The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 165-174


# 以二維共振峰分布建立語者音色模型及其在語者驗證上之應用

# Using 2D Formant Distribution to Build Speaker Models and Its Application in Speaker Verification

呂嘉穀[1]，蕭志濱[2]，李明慶[2]，蒲長恩[3]，吳家隆[2,*]
[1]國立台北大學資訊工程學系
[2]法務部調查局鑑識科學處，[3]法務部調查局通訊監察處
* 通訊作者

**摘要**

語音是重要的生物特徵之一，也是鑑識科學上的重要工具。在鑑識實務上常遭遇到的一個挑戰，就是通訊線路及錄音裝置的多元性。不同的裝置與線路特性會對語音證物的頻譜產生相當的影響，從而也會影響到鑑識的正確性。共振峰是語音中重要的要素，並且較不易受到通道及裝置之頻率響應的影響。在本論文中我們提出一個從分析較長時間語料，所得之二維共振峰的分布，來建立起一個語者之音色模型的方法。這個方法對於相同語詞及相異語詞方式的語者驗證工作均適用。在實驗的部分，我們報告了對約七十人規模的語料分別進行數位錄音及電話錄音的語者驗證測試。

**關鍵詞:** 語者驗證，線性預測方法(LPC)，共振峰，語者音色模型

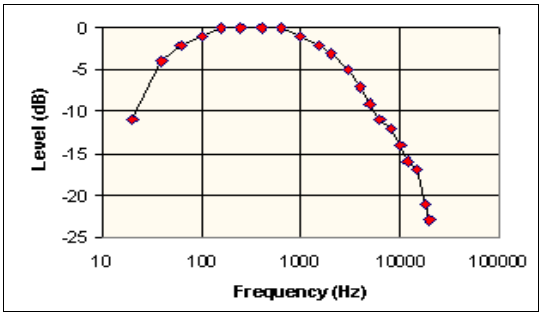
## 1、目的與背景

語音是人類彼此間溝通最方便也最首要的方式。語音不但是用於傳播信息，也是一項重要的生物特徵(biometrics)，可以用來做身份識別之用。對於利用電腦來分析語音這方面的研究，大致可分為兩個領域：一是語詞識別（speech recognition），一是語者識別（speaker recognition）[1-4]。若是要分辨某一個語音樣本是否來自某一個特定的語者，則又稱為語者驗證（speaker verification 或 speaker authentication）。語者驗證又可細分為限定語詞（text dependent）與非限定語詞（text independent）兩種方式[5,6]。在限定語詞的方式中，用來比對的兩段語音樣本，其語音之內容須為相同或相似。而在非限定語詞的方式下，其語句之內容可為不同。後者之處理難度較高，但在取樣上較不受限，其應用也較為廣泛。本研究之內容是屬於語者驗證性質，同時包括了限定語詞與非限定語詞的方式。

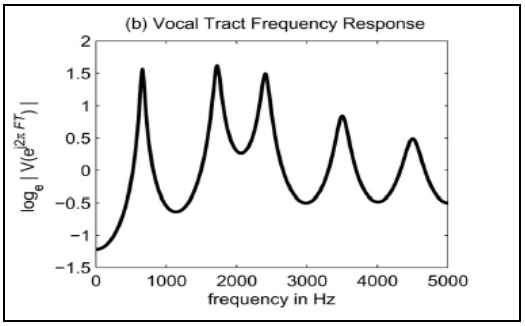
語音分析最基本的技術就是頻譜分析。由於每個人的口腔構造及發音習慣均有所不同，所以發出聲音的共鳴結構就會有所不同，語音中較強的共鳴成分就會形成頻譜中的波峰，稱為共振峰(formants)。不同的語者，因生理構造或口音上的差異，即使是發出同一個字音，其頻譜的形狀也會有所差異。所以藉著分析頻譜，我們可以分辨語者，也可以辨識字音中之韻母。

在語音分析的工作上一個經常遇到的問題就是裝置或是通訊線路(channel)所帶來的影響。如果裝置或線路的頻率響應為已知，我們尚可藉著演算還原出原音訊之頻譜，但是對於鑑識工作方面的實案而言，裝置或線路的特性通常為未知。一般而言，裝置及線路的頻率響應呈現平滑的變化，通常是中頻部分高起而高頻與低頻部分低下的情形。如下圖一所示。相對而言，發聲道的頻率響應，特別是共振良好時，其

