國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

使用性別資訊於語者驗證系統
之研究與實作

A study and implementation on Speaker Verification System using Gender Information

蘇俞睿

Yu-Jui Su

指導教授：張智星 博士

Advisor: Jyh-Shing Roger Jang, Ph.D.

中華民國107年6月

June 2018

doi:10.6342/NTU201801578

側邊

國立臺灣大學
資訊工程系(所)

碩士論文 使用性別資訊於語者驗證系統之研究與實作 蘇俞睿 撰

107
6

國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

使用性別資訊於語者驗證系統之研究與實作

A study and implementation on Speaker Verification System using Gender Information

蘇俞睿

Yu-Jui Su

指導教授：張智星 博士

Advisor: Jyh-Shing Roger Jang, Ph.D.

中華民國107年6月

June 2018

doi:10.6342/NTU201801578

誌謝

能夠順利完成本篇論文與相關研究，首先我要感謝我的導師—張智星老師為我們實驗室向中華電信爭取計畫，並於每週咪挺時在投影片排版上對我們提供了大大小小的意見，實在令我們受益良多；接著，我亦要感謝不時為我的研究方向點出新想法的仲理學長，若不是學長不吝適時提供意見，今天我可能仍然守株待兔、困在自己的觀念盲點裡無法有所突破；另外，葉子與承澤學長則讓我鍛鍊了永不放棄的堅強意志，每當我碰到研究上或其他人為因素造成的困難，他們總會不遺餘力地分享自己或其他學長姐突破困境的經歷與體會的道理，這些陪著我度過種種瓶頸的經驗談與人生哲理，不僅為數度陷入情緒谷底的我注入了滿滿的元氣，也無疑地為我往後的人生點亮了數盞明燈。

除了學長姐與老師，我也必須感謝實驗室裡與我同船共濟，並給了我的許多信心的小夥伴。首先是和我一起被分配到中華電信計劃的宗師和正忞，上天安排我和你們兩位在同一個計畫內打拼，就算說我是一次享盡了三輩子的福氣，也一點都不誇張，有你們這些強大隊友的技術支援，讓我感覺自己彷彿成為了騎上赤兔馬的呂布，面對各種程式或研究上的問題也能過關斬將；另外，我也必須感謝教會我輕鬆面對研究壓力的的不肖子、高血壓以及群暉哥，有你們這三位不可多得的嘴砲戰友，除了讓我能在爆肝的研究生生涯中定期排毒，更使我深刻體會了永生難忘的說話藝術；最後，我也要感謝阿丸和奕瑄，和你們一起打大亂鬥真是生命中的一大樂事，有你們的陪伴，原本枯燥的生活也照樣能充滿歡笑。

一路走來，我真的受到了太多的幫助，需要感謝的人太多了，就感謝天罷！

doi:10.6342/NTU201801578

摘要

在語者驗證領域中，在不改變聲學模型架構之前提下，以男性與女性之語料分別訓練的性別相關模型取代性別不相關模型，是常見的提升系統辨識率作法之一。然而，在實際運用情形中，由於測試語者的性別是未知的，因此性別分類器在此流程下便扮演了非常重要的角色，其準確度更會直接影響語者驗證系統的表現；而確保系統面對不同性別之仿冒者皆能正確拒絕，亦是此作法相當重要的一項訴求。

為探討不同的「語者性別資訊運用方法」對於語者驗證系統所產生的影響，本論文實作了以 i-向量與機率性線性判別分析模型為語者特徵與評分器之語者驗證系統，與 3 種以 i-向量為基礎的性別分類器。本論文在分析一般使用性別相關模型之語者驗證系統的弱點後，分別於「性別分類器表現良好」與「性別分類器表現不良」之兩大狀況下提出其他若干種不同的性別資訊應用方法，並分析各方法在不同的仿冒者性別組成下之表現，最後亦達成了在各種情況下皆能讓系統表現超越傳統作法之目標。

關鍵詞：語者驗證、性別資訊、性別分類器、i-向量、機率性線線性判別分析

doi:10.6342/NTU201801578

Abstract

For speaker verification task, one way to improve system's accuracy without changing the algorithm of acoustic model is to use gender-dependent model instead of gender-independent one. However, since test speakers' gender are not available, gender classifier plays an important role since its accuracy directly affects the performance of the whole speaker verification system; furthermore, ensuring that the system can maintain good performance under different gender composition of test speakers is also an important appeal.

To explore the impact of different gender information's usage on speaker verification system, this paper implemented a speaker verification system using i-vector and PLDA model as speaker feature and scoring model respectively, and 3 i-vector-based gender classifier. After analyzing the weakness of speaker verification system using gender-dependent model in a general way, we proposed several different methods for the application of gender information under the conditions when gender classifier has good and poor performance respectively; moreover, we analysis the performance of each method under different gender composition of test speakers as well. Finally, we reached the goal of making our system achieve better performance than tradition practice under different circumstances.

Keywords: Speaker Verification, Gender Information, Gender Classifier, i-vector, PLDA

doi:10.6342/NTU201801578

目錄

doi:10.6342/NTU201801578

doi:10.6342/NTU201801578

圖表目錄

doi:10.6342/NTU201801578

doi:10.6342/NTU201801578

第 1 章 緒論

近年來，個人電子設備以及網路的普及，使越來越多人習慣將大量個人資料存放於手機、電腦，甚至是雲端裝置。為防止資料外洩，許多設備會在資料被存取前驗證使用者的身份，而由於軟硬體技術之快速發展，諸如指紋、虹膜、人臉辨識等生物識別技術，已逐漸成為重要的身份驗證機制，相較於傳統輸入文字密碼的方式，生物識別兼具了更大的便利性與安全性。其中，由於語言是人類之間彼此交流最重要的手段之一，透過語音進行身份辨識的技術：語者辨識（speaker recognition），乃不可避免地成為生物識別領域的一大研究重點。

1.1 語者辨識簡介

語者辨識，或稱自動語者辨識（automatic speaker recognition），是一種將人類的語音透過電子設備轉換成電子訊號後，利用演算法分析不同語者之間的聲紋特性差異，來進行語者身份辨識的工作。語者辨識的應用領域相當廣泛，除了方才提及的個人設備存取控制之外，舉凡門禁系統、信用交易、犯罪偵測、電子會議輔助、刑事案件舉證等皆屬於其應用範疇；且由於其需求不斷擴大，如何增進語者辨識技術的準確度與或提升系統演算法之速度，至今仍吸引學界與業界投入大量研究心力。

1.1.1 語者識別與語者驗證

語者辨識依據應用情境又可細分為 2 個項子項目：語者識別（speaker identification）與語者驗證（speaker verification）。前者的工作是要在一群已知的語者之中識別出未知的語者，換句話說，當給定一段聲音訊號，語者識別系統必須自其資料庫中，判斷誰的聲音與此段聲音最為相像；而後者的工作則是要判斷兩段聲音是否屬於同一位語者，也就是驗證測試者所宣稱的身份是否為真。

doi:10.6342/NTU201801578

1.1.2 語者辨識之標準流程

典型的語者辨識流程可分為 3 項程序：訓練（training）、註冊（enrollment）與評測（evaluation 或 test）。訓練程序的目的在於訓練出後續程序需使用到的所有模型，例如：能從語者的聲音訊號擷取出特徵的聲學模型（acoustic model），以及為兩段不同的聲音計算聲紋相似度的評分模型（scoring model）等；註冊程序則是要為註冊語者客製化一套專屬於他／她的語者模型（speaker model）；而最後的評測程序，則是要將欲評測的聲音輸入註冊語者的語者模型，以計算出相似度分數，進而完成身份辨識的任務，在此程序中，語者識別系統會選擇分數最高的註冊語者作為識別的答案，而語者驗證系統則會依據分數是否高於一事先給定之門檻值來決定拒絕或接受。

1.1.3 語者辨識之評量指標

評斷語者辨識系統表現的指標有許多種，由於本論文所實作之系統屬於語者驗證之架構，故以下所介紹的各種指標將著重於語者驗證部分。傳統上，二元分類問題最常見的評量標準有準確度（accuracy）或 F-度量（F-measure）等，其定義如圖 1 所示：

		Answer	
		TRUE	FALSE
Predicted	TRUE	TP (True Positive)	FP (False Positive)
	FALSE	FN (False Negative)	TN (True Negative)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + TN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F-measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

圖 1、分類問題的常見指標

doi:10.6342/NTU201801578

而若要同時檢視語者驗證系統在不同的分數門檻下的效能，則可以檢測錯誤權衡曲線圖（Detection Error Rate Curve, DET Curve）或接收者操作特徵曲線圖（Receiver Operating Characteristic Curve, ROC Curve），此兩條曲線分別以正確接受率（true acceptance rate, i.e. $\frac{TP}{TP+FN}$）以及錯誤拒絕率（false rejection rate, i.e. $\frac{FN}{TP+FN}$）為縱軸，並以錯誤接受率（false acceptance rate, i.e. $\frac{FP}{FP+TN}$）為橫軸做圖，分別如圖 2-1 與圖 2-2 所示：

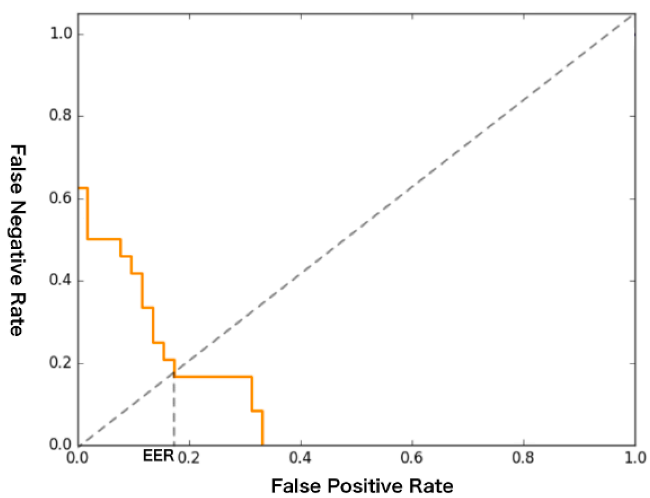


圖 2-1、錯誤權衡曲線圖

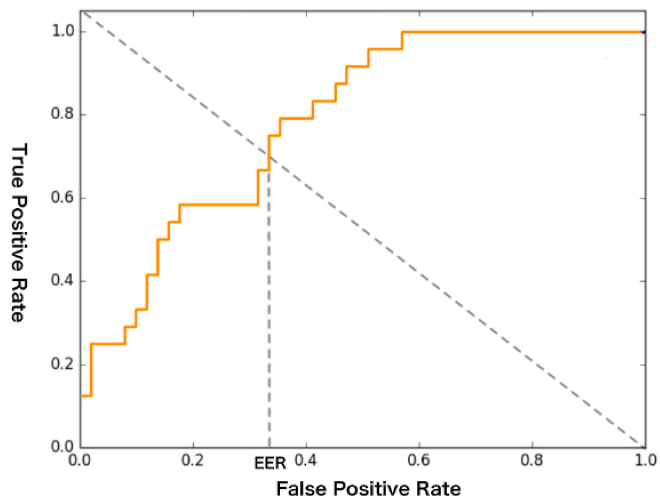


圖 2-2、接收者操作特徵曲線圖

此外，近年來諸多語者驗證相關比賽最常採納的一項評量標準為等錯誤率（Equal Error Rate, EER），其概念為：當分數門檻值能使系統錯分正樣本（positive trial）與負樣本（negative trial）之比例相同，也就是錯誤拒絕率等於錯誤接受率時，此錯誤率即為等錯誤率。由其概念可知，等錯誤率並不需要系統給定分數門檻，其所著重的乃在於系統能否「使正樣本的分數皆較負樣本高」，而非傳統上的「正確地接受正樣本並拒絕負樣本」。

doi:10.6342/NTU201801578

1.2 研究簡介

本碩士論文以語者驗證作為研究主軸，主要方向為探討「語者的性別資訊」對於語者驗證系統所產生的影響，探討的項目包括以下 3 點：

(1) 本論文將探討前人如何利用性別資訊提升語者驗證系統之辨識率，並分析該作法之弱點。

(2) 本論文將提出其他若干種性別資訊應用方法之改良，並分析各種方法適用之情況。

(3) 本輪文將實作若干種性別分類器，並分析各分類器優劣，以實際將其應用於語者驗證系統中。

1.3 章節概要

本論文之第一章為緒論，包含語者辨識與本論文研究方向之簡介；第二章為相關研究方法，包含本論文所設計的實驗中所使用到的各項相關技術之理論與演算法之簡介，包括 MFCC 特徵、高斯混合機率模型、i-向量、線性判別分析、機率性線性判別分析模型等；第三章為實驗設定與結果，包含本論文所使用的語料、資料配置以及參數設定之簡介，並詳述各項實驗結果與討論；第四章為結論，包含本論文研究之總結與未來方向。

doi:10.6342/NTU201801578

第 2 章 語音處理與語者辨識相關技術

2.1 語音訊號之特徵：MFCC

本 2.1 章之內容與圖片素材擷取自 [1] 之課堂內容或投影片。聲音經由電子設備錄製後會轉為電子訊號，若要對其進行後續處理與分析，便必須自其中抽取適當的特徵。本篇論文所抽取之特徵為 MFCC（Mel-Frequency Cepstral Coefficients）特徵，是一種計算複雜度低、在語音訊號處理上最被廣泛使用的特徵之一，圖 3 為抽取 MFCC 特徵之標準流程，以下將對每個步驟進行簡介。

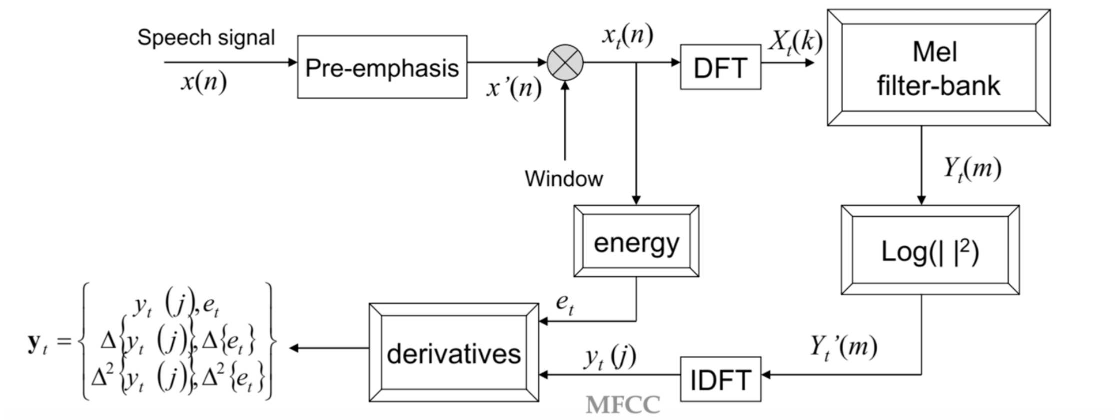


圖 3、MFCC 特徵之抽取流程

MFCC 特徵抽取流程之第一步為預強調（pre-emphasis），其目的是為補償語音訊號中衰減的高頻部分。自人體發出的聲音可分為兩種類型，一種是在時域上具明顯基本週期的有聲音（voiced sound），另一種則是不具明顯基本週期的無聲音（unvoiced sound），圖 4-1 與 圖 4-2 分別則顯示了兩者的時域圖與頻域圖：

doi:10.6342/NTU201801578

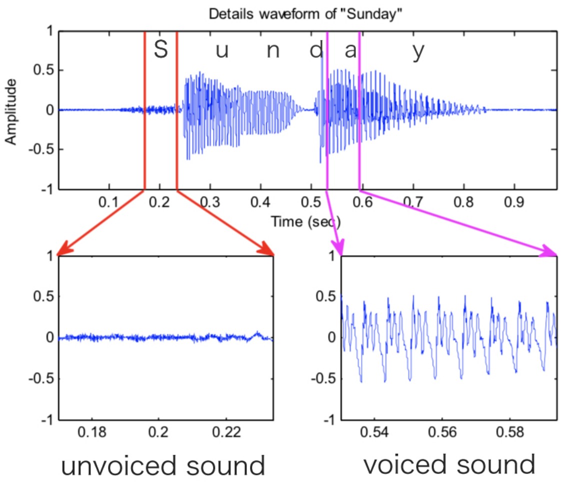


圖 4-1、有聲音與無聲音之時域圖

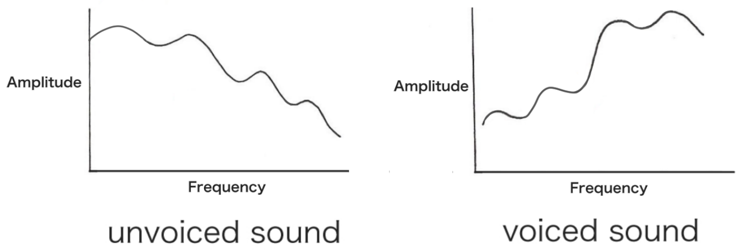


圖 4-2、有聲音與無聲音之頻域圖

其中，有聲音對語音訊號分析來說較能提供有用資訊，但其高頻部分容易在傳輸過程中損耗，導致有聲音在頻域的強度分佈情形呈現低頻強、高頻弱的衰減趨勢。此衰減趨勢容易導致一些重要資訊流失，例如共振峰結構（formant structure）變得不明顯，因此需要透過此預強調流程來彌補。假設原始聲音訊號在時域上以 $x(t)$ 表示，則通過預強調步驟後的新訊號 $x'(t)$ 可表示為：

$$x'(t) = x(t) - a \cdot x(t-1)$$

其中 a 之值通常取在 0.9 到 1 之間。其背後原理為：假設原始訊號 $x(t)$ 通過經 Z 轉換（Z-transform）後可表示為：

$$Z[x(t)] = X(z)$$

則新訊號 $x'(t)$ 通過 Z 轉換後則可表示為：

$$Z\left[x(t) - a \cdot x(t-1)\right] = X(z) \cdot (1 - a \cdot z^{-1})$$

由上式可以看到，新訊號在頻域上即為原始訊號乘上 $(1 - a \cdot z^{-1})$，而此值在低頻較接近 0，越往高頻靠近則約接近 1，因此此流程之原理可視為將原始訊號通過高通濾波器，以補償語音訊號中衰減的的高頻部分。

訊號通過預強調處理後，整段音訊會被切分成許多等長、允許重疊的音幀（frame 或 window），如圖 5 所示：

doi:10.6342/NTU201801578

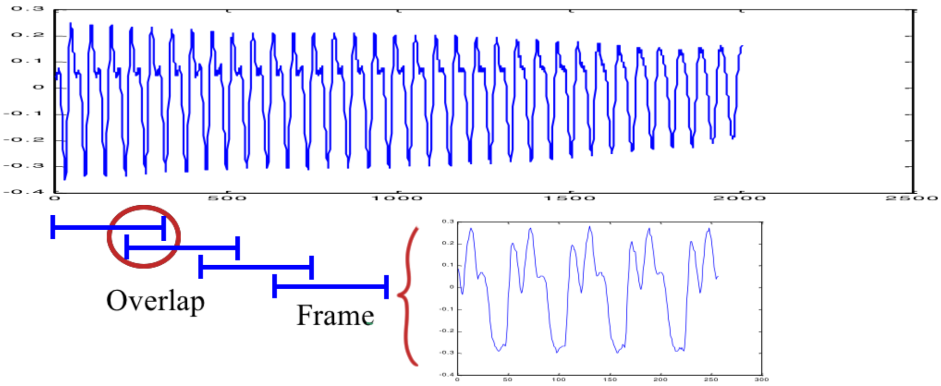


圖 5、聲音訊號與音幀

一個音幀的持續時間相當短，而聲音訊號在極短的時間內，特性是較穩定且一致的，因此較容易擷取出有用的資訊。用來從訊號中切出音幀的函數稱為音窗函數（windowing function），將訊號與音窗函數在時域上進行卷積運算（或在頻域上進行乘積運算）即可得到單一音幀長度之訊號。最常被使用的音窗函數為長方窗（rectangular window）與漢明窗（Hamming window），兩者在時域之數學式可表示為：

長方窗：$w(t) = 1$

漢明窗：$w(t) = \alpha - \beta cos\dfrac{2\pi t}{N-1}, \ \alpha = 0.54, \ \beta = 0.46$

而兩者的時域圖與頻域圖則如圖 6-1 與圖 6-2 所示 [2]：

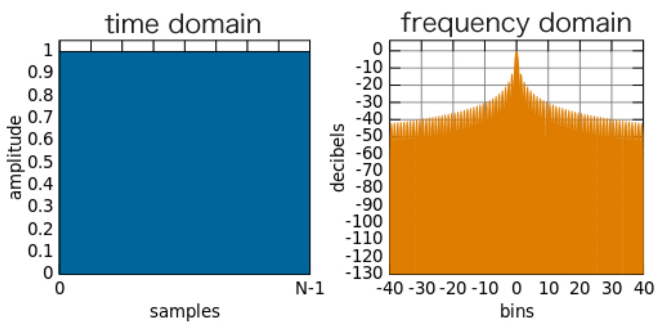


圖 6-1、長方窗之時域圖與頻域圖

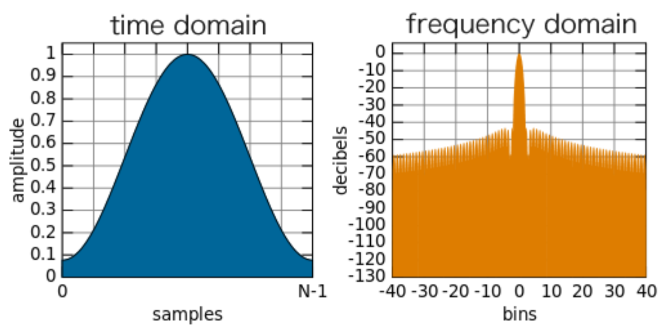


圖 6-2、漢明窗之時域圖與頻域圖

doi:10.6342/NTU201801578

在兩者的頻域圖上，正中間的最高峰稱為主瓣（main lobe），周圍其他較小的峰群稱為旁瓣（side lobe）。在語音訊號分析上，通常希望旁瓣的峰高遠低於主瓣，因此由圖可見，漢明窗是比長方窗更理想的一種窗函數。

訊號被切割成音幀後，會進行離散傅立葉變換（Discrete Fourier Transform, DFT）將訊號從時域轉換到頻域，接著通過梅爾濾波器群（Mel-filter bank），並各自取頻率之係數和，如圖 7 所示：

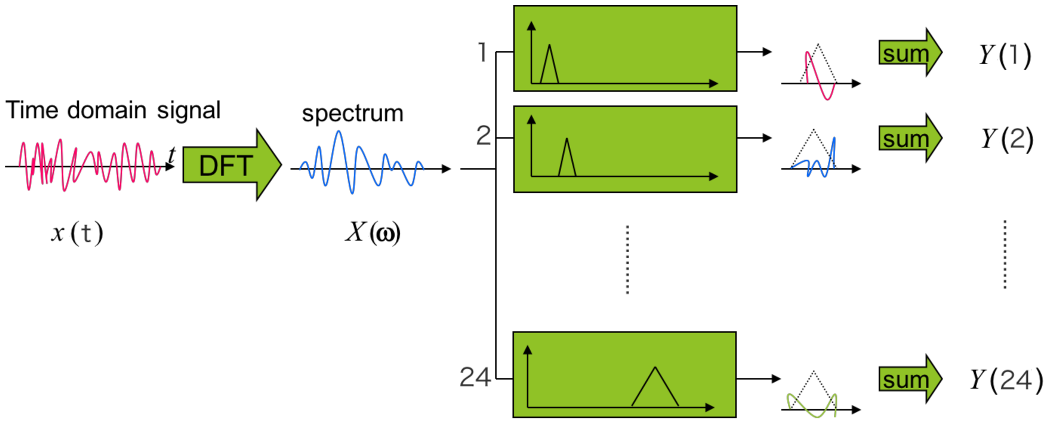


圖 7、離散傅立葉變換與梅爾濾波器

由圖中可見，梅爾濾波器群為 24 個三角帶通濾波器，各濾波器對應之頻帶則稱為關鍵頻帶（critical band）。這些關鍵頻帶是依據人類兩大內耳特性所設計：第一，人類的內耳構造為螺旋狀，其內部的不同區段各自負責不同音訊頻帶之處理，如圖 8 所示：

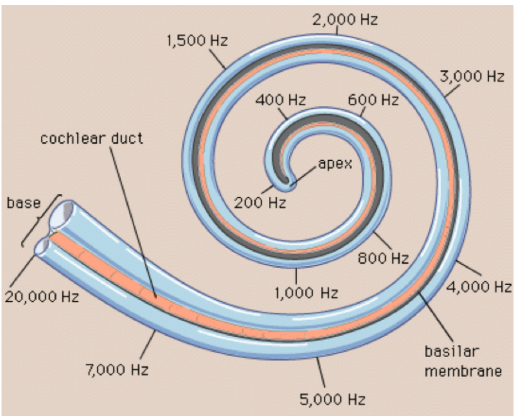


圖 8、人類的內耳構造

doi:10.6342/NTU201801578

並且，內耳相鄰的區段所負責的頻帶是會重疊的，因此相鄰的關鍵頻帶在設計上亦會有所重疊的；第二，內耳越靠近外部的區段所負責的頻帶越高，而且頻率的增長速度不一：較靠近內部的區段（約 200 Hz 至 1000 Hz），頻率增長速率近似等速，但較靠近外部的區段（約 1000 Hz 以後），則近似於指數增長。這代表人類的內耳對高頻率的聲音解析度較差，意即人耳對高頻區段的聲音較不敏感。

接著，由梅爾濾波器群所算出來的 24 個係數，會各自取絕對值與平方，最後進行對數運算；接著，24 個值將當作頻域係數，進行反離散傅立葉變換（Inverse Discrete Fourier Transform, IDFT），如圖 9 所示：

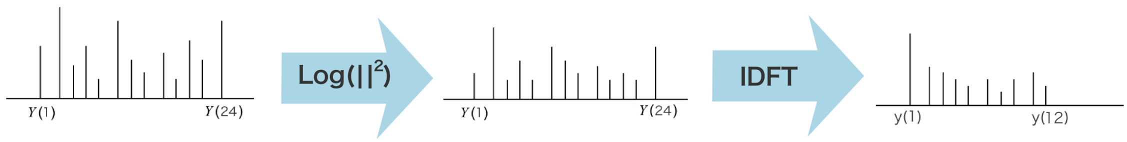


圖 9、絕對值平方、對數運算與反離散傅立葉變換

取絕對值與平方的原因，是為了捨棄在語音處理上較無用處的「向位角」資訊，而只保留「震幅」資訊。而取對數的主要原因有二：第一，人耳對聲音能量的敏感度約與能量大小取對數值成正比，因此取對數較能模擬人內耳之真實情況；第二，取對數後，在進行反離散傅立葉變換時，較利於去除多餘的雜訊。一段聲音在頻域圖中的變化趨勢可拆解為兩部分，一為起伏較大的共振峰結構（formant strucure），另一則為起伏較小但震盪速度較快的激發雜訊（excitation noise），如圖 10 之左側所示：

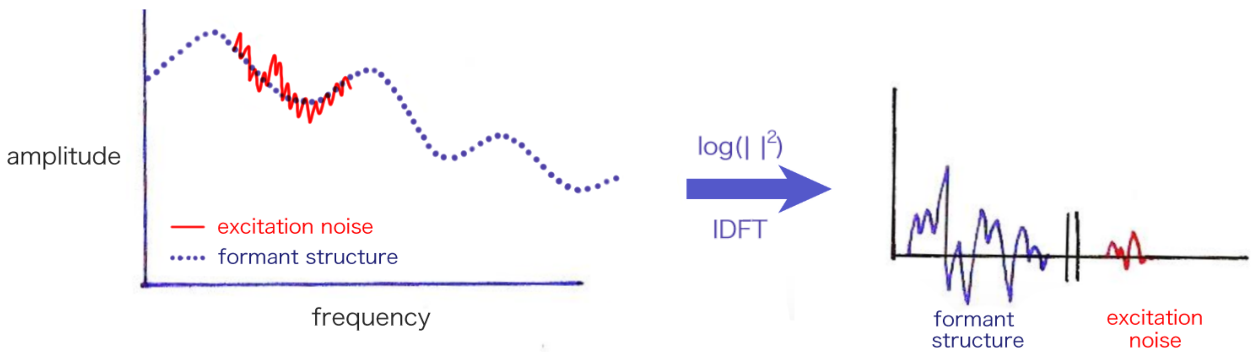


圖 10、去除激發雜訊之原理

doi:10.6342/NTU201801578

其中，單單共振峰結構便幾乎足以決定聲音的音型，因此去除激發雜訊較有利於後續之分析。假設在頻域上，原始聲音訊號 $X(\omega)$ 可表示為共振峰結構 $G(\omega)$ 與激發雜訊 $U(\omega)$ 之乘積：

$$X(\omega) = U(\omega)G(\omega)$$

但進行對數運算後，原式會由乘式轉為加式：

$$log\big(X(\omega)\big) = log\big(U(\omega)\big) + log\big(G(\omega)\big)$$

再者，在進行反離散傅立葉轉換後，更可發現共振峰結構與激發雜訊呈現出左右分離之現象，如圖 10 右側所示，由此可知，越靠左側的訊號，越有可能屬於共振峰結構，因此，如圖 9 所示，在進行反離散傅立葉轉換後，僅須取出前半部的 12 個的係數。

經反離散傅立葉變換所得到 12 個係數，通常會額外加上另一個代表能量的係數。假設訊號在一音幀內表示為 $x(t)$，則一般常見之能量係數 e 之計算方式為：

$$e = 10 \cdot log\Big(\sum_t x(t)^2\Big)$$

12 個係數再加上此 1 能量係數所組成的 13 個係數，即為最簡單的梅爾頻率倒譜係數（Mel-Frequency Cepstral Coefficients, MFCCs）。

MFCC 特徵抽取的最後一階段為導數運算（Derivative operation）。聲音訊號在傳輸過程中可能會混入直流雜訊（DC noise），進而對分析結果造成影響，假設真實訊號 $\overline{x'}$ 可表示為原始訊號 $\overline{x}$ 與直流雜訊 $\overline{h}$ 之加乘，如圖 11 所示：

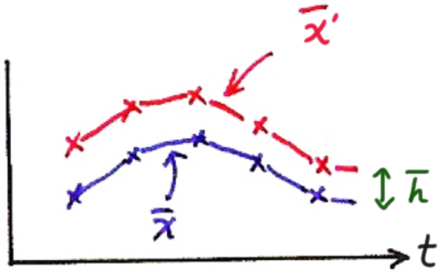


圖 11、直流雜訊示意圖

doi:10.6342/NTU201801578

由圖中可以發現，即使加上直流雜訊，原始訊號與真實訊號的變化趨勢仍是一致的，故兩者之斜率並不會因直流雜訊而有所不同。因此，以微分的方式取得梅爾頻率倒譜係數之導數，並將其合併進原來的梅爾頻率倒譜係數，是減少直流雜訊對特徵影響的最常見方法。由於電子訊號是由不連續的訊號點組成，故必須以前後若干訊號點以不同權重加乘，來模擬導數的計算，如圖 12 所示：

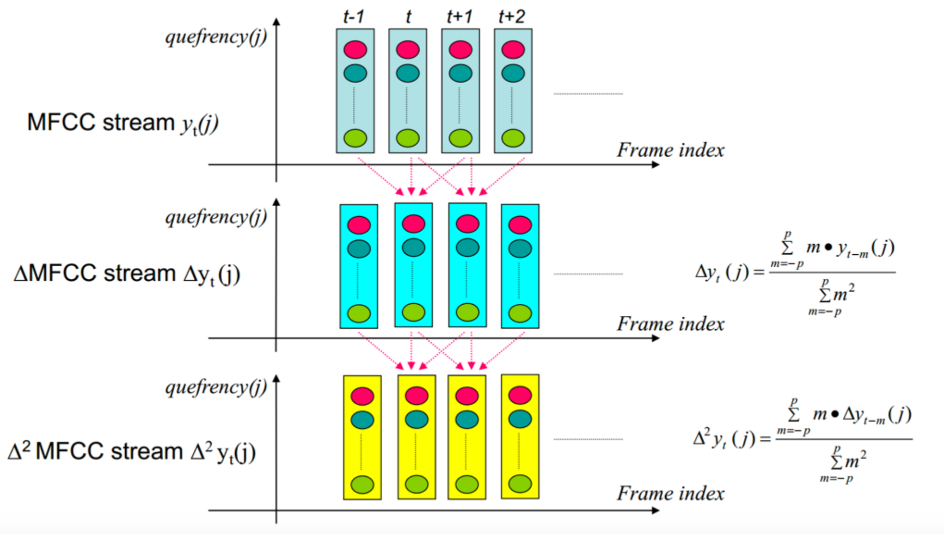


圖 12、MFCC 與其導數之計算

通常，MFCC 特徵會計算 2 階的導數，因此原來的 13 維特徵加上另外兩階導數的 26 維特徵，便可組成 39 維的特徵。

doi:10.6342/NTU201801578

2.2 高斯混合機率模型

為兩段語音訊號抽取出一連串特徵向量（如： MFCC）後，必須有適當的模型來判斷兩段聲音是不是屬於同一位語者。語者辨識領域較早期所流行的是一些使用向量量化（Vector Quantization, VQ）相關技術的方法，而後來取而代之的則是接下來將要介紹的高斯混合模型（Gaussian Mixture Model, GMM）。

2.2.1 高斯混合模型概論

高斯混合模型的目的與一般高斯模型相同，皆是用以模擬多維資料之機率分佈，但其內結構部則是由若干個高斯分量以不同權重加權所組成，而其中每一個高斯分量皆是一個高斯密度函數。圖 13-1 與 圖 13-2 分別為一般高斯模型與高斯混合模型之示意圖：

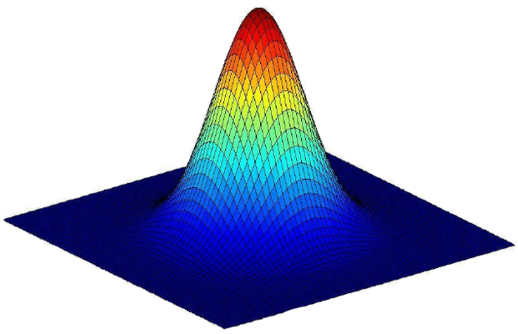


圖 13-1、一般高斯模型

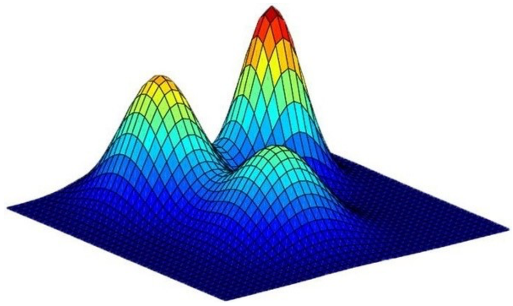


圖 13-2、高斯混合模型

為方便後續說明，此處定義高斯混合模型之高斯分量總數為 C，MFCC 特徵之維度數為 $F(=39)$。一個高斯混合模型的參數可以數學式表示為：

$$\lambda = \{\pi_c, \mu_c, \Sigma_c \mid c = 1,2,...,C\}$$

其中，λ 代表高斯混合模型之所有參數集合，π_c 、μ_c 與 Σ_c 分別代表第 c 個高斯分量之權重、平均向量與共變異數矩陣（covariance matrix），並且滿足：

$$\sum_c \pi_c = 1$$

doi:10.6342/NTU201801578

假設 x 為一高斯混合模型所欲模擬之多維變量，則其機率密度函數可表示為：

$$f(x\,|\,\lambda) = \sum_{c} \pi_c \cdot \mathcal{N}(x\,|\,\mu_c, \Sigma_c)$$

其中，$\mathcal{N}(x\,|\,\mu_c, \Sigma_c)$ 代表平均為 μ_c 、共變異數矩陣為 Σ_c 之高斯機率密度函數，即：

$$\mathcal{N}(x\,|\,\mu_c, \Sigma_c) = \frac{1}{(2\pi)^{F/2} \cdot |\Sigma_c|^{1/2}} \cdot \exp\Big(-\frac{1}{2}(x-\mu_c)^* \cdot \Sigma_c^{-1} \cdot (x-\mu_c) \Big) \tag{1}$$

其中 * 表示取矩陣之共軛轉置。藉由高斯混合模型，我們得以計算「一串特徵向量被觀察到的機率」：今考慮一位語者 s 與其聲音訊號所產生的一連串特徵向量 $W_s = \{W_{st}\,|\,t = 1,2,3,...\}$，則在給定高斯混合模型 λ 之前提下，觀察到此一連串特徵向量之機率為：

$$p(W_s\,|\,\lambda) = \prod_{t} f(W_{s,t}\,|\,\lambda) \tag{2}$$

此值乃為觀測值 W_s 於給定模型參數 λ 下之似然函數（likelihood function）。

以高斯混合模型為基礎進行語者辨識的方法中，以似然率測定（likelihood ratio test）為大宗。此方法會使用到兩種高斯混合模型，一種是語者相關模型（speaker-dependent model），另一種則是通用背景模型（Universal Background Model, UBM）。語者相關模型是以單一註冊語者的語料進行訓練，其所模擬的是該位語者的特徵機率分佈；而通用背景模型則是以大量語者的語料進行訓練，其所模擬的是世界上（除了註冊語者之外的）所有語者的共通特徵機率分佈，因此亦可視其為「語者不相關模型」。假設今欲驗證一串特徵向量 X 是否由語者 s 所產生，則此方法會以下列數學式算出分數：

$$score_{LR} = \frac{p(X\,|\,\lambda_s)}{p(X\,|\,\lambda_0)}$$

doi:10.6342/NTU201801578

其中 λ_s 與 λ_0 分別代表語者相關模型與通用背景模型。此分數的計算概念是「X 是由語者 s 所產生」與「X 是由其他人所產生」的機率比值，若此分數高於或等於一事先設定好的門檻值，則判定 X 屬於語者 s 的聲音，反之則判定 X 不屬於語者 s 的聲音。又為了避免分數之可能數值範圍過於極端，因此通常會對分數進行對數運算，使其成為：

$$score_{LLR} = \log\left(p(X|\lambda_s)\right) - \log\left(p(X|\lambda_0)\right)$$

此方式又稱為對數—似然率測定（log-likelihood ratio test）。以下將分別簡介語者相關模型與通用背景模型之訓練方式。

2.2.2 通用背景模型與其訓練流程

通用背景模型是以最大似然度估計法（Maximum Likelihood Estimation, MLE）為原則進行訓練，其概念是要使模型參數能讓訓練資料的似然函數最大化，即：

$$\lambda_0 = \underset{\lambda}{argmax} \prod_{s} p(W_s|\lambda)$$

其中 s 為訓練資料中的任一語者。但由於上式對於 λ 而言並非線性函數，因此便難以用解析解的方式直接解出此方程式，為此，此處引入一套迭代式的訓練流程：最大期望（Expectation-Maximization, EM）演算法 [3]。最大期望演算法一般分為 3 步驟，分別為模型參數初始化、計算期望（Expectation）與最大化（Maximization），其中後兩步驟可簡稱為 E 步驟與 M 步驟。E 步驟是在給定模型參數之情形下，估計潛在變數之期望值，並計算似然度；M 步驟則是使用 E 步驟之結果，更新模型參數。最大期望演算法的運作流程會令 E 步驟與 M 步驟交替執行，且每次迭代皆會使似然度有所提升，直到收斂為止，如圖 14 所示：

doi:10.6342/NTU201801578

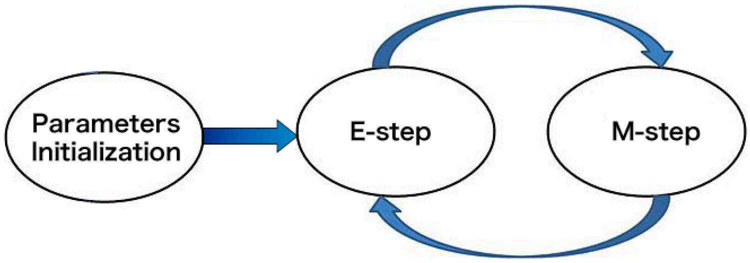


圖 14、最大期望值演算法之流程示意圖

以下將分別簡介通用背景模型訓練流程之 3 步驟。

通用背景模型訓練流程之第一步為初始化模型參數，也就是必須選定其每一個高斯分量的權重、平均向量與共變異數矩陣。最簡單的初始化方法是隨機指定每個參數之初始值，但這裡介紹一個較為常用且效果較佳的方法：k-平均（k-means）演算法。k-平均演算法屬於向量量化方法中的其中一種，是一種非監督式學習的分群方法，其目的為把空間中的若干樣本點劃分為 k 個群落，使得每個樣本點所屬群落之群心皆離此樣本點最近。以 k-平均演算法在訓練資料找出的每個群落皆可作為通用背景模型的一個高斯分量，每個高斯分量的平均向量、共變異數矩陣可由對應群落的樣本點計算而得，而其權重則正比於該群落的樣本點數量。標準的 k-平均演算法分為群心初始化、分配（assignment）與更新（update）等 3 個步驟：初始化步驟為隨機估計 k 個群心的位置；分配步驟則是依據群心位置，為每個樣本點進行分群，使得每個樣本點所屬群落之群心到該樣本點有最短距離；更新步驟則是更新每個群落的群心位置，新的群心位置即為該群落中所有樣本點的平均值。k-平均演算法會令分配與更新步驟交替執行，直到群心位置不再改變為止，其流程如圖 15 所示：

