基於機率線性鑑別分析之強健式語者驗證系統

方偉德、林修德*、廖元甫
國立臺北科技大學電腦與通訊研究所

摘 要 — 本文針對如何消除在語者辨認系統易受不匹配通道與背景雜訊影響的問題， 提出搭配機率線性鑑別分析 (Probabilistic Linear Discriminant Analysis; PLDA))[1]與特徵向量長度正規化之語者驗證系統。主要是將語者語料，先以 i-Vector 分析，求取其特徵超向量，並將特徵超向量做長度正規化[2]，使得語料的模型趨近標準高斯分佈，以配合 PLDA 的模型假設，再用 PLDA 作鑑別性因素分解，去除所有與語者本身無關的干擾因素，達到增進語者辨認效能的目的。實驗結果顯示，在 Core-Core 的驗證項目下， PLDA 加入長度正規化的系統辨識效能比傳統 Linear Discriminant Analysis; LDA[3] 系統的 Min Cost 有 31.13% 的相對效能增益，EER(Equal Error Rate) 有 11.24%的相對效能增益，最後我們將結果跟其它約 40 對參賽隊伍比較，我們的系統的效能還算不錯。[1]

一、 簡介

語者驗證是語音技術中重要的研究之一，其中又以美國國家標準與技術局(National Institute of Standards and Technology , NIST)最熱衷致力於相關的研究，它定期於雙數年舉辦(Speaker Recognition Evaluation;SRE)[4]語者驗證的系統評比，吸引世界各國的相關研究單位一同共襄盛舉，每次的 NIST SRE 所有頂尖的研究團隊便會開發更新更有效能的語者辨認技術並在該評比計畫當中發表其研究成果，每次評比的宗旨便是希望在語者辨認上能發展更有效能的技術進而引導出更多的技術構想。

語者驗證研究方法早期常使用以高斯混合模型(Universal Background Model；UBM/ Gaussian Mixture Model;GMM)[5]為基礎建立語者模型，雖然以 UBM/GMM 的語者辨認系統能提供不錯的辨認效果，但是實際上卻未考慮影響辨認效能的語者或通道環境因素干擾，因此後來發展出聯合因素分析(Joint Favtor Analysis;JFA)[6]技術做為通道補償的方法。透過聯合因素分析可以找出構成該語者語音的各種成分並移除干擾的因素，然而要使用聯合因素分析法必須要有各個語音訊號的語者與通道屬性標註資訊，並透過該標註資訊去準確的移除語者與通道干擾，倘若沒有掌握所有語音的完整標註資訊以至於在分類上的不完全，根本無法訓練出好的 JFA 模型，所以在 2010 年由[3]所提出的 i-Vector 找出語者的特徵空間，然後利用 LDA 進行語者的分類，降低訓練 JFA 模型的複雜度。

[1] 本研究由國科會贊助，計畫編號 NSC 102-2221-E-027-070。

雖然傳統上利用 LDA 可進行語者的分類，但由於其並非機率模型，因此我們在本論文中提出使用 PLDA，利用其精確的因素機率模型，將同個語者的特徵超向量在空間基底的投影量拉近，將不同語者在空間基底的投影量拉開，使得萃取純語者資訊時，能較為精確，此外，由於 PLDA 的模型假設為高斯分佈，故需先將每一句語料的特徵超向量做長度正規化，使得語料的模型趨近標準高斯分佈，以使機率模型更加穩健。最後將此系統應用在 SRE 2012 Core-Core 驗證項目(對話方面分為電話與面談;通道方面分為電話與麥克風…等。)評比中，證明其可有效辨別目標語者和非目標語者。

