The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 1-15

基於數字文本相關之語者驗證系統的研究與實作

Study and Implementation on Digit-related Speaker Verification

周宗鴻　Chung-Hung Chou
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan University
r05922085@ntu.edu.tw

張智星　Jyh-Shing Roger Jang
國立臺灣大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
jang@csie.ntu.edu.tw

蕭善文　Shan-Wen Hsiao
中華電信研究院
Chunghwa Telecom Laboratories, Taoyuan, Taiwan
swhsiao@cht.com.tw

摘要

聲紋驗證為生物辨識中一種重要的驗證方式，此種驗證方式最大的優點即是硬體需求簡單，只需要一般市面上常見的麥克風即可，因此常用於電話及手機的生物辨識。本篇論文目標為建立一套文本相關的聲紋驗證系統並包含三個部分:「動態時間扭曲語者驗證系統」利用強制對齊切開數字後藉由動態時間扭曲比較註冊時數字的梅爾倒頻譜係數與測試時數字的梅爾倒頻譜係數之差異、「語句級語者驗證系統」直接抽取註冊音檔與測試音檔的 i-vector 並使用餘弦相似度或機率線性判別分析來評分這二組 i-vector、「數字級語者驗證系統」利用強制對齊切開數字後抽取註冊音檔與測試音檔中各個數字的 i-vector 並使用餘弦相似度或機率線性判別分析來評分對應數字的 i-vector。

Abstract

Speaker recognition is an important biometric identification method. The biggest advantage of using such method is the simple requirement of its hardware, which only consists of a microphone. Therefore, it is widely implemented in mobile phones and call centers. The purpose of this thesis is to create a text-related speaker verification system, for which we

conduct three different approaches to analyze their result: dynamic time warping compares the differences between the MFCCs for digits at registration and digits at testing after applying forced alignment; sentence-level uses cosine similarity or PLDA to rate the two groups of i-vector retrieved from the audios at registration and testing respectively; digit-level uses cosine similarity or PLDA to rate each i-vector of every digit in the audios after applying forced alignment.

關鍵詞：語者驗證，強制對齊，動態時間扭曲，i-vector，機率線性判別分析

Keywords: Speaker Verification, Forced Alignment, DTW, i-vector, PLDA.

一、緒論

近年來由於智慧型手機及通訊網路的普及，過去許多無法達成的事情已經逐漸變得可行，例如：過去人們都是透過現金進行交易，但現在因為智慧型手機的普及，行動支付及網路購物已經能逐漸取代過去人們傳統的交易行為。然而也因為這些事情帶來許多新的問題，以行動支付及網路購物而言，問題即是要如何辨認使用者之身份，進而確認為本人使用而非盜用便成為現在非常重要的一個課題。

任何的系統都需要辨識、確認使用者之身份，因為如果無法確認使用者其身份導致個人資料外洩是非常嚴重的事情，例如：信用卡資料遭到盜用、家裡遭到小偷入侵、私密照片流出…等。傳統我們在辨識使用者的方法多為：鑰匙、門禁卡、密碼鎖…等，但這些方法往往都會有遺失並且遭到盜用的風險。近年來由於電腦運算能力的大幅進步，陸續開始有不少人提出利用人類生理特徵進行辨識使用者身份，使用人類生理特徵可以避免掉遺失並遭盜用的風險。目前市面上常見的利用人類生理特徵驗證方法有：指紋驗證、臉部驗證、視網膜驗證、聲紋驗證(語者驗證)…等。

本研究目標為建立一套數字文本相關的語者驗證系統並用於行動支付，因為行動支付多應用在智慧型手機且每台智慧型手機皆擁有錄音之功能，又錄製聲音為一件容易達成之事情，因此我們希望能藉由語者驗證來達成確認使用者之身份。

二、語者驗證簡介

語者驗證為依據說話者宣稱之身分及其語音內容判斷是否屬實，此種應用情境常用於：行動支付、門禁系統...等。另外語者驗證又可依照說話者語音之內容分為三大類，分別為：本文獨立（text-independent）、本文相關（text-relative）、本文相依（text-dependent）。本文獨立為語音內容可以為任意的，而本文相關為語音內容必須在某些範圍內，例如：僅能是數字或者顏色…等，而本文相依則為完全限定語音之內容，例如僅能是 Hey, Siri、OK, Google 等遭限制之語音內容。

在語音訊號處理中特徵通常我們會使用梅爾頻率倒譜係數（Mel-scale Frequency Cepstral Coefficients, MFCC）[1]，而語者驗證中除 End-to-End[2]外也多半是先抽取 MFCC 後再進行其它處理，其相關方法有：動態時間扭曲（Dynamic Time Warping, DTW）[3]、高斯混合模型（Gaussian Mixture Model, GMM）[4]、通用背景模型（Universal Background Model, UBM）[5]、聯合因子分析（Joint factor Analysis, JFA）[6]、i-vector[7]。另外值得一提的是，i-vector 在訓練總體變異矩陣時會受到不同通道的干擾，因此一般會使用線性判別分析（Linear Discriminant Analysis, LDA）[8]進行信道補償以及使用機率線性判別分析（Probability Linear Discriminant Analysis, PLDA）[9, 10]進行評分。

三、語料配置及評估標準

（一）語料介紹

本研究之來源語料分為兩大類，第一類為我們在台大資工系張智星教授在課堂上所搜集的錄音，第二類為交大電機系王逸如教授提供給我們的錄音，以下我們將對這兩類語料進行說明。

1. 台大張智星教授之語料

本語料共有 191 位語者，其中男生共有 160 位、女生共有 31 位，室內錄音共有 172 位、室外錄音共有 19 位。每 1 位語者皆有 10 組錄音，每組錄音之內容皆為 0 至 9 的排列組合（長度為 10），錄音裝置皆透過實驗室開發之 Android APP 進行錄音，取樣頻率為單聲道 44.1 KHz，音質為 16bit。此外我們有針對所有的音檔經由人工標記出 0 至 9 在該音檔發音開始的時間以及發音結束的時間，其標記所採用之程式為 Audacity。