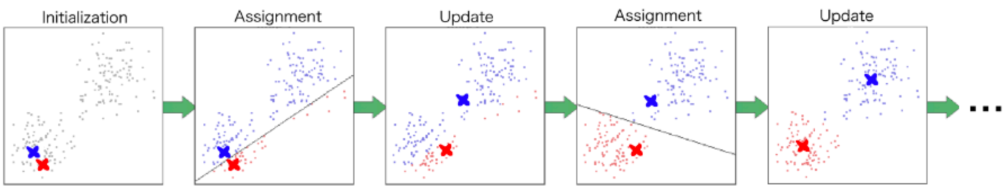


圖 15、k-平均演算法之流程示意圖

doi:10.6342/NTU201801578

E 步驟之主要目的為利用現有的模型參數，計算似然函數值，此值之計算方法與先前提到的 (2) 相同；而 M 步驟之目的為更新模型參數，為方便討論，通常會假設各高斯分量之共變異數矩陣為一純量與單元矩陣之乘積，即：

$$|\Sigma_c| = \sigma_c^2 \cdot I$$

此時 (1) 便可修改為：

$$\mathcal{N}(x\,|\,\mu_c, \Sigma_c) = \frac{1}{(2\pi)^{F/2} \cdot \sigma_c^F} \cdot \exp\Big(-\frac{1}{2\sigma^2}(x-\mu_c)^* \cdot (x-\mu_c)\Big)$$

由於整個最大期望演算法的目的在於將似然度最大化，因此此處將目標函數定義為似然函數取自然對數，即：

$$J(\lambda) = \ln\Bigg(\prod_t \Big(\sum_c \pi_c \cdot \mathcal{N}(x_t\,|\,\mu_c, \Sigma_c)\Big)\Bigg)$$

其中 x_t 代表訓練資料總集中的第 t 個特徵向量。而模型參數的更新必須使目標函數最大化，意即必須滿足此函數對於各模型參數之偏微分值為零的條件，為方便後續說明，此處定義 $\gamma_t(c)$ 為第 c 個高斯分量給定 x_t 的後驗機率（posterior probability），即：

$$\gamma_t(c) = P(c\,|\,x_t) = \frac{\pi_c \cdot \mathcal{N}(x_t\,|\,\mu_c, \Sigma_c)}{\sum_i \pi_i \cdot \mathcal{N}(x_t\,|\,\mu_i, \Sigma_i)} \tag{3}$$

經運算，可得各模型參數的更新值為：

$$\mu_c = \frac{\sum_t \big(\gamma_t(c) \cdot x_t\big)}{\sum_t \gamma_t(c)}$$

$$\sigma_c^2 = \frac{1}{F} \cdot \frac{\sum_t \big(\gamma_t(c) \cdot (x_t-\mu_c)^* \cdot (x_t-\mu_c)\big)}{\sum_t \gamma_t(c)}$$

$$\pi_c = \frac{1}{T}\sum_t \gamma_t(c)$$

doi:10.6342/NTU201801578

其中 T 為訓練資料中特徵向量總數。依照上述的 E 步驟與 M 步驟更新模型參數時，每次迭代皆會確保訓練資料之對數似然度有所提升，以數學符號可以表示為：

$$\sum_s \log\left(p(W_s|\lambda)\right) \geq \sum_s \log\left(P(W_s|\lambda_0)\right)$$

其中模型參數 λ_0 代表模型參數 λ 更新前的狀態。有關上述公式之更詳細數學過程與證明，詳見 [3] 。

2.2.3 語者相關模型與其訓練流程

語者相關模型之訓練方式有兩種，第一種和通用背景模型相同，即取該位語者之所有語料以最大似然度估計法進行訓練，但是最大期望演算法之迭代過程相當費時，尤其註冊語者數量眾多，更會付出龐大的時間代價；另外，當一位語者所能提供的語料過少，此方法未必能建立強健的語者相關模型。因此，此處導入第二種訓練方法：貝氏調適法（Bayesian adaptation） [4] [5]，或稱最大後驗機率調試法（Maximum A Posterior adaptation, MAP adaptation）。

貝氏調適法之概念乃以通用背景模型為基礎，以註冊語者之語料對模型參數進行微量調整，進而形成語者相關模型，圖 16 乃為其過程示意圖：

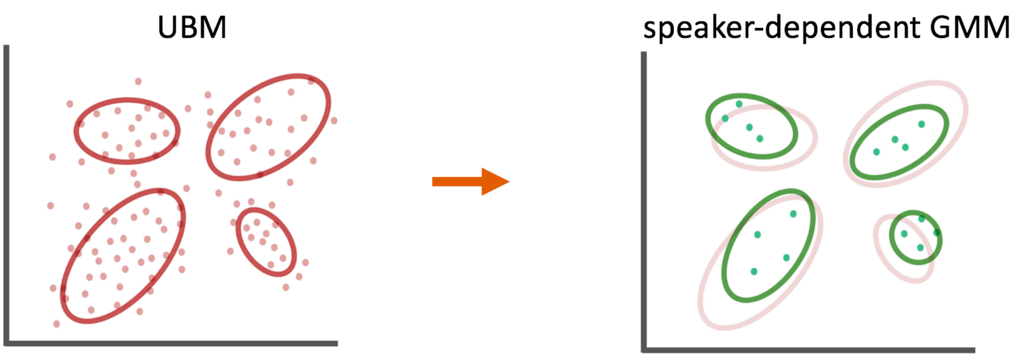


圖 16、貝氏調適法之過程示意圖

doi:10.6342/NTU201801578

為方便後續說明，此處首先定義一些數學符號，首先，給定一位語者 s 與其所產生的一連串 MFCC 特徵向量$W_s = \{W_{st} \mid t = 1,2,3,...\}$，則對於每個高斯分量 c，可定義三項鮑姆韋爾奇統計量（Baum-Welch statistics）：

$$N_c(s) = \sum_t \gamma_{s,t}(c) \tag{4}$$

$$F_c(s) = \sum_t \gamma_{s,t}(c) \cdot W_{s,t} \tag{5}$$

$$S_c(s) = diag\left(\sum_t \gamma_{s,t}(c) \cdot W_{s,t} \cdot W_{s,t}^* \right) \tag{6}$$

其中 $\gamma_{s,t}(c)$ 同 (3)，代表通用背景模型的第 c 個高斯分量給定 $W_{s,t}$ 之後驗機率，$diag()$ 則表示取矩陣中的主對角線元素。三個統計量中，$N_c(s)$ 又稱為零階統計量（0th statistics），格式為純量；$F_c(s)$ 又稱為一階統計量（1st statistics），格式為 $F \times 1$ 之向量；$S_c(s)$ 又稱為二階統計量（2nd statistics），格式為 $F \times F$ 之矩陣。接著，依據這些統計量，可求出特徵向量對於高斯分量 c、給定 W_s 的平均向量與共變異數矩陣之期望值：

$$E_c[x \mid W_s] = \frac{F_c(s)}{N_c(s)}$$

$$E_c[x \cdot x^* \mid W_s] = \frac{S_c(s)}{N_c(s)}$$

假設對於高斯分量 c，通用背景模型之參數以 $\{\pi_{0,c}, \mu_{0,c}, \Sigma_{0,c}\}$ 表示，則語者相關模型之參數 $\{\pi_{s,c}, \mu_{s,c}, \Sigma_{s,c}\}$ 便可以依據以下之公式進行調整：

$$\pi_{s,c} = \left(\frac{\alpha_c \cdot N_c(s)}{T_s} + (1 - \alpha_c) \cdot \pi_{0,c} \right) \cdot \beta_s$$

$$\mu_{s,c} = \alpha_c \cdot E_c[x \mid W_s] + (1 - \alpha_c) \cdot \mu_{0,c} \tag{7}$$

doi:10.6342/NTU201801578

$$\Sigma_{s,c} = \alpha_c \cdot E_c[x \cdot x^* \mid W_s] + (1-\alpha_c) \cdot (\Sigma_{0,c} + \mu_{0,c} \cdot \mu_{0,c}^*) - (\mu_{s,c} \cdot \mu_{s,c}^*)$$

其中，T_s 為語者 s 之特徵向量總數，β_s 之作用是在確保各高斯分量之權重參數 $\pi_{s,c}$ 總和為 1，而 α_c 定義如下：

$$\alpha_c = \frac{N_c(s)}{N_c(s) + r_r}$$

其中 r_r 扮演著控制通用背景模型的模型參數會被多大程度地往註冊語者的方向調整的相關因子。

雖然以似然率測定法計算出的分數可以作為判斷兩位語者相似度之依據，但高斯混合模型在本論文研究中僅扮演中間產物的角色。以下將介紹以 MFCC 特徵與高斯混合模型為基礎、近年來語者辨識領域最強健的特徵之一：i-向量。

doi:10.6342/NTU201801578

2.3 因子分析與 i-向量

i-向量（i-vector）為近年來效果最好、最常被運用在語者辨識相關工作上的特徵之一，亦是用於語者辨識的因子分析（Factor Analysis, FA）方法中集大成者，以下將簡介 i-向量之演變歷程、演算法以及其理論 [6] [7] [8]。

2.3.1 i-向量之概念與其演變歷程

MFCC 特徵是以音幀為單位抽取的特徵，換句話說，不同長度的聲音訊號，會抽出不同個數的 MFCC 特徵，而由於許多模型必須固定輸入特徵的長度，因此相較於 MFCC 特徵，以一整段聲音訊號或一位語者為單位抽取的特徵將更加適用。符合上述條件的辦法之一，是將以貝氏調適法得到的語者相關高斯混合模型的參數（例如每個高斯分量的平均）組成一個高維的超向量（supervector），來作為一位語者的特徵向量。但是，如此高維的特徵向量勢必帶有許多不必要的資訊，在運算上也較費時，面對這些問題，因子分析是一個非常有用的解決方法，它可以從高維且彼此相依的可觀變數（observable variables）中，將我們所感興趣的資訊，濃縮到較低維且彼此不相依的潛在變數（latent variables），進而幫助後續分析。

為方便後續說明，此處定義高斯混合模型之高斯分量總數為 C，MFCC 特徵之維度數為 $F(=39)$。最早成功為語者辨識工作對超向量進行因子分析的方法之一為本徵語音調適法（eigenvoice MAP），其概念為：一個語者相關的超向量，可被分解成「語者不相關項」與「語者相關項」兩部份，數學式表示為：

$$M_s = m + V \cdot y_s$$

其中 M_s 為與語者 s 相關之超向量，是以語者 s 之語料調適而成的可觀變數，大小為 $CF \times 1$；m 為通用背景模型所產生的超向量，大小為 $CF \times 1$；V 為一低秩之矩陣，大小為 $CF \times R_s$，而 R_s 將決定潛在變數之維度，通常設定 $R_s \ll CF$，

doi:10.6342/NTU201801578

而本徵語音就是由 V 的直行所組成；y_s 為潛在變數之向量，大小為 $R_s \times 1$，呈現標準常態分佈，其每一項元素又稱為語者因子（speaker factor）。很明顯地，等號右側的左項 m 即為「語者不相關項」，右項 $V \cdot y_s$ 則為「語者相關項」，且其中的 y_s 便可作為帶有濃縮語者資訊之新特徵。

回顧貝氏調適法用來更新高斯混合模型之平均的 (7) 式，若將等式右側之左項 $\alpha_c \cdot E_c[x \mid W_s]$ 視作「語者相關項」，右項 $(1-\alpha_c) \cdot \mu_{0,c}$ 則視作「語者不相關項」，整條式子便可仿照本徵語音調適法之模式修改為：

$$M_s = m + D \cdot z_s$$

其中 D 為一對角矩陣，大小為 $CF \times CF$，通常令 $D = \frac{1}{r}\Sigma$，Σ 為通用背景模型之超共變異數矩陣（super covariance matrix），由通用背景模型中所有共變異數矩陣以主對角線方向串接而成，即：

$$\Sigma = \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_C \end{bmatrix}$$

其中 Σ_c 為對應到高斯分量 c 之共變異數矩陣，而 r 則為相關因子（relevance factor）；z_s 為潛在變數之向量，大小為 $CF \times 1$，呈現標準常態分佈，其每一項元素又稱為語者因子（speaker factor）。此方法又稱為古典調適法（classical MAP），古典調適法與本徵語音調適法之優缺點具有互相補償的性質：古典調適法需要較大量語料進行語者調適，而當語者調適的語料數量有限時，本徵語音調適法則表現得相對穩健，且訓練之收斂速度快；但本徵語音調適法則有者潛在變數 y_s 之維度 R_s 不能超過訓練語者的人數的缺點。

除了每個人的聲音彼此不相同，諸如錄音設備、錄音環境等因素所造成的信道效應，亦會影響信號之聲紋特性，進而對語者辨識結果產生不良影響。因此，若能消除信道效應，便能使得最後的特徵具有較純粹的語者資訊，而針對信道進

doi:10.6342/NTU201801578

行調試的因子分析方法：本徵信道調適法 （eigenchannel MAP）便由此衍生。其概念為：一個語者、信道相關的超向量，可被分解成「信道不相關項」與「信道不相關項」兩部份，數學式表示為：

$$M_{s,h} = m + D \cdot z_s + U \cdot x_{s,h}$$

其中 $M_{s,h}$ 為與語者 s 、信道 h 相關之超向量，是以語者 s 的信道 h 語料調適而成的可觀變數，大小為 $CF \times 1$；$m + D \cdot z_s$ 與古典調適法相同，在此處則扮演「信道不相關項」之角色；剩下的 $U \cdot x_{s,h}$ 則為「信道相關項」，其中 U 為一低秩之矩陣，大小為 $CF \times R_c$，而 R_c 將決定信道潛在變數之維度，通常設定 $R_c \ll CF$，而本徵信道就是由 U 的直行所組成；$x_{s,h}$ 為信道潛在變數之向量，大小為 $R_c \times 1$，呈現標準常態分佈，其每一項元素又稱為信道因子（channel factor）。

為有效整合上述因子分析方法，聯合因子分析（Joint Factor Analysis, JFA）應運而生，該方法融合了本徵語音模型與本徵信到模型之概念，將語者、信道相關的超向量，分解為「語者、信道不相關項」、「語者相關項」、「信道相關項」與「殘留（residual）項」，數學式表示為：

$$M_{s,h} = m + V \cdot y_s + U \cdot x_{s,h} + D \cdot z_{s,h}$$

聯合因子分析在分析超向量時同時考慮了語者效應與信道效應，可說是將前述因子分析方法的優點集結於一身，而其語者因子向量 y_s 可以直接作為語者特徵輸入語者辨識模型並達到良好的效果。但是，後來的部分研究卻發現其信道因子向量 $x_{s,h}$ 除蘊含信道資訊外，亦帶有部分語者資訊，換言之，聯合因子分析的語者因子本身所帶有的語者資訊並不完整。

綜上所述，為擷取更有效的語者特徵，語者因子與信道因子必須同時被考慮，而建立在此概念上所誕生的新方法：i-向量，成為了目前效果最佳的因子分析方法。i-向量之全名為 identity vector（意為 vector for speaker-identification

doi:10.6342/NTU201801578

application）或 intermediate vector（vector with intermediate dimension between those of a supervector and an acoustic feature vector），不同於聯合因子分析模型以本徵語音矩陣與本徵信道矩陣分別定義了語者因子與信道因子之空間分佈，i-向量模型將語者空間與信道空間合而為一，並只產出一組潛在變數，其數學式表示為：

$$M_{s,h} = m + T \cdot w_{s,h}$$

其中 T 稱為總體變異矩陣（Total variability matrix），為一低秩之矩陣，大小為 $CF \times R_t$，而 R_t 將決定潛在變數之維度，通常設定 $R_t \ll CF$；$w_{s,h}$ 為潛在變數之向量，亦是我們要求的 i-向量，其大小為 $R_t \times 1$，呈現標準常態分佈，其每一項元素又稱為總體因子（total factor）。

2.3.2 i-向量模型之訓練流程

i-向量可說是目前因子分析系列方法之中的集大成者，以下將簡單介紹其總體變異矩陣之訓練方法，以及如何由一段聲音中擷取 i-向量特徵；須特別注意的是，以下過程中，同一語者之不同信道的聲音，會直接被視作不同語者的聲音，因此會將 i-向量原本之符號 $w_{s,h}$ 省略 h 而表示 w_s。在說明總體變異矩陣之訓練流程之前，此處首先定義一些數學符號。首先，給定一位語者 s 與其所產生的可觀變數 $W_s = \{W_{st} \mid t = 1,2,3,...\}$（即 MFCC 特徵向量），則對於每個高斯分量 c，分別定義 $N_c(s)$、$F_c(s)$、$S_c(s)$ 為零階、一階與二階鮑姆韋爾奇統計量，此處定義與 (4)、(5)、(6) 式相同。接著，定義標準化的一階、二階統計量：

$$\tilde{F}_c(s) = \sum_t \gamma_{s,t}(c) \cdot (W_{s,t} - m_c)$$

$$\tilde{S}_c(s) = diag\Big(\sum_t \gamma_{s,t}(c) \cdot (W_{s,t} - m_c) \cdot (W_{s,t} - m_c)^* \Big)$$

doi:10.6342/NTU201801578

其中 m_c 為 m 對應到第 c 個高斯分量之子向量，而 $\tilde{F}_c(s)$、$\tilde{S}_c(s)$ 之格式則分別與 $F_c(s)$、$S_c(s)$ 相同。若進一步推導上述數學式，則可以得到：

$$\tilde{F}_c(s) = F_c(s) - N_c(s) \cdot m_c$$

$$\tilde{S}_c(s) = S_c(s) - diag\left(F_c(s) \cdot m_c^* + m_c \cdot F_c(s)^* - N_c(s) \cdot m_c \cdot m_c^*\right)$$

最後，將上述統計量組合以形成更大的矩陣或向量：

$$\mathbf{N}(s) = \begin{bmatrix} N_1(s) \cdot I & & \\ & \ddots & \\ & & N_C(s) \cdot I \end{bmatrix}$$

$$\tilde{\mathbf{F}}(s) = \begin{bmatrix} \tilde{F}_1(s) \\ \vdots \\ \tilde{F}_C(s) \end{bmatrix}$$

$$\tilde{\mathbf{S}}(s) = \begin{bmatrix} \tilde{S}_1(s) & & \\ & \ddots & \\ & & \tilde{S}_C(s) \end{bmatrix}$$

其中 $\mathbf{N}(s)$ 為一 $CF \times CF$ 之矩陣，其內部由 $N_c(s) \cdot I$ 依主對角線方線串聯而成；$\tilde{\mathbf{F}}(s)$ 為一 $CF \times 1$ 之向量，其內部由 $\tilde{F}_c(s)$ 垂直串接而成；$\tilde{\mathbf{S}}(s)$ 為一 $CF \times CF$ 之矩陣，其內部由 $\tilde{S}_c(s)$ 依主對角線方線串聯而成。到此處即已完成了後續流程所需要用到的數學符號之定義。

總體變異矩陣與通用背景模型相同，亦是以最大似然度估計為原則，使用最大期望演算法進行訓練。在第一步驟中，總體變異矩陣內部之元素值可以隨機的方式進行初始化；而 E 步驟之主要目的為估計潛在變數，也就是 i-向量的期望值，此處定義對於一語者 s 與現有模型參數 Λ，其潛在變數 w_s 對於給定可觀變數 W_s 之後驗機率為 $P_\Lambda\left(w_s | W_s\right)$。已知 w_s 符合標準常態分佈，故依據貝氏理，可以得到：

doi:10.6342/NTU201801578

$$P_{\Lambda}\left(w_s \mid W_s\right) \propto P_{\Lambda}\left(W_s \mid w_s\right) \cdot \mathcal{N}\left(w_s \mid \mathbf{0}, I\right) \tag{8}$$

又依據 [6] 之 Theorem 1 之公式推導，可得到：

$$P_{\Lambda}\left(W_s \mid w_s\right) = G_{\Sigma}\left(s, m\right) + H_{\Lambda}\left(s, w_s\right) \tag{9}$$

其中，

$$G_{\Sigma}\left(s, m\right) = \sum_{c} N_c(s) \cdot \log\left(\frac{1}{(2\pi)^{F/2} \left|\Sigma_c\right|^{1/2}}\right) - \frac{1}{2} tr\left(\Sigma^{-1} \cdot \tilde{\mathbf{S}}(s)\right)$$

$$H_{\Lambda}\left(s, w_s\right) = w_s^{*} \cdot T^{*} \cdot \Sigma^{-1} \cdot \tilde{\mathbf{F}}(s) - \frac{1}{2} w_s^{*} \cdot T^{*} \cdot \mathbf{N}(s) \cdot \Sigma^{-1} \cdot T \cdot w_s$$

其中 $tr(\cdot)$ 代表取矩陣之跡數（trace），而總體變異矩陣 T 隨機設定其初始值。

接著，結合 (8) 與 (9)，並定義：

$$L(s) = I + T^{*} \cdot \Sigma^{-1} \cdot \mathbf{N}(s) \cdot T$$

$$A(s) = L(s)^{-1} \cdot T^{*} \cdot \Sigma^{-1} \cdot \tilde{\mathbf{F}}(s)$$

則可以得到：

$$P_{\Lambda}\left(w_s \mid W_s\right) \propto \exp\left(w_s^{*} \cdot T^{*} \cdot \Sigma^{-1} \cdot \tilde{\mathbf{F}}(s) - \frac{1}{2} w_s^{*} \cdot L(s) \cdot w_s\right)$$

$$\propto \exp(-\frac{1}{2}(A - w_s)^{*} \cdot L(s) \cdot (A(s) - w_s))$$

由此乃可以推得 $P_{\Lambda}\left(w_s \mid W_s\right)$ 呈現平均為 $A(s)$、共變異數矩陣為 $L(s)^{-1}$ 之高斯分佈，而最初欲求之 i-向量之估計期望值即為 $A(s)$。

M 步驟之目的為更新模型，也就是總體變異矩陣之參數 T。首先這裡一樣先定義一些數學符號：

$$\mathfrak{A}_c = \sum_{s} N_c(s) \cdot E[w_s \cdot w_s^{*}]$$

doi:10.6342/NTU201801578

$$\mathfrak{C} = \sum_s \tilde{\mathbf{F}}(s) \cdot E(w_s^*)$$

$$\mathbf{N} = \sum_s \mathbf{N}(s)$$

其中，$E[\cdot]$ 代表求取期望值。又由 E 步驟所得到的結論，便可以進一步推得：

$$\mathfrak{A}_c = \sum_s N_c(s) \cdot L(s)^{-1}$$

$$\mathfrak{C} = \sum_s \tilde{\mathbf{F}}(s) \cdot A^*(s)$$

假設高斯分量 $c = 1,2,...,C$，MFCC 特徵之維度 $f = 1,2,...,F$，而對於每一 $i = (c-1)F + f$，以 T_i 與 $\mathfrak{C}_i$ 分別代表 T 與 $\mathfrak{C}$ 的第 i 列元素，則透過解出以下方程式，便能更新總體變異矩陣之參數 T：

$$T_i \cdot \mathfrak{A}_c = \mathfrak{C}_i$$

另外，超共變異數矩陣 Σ 也可以一併由下列之公式更新：

$$\Sigma = \mathbf{N}^{-1} \cdot \left(\sum_s \tilde{\mathbf{S}}(s) - diag(\mathfrak{C} \cdot T^*) \right)$$

依照上述兩步驟更新模型參數時，每次迭代都會確保訓練資料之對數似然度有所提升，以數學符號可以表示為：

$$\sum_s \log\left(P_\Lambda(W_s)\right) \geq \sum_s \log\left(P_{\Lambda_0}(W_s)\right)$$

其中，

$$\log\left(P_\Lambda(W_s)\right) = G_\Sigma\left(s, m\right) - \frac{1}{2}\log|L(s)| + \frac{1}{2}E[w_s^*] \cdot T^* \cdot \Sigma^{-1}$$

而模型參數 Λ_0 代表模型參數 Λ 更新前的狀態。有關上述所有公式之更詳細數學過程與證明，詳見 [6] 之 Theorem 4 與 Theorem 5。

doi:10.6342/NTU201801578

2.4 i-向量之降維與信道補償措施：線性判別分析

先前 2.3 章提及，總體變異矩陣的訓練過程中，同一語者之不同信道都被視作不同語者，因此擷取出來的 i-向量會仍會受信道效應干擾；而為了消弭信道效應，便必須對 i-向量進行信道補償（channel compensation）。信道補償的方法有非常多種，而最常見的一種就是以下將介紹的線性判別分析（Linear Discriminant Analysis, LDA）[9]。

線性判別分析是一種在統計學、模式識別、機器學習等領域皆經常被使用到的技術，一般是被用來進行特徵的降維，亦可以當成一種線性分類器。它可以藉由將特徵進行線性組合，從而找出一組彼此正交的新方向，而當原特徵被投影到由這些方向為基底組成的空間時，分屬不同類別的樣本能在新空間上被有效分離，而同一類別的樣本則會更加群聚，如圖 17 所示：

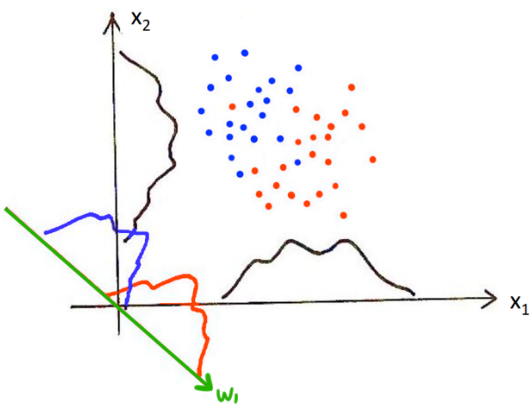


圖 17、線性判別分析示意圖

為方便說明，此處先定義一些符號：假設訓練資料中，共有 S 個類別，而每個類別 s 所包含之樣本點個數為 n_s，而各樣本點則以 $w_{s,h}$ 表示（$h = 1,2,...n_s$），則此處以 $\bar{w}_s$ 代表類別 s 內樣本點之平均、$\bar{w}$ 代表所有樣本點的平均，即：

$$\bar{w}_s = \frac{1}{n_s}\sum_{h=1}^{n_s} w_{s,h}$$

doi:10.6342/NTU201801578

$$\bar{w} = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{n_s}\sum_{i=h}^{n_s} w_{s,h}\right)$$

接著定義類別間變異性矩陣（between-class variance matrix）S_b 與類別內變異性矩陣（within-class variance matrix） S_w：

$$S_b = \frac{1}{S}\sum_{s=1}^{S}(\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})^*$$

$$S_w = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{n_s}\sum_{i=h}^{n_s}(w_{s,h} - \bar{w}_s)(w_{s,h} - \bar{w}_s)^*\right)$$

我們可以推想，若不同類別間的樣本點在空間上彼此越分散，S_b 內的元素會傾向有越大的值；若而同一類別內的樣本點在空間上彼此越分散，S_w 內的元素則會傾向有越大的值。而線性判別分析的原理，是要透過將特徵投影到新空間上，使其類別間變異性與類別內變異性能分別達到最大化與最小化，若以 A 代表用來將特徵投影到新空間上的矩陣，則上述目標可以轉化成以下數學式：

$$A = \underset{W}{argmax}\frac{tr(W^* \cdot S_b \cdot W)}{tr(W^* \cdot S_w \cdot W)}$$

其中 A 為一 $F \times k$ 之矩陣，其中 F 代表原始特徵之維度，k 代表特徵降維後的維度。若將對應到矩陣 $S_w^{-1} \cdot S_b$ 第 i 大的本徵值（eigenvalue）之本徵向量（eigenvector）表示為 a_i，則將 a_1 至 a_k 併排所組成的矩陣，即為 A 之解析解：

$$A = [a_1, a_2, \ldots, a_k]$$

而任意樣本點 w 透過線性判別分析投影到新空間上的位置即為 $A^* \cdot w$ 。

doi:10.6342/NTU201801578

2.5 語者評分模型：機率性線性判別分析

為語者抽出 i-向量特徵後，最後一步即是為其算出分數，方能完成最後的判別。欲判斷兩個 i-向量是否屬於同一語者，最簡單的評分方式莫過於直接取兩者之標準化內積作為分數，即：

$$score_{CS}(w_1, w_2) = \frac{w_1 \cdot w_2}{|w_1||w_2|}$$

其中 w_1 與 w_2 為欲進行評分之兩 i-向量，此方法又稱為餘弦相似度（cosine similarity）；另一個效果較佳的方法是支持向量機（Support Vector Machine, SVM），超向量、i-向量或是其他因子分析之語者特徵搭配支持向量機所構成的語者分類器，在早些時期曾是最常被使用的的架構；而現今以 i-向量為基礎的語者辨識系統中，效果最佳、最主流的評分器，則是以下將要介紹的機率性線性判別分析（Probabilistic Linear Discriminant Analysis, PLDA） [10] [11] [12]。

2.5.1 機率性線性判別分析之模型訓練流程

機率性線性判別分析最早被應用於人臉辨識工作，而後來被發現其配合 i-向量作為評分器也能達到良好的辨識效果。用於其基礎概念與聯合因子分析類似，其假設一 i-向量可被分解成「語者、信道不相關項」、「語者相關項」、「信道相關項」與「殘留項」，數學式表示為：

$$w_{s,h} = \mu + \Phi \cdot \beta_s + \Gamma \cdot \alpha_{s,h} + \epsilon_{s,h} \tag{10}$$

其中 $w_{s,h}$ 為與語者 s、信道 h 相關之 i-向量，大小為 $R_t \times 1$；μ 為訓練資料中 i-向量之總體平均值，Φ 為一低秩之矩陣，與聯合因子分析中之本徵語音扮演相同角色，大小為 $R_t \times R_F$，而 R_F 將決定語者潛在變數之維度；β_s 為語者潛在變數之向量，大小為 $R_F \times 1$，呈現標準常態分佈，其每一項元素又稱為語者因子；Γ 為一低秩之矩陣，與聯合因子分析中之本徵信道扮演相同角色，大小為 $R_t \times R_G$，

doi:10.6342/NTU201801578

而 R_G 將決定信道潛在變數之維度；$\alpha_{s,h}$ 為信道潛在變數之向量，大小為 $R_G \times 1$，呈現標準常態分佈，其每一項元素又稱為信道因子；$\epsilon_{s,h}$ 則為殘留雜訊之向量，大小為 $R_t \times 1$，呈現平均為零向量、共變異數矩陣為 σ 的高斯分佈，其中 σ 為一對角矩陣。

機率性線性判別的模型參數與通用背景模型、聯合因子分析類似，亦是使用最大期望值演算法進行訓練，以下將對其進行簡介。首先一樣先定義數學符號，這邊先將兩項模型參數與兩項潛在變數各自合併為一項：

$$B = [\Phi \quad \Gamma]$$

$$z'_{s,h} = \begin{bmatrix} \beta_s \\ \alpha_{s,h} \end{bmatrix}$$

因此 (10) 可被改寫為：

$$w_{s,h} = \mu + B \cdot z'_{s,h} + \epsilon_{s,h}$$

在第一步驟中，兩項模型參數可以隨機的方式進行初始化；而透過 E 步驟，可以算出潛在變數 z'_s 之期望值與共變異數矩陣，即：

$$E[z'_{s,h}] = (B^* \cdot \sigma^{-1} \cdot B + I)^{-1} \cdot B^* \cdot \sigma^{-1} \cdot (w_{s,h} - \mu)$$

$$E[z'_{s,h} \cdot y'^*_{s,h}] = (B^* \cdot \sigma^{-1} \cdot B^* + I)^{-1} + E[z'_{s,h}] \cdot E[y'_{s,h}]^*$$

而透過 M 步驟，則可以計算出模型參數與殘留雜訊之共變異數矩陣：

$$B = \left(\sum_{s,h} (w_{s,h} - \mu) \cdot E[z'_{s,h}]^* \right) \cdot \left(\sum_{s,h} E[z_{s,h} \cdot z^*_{s,h}] \right)^{-1}$$

$$\sigma = \frac{1}{\sum_s n_s} \sum_{s,h} diag\left((w_{s,h} - \mu) \cdot (w_{s,h} - \mu)^* - B \cdot E[z'_{s,h}](w_{s,h} - \mu)^* \right)$$

其中 n_s 為語者 s 所含有之信道數量。以上述流程輪流執行 E 步驟與 M 步驟直到收斂，即可完成機率性線性判別之模型參數的訓練。關於更詳細的訓練過程數學

doi:10.6342/NTU201801578

推導，可參見 [10]。另外，由於此模型實際在進行評分時不會用到信道相關項，因此若將殘留雜訊之共變異數矩陣 σ 假設為全矩陣，則可以在不影響辨識結果的前提下將信道相關項直接省略，而原來之 (10) 式即變為：

$$w = \mu + \Phi \cdot \beta + \epsilon$$

在前一段落中，語者潛在變數 β_s 與殘留雜訊之共變異數矩陣 σ 被假設為符合高斯分佈，因此此模型又稱為高斯—機率性線性判別分析（Gaussian Probabilistic Linear Discriminant Analysis, G-PLDA）；在前些時期，部分研究顯示若以重尾分佈（Heavy-tailed distribution）來模擬 β_s 與 σ 之機率分佈，將能達到更好的辨識效果，而此模型又稱為重尾—機率性線性判別分析（Heavy-Tailed Gaussian Probabilistic Linear Discriminant Analysis, HT-PLDA）；但隨後，其他研究則又發現：若將 i-向量經過徑向高斯化（radial Gaussianization）處理，高斯—機率性線性判別分析即可與重尾—機率性線性判別分析達到相同的效果，又因為前者之運算速率較高，因此其乃成為現今最主流的評分模型。徑向高斯化分為兩個步驟，第一步是將 i-向量進行線性白化轉換（linear whitening transformation ），由於重尾分佈屬於橢圓對稱密度（Elliptically Symmetric Densities, ESD）家族，而高斯分佈屬於球形對稱密度（Spherically Symmetric Density, SSD）家族，因此透過線性白化轉換即可將橢圓對稱密度轉換為球形對稱密度；第二步是則是將其長度分佈進行非線性柱狀圖變形（non-linear histogram warping），其過程是將 i-向量之長度（$|w|$）之累積分佈通過反向累積卡方分佈（inverse cumulative Chi distribution）函數，而此步驟之簡化版的則是直接將各 i-向量之長度統一標準化為 1 即可。關於徑向高斯化之更詳細原理可參見 [13]。

doi:10.6342/NTU201801578

2.5.2 機率性線性判別分析之評分流程

在為兩個 i-向量 w_1 與 w_2 抽出語者潛在變數後，機率性線性判別分析進行評分的原理是以「兩個潛在變數來自同一 i-向量之機率」與「兩個潛在變數來自不同 i-向量之機率」之比值來進行評分，其數學式表示為：

$$score_{PLDA}(w_1, w_2) = \frac{p(w_1, w_2 \mid \mathscr{H}_s)}{p(w_1 \mid \mathscr{H}_d) \cdot p(w_2 \mid \mathscr{H}_d)} \tag{11}$$

其中 $\mathscr{H}_s$ 代表兩個 i-向量會產生相同潛在變數之假說，$\mathscr{H}_d$ 則代表兩個 i-向量會產生不同潛在變數之假說。而在 i-向量符合常態分佈之前提下，則可將方才之分數取對數而改寫為：

$$score_{PLDA}(w_1, w_2) = \log \mathscr{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_{tot} & \sigma_{ac} \\ \sigma_{ac} & \sigma_{tot} \end{bmatrix}\right) - \log \mathscr{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_{tot} & \mathbf{0} \\ \mathbf{0} & \sigma_{tot} \end{bmatrix}\right)$$

其中，

$$\sigma_{tot} = \Phi \cdot \Phi^* + \sigma$$

$$\sigma_{tot} = \Phi \cdot \Phi^*$$

又若假設 μ 為零向量，且令：

$$Q = \sigma_{tot}^{-1} - (\sigma_{tot} - \sigma_{ac} \cdot \sigma_{tot}^{-1} \cdot \sigma_{ac})^{-1}$$

$$P = \sigma_{tot}^{-1} \cdot \sigma_{ac} \cdot (\sigma_{tot} - \sigma_{ac} \cdot \sigma_{tot}^{-1} \cdot \sigma_{ac})^{-1}$$

則可以將 (11) 式展開為：

$$score_{PLDA}(w_1, w_2) = w_1^* \cdot Q \cdot w_1 + w_2^* \cdot Q \cdot w_2 + 2w_1^* \cdot P \cdot w_2 + const \tag{12}$$

假設 Φ 為一大小為 $D \times K$ 之矩陣（$D > K$），P 便可基於其對稱性用奇異值分解（Singular Value Decomposition, SVD）分解為三個矩陣之乘積，即：

doi:10.6342/NTU201801578

$$P = U_K \cdot \tilde{P} \cdot U_K^*$$

其中 U_K 是由 K 行互相正交的向量所組成，這些向量乃是 P 的本徵向量；而 $\tilde{P}$ 則是由 P 的本徵值由主對角線方向串接而成的對角矩陣，即：

$$\tilde{P} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix}$$

其中 λ_k 乃是矩陣 P 的第 k 大本徵值。又因為 P 與 Q 之秩數與 Φ 相同，(12) 式可以用更快的方法計算而得：

$$score_{PLDA}(w_1, w_2) = \tilde{w}_1^* \cdot \tilde{Q} \cdot \tilde{w}_1 + \tilde{w}_2^* \cdot \tilde{Q} \cdot \tilde{w}_2 + 2\tilde{w}_1^* \cdot \tilde{P} \cdot \tilde{w}_2 + const \tag{13}$$

其中，

$$\tilde{Q} = U_K^* \cdot Q \cdot U_K$$

$$\tilde{w}_i = U_K^* \cdot w_i,\ i = 1,2$$

並且，當 K/D 之值越小，(13) 式便可較 (12) 式擁有更快的運算速率。關於機率性線性判別分析之評分方式的更詳細原理，可參見 [12]。

doi:10.6342/NTU201801578

第 3 章 實驗結果

3.1 實驗語料與資料配置

本論文使用了 NIST SRE 2010 與 TIMIT 兩份語料庫分別進行實驗，以下將簡介此二語料庫與資料配置。

3.1.1 NIST SRE 2010 語料庫與資料配置簡介

NIST SRE 2010 [14] 是一場在 2010 年由美國國家標準技術研究所（National Institute of Standards and Technology, NIST）所舉辦的語者辨識（Speaker Recognition Evaluation, SRE）比賽，本論文取其訓練資料中的 3 種資料類型：「10sec」、「core」與「8conv」，並除去重複的語者後進行實驗，其中「10sec」為長度約 10 秒的電話對話錄音，一共有 2,257 位（男 959／女 1,298）語者，每位語者提供 1 份錄音檔；「core」為長度約 3 至 15 分鐘的電話對話錄音或麥克風錄製的面試錄音，一共有 4,004 位（男 1,817／女 2,187）語者，每位語者提供 1 份錄音檔；「8conv」為長度約 2 至 3 分鐘的電話對話錄音，一共有 445 位（男 194／女 251）語者，每位語者提供 8 份錄音檔。3 種資料類型之音檔取樣頻率為 8,000 赫茲、樣本位元數皆為 16。

資料配置方面，訓練資料是以 3 種資料類型各取 2／3 語者數量，一共 4,469 位（男 1,965／女 2,504）語者的集合；測試資料則是剩下的 1／3 語者數量，一共 2,237 位（男 1,005／女 1,232）語者的集合。而關於測試資料中註冊語料與驗證語料的分配，每位語者皆僅會註冊 1 組語者模型，其中「8conv」的每組模型皆以 7 份音檔註冊而成，剩下 1 份音檔作為驗證音檔；而由於「10sec」與「core」的每位語者皆只有 1 份音檔，因此這裡將每份音檔分割為 4 等份，每

doi:10.6342/NTU201801578

組模型皆以 3 份音檔註冊而成，剩下 1 份音檔作為驗證音檔。測試資料的每一位語者皆會自驗證語料中挑選 1,000 份音檔來測試自己的語者模型，1,000 份音檔中其中 1 份是會是正樣本（屬於此語者的聲音），而其他 999 份則是隨機挑選的負樣本（不屬於此語者的聲音），因此所有欲評測的樣本一共有
$2{,}237 \times 1{,}000 = 2{,}237{,}000$ 份。

3.1.2 TIMIT 語料庫與資料配置簡介

TIMIT [15] 是由德州儀器（Texas Instruments, TI）、麻省理工學院（Massachusetts Institute of Technology, MIT）與斯坦福國際研究院（SRI International）聯合提供的語料庫，其一共包含了 630 位（男 438／女 192）語者。這些語者的腔調包含了美國 8 種主要地方方言腔調，每位語者皆提供 10 份錄音檔，每份錄音檔之取樣頻率為 16,000 赫茲、樣本位元數為 16，錄音內容則為指定內容、包含豐富音素（phoneme）的一句話，每句話平均長度約為 3 秒，最長不超過 8 秒。

資料配置方面，訓練資料是以男性與女性各取 2／3 語者數量，一共 420 位（男 292／女 128）語者的集合；測試資料則是剩下的 1／3 語者數量，一共 210 位（男 146／女 64）語者的集合。而關於測試資料中註冊語料與驗證語料的分配，每位語者皆會註冊 2 組語者模型，每組模型皆以 4 份音檔註冊而成，而剩下的 2 份音檔則作為測試音檔。測試資料的每一位語者皆會自驗證語料中挑選 100 份音檔來測試自己的語者模型，100 份音檔中其中 1 份是會是正樣本，而其他 99 份則是隨機挑選的負樣本，因此所有欲評測的樣本一共有
$210 \times 2 \times 100 = 42{,}000$ 份。

doi:10.6342/NTU201801578

3.2 實驗一：對照組

實驗一將介紹本論文設定的對照組流程與實驗結果，此實驗為完全不使用語者性別資訊的傳統語者驗證方法，亦為本論文之後所有實驗之比較依據。本實驗之流程架構如圖 18 所示：