變化較為複雜，從低頻到高頻會出現若干個高峰點，如圖二所示。




圖一、裝置或線路之頻率響應示意圖




圖二、發聲道頻率響應例圖

當我們把較緩慢變化的裝置或通道的頻率響應，與較快速變化的發聲道頻率響應相乘(或相加)時，其結果就是前者會影響到後者的整體起伏變化，但是較不會影響到個別高峰點的出現及位置。也就是說，雖然語音之頻譜會受到裝置的影響，但是其中共振峰的位置相對穩定。本研究所提出之語者音色模型將以共振峰的位置為主。我們將從語音中擷取出共振峰，然後以共振峰的分布來建立一語者的音色模型。

近年來有越來越多的研究指出觀察長時間共振峰分析(LTF, Long term formant analysis)在語者驗證上的重要性。所謂長時間共振峰分析就是累計整段語料中各共振峰出現的位置，通常是前四個共振峰。因為是包含眾多不同的字音在內，所以各共振峰的位置並非固定，一段時間累積下來，就會得到一個分布曲線。因為是來自於相同語者，所以語者的因素自然包含在其中。又因為是來自於許多不同的字音的混合，所以字音的因素就會被淡化掉。因為語者驗證的重點是在語者的音色特徵而非語詞內容，所以長時間共振峰分析會是一個可利用的工具。

英國的 Nolan 與 Grigoras 在 2005 年的一篇論文中報告，紀錄語音中前四共振峰的長時間分布，在語者鑑識實案上十分有效[7]。在後續的研究中他們進一步報告各共振峰的長時間分布多呈現出不對稱(skewed)的情形，並且其分布最高點(mode)之位置在鑑識上的重要性超過其平均位置[8]。歐洲學者 Becker、Jessen、及 Grigoras 在 2008 年提出將長時間共振峰分析所得之參數值套用到高斯混合模型(Gaussian mixture model)來進行語者識別[9]。他們假定各共振峰的長時間分布為高斯分布，並自各段語料估計出高斯分布的平均值與標準差值，以進行 likelihood 計算。他們對 68 位男性語者的語料，以前三個共振峰的位置及頻寬為參數(共六個)，達到了 EER 為 0.03 的驗證成績。

德國學者 Moos 對 71 位男性語者的行動電話錄音語料進行 LTF 分析[10]，他發現 F2 與 F3 合用時有優良的語者鑑別效果。他同時發現，F3 較 F2 有著更好的穩定性，也就是對同一個語者其變異性較低。在

文章中也指出，LTF 也具有一些其他良好的特性，例如不易受到說話速度快慢及音調高低等因素的影響。中國學者 Xu 與 Kong 在 2012 年的一篇文章中報告他們以 LTF 分析進行跨語言的語者驗證[11]。他們以前四個共振峰的分布之 peak, kurtosis, 與 skewness 作為特徵值，發現能夠成功的以三種不同語言(中、英、韓)的語料進行跨語言的語者驗證。歐洲學者 Jessen 與 Becker 在 2010 也曾報告他們對德語、俄語、及阿爾巴尼亞語所進行的實驗，也有著相似的結論[12]。

前述學者所提出的長時間共振峰分析，多係對個別共振峰的分布一一的來進行，也就是屬於一個維度的分析。本論文提出的方法是將前幾個共振峰做成對的分析，也就是求得二維的共振峰分布來進行分析。又因為前二共振峰的分布與幾個主要單音韻母有很明確的對應，我們進一步將 F1-F2 平面分割為若干區域，並分別分析落在這些區域中的音框以建立更細緻的語者音色模型。在下一節中我們將詳細介紹本研究所提出建立音色模型的方法。在第三部分中我們會將這個音色模型應用到語者驗證的實驗上。

## 2、研究方法

本論文所提出的方法大致可分為以下幾個步驟。首先我們找出一段語音中具有共振的部分，也就是其中的有聲字音(voiced sounds)部分。其次我們以線性預測方法，逐一分析這些有聲字音的音框，找出其中的共振峰。再根據所找出的共振峰的分布建立起該位語者的音色模型。最後，我們藉比對兩組共振峰分布的相似度，來比對兩段語料之音色相似度。這些步驟分別敘述如下。

### 2.1、找出語料中之有聲字音部分

因為本方法是要找出語料中的共振峰分布，以建立起一語者之音色模型，所以首先我們就要找出語料中具有明顯共振的部分，即是語料中的有聲字音。在本研究中，我們先將語料切割為 20ms 大小的音框，相鄰的音框有 10ms (即為 50%)的重疊。我們對每一個音框計算出一個音量大小值，以及求取其 autocorrelation function (ACF)曲線，並找出其在合理之週期範圍所能達到的最高值。如果一個音框具有足夠的音量大小以及夠大的 ACF 峰值，我們就接受此一音框為一個有聲字音的音框。在語料量足夠的情形下，上述兩項門檻值可以做較嚴格之設定，以確保所找出的音框均有不錯的共振品質。

