The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 31-41

基於稀疏表示之語者識別

Sparse Representation Based Speaker Identification

王光耀　Kuang-Yao Wang
國立中央大學資訊工程學系
Department of Computer Science and information Engineering
National Central University

王家慶　Jia-Ching Wang
國立中央大學資訊工程學系
Department of Computer Science and information Engineering
National Central University
jcw@csie.ncu.edu.tw

摘要

稀疏表示分類器(Sparse Representation Classifier, SRC)是一種基於影像稀疏表示(Sparse Representation)的機器學習方法。在影像以及人臉辨識上的研究上，稀疏表示分類器具有非常好的辨識效果以及強健性。有鑑於 SRC 在影像辨識上的高鑑別能力，近幾年已有許多基於稀疏表示的語者識別(Speaker Identification)方法相繼被提出。本論文提出一套基於稀疏表示的辨識系統，我們提出以機率型主成份分析(Probabilistic Principle Component Analysis, PPCA)建構超級向量(Supervector)，並加入檢定的方式調整特徵值選取，使語者高斯混合模型(Gaussian Mixture Model, GMM)中每個高斯的維度可以針對資料的不同作調整。接著，我們在稀疏字典上加強，透過低秩矩陣還原(Low-Rank Matrix Recovery)以及核化降維(Kernel Dimension Reduction)對說話內容(Session)以及通道(Channel)變異補償，使字典增加鑑別性。最後利用稀疏表示分類器進行分類。根據實驗結果顯示，不論是參數的改進、字典的處理，對辨識率都有一定程度的提升。此外，與傳統的 *i*-vector 語者識別系統相比，提出的系統則具有更佳的辨識率表現。

關鍵詞：稀疏表示、語者識別、核化稀疏表示、超級向量

一、簡介

以生物特徵作為辨識基礎的研究已經有長達數十年的歷史，包含人臉、語音和指紋。其中聲音因具有取得容易、非侵入性、運算量少、輸入具有便利性等優點，語者識別一直是近年來熱門的主題。語者識別利用語者語音的特性來識別使用者身份，大致上的研究方向分成語音特徵擷取以及分類演算法兩部分，其中以分類演算法的研究最多，

包含的層面也最廣，包括語者模型的建立、分類機制、因為周邊環境的影響所需的補償辦法等研究。目前來說高斯混和模型 [1][2]和支持向量機(Support Vector Machine, SVM)[3]是經典的分類方法。特徵參數方面，會將一段語音切割成數個音框，音框代表此語者在短時間內的發聲特徵，幾種常見的有梅爾倒頻譜係數(Mel-scale Frequency Cepstral Coefficients, MFCC)[5] 、 感 知 線 性 預 測 係 數 (Perceptual Linear Prediction Coefficients, PLPC)[6])等。

基於 GMM 超級向量(GMM-Supervector)[7-10]的辨識方法是一種綜合 GMM 模型表示以及 SVM 分類方法優點的語者識別演算法。GMM 超級向量是一種語者模型的向量表示法，通用背景模型(Universal Background Model, UBM)經過輸入語音調適後成為 GMM 語者模型，其中各個高斯成份(Components) 的期望值被串在一起形成一個超級向量。此向量表示可以用來代表整個語音檔的特徵向量，最後，再以 SVM 作為分類方法來達到辨識語者的目的。然而語者識別系統會因為錄製工具和背景不同，造成通道差異、說話內容差異和環境差異的干擾，導致辨識系統的辨識率低落。擾動屬性投影(Nuisance Attribute Projection, NAP)是一種有效補償通道干擾的方式[7][11]，藉由設計一個最佳化問題使干擾最小，配合著前面提到的方法，更有效地提高辨識系統的強健性。在 NAP 之後陸續有聯合因素分析(Joint Factor Analysis , JFA)[12][13]、*i*-vector[14-16]等技術被用來補償上述提到的差異，透過對語音參數的拆解，將一些語者不相關的資訊刪除，得到去除干擾後最能代表語者的特徵。