2. 交大王逸如教授之語料

本語料共有 100 位語者，其中男生共有 50 位、女生共有 50 位，皆為室內錄音。每 1 位語者皆有 10 組錄音，每組錄音之內容皆為長度 4 – 12 的隨機數字，錄音裝置為麥克風，取樣頻率為單聲道 16KHz，音質為 16bit。此外該語料包含標記檔對應到每個音檔中各個數字發音開始與結束的時間。

（二）資料配置

我們將第三章第一節之二組語料拆成訓練資料與測試資料。訓練資料部分會依照原性別比例從台大語料中取 41 位語者（其中男生 34 位、女生 7 位）、交大語料中之所有語者（其中男生 50 位、女生 50 位），測試資料部分取剩餘之台大語料（其中男生 126 位、女生 24 位）。

訓練資料會全部用於訓練模型，而測試資料會分為語者正確（接受）及語者錯誤（拒絕）這二種情況。現在我們假設註冊音檔使用二組音檔，則接受之筆數會共有 150 位語者 * 每一位語者有剩餘 10 – 2（8）組音檔可用於驗證，所以接受筆數共有 1200 筆；而拒絕之筆數會共有 150 位語者 * 149 位非正確語者 * 每一位語者有剩餘 10 – 2（8）組音檔可用於驗證，所以拒絕筆數共有 178800 筆。其正確錯誤比為 1200:178800 = 1:149。

（三）評估標準

語者驗證的錯誤可以分成錯誤接受（False Acceptance）及錯誤拒絕（False Rejection）二類。錯誤接受為聲音片段並非為其宣稱之語者但系統誤判為是其宣稱之語者，我們稱之為錯誤接受，而錯誤拒絕為聲音片段為其宣稱之語者但系統誤判為不是其宣稱之語者，我們稱之為錯誤拒絕。

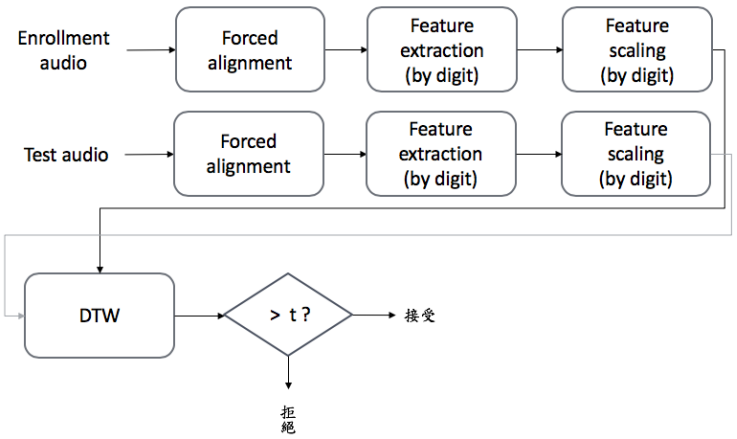
我們希望能夠從全部的測試中找一個門檻值，使得錯誤接受比例（False Acceptance Rate, FAR）等於錯誤拒絕比例（False Rejection rate, FRR），該比例我們稱之為相同錯誤比例 (Equal Error Rate, EER)，而 EER 即為我們用來評估系統的標準。另外一提的是，雖然一般來說 EER 即能夠反應系統的效能，但當系統要實際上線時仍需要依照應用情境對門檻值進行抽換。舉例來說當用於高安全性的系統時，我們更在乎的會是 FAR 而非 EER，從上述中即會發現會因為情境的不同而對於 FAR 及 FRR 有不同的要求，而這部分則可以經過錯誤權衡圖觀察後選擇適合的門檻值。

四、實驗與分析

所有的實驗採用之作業系統為 Ubuntu 16.04、處理器為 Intel i7 8700、記憶體為 DDR 2400 16GB 4 條、程式語言為 Python3.6，語者驗證工具箱採用 SPEAR [11]，預設使用人工標記之結果。

（一）動態時間扭曲語者驗證

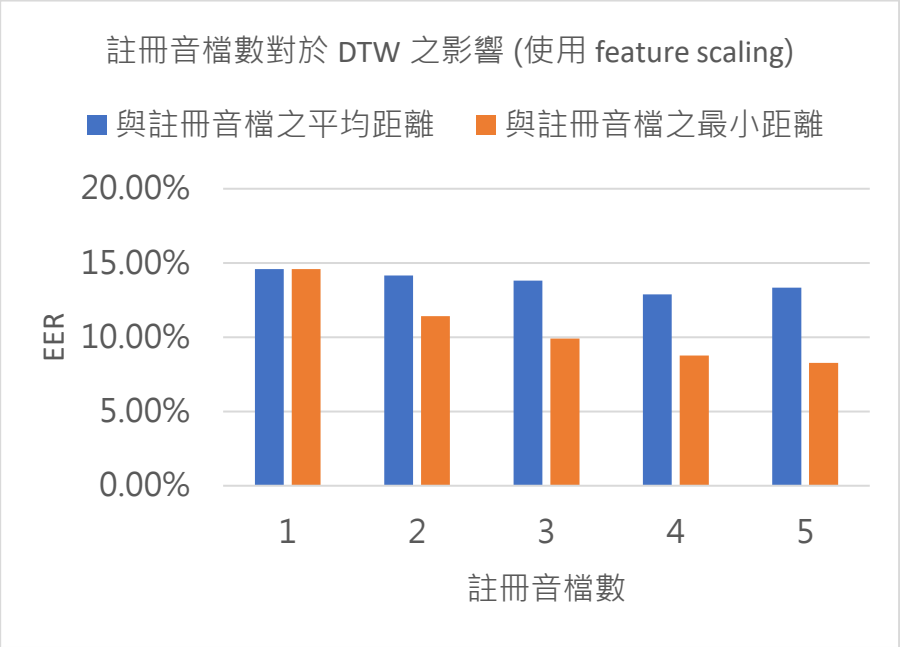
我們依照第三章第二節的資料配置對於所有測試利用動態時間扭曲比較聲音之間的距離。其對一筆測試之詳細作法，我們先使用人工標記（強制對齊）精準地切出各個數字之片段，再使用動態時間扭曲計算註冊與測試對應數字間的距離，而這些數字之距離加總即為這筆測試之距離。由於可能有多個註冊音檔之情況，因此我們分別可以使用平均距離或是最短距離作為最終距離。反覆對於所有測試做完後，我們可以得到非常多的距離，再利用這些距離我們即可計算 EER，圖一即為本系統之流程圖。