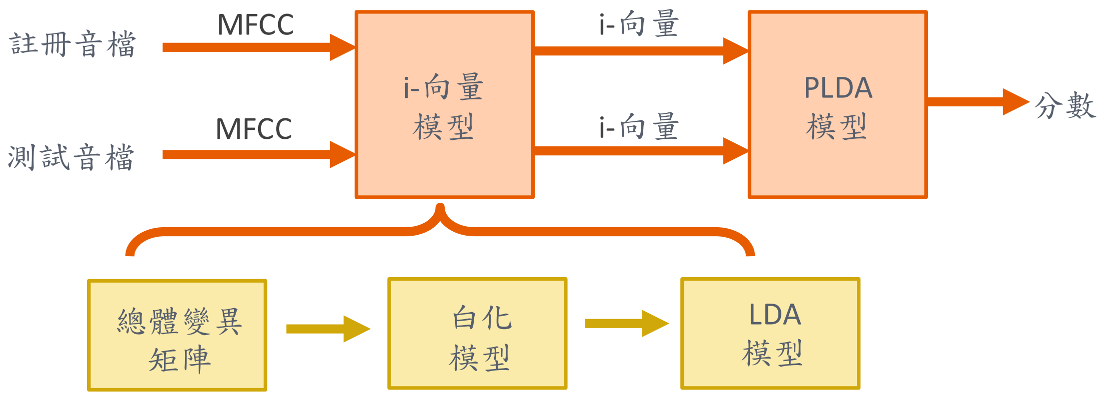


圖 18、對照組之流程架構

流程圖中之各項模型以 Bob toolbox [16] 進行實作。由圖中可見，註冊音檔與測試音檔分別抽取 MFCC 特徵後，會經由 i-向量模型各自抽取 i-向量，最後經由 PLDA 模型算出分數；其中，此處所述之 i-向量模型裡除了包含原本用以將超向量轉化為 i-向量的總體變異矩陣之外，亦包涵了白化轉換模型與線性判別分析模型。表 1 為對照組流程所使用的模型參數，而本論文其他語者驗證相關流程之模型參數設定亦與對照組相同：

表 1、語者驗證系統模型參數設

Model Parameter		Value
MFCC extraction	Frame size	512 sample points
	Hop size	128 sample points
UBM model	Number of mixture	128
	Number of training iteration	10
i-vector model	Dimension of i-vector	300
	Number of training iteration	5
LDA model	Reduced Dimension of i-vector	50
PLDA model	Dimension of speaker factor	50
	Dimension of channel factor	50
	Number of training iteration	10

doi:10.6342/NTU201801578

對照組流程的實驗結果如圖 19 所示：

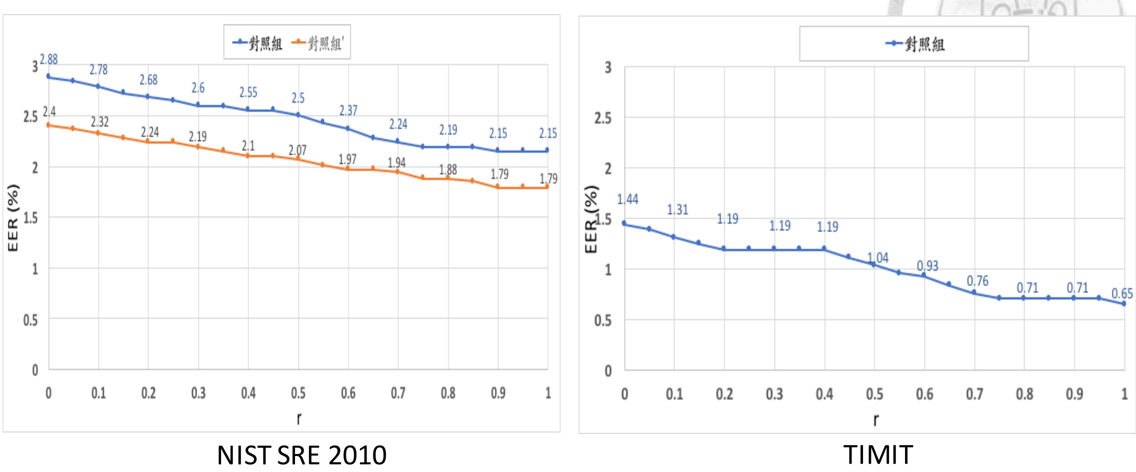


圖 19、對照組流程之結果

圖之左側與右側分別為 NIST SRE 2010 語料庫與 TIMIT 語料庫之實驗結果，兩張圖的縱軸為語者驗證系統之等錯誤率，單位為百分比；橫軸為 r，此 r 值代表的是「與註冊語者不同性別的仿冒者比例」，以下將簡單說明此實驗設計之意義。

在章節 3.2 中提及，本論文之資料配置形式為每一位註冊語者配對 1 份自己的音檔作為正樣本，並搭配許多其他仿冒者的聲音作為負樣本，而挑選仿冒者的一項重要考量，便是這些語者的性別。今假設語者驗證系統架設於使用者之行動裝置上，則可以考慮兩種情景，如圖 20 所示：

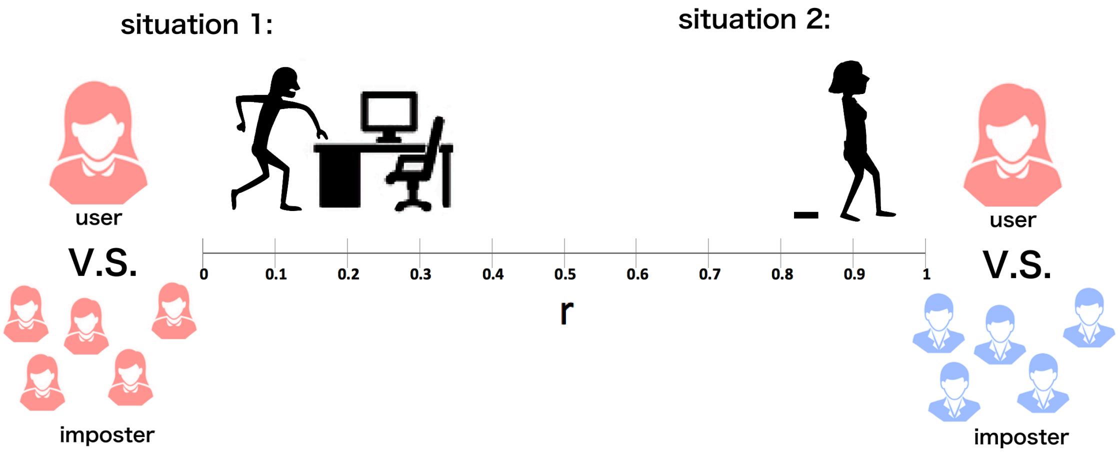


圖 20、不同的 r 值與系統應用情境

doi:10.6342/NTU201801578

第一，當使用者到了教室或是工作場所時，若有同事或同學意圖在未經該使用者同意下存取其行動裝置，由於他們皆熟悉該使用者的聲音，因此勢必會找聲音與該使用者相像的人—至少會與該使用者同性別—來解鎖其語者驗證系統，在此情境下，我們可以假設絕大部分的仿冒者皆會和使用者同性別（事實上，NIST SRE 2010 比賽亦是設計成每個樣本的註冊方與測試方的性別皆相同，因此與此情境相同）。第二，當使用者的行動裝置（被陌生人）偷竊或不慎遺失時，由於拿到其行動裝置的人並不知曉該使用者的性別或聲音特性，因此此種情境下便不能假設仿冒者會和該使用者之性別相同，仿冒者和該使用者的性別相同的機率在此時只有 1／2，甚至更低。綜合以上兩種情境，我們便必須確保語者驗證系統在不同的 r 值，也就是不同的「仿冒者與使用者不同性別的機率」的情況下，皆應有良好的表現。因此，本論文所有的語者驗證相關實驗，皆會在不同的 r 值之下測量系統的表現。

在圖 19 左圖中有兩條線，兩者之差別在於音幀大小的不同。藍線是在音幀的樣本點數上與 TIMIT 語料統一，即一個音幀中含有 512 個樣本點，時間長度為 64 ms；橘線則是在音幀的時間長度上與 TIMIT 語料統一，即一個音幀中含有 256 個樣本點，時間長度為 32 ms。由圖中可見，由於橘線的總音幀數會較多，因此各模型之訓練效果較好、實驗結果亦較佳，但由於音幀數較多導致訓練過程曠日費時，因此往後的實驗皆使用與藍線相同的設定。

由圖 19 中可見，雖然對照組流程於 TIMIT 語料與 NIST SRE 2010 語料上的表現有明顯差距，但兩者的變化趨勢是相同的：從最左側的 r 等於 0 慢慢往最右側的 r 等於 1 移動時，系統的等錯誤率皆呈現近似線性下降的趨勢。造成此下降趨勢之原因其實非常直觀：與使用者不同性別的仿冒者，相較於同性別者，大多會與該使用者在聲紋特性上有較大的差別，造成其被系統拒絕的機率也相對高出許多，因此當 r 值越大，正樣本所得到的分數與其他負樣本相比便會相對高出許多，進而得到較低的等錯誤率。

doi:10.6342/NTU201801578

3.3 實驗二：性別資訊應用方法—當性別分類器表現良好

實驗二將介紹若干種將性別資訊應用於語者驗證系統以提升其辨識率之方法，包括 [17] 所介紹的一般最常見作法，以及本論文提出的另外 2 種針對前者弱點所進行之改良方法。另外，本實驗之目的在於證明在性別分類器表現良好之前提下，藉由有效運用性別資訊，便能提升系統辨識率；因此本實驗中對於所有原本應假定為未知性別的測試語者，皆以其真實性別代替性別分類器的預測結果。

3.3.1 前人作法（一般常見作法）

本方法是參考了論文 [17] 而實作的流程，其流程架構如圖 21 所示：

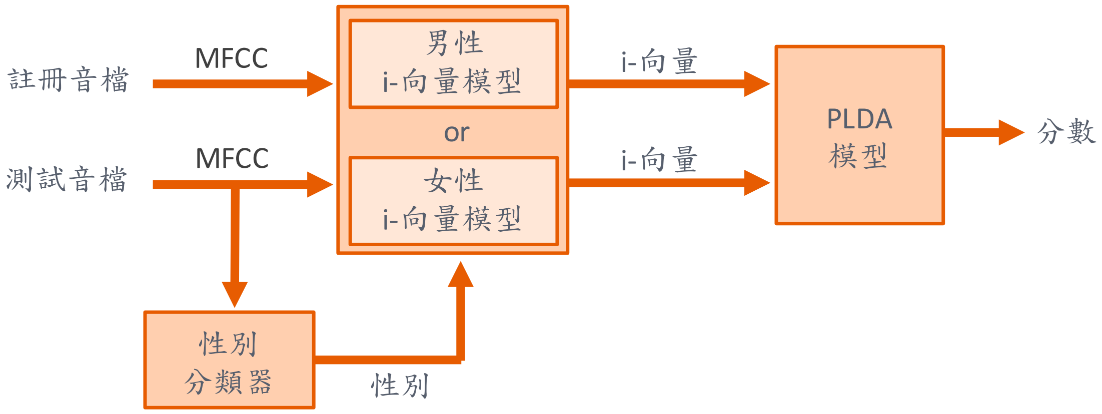
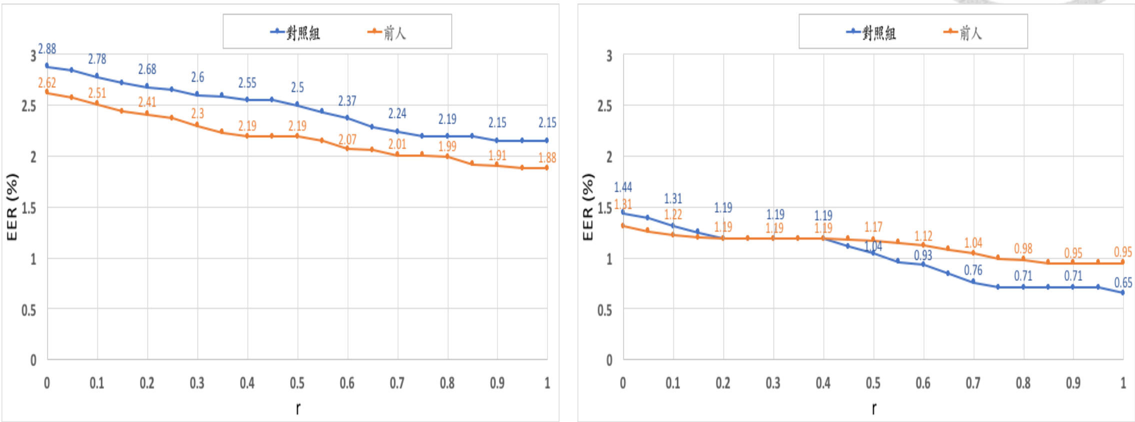


圖 21、前人作法之流程架構圖

其流程基本上與對照組流程類似，兩者不一樣的地方為：對照組之 i-向量模型是使用訓練資料中所有語者共同訓練而得來，而本作法之 i-向量模型則分為 2 個性別相關的模型，分別是使用訓練語者中的男性語者與女性語者進行訓練而得來，而所有註冊音檔或測試音檔，皆必須使用對應其語者性別之模型來抽取 i-向量。其中，註冊語者的性別可以假設是已知的，因為在實際運用情境中，我們可以要求所有使用者要註冊系統時提供自己的性別資訊；但是由於測試語者的性別是未知的，因此測試音檔必須先藉由性別分類器預測其性別，才能套入正確性別的 i-向量模型來抽取 i-向量。

doi:10.6342/NTU201801578

本做法的實驗結果如圖 22 所示：



NIST SRE 2010　　　　TIMIT

圖 22、前人作法之結果

同樣地，圖之左側與右側分別為 NIST SRE 2010 語料庫與 TIMIT 語料庫之實驗結果，兩張圖的藍線為方才對照組流程的結果，橘線則是本做法的結果。由圖中可以看到，在 NIST SRE 2010 語料中，本做法的結果從 r = 0 至 r = 1 始終都較對照組作法要進步，證明了其實用性；但反觀 TIMIT 語料，其實驗結果在最左側 r = 0 時尚較對照組進步一些，但慢慢往右側 r = 1 方向移動時，由於其等錯誤率的下降速率較對照組慢上許多，導致其結果逐漸被對照組超越，甚至被拉開距離。

由於許多語者驗證比賽的樣本（如 NIST SRE 2010）並不包含註冊語者與測試語者性別不同的情形，因此許多以本作法實作的語者驗證系統（如 [17]），便沒能對此現象做進一部探討或進行補償措施。而本研究推測，造成上述現象的主因，是來自於本做法的一個缺陷：使用性別相關模型抽取的男性 i-向量與女性 i-向量，在空間分佈上可能存在重疊現象。圖 23 為 i-向量之理想二維點分佈圖：

doi:10.6342/NTU201801578

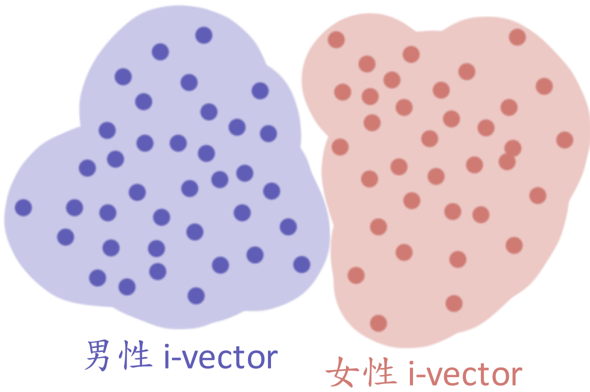


圖 23、i-向量之理想二維點分佈圖

在理想上，因為男性與女性的聲紋特性存在很大的差異，容易讓人以為 i-向量會在空間分佈上呈現男女各自群聚，且容易界定出兩群 i-向量之邊界的情形，正如同圖 23 所示。在此假設下，當一個樣本的註冊方與測試方擁有不同性別，因為兩者的 i-向量之間的距離很大，因此這個樣本便相當容易被系統拒絕。但實際情況中卻並非一定如此，尤其就男／女性 i-向量模型在訓練過程中並沒有接觸過女／男性的資料這個事實來看，我們更必須懷疑此假設之合理性。圖 24 左右兩圖是分別使用主成分分析（Principal Components Analysis, PCA）和 t-分佈隨機相鄰嵌入（T-distributed Stochastic Neighbor Embedding, TSNE），將訓練資料中的 i-向量降至二維平面所繪製的點分佈圖：

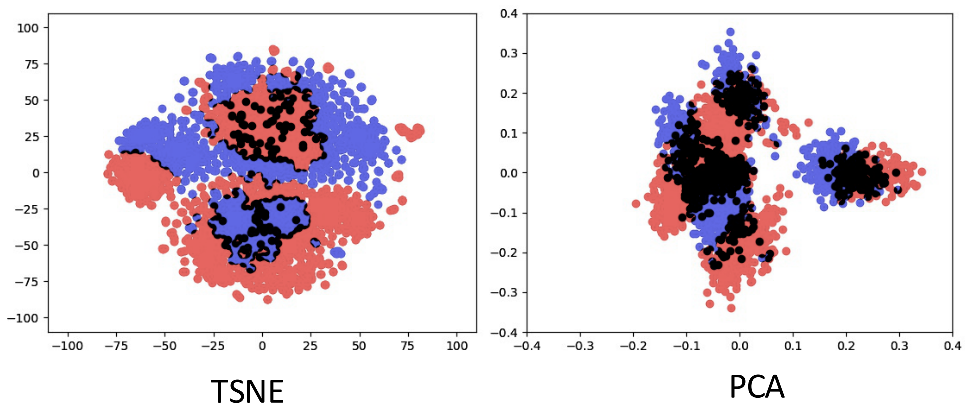


圖 24、i-向量之實際二維點分佈圖－以 TSNE 與 PCA 將 i-向量降為二維

doi:10.6342/NTU201801578

由兩張圖我們能清楚看到，兩個性別的 i-向量群在空間分佈上確實存在重疊的情形。這種現象會造成語者驗證系統的錯誤率上升，尤其在當欲測試的樣本中含有大量註冊方與測試方不同性別的情形，即 r 有較大的值時更為顯著。為方便討論，此處假設系統之評分器僅以註冊方與測試方 i-向量之間的歐氏距離作為判斷是否為同一語者之唯一標準。在不分男、女資料所訓練出來的混合 i-向量模型，在訓練資料足夠的前提下，我們可以假定其會盡量將不同語者的 i-向量在空間上彼此分開；但是，正如方才所述，由於男／女性 i-向量模型在訓練過程中並沒有接觸過女／男性的資料，因此便很有可能產生特定男性語者與女性語者的 i-向量距離過度相近的現象，使得系統容易將他們誤判為同一語者。此現象在訓練流程中，會使得系統的評分器—機率線性判別分析模型容易被混淆；在測試流程中，也容易會造成不同性別間語者被錯誤判斷成同一語者的比例較對照組流程還要高的現象。因此，針對因不同性別 i-向量空間重疊現象所造成的上述缺陷，本論文接下來將提出兩種簡單的改良機制，以令結果獲得改善。

3.3.2 方法一：拒絕與註冊者性別不符之測試者

第一種改良方法之流程架構如圖 25 所示：

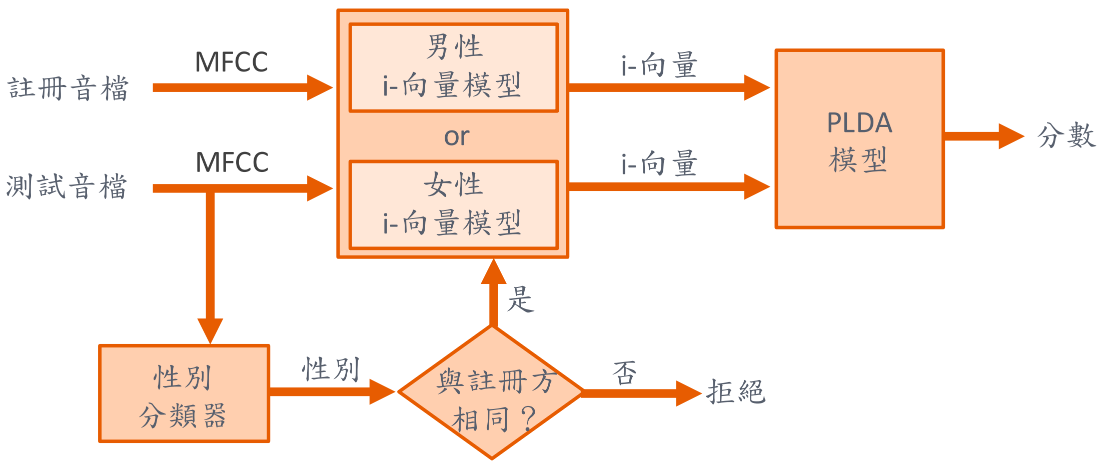
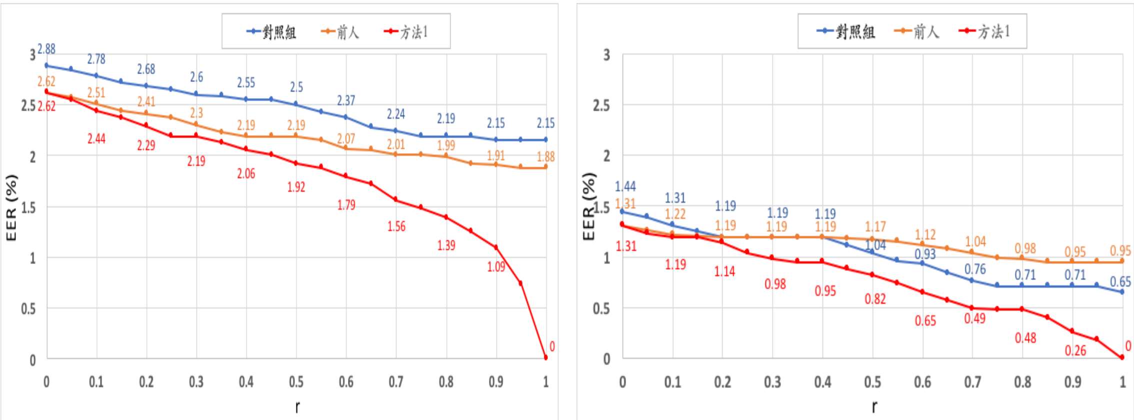


圖 25、方法一之流程架構圖

doi:10.6342/NTU201801578

此版本之流程與前人作法類似，不同之處在於其多出了一個步驟：在以性別分類器判斷出測試語者之性別後，須檢視其是否與註冊語者之性別相同，若相同，則繼續進行原本的流程—以性別相關模型抽出雙方的 i-向量之後，以機率性線性判別分析模型計算出分數；若不同，則直接拒絕此測試語者（打上非常低的分數），因為不同性別的聲音不可能屬於同一位語者。如此一來，便能消除在前人作法中因不同性別的 i-向量在空間分佈上的重疊現象而造成的錯誤接受率攀升之問題。

方法一流程的實驗結果如圖 26 所示：



NIST SRE 2010　　　　TIMIT

圖 26、方法一流程之結果

同樣地，藍線與橘線分別為對照組流程與前人作法流程之實驗結果，而新增的紅線則是本方法一流程之實驗結果。在圖中，我們可以看到在最左側 r = 0 的時候，因為沒有註冊語者與測試語者性別不同的樣本，因此方法一與前人做法的表現是相同的；但慢慢往右側 r = 1 方向移動時，由於註冊方與測試方性別不同的樣本皆被系統拒絕，因此方法一在兩種語料的等錯誤率始終低於對照組，甚至越往右便越與對照組的結果拉開距離；到最後 r = 1 時，由於所有錯誤樣本皆成功被系統拒絕，等錯誤率便可達到 0%。此實驗結果除證明了前人做法確實存在著本論文所預測的缺陷與造成它的原因，亦證明了方法一流程在改善此種缺陷上的效果。但是，方才提及前人做法的缺陷來自於男女 i-向量的重疊現象，而此方法實際上並沒有在 i-向量之空間分布上將分離男性語者與女性語者，因此仍可能會

doi:10.6342/NTU201801578

使機率性線性判別分析模型在訓練過程中被混淆，造成訓練成效下降，可說是治標不治本的改善方法。

3.3.3 方法二：將性別作為新的一維特徵

為了從根本上解決前人作法之缺陷，本論文提出了第二種改良方法，其流程架構如圖 27 所示：

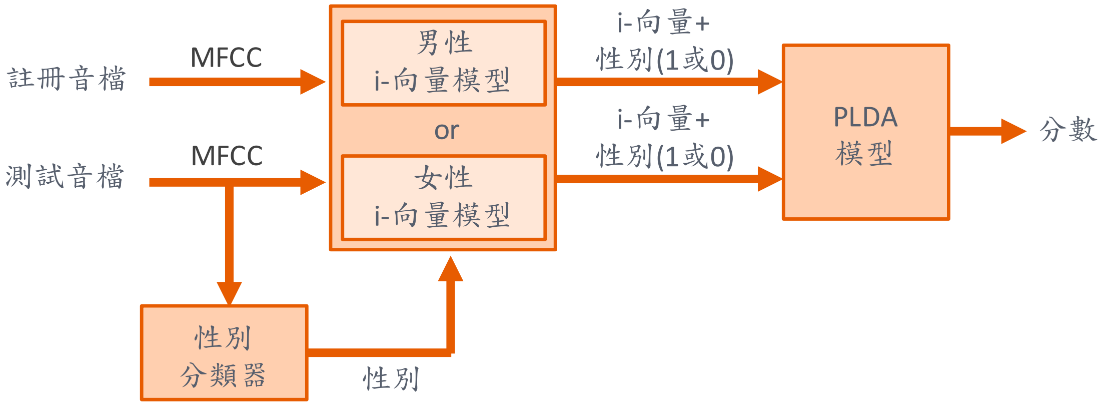
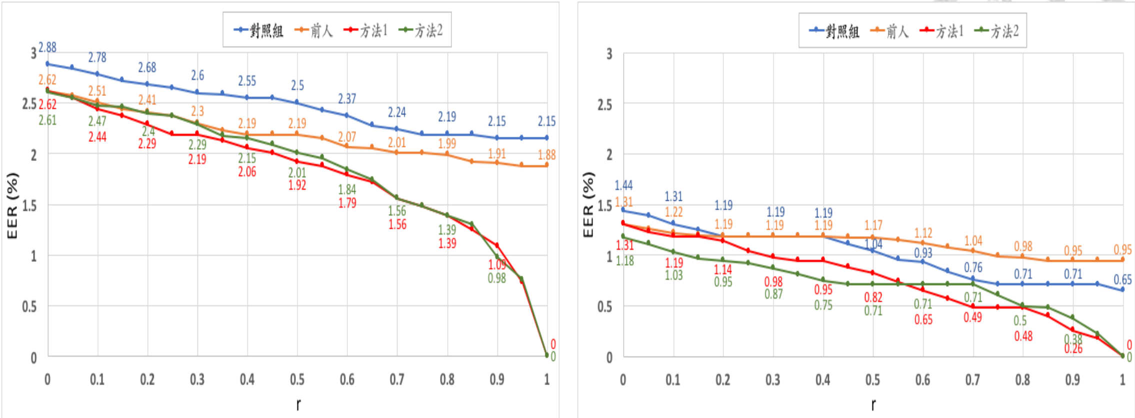


圖 27、方法二之流程架構圖

此版本之流程語前人作法亦十分相似，不同之處在於其在 i-向量上多加了新的一維特徵，使得 i-向量由原本的 50 維變為 51 維。此新一維特徵之目的是用來代表性別資訊，實際表示方法為：若此 i-向量的來源為男性，則此維特徵之值為 1，反之若為女性則為 0。由於新增的此維特徵，男性和女性的 i-向量在空間分佈上便完全沒有互相重疊的現象，因此無論是前人做法之缺陷造成的錯誤接受或錯誤拒絕升高的問題，在此流程下都能同時獲得解決。

方法二流程的實驗結果如圖 28 所示：

doi:10.6342/NTU201801578



NIST SRE 2010　　　　TIMIT

圖 28、方法二流程之結果

同樣地，藍線、橘線與紅線分別代表對照組、前人作法與方法一的結果，而新增的綠線則是本方法二流程之實驗結果。由圖中可以見，方法二流程在兩種語料上相較於方法一的表現並不相同：在 NIST SRE 2010 語料上，方法二在 r 值較小的時候，幾乎都表現得比方法一略差，在 r 值較大時才差不多追平；但在 TIMIT 語料上，方法二在 r 值較小時的表現反倒略優於方法一，但在 r 值較大時卻被方法一超越。此兩種相異的表現，此處推測是資料特性不同或是隨機誤差所致，至於本作法沒能如預期地大幅較方法一進步，此處推測可能是因為訓練資料中距離過於相近之異性別 i-向量畢竟只佔少部分之比例，因此並沒有大幅地影響到機率性線性判別分析模型的訓練成效。不過就結論而言，方法一和方法二在不同 r 值之整體表現上，等錯誤率皆較對照組與前人做法降低許多，因此此兩種流程在「性別分類器表現良好」的前提下作為修正前人作法缺陷的方法，確實皆被證明是可行的。

doi:10.6342/NTU201801578

3.4 實驗三：性別分類器實作

在實驗二中，所有實驗流程皆是基於性別分類器表現良好為前提所設計，因此原本應假定為未知性別的測試語者，皆以其真實性別代替性別分類器的預測結果；而本章節將實際進行性別分類器之實作，並分析各分類器之優劣。本實驗所實作的性別分類器共有 3 種，每種分類器皆是以性別不相關之 i-向量（即章節 3.2 中對照組流程所使用的不分性別模型所產生之 i-向量）作為特徵，以下將各別簡介其原理與結果。

3.4.1 線性判別分析性別分類器

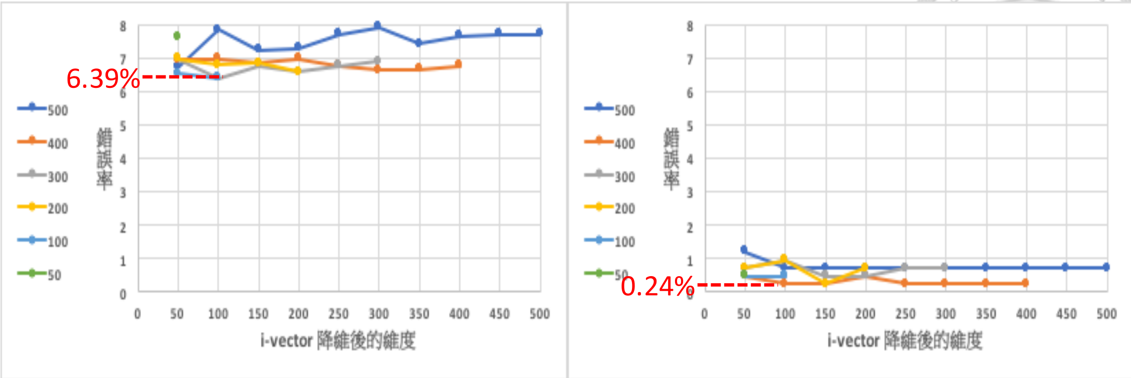
此分類器之作法是參考了論文 [18] 所實作的流程，其流程圖如圖 29 所示：

圖 29、線性判別分析性別分類器之流程架構

由圖中可見，經由 MFCC 特徵轉換而來的 i-向量，會以線性判別分析模型作為二元分類器，判斷音檔的性別；其中此線性判別分析模型的訓練過程，是以性別作為標籤（意即將所有訓練資料分為男、女兩個類別），將 i-向量降至單維作為性別分數，而性別分數之門檻值則定為令訓練資料達到等錯誤率，即男性與女性之錯誤率相同時之分數。需注意的是，本流程所使用的線性判別分析模型，與 i-向量模型本身所包含的線性判別分析模型是不同的模型，兩者之差別在於前者在訓練過程中是以「性別」為標籤，並作為分類器使用；而後者則是以「語者」作為標籤，並作為降維器使用。

本分類器之實驗結果如圖 30 所示：

doi:10.6342/NTU201801578



NIST SRE 2010　　TIMIT

圖 30、線性判別分析性別分類器之結果

同樣地，圖之左側與右側分別為 NIST SRE 2010 與 TIMIT 語料庫之實驗結果，而本分類器之實驗過程調整了包含於 i-向量模型中的兩個參數：i-向量之初始維度與經（i-向量模型中的）線性判別分析模型降維後的維度，兩者分別在圖 30 中以不同顏色的線與圖之橫軸表示，而圖之縱軸則為錯誤率，單位為百分比。圖中的紅色虛線所標記的是兩種語料庫中最佳參數所表現出的結果，分別為 6.39% 與 0.24%。值得一提的是，在 TIMIT 語料庫中，0.24% 的錯誤率代表在其 420 份測試音檔中，只有一份音檔的性別被分類錯誤，其他音檔的性別分類結果則皆是正確的。

3.4.2 支持向量機性別分類器

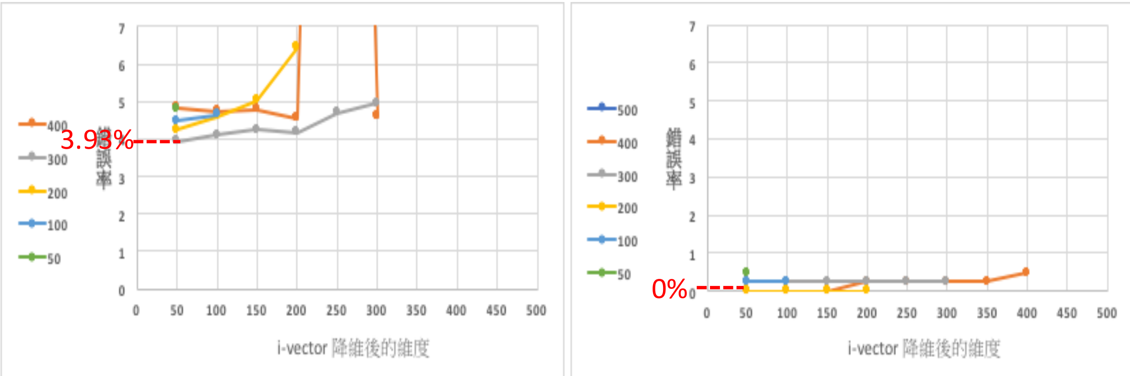
此分類器之流程圖如圖 31 所示：

圖 31、支持向量機性別分類器之流程架構圖

doi:10.6342/NTU201801578

由圖中可見，經由 MFCC 特徵轉換而來的 i-向量，會以支持向量機作為二元分類器，判斷音檔的性別；其中，支持向量機類別權重設定為平衡（balenced class weight），此設定會讓訓練過程中每個樣本的成本（cost）之權重反比於其所屬之類別的樣本總數。

本分類器之實驗結果如圖 32 所示：



NIST SRE 2010　　TIMIT

圖 32、支持向量機性別分類器之結果

本分類器之實驗與前者相同，調整了 i-向量模型中的 i-向量之初始維度與其降維後的維度，兩者分別在圖 32 中以不同顏色的線與圖之橫軸表示；此外，亦調整了支持向量機中最重要的 2 個參數：cost 與 gamma。在圖中，每一個數據點皆是將 cost 與 gamma 分別在集合 {0.1, 1, 10, 100, 1000} 與 {0.01, 0.1, 1, 10, 100} 內進行排列組合，並以網格搜尋法（grid search）所尋找到的最佳參數組合之結果。同樣地，圖中的紅色虛線所標記的是兩種語料庫中最佳參數所表現出的結果，NIST SRE 2010 語料庫的錯誤率可達 3.93%，而 TIMIT 語料庫則可以達到完全分類正確，即 0% 錯誤率的結果。

doi:10.6342/NTU201801578

3.4.3 神經網路性別分類器

此分類器之流程圖如圖 33 所示：

圖 33、神經網路性別分類器之流程架構

由圖中可見，經由 MFCC 特徵轉換而來的 i-向量，會以神經網路模型作為二元分類器，判斷音檔的性別；而神經網路模型的參數（keras API）如表 2 所示：

表 2、神經網路性別分類器之模型參數

模型參數	層參數
batch _size = 128	kernel_initializer='glorot_normal'
epochs = 200	bias_initializer='zeros'
loss = binary_crossentropy	BatchNormalization()
optimizer = Adam(lr=3e-4)	Activation('relu')
metrics = 'accuracy'	Dropout(0.35)

其中，模型的每層節點數量由輸入至輸出端線性減少，舉例來說，若 i-向量維度數為 n，神經網路維中間層數為 4 層，則模型之各層結點數則如圖 34 所示：

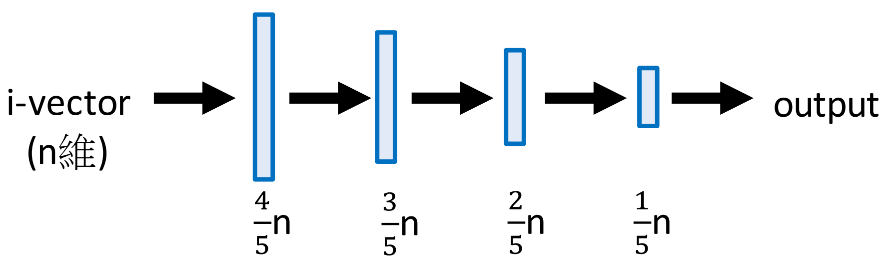


圖 34、神經網路模型之形狀

本分類器之實驗結果如圖 35 所示：

doi:10.6342/NTU201801578

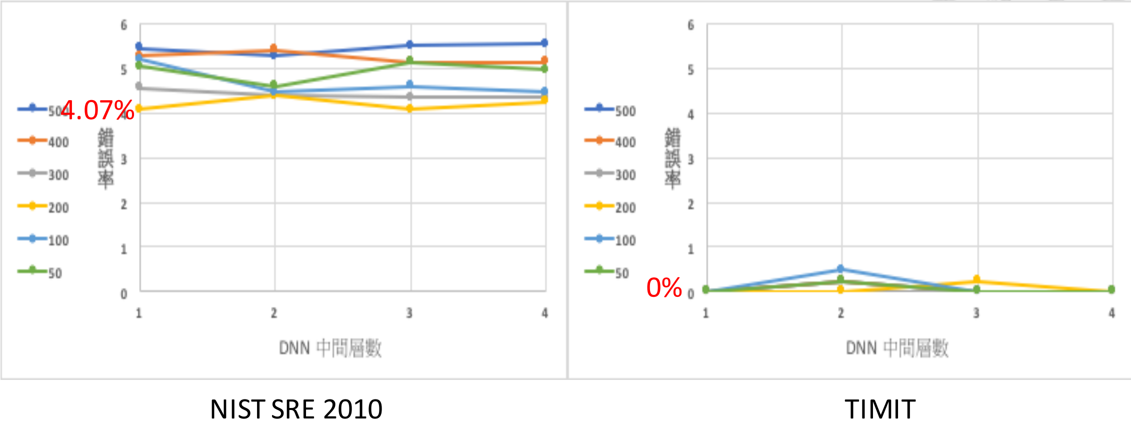


圖 35、神經網路性別分類器之結果

與之前的實驗不同的是，本實驗所使用的 i-向量一律維持初始維度，而不使用線性判別分析模型進行降維，而本分類器之實驗過程調整的兩個參數則改為是 i-向量之初始維度與神經網路模型的中間層數，兩者分別在圖 35 中以不同顏色的線與圖之橫軸表示。同樣地，圖中的紅色虛線所標記的是兩種語料庫中最佳參數所表現出的結果，NIST SRE 2010 語料庫的錯誤率可達 4.07%，而 TIMIT 語料庫則與支持向量機性別分類器一樣，可以達到完全分類正確，即 0% 錯誤率的結果。

doi:10.6342/NTU201801578

3.5 實驗四：性別資訊應用方法 — 當性別分類器表現不良

實驗四同樣將介紹若干種性別資訊應用於語者驗證系統之方法，但不同於實驗二以其真實性別代替性別分類器的預測結果，本實驗將使用在實驗三中實作的性別分類器，並特別針對當性別分類器表現不良的情況，介紹其他不同於實驗二的性別資訊應用方法。

在實驗三中，因 TIMIT 語料庫在性別分類器上可以達到零錯誤率的效果，故其於語者驗證流程中套用實驗三所實作的性別分類器之結果，將與實驗二無異，因此本章節之實驗將僅包含 NIST SRE 2010 語料的部分。圖 36 是對照組流程與將在實驗三中於 NIST SRE 2010 語料中表現最好的支持向量機性別分類器套用於實驗二各流程之實驗結果：