近年來對於複雜的資訊(例如訊號、圖像)，往往希望可以用較簡化的方式呈現，特別在訊號處理的部分，分析時經常先將資料轉換至不同的定義域，並且假設其在轉換後，會呈現稀疏分布[24]。在近期稀疏表示的發展中，SRC 在[25][26]中被提出，SRC 是一個 Nonparametric 學習方法，不需訓練過程但是需要訓練資料，以及可以直接參考訓練資料對策是資料進行分類的動作。實驗結果顯示出在人臉辨識的應用上，SRC 有著比 KNN (K-Nearest Neighbors)[27]以及 Nearest Subspace(NS)[28][29]更好的辨識率。近幾年也有少數研究將 SRC 應用於語者驗證[16-19]的問題上，然而此方面的研究仍屬於剛起步的階段，仍有許多問題值得我們探討。

本論文提出一套基於核化稀疏表示的語者識別系統，文章的編排共分成七個部分:第一部分為簡介，第二部分為參數擷取，我們以 PPCA 建構超級向量(PPCA 超級向量)取代 GMM 超級向量[20][31]，並且以巴雷特檢定(Bartlett Test)的方式調整特徵值的選取，第三部分則是描述如何利用特徵參數建立稀疏表示分類器。第四部分描述我們提出的兩種變異補償方法，低秩矩陣還原以及核化稀疏表示分類器(Kernel Sparse Representation Classifier, KSRC)。第五部分為實驗部分，展示提出之改良方法是否具有其必要性。最後在第六部分則是結論。

二、參數擷取

在本論文中，我們利用機率型主成分分析建構超級向量[20][33]，並以提出巴雷特檢定(Bartlett Test)主成分的個數。傳統上，超級向量是由高斯混和模型建構而成，這裡

加入主成分分析的概念，並希望能以機率分布模型的形式與高斯模型對應，使得資料點由原本高斯混和模型轉成 PPCA 混和模型，再透過 Latent Factor 的轉換，形成新的超級向量，稱作 PPCA 超級向量[20][21][33]，其中，主軸的挑選，我們引入巴雷特檢定(Bartlett Test)[22][23]的概念，建立假說，找到臨界的特徵值。而目前語者識別問題中，*i*-vector 是表現最好的參數之一。藉由訓練出總體變異矩陣，將原本的超級向量轉到更低維的空間，使 *i*-vector 更加表現出語者及通道的資訊，因此，我們將 PPCA 超級向量轉換到總體變異空間上，希望得到更具鑑別力的 *i*-vector 使辨識效能提升。

2.1 基於機率型主成分分析之因素分析模型

傳統的 GMM 超級向量，並沒有考量到其聲學特徵參數有高度的冗餘性[21][31]，因此應該採用更低的子空間來表示。在[21][31]中比較成份分析(Factor Analysis)與主成分分析(PCA)的關聯性，觀察成份分析的數學式：

$$\mathbf{x}=\mathbf{W}\mathbf{z}+\boldsymbol{\mu}+\boldsymbol{\varepsilon} \tag{1}$$

其中 **x** 代表高維的資料，**z** 代表低維變數，又稱 Latent Factor。回想主成分分析的數學式，找出主成分 **V** 使資料降維：

$$\mathbf{x}=\boldsymbol{\mu}+\mathbf{V}\mathbf{z} \tag{2}$$

我們可將成份分析視成對資料 **x** 做 PCA 加上一個噪音項，並且導入機率的概念解釋。

在式(1)中，**x** 是 $K\times1$的原始資料， **W** 為 $K\times J$ 的轉換矩陣，其中 $J<K$ ，而 Latent Factor **z** 假設為一高斯分布 $N(0,\mathrm{I})$，噪音 $\boldsymbol{\varepsilon}\sim N(0,\sigma^2\mathrm{I})$，根據以上的假設可以知道原始資料 x 能建模(Model)成 $N(\boldsymbol{\mu},\sigma^2\mathbf{I}+\mathbf{W}\mathbf{W}^{\mathrm{T}})$ ，稱作 PPCA 模型。當遇到更複雜的資料時，希望藉由混合多組 PPCA 模型，如下式：

$$p(\mathbf{x})=\sum_{c=1}^{M} w_c\, p(\mathbf{x}|c) \tag{3}$$

$$p(\mathbf{x}|c)=N(\boldsymbol{\mu}_c,\sigma_c{}^2\mathbf{I}+\mathbf{W}_c\mathbf{W}_c{}^{\mathrm{T}}) \tag{4}$$