圖一、動態時間扭曲流程圖

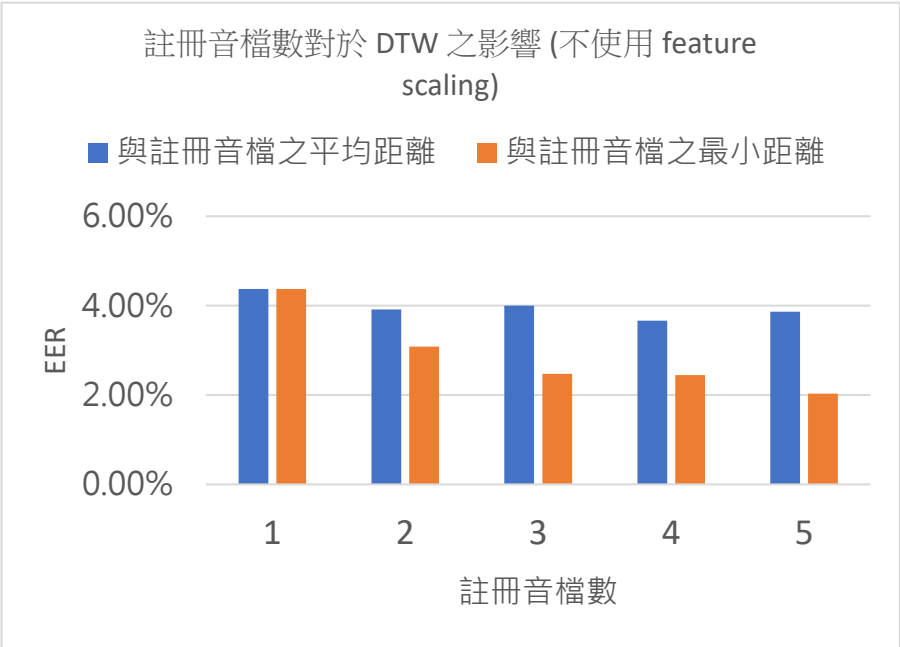
圖二為我們使用 feature scaling 時註冊音檔數對於動態時間扭曲之比較直條圖，我們可以發現使用平均距離的 EER 並不會因為註冊音檔數的增加而有明顯的穩定下降，但使用最短距離則有明顯的穩定下降之趨勢。經過分析發現這種現象是因為「平均」用於距離並不公平，舉例來說：小明在早上時錄音註冊，在中午時又進行錄音註冊，晚上進行最後的錄音註冊，因此我們有小明的三個註冊音檔。而現在有一個小明的聲音要驗證，則該聲音會與小明的三個註冊音檔進行 DTW 算出距離，如果該聲音是在早上錄製則肯定會跟小明在早上註冊的聲音最像，但會跟中午、晚上的聲音則較不像，如果我們在這邊使用平均距離則會被中午吃飯及晚上刷牙所影響，但使用最小距離則不會受到影響。

這說明了為什麼使用平均距離錯誤率比較不會穩定下降的原因。



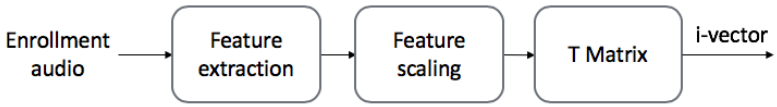
圖二、使用 feature scaling 時註冊音檔數對於動態時間扭曲之比較

此外可以觀察到動態時間扭曲應用於語者驗證上效果非常的不理想，這與我們過去的認知大不相同，經過一些分析及猜測後發現可能是進行 feature scaling 所造成的，這樣猜測的原因在於 feature scaling 會導致每一維中的 MFCC 距離影響變相同，但語者及數字特性可能主要包含於 MFCC 中的某些特定維度進而導致區分語者的能力下降，圖三即為我們為了驗證猜測所進行實驗之結果，可以發現 EER 確實有大幅降低。



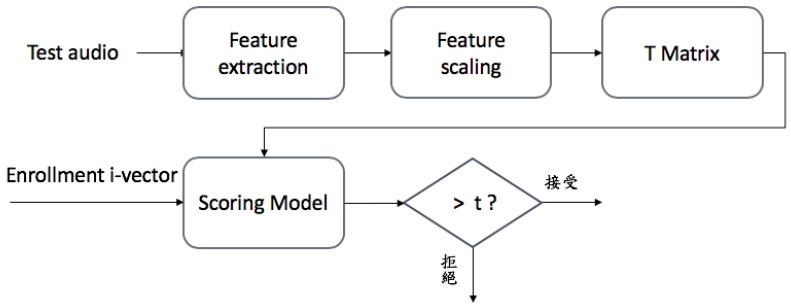
圖三、不使用 feature scaling 時註冊音檔數對於動態時間扭曲之比較