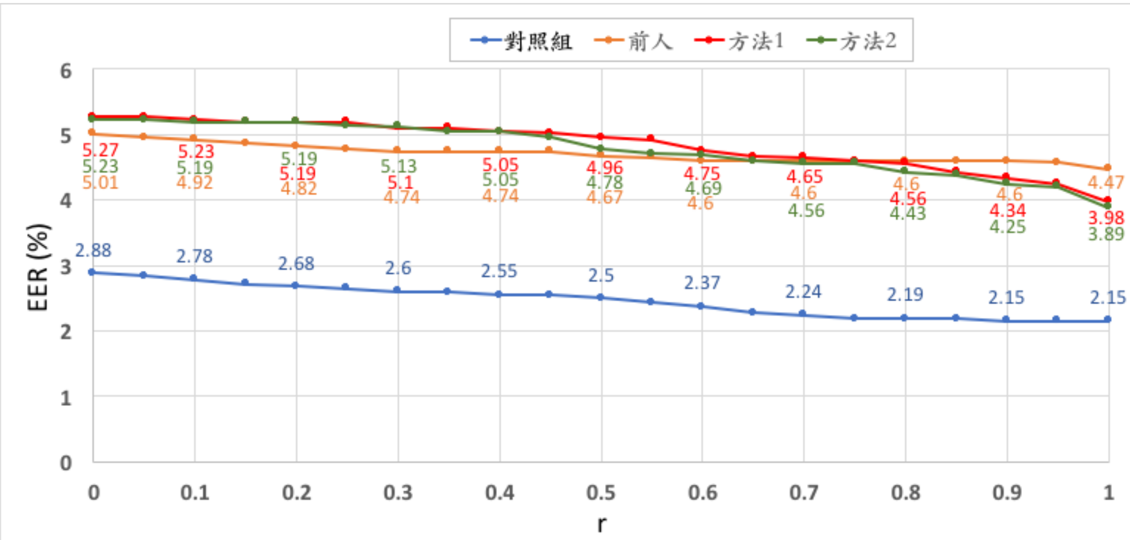


圖 36、各項語者驗證流程使用支持向量機性別分類器之實驗結果

同樣地，藍線、橘線、紅線與綠線分別代表對照組、前人作法、方法一與方法二的結果，由圖中可以見到，在實際套用真實性別分類器之後，實驗二中的所有流程在不同 r 值下的實驗結果產生了明顯的退步，甚至比對照組流程之結果更差。雖支持向量機性別分類器之錯誤率僅有 3.93%，但這些少量的錯誤卻會造成語者驗證系統之表現造成巨大影響，其原因一樣可分為錯誤接受與錯誤拒絕兩種情形來探討：我們可以想像，當註冊語者與測試語者為不同人時，即使性別分類器誤判了測試語者的性別，系統錯誤接受此負樣本的機率仍然不高。但當註冊語者與

doi:10.6342/NTU201801578

測試語者為同一人，一旦測試語者被誤判性別，使用前人作法的系統會分別以不同性別之模型為註冊方與測試方抽取 i-向量，此種情形幾乎必然會造成系統為此正樣本評出較非常低的分數；而使用方法一流程之系統，在檢測出註冊語者與測試語者分屬不同性別後，便會對此樣本予以拒絕；最後，使用方法二流程之系統，由於訓練資料中同一語者的 i-向量中用以代表性別的最後一維特徵皆有相同的數值，因此當評分器見到註冊語者與測試語者在該維特徵之數值相異，自然同樣會評出非常低的分數。綜上所述，實驗二的 3 種方法皆會為大量正樣本評出低落的分數，進而使得等錯誤率跟著大幅提升。

3.5.1 性別分類器的合成

在「使用真實性別」與「使用預測性別」的實驗結果之間所存在的巨大差距，顯示當前所實作的性別分類器在準確度上尚未達到能使用的標準，而欲使實驗二的各項方法真正具備實用性，勢必需要效果更精良的性別分類器。最簡單的提升性別分類器準確度方法之一，便是結合多個性別分類器，共同產生預測結果。此方法能否發揮效果與各性別分類器之間是否具互補性有關，換句話說，當各個性別分類器所誤判的樣本越少重複，此方法越有可能表現出超越單一分類器之效能。圖 37 為各單一分類器誤判樣本的比例分佈：

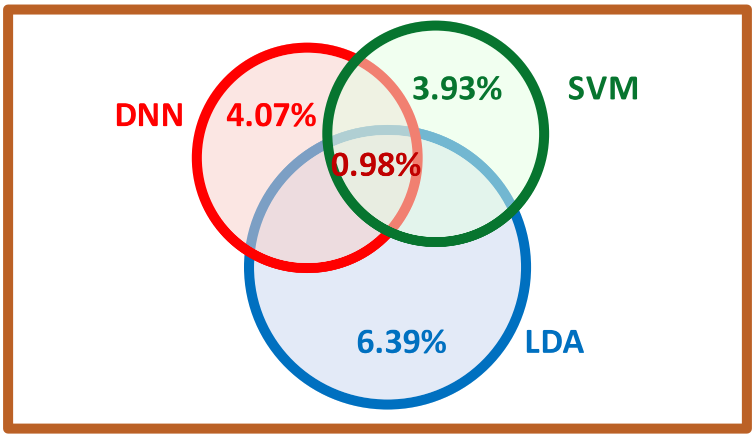


圖 37、各性別分類器之誤判樣本比例分佈

由圖中可見，在測試語者中，線性判別分析（藍色圓圈）、支持向量機（綠色圓圈）與神經網路（紅色圓圈）性別分類器之錯誤率各為 6.39%、3.93% 與

doi:10.6342/NTU201801578

4.07%，而同時被此 3 種性別分類器誤判的語者，僅佔全體之 0.98%，遠小於其他單一性別分類器之錯誤率，說明此種合成方法在理論上是可行的。

此處一共嘗試了 3 種不同的性別分類器合成方法：第一種作法為「投票」，對於一位測試語者，選擇最多性別分類器對其預測的性別作為預測結果，其數學式可表示為：

$$score_{tot}(s) = \frac{1}{3} prd_{lda}(s) + \frac{1}{3} prd_{svm}(s) + \frac{1}{3} prd_{dnn}(s)$$

其中，$score_{tot}(s)$ 為合成性別分類器對測試語者 s 所計算之性別總分數，$prd_m(s)$ 為分類器 m 對其預測之性別（男性為 1、女性為 0），若性別總分數大於 0.5，則判定此語者 s 為男性，否則判定為女性。第二種作法為「權重投票」，其原理類似於前者，但每種分類器的預測結果會以不同權重進行加權以決定最終預測結果，其數學式可表示為：

$$score_{tot}(s) = w_{lda} \cdot prd_{lda}(s) + w_{svm} \cdot prd_{svm}(s) + w_{dnn} \cdot prd_{dnn}(s)$$

其中 w_m 為分類器 m 所佔有之權重，每個權重是以邏輯回歸法（logistic regression）訓練而來。第三種方法為「權重分數投票」，其原理類似於前者，會將每種分類器所提供的性別分數組以不同權重進行加權而決定最終結果，其中，線性判別性別分類器所提供的性別分數即為被降維至單維的 i-向量，而支持向量機與神經網路性別分類器的性別分數則是其預測一測試語者屬於男性與屬於女性的兩個機率值，其數學式可表示為：

$$score_{tot}(s) = \sum_i w_{lda_i} \cdot score_{lda_i}(s) + \sum_j w_{svm_j} \cdot score_{svm_j}(s) + \sum_k w_{dnn_k} \cdot score_{dnn_k}(s)$$

其中，$score_{m_i}(s)$ 代表分類器 m 對測試語者 s 所計算的第 i 份性別分數，w_{m_i} 則代表其對應之權重，同樣地，權重亦是以邏輯回歸訓練而來。此三種合成性別分類器方法之結果如表 3 所示：

doi:10.6342/NTU201801578

合成方法	錯誤率
投票	3.80 %
權重投票	3.80 %
權重分數投票	3.49 %

表 3、各性別分類器合成方法之實驗節果

由表中可見，雖然三種方法的錯誤率皆較單一性別分類器還低，但進步效果十分有限，若將其中表現最佳的權重分數投票版本之性別分類器套用至實驗二的各方法，則可以得到如圖 38 之結果：

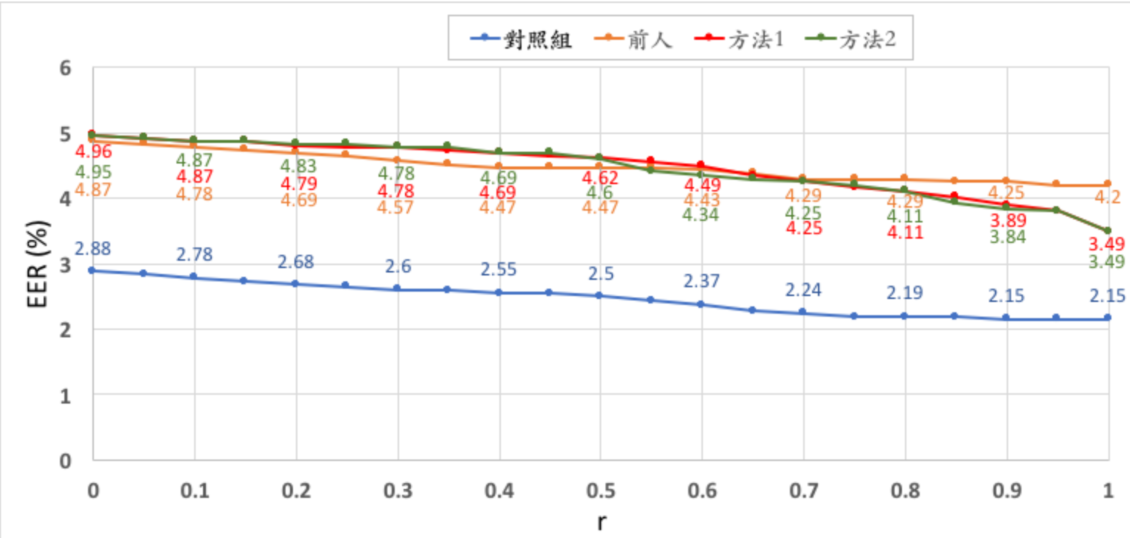


圖 38、各項語者驗證流程使用「權重分數投票」性別分類器之實驗結果

由圖中可見，雖然各流程之等錯誤率隨著性別分類器之辨識率上升而有所下降，但仍然與對照組的結果有一大段差距。此外，此種合成性別分類器之作法還存在著一大爭議：性別分類器在語者驗證系統當中只扮演著輔助的角色，但此種合成性別分類器必須額外使用 3 個 i-向量模型與分類器各自抽取 i-向量並進行性別預測，換句話說，此種做法花在性別分類工作的時間將會是語者辨識工作本身的 3 倍，可說是本末倒置，因此，就算存在其他更好的性別分類器合成方法能使辨識率進一步提高，此種作法也勢必會引人詬病。綜上所述，本研究將在以下提出不同於實驗二的另外 3 種改良流程，試圖讓系統在單一性別分類器表現不良的情況之下，仍能超越對照組作法之結果。

doi:10.6342/NTU201801578

3.5.2 方法三：將「性別分數」作為新的特徵

在先前提及的方法二流程裡，一旦系統偵測到註冊方與測試方用以代表性別資訊的 i-向量最後一維特徵之數值（0 或 1）不同，便會為此樣本評出非常低之分數，進而造成了被誤判性別的正樣本皆會被錯誤拒絕的情形。這不禁令人好奇，除了 0 或 1，性別資訊的數值沒有其他種表示方式嗎？此問題衍生了接下來將要介紹的方法三流程，其流程架構如圖 39 所示：

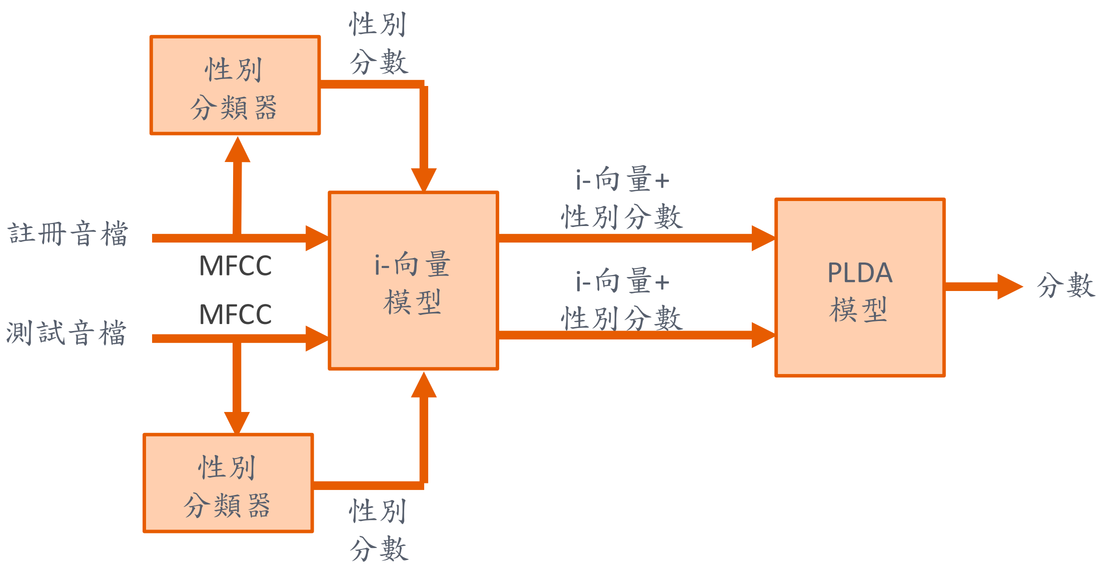


圖 39、方法三之流程架構圖

由圖中可見，本流程以前人作法流程為基礎，並模仿方法二之方式於 i-向量尾端附加性別資訊，但此處之性別資訊並不使用 0 或 1，而是仿照權重分數投票版本之合成性別分類器當中所使用的性別分數，如此一來，便能避免掉方才提及的錯誤拒絕之問題。另外，為避免重蹈實驗二各流程的覆轍，本流程亦避開了前人做法的「性別相關模型」和方法 1 的「對於雙方性別不同之樣本予以拒絕」等作法，回到與對照組類似的架構。

方法三流程的實驗結果如圖 40 所示：

doi:10.6342/NTU201801578

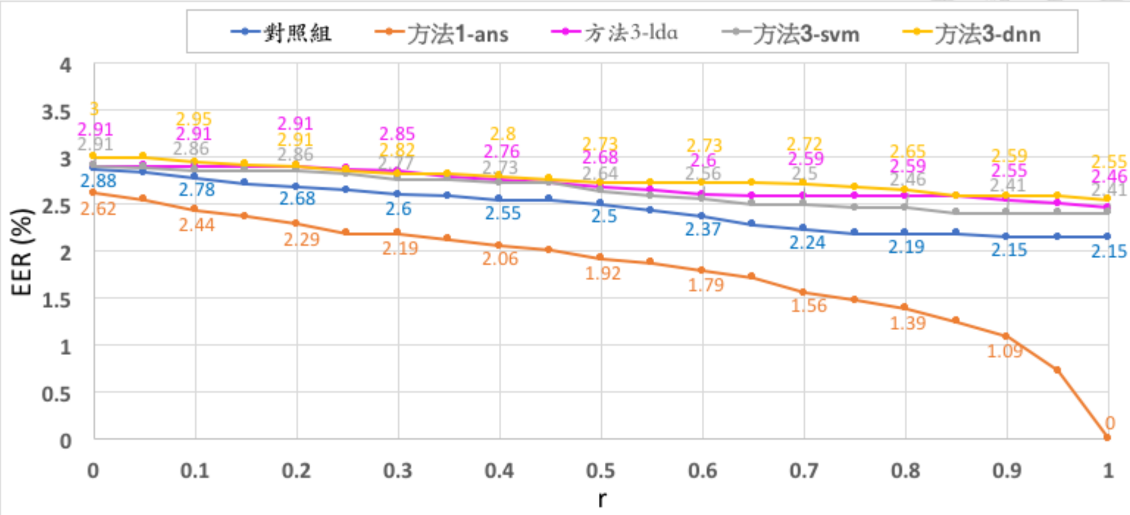


圖 40、方法三流程之結果

圖中，藍線代表對照組流程之結果，橘線則代表方法一流程在使用真實性別代替性別分類器預測之結果。原則上，實驗四之目標是在使用真實性別分類器的前提下，實作出比藍線更加進步的改良版本，但是由於本實驗接下來的改良流程皆是使用效果不精良之性別分類器對測試語者進行性別預測，因此這些改良流程的結果在預期上將難以超越圖中的橘線，意即橘線可被視為實驗四所有改良流程之上限。圖中另外 3 條粉紅色、灰色與黃色的線，分別代表以此方法三流程套用線性判別分析、支持向量機、神經網路性別分類器之結果，雖然此 3 條線已較先前各改良流程進步許多，但遺憾地是，這些實驗結果仍舊比對照組流程要差。本流程失敗的原因，除了推測是性別分類器在產生性別分數上的成效並不理想之外，亦必須考慮使用性別不相關模型取代性別相關模型所帶來的退步。

3.5.3 方法四：檢查註冊音檔的性別分類準確率

在先前的實驗中，我們可以見到圖 36 與圖 26 之差距來自於少數遭誤判性別的測試語者被實驗二之各流程系統錯誤拒絕。倘若我們能將這些測試語者遭誤判性別的樣本與其他樣本分別交由對照組流程與實驗二之改良流程進行判別，理論上便能得到近似於實驗二之效果。若欲實際執行此概念，要如何預測性別分類器能否正確預測測試語者之性別，必然是此方法必須面對的第一個問題，而方法四流程將提供一個較簡單的預測方法：藉由同一語者之註冊音檔來判斷測試音檔被

doi:10.6342/NTU201801578

誤判性別之可能性，以下為其流程架構圖：

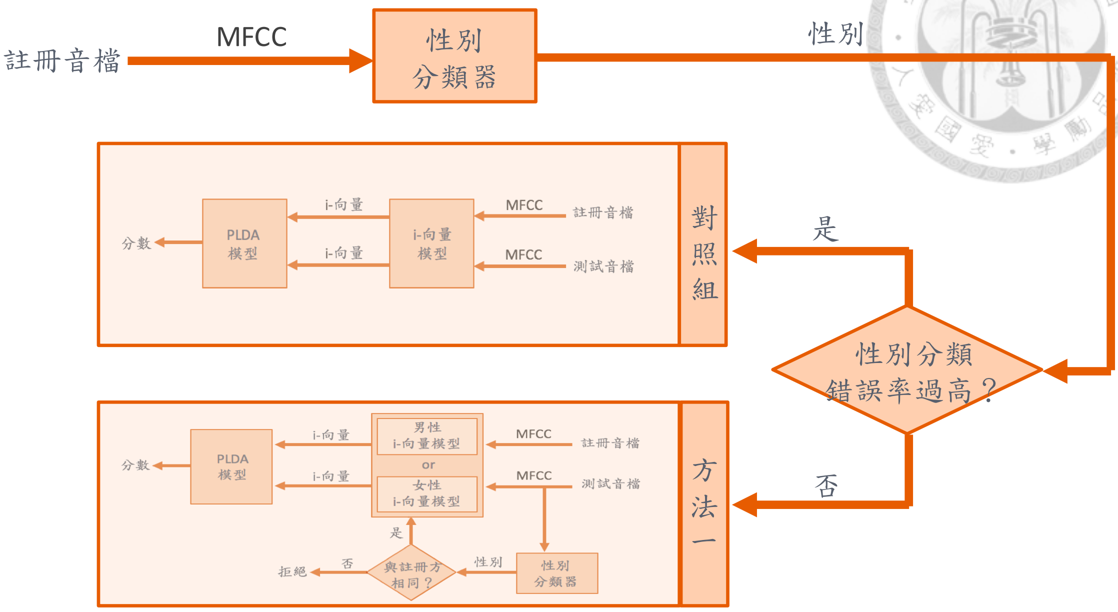


圖 41、方法四之流程架構圖

由圖中可見，當一位語者使用本系統進行註冊，系統會以性別分類器預測其註冊音檔之性別，而此處所使用的性別分類器為實驗三當中表現最好的支持向量機性別分類器。由於此位語者必須提供其真實性別，我們便能藉此測定性別分類器在此位語者身上之表現：若性別分類器在此語者之註冊音檔中出現過多次誤判情形，則代表性別分類器在此位語者身上表現較差，或者此位語者之「聲音性別」並不典型，例如此語者可能是一位音調偏低的女性，或音調偏高的男性。針對此類型之語者，我們可以預料到其未來在作為測試語者時，亦會有很高的機率會被誤判性別，進而無法通過自己以正確性別註冊的語者模型，即錯誤拒絕的情況。因此，當這類型的語者在樣本中作為註冊方時，便將其交由對照組流程進行判別，而其他情況則使用在實驗二中表現最好的方法一流程進行判別。

方法四之前半部流程基本上可被視為一種二元分類器，此分類器之目的在於預測測試語者之性別是否會被誤判，而其演算法則為：只要一語者滿足「註冊音檔之性別誤判率過高」之條件，便判斷其測試音檔將被誤判性別。既然是二元分類器，我們便能測定其召回率（recall）、精確率（precision）與 F-度量，如圖 42 所示：

doi:10.6342/NTU201801578

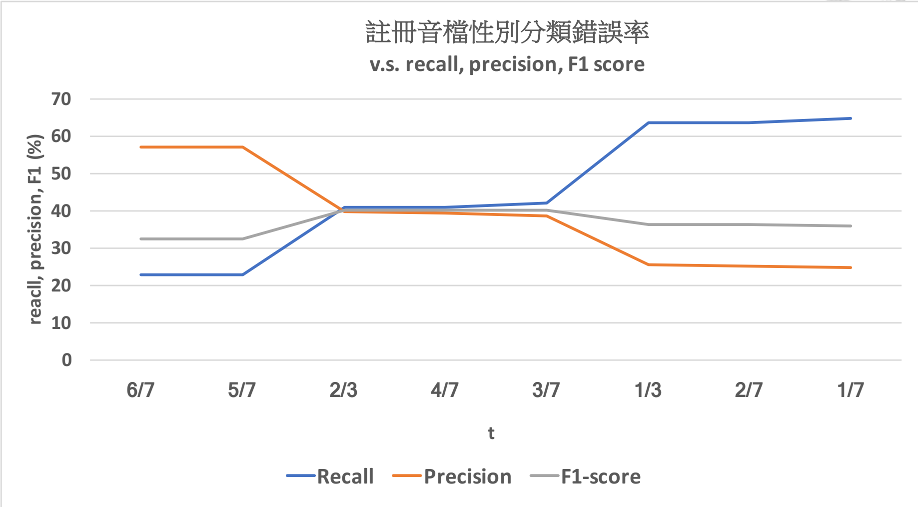


圖 42、方法四流程之「測試音檔性別誤判預測器」之效果

圖 42 之橫軸 t ，此 t 值代表的意義為註冊音檔性別分類錯誤率之門檻值，意即：若一語者在註冊音檔之性別分類錯誤率大於或等於此 t 值，便導入對照組流程，否則導入方法一流程。其中，因本實驗所使用的資料配置中，註冊語者只有 3 或 7 份註冊音檔，因此圖中橫軸的 t 值便只有 n／3 或 n／7 所組合起來的有限種可能，其中 n 為小於等於其分母之正整數；而此「測試音檔性別誤判預測器」之召回率、精確率與 F-度量則分別以藍線、橘線與灰線表示，圖之縱軸則為其百分比。我們可以想像，當 t 值減小時，真陽性（true positive）之樣本增加，但真正被誤判性別的語者總數是不變的，因此圖 42 中的召回率便呈現出嚴格遞增之現象；但偽陽性（false positive）之樣本增加速率會比真陽性更快，因此圖 42 中的精確率便呈現出嚴格遞減之現象；至於 F-度量，則是在 t 之值偏中間的時候有較大的值。

然而，該如何選定最合適之 t 值，則必須實際檢視系統之整體表現，如圖 43 所示：

doi:10.6342/NTU201801578

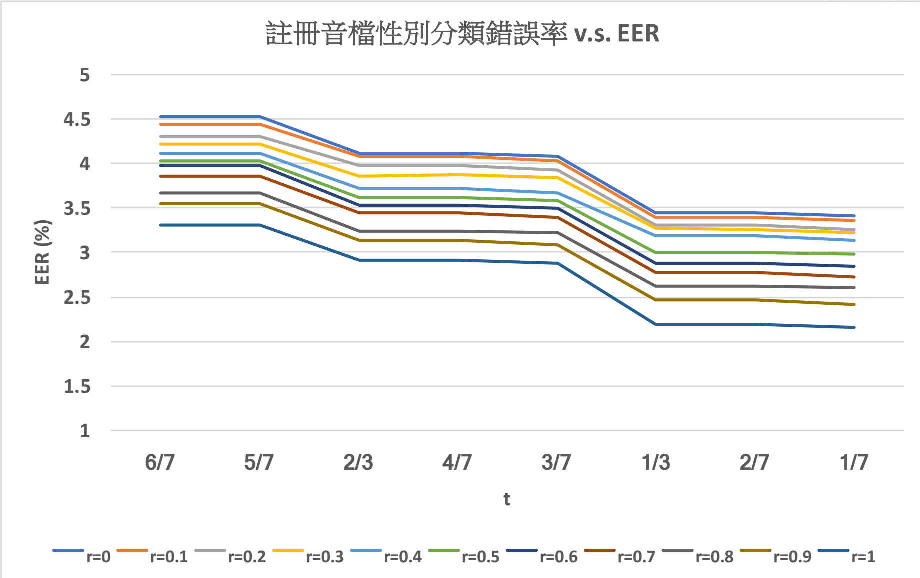


圖 43、方法四流程之選定註冊音檔性別誤判率門檻

圖 43 之橫軸同樣 t，縱軸為等錯誤率，至於不同的 r 值，即與註冊語者不同性別的仿冒者比例，則以不同顏色的線表示。在理論上，若註冊音檔之性別誤判率之門檻被設定太高，便無法有效選出測試音檔被誤判性別的語者；反之若設定得太低，雖然可以篩選出大量測試音檔被誤判性別的語者，但因為有過多沒有被誤判的語者將同時被導入對照組流程，反而會造成進步效果不彰。依此論點，圖 43 的各條曲線在預期上應該會與圖 42 中的 F-度量成反比，呈現出兩側較高、中間較低的現象，而最低點則是最合適之門檻，並且不同曲線應該能得到不同的最佳門檻值。但實際上，實驗結果顯示，每條曲線皆呈現出嚴格遞減之現象，並在最右側達到最低點，也就是說，在真實情況中，任何語者只要有一份註冊音檔被誤判性別，便必須導入對照組流程，才能達到最佳效果。

圖 44 為方法四流程指定 t 值為 1／7 時之實驗結果：

doi:10.6342/NTU201801578

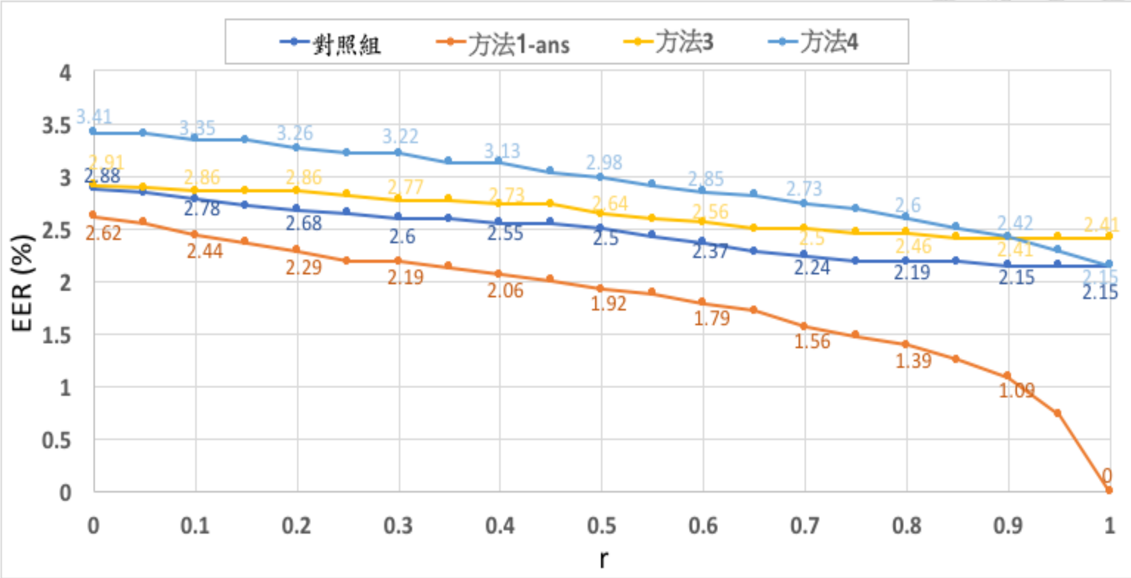


圖 44、方法四流程之結果

同樣地，藍線與橘線分別代表對照組與方法一流程使用真實性別代替性別分類器預測之結果，黃線為先前方法三使用支持向量機性別分類器之結果，而新增的天藍色線條則是本方法四流程之實驗結果。相當可惜地，此方法四流程之結果仍然無法超越對照組之結果，顯然地，以註冊音檔來預測同一語者之測試音檔被誤判性別之可能性，並不是一套理想的方法。

3.5.4 方法五：檢查測試音檔的男女性別機率差

方法五與方法四類似，亦是對照組流程與方法一流程的組合，而其不同於前者之處，在於其流程是以測試音檔本身來預測性別分類器能否正確預測其性別。實驗三所實做的性別分類器，除了能預測一位語者之性別，更可以精確地計算出該語者屬於男性或女性的機率，而若一份音檔的兩個機率值相差非常多，代表性別分類器對此音檔之性別預測結果有較大的信心，反之，若一份音檔的兩個機率值相差過小，則代表性別分類器的判斷有較大的機會出錯。依此概念，便可以衍生出如圖 45 之流程：

doi:10.6342/NTU201801578

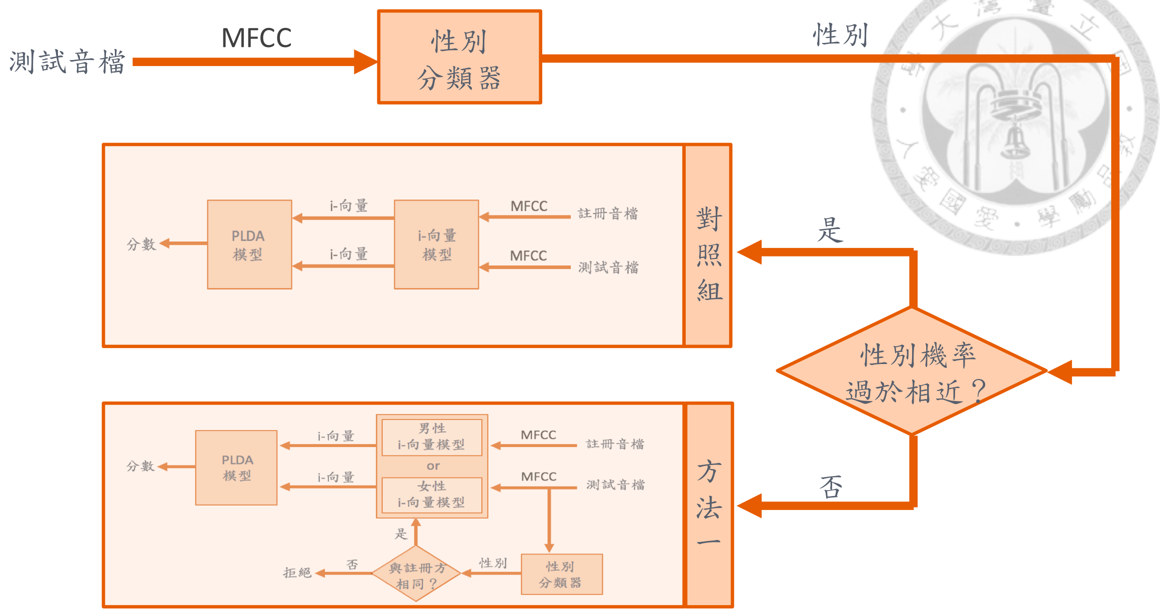


圖 45、方法五之流程架構圖

本流程同樣使用支持向量機性別分類器為測試音檔進行性別預測。當性別分類器於測試音檔計算出來的兩個性別機率值過於相近，便將此樣本交由對照組流程進行判別；反之，則交由方法一流程進行判別。

同樣地，方法五之前半部流程可被視為預測測試語者之性別是否會被誤判的二元分類器，其召回率、精確率與 F-度量則如圖 46 所示：

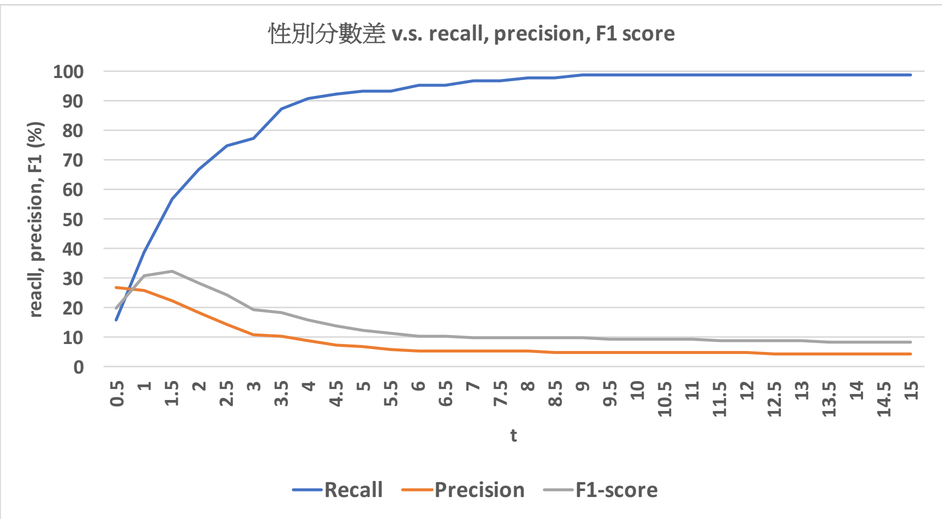


圖 46、方法五流程之「測試音檔性別誤判預測器」之效果

圖 46 之橫軸同樣為 t ，此 t 值代表的意義為測試音檔性別對數機率差之門檻值，意即：若性別分類器在測試音檔上所計算出之兩個性別對數機率差小於此值，

doi:10.6342/NTU201801578

便導入對照組流程，否則導入方法一流程。同樣地，召回率、精確率與 F-度量分別以藍線、橘線與灰線表示，圖之縱軸則為其百分比。由於本流程的門檻概念與方法四相反，因此圖 46 中的召回率與準確率亦與方法四相反，分別隨著 t 值之增加呈現出嚴格遞增與遞減之現象，而 F-度量則一樣在 t 之值偏中間（約接近 1.5 ）時有較大的值。

系統之整體表現與 t 值之關係則如圖 47 所示：

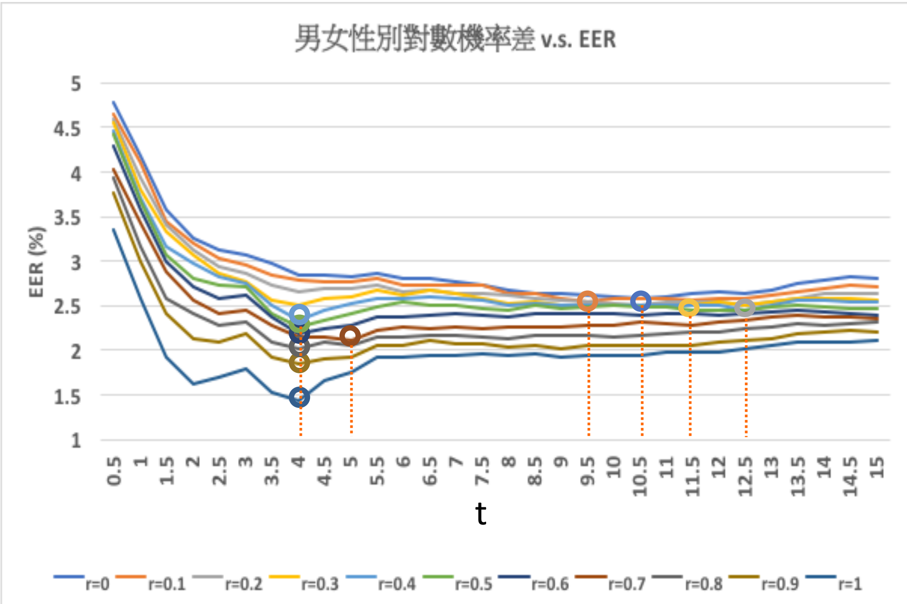


圖 47、方法五流程之選定測試音檔性別機率差門檻

與方才方法四不同，圖 47 中的每條線皆呈現出兩側高、中間低的現象，而每條線的最低點位置（在圖中以圓圈標出）亦不相同，例如：當 r 值較小（0～0.3）時，這些線條之最低點出現在圖的右半部，顯示其最佳門檻值偏大；而當 r 值較大（0.4～1）時，這些線條之最低點出現在圖的左半部，顯示其最佳門檻值偏小。由此現象可以推斷，男女性別機率差之門檻並沒有最佳解，而是會隨著資料的 r 值而變動。

圖 48 為方法五流程指定 t 值為 4 時之實驗結果：

doi:10.6342/NTU201801578

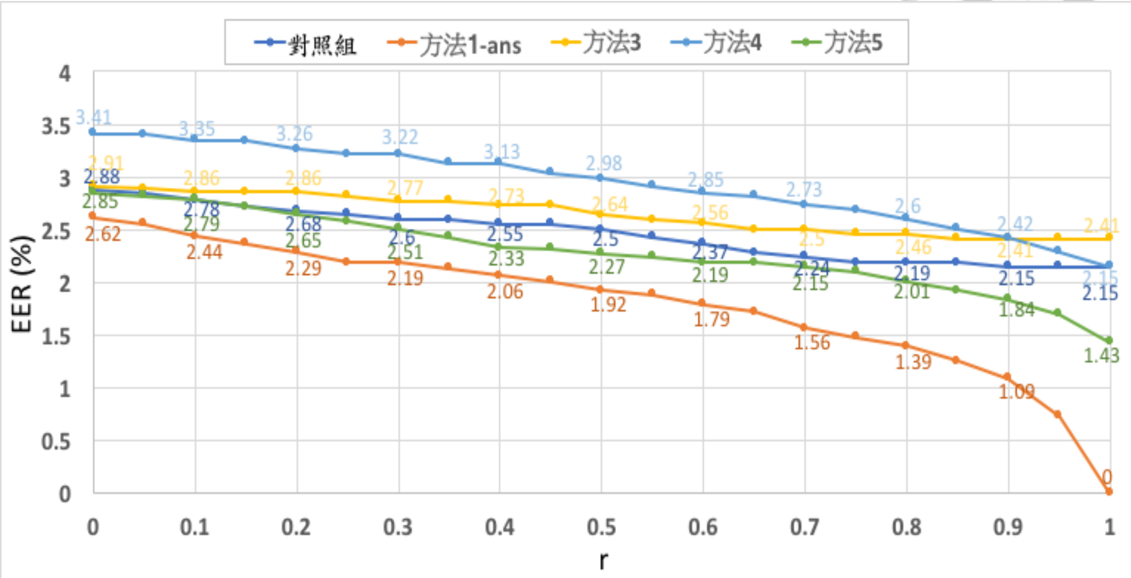


圖 48、方法五流程之結果

同樣地，藍線、橘線、黃線與天藍線分別代表對照組、方法一流程使用真實性別代替性別分類器、方法三與方法四之結果，而新增的綠線則是本方法五流程之結果。我們可以看到，此方法五流程終於達到了最初的預期，也就是在性別分類器表現不良之前提下，仍然超越對照組流程的目標。

doi:10.6342/NTU201801578

第 4 章 結論與未來方向

在語者驗證領域中，有效利用語者性別將能提升系統之準確度，但前一章節的各項實驗結果證實，無論性別分類器的表現或仿冒者的性別組成，皆會大大小小地影響系統的表現，因此不同的性別資訊運用方法，皆各有其適用的前提。以下整理本論文所使用的各種方法與其對應的適用情形作為總結：

(1) 當我們能確保系統的使用者可以提供品質良好的錄音，則性別分類器有較大機會表現良好，因此可選擇方法一或方法二；甚至當我們能進一步確保測試語者性別必定與註冊者相同，前人作法亦為很好的選擇。

(2) 當我們不能保證使用者能提供優良品質之錄音，則性別分類器有較大機會表現不良，因此在這種情況下，改良五為較合適之選項。

(3) 本論文提出的許多改良方法皆能達到較傳統作法低的錯誤率，但傳統作法的優點則是不必花費額外時間進行性別預測，因此當我們對於系統的整體運算速度有較高的需求，則可以選擇維持原來的傳統作法。

關於未來方向，若要持續往使用性別資訊的方向鑽研，則可以繼續改進性別分類器之辨識率，例如新增對於性別分類有用的資訊或特徵（例如語者之音高）、嘗試不同的特徵或分類器等；若系統之運用情境無法令性別分類器有更好的表現，則可以繼續嘗試其他結合傳統對照組流程與本論文改良流程的方法。而除性別以外，亦可考量將其他資訊（例如語者年齡、音高等）作為劃分聲學模型之依據，不同架構流程的結合也有很大機會使系統之辨識率進一步提高。

doi:10.6342/NTU201801578

參考文獻

[1] http://speech.ee.ntu.edu.tw/DSP2017Autumn/

[2] https://en.wikipedia.org/wiki/Window_function

[3] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society. Series B (methodological)* (1977): 1-38.

[4] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital signal processing* 10.1-3 (2000): 19-41.

[5] Gauvain, J-L., and Chin-Hui Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE transactions on speech and audio processing* 2.2 (1994): 291-298.

[6] Kenny, Patrick. "Joint factor analysis of speaker and session variability: Theory and algorithms." *CRIM, Montreal,(Report) CRIM-06/08-13* 14 (2005): 28-29.

[7] Kenny, Patrick, et al. "A study of interspeaker variability in speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (2008): 980-988.

[8] Dehak, Najim, et al. "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011): 788-798.

[9] Mika, Sebastian, et al. "Fisher discriminant analysis with kernels." Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.. Ieee, 1999.

[10] Prince, Simon JD, and James H. Elder. "Probabilistic linear discriminant analysis for inferences about identity." Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007.

[11] Kenny, Patrick. "Bayesian speaker verification with heavy-tailed priors." Odyssey. 2010.

[12] Garcia-Romero, Daniel, and Carol Y. Espy-Wilson. "Analysis of i-vector length normalization in speaker recognition systems." *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

doi:10.6342/NTU201801578

[13] Lyu, Siwei, and Eero P. Simoncelli. "Nonlinear extraction of independent components of natural images using radial gaussianization." *Neural computation* 21.6 (2009): 1485-1519.

[14] https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2010

[15] Garofolo, John S., et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[16] Anjos, André, et al. "Continuously reproducing toolchains in pattern recognition and machine learning experiments." (2017).

[17] Kanervisto, Anssi, et al. "Effects of gender information in text-independent and text-dependent speaker verification." *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017.

[18] M. Wang, Y. Chen, Z. Tang, and E. Zhang, "I-vector based speaker gender recognition," in 2015 IEEE Advanced Information Technology, Electronic and Au- tomation Control Conference (IAEAC), 2015, pp. 729- 732.

doi:10.6342/NTU201801578

摘要

在語者驗證領域中，在不改變聲學模型架構之前提下，以男性與女性之語料分別訓練的性別相關模型取代性別不相關模型，是常見的提升系統辨識率作法之一。然而，在實際運用情形中，由於測試語者的性別是未知的，因此性別分類器在此流程下便扮演了非常重要的角色，其準確度更會直接影響語者驗證系統的表現；而確保系統面對不同性別之仿冒者皆能正確拒絕，亦是此作法相當重要的一項訴求。

為探討不同的「語者性別資訊運用方法」對於語者驗證系統所產生的影響，本論文實作了以 i-向量與機率性線性判別分析模型為語者特徵與評分器之語者驗證系統，與 3 種以 i-向量為基礎的性別分類器。本論文在分析一般使用性別相關模型之語者驗證系統的弱點後，分別於「性別分類器表現良好」與「性別分類器表現不良」之兩大狀況下提出其他若干種不同的性別資訊應用方法，並分析各方法在不同的仿冒者性別組成下之表現，最後亦達成了在各種情況下皆能讓系統表現超越傳統作法之目標。

關鍵詞：語者驗證、性別資訊、性別分類器、i-向量、機率性線性判別分析

doi:10.6342/NTU201801578

Abstract

For speaker verification task, one way to improve system's accuracy without changing the algorithm of acoustic model is to use gender-dependent model instead of gender-independent one. However, since test speakers' gender are not available, gender classifier plays an important role since its accuracy directly affects the performance of the whole speaker verification system; furthermore, ensuring that the system can maintain good performance under different gender composition of test speakers is also an important appeal.

To explore the impact of different gender information's usage on speaker verification system, this paper implemented a speaker verification system using i-vector and PLDA model as speaker feature and scoring model respectively, and 3 i-vector-based gender classifier. After analyzing the weakness of speaker verification system using gender-dependent model in a general way, we proposed several different methods for the application of gender information under the conditions when gender classifier has good and poor performance respectively; moreover, we analysis the performance of each method under different gender composition of test speakers as well. Finally, we reached the goal of making our system achieve better performance than tradition practice under different circumstances.

Keywords: Speaker Verification, Gender Information, Gender Classifier, i-vector, PLDA

doi:10.6342/NTU201801578

目錄

doi:10.6342/NTU201801578

doi:10.6342/NTU201801578

圖表目錄

doi:10.6342/NTU201801578

doi:10.6342/NTU201801578

第 1 章 緒論

近年來，個人電子設備以及網路的普及，使越來越多人習慣將大量個人資料存放於手機、電腦，甚至是雲端裝置。為防止資料外洩，許多設備會在資料被存取前驗證使用者的身份，而由於軟硬體技術之快速發展，諸如指紋、虹膜、人臉辨識等生物識別技術，已逐漸成為重要的身份驗證機制，相較於傳統輸入文字密碼的方式，生物識別兼具了更大的便利性與安全性。其中，由於語言是人類之間彼此交流最重要的手段之一，透過語音進行身份辨識的技術：語者辨識（speaker recognition），乃不可避免地成為生物識別領域的一大研究重點。

1.1 語者辨識簡介

語者辨識，或稱自動語者辨識（automatic speaker recognition），是一種將人類的語音透過電子設備轉換成電子訊號後，利用演算法分析不同語者之間的聲紋特性差異，來進行語者身份辨識的工作。語者辨識的應用領域相當廣泛，除了方才提及的個人設備存取控制之外，舉凡門禁系統、信用交易、犯罪偵測、電子會議輔助、刑事案件舉證等皆屬於其應用範疇；且由於其需求不斷擴大，如何增進語者辨識技術的準確度與或提升系統演算法之速度，至今仍吸引學界與業界投入大量研究心力。

1.1.1 語者識別與語者驗證

語者辨識依據應用情境又可細分為 2 個項子項目：語者識別（speaker identification）與語者驗證（speaker verification）。前者的工作是要在一群已知的語者之中識別出未知的語者，換句話說，當給定一段聲音訊號，語者識別系統必須自其資料庫中，判斷誰的聲音與此段聲音最為相像；而後者的工作則是要判斷兩段聲音是否屬於同一位語者，也就是驗證測試者所宣稱的身份是否為真。

doi:10.6342/NTU201801578

1.1.2 語者辨識之標準流程

典型的語者辨識流程可分為 3 項程序：訓練（training）、註冊（enrollment）與評測（evaluation 或 test）。訓練程序的目的在於訓練出後續程序需使用到的所有模型，例如：能從語者的聲音訊號擷取出特徵的聲學模型（acoustic model），以及為兩段不同的聲音計算聲紋相似度的評分模型（scoring model）等；註冊程序則是要為註冊語者客製化一套專屬於他／她的語者模型（speaker model）；而最後的評測程序，則是要將欲評測的聲音輸入註冊語者的語者模型，以計算出相似度分數，進而完成身份辨識的任務，在此程序中，語者識別系統會選擇分數最高的註冊語者作為識別的答案，而語者驗證系統則會依據分數是否高於一事先給定之門檻值來決定拒絕或接受。

1.1.3 語者辨識之評量指標

評斷語者辨識系統表現的指標有許多種，由於本論文所實作之系統屬於語者驗證之架構，故以下所介紹的各種指標將著重於語者驗證部分。傳統上，二元分類問題最常見的評量標準有準確度（accuracy）或 F-度量（F-measure）等，其定義如圖 1 所示：

		Answer	
		TRUE	FALSE
Predicted	TRUE	TP (True Positive)	FP (False Positive)
	FALSE	FN (False Negative)	TN (True Negative)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+TN}$$

$$Specificity = \frac{TN}{TN+FP}$$

$$F-measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

圖 1、分類問題的常見指標

doi:10.6342/NTU201801578

而若要同時檢視語者驗證系統在不同的分數門檻下的效能，則可以檢測錯誤權衡曲線圖（Detection Error Rate Curve, DET Curve）或接收者操作特徵曲線圖（Receiver Operating Characteristic Curve, ROC Curve），此兩條曲線分別以正確接受率（true acceptance rate, i.e. $\frac{TP}{TP+FN}$）以及錯誤拒絕率（false rejection rate, i.e. $\frac{FN}{TP+FN}$）為縱軸，並以錯誤接受率（false acceptance rate, i.e. $\frac{FP}{FP+TN}$）為橫軸做圖，分別如圖 2-1 與圖 2-2 所示：

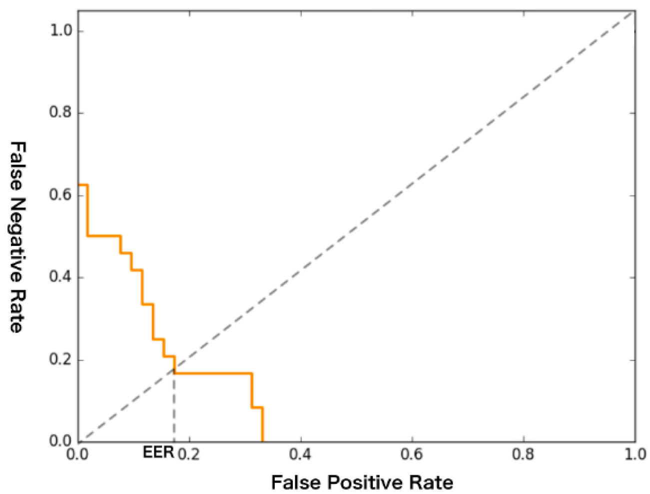


圖 2-1、錯誤權衡曲線圖

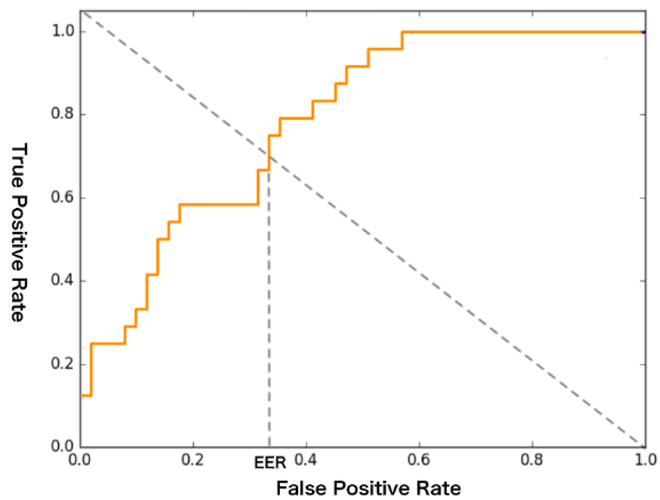


圖 2-2、接收者操作特徵曲線圖

此外，近年來諸多語者驗證相關比賽最常採納的一項評量標準為等錯誤率（Equal Error Rate, EER），其概念為：當分數門檻值能使系統錯分正樣本（positive trial）與負樣本（negative trial）之比例相同，也就是錯誤拒絕率等於錯誤接受率時，此錯誤率即為等錯誤率。由其概念可知，等錯誤率並不需要系統給定分數門檻，其所著重的乃在於系統能否「使正樣本的分數皆較負樣本高」，而非傳統上的「正確地接受正樣本並拒絕負樣本」。

doi:10.6342/NTU201801578

1.2 研究簡介

本碩士論文以語者驗證作為研究主軸，主要方向為探討「語者的性別資訊」對於語者驗證系統所產生的影響，探討的項目包括以下 3 點：

(1) 本論文將探討前人如何利用性別資訊提升語者驗證系統之辨識率，並分析該作法之弱點。

(2) 本論文將提出其他若干種性別資訊應用方法之改良，並分析各種方法適用之情況。

(3) 本輪文將實作若干種性別分類器，並分析各分類器優劣，以實際將其應用於語者驗證系統中。

1.3 章節概要

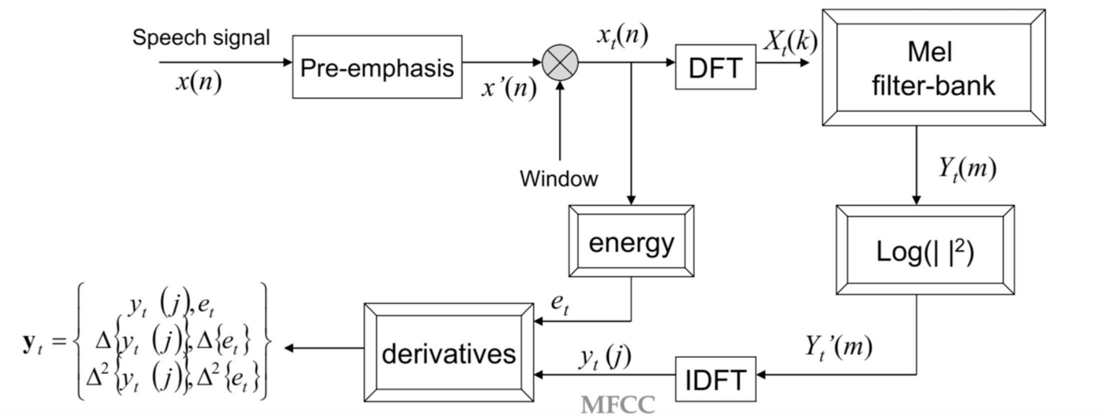
本論文之第一章為緒論，包含語者辨識與本論文研究方向之簡介；第二章為相關研究方法，包含本論文所設計的實驗中所使用到的各項相關技術之理論與演算法之簡介，包括 MFCC 特徵、高斯混合機率模型、i-向量、線性判別分析、機率性線性判別分析模型等；第三章為實驗設定與結果，包含本論文所使用的語料、資料配置以及參數設定之簡介，並詳述各項實驗結果與討論；第四章為結論，包含本論文研究之總結與未來方向。

doi:10.6342/NTU201801578

第 2 章 語音處理與語者辨識相關技術

2.1 語音訊號之特徵：MFCC

本 2.1 章之內容與圖片素材擷取自 [1] 之課堂內容或投影片。聲音經由電子設備錄製後會轉為電子訊號，若要對其進行後續處理與分析，便必須自其中抽取適當的特徵。本篇論文所抽取之特徵為 MFCC（Mel-Frequency Cepstral Coefficients）特徵，是一種計算複雜度低、在語音訊號處理上最被廣泛使用的特徵之一，圖 3 為抽取 MFCC 特徵之標準流程，以下將對每個步驟進行簡介。



$$\mathbf{y}_t = \left\{ \begin{array}{c} y_t\ (j), e_t \\ \Delta\{y_t\ (j)\}, \Delta\{e_t\} \\ \Delta^2\{y_t\ (j)\}, \Delta^2\{e_t\} \end{array} \right\}$$

圖 3、MFCC 特徵之抽取流程

MFCC 特徵抽取流程之第一步為預強調（pre-emphasis），其目的是為補償語音訊號中衰減的高頻部分。自人體發出的聲音可分為兩種類型，一種是在時域上具明顯基本週期的有聲音（voiced sound），另一種則是不具明顯基本週期的無聲音 （unvoiced sound），圖 4-1 與 圖 4-2 分別則顯示了兩者的時域圖與頻域圖：

doi:10.6342/NTU201801578

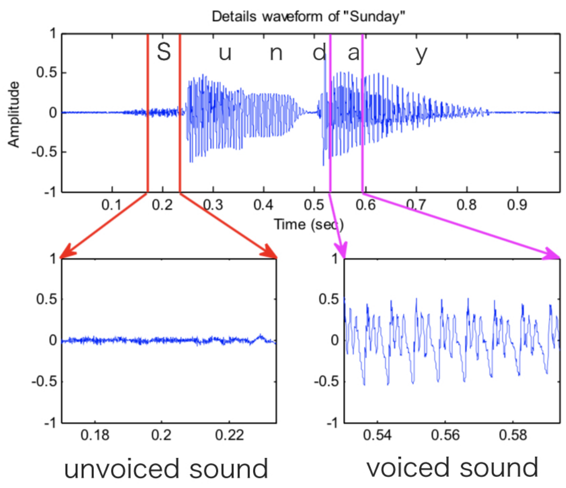


圖 4-1、有聲音與無聲音之時域圖

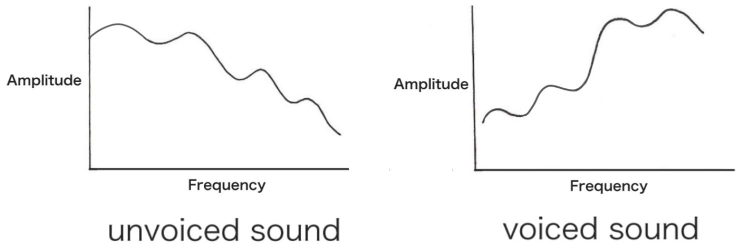


圖 4-2、有聲音與無聲音之頻域圖

其中，有聲音對語音訊號分析來說較能提供有用資訊，但其高頻部分容易在傳輸過程中損耗，導致有聲音在頻域的強度分佈情形呈現低頻強、高頻弱的衰減趨勢。此衰減趨勢容易導致一些重要資訊流失，例如共振峰結構（formant structure）變得不明顯，因此需要透過此預強調流程來彌補。假設原始聲音訊號在時域上以 $x(t)$ 表示，則通過預強調步驟後的新訊號 $x'(t)$ 可表示為：

$$x'(t) = x(t) - a \cdot x(t-1)$$

其中 a 之值通常取在 0.9 到 1 之間。其背後原理為：假設原始訊號 $x(t)$ 通過經 Z 轉換（Z-transform）後可表示為：

$$Z[x(t)] = X(z)$$

則新訊號 $x'(t)$ 通過 Z 轉換後則可表示為：

$$Z\left[x(t) - a \cdot x(t-1)\right] = X(z) \cdot (1 - a \cdot z^{-1})$$

由上式可以看到，新訊號在頻域上即為原始訊號乘上 $(1 - a \cdot z^{-1})$，而此值在低頻較接近 0，越往高頻靠近則約接近 1，因此此流程之原理可視為將原始訊號通過高通濾波器，以補償語音訊號中衰減的的高頻部分。

訊號通過預強調處理後，整段音訊會被切分成許多等長、允許重疊的音幀（frame 或 window），如圖 5 所示：

doi:10.6342/NTU201801578

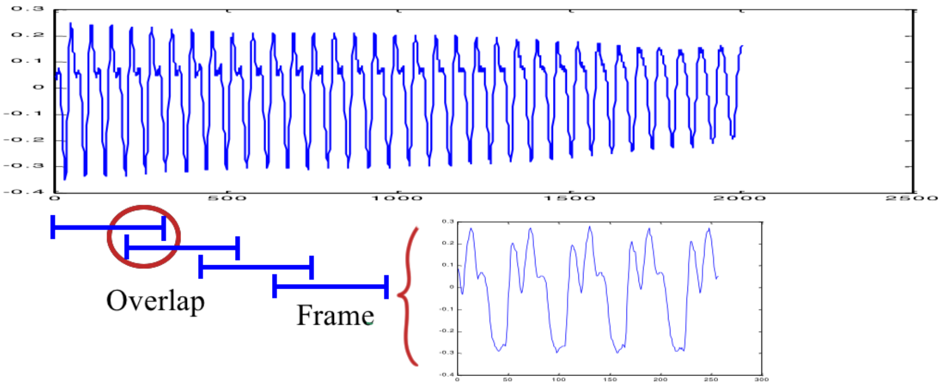


圖 5、聲音訊號與音幀

一個音幀的持續時間相當短，而聲音訊號在極短的時間內，特性是較穩定且一致的，因此較容易擷取出有用的資訊。用來從訊號中切出音幀的函數稱為音窗函數（windowing function），將訊號與音窗函數在時域上進行卷積運算（或在頻域上進行乘積運算）即可得到單一音幀長度之訊號。最常被使用的音窗函數為長方窗（rectangular window）與漢明窗（Hamming window），兩者在時域之數學式可表示為：

$$長方窗：w(t) = 1$$

$$漢明窗：w(t) = \alpha - \beta cos\frac{2\pi t}{N-1}, \alpha = 0.54, \beta = 0.46$$

而兩者的時域圖與頻域圖則如圖 6-1 與圖 6-2 所示 [2]：

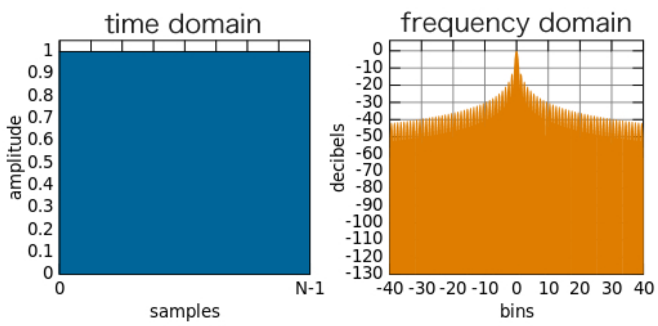


圖 6-1、長方窗之時域圖與頻域圖

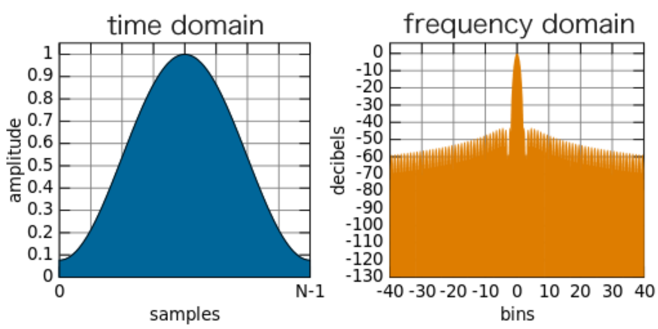


圖 6-2、漢明窗之時域圖與頻域圖

doi:10.6342/NTU201801578

在兩者的頻域圖上，正中間的最高峰稱為主辦（main lobe），周圍其他較小的峰群稱為旁辦（side lobe）。在語音訊號分析上，通常希望旁辦的峰高遠低於主辦，因此由圖可見，漢明窗是比長方窗更理想的一種窗函數。

訊號被切割成音幀後，會進行離散傅立葉變換（Discrete Fourier Transform, DFT）將訊號從時域轉換到頻域，接著通過梅爾濾波器群（Mel-filter bank），並各自取頻率之係數和，如圖 7 所示：

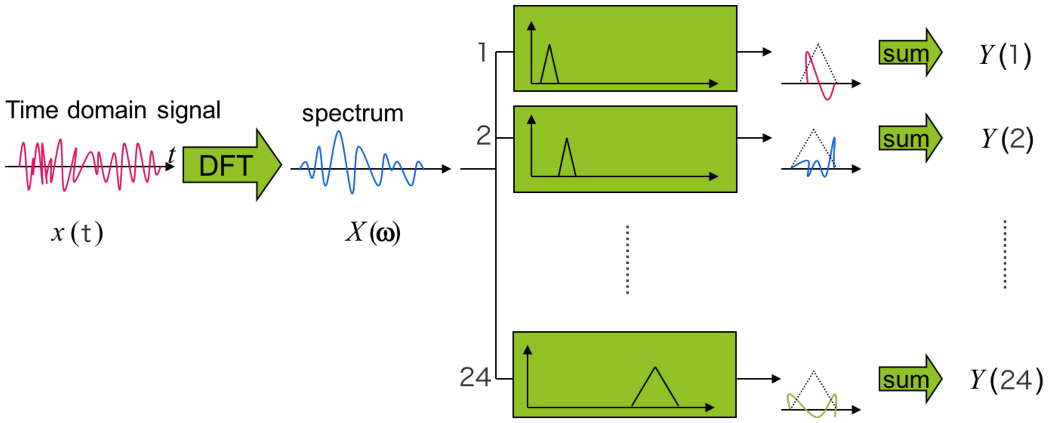


圖 7、離散傅立葉變換與梅爾濾波器

由圖中可見，梅爾濾波器群為 24 個三角帶通濾波器，各濾波器對應之頻帶則稱為關鍵頻帶（critical band）。這些關鍵頻帶是依據人類兩大內耳特性所設計：第一，人類的內耳構造為螺旋狀，其內部的不同區段各自負責不同音訊頻帶之處理，如圖 8 所示：

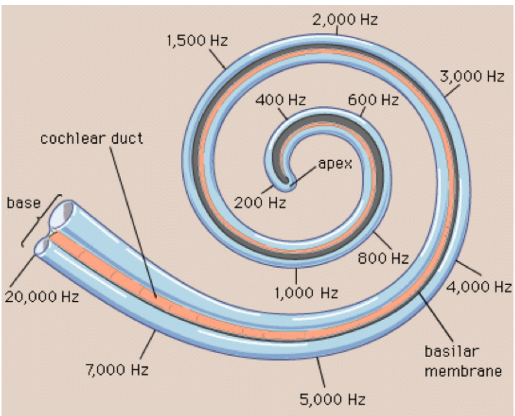


圖 8、人類的內耳構造

doi:10.6342/NTU201801578

並且，內耳相鄰的區段所負責的頻帶是會重疊的，因此相鄰的關鍵頻帶在設計上亦會有所重疊的；第二，內耳越靠近外部的區段所負責的頻帶越高，而且頻率的增長速度不一：較靠近內部的區段（約 200 Hz 至 1000 Hz），頻率增長速率近似等速，但較靠近外部的區段（約 1000 Hz 以後），則近似於指數增長。這代表人類的內耳對高頻率的聲音解析度較差，意即人耳對高頻區段的聲音較不敏感。

接著，由梅爾濾波器群所算出來的 24 個係數，會各自取絕對值與平方，最後進行對數運算；接著，24 個值將當作頻域係數，進行反離散傅立葉變換（Inverse Discrete Fourier Transform, IDFT），如圖 9 所示：

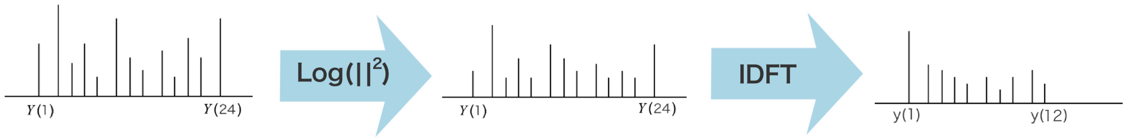


圖 9、絕對值平方、對數運算與反離散傅立葉變換

取絕對值與平方的原因，是為了捨棄在語音處理上較無用處的「向位角」資訊，而只保留「震幅」資訊。而取對數的主要原因有二：第一，人耳對聲音能量的敏感度約與能量大小取對數值成正比，因此取對數較能模擬人內耳之真實情況；第二，取對數後，在進行反離散傅立葉變換時，較利於去除多餘的雜訊。一段聲音在頻域圖中的變化趨勢可拆解為兩部分，一為起伏較大的共振峰結構（formant strucure），另一則為起伏較小但震盪速度較快的激發雜訊（excitation noise），如圖 10 之左側所示：

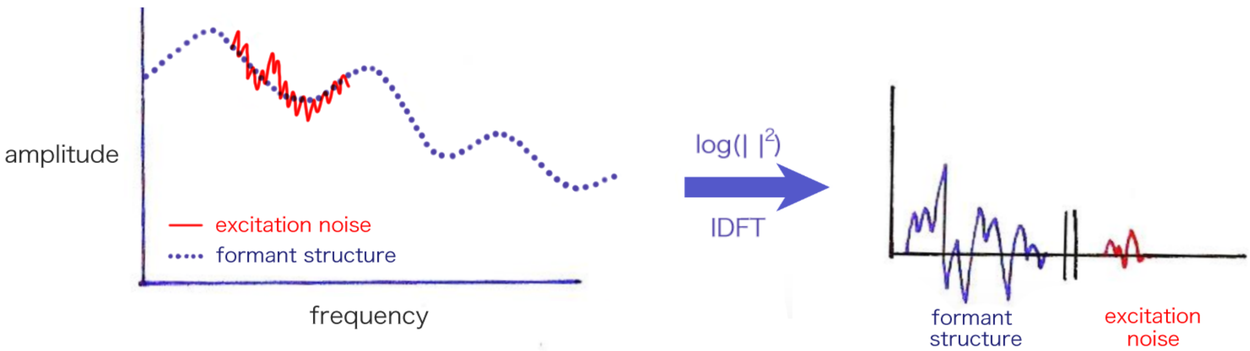


圖 10、去除激發雜訊之原理

doi:10.6342/NTU201801578

其中，單單共振峰結構便幾乎足以決定聲音的音型，因此去除激發雜訊較有利於後續之分析。假設在頻域上，原始聲音訊號 $X(\omega)$ 可表示為共振峰結構 $G(\omega)$ 與激發雜訊 $U(\omega)$ 之乘積：

$$X(\omega) = U(\omega)G(\omega)$$

但進行對數運算後，原式會由乘式轉為加式：

$$log\big(X(\omega)\big) = log\big(U(\omega)\big) + log\big(G(\omega)\big)$$

再者，在進行反離散傅立葉轉換後，更可發現共振峰結構與激發雜訊呈現出左右分離之現象，如圖 10 右側所示，由此可知，越靠左側的訊號，越有可能屬於共振峰結構，因此，如圖 9 所示，在進行反離散傅立葉轉換後，僅須取出前半部的 12 個的係數。

經反離散傅立葉變換所得到 12 個係數，通常會額外加上另一個代表能量的係數。假設訊號在一音幀內表示為 $x(t)$，則一般常見之能量係數 e 之計算方式為：

$$e = 10 \cdot log\Big(\sum_t x(t)^2\Big)$$

12 個係數再加上此 1 能量係數所組成的 13 個係數，即為最簡單的梅爾頻率倒譜係數（Mel-Frequency Cepstral Coefficients, MFCCs）。

MFCC 特徵抽取的最後一階段為導數運算（Derivative operation）。聲音訊號在傳輸過程中可能會混入直流雜訊（DC noise），進而對分析結果造成影響，假設真實訊號 $\overline{x'}$ 可表示為原始訊號 $\overline{x}$ 與直流雜訊 $\overline{h}$ 之加乘，如圖 11 所示：

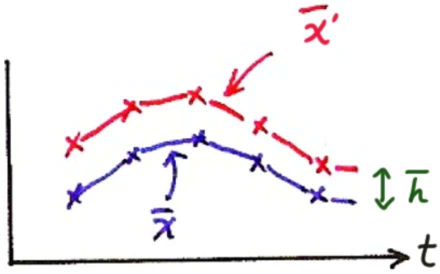


圖 11、直流雜訊示意圖

doi:10.6342/NTU201801578

由圖中可以發現，即使加上直流雜訊，原始訊號與真實訊號的變化趨勢仍是一致的，故兩者之斜率並不會因直流雜訊而有所不同。因此，以微分的方式取得梅爾頻率倒譜係數之導數，並將其合併進原來的梅爾頻率倒譜係數，是減少直流雜訊對特徵影響的最常見方法。由於電子訊號是由不連續的訊號點組成，故必須以前後若干訊號點以不同權重加乘，來模擬導數的計算，如圖 12 所示：

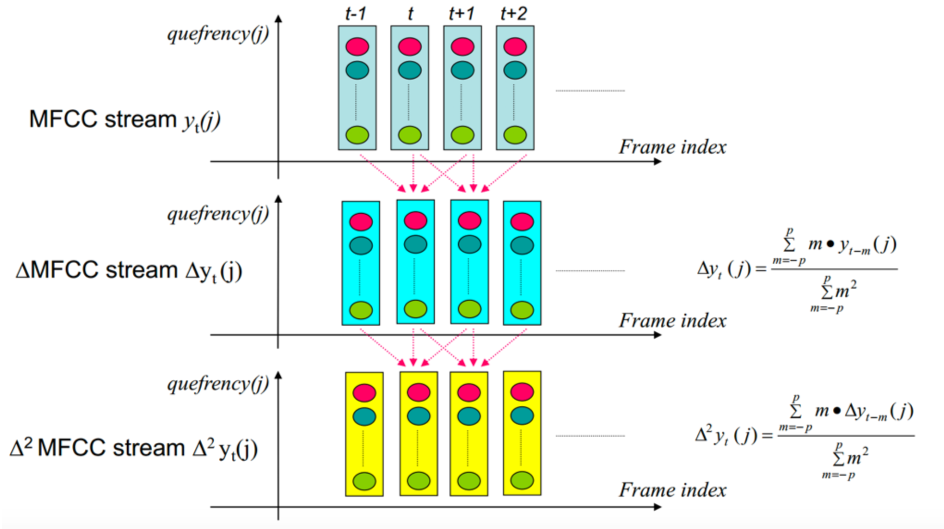


圖 12、MFCC 與其導數之計算

通常，MFCC 特徵會計算 2 階的導數，因此原來的 13 維特徵加上另外兩階導數的 26 維特徵，便可組成 39 維的特徵。

doi:10.6342/NTU201801578

2.2 高斯混合機率模型

為兩段語音訊號抽取出一連串特徵向量（如： MFCC）後，必須有適當的模型來判斷兩段聲音是不是屬於同一位語者。語者辨識領域較早期所流行的是一些使用向量量化（Vector Quantization, VQ）相關技術的方法，而後來取而代之的則是接下來將要介紹的高斯混合模型（Gaussian Mixture Model, GMM）。

2.2.1 高斯混合模型概論

高斯混合模型的目的與一般高斯模型相同，皆是用以模擬多維資料之機率分佈，但其內結構部則是由若干個高斯分量以不同權重加權所組成，而其中每一個高斯分量皆是一個高斯密度函數。圖 13-1 與 圖 13-2 分別為一般高斯模型與高斯混合模型之示意圖：

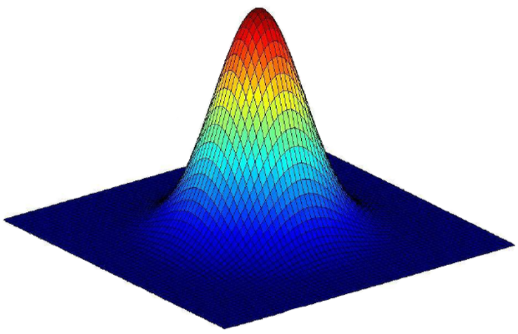


圖 13-1、一般高斯模型

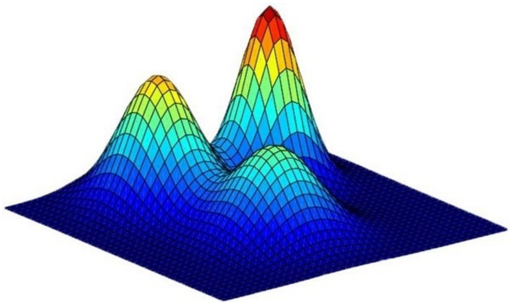


圖 13-2、高斯混合模型

為方便後續說明，此處定義高斯混合模型之高斯分量總數為 C，MFCC 特徵之維度數為 $F(=39)$。一個高斯混合模型的參數可以數學式表示為：

$$\lambda = \{\pi_c, \mu_c, \Sigma_c \mid c = 1,2,...,C\}$$

其中，λ 代表高斯混合模型之所有參數集合，π_c 、μ_c 與 Σ_c 分別代表第 c 個高斯分量之權重、平均向量與共變異數矩陣（covariance matrix），並且滿足：

$$\sum_c \pi_c = 1$$

doi:10.6342/NTU201801578

假設 x 為一高斯混合模型所欲模擬之多維變量，則其機率密度函數可表示為：

$$f(x\,|\,\lambda) = \sum_{c} \pi_c \cdot \mathcal{N}(x\,|\,\mu_c, \Sigma_c)$$

其中，$\mathcal{N}(x\,|\,\mu_c, \Sigma_c)$ 代表平均為 μ_c 、共變異數矩陣為 Σ_c 之高斯機率密度函數，即：

$$\mathcal{N}(x\,|\,\mu_c, \Sigma_c) = \frac{1}{(2\pi)^{F/2} \cdot |\Sigma_c|^{1/2}} \cdot \exp\Big(-\frac{1}{2}(x-\mu_c)^* \cdot \Sigma_c^{-1} \cdot (x-\mu_c) \Big) \tag{1}$$

其中 * 表示取矩陣之共軛轉置。藉由高斯混合模型，我們得以計算「一串特徵向量被觀察到的機率」：今考慮一位語者 s 與其聲音訊號所產生的一連串特徵向量 $W_s = \{W_{st}\,|\,t = 1,2,3,...\}$，則在給定高斯混合模型 λ 之前提下，觀察到此一連串特徵向量之機率為：

$$p(W_s\,|\,\lambda) = \prod_{t} f(W_{s,t}\,|\,\lambda) \tag{2}$$

此值乃為觀測值 W_s 於給定模型參數 λ 下之似然函數（likelihood function）。

以高斯混合模型為基礎進行語者辨識的方法中，以似然率測定（likelihood ratio test）為大宗。此方法會使用到兩種高斯混合模型，一種是語者相關模型（speaker-dependent model），另一種則是通用背景模型（Universal Background Model, UBM）。語者相關模型是以單一註冊語者的語料進行訓練，其所模擬的是該位語者的特徵機率分佈；而通用背景模型則是以大量語者的語料進行訓練，其所模擬的是世界上（除了註冊語者之外的）所有語者的共通特徵機率分佈，因此亦可視其為「語者不相關模型」。假設今欲驗證一串特徵向量 X 是否由語者 s 所產生，則此方法會以下列數學式算出分數：

$$score_{LR} = \frac{p(X\,|\,\lambda_s)}{p(X\,|\,\lambda_0)}$$

doi:10.6342/NTU201801578

其中 λ_s 與 λ_0 分別代表語者相關模型與通用背景模型。此分數的計算概念是「X 是由語者 s 所產生」與「X 是由其他人所產生」的機率比值，若此分數高於或等於一事先設定好的門檻值，則判定 X 屬於語者 s 的聲音，反之則判定 X 不屬於語者 s 的聲音。又為了避免分數之可能數值範圍過於極端，因此通常會對分數進行對數運算，使其成為：

$$score_{LLR} = \log\big(p(X\,|\,\lambda_s)\big) - \log\big(p(X\,|\,\lambda_0)\big)$$

此方式又稱為對數—似然率測定（log-likelihood ratio test）。以下將分別簡介語者相關模型與通用背景模型之訓練方式。

2.2.2 通用背景模型與其訓練流程

通用背景模型是以最大似然度估計法（Maximum Likelihood Estimation, MLE）為原則進行訓練，其概念是要使模型參數能讓訓練資料的似然函數最大化，即：

$$\lambda_0 = \underset{\lambda}{argmax} \prod_{s} p(W_s|\lambda)$$

其中 s 為訓練資料中的任一語者。但由於上式對於 λ 而言並非線性函數，因此便難以用解析解的方式直接解出此方程式，為此，此處引入一套迭代式的訓練流程：最大期望（Expectation-Maximization, EM）演算法 [3]。最大期望演算法一般分為 3 步驟，分別為模型參數初始化、計算期望（Expectation）與最大化（Maximization），其中後兩步驟可簡稱為 E 步驟與 M 步驟。E 步驟是在給定模型參數之情形下，估計潛在變數之期望值，並計算似然度；M 步驟則是使用 E 步驟之結果，更新模型參數。最大期望演算法的運作流程會令 E 步驟與 M 步驟交替執行，且每次迭代皆會使似然度有所提升，直到收斂為止，如圖 14 所示：

doi:10.6342/NTU201801578

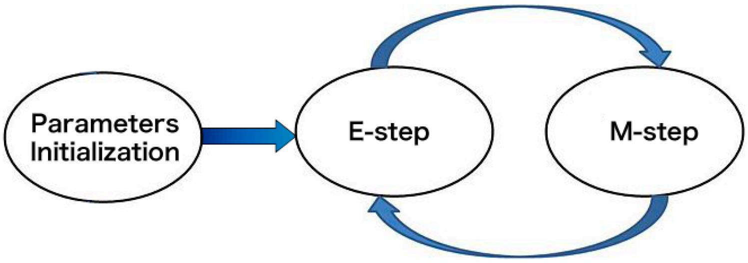


圖 14、最大期望值演算法之流程示意圖

以下將分別簡介通用背景模型訓練流程之 3 步驟。

通用背景模型訓練流程之第一步為初始化模型參數，也就是必須選定其每一個高斯分量的權重、平均向量與共變異數矩陣。最簡單的初始化方法是隨機指定每個參數之初始值，但這裡介紹一個較為常用且效果較佳的方法：k-平均（k-means）演算法。k-平均演算法屬於向量量化方法中的其中一種，是一種非監督式學習的分群方法，其目的為把空間中的若干樣本點劃分為 k 個群落，使得每個樣本點所屬群落之群心皆離此樣本點最近。以 k-平均演算法在訓練資料找出的每個群落皆可作為通用背景模型的一個高斯分量，每個高斯分量的平均向量、共變異數矩陣可由對應群落的樣本點計算而得，而其權重則正比於該群落的樣本點數量。標準的 k-平均演算法分為群心初始化、分配（assignment）與更新（update）等 3 個步驟：初始化步驟為隨機估計 k 個群心的位置；分配步驟則是依據群心位置，為每個樣本點進行分群，使得每個樣本點所屬群落之群心到該樣本點有最短距離；更新步驟則是更新每個群落的群心位置，新的群心位置即為該群落中所有樣本點的平均值。k-平均演算法會令分配與更新步驟交替執行，直到群心位置不再改變為止，其流程如圖 15 所示：

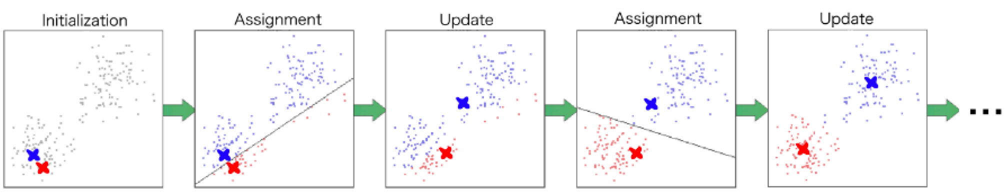


圖 15、k-平均演算法之流程示意圖

doi:10.6342/NTU201801578

E 步驟之主要目的為利用現有的模型參數，計算似然函數值，此值之計算方法與先前提到的 (2) 相同；而 M 步驟之目的為更新模型參數，為方便討論，通常會假設各高斯分量之共變異數矩陣為一純量與單元矩陣之乘積，即：

$$|\Sigma_c| = \sigma_c^2 \cdot I$$

此時 (1) 便可修改為：

$$\mathcal{N}(x \mid \mu_c, \Sigma_c) = \frac{1}{(2\pi)^{F/2} \cdot \sigma_c^F} \cdot \exp\Big(-\frac{1}{2\sigma^2}(x-\mu_c)^* \cdot (x-\mu_c)\Big)$$