二、 PLDA 與特徵向量長度正規化

本論文的主要的想法是先利用 UBM 求取高維的語者特徵超向量，再利用 i-Vector 把語者特徵超向量降到低維度，然後經由 PLDA 的機率模型，將特徵超向量，分解成語者、通道與其餘因素，然後只萃取我們要的純語者的資訊。且因為 PLDA 的機率模型，一般假設是高斯分布，可是一個登記語者的語料量不一定夠多，所以導致機率模型的變異量可能不是高斯分佈，反而是 heavy-tail 的 Student’t 分布。然而若採用 heavy-tail 的機率模型，會造成系統過於複雜，因此我們進一步改進使用簡單的長度正規化(Length Normalization)做補償，讓 i-Vector 求出的語者投影量從原本的 Student’t 機率分布，正規化成高斯分布，使 PLDA 的機率模型更正確。

以下進一步敘述 PLDA 模型與系統實現架構：

2.1 PLDA 模型

首先將所有語料，利用 i-Vector 的模型把語料的特徵向量作轉換。這是一種能將資料群由高維度的空間投影到低維度的空間，且 i-Vector 的優點在於當我們在訓練模型時，不用標註每一句訓練語料是由何種語者跟何種環境(通道，雜訊等)所構成，換句話說，i-Vector 的模型可以包含所有語者跟語者以外的因素。

在這個模型下，我們假設語者的超向量包含語者、通道跟環境等因素的影響，並且將 GMM 的超向量表示為 M 如方程式(1)所示：

$$M = m + Tw(L) \quad (1)$$

其中， m 是由非特定語者、非特定通道語與非特定環境雜訊所訓練出的 UBM 模型，由 C 個維度為 F 的平均值串接而成的超向量，維度是 $CF \times 1$。T 是一個低秩

Corresponding author:gue991231@yahoo.com.tw
DOI:10.6159/IJSE.2014.(4-1).85

的矩陣，是構成 TV 空間的元素，也就是構成模型的基底，維度是$CF \times R_L$。$w(L)$是構成空間的投影量，屬於隱藏的變數，初始設定為標準的高斯機率分佈，並將其定義成 i-Vectors，而 L 代表 24 種語言。

接著我們使用 PLDA 模型將 i-Vector 的特徵空間分成語者、通道、其餘因素，以排除其它不必要的特徵參數(通道、語者因素)。

首先我們使用 F 來表示特徵向量的維度，先假設訓練語料包含 i 個語者，每一個語者有 j 個句子，用w_{ij}表示第 i 個語者的第 j 個句子。因此機率線性鑑別分析模型的方程式如(2)所示：

$$w_{ij} = m + Vy_i + Ux_{ij} + \varepsilon_{ij} \quad (2)$$

其中w_{ij}為 UBM 經過 i-Vector extractor 降維後的投影量；m 為模型參數的平均值向量，維度是 F × 1；V 為 eigen-voice，維度是 F × N_1，；U 為 eigen-channel，維度是 F × N_2；ε_{ij}為其餘因素，維度是 F × F，是一個 diagonal covariance ∑；y_i為 eigen-voice 的投影量，x_{ij}為 eigen-channel 的投影量，且y_i和x_{ij}我們假設為高斯分布$\mathcal{g}[0, I]$。

因此，此模型包含了二個部分：

(1)語者成分：m + Vy_i，此部分只包含語者的因素，不包含語者之外的干擾因素，我們又稱為不同語者之間的變化性(between-individual variation)。

(2)通道與雜訊成分：Ux_{ij} + ε_{ij}，此部分只包含通道與雜訊的影響因素，我們又稱為同語者彼此之間的雜訊差異性(within-individual noise)。

以下進一步敘述系統中的 PLDA 模型參數估計與特徵長度正規化演算法：

2.1.1 PLDA 語者模型參數估計

PLDA 機率模型可以用亂數先初始化，然後使用 EM 演算法並且遞回(iteration)好幾次來更新 PLDA 的模型，最終我們要更新到最佳的投影量y_i和x_{ij}。

- M-Step：在這個部分我們要更新的參數是 m、V、U、∑。首先先將(2)重新改寫成(3)所示：

$$w_{ij} = m + [V \quad U] \begin{bmatrix} y_i \\ x_{ij} \end{bmatrix} + \varepsilon_{ij} = m + B z_{ij} + \varepsilon_{ij} \quad (3)$$