（二）語句級語者驗證



圖四、語句級註冊流程圖

圖四為語句級註冊流程圖，輸入為一位語者的 n 個音檔，抽出音檔的 MFCC 後對 MFCC 進行倒頻譜平均值與變異數正規化，最後將這些經過倒頻譜平均值與變異數正規化後的 MFCC 丟進全變異空間模型後我們即可得到該代表語者之 i-vector。



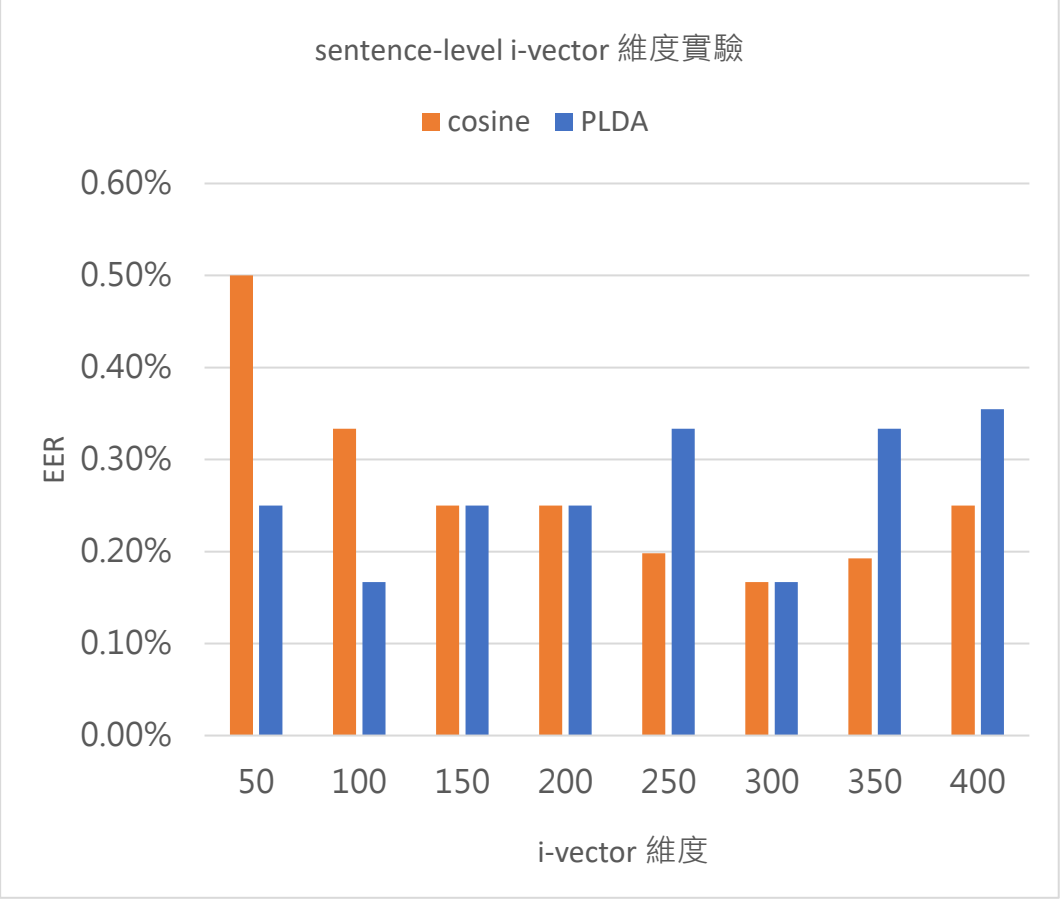
圖五、語句級測試流程圖

圖五為語句級測試流程圖，輸入為一個測試的音檔以及該測試音檔宣稱屬於語者的 i-vector，抽出測試音檔的 MFCC 並對 MFCC 進行倒頻譜平均值與變異數正規化，並將這些經過倒頻譜平均值與變異數正規化後的 MFCC 丟進全變異空間模型後我們即可得到該測試音檔之 i-vector。最終我們將測試音檔之 i-vector 以及宣稱屬於語者的 i-vector 利用評分模型即可輸出分數，反覆對於所有測試做完後，我們即可計算 EER 作為評量標準。

實驗設定部分，我們特徵參數使用 39 維的 MFCC，抽取 MFCC 使用的窗大小為 1024、重疊大小為 512，通用背景模型部分我們的高斯混合模型使用 128 個高斯分佈最多迭代 25 次，初始化使用 k-平均演算法最多迭代 25 次、i-vector 部分我們最多迭代 10 次，餘弦相似度部分我們不對 i-vector 做後處理，直接使用得到之 i-vector 進行計算。機率線

性判別分析部分，我們會先對 i-vector 使用語者之標籤做 LDA 降到 50 維，而機率線性判別分析的語者空間以及通道空間之矩陣維度皆設定為 50 維，最多迭代 50 次

圖六為我們依照上述實驗設定所跑出來的實驗結果，我們可以發現在我們的資料上評分方式使用餘弦相似度在 i-vector 維度較低的時候表現都比使用機率線性判別分析來得差，但是隨著 i-vector 維度上升，餘弦相似度之表現反而比機率線性判別分析來得好。對於這個現象之看法是：由於我們的資料僅有數字且語者數目為幾百人，因此使用機率線性判別分析在較高維度的 i-vector 時反而有可能造成過度擬合的情形，最後亦可以發現餘弦相似度及機率線性判別分析都在 i-vector 維度 300 時有最佳的表現。



圖六、語句級使用不同評分模型以及不同 i-vector 維度之比較

（三）數字級語者驗證

數字級語者驗證之註冊流程及測試流程多半與語句級語者驗證系統相同，最大差異在於數字級語者驗證系統會將一段序列如：1029384756 拆成 1、0、2、9、3、8、4、7、5、6 後，再將個別數字當成「語句」送入語句級註冊（其流程圖如圖四）。以上述例子而