由於整個最大期望演算法的目的在於將似然度最大化，因此此處將目標函數定義為似然函數取自然對數，即：

$$J(\lambda) = \ln\left(\prod_t \Big(\sum_c \pi_c \cdot \mathcal{N}(x_t \mid \mu_c, \Sigma_c)\Big)\right)$$

其中 x_t 代表訓練資料總集中的第 t 個特徵向量。而模型參數的更新必須使目標函數最大化，意即必須滿足此函數對於各模型參數之偏微分值為零的條件，為方便後續說明，此處定義 $\gamma_t(c)$ 為第 c 個高斯分量給定 x_t 的後驗機率（posterior probability），即：

$$\gamma_t(c) = P(c \mid x_t) = \frac{\pi_c \cdot \mathcal{N}(x_t \mid \mu_c, \Sigma_c)}{\sum_i \pi_i \cdot \mathcal{N}(x_t \mid \mu_i, \Sigma_i)} \tag{3}$$

經運算，可得各模型參數的更新值為：

$$\mu_c = \frac{\sum_t \big(\gamma_t(c) \cdot x_t\big)}{\sum_t \gamma_t(c)}$$

$$\sigma_c^2 = \frac{1}{F} \cdot \frac{\sum_t \big(\gamma_t(c) \cdot (x_t - \mu_c)^* \cdot (x_t - \mu_c)\big)}{\sum_t \gamma_t(c)}$$

$$\pi_c = \frac{1}{T}\sum_t \gamma_t(c)$$

doi:10.6342/NTU201801578

其中 T 為訓練資料中特徵向量總數。依照上述的 E 步驟與 M 步驟更新模型參數時，每次迭代皆會確保訓練資料之對數似然度有所提升，以數學符號可以表示為：

$$\sum_s \log\left(p(W_s|\lambda)\right) \geq \sum_s \log\left(P(W_s|\lambda_0)\right)$$

其中模型參數 λ_0 代表模型參數 λ 更新前的狀態。有關上述公式之更詳細數學過程與證明，詳見 [3] 。

2.2.3 語者相關模型與其訓練流程

語者相關模型之訓練方式有兩種，第一種和通用背景模型相同，即取該位語者之所有語料以最大似然度估計法進行訓練，但是最大期望演算法之迭代過程相當費時，尤其註冊語者數量眾多，更會付出龐大的時間代價；另外，當一位語者所能提供的語料過少，此方法未必能建立強健的語者相關模型。因此，此處導入第二種訓練方法：貝氏調適法（Bayesian adaptation） [4] [5]，或稱最大後驗機率調試法（Maximum A Posterior adaptation, MAP adaptation）。

貝氏調適法之概念乃以通用背景模型為基礎，以註冊語者之語料對模型參數進行微量調整，進而形成語者相關模型，圖 16 乃為其過程示意圖：

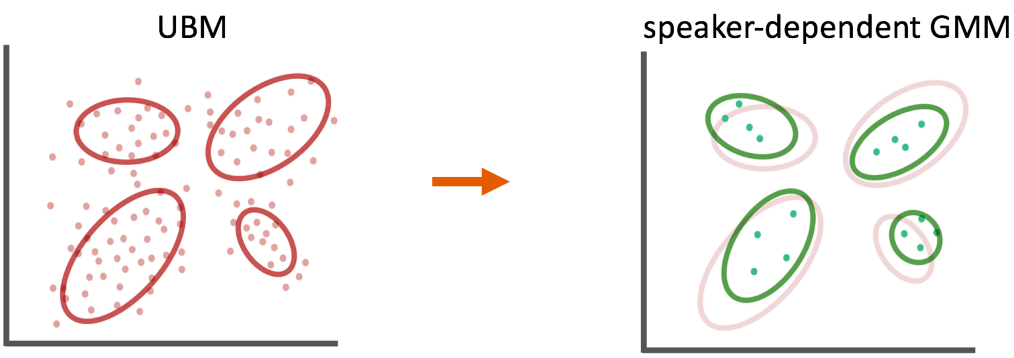


圖 16、貝氏調適法之過程示意圖

doi:10.6342/NTU201801578

為方便後續說明，此處首先定義一些數學符號，首先，給定一位語者 s 與其所產生的一連串 MFCC 特徵向量$W_s = \{W_{st} \mid t = 1,2,3,...\}$，則對於每個高斯分量 c，可定義三項鮑姆韋爾奇統計量（Baum-Welch statistics）：

$$N_c(s) = \sum_t \gamma_{s,t}(c) \tag{4}$$

$$F_c(s) = \sum_t \gamma_{s,t}(c) \cdot W_{s,t} \tag{5}$$

$$S_c(s) = diag\left(\sum_t \gamma_{s,t}(c) \cdot W_{s,t} \cdot W_{s,t}^*\right) \tag{6}$$

其中 $\gamma_{s,t}(c)$ 同 (3)，代表通用背景模型的第 c 個高斯分量給定 $W_{s,t}$ 之後驗機率，$diag()$ 則表示取矩陣中的主對角線元素。三個統計量中，$N_c(s)$ 又稱為零階統計量（0th statistics），格式為純量；$F_c(s)$ 又稱為一階統計量（1st statistics），格式為 $F \times 1$ 之向量；$S_c(s)$ 又稱為二階統計量（2nd statistics），格式為 $F \times F$ 之矩陣。接著，依據這些統計量，可求出特徵向量對於高斯分量 c、給定 W_s 的平均向量與共變異數矩陣之期望值：

$$E_c[x \mid W_s] = \frac{F_c(s)}{N_c(s)}$$

$$E_c[x \cdot x^* \mid W_s] = \frac{S_c(s)}{N_c(s)}$$

假設對於高斯分量 c，通用背景模型之參數以 $\{\pi_{0,c}, \mu_{0,c}, \Sigma_{0,c}\}$ 表示，則語者相關模型之參數 $\{\pi_{s,c}, \mu_{s,c}, \Sigma_{s,c}\}$ 便可以依據以下之公式進行調整：

$$\pi_{s,c} = \left(\frac{\alpha_c \cdot N_c(s)}{T_s} + (1-\alpha_c) \cdot \pi_{0,c}\right) \cdot \beta_s$$

$$\mu_{s,c} = \alpha_c \cdot E_c[x \mid W_s] + (1-\alpha_c) \cdot \mu_{0,c} \tag{7}$$

doi:10.6342/NTU201801578

$$\Sigma_{s,c} = \alpha_c \cdot E_c[x \cdot x^* \mid W_s] + (1-\alpha_c) \cdot (\Sigma_{0,c} + \mu_{0,c} \cdot \mu_{0,c}^*) - (\mu_{s,c} \cdot \mu_{s,c}^*)$$

其中，T_s 為語者 s 之特徵向量總數，β_s 之作用是在確保各高斯分量之權重參數 $\pi_{s,c}$ 總和為 1，而 α_c 定義如下：

$$\alpha_c = \frac{N_c(s)}{N_c(s) + r_r}$$

其中 r_r 扮演著控制通用背景模型的模型參數會被多大程度地往註冊語者的方向調整的相關因子。

雖然以似然率測定法計算出的分數可以作為判斷兩位語者相似度之依據，但高斯混合模型在本論文研究中僅扮演中間產物的角色。以下將介紹以 MFCC 特徵與高斯混合模型為基礎、近年來語者辨識領域最強健的特徵之一：i-向量。

doi:10.6342/NTU201801578

2.3 因子分析與 i-向量

i-向量（i-vector）為近年來效果最好、最常被運用在語者辨識相關工作上的特徵之一，亦是用於語者辨識的因子分析（Factor Analysis, FA）方法中集大成者，以下將簡介 i-向量之演變歷程、演算法以及其理論 [6] [7] [8]。

2.3.1 i-向量之概念與其演變歷程

MFCC 特徵是以音幀為單位抽取的特徵，換句話說，不同長度的聲音訊號，會抽出不同個數的 MFCC 特徵，而由於許多模型必須固定輸入特徵的長度，因此相較於 MFCC 特徵，以一整段聲音訊號或一位語者為單位抽取的特徵將更加適用。符合上述條件的辦法之一，是將以貝氏調適法得到的語者相關高斯混合模型的參數（例如每個高斯分量的平均）組成一個高維的超向量（supervector），來作為一位語者的特徵向量。但是，如此高維的特徵向量勢必帶有許多不必要的資訊，在運算上也較費時，面對這些問題，因子分析是一個非常有用的解決方法，它可以從高維且彼此相依的可觀變數（observable variables）中，將我們所感興趣的資訊，濃縮到較低維且彼此不相依的潛在變數（latent variables），進而幫助後續分析。

為方便後續說明，此處定義高斯混合模型之高斯分量總數為 C，MFCC 特徵之維度數為 $F(=39)$。最早成功為語者辨識工作對超向量進行因子分析的方法之一為本徵語音調適法（eigenvoice MAP），其概念為：一個語者相關的超向量，可被分解成「語者不相關項」與「語者相關項」兩部份，數學式表示為：

$$M_s = m + V \cdot y_s$$

其中 M_s 為與語者 s 相關之超向量，是以語者 s 之語料調適而成的可觀變數，大小為 $CF \times 1$；m 為通用背景模型所產生的超向量，大小為 $CF \times 1$；V 為一低秩之矩陣，大小為 $CF \times R_s$，而 R_s 將決定潛在變數之維度，通常設定 $R_s \ll CF$，

doi:10.6342/NTU201801578

而本徵語音就是由 V 的直行所組成；y_s 為潛在變數之向量，大小為 $R_s \times 1$，呈現標準常態分佈，其每一項元素又稱為語者因子（speaker factor）。很明顯地，等號右側的左項 m 即為「語者不相關項」，右項 $V \cdot y_s$ 則為「語者相關項」，且其中的 y_s 便可作為帶有濃縮語者資訊之新特徵。

回顧貝氏調適法用來更新高斯混合模型之平均的 (7) 式，若將等式右側之左項 $\alpha_c \cdot E_c[x \mid W_s]$ 視作「語者相關項」，右項 $(1-\alpha_c) \cdot \mu_{0,c}$ 則視作「語者不相關項」，整條式子便可仿照本徵語音調適法之模式修改為：

$$M_s = m + D \cdot z_s$$

其中 D 為一對角矩陣，大小為 $CF \times CF$，通常令 $D = \frac{1}{r}\Sigma$，Σ 為通用背景模型之超共變異數矩陣（super covariance matrix），由通用背景模型中所有共變異數矩陣以主對角線方向串接而成，即：

$$\Sigma = \begin{bmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_C \end{bmatrix}$$

其中 Σ_c 為對應到高斯分量 c 之共變異數矩陣，而 r 則為相關因子（relevance factor）；z_s 為潛在變數之向量，大小為 $CF \times 1$，呈現標準常態分佈，其每一項元素又稱為語者因子（speaker factor）。此方法又稱為古典調適法（classical MAP），古典調適法與本徵語音調適法之優缺點具有互相補償的性質：古典調適法需要較大量語料進行語者調適，而當語者調適的語料數量有限時，本徵語音調適法則表現得相對穩健，且訓練之收斂速度快；但本徵語音調適法則有者潛在變數 y_s 之維度 R_s 不能超過訓練語者的人數的缺點。

除了每個人的聲音彼此不相同，諸如錄音設備、錄音環境等因素所造成的信道效應，亦會影響信號之聲紋特性，進而對語者辨識結果產生不良影響。因此，若能消除信道效應，便能使得最後的特徵具有較純粹的語者資訊，而針對信道進

doi:10.6342/NTU201801578

行調試的因子分析方法：本徵信道調適法（eigenchannel MAP）便由此衍生。其概念為：一個語者、信道相關的超向量，可被分解成「信道不相關項」與「信道不相關項」兩部份，數學式表示為：

$$M_{s,h} = m + D \cdot z_s + U \cdot x_{s,h}$$

其中 $M_{s,h}$ 為與語者 s 、信道 h 相關之超向量，是以語者 s 的信道 h 語料調適而成的可觀變數，大小為 $CF \times 1$；$m + D \cdot z_s$ 與古典調適法相同，在此處則扮演「信道不相關項」之角色；剩下的 $U \cdot x_{s,h}$ 則為「信道相關項」，其中 U 為一低秩之矩陣，大小為 $CF \times R_c$，而 R_c 將決定信道潛在變數之維度，通常設定 $R_c \ll CF$，而本徵信道就是由 U 的直行所組成；$x_{s,h}$ 為信道潛在變數之向量，大小為 $R_c \times 1$，呈現標準常態分佈，其每一項元素又稱為信道因子（channel factor）。

為有效整合上述因子分析方法，聯合因子分析（Joint Factor Analysis, JFA）應運而生，該方法融合了本徵語音模型與本徵信到模型之概念，將語者、信道相關的超向量，分解為「語者、信道不相關項」、「語者相關項」、「信道相關項」與「殘留（residual）項」，數學式表示為：

$$M_{s,h} = m + V \cdot y_s + U \cdot x_{s,h} + D \cdot z_{s,h}$$

聯合因子分析在分析超向量時同時考慮了語者效應與信道效應，可說是將前述因子分析方法的優點集結於一身，而其語者因子向量 y_s 可以直接作為語者特徵輸入語者辨識模型並達到良好的效果。但是，後來的部分研究卻發現其信道因子向量 $x_{s,h}$ 除蘊含信道資訊外，亦帶有部分語者資訊，換言之，聯合因子分析的語者因子本身所帶有的語者資訊並不完整。

綜上所述，為擷取更有效的語者特徵，語者因子與信道因子必須同時被考慮，而建立在此概念上所誕生的新方法：i-向量，成為了目前效果最佳的因子分析方法。i-向量之全名為 identity vector（意為 vector for speaker-identification

doi:10.6342/NTU201801578

application）或 intermediate vector（vector with intermediate dimension between those of a supervector and an acoustic feature vector），不同於聯合因子分析模型以本徵語音矩陣與本徵信道矩陣分別定義了語者因子與信道因子之空間分佈，i-向量模型將語者空間與信道空間合而為一，並只產出一組潛在變數，其數學式表示為：

$$M_{s,h} = m + T \cdot w_{s,h}$$

其中 T 稱為總體變異矩陣（Total variability matrix），為一低秩之矩陣，大小為 $CF \times R_t$，而 R_t 將決定潛在變數之維度，通常設定 $R_t \ll CF$；$w_{s,h}$ 為潛在變數之向量，亦是我們要求的 i-向量，其大小為 $R_t \times 1$，呈現標準常態分佈，其每一項元素又稱為總體因子（total factor）。

2.3.2 i-向量模型之訓練流程

i-向量可說是目前因子分析系列方法之中的集大成者，以下將簡單介紹其總體變異矩陣之訓練方法，以及如何由一段聲音中擷取 i-向量特徵；須特別注意的是，以下過程中，同一語者之不同信道的聲音，會直接被視作不同語者的聲音，因此會將 i-向量原本之符號 $w_{s,h}$ 省略 h 而表示 w_s。在說明總體變異矩陣之訓練流程之前，此處首先定義一些數學符號。首先，給定一位語者 s 與其所產生的可觀變數 $W_s = \{W_{st} \mid t = 1,2,3,...\}$（即 MFCC 特徵向量），則對於每個高斯分量 c，分別定義 $N_c(s)$、$F_c(s)$、$S_c(s)$ 為零階、一階與二階鮑姆韋爾奇統計量，此處定義與 (4)、(5)、(6) 式相同。接著，定義標準化的一階、二階統計量：

$$\tilde{F}_c(s) = \sum_t \gamma_{s,t}(c) \cdot (W_{s,t} - m_c)$$

$$\tilde{S}_c(s) = diag\left(\sum_t \gamma_{s,t}(c) \cdot (W_{s,t} - m_c) \cdot (W_{s,t} - m_c)^* \right)$$

doi:10.6342/NTU201801578

其中 m_c 為 m 對應到第 c 個高斯分量之子向量，而 $\tilde{F}_c(s)$、$\tilde{S}_c(s)$ 之格式則分別與 $F_c(s)$、$S_c(s)$ 相同。若進一步推導上述數學式，則可以得到：

$$\tilde{F}_c(s) = F_c(s) - N_c(s) \cdot m_c$$

$$\tilde{S}_c(s) = S_c(s) - diag\left(F_c(s) \cdot m_c^* + m_c \cdot F_c(s)^* - N_c(s) \cdot m_c \cdot m_c^*\right)$$

最後，將上述統計量組合以形成更大的矩陣或向量：

$$\mathbf{N}(s) = \begin{bmatrix} N_1(s) \cdot I & & \\ & \ddots & \\ & & N_C(s) \cdot I \end{bmatrix}$$

$$\tilde{\mathbf{F}}(s) = \begin{bmatrix} \tilde{F}_1(s) \\ \vdots \\ \tilde{F}_C(s) \end{bmatrix}$$

$$\tilde{\mathbf{S}}(s) = \begin{bmatrix} \tilde{S}_1(s) & & \\ & \ddots & \\ & & \tilde{S}_C(s) \end{bmatrix}$$

其中 $\mathbf{N}(s)$ 為一 $CF \times CF$ 之矩陣，其內部由 $N_c(s) \cdot I$ 依主對角線方線串聯而成；$\tilde{\mathbf{F}}(s)$ 為一 $CF \times 1$ 之向量，其內部由 $\tilde{F}_c(s)$ 垂直串接而成；$\tilde{\mathbf{S}}(s)$ 為一 $CF \times CF$ 之矩陣，其內部由 $\tilde{S}_c(s)$ 依主對角線方線串聯而成。到此處即已完成了後續流程所需要用到的數學符號之定義。

總體變異矩陣與通用背景模型相同，亦是以最大似然度估計為原則，使用最大期望演算法進行訓練。在第一步驟中，總體變異矩陣內部之元素值可以隨機的方式進行初始化；而 E 步驟之主要目的為估計潛在變數，也就是 i-向量的期望值，此處定義對於一語者 s 與現有模型參數 Λ，其潛在變數 w_s 對於給定可觀變數 W_s 之後驗機率為 $P_\Lambda\left(w_s | W_s\right)$。已知 w_s 符合標準常態分佈，故依據貝氏理，可以得到：

doi:10.6342/NTU201801578

$$P_\Lambda\left(w_s \mid W_s\right) \propto P_\Lambda\left(W_s \mid w_s\right) \cdot \mathcal{N}\left(w_s \mid \mathbf{0}, I\right) \tag{8}$$

又依據 [6] 之 Theorem 1 之公式推導，可得到：

$$P_\Lambda\left(W_s \mid w_s\right) = G_\Sigma\left(s, m\right) + H_\Lambda\left(s, w_s\right) \tag{9}$$

其中，

$$G_\Sigma\left(s, m\right) = \sum_c N_c(s) \cdot \log\left(\frac{1}{(2\pi)^{F/2} \mid \Sigma_c \mid^{1/2}}\right) - \frac{1}{2} tr\left(\Sigma^{-1} \cdot \tilde{\mathbf{S}}(s)\right)$$

$$H_\Lambda\left(s, w_s\right) = w_s^* \cdot T^* \cdot \Sigma^{-1} \cdot \tilde{\mathbf{F}}(s) - \frac{1}{2} w_s^* \cdot T^* \cdot \mathbf{N}(s) \cdot \Sigma^{-1} \cdot T \cdot w_s$$

其中 $tr(\cdot)$ 代表取矩陣之跡數（trace），而總體變異矩陣 T 隨機設定其初始值。接著，結合 (8) 與 (9)，並定義：

$$L(s) = I + T^* \cdot \Sigma^{-1} \cdot \mathbf{N}(s) \cdot T$$

$$A(s) = L(s)^{-1} \cdot T^* \cdot \Sigma^{-1} \cdot \tilde{\mathbf{F}}(s)$$

則可以得到：

$$P_\Lambda\left(w_s \mid W_s\right) \propto \exp\left(w_s^* \cdot T^* \cdot \Sigma^{-1} \cdot \tilde{\mathbf{F}}(s) - \frac{1}{2} w_s^* \cdot L(s) \cdot w_s\right)$$

$$\propto \exp(-\frac{1}{2}(A - w_s)^* \cdot L(s) \cdot (A(s) - w_s))$$

由此乃可以推得 $P_\Lambda\left(w_s \mid W_s\right)$ 呈現平均為 $A(s)$、共變異數矩陣為 $L(s)^{-1}$ 之高斯分佈，而最初欲求之 i-向量之估計期望值即為 $A(s)$。

M 步驟之目的為更新模型，也就是總體變異矩陣之參數 T。首先這裡一樣先定義一些數學符號：

$$\mathfrak{A}_c = \sum_s N_c(s) \cdot E[w_s \cdot w_s^*]$$

doi:10.6342/NTU201801578

$$\mathfrak{C} = \sum_s \tilde{\mathbf{F}}(s) \cdot E(w_s^*)$$

$$\mathbf{N} = \sum_s \mathbf{N}(s)$$

其中，$E[\cdot]$ 代表求取期望值。又由 E 步驟所得到的結論，便可以進一步推得：

$$\mathfrak{A}_c = \sum_s N_c(s) \cdot L(s)^{-1}$$

$$\mathfrak{C} = \sum_s \tilde{\mathbf{F}}(s) \cdot A^*(s)$$

假設高斯分量 $c = 1,2,...,C$，MFCC 特徵之維度 $f = 1,2,...,F$，而對於每一 $i = (c-1)F + f$，以 T_i 與 $\mathfrak{C}_i$ 分別代表 T 與 $\mathfrak{C}$ 的第 i 列元素，則透過解出以下方程式，便能更新總體變異矩陣之參數 T：

$$T_i \cdot \mathfrak{A}_c = \mathfrak{C}_i$$

另外，超共變異數矩陣 Σ 也可以一併由下列之公式更新：

$$\Sigma = \mathbf{N}^{-1} \cdot \left(\sum_s \tilde{\mathbf{S}}(s) - diag(\mathfrak{C} \cdot T^*) \right)$$

依照上述兩步驟更新模型參數時，每次迭代都會確保訓練資料之對數似然度有所提升，以數學符號可以表示為：

$$\sum_s \log\left(P_\Lambda(W_s)\right) \geq \sum_s \log\left(P_{\Lambda_0}(W_s)\right)$$

其中，

$$\log\left(P_\Lambda(W_s)\right) = G_\Sigma\left(s, m\right) - \frac{1}{2}\log|L(s)| + \frac{1}{2}E[w_s^*] \cdot T^* \cdot \Sigma^{-1}$$

而模型參數 Λ_0 代表模型參數 Λ 更新前的狀態。有關上述所有公式之更詳細數學過程與證明，詳見 [6] 之 Theorem 4 與 Theorem 5。

doi:10.6342/NTU201801578

2.4 i-向量之降維與信道補償措施：線性判別分析

先前 2.3 章提及，總體變異矩陣的訓練過程中，同一語者之不同信道都被視作不同語者，因此擷取出來的 i-向量會仍會受信道效應干擾；而為了消弭信道效應，便必須對 i-向量進行信道補償（channel compensation）。信道補償的方法有非常多種，而最常見的一種就是以下將介紹的線性判別分析（Linear Discriminant Analysis, LDA）[9]。

線性判別分析是一種在統計學、模式識別、機器學習等領域皆經常被使用到的技術，一般是被用來進行特徵的降維，亦可以當成一種線性分類器。它可以藉由將特徵進行線性組合，從而找出一組彼此正交的新方向，而當原特徵被投影到由這些方向為基底組成的空間時，分屬不同類別的樣本能在新空間上被有效分離，而同一類別的樣本則會更加群聚，如圖 17 所示：

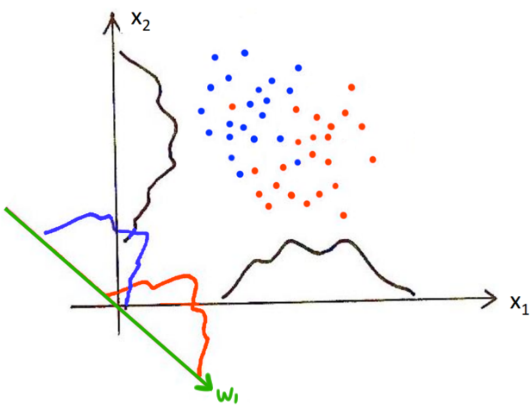


圖 17、線性判別分析示意圖

為方便說明，此處先定義一些符號：假設訓練資料中，共有 S 個類別，而每個類別 s 所包含之樣本點個數為 n_s，而各樣本點則以 $w_{s,h}$ 表示（$h = 1,2,...n_s$），則此處以 $\bar{w}_s$ 代表類別 s 內樣本點之平均、$\bar{w}$ 代表所有樣本點的平均，即：

$$\bar{w}_s = \frac{1}{n_s}\sum_{h=1}^{n_s} w_{s,h}$$

doi:10.6342/NTU201801578

$$\bar{w} = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{n_s}\sum_{i=h}^{n_s} w_{s,h}\right)$$

接著定義類別間變異性矩陣（between-class variance matrix）S_b 與類別內變異性矩陣（within-class variance matrix） S_w：

$$S_b = \frac{1}{S}\sum_{s=1}^{S}(\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})^*$$

$$S_w = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{1}{n_s}\sum_{i=h}^{n_s}(w_{s,h} - \bar{w}_s)(w_{s,h} - \bar{w}_s)^*\right)$$

我們可以推想，若不同類別間的樣本點在空間上彼此越分散，S_b 內的元素會傾向有越大的值；若而同一類別內的樣本點在空間上彼此越分散，S_w 內的元素則會傾向有越大的值。而線性判別分析的原理，是要透過將特徵投影到新空間上，使其類別間變異性與類別內變異性能分別達到最大化與最小化，若以 A 代表用來將特徵投影到新空間上的矩陣，則上述目標可以轉化成以下數學式：

$$A = \underset{W}{argmax}\frac{tr(W^* \cdot S_b \cdot W)}{tr(W^* \cdot S_w \cdot W)}$$

其中 A 為一 $F \times k$ 之矩陣，其中 F 代表原始特徵之維度，k 代表特徵降維後的維度。若將對應到矩陣 $S_w^{-1} \cdot S_b$ 第 i 大的本徵值（eigenvalue）之本徵向量（eigenvector）表示為 a_i，則將 a_1 至 a_k 併排所組成的矩陣，即為 A 之解析解：

$$A = [a_1, a_2, ..., a_k]$$

而任意樣本點 w 透過線性判別分析投影到新空間上的位置即為 $A^* \cdot w$ 。

doi:10.6342/NTU201801578

2.5 語者評分模型：機率性線性判別分析

為語者抽出 i-向量特徵後，最後一步即是為其算出分數，方能完成最後的判別。欲判斷兩個 i-向量是否屬於同一語者，最簡單的評分方式莫過於直接取兩者之標準化內積作為分數，即：

$$score_{CS}(w_1, w_2) = \frac{w_1 \cdot w_2}{|w_1||w_2|}$$

其中 w_1 與 w_2 為欲進行評分之兩 i-向量，此方法又稱為餘弦相似度（cosine similarity）；另一個效果較佳的方法是支持向量機（Support Vector Machine, SVM），超向量、i-向量或是其他因子分析之語者特徵搭配支持向量機所構成的語者分類器，在早些時期曾是最常被使用的的架構；而現今以 i-向量為基礎的語者辨識系統中，效果最佳、最主流的評分器，則是以下將要介紹的機率性線性判別分析（Probabilistic Linear Discriminant Analysis, PLDA） [10] [11] [12]。

2.5.1 機率性線性判別分析之模型訓練流程

機率性線性判別分析最早被應用於人臉辨識工作，而後來被發現其配合 i-向量作為評分器也能達到良好的辨識效果。用於其基礎概念與聯合因子分析類似，其假設一 i-向量可被分解成「語者、信道不相關項」、「語者相關項」、「信道相關項」與「殘留項」，數學式表示為：

$$w_{s,h} = \mu + \Phi \cdot \beta_s + \Gamma \cdot \alpha_{s,h} + \epsilon_{s,h} \tag{10}$$

其中 $w_{s,h}$ 為與語者 s、信道 h 相關之 i-向量，大小為 $R_t \times 1$；μ 為訓練資料中 i-向量之總體平均值，Φ 為一低秩之矩陣，與聯合因子分析中之本徵語音扮演相同角色，大小為 $R_t \times R_F$，而 R_F 將決定語者潛在變數之維度；β_s 為語者潛在變數之向量，大小為 $R_F \times 1$，呈現標準常態分佈，其每一項元素又稱為語者因子；Γ 為一低秩之矩陣，與聯合因子分析中之本徵信道扮演相同角色，大小為 $R_t \times R_G$，

doi:10.6342/NTU201801578

而 R_G 將決定信道潛在變數之維度；$\alpha_{s,h}$ 為信道潛在變數之向量，大小為 $R_G \times 1$，呈現標準常態分佈，其每一項元素又稱為信道因子；$\epsilon_{s,h}$ 則為殘留雜訊之向量，大小為 $R_t \times 1$，呈現平均為零向量、共變異數矩陣為 σ 的高斯分佈，其中 σ 為一對角矩陣。

機率性線性判別的模型參數與通用背景模型、聯合因子分析類似，亦是使用最大期望值演算法進行訓練，以下將對其進行簡介。首先一樣先定義數學符號，這邊先將兩項模型參數與兩項潛在變數各自合併為一項：

$$B = [\Phi \quad \Gamma]$$

$$z'_{s,h} = \begin{bmatrix} \beta_s \\ \alpha_{s,h} \end{bmatrix}$$

因此 (10) 可被改寫為：

$$w_{s,h} = \mu + B \cdot z'_{s,h} + \epsilon_{s,h}$$

在第一步驟中，兩項模型參數可以隨機的方式進行初始化；而透過 E 步驟，可以算出潛在變數 z'_s 之期望值與共變異數矩陣，即：

$$E[z'_{s,h}] = (B^* \cdot \sigma^{-1} \cdot B + I)^{-1} \cdot B^* \cdot \sigma^{-1} \cdot (w_{s,h} - \mu)$$

$$E[z'_{s,h} \cdot y'^{*}_{s,h}] = (B^* \cdot \sigma^{-1} \cdot B^* + I)^{-1} + E[z'_{s,h}] \cdot E[y'_{s,h}]^*$$

而透過 M 步驟，則可以計算出模型參數與殘留雜訊之共變異數矩陣：

$$B = \left(\sum_{s,h} (w_{s,h} - \mu) \cdot E[z'_{s,h}]^* \right) \cdot \left(\sum_{s,h} E[z_{s,h} \cdot z^*_{s,h}] \right)^{-1}$$

$$\sigma = \frac{1}{\sum_s n_s} \sum_{s,h} diag\left((w_{s,h} - \mu) \cdot (w_{s,h} - \mu)^* - B \cdot E[z'_{s,h}](w_{s,h} - \mu)^* \right)$$

其中 n_s 為語者 s 所含有之信道數量。以上述流程輪流執行 E 步驟與 M 步驟直到收斂，即可完成機率性線性判別之模型參數的訓練。關於更詳細的訓練過程數學

doi:10.6342/NTU201801578

推導，可參見 [10]。另外，由於此模型實際在進行評分時不會用到信道相關項，因此若將殘留雜訊之共變異數矩陣 σ 假設為全矩陣，則可以在不影響辨識結果的前提下將信道相關項直接省略，而原來之 (10) 式即變為：

$$w = \mu + \Phi \cdot \beta + \epsilon$$

在前一段落中，語者潛在變數 β_s 與殘留雜訊之共變異數矩陣 σ 被假設為符合高斯分佈，因此此模型又稱為高斯—機率性線性判別分析（Gaussian Probabilistic Linear Discriminant Analysis, G-PLDA）；在前些時期，部分研究顯示若以重尾分佈（Heavy-tailed distribution）來模擬 β_s 與 σ 之機率分佈，將能達到更好的辨識效果，而此模型又稱為重尾—機率性線性判別分析（Heavy-Tailed Gaussian Probabilistic Linear Discriminant Analysis, HT-PLDA）；但隨後，其他研究則又發現：若將 i-向量經過徑向高斯化（radial Gaussianization）處理，高斯—機率性線性判別分析即可與重尾—機率性線性判別分析達到相同的效果，又因為前者之運算速率較高，因此其乃成為現今最主流的評分模型。徑向高斯化分為兩個步驟，第一步是將 i-向量進行線性白化轉換（linear whitening transformation ），由於重尾分佈屬於橢圓對稱密度（Elliptically Symmetric Densities, ESD）家族，而高斯分佈屬於球形對稱密度（Spherically Symmetric Density, SSD）家族，因此透過線性白化轉換即可將橢圓對稱密度轉換為球形對稱密度；第二步是則是將其長度分佈進行非線性柱狀圖變形（non-linear histogram warping），其過程是將 i-向量之長度（$|w|$）之累積分佈通過反向累積卡方分佈（inverse cumulative Chi distribution）函數，而此步驟之簡化版的則是直接將各 i-向量之長度統一標準化為 1 即可。關於徑向高斯化之更詳細原理可參見 [13]。

doi:10.6342/NTU201801578

2.5.2 機率性線性判別分析之評分流程

在為兩個 i-向量 w_1 與 w_2 抽出語者潛在變數後，機率性線性判別分析進行評分的原理是以「兩個潛在變數來自同一 i-向量之機率」與「兩個潛在變數來自不同 i-向量之機率」之比值來進行評分，其數學式表示為：

$$score_{PLDA}(w_1, w_2) = \frac{p(w_1, w_2 \mid \mathscr{H}_s)}{p(w_1 \mid \mathscr{H}_d) \cdot p(w_2 \mid \mathscr{H}_d)} \tag{11}$$

其中 $\mathscr{H}_s$ 代表兩個 i-向量會產生相同潛在變數之假說，$\mathscr{H}_d$ 則代表兩個 i-向量會產生不同潛在變數之假說。而在 i-向量符合常態分佈之前提下，則可將方才之分數取對數而改寫為：

$$score_{PLDA}(w_1, w_2) = \log \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_{tot} & \sigma_{ac} \\ \sigma_{ac} & \sigma_{tot} \end{bmatrix} \right) - \log \mathcal{N}\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_{tot} & \mathbf{0} \\ \mathbf{0} & \sigma_{tot} \end{bmatrix} \right)$$

其中，

$$\sigma_{tot} = \Phi \cdot \Phi^* + \sigma$$

$$\sigma_{tot} = \Phi \cdot \Phi^*$$

又若假設 μ 為零向量，且令：

$$Q = \sigma_{tot}^{-1} - (\sigma_{tot} - \sigma_{ac} \cdot \sigma_{tot}^{-1} \cdot \sigma_{ac})^{-1}$$

$$P = \sigma_{tot}^{-1} \cdot \sigma_{ac} \cdot (\sigma_{tot} - \sigma_{ac} \cdot \sigma_{tot}^{-1} \cdot \sigma_{ac})^{-1}$$

則可以將 (11) 式展開為：

$$score_{PLDA}(w_1, w_2) = w_1^* \cdot Q \cdot w_1 + w_2^* \cdot Q \cdot w_2 + 2w_1^* \cdot P \cdot w_2 + const \tag{12}$$

假設 Φ 為一大小為 $D \times K$ 之矩陣（$D > K$），P 便可基於其對稱性用奇異值分解（Singular Value Decomposition, SVD）分解為三個矩陣之乘積，即：

doi:10.6342/NTU201801578

$$P = U_K \cdot \tilde{P} \cdot U_K^*$$

其中 U_K 是由 K 行互相正交的向量所組成，這些向量乃是 P 的本徵向量；而 $\tilde{P}$ 則是由 P 的本徵值由主對角線方向串接而成的對角矩陣，即：

$$\tilde{P} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{bmatrix}$$

其中 λ_k 乃是矩陣 P 的第 k 大本徵值。又因為 P 與 Q 之秩數與 Φ 相同，(12) 式可以用更快的方法計算而得：

$$score_{PLDA}(w_1, w_2) = \tilde{w}_1^* \cdot \tilde{Q} \cdot \tilde{w}_1 + \tilde{w}_2^* \cdot \tilde{Q} \cdot \tilde{w}_2 + 2\tilde{w}_1^* \cdot \tilde{P} \cdot \tilde{w}_2 + const \quad (13)$$

其中，

$$\tilde{Q} = U_K^* \cdot Q \cdot U_K$$

$$\tilde{w}_i = U_K^* \cdot w_i,\ i = 1,2$$

並且，當 K/D 之值越小，(13) 式便可較 (12) 式擁有更快的運算速率。關於機率性線性判別分析之評分方式的更詳細原理，可參見 [12]。

doi:10.6342/NTU201801578

第 3 章 實驗結果

3.1 實驗語料與資料配置

本論文使用了 NIST SRE 2010 與 TIMIT 兩份語料庫分別進行實驗，以下將簡介此二語料庫與資料配置。

3.1.1 NIST SRE 2010 語料庫與資料配置簡介

NIST SRE 2010 [14] 是一場在 2010 年由美國國家標準技術研究所（National Institute of Standards and Technology, NIST）所舉辦的語者辨識（Speaker Recognition Evaluation, SRE）比賽，本論文取其訓練資料中的 3 種資料類型：「10sec」、「core」與「8conv」，並除去重複的語者後進行實驗，其中「10sec」為長度約 10 秒的電話對話錄音，一共有 2,257 位（男 959／女 1,298）語者，每位語者提供 1 份錄音檔；「core」為長度約 3 至 15 分鐘的電話對話錄音或麥克風錄製的面試錄音，一共有 4,004 位（男 1,817／女 2,187）語者，每位語者提供 1 份錄音檔；「8conv」為長度約 2 至 3 分鐘的電話對話錄音，一共有 445 位（男 194／女 251）語者，每位語者提供 8 份錄音檔。3 種資料類型之音檔取樣頻率為 8,000 赫茲、樣本位元數皆為 16。

資料配置方面，訓練資料是以 3 種資料類型各取 2／3 語者數量，一共 4,469 位（男 1,965／女 2,504）語者的集合；測試資料則是剩下的 1／3 語者數量，一共 2,237 位（男 1,005／女 1,232）語者的集合。而關於測試資料中註冊語料與驗證語料的分配，每位語者皆僅會註冊 1 組語者模型，其中「8conv」的每組模型皆以 7 份音檔註冊而成，剩下 1 份音檔作為驗證音檔；而由於「10sec」與「core」的每位語者皆只有 1 份音檔，因此這裡將每份音檔分割為 4 等份，每

doi:10.6342/NTU201801578

組模型皆以 3 份音檔註冊而成，剩下 1 份音檔作為驗證音檔。測試資料的每一位語者皆會自驗證語料中挑選 1,000 份音檔來測試自己的語者模型，1,000 份音檔中其中 1 份是會是正樣本（屬於此語者的聲音），而其他 999 份則是隨機挑選的負樣本（不屬於此語者的聲音），因此所有欲評測的樣本一共有

$2{,}237 \times 1{,}000 = 2{,}237{,}000$ 份。

3.1.2 TIMIT 語料庫與資料配置簡介

TIMIT [15] 是由德州儀器（Texas Instruments, TI）、麻省理工學院（Massachusetts Institute of Technology, MIT）與斯坦福國際研究院（SRI International）聯合提供的語料庫，其一共包含了 630 位（男 438／女 192）語者。這些語者的腔調包含了美國 8 種主要地方方言腔調，每位語者皆提供 10 份錄音檔，每份錄音檔之取樣頻率為 16,000 赫茲、樣本位元數為 16，錄音內容則為指定內容、包含豐富音素（phoneme）的一句話，每句話平均長度約為 3 秒，最長不超過 8 秒。

資料配置方面，訓練資料是以男性與女性各取 2／3 語者數量，一共 420 位（男 292／女 128）語者的集合；測試資料則是剩下的 1／3 語者數量，一共 210 位（男 146／女 64）語者的集合。而關於測試資料中註冊語料與驗證語料的分配，每位語者皆會註冊 2 組語者模型，每組模型皆以 4 份音檔註冊而成，而剩下的 2 份音檔則作為測試音檔。測試資料的每一位語者皆會自驗證語料中挑選 100 份音檔來測試自己的語者模型，100 份音檔中其中 1 份是會是正樣本，而其他 99 份則是隨機挑選的負樣本，因此所有欲評測的樣本一共有

$210 \times 2 \times 100 = 42{,}000$ 份。

doi:10.6342/NTU201801578

3.2 實驗一：對照組

實驗一將介紹本論文設定的對照組流程與實驗結果，此實驗為完全不使用語者性別資訊的傳統語者驗證方法，亦為本論文之後所有實驗之比較依據。本實驗之流程架構如圖 18 所示：

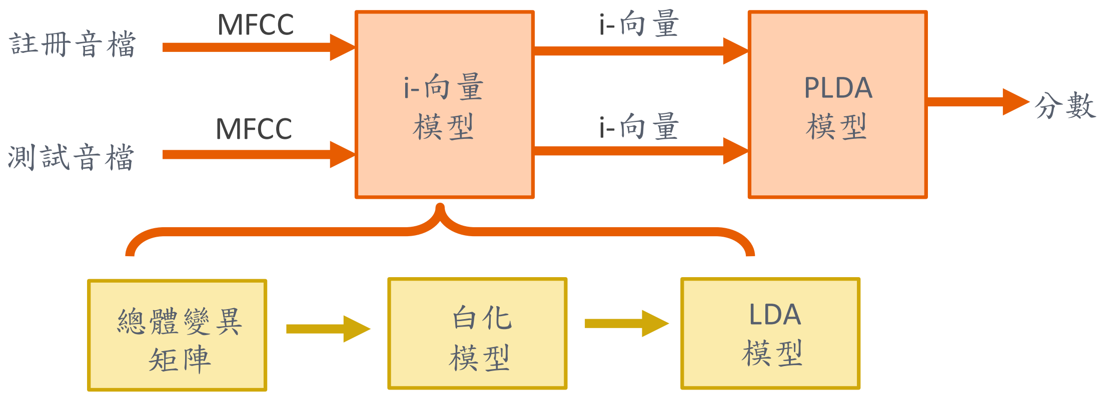


圖 18、對照組之流程架構

流程圖中之各項模型以 Bob toolbox [16] 進行實作。由圖中可見，註冊音檔與測試音檔分別抽取 MFCC 特徵後，會經由 i-向量模型各自抽取 i-向量，最後經由 PLDA 模型算出分數；其中，此處所述之 i-向量模型裡除了包含原本用以將超向量轉化為 i-向量的總體變異矩陣之外，亦包涵了白化轉換模型與線性判別分析模型。表 1 為對照組流程所使用的模型參數，而本論文其他語者驗證相關流程之模型參數設定亦與對照組相同：