### 2.2、以線性預測法(LPC)找出音框之共振峰

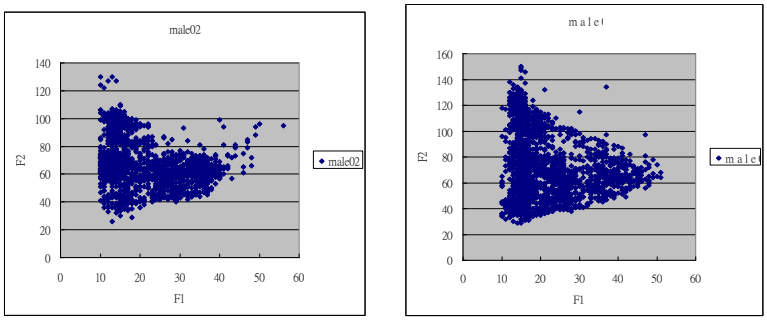
本研究聲音樣本的取樣率(sampling rate)為 11,025 Hz，依據文獻的建議，階數 p 的設定大約是在 14 至 16 左右。在這個取樣率之下，音訊中的頻寬大約保留到 5K Hz，其中共振峰的數目因字音而異，大約有五到六個。本研究在建立語者音色模型時，將會利用到前四個共振峰，即 F1 到 F4。在推導這 p 個 LPC 係數方面我們所用的方法是常見的 Levinson- Durbin 演算法。這個演算法首先自一音框求出 p 個 autocorrelation function 的值，然後再藉由一個遞迴式的演算法解出模型的 p 個係數值來。在對一個音框求出一組係數值之後，我們會再將音框之音訊值帶入模型，並計算出預測值與實際值之誤差。倘若誤差值過大，則表示其所找出的共振峰並不準確，或是該音框的共振仍為不佳，或是該音框受到了較大的雜訊。當有此情形發生時，我們就會略過這些音框不用。一般而言在此一階段會被淘汰的音框大約占有聲字音的 5%以內。

### 2.3、建立語者之音色模型

自一語者之語料找出有聲字音之音框，並推算出其中的共振峰之後，我們就可開始建立該語者之音色模型。如果是使用於相同語詞式的語者驗證工作，無論語料多寡，我們都可以建立起音色模型，只是在語料量少時，所建立起的音色模型也是與語詞內容相關。而當使用於相異語詞式的語者驗證時，我們就需要有較多的語料才能建立起一個較為完整的音色模型。

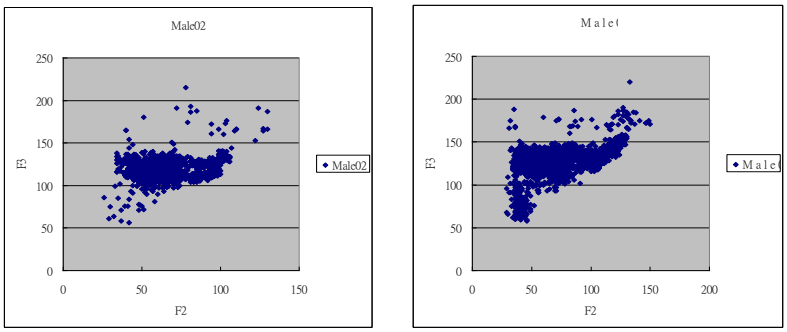
在我們所建立的語者模型中，第一部分就是共振峰分布的情形。在上一個小節中我們提到，我們自每一個音框找出其前四道的共振峰(F1- F4)。在此我們將自語料產生出三個二維(2D)的共振峰分布圖，分別是 F1 對 F2，F2 對 F3，以及 F3 對 F4。每一個有效的音框將會對應到這些圖中的一個點，語料的時間越長，則分布圖中的點也會越多。因為每個語者的音色有所不同，即使是在發出同一個音，其共振峰的位置也會有所差異，反映在這些二維的分布圖上，就是這些點的集中位置會有所不同。

在圖二中我們顯示了兩位不同男性語者在約 60 秒的相同語詞語料所呈現出的 F1-F2 共振峰分布圖。從圖中我們可以很清楚的看出兩位語者的 F1 對 F2 在分布上的差異。因為是依據相同的語詞，所以這裡所反映出的差異主要是來自於二人在音色上的不同。圖四及圖五分別顯示出此二位語者之 F2-F3 與 F3-F4 的分布圖。