其中 M 是高斯成份的個數，這可以跟高斯混合模型做對應，一般表示資料以多個高斯模型描述，這裡的基本單位以 PPCA 模型取代。

因此，我們希望找出 $\mathbf{W}_c$ 及 σ_c 來推算 Latent Factor，方法是利用最大概似估計(Maximum Likelihood Estimation, MLE)，因為已經知道 **x** 的分布，所以藉由 MLE 得到的 $\mathbf{W}_c$ 及 σ_c 可進而推出 Latent Factor $\mathbf{z}_c$

$$\mathbf{z}_c=(\mathbf{W}_c{}^T\mathbf{W}_c+\sigma_c{}^2\mathbf{I})^{-1}\mathbf{W}_c{}^T(\mathbf{x}-\boldsymbol{\mu}_c) \tag{5}$$

當原始資料 **x** 進入系統後，會先以 PPCA 混合模型，並且透過式(5)的轉換得到屬於

每個高斯成份的 Latent Factor。當所有的參數都做了 PPCA 後，則原始以 MFCC 為基礎的 UBM 也必須調整，形成以 Latent Factor 為參數下構建的新 UBM：

$$p(\mathbf{z}|c)=\sum_{c=1}^{M} w_c N(\boldsymbol{\mu}_{z,c},\boldsymbol{\Sigma}_{z,c})=\sum_{c=1}^{M} w_c N(0,\mathbf{I}-\sigma_c{}^2\boldsymbol{\Lambda}_c{}^{-1}) \tag{6}$$

$$\text{其中}\quad \boldsymbol{\mu}_{z,c}=E\{(\mathbf{W}_c{}^T\mathbf{W}_c+\sigma_c{}^2\mathbf{I})^{-1}\mathbf{W}_c{}^T(\mathbf{x}-\boldsymbol{\mu}_c)\}=0 \tag{7}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{z,c}&=E\{\mathbf{z}_c\mathbf{z}_c{}^T\}-\boldsymbol{\mu}_{z,c}\boldsymbol{\mu}_{z,c}{}^T\\&=(\mathbf{W}_c{}^T\mathbf{W}_c+\sigma_c{}^2\mathbf{I})^{-1}\mathbf{W}_c{}^T E\{(\mathbf{x}-\boldsymbol{\mu}_c)(\mathbf{x}-\boldsymbol{\mu}_c)^T\}\mathbf{W}_c(\mathbf{W}_c{}^T\mathbf{W}_c+\sigma_c{}^2\mathbf{I})^{-T}\\&=\boldsymbol{\Lambda}_c{}^{-1}(\boldsymbol{\Lambda}_c-\sigma_c{}^2\mathbf{I})^{T/2}\boldsymbol{\Lambda}_c(\boldsymbol{\Lambda}_c-\sigma_c{}^2\mathbf{I})^{1/2}\boldsymbol{\Lambda}_c{}^{-T}=\mathbf{I}-\sigma_c{}^2\boldsymbol{\Lambda}_c{}^{-1}\end{aligned} \tag{8}$$

2.2 巴雷特檢定

在上個小節中提到的 PPCA 中，有一個問題是需要討論的，那就是主軸個數的挑選。我們假設後面不要的特徵值，其數值小到足以視為相同，因此，定義假說 $H_0:\lambda_{k+1}=\lambda_{K+2}=...=\lambda_n$，希望迴圈由 *n*-1 往前找到門檻值 *k*。而如果我們將特徵值的大小視為高斯分布，則巴雷特檢定[22][23]可以整理成相似度的檢驗，如下式：

$$T=\frac{\prod_{q=k+1}^{n}\lambda_q}{\left(\frac{\sum_{q=k+1}^{n}\lambda_q}{n-k}\right)^{n-k}} \tag{9}$$

當觀測資料數 *m* 夠多的話，可將 T 視為一個 X^2 分布，且將上式近似於下式：

$$T\approx\left(m-(2n-11)/6\right)\left((n-k)\log\bar{\lambda}-\sum_{q=k+1}^{n}\log\lambda_q\right) \tag{10}$$

其中 $\bar{\lambda}$ 是後 *n-k* 個特徵值的平均，在檢定過程中，當 $T>X^2{}_{\alpha,n}$ 時，則否決假說 H_0，檢定終止，*q* 即為我們挑選的主軸個數，其中 α 為 X^2 的顯著水平。