表 1、語者驗證系統模型參數設

Model Parameter		Value
MFCC extraction	Frame size	512 sample points
	Hop size	128 sample points
UBM model	Number of mixture	128
	Number of training iteration	10
i-vector model	Dimension of i-vector	300
	Number of training iteration	5
LDA model	Reduced Dimension of i-vector	50
PLDA model	Dimension of speaker factor	50
	Dimension of channel factor	50
	Number of training iteration	10

doi:10.6342/NTU201801578

對照組流程的實驗結果如圖 19 所示：

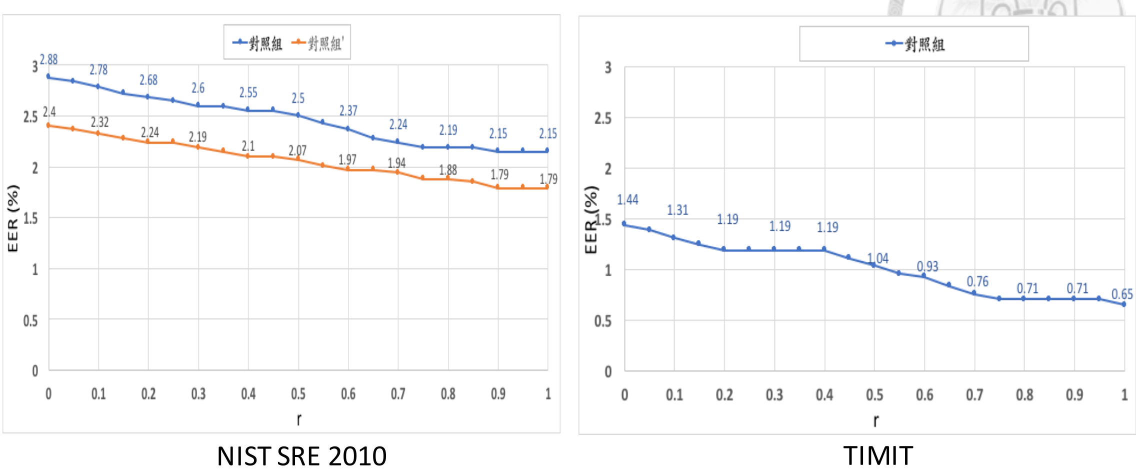


圖 19、對照組流程之結果

圖之左側與右側分別為 NIST SRE 2010 語料庫與 TIMIT 語料庫之實驗結果，兩張圖的縱軸為語者驗證系統之等錯誤率，單位為百分比；橫軸為 r，此 r 值代表的是「與註冊語者不同性別的仿冒者比例」，以下將簡單說明此實驗設計之意義。

在章節 3.2 中提及，本論文之資料配置形式為每一位註冊語者配對 1 份自己的音檔作為正樣本，並搭配許多其他仿冒者的聲音作為負樣本，而挑選仿冒者的一項重要考量，便是這些語者的性別。今假設語者驗證系統架設於使用者之行動裝置上，則可以考慮兩種情景，如圖 20 所示：

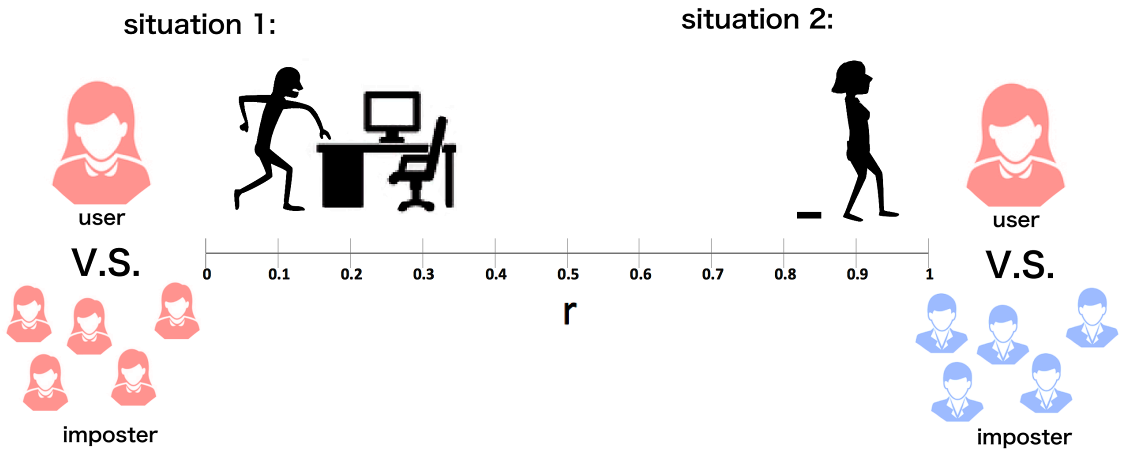


圖 20、不同的 r 值與系統應用情境

doi:10.6342/NTU201801578

第一，當使用者到了教室或是工作場所時，若有同事或同學意圖在未經該使用者同意下存取其行動裝置，由於他們皆熟悉該使用者的聲音，因此勢必會找聲音與該使用者相像的人—至少會與該使用者同性別—來解鎖其語者驗證系統，在此情境下，我們可以假設絕大部分的仿冒者皆會和使用者同性別（事實上，NIST SRE 2010 比賽亦是設計成每個樣本的註冊方與測試方的性別皆相同，因此與此情境相同）。第二，當使用者的行動裝置（被陌生人）偷竊或不慎遺失時，由於拿到其行動裝置的人並不知曉該使用者的性別或聲音特性，因此此種情境下便不能假設仿冒者會和該使用者之性別相同，仿冒者和該使用者的性別相同的機率在此時只有 1／2，甚至更低。綜合以上兩種情境，我們便必須確保語者驗證系統在不同的 r 值，也就是不同的「仿冒者與使用者不同性別的機率」的情況下，皆應有良好的表現。因此，本論文所有的語者驗證相關實驗，皆會在不同的 r 值之下測量系統的表現。

在圖 19 左圖中有兩條線，兩者之差別在於音幀大小的不同。藍線是在音幀的樣本點數上與 TIMIT 語料統一，即一個音幀中含有 512 個樣本點，時間長度為 64 ms；橘線則是在音幀的時間長度上與 TIMIT 語料統一，即一個音幀中含有 256 個樣本點，時間長度為 32 ms。由圖中可見，由於橘線的總音幀數會較多，因此各模型之訓練效果較好、實驗結果亦較佳，但由於音幀數較多導致訓練過程曠日費時，因此往後的實驗皆使用與藍線相同的設定。

由圖 19 中可見，雖然對照組流程於 TIMIT 語料與 NIST SRE 2010 語料上的表現有明顯差距，但兩者的變化趨勢是相同的：從最左側的 r 等於 0 慢慢往最右側的 r 等於 1 移動時，系統的等錯誤率皆呈現近似線性下降的趨勢。造成此下降趨勢之原因其實非常直觀：與使用者不同性別的仿冒者，相較於同性別者，大多會與該使用者在聲紋特性上有較大的差別，造成其被系統拒絕的機率也相對高出許多，因此當 r 值越大，正樣本所得到的分數與其他負樣本相比便會相對高出許多，進而得到較低的等錯誤率。

doi:10.6342/NTU201801578

3.3 實驗二：性別資訊應用方法—當性別分類器表現良好

實驗二將介紹若干種將性別資訊應用於語者驗證系統以提升其辨識率之方法，包括 [17] 所介紹的一般最常見作法，以及本論文提出的另外 2 種針對前者弱點所進行之改良方法。另外，本實驗之目的在於證明在性別分類器表現良好之前提下，藉由有效運用性別資訊，便能提升系統辨識率；因此本實驗中對於所有原本應假定為未知性別的測試語者，皆以其真實性別代替性別分類器的預測結果。

3.3.1 前人作法（一般常見作法）

本方法是參考了論文 [17] 而實作的流程，其流程架構如圖 21 所示：

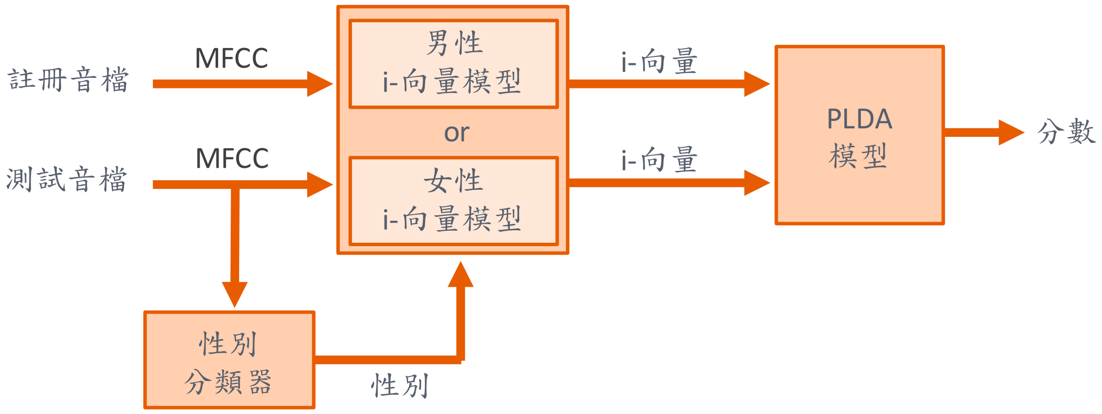


圖 21、前人作法之流程架構圖

其流程基本上與對照組流程類似，兩者不一樣的地方為：對照組之 i-向量模型是使用訓練資料中所有語者共同訓練而得來，而本作法之 i-向量模型則分為 2 個性別相關的模型，分別是使用訓練語者中的男性語者與女性語者進行訓練而得來，而所有註冊音檔或測試音檔，皆必須使用對應其語者性別之模型來抽取 i-向量。其中，註冊語者的性別可以假設是已知的，因為在實際運用情境中，我們可以要求所有使用者要註冊系統時提供自己的性別資訊；但是由於測試語者的性別是未知的，因此測試音檔必須先藉由性別分類器預測其性別，才能套入正確性別的 i-向量模型來抽取 i-向量。

doi:10.6342/NTU201801578

本做法的實驗結果如圖 22 所示：

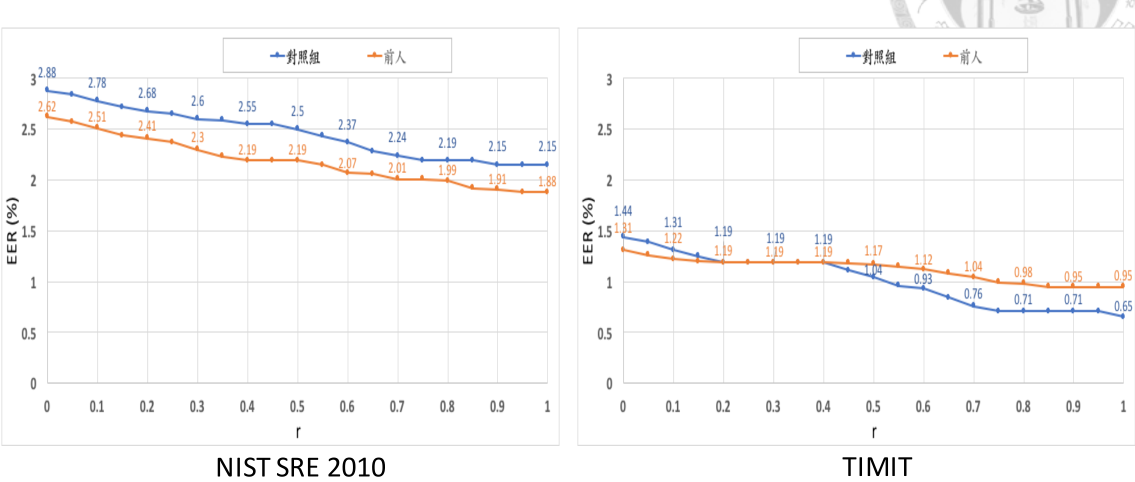


圖 22、前人作法之結果

同樣地，圖之左側與右側分別為 NIST SRE 2010 語料庫與 TIMIT 語料庫之實驗結果，兩張圖的藍線為方才對照組流程的結果，橘線則是本做法的結果。由圖中可以看到，在 NIST SRE 2010 語料中，本做法的結果從 r = 0 至 r = 1 始終都較對照組作法要進步，證明了其實用性；但反觀 TIMIT 語料，其實驗結果在最左側 r = 0 時尚較對照組進步一些，但慢慢往右側 r = 1 方向移動時，由於其等錯誤率的下降速率較對照組慢上許多，導致其結果逐漸被對照組超越，甚至被拉開距離。

由於許多語者驗證比賽的樣本（如 NIST SRE 2010）並不包含註冊語者與測試語者性別不同的情形，因此許多以本作法實作的語者驗證系統（如 [17]），便沒能對此現象做進一部探討或進行補償措施。而本研究推測，造成上述現象的主因，是來自於本做法的一個缺陷：使用性別相關模型抽取的男性 i-向量與女性 i-向量，在空間分佈上可能存在重疊現象。圖 23 為 i-向量之理想二維點分佈圖：

doi:10.6342/NTU201801578

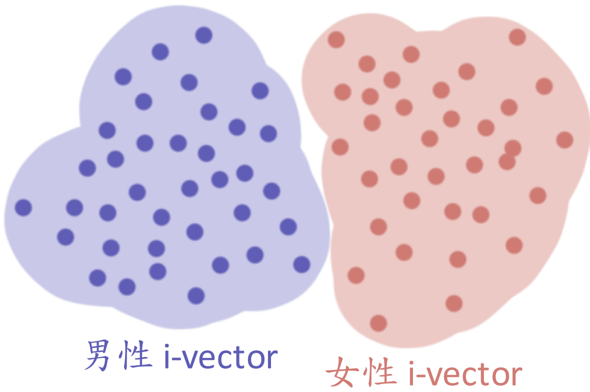


圖 23、i-向量之理想二維點分佈圖

在理想上，因為男性與女性的聲紋特性存在很大的差異，容易讓人以為 i-向量會在空間分佈上呈現男女各自群聚，且容易界定出兩群 i-向量之邊界的情形，正如同圖 23 所示。在此假設下，當一個樣本的註冊方與測試方擁有不同性別，因為兩者的 i-向量之間的距離很大，因此這個樣本便相當容易被系統拒絕。但實際情況中卻並非一定如此，尤其就男／女性 i-向量模型在訓練過程中並沒有接觸過女／男性的資料這個事實來看，我們更必須懷疑此假設之合理性。圖 24 左右兩圖是分別使用主成分分析（Principal Components Analysis, PCA）和 t-分佈隨機相鄰嵌入（T-distributed Stochastic Neighbor Embedding, TSNE），將訓練資料中的 i-向量降至二維平面所繪製的點分佈圖：

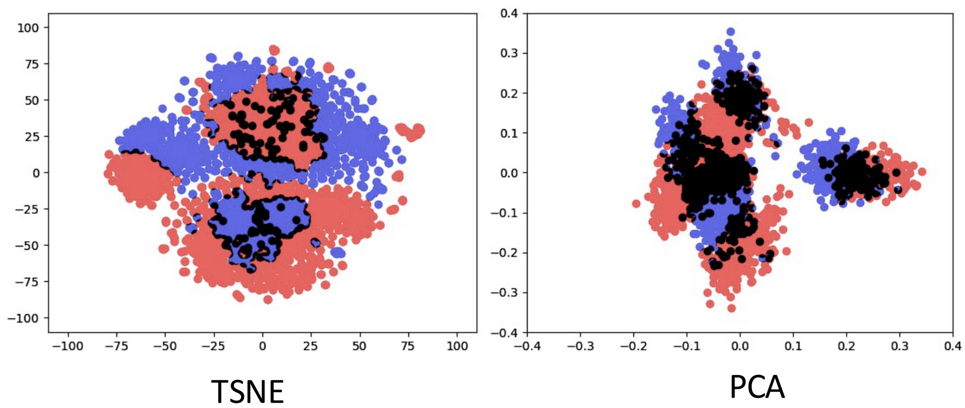


圖 24、i-向量之實際二維點分佈圖—以 TSNE 與 PCA 將 i-向量降為二維

doi:10.6342/NTU201801578

由兩張圖我們能清楚看到，兩個性別的 i-向量群在空間分佈上確實存在重疊的情形。這種現象會造成語者驗證系統的錯誤率上升，尤其在當欲測試的樣本中含有大量註冊方與測試方不同性別的情形，即 r 有較大的值時更為顯著。為方便討論，此處假設系統之評分器僅以註冊方與測試方 i-向量之間的歐氏距離作為判斷是否為同一語者之唯一標準。在不分男、女資料所訓練出來的混合 i-向量模型，在訓練資料足夠的前提下，我們可以假定其會盡量將不同語者的 i-向量在空間上彼此分開；但是，正如方才所述，由於男／女性 i-向量模型在訓練過程中並沒有接觸過女／男性的資料，因此便很有可能產生特定男性語者與女性語者的 i-向量距離過度相近的現象，使得系統容易將他們誤判為同一語者。此現象在訓練流程中，會使得系統的評分器—機率線性判別分析模型容易被混淆；在測試流程中，也容易會造成不同性別間語者被錯誤判斷成同一語者的比例較對照組流程還要高的現象。因此，針對因不同性別 i-向量空間重疊現象所造成的上述缺陷，本論文接下來將提出兩種簡單的改良機制，以令結果獲得改善。

3.3.2 方法一：拒絕與註冊者性別不符之測試者

第一種改良方法之流程架構如圖 25 所示：

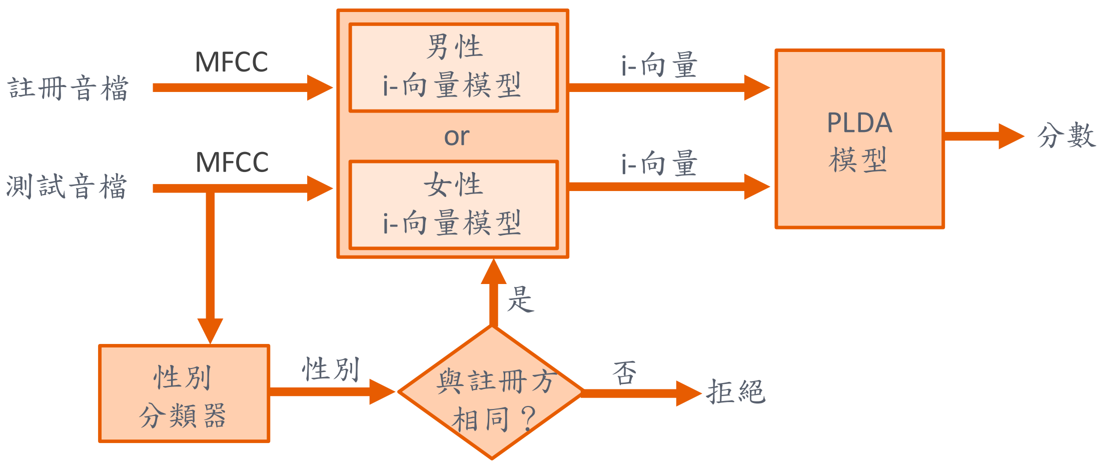
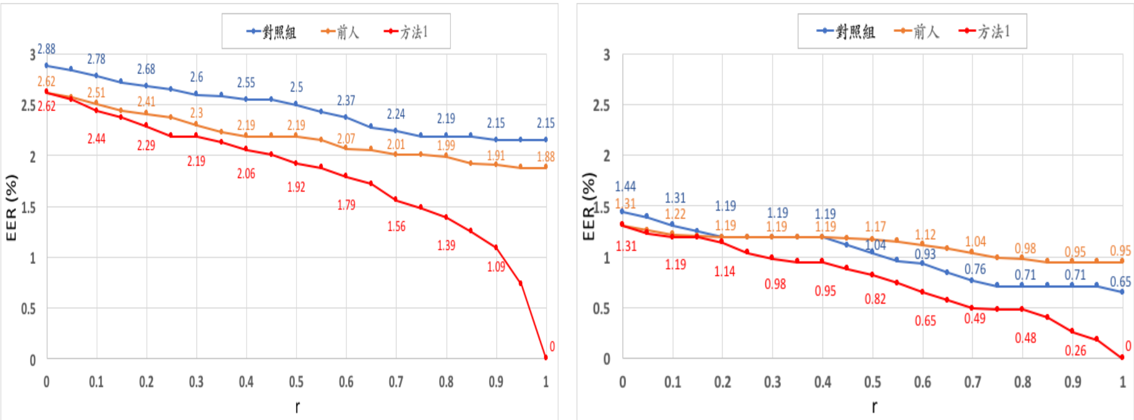


圖 25、方法一之流程架構圖

doi:10.6342/NTU201801578

此版本之流程與前人作法類似，不同之處在於其多出了一個步驟：在以性別分類器判斷出測試語者之性別後，須檢視其是否與註冊語者之性別相同，若相同，則繼續進行原本的流程—以性別相關模型抽出雙方的 i-向量之後，以機率性線性判別分析模型計算出分數；若不同，則直接拒絕此測試語者（打上非常低的分數），因為不同性別的聲音不可能屬於同一位語者。如此一來，便能消除在前人作法中因不同性別的 i-向量在空間分佈上的重疊現象而造成的錯誤接受率攀升之問題。

方法一流程的實驗結果如圖 26 所示：



NIST SRE 2010　　TIMIT

圖 26、方法一流程之結果

同樣地，藍線與橘線分別為對照組流程與前人作法流程之實驗結果，而新增的紅線則是本方法一流程之實驗結果。在圖中，我們可以看到在最左側 r＝0 的時候，因為沒有註冊語者與測試語者性別不同的樣本，因此方法一與前人做法的表現是相同的；但慢慢往右側 r＝1 方向移動時，由於註冊方與測試方性別不同的樣本皆被系統拒絕，因此方法一在兩種語料的等錯誤率始終低於對照組，甚至越往右便越與對照組的結果拉開距離；到最後 r＝1 時，由於所有錯誤樣本皆成功被系統拒絕，等錯誤率便可達到 0%。此實驗結果除證明了前人做法確實存在著本論文所預測的缺陷與造成它的原因，亦證明了方法一流程在改善此種缺陷上的效果。但是，方才提及前人做法的缺陷來自於男女 i-向量的重疊現象，而此方法實際上並沒有在 i-向量之空間分布上將分離男性語者與女性語者，因此仍可能會

doi:10.6342/NTU201801578

使機率性線性判別分析模型在訓練過程中被混淆，造成訓練成效下降，可說是治標不治本的改善方法。

3.3.3 方法二：將性別作為新的一維特徵

為了從根本上解決前人作法之缺陷，本論文提出了第二種改良方法，其流程架構如圖 27 所示：

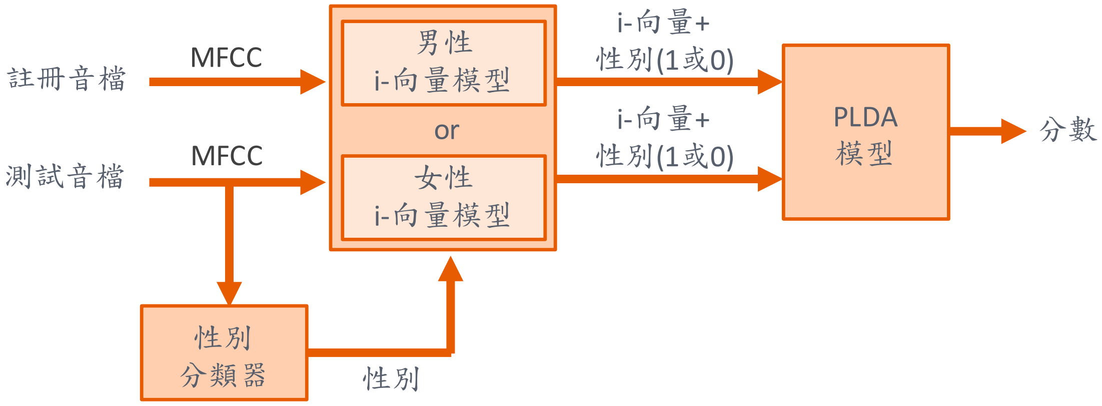
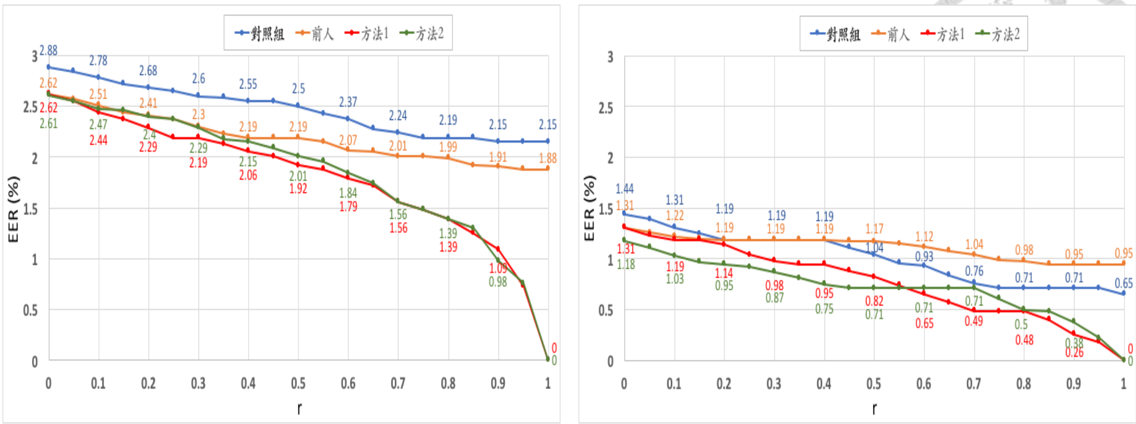


圖 27、方法二之流程架構圖

此版本之流程語前人作法亦十分相似，不同之處在於其在 i-向量上多加了新的一維特徵，使得 i-向量由原本的 50 維變為 51 維。此新一維特徵之目的是用來代表性別資訊，實際表示方法為：若此 i-向量的來源為男性，則此維特徵之值為 1，反之若為女性則為 0。由於新增的此維特徵，男性和女性的 i-向量在空間分佈上便完全沒有互相重疊的現象，因此無論是前人做法之缺陷造成的錯誤接受或錯誤拒絕升高的問題，在此流程下都能同時獲得解決。

方法二流程的實驗結果如圖 28 所示：

doi:10.6342/NTU201801578



NIST SRE 2010　　　　TIMIT

圖 28、方法二流程之結果

同樣地，藍線、橘線與紅線分別代表對照組、前人作法與方法一的結果，而新增的綠線則是本方法二流程之實驗結果。由圖中可以見，方法二流程在兩種語料上相較於方法一的表現並不相同：在 NIST SRE 2010 語料上，方法二在 r 值較小的時候，幾乎都表現得比方法一略差，在 r 值較大時才差不多追平；但在 TIMIT 語料上，方法二在 r 值較小時的表現反倒略優於方法一，但在 r 值較大時卻被方法一超越。此兩種相異的表現，此處推測是資料特性不同或是隨機誤差所致，至於本作法沒能如預期地大幅較方法一進步，此處推測可能是因為訓練資料中距離過於相近之異性別 i-向量畢竟只佔少部分之比例，因此並沒有大幅地影響到機率性線性判別分析模型的訓練成效。不過就結論而言，方法一和方法二在不同 r 值之整體表現上，等錯誤率皆較對照組與前人做法降低許多，因此此兩種流程在「性別分類器表現良好」的前提下作為修正前人作法缺陷的方法，確實皆被證明是可行的。

doi:10.6342/NTU201801578

3.4 實驗三：性別分類器實作

在實驗二中，所有實驗流程皆是基於性別分類器表現良好為前提所設計，因此原本應假定為未知性別的測試語者，皆以其真實性別代替性別分類器的預測結果；而本章節將實際進行性別分類器之實作，並分析各分類器之優劣。本實驗所實作的性別分類器共有 3 種，每種分類器皆是以性別不相關之 i-向量（即章節 3.2 中對照組流程所使用的不分性別模型所產生之 i-向量）作為特徵，以下將各別簡介其原理與結果。

3.4.1 線性判別分析性別分類器

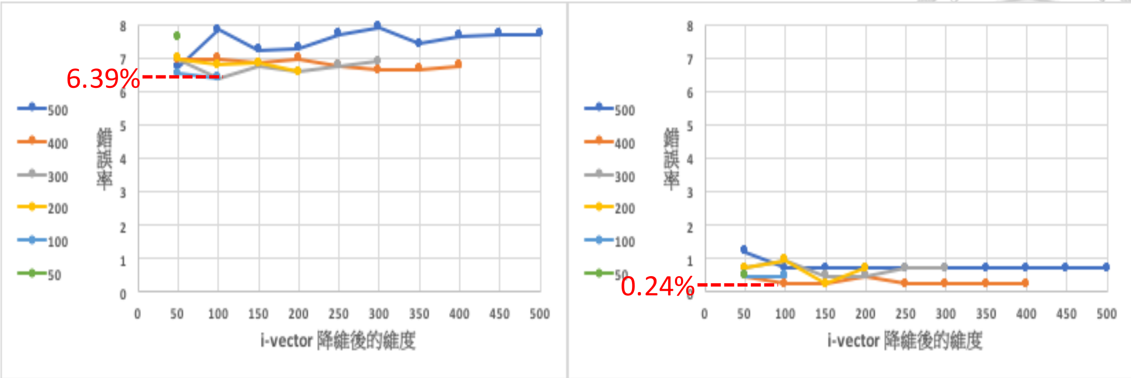
此分類器之作法是參考了論文 [18] 所實作的流程，其流程圖如圖 29 所示：

圖 29、線性判別分析性別分類器之流程架構

由圖中可見，經由 MFCC 特徵轉換而來的 i-向量，會以線性判別分析模型作為二元分類器，判斷音檔的性別；其中此線性判別分析模型的訓練過程，是以性別作為標籤（意即將所有訓練資料分為男、女兩個類別），將 i-向量降至單維作為性別分數，而性別分數之門檻值則定為令訓練資料達到等錯誤率，即男性與女性之錯誤率相同時之分數。需注意的是，本流程所使用的線性判別分析模型，與 i-向量模型本身所包含的線性判別分析模型是不同的模型，兩者之差別在於前者在訓練過程中是以「性別」為標籤，並作為分類器使用；而後者則是以「語者」作為標籤，並作為降維器使用。

本分類器之實驗結果如圖 30 所示：

doi:10.6342/NTU201801578



NIST SRE 2010　　　　TIMIT

圖 30、線性判別分析性別分類器之結果

同樣地，圖之左側與右側分別為 NIST SRE 2010 與 TIMIT 語料庫之實驗結果，而本分類器之實驗過程調整了包含於 i-向量模型中的兩個參數：i-向量之初始維度與經（i-向量模型中的）線性判別分析模型降維後的維度，兩者分別在圖 30 中以不同顏色的線與圖之橫軸表示，而圖之縱軸則為錯誤率，單位為百分比。圖中的紅色虛線所標記的是兩種語料庫中最佳參數所表現出的結果，分別為 6.39% 與 0.24%。值得一提的是，在 TIMIT 語料庫中，0.24% 的錯誤率代表在其 420 份測試音檔中，只有一份音檔的性別被分類錯誤，其他音檔的性別分類結果則皆是正確的。

3.4.2 支持向量機性別分類器

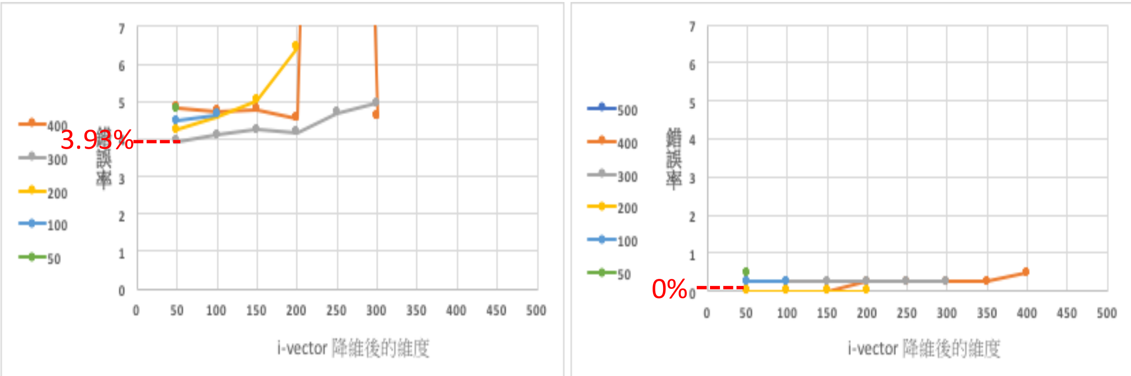
此分類器之流程圖如圖 31 所示：

圖 31、支持向量機性別分類器之流程架構圖

doi:10.6342/NTU201801578

由圖中可見，經由 MFCC 特徵轉換而來的 i-向量，會以支持向量機作為二元分類器，判斷音檔的性別；其中，支持向量機類別權重設定為平衡（balenced class weight），此設定會讓訓練過程中每個樣本的成本（cost）之權重反比於其所屬之類別的樣本總數。

本分類器之實驗結果如圖 32 所示：



NIST SRE 2010　　　　TIMIT

圖 32、支持向量機性別分類器之結果

本分類器之實驗與前者相同，調整了 i-向量模型中的 i-向量之初始維度與其降維後的維度，兩者分別在圖 32 中以不同顏色的線與圖之橫軸表示；此外，亦調整了支持向量機中最重要的 2 個參數：cost 與 gamma。在圖中，每一個數據點皆是將 cost 與 gamma 分別在集合 {0.1, 1, 10, 100, 1000} 與 {0.01, 0.1, 1, 10, 100} 內進行排列組合，並以網格搜尋法（grid search）所尋找到的最佳參數組合之結果。同樣地，圖中的紅色虛線所標記的是兩種語料庫中最佳參數所表現出的結果，NIST SRE 2010 語料庫的錯誤率可達 3.93%，而 TIMIT 語料庫則可以達到完全分類正確，即 0% 錯誤率的結果。

doi:10.6342/NTU201801578

3.4.3 神經網路性別分類器

此分類器之流程圖如圖 33 所示：

圖 33、神經網路性別分類器之流程架構

由圖中可見，經由 MFCC 特徵轉換而來的 i-向量，會以神經網路模型作為二元分類器，判斷音檔的性別；而神經網路模型的參數（keras API）如表 2 所示：

表 2、神經網路性別分類器之模型參數

模型參數	層參數
batch _size = 128	kernel_initializer='glorot_normal'
epochs = 200	bias_initializer='zeros'
loss = binary_crossentropy	BatchNormalization()
optimizer = Adam(lr=3e-4)	Activation('relu')
metrics = 'accuracy'	Dropout(0.35)

其中，模型的每層節點數量由輸入至輸出端線性減少，舉例來說，若 i-向量維度數為 n，神經網路維中間層數為 4 層，則模型之各層結點數則如圖 34 所示：

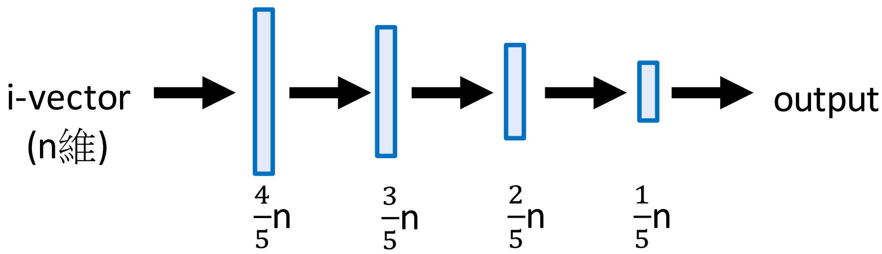
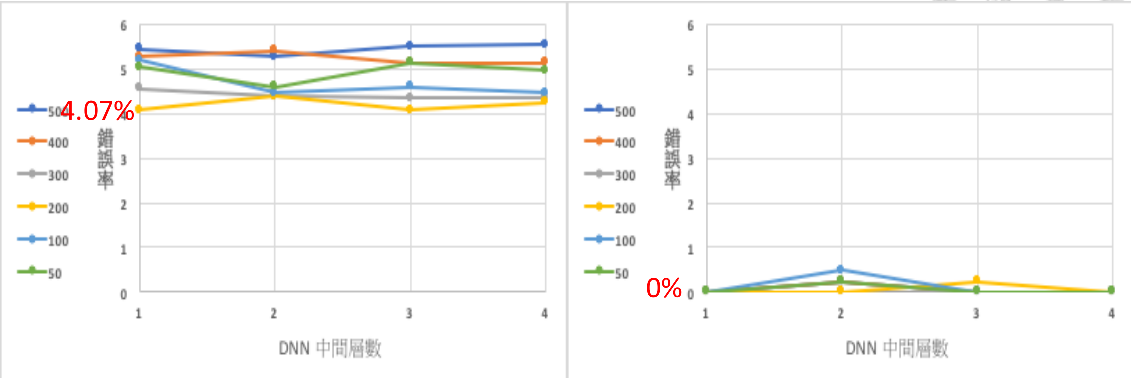


圖 34、神經網路模型之形狀

本分類器之實驗結果如圖 35 所示：

doi:10.6342/NTU201801578



NIST SRE 2010　　　　TIMIT

圖 35、神經網路性別分類器之結果

與之前的實驗不同的是，本實驗所使用的 i-向量一律維持初始維度，而不使用線性判別分析模型進行降維，而本分類器之實驗過程調整的兩個參數則改為是 i-向量之初始維度與神經網路模型的中間層數，兩者分別在圖 35 中以不同顏色的線與圖之橫軸表示。同樣地，圖中的紅色虛線所標記的是兩種語料庫中最佳參數所表現出的結果，NIST SRE 2010 語料庫的錯誤率可達 4.07%，而 TIMIT 語料庫則與支持向量機性別分類器一樣，可以達到完全分類正確，即 0% 錯誤率的結果。

doi:10.6342/NTU201801578

3.5 實驗四：性別資訊應用方法 — 當性別分類器表現不良

實驗四同樣將介紹若干種性別資訊應用於語者驗證系統之方法，但不同於實驗二以其真實性別代替性別分類器的預測結果，本實驗將使用在實驗三中實作的性別分類器，並特別針對當性別分類器表現不良的情況，介紹其他不同於實驗二的性別資訊應用方法。

在實驗三中，因 TIMIT 語料庫在性別分類器上可以達到零錯誤率的效果，故其於語者驗證流程中套用實驗三所實作的性別分類器之結果，將與實驗二無異，因此本章節之實驗將僅包含 NIST SRE 2010 語料的部分。圖 36 是對照組流程與將在實驗三中於 NIST SRE 2010 語料中表現最好的支持向量機性別分類器套用於實驗二各流程之實驗結果：

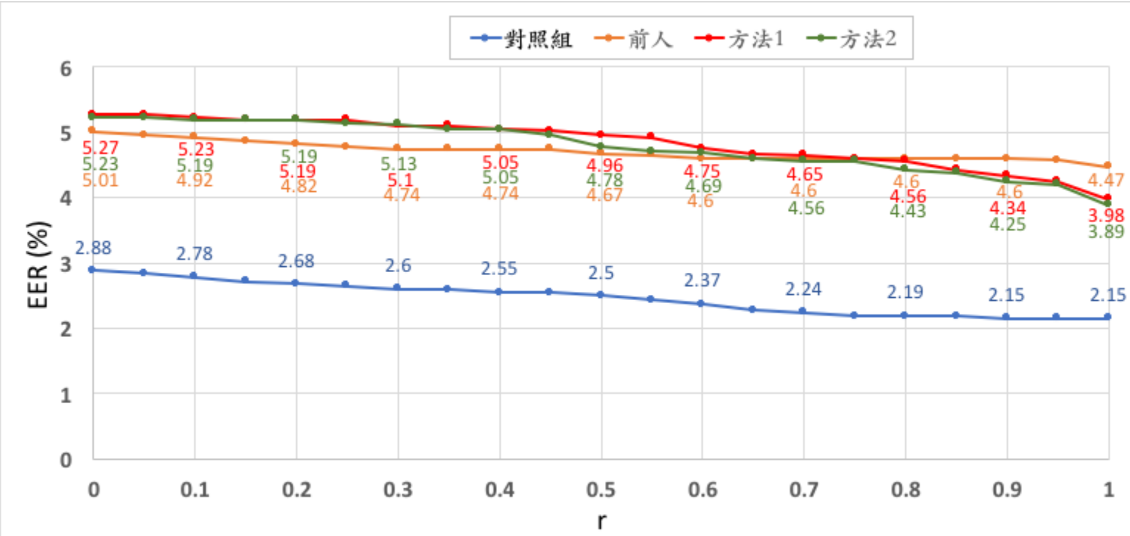


圖 36、各項語者驗證流程使用支持向量機性別分類器之實驗結果

同樣地，藍線、橘線、紅線與綠線分別代表對照組、前人作法、方法一與方法二的結果，由圖中可以見到，在實際套用真實性別分類器之後，實驗二中的所有流程在不同 r 值下的實驗結果產生了明顯的退步，甚至比對照組流程之結果更差。雖支持向量機性別分類器之錯誤率僅有 3.93%，但這些少量的錯誤卻會造成語者驗證系統之表現造成巨大影響，其原因一樣可分為錯誤接受與錯誤拒絕兩種情形來探討：我們可以想像，當註冊語者與測試語者為不同人時，即使性別分類器誤判了測試語者的性別，系統錯誤接受此負樣本的機率仍然不高。但當註冊語者與

doi:10.6342/NTU201801578

測試語者為同一人，一旦測試語者被誤判性別，使用前人作法的系統會分別以不同性別之模型為註冊方與測試方抽取 i-向量，此種情形幾乎必然會造成系統為此正樣本評出較非常低的分數；而使用方法一流程之系統，在檢測出註冊語者與測試語者分屬不同性別後，便會對此樣本予以拒絕；最後，使用方法二流程之系統，由於訓練資料中同一語者的 i-向量中用以代表性別的最後一維特徵皆有相同的數值，因此當評分器見到註冊語者與測試語者在該維特徵之數值相異，自然同樣會評出非常低的分數。綜上所述，實驗二的 3 種方法皆會為大量正樣本評出低落的分數，進而使得等錯誤率跟著大幅提升。