接下來用最大似然(Maximum likelihood)估測如(4)：

$$Q(\theta_t, \theta_{t-1}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\int \Pr(z_i \mid w_{i1\ldots iJ}, \theta_{t-1}) \log[\Pr(w_{ij} \mid z_i)\Pr(z_i)]dz_i \quad (4)$$

其中 t 為遞回(iteration)的次數。(4)式裡的 log probability 第一項我們可以改寫成：

$$\log[Pr(w_{ij}|z_i, \theta_t) = K - 0.5 (\log|\Sigma^{-1}| + (w_{ij} - m - Bz_{ij})^T \Sigma^{-1} (w_{ij} - m - Bz_{ij}) \quad (5)$$

其中 K 為不重要的常數。最後我們要更新以下三個參數：

$$m = \frac{1}{IJ}\sum_{i,j} w_{ij} \quad (6)$$

$$B = \left(\sum_{i,j}(w_{ij} - m) \quad E[z_i]^T\right)\left(\sum_{i,j} E[z_i z_i^T]\right)^{-1} \quad (7)$$

$$\Sigma = \frac{1}{IJ}\sum_{i,j} Diag[(w_{ij} - m)(w_{ij} - m)^T - BE[z_i](w_{ij} - m)^T] \quad (8)$$

E-Step：在這個部分我們要同時估算的參數是所有隱藏變數y_i、$x_{i1\ldots iJ}$，因此我們要更新以下二個參數：

$$E[z_i] = (B^T\Sigma^{-1}B + I)^{-1}B^T\Sigma^{-1}(w_i - m) \quad (9)$$

$$E[z_i z_i^T] = (B^T\Sigma^{-1}B^T + I)^{-1} + E[z_i]E[z_i]^T \quad (10)$$

經過多次的遞回(iteration)之後，最後將$E[z_i]$裡面的y_i投影量取出成為我們要的語者因素(speaker factor)。

2.2.2 特徵向量長度正規化

此外，礙於登記語者的語料量不足，以致在(1)式中:w(L)(為包含語者與通道的因素)呈現 Student' t 分布，其共變數屬於 Elliptically Symmetric Densities (ESD)類別[7]，因此我們要利用長度正規化將 i-Vector 的每一句語料轉為高斯分佈，因具有統計獨立的特性，故可去除語料間的相關性，利於之後 PLDA 模型的建立。

以下進一步敘述長度正規化的演算法:

- Step I:藉由 linear whitening transformation 將 i-Vector 後的每一句語料從 ESD 類別轉為 Spherically Symmetric Density (SSD)類別。
- Step II:將每個語句的長度向量藉由 Chi distribution 的自由度等同於其向量的維度的性質，轉為一個標準的高斯分佈的長度向量，而長度分佈的 whitened variable 以 η_{wht} 表示，其長度正規化的方程式如(11)所示：

$$\mathcal{g}(||\eta_{wht}||) = F_X^{-1}F_r(||\eta_{wht}||) \quad (11)$$

(11)式為 inverse cumulative Chi distribution 與長度隨機變數η_{wht}累積分佈的結合運算，此外F_r需要從所有語料中估計其機率分佈。

由於估計隨機長度變數的累積分佈構成了潛在的問題，且依照 NIST SRE 的評估方式，如果這些被評估的語料的長度分佈，要精確的建立其機率模型，必須使用全部的語料。而這種方式違反只能使用所涉及的兩種語料的限制，以產生一個驗證分數。出於這個原因，我們簡化第二步驟，而採用簡單地縮放每個 i-Vector 的向量的長度至單位長度。最後可將其經過 i-Vector 的語料的機率分佈情形由 Student' t 分布轉成高斯分佈，使得我們可將其它非語