2.3 *i*-vector

啟發於早期 JFA 在語者識別上的應用，Dehak *et al*. 提出的一個新的分析方法 *i*-vector [14-16]，不像 JFA 將語者和通道分開，*i*-vector 僅用一個總體變異性空間，他發現 JFA 中的通道部分仍包含了能用來識別語者的資訊，所以將 JFA 中分開的變異部分，合併成一個單一的總體變異性空間超級向量，藉由合併錄音方式的變異性，提升其辨識性，表示如下：

$$\boldsymbol{\mu}=\mathbf{m}+\mathbf{T}\mathbf{w} \tag{11}$$

m 是 UBM 超級向量，與 JFA 中使用的相同；**T** 代表的是所有變異性的矩陣；*i*-vector 代表的是總體變異性元素 **w**。

最後，總結參數擷取的整體架構，參數擷取包含三個部分，第一，由 Universal Background Data 首先訓練出 UBM，並透過 PPCA 將 UBM 轉換成以 Latent Factor 為參數的 UBM，第二，當輸入語音進入系統後，擷取其 Latent Factor，接著，對新的 UBM 調適產生 PPCA 超級向量。最後，在基於 PPCA 超級向量參數下訓練總變異矩陣，並將輸入語音轉換成 *i*-vector。

三、稀疏表示分類器

我們利用稀疏表示具鑑別力的特性，應用於語者識別問題上，利用 *i*-vector 作為特徵參數，以固定維度的向量表示語音訊號。假設有 *C* 個不同語者類別的訓練資料，每一筆訓練資料為一個 *i*-vector，我們需要建構一個跨類別的(Global)字典 $\mathbf{D}\in\mathbb{R}^{P\times Q}$，作法是將所有語者類別的字典組在一起

$$\mathbf{D}=[\mathbf{D}_1\ \mathbf{D}_2.....\mathbf{D}_C] \tag{12}$$

其中，$\mathbf{D}_j$ 表示第 *j* 個類別的字典，由類別 *j* 的 *i*-vector 串接而成。此外，*P* 代表參數維度，*Q* 是訓練資料個數。在測試資料 *y* 與字典 *D* 已知的情況下，我們希望求出稀疏係數 $\mathbf{x}=[x_1\ x_2\ ...\ x_k]$，使原訊號與重建訊號之間的誤差能越小越好，且 **x** 要符合稀疏特性，如下式：

$$\min_x \|\mathbf{y}-\mathbf{D}\mathbf{x}\|_2^2+\lambda\|\mathbf{x}\|_1 \tag{13}$$

在解出稀疏係數 *x* 後，決策上，將 *k* 類別各自的字典 $\mathbf{D}_j$ 及係數 $\mathbf{x}_j$ 還原 **y**，並與 **y** 計算誤差，獲得最小誤差者即為所屬類別：

$$j^*=\arg\min_j \|\mathbf{y}-\mathbf{D}_j\mathbf{x}_j\|_2^2 \tag{14}$$

四、字典處理及變異補償

在以 *i*-vector 為參數的架構下，*i*-vector 的總變異矩陣是假設訓練資料標籤不同所訓練而成，因此，它存在與語者相關的變異，這是我們想要的，同時它存在一部分錄音通道及說話內容造成的變異，因此，以往的研究都會加入 Linear Discriminant Analysis(LDA)及 Within-Class Covariance Normalization (WCCN)去補償錄音通道的變異[33]，希望即使是在錄音通道不同的狀態下，找出類別間的變異最大化，類別內的變異最小化的空間，使變異排除，另外說話內容的變異也可以由 LDA 的處理解釋，因為它將語者的說話內容做變異最小的假設，消除說話內容變異造成的問題。

本節提出對 *i*-vector 字典的處理及補償辦法，包括以低秩矩陣還原以及核化稀疏表示分類器(Kernel SRC)分別對說話內容及錄音通道變異做補償，並增加字典的鑑別性。

4.1 低秩矩陣還原字典處理及變異補償

低秩矩陣還原[34]提供一種訊號拆解的方式，假設訊號可以被拆解成一個低秩(Low Rank)矩陣 **A** 及稀疏誤差(Sparse Error) **E** 的和，如下式：

$$\mathbf{D} = \mathbf{A} + \mathbf{E} \tag{15}$$

在 **A** 與 **E** 皆未知的情況下，我們希望讓 **A** 能以最少的 rank 還原原訊號，並且 **E** 能符合 Sparse 的特性，因此整理成以下的最佳化式：

$$\min_{\mathbf{A},\mathbf{E}} rank(\mathbf{A}) + \gamma \|\mathbf{E}\|_0 \tag{16}$$

進而，整理成一個凸最佳化的問題，並用 Augmented Lagrange Multiplier (ALM)求解。

透過低秩矩陣還原，我們將每個語者類別的字典分別作低秩矩陣還原分解，得到與原來字典大小相同的低秩矩陣 **A**，取代原來的字典，讓字典中每個語者的特徵更為凸顯。

4.2 核化稀疏表示分類器

藉由核化方法(Kernel Trick)，稀疏表示分類器可進一步非線性化，稱作核化稀疏表式分類器 [32]。輸入空間的資料經過非線性核化映射(Kernel Mapping)投射至高維參數空間，讓原本在輸入空間混淆的參數資料在高維空間變成可分離的。透過更進一步的核化降維方法，我們可得到測試資料的稀疏組合係數(Sparse Combination Coefficients)來進行分類的動作。

我們利用稀疏表示具鑑別力的特性，應用於語者識別問題上，利用 *i*-vector 作為特

徵參數，以固定維度的向量表示語音訊號。假設有 c 個不同語者類別的訓練資料，每一筆訓練資料為一個 i-vector，我們需要建構一個跨類別的(Global)字典 D，作法是將所有語者類別的字典組在一起 $\{\mathbf{a}_i, y_i\}_{i=1}^{Q}$，其中 $\mathbf{a}_i \in \mathbb{R}^P$, $y_i \in \{1,2,...,c\}$, 一筆測試資料 $\mathbf{a} \in \mathbb{R}^P$. 令 Φ 表示 $k(\cdot,\cdot)$ 的非線性映射，可將輸入空間 Input Space χ 的資料投射到高維空間 F：

$$\Phi : \mathbf{a} \in \chi \to \Phi(\mathbf{a}) \in F \tag{17}$$

與稀疏表示分類器相同，參數空間的測試資料可表示為訓練資料的線性組合：

$$\Phi(\mathbf{a}) = \sum_{i=1}^{Q} x_i \Phi(\mathbf{a}_i) = \mathbf{\Phi_A x} \tag{18}$$

其中 $\mathbf{\Phi_A} = \left[\Phi(\mathbf{a}_1), \Phi(\mathbf{a}_2), ..., \Phi(\mathbf{a}_Q)\right]$，且 $\mathbf{x} = [x_1, x_2, \ldots, x_Q]^T$。

根據稀疏表示的概念，稀疏係數向量 $\mathbf{x}$ 可以從下式最佳化問題解出：

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \|\Phi(\mathbf{a}) - \mathbf{\Phi_A x}\|_2 \le \varepsilon \tag{19}$$

當核化空間的維度是未知且遠高於訓練資料的個數時，會導致式子(14)的解會不夠稀疏，因此會需要在參數空間進行降低維度的動作。令 $\mathbf{P}$ 表示投射矩陣 (Projection Matrix)。則(19)可改成下式：

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to} \left\|\mathbf{P}^T \Phi(\mathbf{a}) - \mathbf{P}^T \mathbf{\Phi_A x}\right\|_2 \le \varepsilon \tag{20}$$

具有最小重建誤差(Reconstruction Residual)的類別則為核化稀疏表示分類器的分類結果：

$$\hat{i} = \arg\min_{i} r_i(\mathbf{a}) = \left\|\mathbf{P}^T \Phi(\mathbf{a}) - \mathbf{P}^T \mathbf{\Phi_A} \delta(\mathbf{x}_i)\right\| \tag{21}$$

其中 $\delta(\mathbf{x}_i)$ 是 $\hat{\mathbf{x}}$ 的第 i 個語者的非零稀疏系數。

Kernel SRC(KSRC)即結合 Kernel LDA 及 SRC 兩個方法，將原始特徵向量由 Kernel 投影至高維後，再由 LDA 降維，在先前的文獻有提到，i-vector 存在 Channel 的變異，因此，我們透過 Kernel LDA 補償，希望錄製方式不同產生的干擾能藉由最大化 Between-Class 變異及最小化 Within-Class 變異得到補償。

五、實驗結果

我們以 NIST2005 [30] 做為語者資料庫，NIST 每年都會錄製語者資料庫，許多產業界、學術界都會以此作為評估效能的標準，而 2005 年發布的資料庫以電話對話語音數據為主，同時收集一些輔助麥克風接收的數據，這些數據主要來自英語演講，並包括四種額外語言。我們挑選其中 Male 8con-1con 的 Condition 做為測試資料庫，其中 8con 為訓練資料，而 1con 是測試資料。