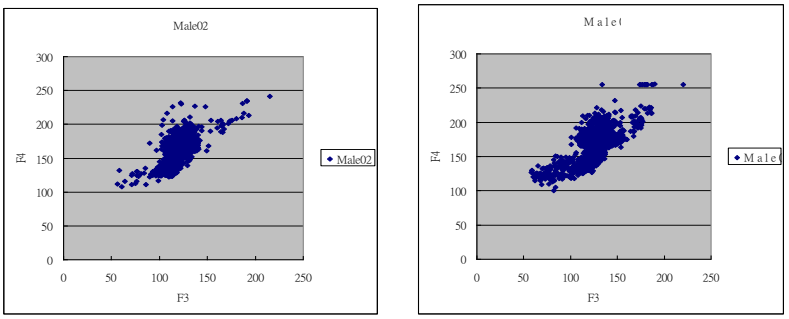

(a) (b)

圖三、兩位男性語者之 F1-F2 分布圖




(a) (b)

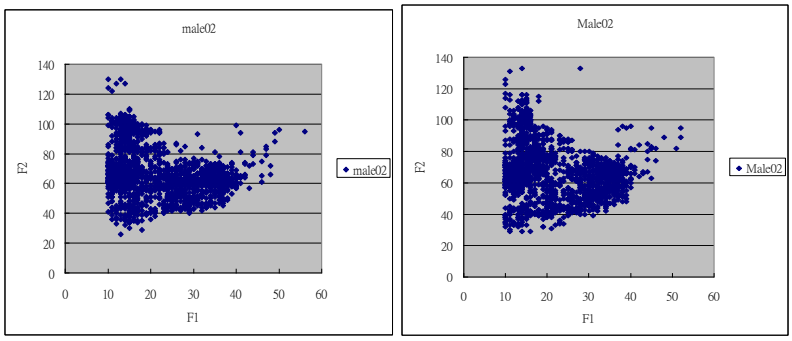
圖四、兩位男性語者之 F2-F3 分布圖




(a) (b)

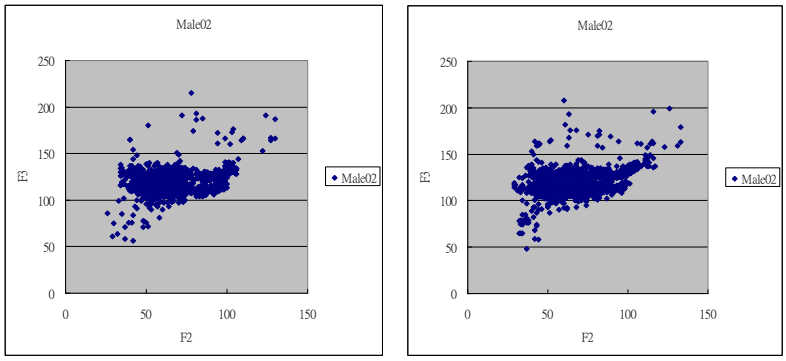
圖五、兩位男性語者之 F3-F4 分布圖

在上二圖中我們同樣可以觀察到，此二位語者在共振峰分布上的差異。接下來我們要顯示的是，相同一位語者對於不同的語詞其所呈現之共振峰的分布情形。在這裡我們將以 Male02 這位男性語者在兩段各約 60 秒，但語詞內容為不同的語料所得之共振峰分布圖來相比較。在下面的三個圖中我們分別比較此位語者在這兩段語料在 F1-F2，F2-F3，與 F3-F4 分布上的差異。



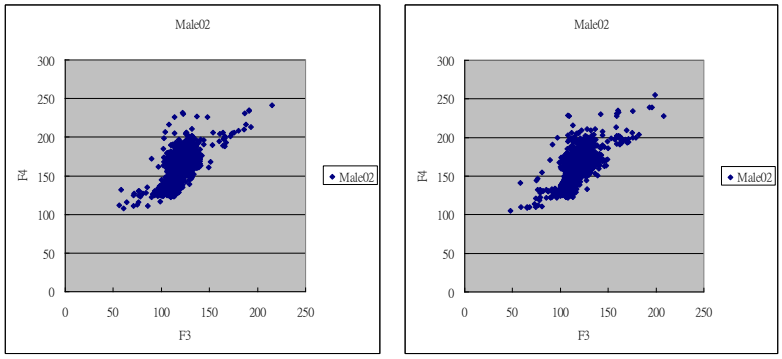

(a) (b)

圖六、同一位男性語者不同語詞語料之 F1-F2 分布圖




(a) (b)

圖七、同一位男性語者不同語詞語料之 F2-F3 分布圖




(a) (b)