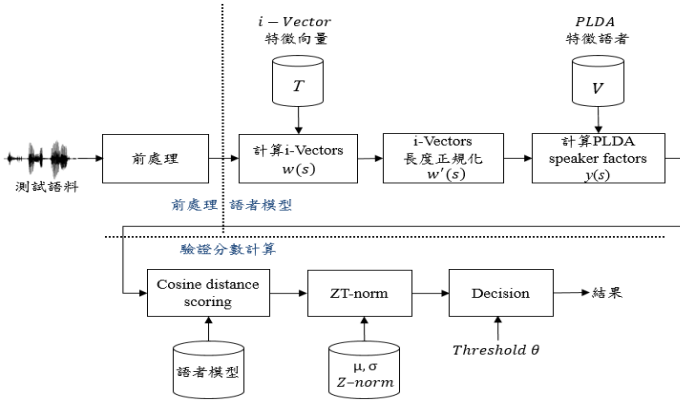
者及通道的因素方便去除。

2.2 系統架構

圖一是我們發展出來的PLDA+Length Normalization系統架構圖，系統中使用 i-Vector，找出建構語者空間的低維度基底: T ，有了空間基底後，就能估算出所有語者在空間基底上的投影量 i-Vectors: $w(L)$ ，然後針對每一句語料個別做 Length Normalization $w^{'}(s)$ ，最後利用 PLDA 機率模型分類資訊，並將同一語者間投影量拉近、不同語者間的投影量拉開，產生新的投影量:y(L)，最後將y(L)新的投影量進行目標語者和測試語料 cosine distance scoring [8]分數計算，其計算公式如下：

$$\text{Score}(w''_{target}, w''_{test}) = \frac{\langle w''_{target}, w''_{test} \rangle}{\|w''_{target}\| \|w''_{test}\|} \quad (12)$$

由於我們將 PLDA 此系統與 LDA 系統進行比較，所以我們也測試 LDA 系統使用長度正規化效能是否可以大大提升，因此我們也提出 LDA 經過長度正規化語者系統。



圖一：PLDA+Length Normalization 系統架構圖

三、 實驗設定與結果分析

3.1 SRE2012 語料庫

我們希望藉由參加 SRE2012 的比賽來檢驗系統效能，因此我們使用了該組織所提供的語料庫進行實驗，NIST SRE12 的測試語料中，首次嘗試隨機加入了雜訊比 SNR15 和 SNR6 的人聲和環境雜訊，非常符合我們日常生活的測試環境，因此也增加語者辨識的困難度。此部分我們希望藉由NIST SRE12的評比來檢驗LDA和PLDA系統效能，本研究中所使用的發展語料庫包括 SRE04-1side、SRE04-3side、SRE04-8side、SRE04-16side、SRE05-1conv4w、SRE05-3conv4w、SRE05-8conv4w、SRE05-mic、SRE06-1conv4w、SRE06-3conv4w、SRE06-8conv4w、SRE06-mic、SRE08-8conv4w、SRE08-followup。

3.2 2012 Speaker Verification:Closed-Set 評比

在 NIST SRE 2012 的任務中，語者辨認系統的效能量測是由決策成本函數(Decision Cost Function;DCF)作為衡量標準，DCF 是錯誤機率的加權總合，加權總和後的值可用來評估系統的好壞，如方程式(13)所示：

$$\text{DCF} = C_{Miss} \times P_{Target} \times P_{Miss|Target} + C_{FlaseAlram} \times P_{Impostor} \times (P_{FlaseAlram|KnownImpostor} \times P_{Known} + P_{FlaseAlram|UnknownImpostor} \times (1 - P_{Known})) \quad (13)$$