為了驗證巴雷特檢定是否能進一步改善辨識系統，我們定義另一套主軸個數選取方式:各個高斯成份在統一主軸個數下與透過檢定方式挑選適合主軸個數下的差別，最後

再經由轉換到 *i*-vector 上。藉由比較巴雷特檢定與我們定義的主軸選取方式來判斷巴雷特檢定的必要性。在我們的實驗中，如表一所示，固定每個高斯成份的主軸個數為 27 時來到最高 76.91%，不過隨著主軸個數的減少辨識率也降低。而加入巴雷特檢定後，我們設定巴雷特檢定的 $\alpha = 0.05$。各個高斯成份降維的數目不固定，有效的提升辨識率到 77.01%。在接下來的實驗中，我們會以這個辨識率為 77.01%的系統作為後續實驗的基準(Baseline)，測試提出/採用的改進方法是否有效。

《表一》不同主軸個數與巴雷特檢定選取之辨識率比較。

方法			辨識率
PPCA-SV	30個主軸	+ *i*-vector + SRC	76.60%
	27個主軸		76.91%
	24個主軸		76.40%
	21個主軸		76.20%
	18個主軸		75.99%
	15個主軸		72.23%
PPCA-SV + Bartlett Test + *i*-vector + SRC			77.01%

在以下實驗中，我們比較了四種不同語者識別系統。Baseline 我們使用目前 state-of-the-art 辨識方法，*i*-vector based cosine distance，即以 *i*-vector 為參數下，將輸入特徵與每個 Class 字典的 8 筆 *i*-vector 做內積求平均，挑選內積最大的語者類別，表示相似度為最大，通常在參數後面會加上 LDA 與 WCCN 對錄音通道變異補償。其餘三個系統分別為基於稀疏表示器的系統、基於核化稀疏表示器作字典補償之識別系統以及基於低秩矩陣還原之識別系統。實驗數據如表二所示，透過低秩矩陣還原方法建構的字典具有最高的辨識率 80.57%，比 Baseline 以及沒有變異補償的系統分別多了 13.63%以及 3.56%的辨識率。與沒有變異補償的系統相比，Kernel LDA 達到變異性補償的效果，增加了字典的鑑別性，辨識率提升了 2.24%。

《表二》不同語者識別系統之辨識率比較。

方法	辨識率
GMM-SV + *i*-vector + LDA + WCCN + CD (Baseline)	66.94%
PPCA-SV + Bartlett Test + *i*-vector + SRC	77.01%
PPCA-SV + Bartlett Test + *i*-vector + Kernel SRC	79.25%
PPCA-SV + Bartlett Test + *i*-vector + SRC (Low-Rank Matrix Recovery Based Dictionary)	80.57%

六、結論

這篇論文提出一套基於稀疏表示分類器為基礎的辨識系統，在前端以 PPCA-Supervector 為參數，加入巴雷特檢定作為準則，決定每個高斯 Component 挑選主軸的辦法，使每個高斯 Component 的維度可以針對資料的不同，決定適當的維度，接著，訓練出總變異矩陣，將 Supervector 投映至總變異空間上，以 *i*-vector 作為辨識參數。在字典的建構上，我們提出以低秩矩陣還原以及 Kernel SRC 進行變異補償，去除說話內容以及通道變異造成的干擾。從實驗結果看來，PPCA-Supervector 在未做巴雷特檢定前就有比 GMM-Supervector 好的效果，而再加入巴雷特檢定後，效果更加提高，與傳統基於 *i*-vector 的識別系統相比，辨識率提升了 10.07%。此外，再加入兩種變異補償方法後，低秩矩陣還原以及 Kernel SRC 分別可以再提升系統的辨識率 3.56%以及 2.24%。

參考文獻

[1] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture models," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[2] B. L. Pellom and J. H. L. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Process. Lett.*, vol. 5, no. 11, pp. 281–284, Nov. 1998.

[3] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, pp. 210–229, 2006.

[4] J. L. Gauvain and C. H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[5] M. R. Hasan, M. Jamil, M. G. Rabbani, and M. S. Rahman, "Speaker identification using Mel frequency cepstral coefficients," *3rd international Conference on Electrical & Computer Engineering ICECE 2004*, 28-30 December 2004, Dhaka, Bangladesh.

[6] H. Hermansky. "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoust. Society ofAmer.*, 87: 1738- 1752, April, 1990.

[7] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and nap variability compensation," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 97–100.

[8] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Process. Lett.*, vol. 13, no. 5, pp. 308–311, May 2006.

[9] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2010.

[10] B. G. B. Fauve, D. Matrouf, N. Scheffer, J.-F. Bonastre, and J. S. D. Mason, "State-of-the-art performance in text-independent speaker verification through open-source software," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 1960–1968, 2007.

[11] A. Solomonoff, W. M. Campbell, and C. Quillen, “Channel compensation for SVM speaker recognition,” *in Proc. Odyssey04*, 2004, pp. 57–62.

[12] O. Glembek, L. Burget, N. Brummer, and P. Kenny, “Comparison of scoring methods used in speaker recognition with joint factor analysis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 4057–4060.

[13] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of interspeaker variability in speaker verification,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 980–988, Jul. 2008.

[14] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, “i-vector based Speaker Recognition on Short Utterances,” in *Interspeech*, 2011.

[15] P. Matejka, O. Glembek, F. Castaldo, O. Plchot, P. Kenny, L. Burget, and J. Cernocky, “Full-covariance ubm and heavy-tailed plda in i-vector speaker verification,” *Proc. ICASSP ’11*, pp. 4828–4831, 2011.

[16] J.M.K. Kua, J. Epps, E. Ambikairajah, “i-vector with sparse representation classification for speaker verification, ” *Speech Commun*, 2013.

[17] K. Huang and S. Aviyente, “Sparse Representation for Signal Classification,” *Neural Information Processing Systems*, 2006.

[18] J. M. K. Kua, E. Ambikairajah, J. Epps, and R. Togneri, “Speaker verification using sparse representation classification,” in *Proc. ICASSP*, May 2011, pp. 4548–4551.

[19] R. Saeidi, A. Hurmalainen, T. Virtanen, and D. A. van Leeuwen, “Exemplar-based Sparse Representation and Sparse Discrimination for Noise Robust Speaker Identification,” in *Odyssey speaker and language recognition workshop*, Singapore, 2012.

[20] T. Hasan and J. H. L. Hansen, “Factor analysis of acoustic features using a mixture of probabilistic principal component analyzers for robust speaker verification,” *in Proc. Odyssey*, Singapore, Jun. 2012.

[21] M. Tipping and C. Bishop, “Mixtures of probabilistic principal component analyzers,” *Neural Computation*, vol. 11, no. 2, pp. 443–482, 1999.

[22] M. Y. Lee, Perceptual Factor Analysis for Speech Enhancement, Master Thesis, Institute of Computer Science and Information Engineering, National Cheng Kung University (2004).

[23] M.S. Bartlett, “Tests of significance in factor analysis,” *British Journal of Psychology*, Statistical Section 3, 77–85, 1950

[24] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: an algorithm for designing overcomplete dictionaries for Sparse Representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[25] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, “Feature selection in face recognition: A sparse representation perspective,” EECS Dept., Univ. California, Berkeley, CA, Tech. Report UCB/EECS-2007-99, 2007.

[26] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–226, 2009.

[27] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd edition New York: Wiley, 2000.

[28] S. Z. Li, “Face recognition based on nearest linear combinations,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 839–844, 1998.

[29] K. C. Lee, J. Ho, and D. Kriegman, “Acquiring linear subspaces for face recognition under variable lighting,” *IEEE Transaction Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, pp. 684–698, 2005.

[30] NIST 2005 Speaker Recognition Evaluation plan, http://www.nist.gov/speech/tests/spk/2005/sre-05_evalplan-v6.pdf.

[31] T. Hasan and J. H. L. Hansen, “Acoustic factor analysis for robust speaker verification,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no.4, pp.842-853, Apr. 2013.

[32] L. Zhang, W. D. Zhou, P. C. Chang, J. Liu, Z. Yan, T. Wang, and F. Z. Li, “Kernel sparse representation-based classifier,” *IEEE Trans. Signal Processing*, vol. 60, no. 4, pp. 1684–1695, Apr. 2012.

[33] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, M. Mason, “Weighted LDA techniques for i-vector based speaker verification,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4781–4784, 2012.

[34] J. Wright, A. Ganesh, S. Rao, and Y. Ma, “Robust Principal Component Analysis: Exact Recovery of Corrupted Low-Rank Matrices via Convex Optimization,” Submitted to the Journal of the ACM, 2009.