圖八、同一位男性語者不同語詞語料之 F3-F4 分布圖

自以上三對共振峰分布圖我們可以清楚看出它們之間的相似性。即使在語詞內容為不同的情形下，因著語料長度夠長，亦即字音重疊程度提高，相同語者之共振峰的累進分布也就趨近於相似。基於以上對不

同語者相同語詞，及相同語者不同語詞之共振峰分布所做的觀察，我們認為以共振峰之分布來建立語者音色模型，將會是一個有效區分語者的工具。

除了以共振峰來建立語者音色模型外，我們也會為語者建立若干個頻譜模型。其原因為頻譜中能量變化上下起伏，但共振峰只標示出其中高峰點的位置，而忽略了頻譜中下凹部分變化的資訊。這些下凹的部分包含了實際發聲道系統中的零點(zeros)，但是 LPC 方法假定發聲道模型中只有極點，所以其所包含的資訊並不完整。

我們為每個語者建立起十個平均頻譜作為其音色之頻譜模型。在此我們先將整個 F1-F2 平面分割為十個區域，其中之九個區域大致對應到一般(F1,F2)會落入的區域，而第十個區域就是對應(F1,F2)不可能出現的區域，對應到此一區域即表示在共振峰的推算上發生了錯誤。對於每一個有聲字音的音框，我們除了求取其前四個共振峰之外，我們也求出其對數功率頻譜(log power spectrum)。我們依其 F1 與 F2 的值將其頻譜累計至前述之十個區域的平均頻譜之中。

本研究的目標是要發展出，一種能夠同時適用於相同語詞式、與相異語詞式語者驗證的方法。共振峰的分布圖仍多少會受到語詞內容的影響，特別是當語料的量較少時。但是這九個區域中的平均頻譜，則較不會受到語詞內容不同的影響。其原因為，當語詞內容為不同時，只會影響到落到各區域中音框的數目，但是對於取自各區中音框之平均頻譜的影響則較輕。也就是說，由這些平均頻譜所組成的模型，能夠更全面性的表現出一位語者的音色特徵。

下表顯示出對男聲我們分割 F1-F2 平面的方式。因為男女聲在共振峰分布上有著相當顯著的差異，所以在區域的分割上也有所不同。在我們的分割方式中，有部分的區域剛好與幾個主要的韻母有所對應，我們也一併標示於表中。

表一、將男性語者之 F1-F2 平面分為 10 個區域

| 區域編號 | F1 範圍 (Hz) | F2 範圍 (Hz) | 所對應之單韻母音 |
|---|---|---|---|
| 1 | 258 – 387 | 580 – 1010 | ㄨ |
| 2 | 258 – 387 | 1010 – 1913 | |
| 3 | 258 – 387 | 1913 – 2687 | ㄧ |
| 4 | 387 – 688 | 580 – 1075 | ㄛ |
| 5 | 387 – 688 | 1075 – 1720 | ㄜ |
| 6 | 387 – 688 | 1720 – 2472 | ㄟ |
| 7 | 688- 1075 | 580 – 1075 | |
| 8 | 688- 1075 | 1075 – 1483 | ㄚ |
| 9 | 688- 1075 | 1483 – 2257 | |
| 0 | 其他頻率 | 其他頻率 | 屬錯誤情形 |

再接下來我們就對落在每一個區域內的音框進行分析，並擷取出一些描述語者聲音特色的參數。計有以下各項：

A、各區域內音框之 FFT 平均頻譜

我們利用 LPC 頻譜來找出 F1 及 F2 共振峰，藉以將音框分群，其原因為 LPC 頻譜的長處就是在於找出主要之共振峰。但是聲音中除了共振峰外尚有其他的特徵及變化，FFT 頻譜就有較全面的紀錄。在此我們先求出各音框的 FFT 頻譜，再將這些頻譜作平均，以得到一個平均 FFT 頻譜。因為音框已經大致依照韻母經過分類，所以落在同一區域中的音框，其發音大致相近，頻譜也會相似。當我們將這些頻譜加以平均時，其個別之差異將會淡化，而共有之特徵將會得到增強。未來在比對兩個語者模型時，我們會將對應區域中之平均頻譜加以比對，以計算其相似度。

B、各區域內音框之 LPC 平均頻譜

LPC 頻譜係由線性預測之係數所推出，其特色在於較能顯出共振峰之位置。在比對語者之音色時，共振峰仍為最主要之資訊，因為共振峰位置是由發聲器官及發音習慣所決定。我們在此將一個區域內所有音框之 LPC 頻譜加以平均，以得到一個平均頻譜。這個頻譜在高峰的部分與 FFT 平均頻譜十分相近，但是在波谷的部分則有較大的差異。我們可藉比對 LPC 頻譜特別來檢視兩位語者在共振峰位置上的相似程度。