其中C_{Miss}和$C_{FlaseAlram}$分別代表錯誤的拒絕與錯誤的接受的代價，P_{Target}和$P_{Impostor}$分別代表目標語者與非目標語者出現的機率，$P_{Miss|Target}$代表錯誤拒絕機率，$P_{FlaseAlram|KnownImpostor}$是錯誤接受機率而且是已知的非目標語者，$P_{FlaseAlram|UnknownImpostor}$是錯誤接受機率而且是未知的非目標語者。為了符合實際的需要，NIST 給定的成本權重設定為 $C_{Miss} = 1$ ， $C_{FlaseAlram} = 1$ ，$P_{Target-A1} = 0.01$，$P_{Target-A2} = 0.001$，$P_{Known} = 0.5$。因為 SRE12 評比中P_{Target}給定測試語者與目標語者，故其決策函數分別為DCF_{A1}及DCF_{A2}，因此最後的 DCF 判斷如方程式(14)所示:

$$\text{Final DCF} = \frac{DCF_{A1} + DCF_{A2}}{2} \quad (14)$$

3.3 評比方式

前處理的部分，系統使用的聲特徵參數皆為 39 維的梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCC)，在求得 MFCCs 後將其參數移除靜音部分，並將移除靜音後得聲特徵參數利用(ARMA Filtering;MVA)來過濾環境所造成的雜訊，最後的參數正規化我們採用(Histogram Equalization;HEQ)來改善訓練語料及測試語料因通道所造成不匹配的情形，進而產生強健之語料。

我們利用使用兩套系統分別做實驗，系統 A 為 LDA+Length Normalizatin，系統 B 為 PLDA+Length Normalizatin，並分別測試沒加 Length Normalizatin。

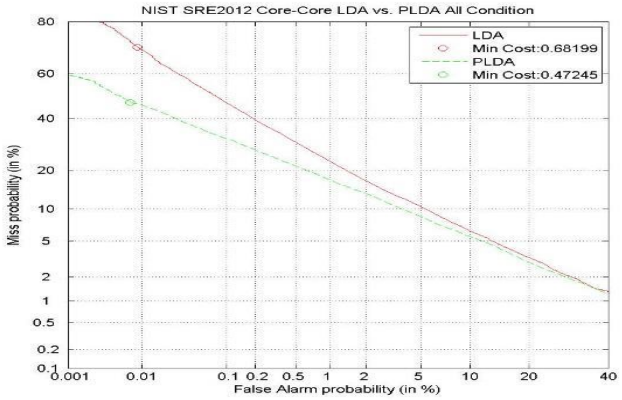
3.4 SRE2012 Speaker Verification 實驗結果與比較

表 I 為無正規化條件下的 LDA 與 PLDA 之 Core-Core 之 Min Cost 與 EER；圖二則為詳細的 PLDA 與 LDA(無長度正規化)的 DET Curve(all-condition)的系統比較圖，可以發現在沒有做長度正規化，PLDA 的系統效能較 LDA 系統效能顯著提升；表 II 為有正規化條件下的 LDA 與 PLDA 之 Core-Core 之 Min Cost 與 EER。圖三則詳細的 PLDA 與 LDA(長度正規化)的 DET Curve(all-condition)的系統比較圖，從表一，表二，圖二與圖三的結果，可証實有加入長度正規化的結果較優；

圖四為 SRE2012 Core-Core 各參賽組織(共約 40 隊)之 DET 曲線比較圖，可跟圖三本系統之 DET 曲線比較圖做比較，此外，表III為在各種試驗下，本系統與此次競賽的最佳結果的 Min Cost 的詳細比較。可發現我們提出的系統與國外最優秀的驗證系統雖然還有一定程度的差距，但若與全部約 40 組隊伍來比較，我們的成績還算不錯。

表 I LDA vs. PLDA 於 SRE12 Core-Core 之 Min Cost 與 EER

Condition	LDA		PLDA	
	Min Cost	EER	Min Cost	EER
All condition	0.682	7.65%	0.4725	6.92%
Interview-NoAddedNoise	0.4428	6.69%	0.3601	6.38%
Telephone-NoAddedNoise	0.6884	7.46%	0.4958	6.71%
Interview-AddedNoise	0.5396	6.95%	0.3237	5.64%
Telephone-AddedNoise	0.6653	7.96%	0.5472	7.66%