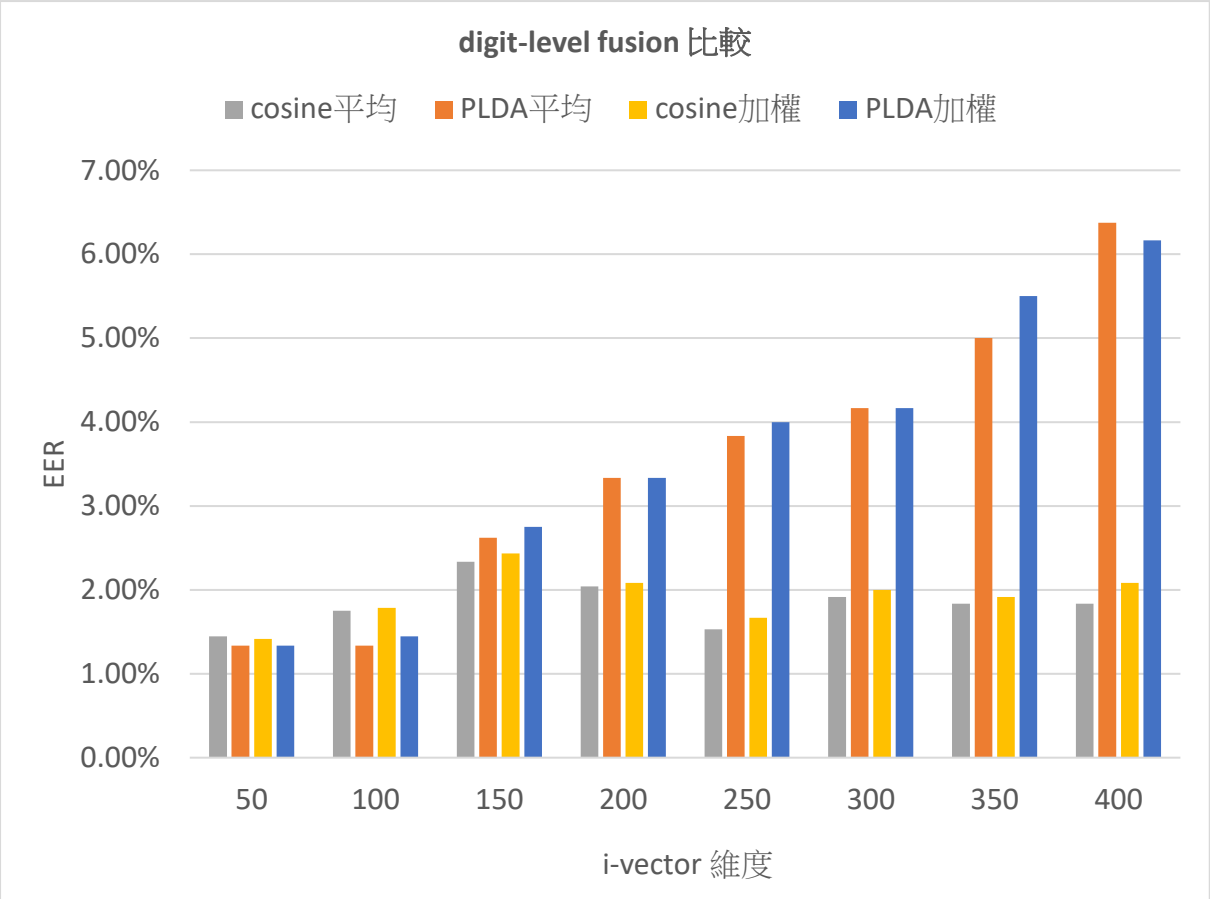
言，數字級語者驗證在註冊時即可得到 10 組代表語者的 i-vector（分別為 1、0、2、9、3、8、4、7、5、6 的 i-vector），測試時將註冊時得到之 10 組代表語者的 i-vector 與切好的數字一一對應送入語句級測試（其流程圖如圖五）即可得到 10 組分數，最後再將這些分數進行組合即可形成最終之分數，下面為我們的二種組合方法。

1. 使用平均分數

$$score = \sum_{i=0}^{9} digit\ i's\ score\ * \frac{1}{10}$$

2. 使用加權分數

$$score = \frac{frames\ of\ dgiti\ i}{\sum_{i=0}^{9} frames\ of\ digit\ i} * digit\ i's\ score$$