C、各區域內音框之共振峰的累加曲線

我們做完 LPC 分析之後，對於每個音框我們得到了一組的共振峰。目前我們是找出前五至六個，就是 F1 到 F6。不過在有效的頻寬範圍中我們可能只會看到四個共振峰，此時第五個共振峰就經常在有效頻寬之外，在比對時可忽略不看。在這裡我們將一個區域內，所有音框的所有共振峰，全部投到同一個頻率軸線上。因為區域本身就是按 F1 與 F2 來劃分的，自然這些音框在 F1 與 F2 的頻率範圍會有相當高的一致性。但是我們發現這些來自同一語者的聲音，在 F3, F4, 甚至於 F5 也會有著相當好的一致性。這一項特徵參數也是用來反映出一位語者在發出不同字音時，其共振峰分布的情形。但是與前一項不同之處是，在前一項中，我們是對 LPC 頻譜之強度值做平均，所以一個音強的音框，會較一個音弱的音框有著更大的影響。但是在這裡我們對每一個音框的共振峰都是以同一強度值紀錄，所以在意義上略有不同。

## 2.4、藉比對音色模型決定二段語料之音色相似度

在前面的部分我們說明了如何自一段語料建立起其語者的音色模型，一個模型中包含了三個二維的共振峰分布圖，以及將 F1-F2 平面分為十個區域後，在每個區域所累積出的 FFT 平均頻譜、LPC 平均頻譜、以及共振峰分布累加曲線。在比，我們將比對兩個音色模型的內容，以估計兩段語料之語者在音色上的相似程度。

我們進行比對兩個音色模型的基本方法為計算其相關係數值。在比對二維共振峰分布時，我們分別就三對的二維共振峰分布(F1-F2, F2-F3, F3-F4)兩兩計算出其間之相關係數值。在比對兩個音色模型中對應之 FFT 頻譜、LPC 頻譜、及一維之共振峰累加曲線時，我們則兩兩計算其間之一維相關係數。在完成以上之計算後，我們將得到六個相關係數值。因為這些相關係數均具有不同的特性，之後我們可以依語料的特性顯選擇使用，或是將這些相關係數做加權平均，以得到一個綜合相似指標值。

## 3、實驗與結果

在以相異語詞進行語者驗證之實驗部分，我們分別針對了男聲及女聲，並以數位錄音及電話錄音兩種方式進行實驗。每一種的錄音方式又可分為比對同次錄音中之不同語句，以及比對不同次錄音中之不同語句兩種方式。在實驗中，我們使用了 72 人的語音樣本，其中有男生 38 人及女生 34 人，均為 18 歲以上之成年人。採樣分兩次進行，時間上的間隔為兩個月。實驗中所用到的國語語句每組的句數均是六十句，每句有六至十個字不等。

每次的錄音因語者說話速度快慢不同，大約有三分鐘的長度。我們再將每份語料分為前後兩段，每段的長度大約在 90 秒左右，其中包含一句與一句之間停頓的時間。倘若扣除掉語句間停頓所花的時間，每段錄音的長度約為 60 秒左右。因為前段與後段有著不同的語詞內容，所以我們可以用同次錄音中的前段與後段進行相異語詞之語者驗證。因為是取自於同一次的錄音，無論是錄音裝置、錄音環境、或是語者的生理狀況都會極為相似，所以我們預期比對的結果(即驗證正確率)將會較好。

我們也將利用不同次錄音中的前後段落進行交叉的比對，例如將第一次錄音中的前半段，與第二次錄音

中的後半段進行比對；或是將第一次錄音的後半段與第二次錄音的前半段進行比對。這種的比對方式不單是語詞不相同，就連錄音的裝置、錄音的環境、通訊的線路、以及語者的生理狀況都有可能不同。我們預期驗證的正確率也將會下降。

在驗證所用的參數部分，如在前面一節中所述，我們有以下幾個，分別給予編號 P1 到 P9：

| | |
|---|---|
| P1 | 九個特徵區域中音框之 FFT 平均頻譜 |
| P2 | 九個特徵區域中音框之 LPC 平均頻譜 |
| P3 | 九個特徵區域中音框之共振峰分布曲線 |
| P4 | 全域之 F1-F2 分布 |
| P5 | 全域之 F2-F3 分布 |
| P6 | 全域之 F3-F4 分布 |
| P7 | P1-P3 之綜合 |
| P8 | P4-P6 之綜合 |
| P9 | P1-P6 之綜合 |

其中 P1 到 P6 是個別的參數曲線或是分布圖，P7 是把前三個特徵曲線(P1-P3)加以綜合的結果。而 P8 是將 P4-P6 這三個分布圖加以綜合的結果，而 P9 則是再進一步把 P7 和 P8 加以綜合。接下來我們就依序表列不同條件下所得之驗證正確率。

A、以同一次數位錄音中之不同段落進行語者驗證

相對於電話錄音，數位錄音有著較大的頻寬，保存語者音色的能力較佳。又因為比對用的語料取自同一次的錄音，在錄音裝置、錄音環境、通訊線路、以及語者生理狀態等多方面均為最相近，所以驗證的準確率最高。

表二、以同次數位錄音中之相異段落進行語者驗證所得之驗證等錯誤率 EER (%)

| 參數 | 男聲 | 女聲 |
|---|---|---|
| P1 | 0.3 | 0.0 |
| P2 | 0.3 | 0.0 |
| P3 | 1.3 | 0.1 |
| P4 | 0.3 | 0.7 |
| P5 | 5.4 | 4.3 |
| P6 | 5.6 | 4.5 |
| P7 | 0.2 | 0.1 |
| P8 | 1.8 | 0.1 |
| P9 | 1.1 | 0.0 |

從上表中可以看到，無論是男聲或是女聲，我們都達到了相當高的正確率。這就反映出這些特徵值確實能夠掌握到一個語者的音色特徵。仔細的比較，我們可以發現 P1-P3 的表現略為優於 P4-P6，但是 P4 仍然是有著相當高的驗證正確率(EER 1%以下)。女聲的正確率略優於男聲，這也是因為女聲所使用的頻域較男聲為寬，其音色可有較大的差異度。

B、以不同次之數位錄音中之不同段落進行語者驗證

如前所述，兩次錄音之間間隔了兩個月的時間，在裝置及人員方面均有所變化，所得到的語者驗證正確率也就有所下降。

表三、 以不同次數位錄音中之不同段落進行語者驗證所得之驗證等錯誤率 EER (%)

| 參數 | 男聲 | 女聲 |
|---|---|---|
| P1 | 15.1 | 14.3 |
| P2 | 6.2 | 9.2 |
| P3 | 7.0 | 11.2 |
| P4 | 19.4 | 21.3 |
| P5 | 24.2 | 16.0 |
| P6 | 18.8 | 21.0 |
| P7 | 8.4 | 10.5 |
| P8 | 12.4 | 11.5 |
| P9 | 6.9 | 10.6 |

從上表中可以看到，所有特徵值的驗證 EER 值都有明顯的上升。但其中部分的參數，尤其是 P2 與 P3，表現相對穩定。因此之故，P7 與 P9 也有著較佳的表現。在這個部份我們看到對男聲的正確率略為優於女聲，一個可能的原因就是男聲相對受到隨時間變化因素的影響較小。

C、以同次之電話錄音中之不同段落進行語者驗證

電話錄音之頻寬為 3.5k Hz 左右，相當程度低於數位錄音之約 5.5k Hz 的頻寬。這個減少的頻寬對於語者的音色會產生一定程度的影響，這樣的影響也些微的反映在驗證的 EER 之上。

表四、 以同次電話錄音中之相異段落進行語者驗證所得之驗證等錯誤率 EER (%)

| 參數 | 男聲 | 女聲 |
|---|---|---|
| P1 | 0.1 | 0.2 |
| P2 | 0.3 | 0.4 |
| P3 | 0.3 | 0.1 |
| P4 | 1.9 | 2.1 |
| P5 | 1.8 | 1.4 |
| P6 | 3.0 | 1.5 |
| P7 | 0.2 | 0.2 |
| P8 | 0.2 | 0.4 |
| P9 | 0.1 | 0.2 |

比較上表與表二，我們看到在男聲的部分差異不大，但是在女聲的部分正確率有略為下降。這個下降的情形主要是受到頻寬被壓縮的緣故，對女聲音色的影響較比對男聲明顯。與前兩表(表二及表三)相似的是，P1-P3 的表現仍是優於 P4-P6，但是當我們將 P4-P6 綜合起來用(即為 P8)，仍然是有著不錯的正確率。

D、以不同次之電話錄音中之不同段落進行語者驗證

在四種組合之中，此一組合之情形最接近鑑識工作的實際情況。此處所利用到的電話線路十分多元，包括固網、手機、長途等等。語者的發話環境也各為不同，年齡的範圍也較廣。所以這個部分的語料的品質較為接近實案中的情形。

表五、 以不同次電話錄音中之不同段落進行語者驗證所得之驗證等錯誤率 EER (%)

| 參數 | 男聲 | 女聲 |
|---|---|---|
| P1 | 11.5 | 8.1 |
| P2 | 12.5 | 9.0 |
| P3 | 12.5 | 8.4 |
| P4 | 7.7 | 11.4 |
| P5 | 16.7 | 8.9 |
| P6 | 24.6 | 15.6 |

| | | |
|---|---|---|
| P7 | 12.2 | 8.7 |
| P8 | 15.5 | 6.3 |
| P9 | 14.4 | 6.2 |

將上表與表四比較，我們可以發現驗證之 EER 有所上升。從 P9 可以看出來，女聲部分大約上升了六個百分點，但是男聲則上升約十個百分點。值得注意的是女聲部分 P8 的表現超越 P7，這表示出當錄音之線路及裝置多元時，共振峰分布之特徵比頻譜特徵有著較佳的表現。此外男聲部分 P4 的表現也相對較佳。P6 在電話錄音的部分表現較差的原因是，共振峰 F4 常常是高於截止頻。所以找到的 F4 經常並非真正的 F4。不同次(相隔兩個月)的電話錄音，又是在非實驗室的環境之下錄製，能夠達到 90%以上的驗證正確率，顯示本方法具有實用之潛力。

## 4、結論

在本論文中我們提出了一種可以以相異語詞語料進行語者驗證的方法。因為用來比對之語料的語詞可能為不同，我們無法逐句地來進行比對。我們所提出的方法，乃是自語料分析長時間共振峰的分布以建立起語者音色模型，然後再就兩個音色模型加以比對。這樣的一個音色模型，我們認為至少具有以下兩方面的優點。第一是頻譜本身的形狀相當容易受到通訊線路或錄音裝置的影響，從而影響到比對的結果，但是共振峰的位置卻是相對較為不會受到裝置或線路的影響。因為在鑑識實務上，線路及錄音裝置相當多樣化，也難以取得其頻率特性資料。使用共振峰特徵將有助於提升鑑識的穩定性。第二方面的優點是來自於我們將 F1-F2 平面分區之後，再就落在各區之中的音框，分別求取其平均頻譜和共振峰分布曲線，以建立語者音色模型。這些區域大致與不同的單韻母音對應，在語詞不同的情形下，各韻母出現的次數或為不同，但所建立起的語者音色模型仍屬完備。

## 5、參考文獻


1. R.D. Peacocke and D.H. Graf, “An introduction to Speech and Speaker Recognition,” *IEEE Computer Magazine*, pp. 26-33, August 1990.
2. J.P. Campbell, “Speaker Recognition: A Tutorial,” *Proceedings of the IEEE*, Vol. 85, pp. 1437-1462, September 1997.
3. Sadaoki Furui, *Digital Speech: Processing, Synthesis, and Recognition*, 2nd. Edition, Marcel Dekker, New York, New York, 2001
4. Thomas F. Quatieri, *Discrete-Time Speech Signal Processing principles and Practice*, Prentice Hall, 2002.
5. T. Dutta, “Text Dependent Speaker Identification based on Spectrograms,” *Proceedings of Image and vision Computing New Zealand 2007*, pp. 238-243, December 2007.
6. F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcıa, D. Petrovska-Delacretaz, and D. A. Reynolds, “A Tutorial on Text-Independent Speaker Verification,” *EURASIP Journal on Applied Signal Processing*, Vol. 4, pp. 430–451, 2004.
7. F. Nolan and C. Grigoras, “A case for formant analysis in forensic speaker identification,” International Journal of Speech Language and the Law, Vol. 12, No. 2, pp. 143-173, 2005.
8. K. McDougall, P. Harrison, F. Nolan, and C. Kirchhubel, “Voice Similarity and Long Trem Formant Analysis,” University of Cambridge report.
9. T. Becker, M. Jessen, and C. Grigoras, “Forensic Speaker Verification Using Formant Features and Gaussian Mixture Model,” Proceedings of Interspeech 2008 Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches, Brisbane, Queensland, Australia, Sepember, 2008.
10. A. Moos, “Long-Term Formant Distribution (LTF) based on German spontaneous and read speech,” Proceedings of IAFPA 2008, Swiss Federal Institute of Technology, Lausanne, 2008.
11. Y. Xi, “Vocal tract characteristics on long-term formant distribution,” Proceedings of the 2012 International Conference on Computer Science and Network Tecnology (2012 ICCSNT), pp. 207-211, Dec. 29-31, 2012.
12. M. Jessen and T. Becker,” Long-term formant distribution as a forensic-phonetic feature,” Journal of Acoustical Society of America, Vol. 128, No. 4, p. 2378, 2010.