圖二:PLDA 與 LDA(non-normalization)之 DET 曲線圖

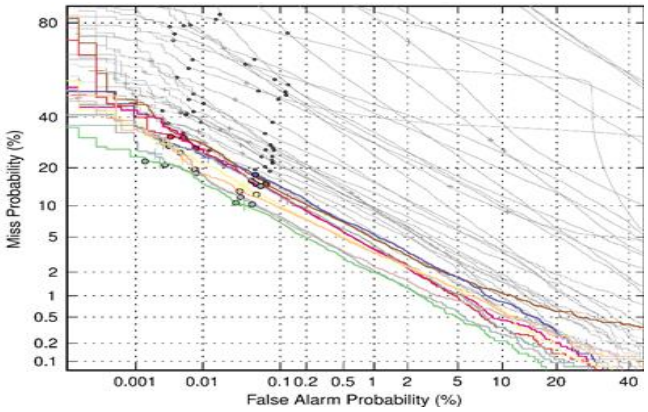
表Ⅱ LDA vs. PLDA 於 SRE12 Core-Core 之 Min Cost 與 EER(長度正規化)

Condition	LDA		PLDA	
	Min Cost	EER	Min Cost	EER
All condition	0.5692	6.93%	0.4697	6.79%
Interview-NoAddedNoise	0.3462	6.14%	0.3615	6.31%
Telephone-NoAddedNoise	0.5750	6.71%	0.4857	6.48%
Interview-AddedNoise	0.4300	6.11%	0.3220	5.64%
Telephone-AddedNoise	0.6183	7.44%	0.5492	7.50%

圖三:PLDA 與 LDA(normalization)之 DET 曲線圖

表 Ⅲ SRE12 最佳效能紀錄與本系統之 Min Cost 比較

Condition	SRE12 Best	Our System
All condition	0.1878	0.4697
Interview-NoAddedNoise	0.1836	0.3615
Telephone-NoAddedNoise	0.1846	0.4857
Interview-AddedNoise	0.1645	0.3220
Teleph-AddedNoise	0.1394	0.5492



圖四: 所有參賽團隊之 DET 曲線圖

結論

本文提出的 PLDA 搭配特徵長度正規化發法，建立語者驗證系統，並實踐於 NIST SRE2012 語料庫上。實驗結果顯示，若使用長度正規化，PLDA 和 LDA 相比，PLDA 的 Min Cost 相對效能增益為 17.48%，EER 相對效能增益為 2.02%。若和所有參加 SRE 2012 的參賽團隊（約 40 隊）的系統比較，我們的系統表現還不錯，因此使用 PLDA 機率模型並搭配特徵長度正規化，的確能有效對抗通道與雜訊干擾、提升語者辨認效能。

參考文獻

[1] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in Proc. ICCV’07, Oct. 2007, pp. 1-8

[2] Daniel Garcia-Romero and Carol Y. Espy-Wilson., “Factor Analysis of i-Vector Length Normalization in Speaker Recognition Systems,”2011 Department of Electrical and Computer Engineering, University of Maryland, College Park, MD .2010

[3] A. Hatch, S. Kajarekar, and A. Stolcke, “Within-Class Covariance Normalization for SVM-Based Speaker Recognition,” in International Conference on Spoken Language Processing, Pittsburgh, PA, USA, Sept. 2006, pp. 1471-1474.

[4] NIST Speaker Recognition Evaluation, http://www.nist.gov/itl/iad/mig/sre12.cfm

[5] Dehak, N., “Discriminative and Generative Approches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Verification,” Ph.D. thesis, ´Ecole de Technologie Sup´erieure, Montreal, 2009

[6] P Kenny Joint factor analysis of speaker and session variability: Theory and algorithms, CRIM, Montreal,(Report) CRIM-06/08-13, 2005

[7] S. Lyu and E. P. Simoncelli, "Nonlinear Extraction of Independent Components of Natural Images using Radial Gaussianization," Neural Computation, vol. 21, no. 6, June 2009.

[8] Dehak, N., “Discriminative and Generative Approches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker..