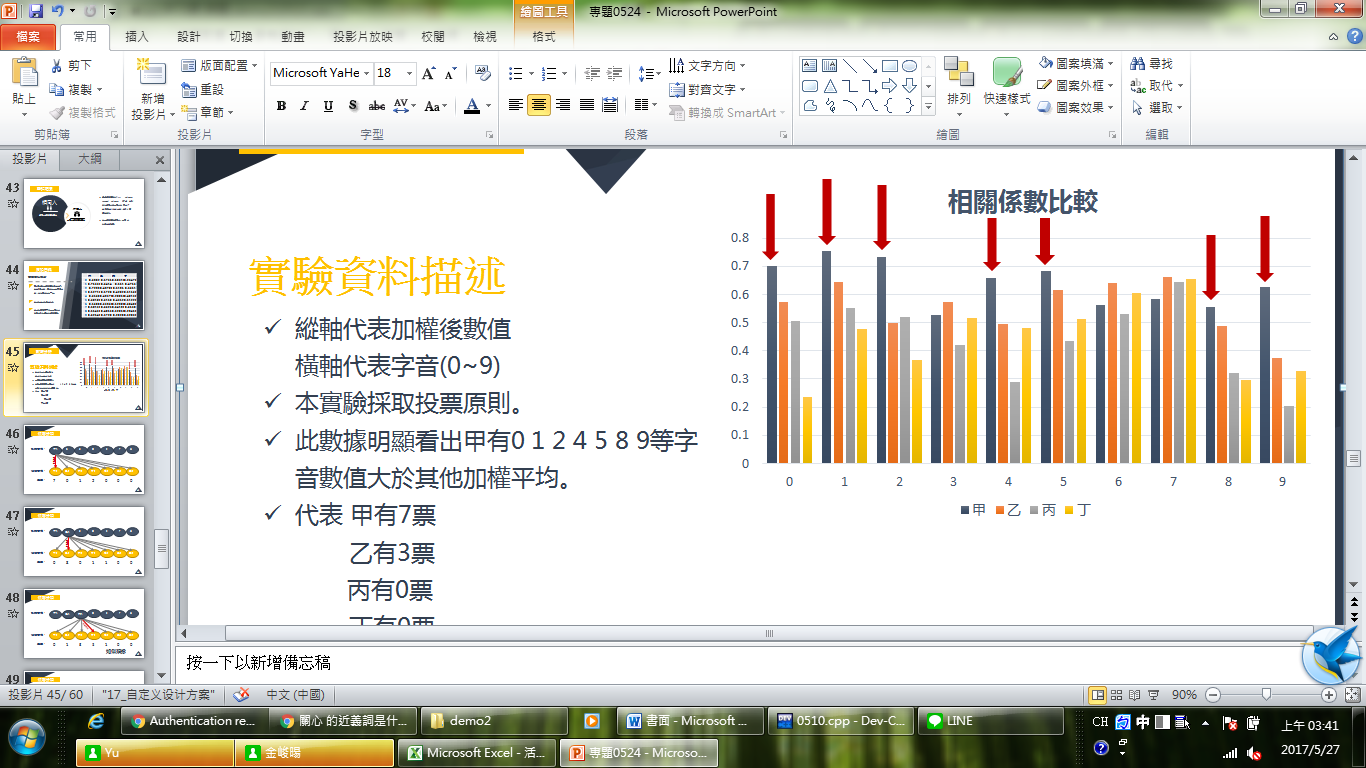


圖表 16 近鄰分析示意圖

假設現在有一筆未知說話人的測試音檔A，利用圖表16所描述之近鄰分析方式，將音檔A與四位受試者(甲乙丙丁)進行相關係數的計算，相關係數為表格1。



表格 1 相關係數