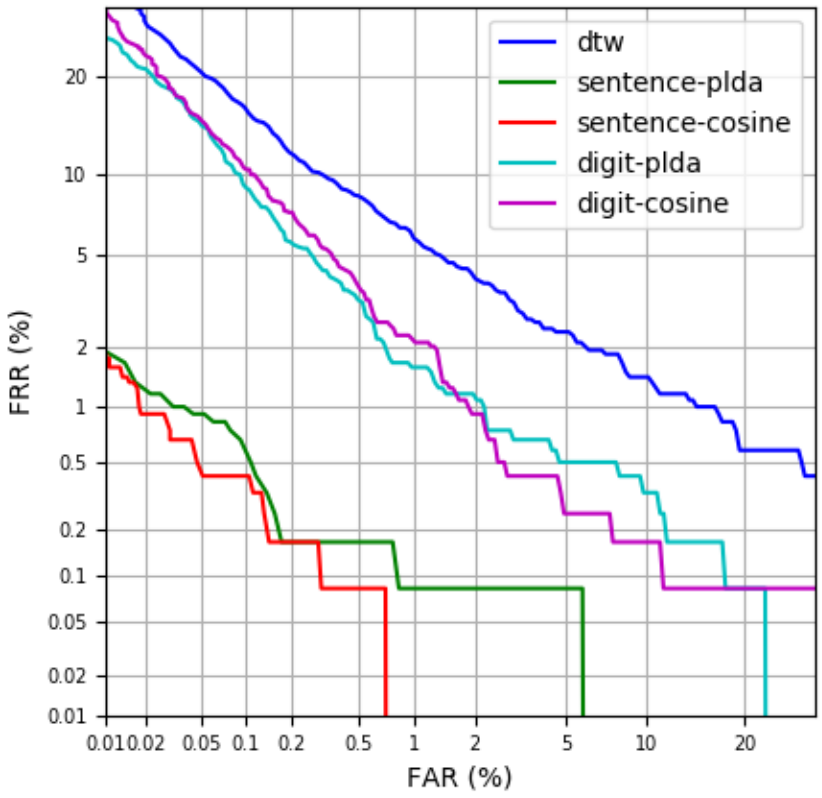


圖七、數字級使用不同評分模型、不同 i-vector 維度、不同分數組合之比較

圖七為數字級使用不同評分模型、不同 i-vector 維度、不同分數組合之比較直條圖，我們可以從圖中發現使用平均分數以及加權分數對於錯誤率並不會有太大的影響。此外可以發現 PLDA 隨著維度上升而整體錯誤率也有明顯上升的趨勢，我們認為其原因相同

於語句級。最後亦可以發現整體最好的表現大概落在 PLDA 使用平均時，其錯誤率落在 1.33%。

（四）錯誤分析

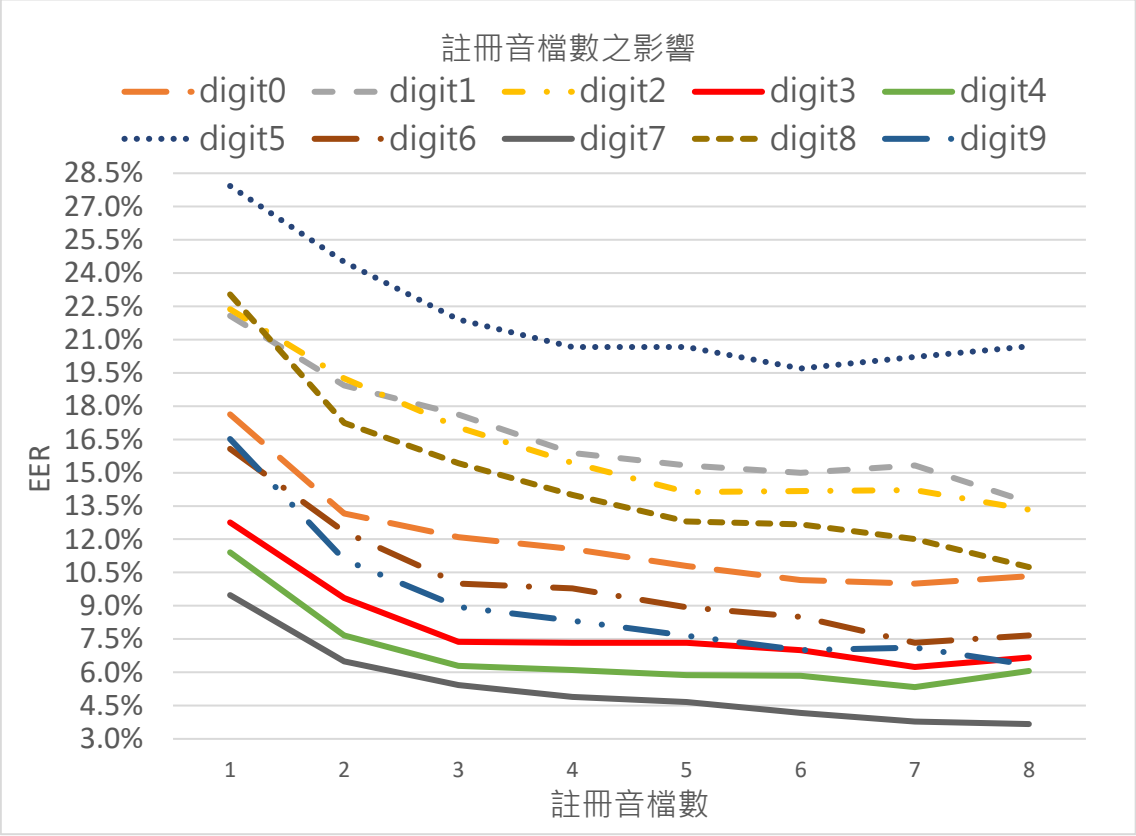


圖八、錯誤權衡圖

圖八為錯誤權衡圖，從圖中我們可以發現語句級語者驗證系統最佳，數字級語者驗證系統次佳，動態時間扭曲語者驗證系統最差。動態時間扭曲為最差的原因我們可以很直觀的猜到是因為 MFCC 中包含著雜訊、音量...等特徵，然而這些特徵並非語者的特徵，因此當我們使用動態時間扭曲進行比較距離時會受到這些特徵所影響，進而影響整體的錯誤率。然而數字級語者驗證系統效果較差並不直觀，因為數字級語者驗證系統採用的方式是文本相依的方式，直觀上我們會認為此種方式效果應該較佳，但實際上卻並非如此。因此下面我們將探討數字級語者驗證系統較差的可能原因。

首先我們檢查使用單獨數字的錯誤率，如圖九所示。我們可以發現增加註冊音檔數時每個數字的錯誤率都有明顯下降，但每個數字的錯誤率差異非常明顯，錯誤率最高的前四條線依序為：digit5、digit2、digit1、digit8。我們對於有這種現象的猜測是每個數字的發

音特質不同造成，原因是我們在人工標記音檔時發現有些數字聽起來非常像是雜音，此外亦有些數字再連續唸起來時會被省略部分音或者唸過快的現象，經過觀察發現這些數字多半只有母音，若同時包含母音及子音則比較不會有此種現象。

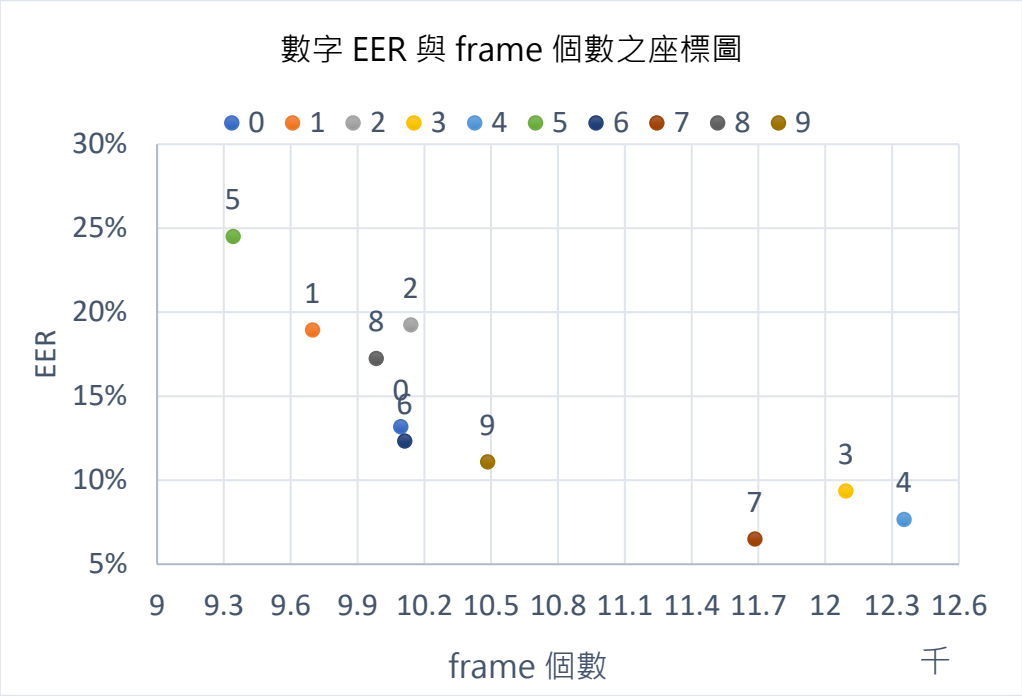


圖九、單獨數字使用 50 維 i-vector 於不同註冊音檔數之錯誤率趨勢圖

為了驗證每個數字錯誤率差異很大是否如我們猜測，我們觀察數字 0 到 9 在 Audacity 的訊號圖，確實發現到錯誤較高的數字發音較為簡單，而錯誤率較低的數字發音則較有變化。此外我們亦對註冊音檔數與錯誤率的關係做出座標圖如圖九所示，從座標圖中我們可以明顯發現有呈現反比之趨勢。圖十之結果由個別數字的錯誤率來觀察 EER 與音窗的個數之相關性，此間接可以證實發音的長度與難度與錯誤率成正比的關係。

從上述實驗中可以發現某些數字錯誤率確實比較高，也發現了這些錯誤率較高可能是因為發音特質以及長度造成的影響，因此我們認為很有可能是這些錯誤率較高的數字加入計算反而導致整體錯誤率上升。圖十一為我們使用 PLDA 評分模型在 50 維 i-vector 拿掉錯誤率最高的前 n 個數字之直條圖，從圖中可以發現當我們拿掉 1 個數字後錯誤略確實有些微提升，拿掉 2 個數字時錯誤率又再度回到不拿掉任何數字之錯誤率，拿掉 3 個以上數字後錯誤率開始逐漸升高。因此我們可以從實驗中推斷這 2 個錯誤率最高的數

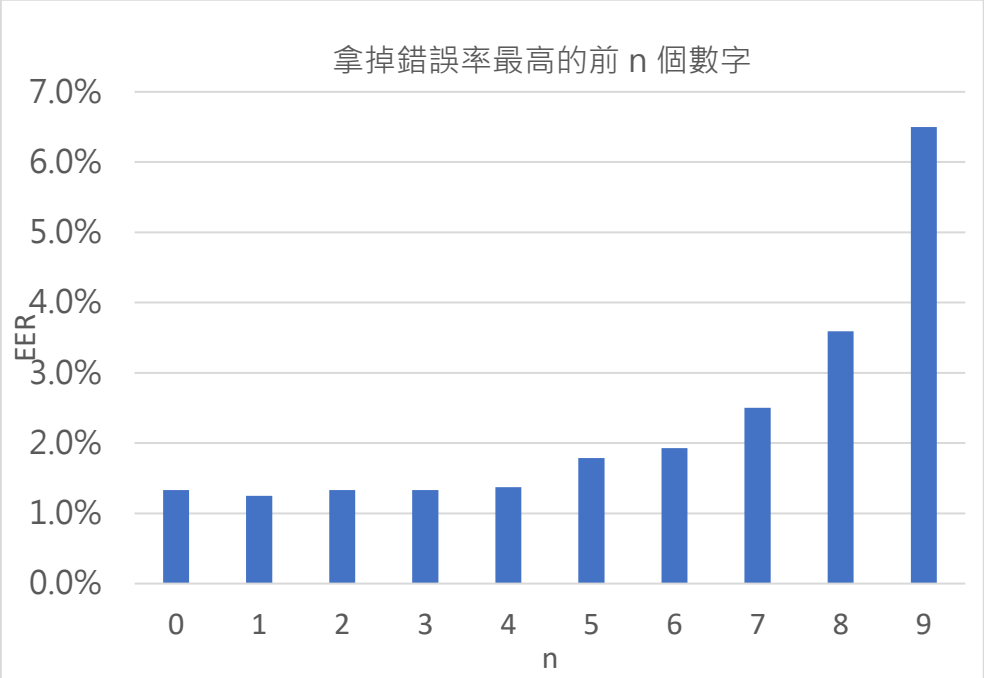
字對於錯誤率並無明顯助益，此外雖然拿掉這 2 個錯誤率最高的數字對於錯誤率無明顯提升，但卻也可以減少使用者所需花費的時間。



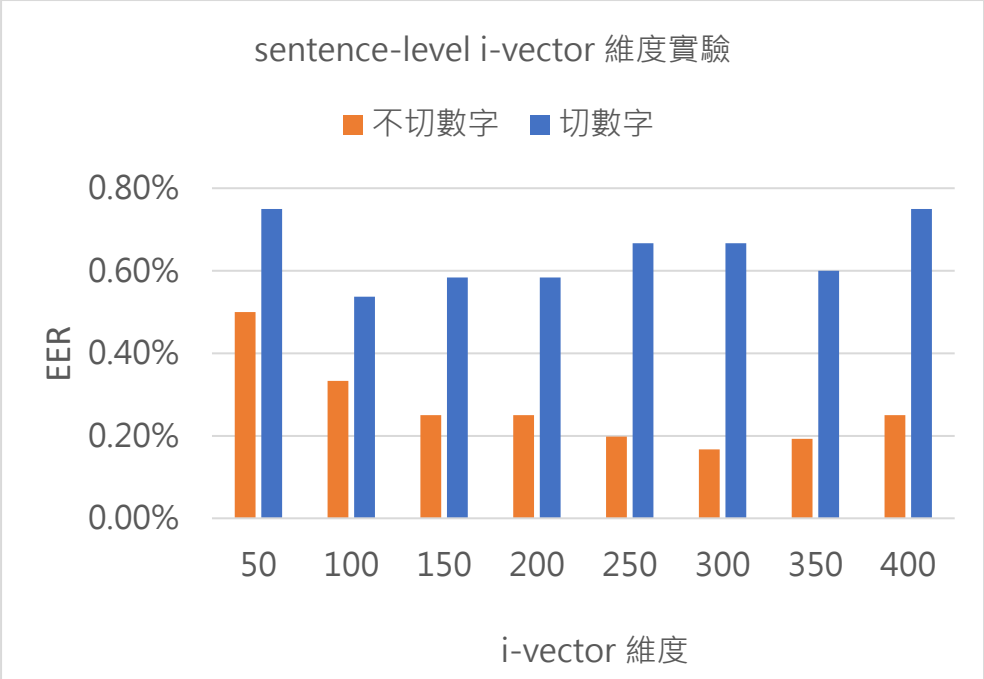
圖十、個別數字錯誤率與窗的個數之關係座標圖

由於拿掉錯誤率較高的數字後我們的錯誤率仍然比語句級高出許多，我們進而猜測連續數字發音中間的一些特徵是否有可能被移除了，舉例來說：多數人的發音時常帶著一些習慣的腔調，但因為我們要把數字強制拆開所以導致夾雜在數字中間的這些腔調也消失了。因此我們決定將語句級的音檔也進行相同作法，其做法是先抽取整句話之 MFCC 再依照數字發音時間只取數字發音時間內對應窗的 MFCC。

圖十二即為不切數字與切數字於語句級之比較直條圖，我們可以發現切數字後合併錯誤率平均落在 0.64%附近、不切數字錯誤率大致落在 0.27%附近，不切數字在任意的 i-vector 維度上表現都比切數字來得好許多，因此我們可以推斷連續數字發音中間的特徵確實包含著語者特徵。



圖十一、拿掉錯誤率最高的前 n 個數字之直條圖



圖十二、不切數字與切數字於語句級之比較直條圖

五、結論

我們從實驗中得知 i-vector 的方法使用二個音檔註冊時 EER 語句級可以落在 0.2% ，而數字級 EER 落在 1.33%，但使用動態時間扭曲方法則 EER 無法低於 3%，由此可見動態時間扭曲方法之效果顯然比 i-vector 方法來得差。另外比較語句級語者驗證系統與數

字級語者驗證系統，我們發現語句級語者驗證系統效果明顯比數字級語者驗證系統來得好。此外經過近一步分析數字級語者驗證系統我們發現以下現象：

(1) 移除錯誤較高的數字對於錯誤率影響不大。

(2) 發音較長的數字比發音較短的數字用於語者驗證效果較佳。

(3) 發音較為複雜之數字的語者驗證效果也比發音簡單的數字來得好。

(4) 數字級語者驗證系統效果較差的原因之一是由於強制切開導致連續數字發音間的語者特徵消失所致。

(5) 文本內容較複雜時 i-vector 維度應該增加，文本內容較簡易時 i-vector 維度應該減少。

參考文獻

[1] Rabiner, Lawrence R., and Biing-Hwang Juang. Fundamentals of speech recognition. Vol. 14. Englewood Cliffs: PTR Prentice Hall, 1993..

[2] Heigold, Georg, et al. "End-to-end text-dependent speaker verification." Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016.

[3] Muda, Lindasalwa, Mumtaj Begam, and Irraivan Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[4] Reynolds, Douglas A. "Speaker identification and verification using Gaussian mixture speaker models." Speech communication 17.1-2 (1995): 91-108.

[5] 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.

[6] Kenny, Patrick, et al. "Joint factor analysis versus eigenchannels in speaker recognition." IEEE Transactions on Audio, Speech, and Language Processing 15.4 (2007): 1435-1447.

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

[8] 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.

[9] 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.

[10] 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

[11] Khoury, Elie, Laurent El Shafey, and Sébastien Marcel. "Spear: An open source toolbox for speaker recognition based on Bob." Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014.