3.5.1 性別分類器的合成

在「使用真實性別」與「使用預測性別」的實驗結果之間所存在的巨大差距，顯示當前所實作的性別分類器在準確度上尚未達到能使用的標準，而欲使實驗二的各項方法真正具備實用性，勢必需要效果更精良的性別分類器。最簡單的提升性別分類器準確度方法之一，便是結合多個性別分類器，共同產生預測結果。此方法能否發揮效果與各性別分類器之間是否具互補性有關，換句話說，當各個性別分類器所誤判的樣本越少重複，此方法越有可能表現出超越單一分類器之效能。圖 37 為各單一分類器誤判樣本的比例分佈：

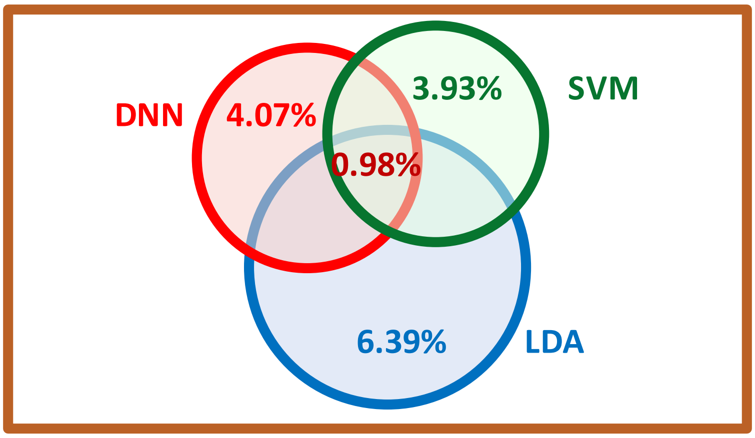


圖 37、各性別分類器之誤判樣本比例分佈

由圖中可見，在測試語者中，線性判別分析（藍色圓圈）、支持向量機（綠色圓圈）與神經網路（紅色圓圈）性別分類器之錯誤率各為 6.39%、3.93% 與

doi:10.6342/NTU201801578

4.07%，而同時被此 3 種性別分類器誤判的語者，僅佔全體之 0.98%，遠小於其他單一性別分類器之錯誤率，說明此種合成方法在理論上是可行的。

此處一共嘗試了 3 種不同的性別分類器合成方法：第一種作法為「投票」，對於一位測試語者，選擇最多性別分類器對其預測的性別作為預測結果，其數學式可表示為：

$$score_{tot}(s) = \frac{1}{3} prd_{lda}(s) + \frac{1}{3} prd_{svm}(s) + \frac{1}{3} prd_{dnn}(s)$$

其中，$score_{tot}(s)$ 為合成性別分類器對測試語者 s 所計算之性別總分數，$prd_m(s)$ 為分類器 m 對其預測之性別（男性為 1、女性為 0），若性別總分數大於 0.5，則判定此語者 s 為男性，否則判定為女性。第二種作法為「權重投票」，其原理類似於前者，但每種分類器的預測結果會以不同權重進行加權以決定最終預測結果，其數學式可表示為：

$$score_{tot}(s) = w_{lda} \cdot prd_{lda}(s) + w_{svm} \cdot prd_{svm}(s) + w_{dnn} \cdot prd_{dnn}(s)$$

其中 w_m 為分類器 m 所佔有之權重，每個權重是以邏輯回歸法（logistic regression）訓練而來。第三種方法為「權重分數投票」，其原理類似於前者，會將每種分類器所提供的性別分數組以不同權重進行加權而決定最終結果，其中，線性判別性別分類器所提供的性別分數即為被降維至單維的 i-向量，而支持向量機與神經網路性別分類器的性別分數則是其預測一測試語者屬於男性與屬於女性的兩個機率值，其數學式可表示為：

$$score_{tot}(s) = \sum_i w_{lda_i} \cdot score_{lda_i}(s) + \sum_j w_{svm_j} \cdot score_{svm_j}(s) + \sum_k w_{dnn_k} \cdot score_{dnn_k}(s)$$

其中，$score_{m_i}(s)$ 代表分類器 m 對測試語者 s 所計算的第 i 份性別分數，w_{m_i} 則代表其對應之權重，同樣地，權重亦是以邏輯回歸訓練而來。此三種合成性別分類器方法之結果如表 3 所示：

doi:10.6342/NTU201801578

合成方法	錯誤率
投票	3.80 %
權重投票	3.80 %
權重分數投票	3.49 %

表 3、各性別分類器合成方法之實驗節果

由表中可見，雖然三種方法的錯誤率皆較單一性別分類器還低，但進步效果十分有限，若將其中表現最佳的權重分數投票版本之性別分類器套用至實驗二的各方法，則可以得到如圖 38 之結果：

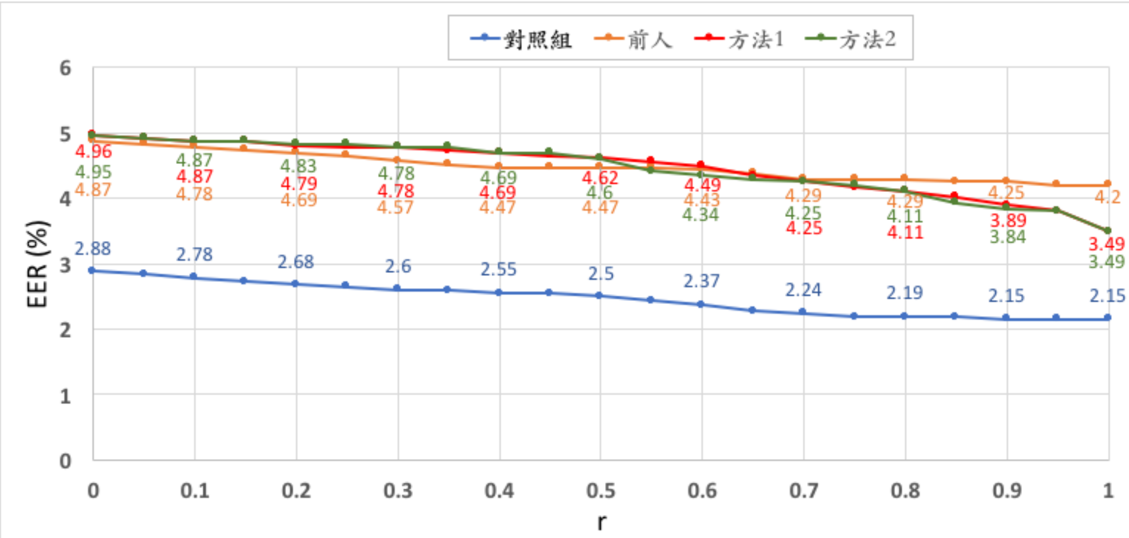


圖 38、各項語者驗證流程使用「權重分數投票」性別分類器之實驗結果

由圖中可見，雖然各流程之等錯誤率隨著性別分類器之辨識率上升而有所下降，但仍然與對照組的結果有一大段差距。此外，此種合成性別分類器之作法還存在著一大爭議：性別分類器在語者驗證系統當中只扮演著輔助的角色，但此種合成性別分類器必須額外使用 3 個 i-向量模型與分類器各自抽取 i-向量並進行性別預測，換句話說，此種做法花在性別分類工作的時間將會是語者辨識工作本身的 3 倍，可說是本末倒置，因此，就算存在其他更好的性別分類器合成方法能使辨識率進一步提高，此種作法也勢必會引人詬病。綜上所述，本研究將在以下提出不同於實驗二的另外 3 種改良流程，試圖讓系統在單一性別分類器表現不良的情況之下，仍能超越對照組作法之結果。

doi:10.6342/NTU201801578

3.5.2 方法三：將「性別分數」作為新的特徵

在先前提及的方法二流程裡，一旦系統偵測到註冊方與測試方用以代表性別資訊的 i-向量最後一維特徵之數值（0 或 1）不同，便會為此樣本評出非常低之分數，進而造成了被誤判性別的正樣本皆會被錯誤拒絕的情形。這不禁令人好奇，除了 0 或 1，性別資訊的數值沒有其他種表示方式嗎？此問題衍生了接下來將要介紹的方法三流程，其流程架構如圖 39 所示：

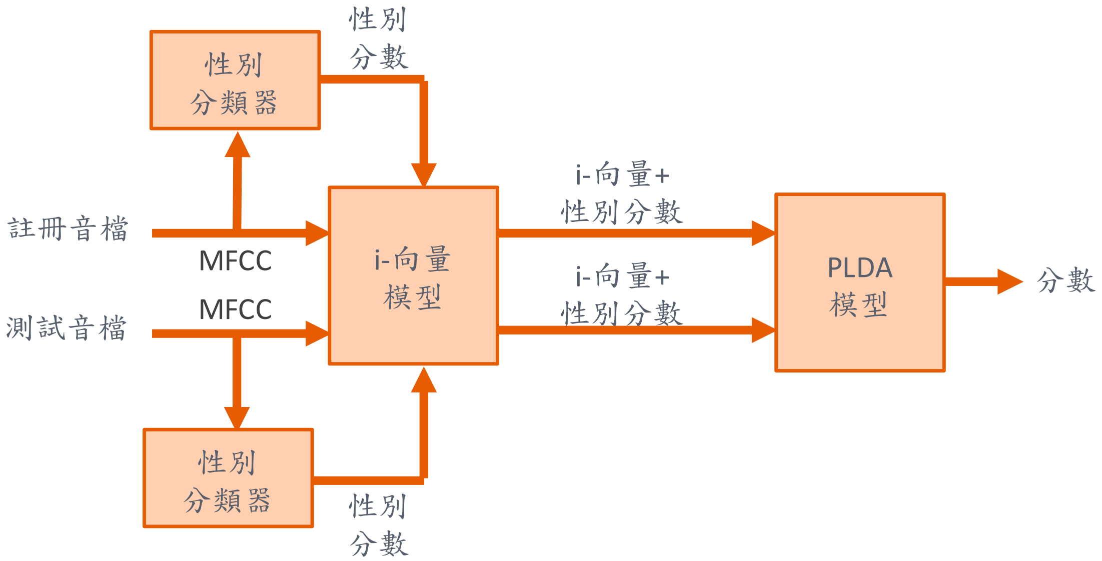


圖 39、方法三之流程架構圖

由圖中可見，本流程以前人作法流程為基礎，並模仿方法二之方式於 i-向量尾端附加性別資訊，但此處之性別資訊並不使用 0 或 1，而是仿照權重分數投票版本之合成性別分類器當中所使用的性別分數，如此一來，便能避免掉方才提及的錯誤拒絕之問題。另外，為避免重蹈實驗二各流程的覆轍，本流程亦避開了前人做法的「性別相關模型」和方法 1 的「對於雙方性別不同之樣本予以拒絕」等作法，回到與對照組類似的架構。

方法三流程的實驗結果如圖 40 所示：

doi:10.6342/NTU201801578

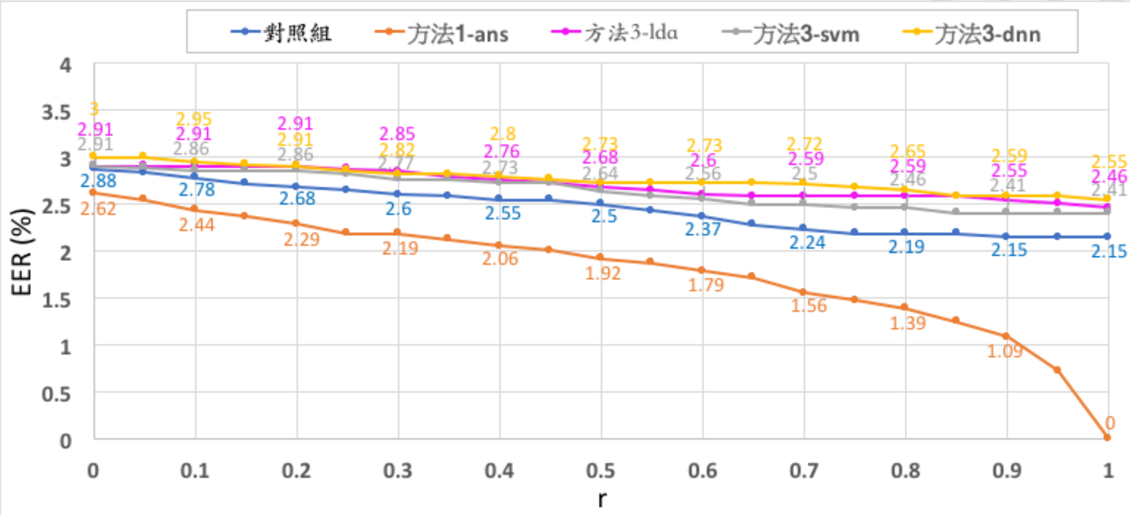


圖 40、方法三流程之結果

圖中，藍線代表對照組流程之結果，橘線則代表方法一流程在使用真實性別代替性別分類器預測之結果。原則上，實驗四之目標是在使用真實性別分類器的前提下，實作出比藍線更加進步的改良版本，但是由於本實驗接下來的改良流程皆是使用效果不精良之性別分類器對測試語者進行性別預測，因此這些改良流程的結果在預期上將難以超越圖中的橘線，意即橘線可被視為實驗四所有改良流程之上限。圖中另外 3 條粉紅色、灰色與黃色的線，分別代表以此方法三流程套用線性判別分析、支持向量機、神經網路性別分類器之結果，雖然此 3 條線已較先前各改良流程進步許多，但遺憾地是，這些實驗結果仍舊比對照組流程要差。本流程失敗的原因，除了推測是性別分類器在產生性別分數上的成效並不理想之外，亦必須考慮使用性別不相關模型取代性別相關模型所帶來的退步。

3.5.3 方法四：檢查註冊音檔的性別分類準確率

在先前的實驗中，我們可以見到圖 36 與圖 26 之差距來自於少數遭誤判性別的測試語者被實驗二之各流程系統錯誤拒絕。倘若我們能將這些測試語者遭誤判性別的樣本與其他樣本分別交由對照組流程與實驗二之改良流程進行判別，理論上便能得到近似於實驗二之效果。若欲實際執行此概念，要如何預測性別分類器能否正確預測測試語者之性別，必然是此方法必須面對的第一個問題，而方法四流程將提供一個較簡單的預測方法：藉由同一語者之註冊音檔來判斷測試音檔被

doi:10.6342/NTU201801578

誤判性別之可能性，以下為其流程架構圖：

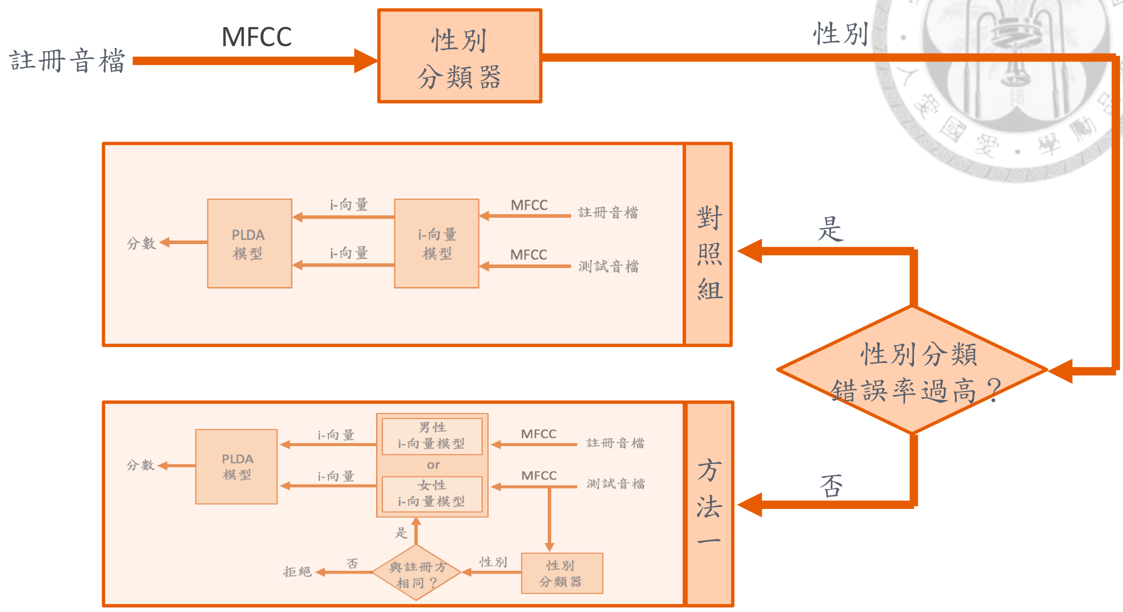


圖 41、方法四之流程架構圖

由圖中可見，當一位語者使用本系統進行註冊，系統會以性別分類器預測其註冊音檔之性別，而此處所使用的性別分類器為實驗三當中表現最好的支持向量機性別分類器。由於此位語者必須提供其真實性別，我們便能藉此測定性別分類器在此位語者身上之表現：若性別分類器在此語者之註冊音檔中出現過多次誤判情形，則代表性別分類器在此位語者身上表現較差，或者此位語者之「聲音性別」並不典型，例如此語者可能是一位音調偏低的女性，或音調偏高的男性。針對此類型之語者，我們可以預料到其未來在作為測試語者時，亦會有很高的機率會被誤判性別，進而無法通過自己以正確性別註冊的語者模型，即錯誤拒絕的情況。因此，當這類型的語者在樣本中作為註冊方時，便將其交由對照組流程進行判別，而其他情況則使用在實驗二中表現最好的方法一流程進行判別。

方法四之前半部流程基本上可被視為一種二元分類器，此分類器之目的在於預測測試語者之性別是否會被誤判，而其演算法則為：只要一語者滿足「註冊音檔之性別誤判率過高」之條件，便判斷其測試音檔將被誤判性別。既然是二元分類器，我們便能測定其召回率（recall）、精確率（precision）與 F-度量，如圖 42 所示：

doi:10.6342/NTU201801578

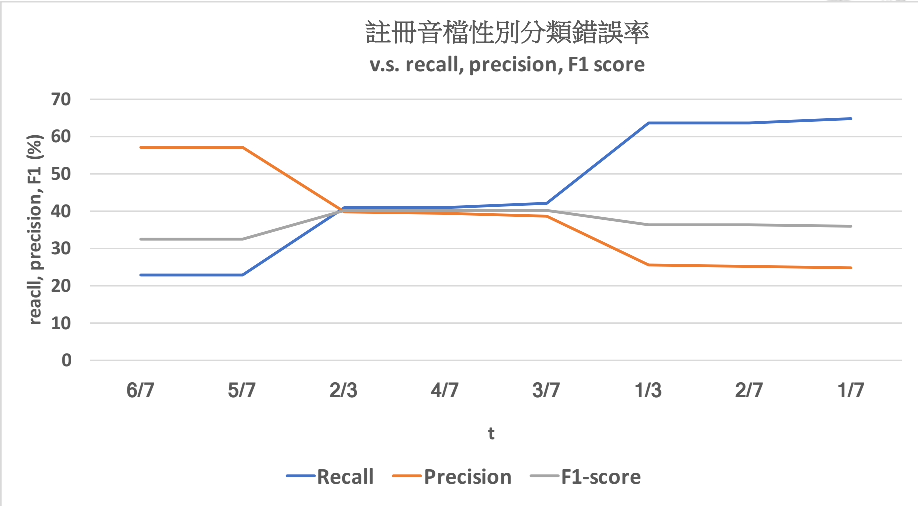


圖 42、方法四流程之「測試音檔性別誤判預測器」之效果

圖 42 之橫軸 t ，此 t 值代表的意義為註冊音檔性別分類錯誤率之門檻值，意即：若一語者在註冊音檔之性別分類錯誤率大於或等於此 t 值，便導入對照組流程，否則導入方法一流程。其中，因本實驗所使用的資料配置中，註冊語者只有 3 或 7 份註冊音檔，因此圖中橫軸的 t 值便只有 n／3 或 n／7 所組合起來的有限種可能，其中 n 為小於等於其分母之正整數；而此「測試音檔性別誤判預測器」之召回率、精確率與 F-度量則分別以藍線、橘線與灰線表示，圖之縱軸則為其百分比。我們可以想像，當 t 值減小時，真陽性（true positive）之樣本增加，但真正被誤判性別的語者總數是不變的，因此圖 42 中的召回率便呈現出嚴格遞增之現象；但偽陽性（false positive）之樣本增加速率會比真陽性更快，因此圖 42 中的精確率便呈現出嚴格遞減之現象；至於 F-度量，則是在 t 之值偏中間的時候有較大的值。

然而，該如何選定最合適之 t 值，則必須實際檢視系統之整體表現，如圖 43 所示：

doi:10.6342/NTU201801578

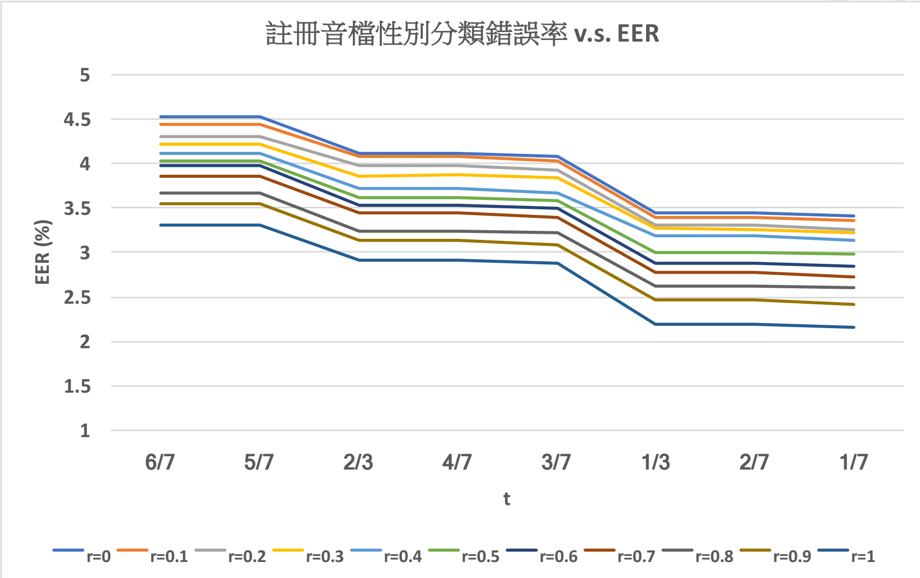


圖 43、方法四流程之選定註冊音檔性別誤判率門檻

圖 43 之橫軸同樣 t，縱軸為等錯誤率，至於不同的 r 值，即與註冊語者不同性別的仿冒者比例，則以不同顏色的線表示。在理論上，若註冊音檔之性別誤判率之門檻被設定太高，便無法有效選出測試音檔被誤判性別的語者；反之若設定得太低，雖然可以篩選出大量測試音檔被誤判性別的語者，但因為有過多沒有被誤判的語者將同時被導入對照組流程，反而會造成進步效果不彰。依此論點，圖 43 的各條曲線在預期上應該會與圖 42 中的 F-度量成反比，呈現出兩側較高、中間較低的現象，而最低點則是最合適之門檻，並且不同曲線應該能得到不同的最佳門檻值。但實際上，實驗結果顯示，每條曲線皆呈現出嚴格遞減之現象，並在最右側達到最低點，也就是說，在真實情況中，任何語者只要有一份註冊音檔被誤判性別，便必須導入對照組流程，才能達到最佳效果。

圖 44 為方法四流程指定 t 值為 1／7 時之實驗結果：

doi:10.6342/NTU201801578

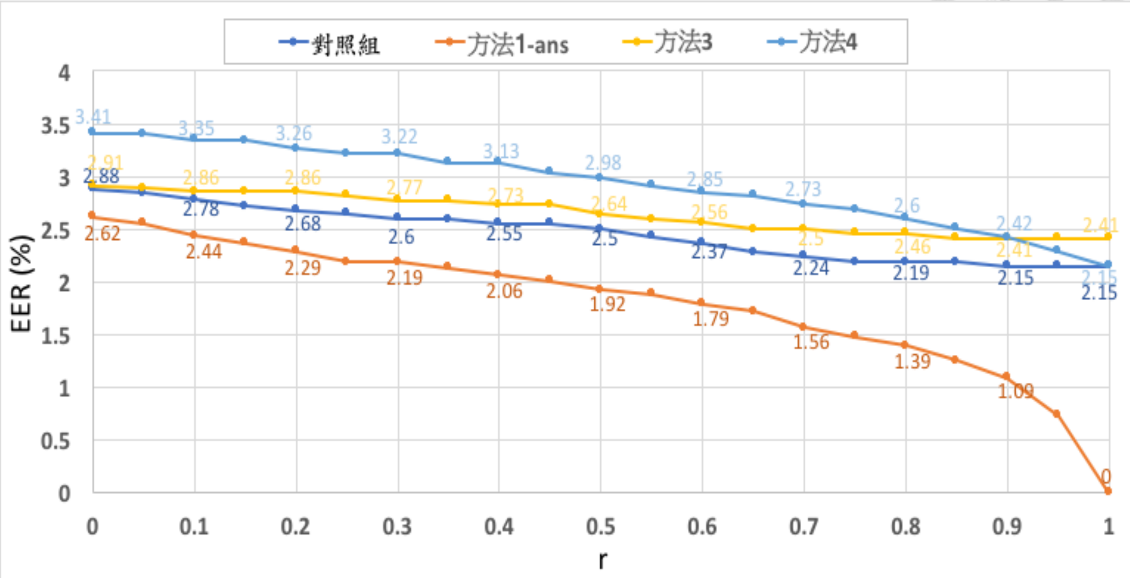


圖 44、方法四流程之結果

同樣地，藍線與橘線分別代表對照組與方法一流程使用真實性別代替性別分類器預測之結果，黃線為先前方法三使用支持向量機性別分類器之結果，而新增的天藍色線條則是本方法四流程之實驗結果。相當可惜地，此方法四流程之結果仍然無法超越對照組之結果，顯然地，以註冊音檔來預測同一語者之測試音檔被誤判性別之可能性，並不是一套理想的方法。

3.5.4 方法五：檢查測試音檔的男女性別機率差

方法五與方法四類似，亦是對照組流程與方法一流程的組合，而其不同於前者之處，在於其流程是以測試音檔本身來預測性別分類器能否正確預測其性別。實驗三所實做的性別分類器，除了能預測一位語者之性別，更可以精確地計算出該語者屬於男性或女性的機率，而若一份音檔的兩個機率值相差非常多，代表性別分類器對此音檔之性別預測結果有較大的信心，反之，若一份音檔的兩個機率值相差過小，則代表性別分類器的判斷有較大的機會出錯。依此概念，便可以衍生出如圖 45 之流程：

doi:10.6342/NTU201801578

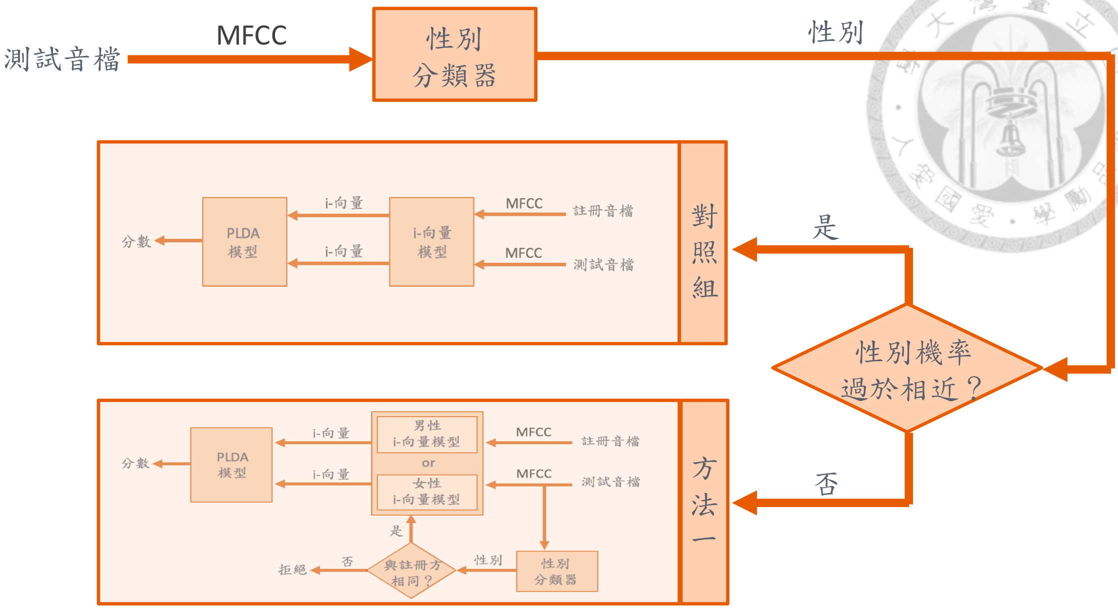


圖 45、方法五之流程架構圖

本流程同樣使用支持向量機性別分類器為測試音檔進行性別預測。當性別分類器於測試音檔計算出來的兩個性別機率值過於相近，便將此樣本交由對照組流程進行判別；反之，則交由方法一流程進行判別。

同樣地，方法五之前半部流程可被視為預測測試語者之性別是否會被誤判的二元分類器，其召回率、精確率與 F-度量則如圖 46 所示：

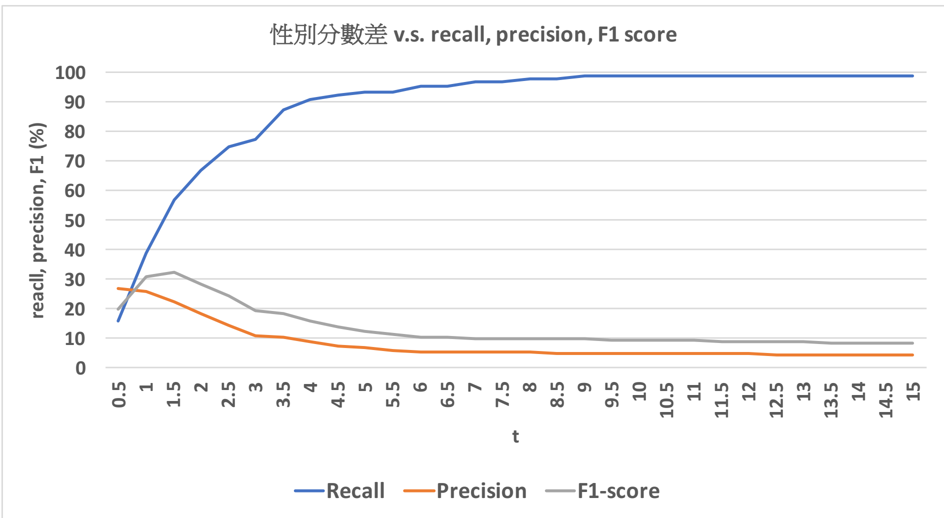


圖 46、方法五流程之「測試音檔性別誤判預測器」之效果

圖 46 之橫軸同樣為 t ，此 t 值代表的意義為測試音檔性別對數機率差之門檻值，意即：若性別分類器在測試音檔上所計算出之兩個性別對數機率差小於此值，

doi:10.6342/NTU201801578

便導入對照組流程，否則導入方法一流程。同樣地，召回率、精確率與 F-度量分別以藍線、橘線與灰線表示，圖之縱軸則為其百分比。由於本流程的門檻概念與方法四相反，因此圖 46 中的召回率與準確率亦與方法四相反，分別隨著 t 值之增加呈現出嚴格遞增與遞減之現象，而 F-度量則一樣在 t 之值偏中間（約接近 1.5 ）時有較大的值。

系統之整體表現與 t 值之關係則如圖 47 所示：

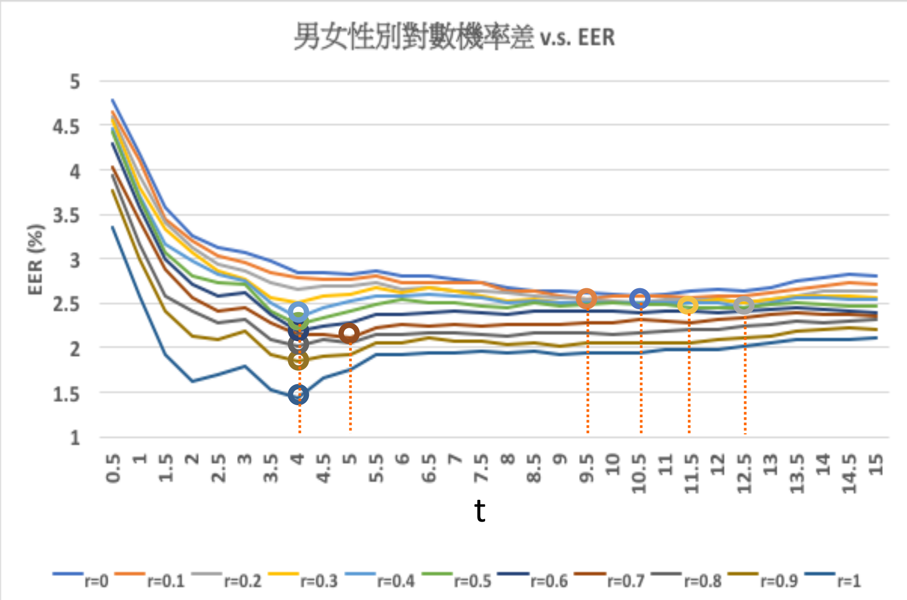


圖 47、方法五流程之選定測試音檔性別機率差門檻

與方才方法四不同，圖 47 中的每條線皆呈現出兩側高、中間低的現象，而每條線的最低點位置（在圖中以圓圈標出）亦不相同，例如：當 r 值較小（0～0.3）時，這些線條之最低點出現在圖的右半部，顯示其最佳門檻值偏大；而當 r 值較大（0.4～1）時，這些線條之最低點出現在圖的左半部，顯示其最佳門檻值偏小。由此現象可以推斷，男女性別機率差之門檻並沒有最佳解，而是會隨著資料的 r 值而變動。

圖 48 為方法五流程指定 t 值為 4 時之實驗結果：

doi:10.6342/NTU201801578

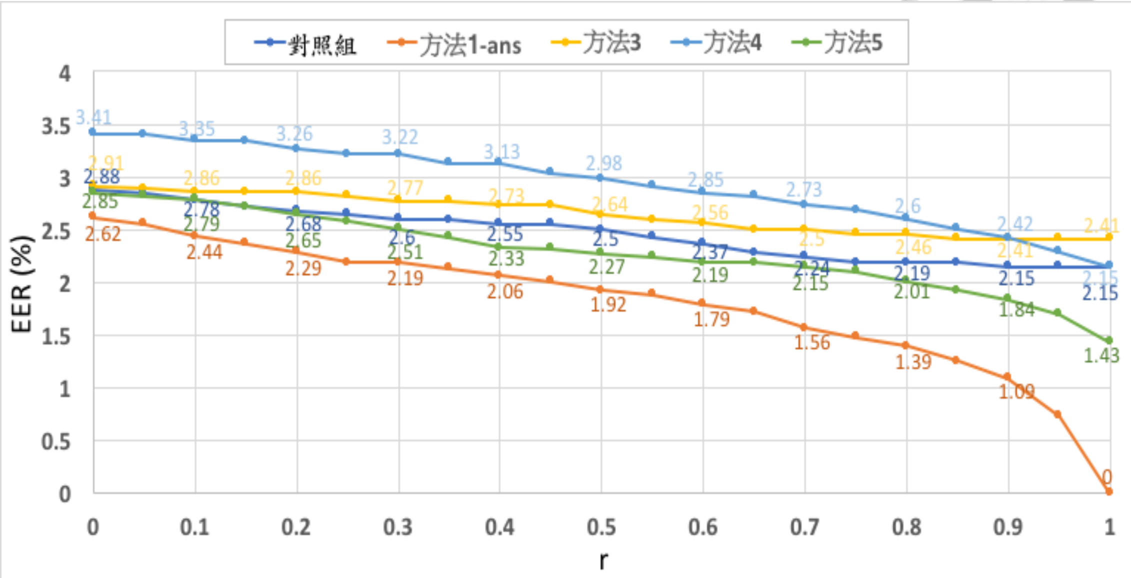


圖 48、方法五流程之結果

同樣地，藍線、橘線、黃線與天藍線分別代表對照組、方法一流程使用真實性別代替性別分類器、方法三與方法四之結果，而新增的綠線則是本方法五流程之結果。我們可以看到，此方法五流程終於達到了最初的預期，也就是在性別分類器表現不良之前提下，仍然超越對照組流程的目標。

doi:10.6342/NTU201801578

第 4 章 結論與未來方向

在語者驗證領域中，有效利用語者性別將能提升系統之準確度，但前一章節的各項實驗結果證實，無論性別分類器的表現或仿冒者的性別組成，皆會大大小小地影響系統的表現，因此不同的性別資訊運用方法，皆各有其適用的前提。以下整理本論文所使用的各種方法與其對應的適用情形作為總結：

(1) 當我們能確保系統的使用者可以提供品質良好的錄音，則性別分類器有較大機會表現良好，因此可選擇方法一或方法二；甚至當我們能進一步確保測試語者性別必定與註冊者相同，前人作法亦為很好的選擇。

(2) 當我們不能保證使用者能提供優良品質之錄音，則性別分類器有較大機會表現不良，因此在這種情況下，改良五為較合適之選項。

(3) 本論文提出的許多改良方法皆能達到較傳統作法低的錯誤率，但傳統作法的優點則是不必花費額外時間進行性別預測，因此當我們對於系統的整體運算速度有較高的需求，則可以選擇維持原來的傳統作法。

關於未來方向，若要持續往使用性別資訊的方向鑽研，則可以繼續改進性別分類器之辨識率，例如新增對於性別分類有用的資訊或特徵（例如語者之音高）、嘗試不同的特徵或分類器等；若系統之運用情境無法令性別分類器有更好的表現，則可以繼續嘗試其他結合傳統對照組流程與本論文改良流程的方法。而除性別以外，亦可考量將其他資訊（例如語者年齡、音高等）作為劃分聲學模型之依據，不同架構流程的結合也有很大機會使系統之辨識率進一步提高。

doi:10.6342/NTU201801578

參考文獻

[1] http://speech.ee.ntu.edu.tw/DSP2017Autumn/

[2] https://en.wikipedia.org/wiki/Window_function

[3] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the royal statistical society. Series B (methodological)* (1977): 1-38.

[4] Reynolds, Douglas A., Thomas F. Quatieri, and Robert B. Dunn. "Speaker verification using adapted Gaussian mixture models." *Digital signal processing* 10.1-3 (2000): 19-41.

[5] Gauvain, J-L., and Chin-Hui Lee. "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains." *IEEE transactions on speech and audio processing* 2.2 (1994): 291-298.

[6] Kenny, Patrick. "Joint factor analysis of speaker and session variability: Theory and algorithms." *CRIM, Montreal,(Report) CRIM-06/08-13* 14 (2005): 28-29.

[7] Kenny, Patrick, et al. "A study of interspeaker variability in speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (2008): 980-988.

[8] Dehak, Najim, et al. "Front-end factor analysis for speaker verification." *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011): 788-798.

[9] Mika, Sebastian, et al. "Fisher discriminant analysis with kernels." Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop.. Ieee, 1999.

[10] Prince, Simon JD, and James H. Elder. "Probabilistic linear discriminant analysis for inferences about identity." Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007.

[11] Kenny, Patrick. "Bayesian speaker verification with heavy-tailed priors." Odyssey. 2010.

[12] Garcia-Romero, Daniel, and Carol Y. Espy-Wilson. "Analysis of i-vector length normalization in speaker recognition systems." *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

doi:10.6342/NTU201801578

[13] Lyu, Siwei, and Eero P. Simoncelli. "Nonlinear extraction of independent components of natural images using radial gaussianization." *Neural computation* 21.6 (2009): 1485-1519.

[14] https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2010

[15] Garofolo, John S., et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[16] Anjos, André, et al. "Continuously reproducing toolchains in pattern recognition and machine learning experiments." (2017).

[17] Kanervisto, Anssi, et al. "Effects of gender information in text-independent and text-dependent speaker verification." *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017.

[18] M. Wang, Y. Chen, Z. Tang, and E. Zhang, "I-vector based speaker gender recognition," in 2015 IEEE Advanced Information Technology, Electronic and Au- tomation Control Conference (IAEAC), 2015, pp. 729- 732.

doi:10.6342/NTU201801578