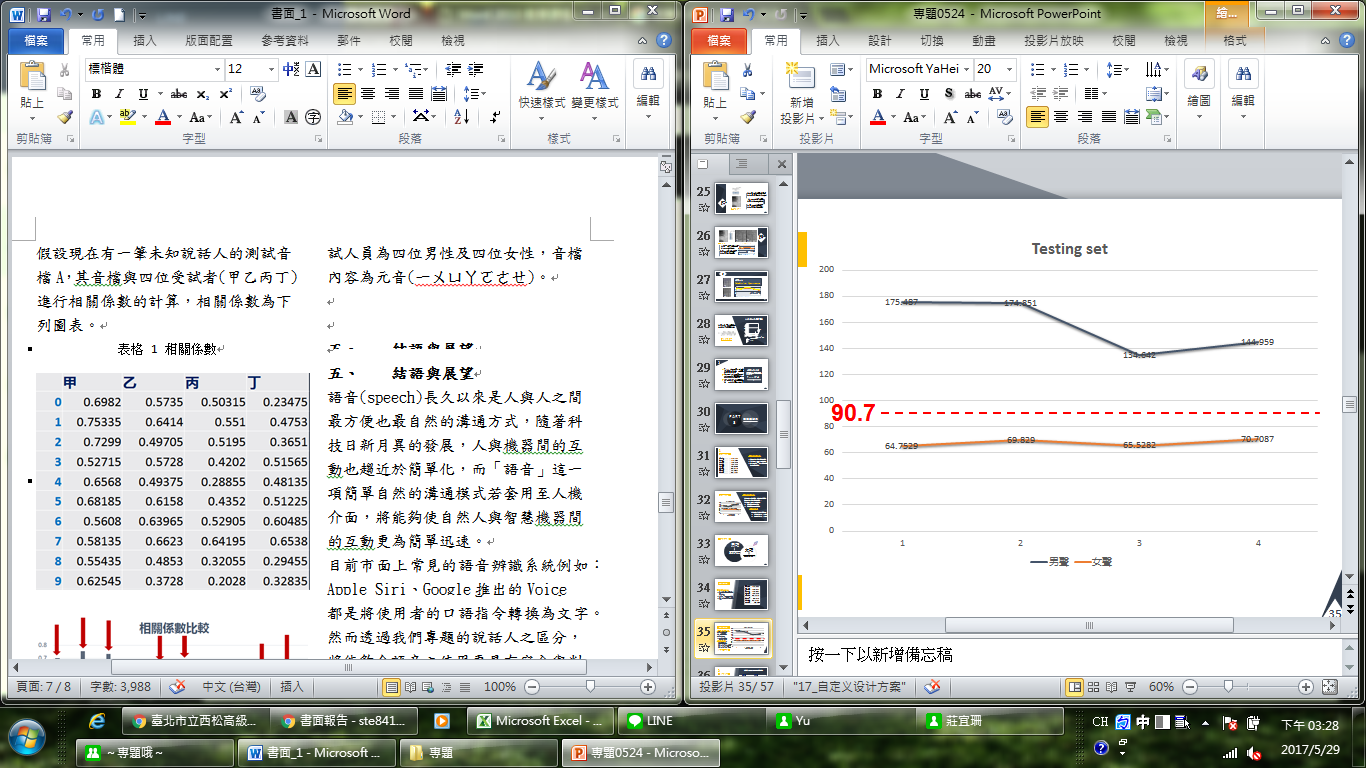


圖表 17 相關係數比較

圖表17中，橫軸為帶聲音的內容(數字0~9)，縱軸為相關係數。從圖表中可以看出此音檔與甲受試者有七個相關係數最高，所以甲有七票；與乙受試者有三個相關係數最高，所以乙有三票；丙與丁各為零票。因此我們判斷此音檔的說話人為甲。

1. **實驗與結果**

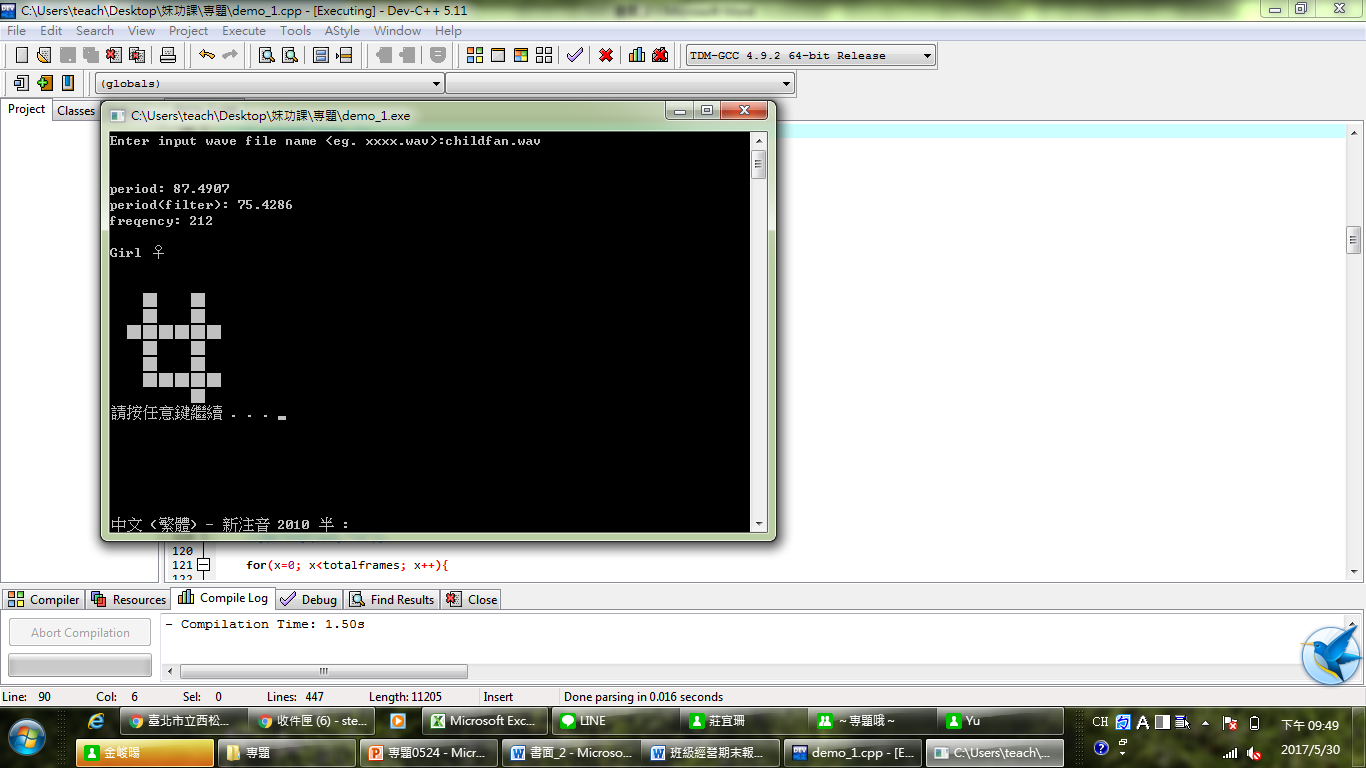
利用專題第一部分，藉由音高做男女性聲音之區分，做了實驗測試。我們隨機找了八位測試人員，其中測試人員為四位男性及四位女性，音檔內容為元音(ㄧㄨㄩㄚㄛㄜㄝ)。經過我們專題測試後得到，男性的聲音的基本週期皆大於90.7；女性的聲音的基本週期皆小於90.7，由此得出，我們用音高判別男女性的正確率為100%。



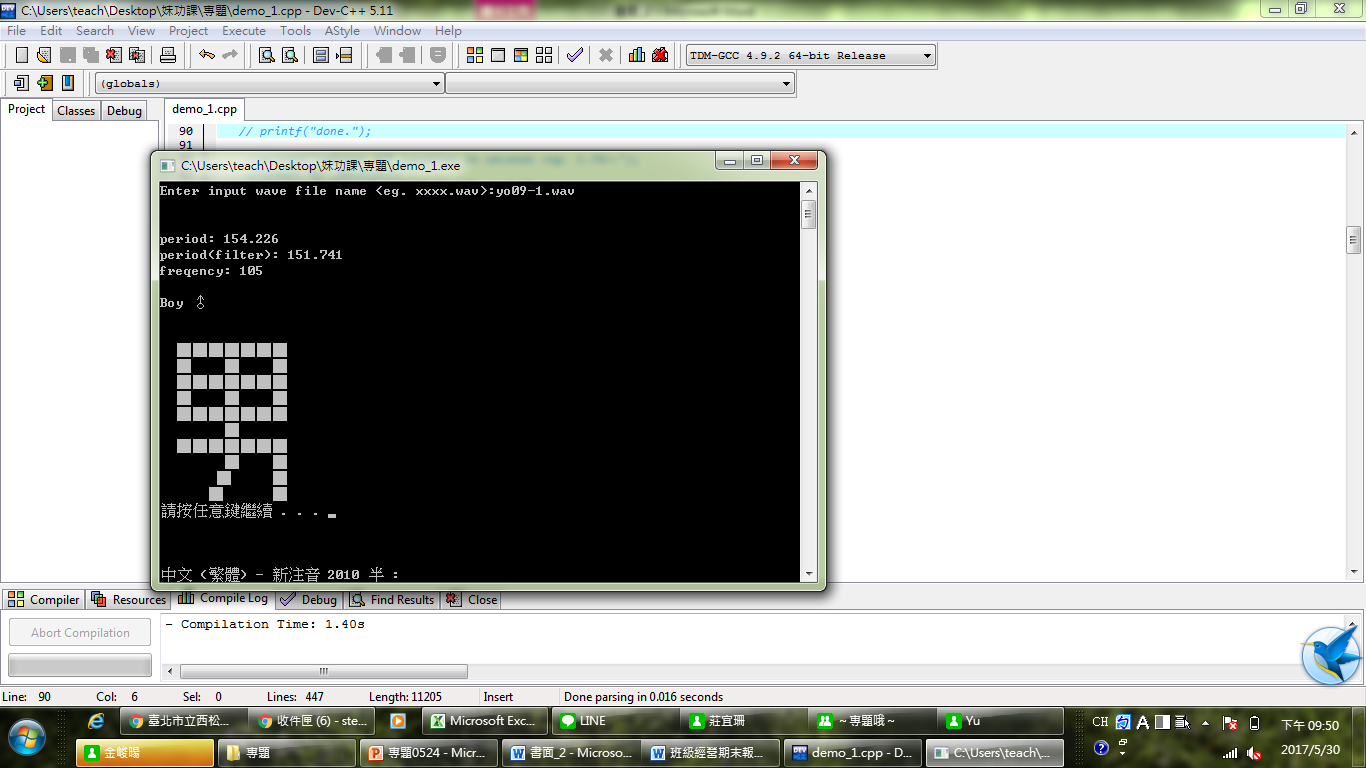
圖表 18 音高判別實驗

另外，除了元音，輸入音檔內容不限定，可以為任意說話內容，執行判別男性與女性之結果的正確率依然是100%。

執行畫面為下圖表19以及圖表20所示。



圖表 19 音高判別執行畫面(女性之聲音)



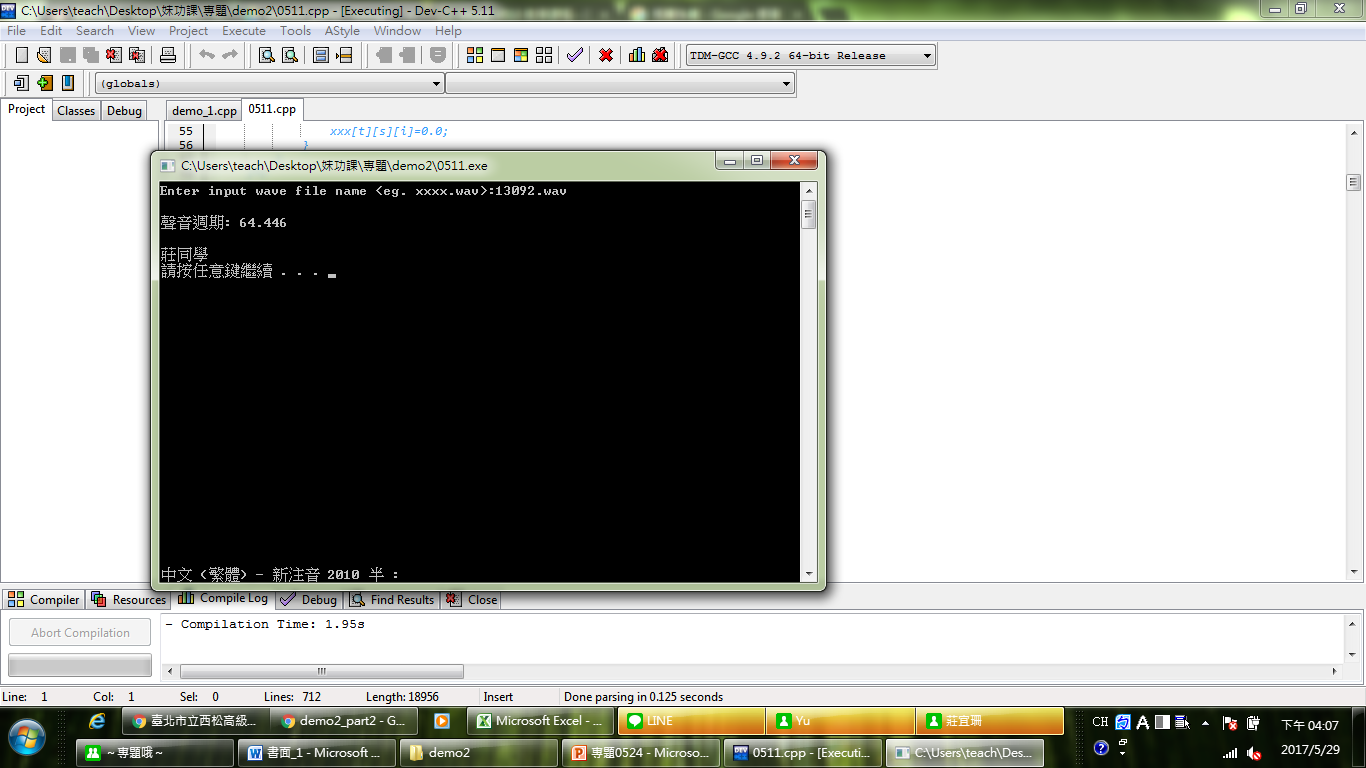
圖表 20 音高判別執行畫面(男性之聲音)

專題第二部分，利用調變聲譜圖同時做音高與音色之區分。我們隨機找了七位受試者，每個人錄製兩份音檔，音檔內容為數字0~9，總共十四筆錄音檔，再利用近鄰分析的方式計算出相關係數。如下圖表21，我們發現，每位受試者的測試音檔皆能正確的比對到相同人的樣板音檔。



圖表 21 相同人辨識

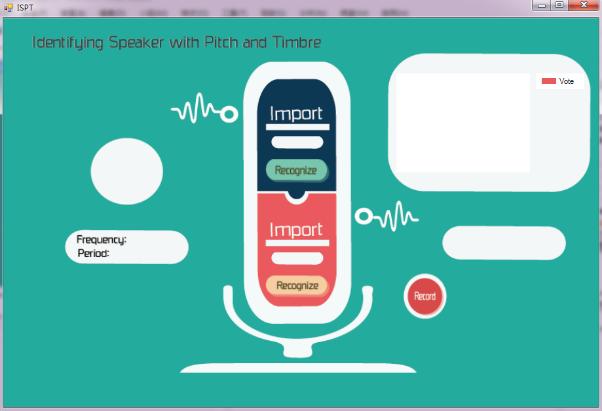
因此得出正確率為100%。執行畫面為以下圖表。



圖表 22相同人分析執行畫面

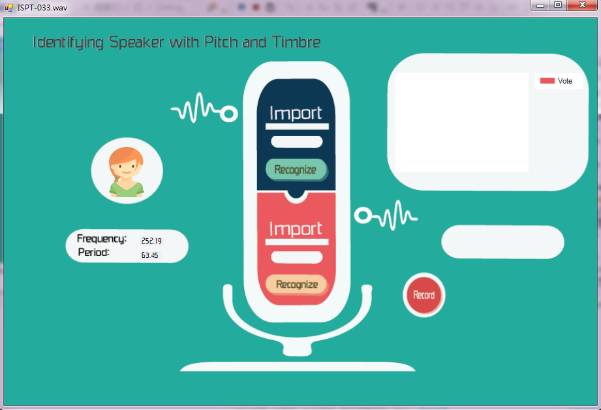
最後我們將專題的兩個部分做了結合的動作，並且為了執行畫面的排版以及美感程度，我們將原本使用的語言C++，改寫成C#。最終執行畫面如圖表23至圖表25所示。

圖表23為開始執行的畫面，麥克風上下兩部份分別為專題的第一以及第二部分。



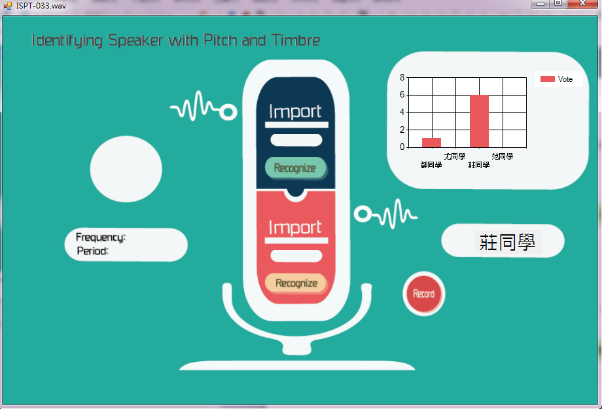
圖表 23 合併後執行畫面

先按下麥克風右方的紅色「Record」開始進行錄音，再按下麥克風上半部(深藍色底)「IMPORT」的白色框框進行資料匯入，最後按下綠色的「RECONGIZE」執行判別男女性之音高辨識。



圖表 24 辨識男女性聲音之執行結果

圖表24為專題第一部分，也就是利用音高來判別男女性的執行結果，很明顯看見測試者為女生，頻率周期都顯示在女性圖像的下方。



圖表 25辨識說話人之執行結果

圖表25為專題第二部分，判別相同及不同人的執行結果，操作方法也如上述，錄音完後按下麥克風下半部(粉紅色底)「IMPORT」的白色框框進行資料匯入，最後按下淡橘色的「RECONGIZE」進行辨識，右上方可以明顯看出投票比數，並在圖表下方可以看出票數最高、最相似的說話人為誰。

1. **結語與展望**

語音(speech)長久以來是人與人之間最方便也最自然的溝通方式，隨著科技日新月異的發展，人與機器間的互動也趨近於簡單化，而「語音」這一項簡單自然的溝通模式若套用至人機介面，將能夠使自然人與智慧機器間的互動更為簡單迅速。

目前市面上常見的語音辨識系統例如：Apple Siri、Google推出的Voice search、以及Samsung S Voice等，都是將使用者的口語指令轉換為文字。然而透過我們專題的說話人之區分，將能夠令語音之使用更具有安全與判別性，未來可應用於語音解鎖、特定人聲擷取等。

1. **銘謝**

首先我們要感謝我們的專題指導教授，在這一年裏老師一直都很有耐心的指導並幫助我們。不論是幫我們將實驗室電腦從XP升級至win10或是提供相關學術論文，老師都會盡他最大能力來幫我們。即使不使用通訊軟體聯繫，老師會一天開三次信箱，即時回應我們的任何問題，並隨時鼓勵以及提供建議。在這一年中，指導老師除了引導我們專題進行方向外，老師也像是我們的親人一般，每個星期開會結束後，都會關心我們學生的健康及課業狀況。專題能與這樣的老師一起研究實在是太幸運了！

另外也要感謝在專題進行中協助過我們的同學們，也許只有一句話，你們也給了我們無限的靈感。最後還要感謝支持我們專題的所有人，謝謝你們讓我們的專題製作可以順利完成。

1. **參考文獻**

[1] WAVE PCM soundfile format. (n.d.). Retrieved October 13, 2016, from http://soundfile.sapp.org/doc/WaveFormat/

[2] Steven M. Schimmel, Les E. Atlas and Kaibao Nie(2007, April). Feasibility of single channel speaker separation based on modulation frequency analysis. Paper presented at IEEE International Conference, Honolulu, HI, USA.