資料導向線性特徵轉換於中文大詞彙連續語音辨識之應用

陳鴻彬
國立台灣師範大學
資訊工程研究所
james@csie.ntnu.edu.tw

張志豪
國立台灣師範大學
資訊工程研究所
g92470205@csie.ntnu.edu.tw

陳柏琳
國立台灣師範大學
資訊工程研究所
berlin@csie.ntnu.edu.tw

摘要

本論文探討各種資料導向線性特徵轉換和各種結合頻域-時域(Spatial-Temporal)資訊方法於中文大詞彙連續語音辨識(Large Vocabulary Continuous Speech Recognition, LVCSR)之應用。首先，本論文將線性鑑別分析(Linear Discriminant Analysis, LDA)應用在語音特徵空間轉換以及時域與頻域資訊的結合，並且與傳統語音特徵擷取方式作一系列的比較。再者，本論文研究幾種線性鑑別分析的改進方法，諸如異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA)和異質性鑑別分析(Heteroscedastic Discriminant Analysis, HDA)。最後，我們探討噪音對於上述各種資料導向線性特徵轉換技術的影響。本論文以中文廣播新聞為實驗語料，實驗內容除了中文大詞彙連續語音辨識外，亦包括中文自由音節辨識(Free Syllable Decoding)。

In this paper, we studied several kinds of data-driven linear feature transformation techniques, as well as the integration of spatial-temporal information, for Mandarin Large Vocabulary Continuous Speech Recognition (LVCSR). First, we explored the use of linear discriminant analysis (LDA) for the transformation of speech features as well as the integration of temporal-spatial information, and compared such an approach with the conventional feature extraction approaches. Then, several improved approaches for linear transformation of speech features, such as heteroscedastic linear discriminant analysis (HLDA) and heteroscedastic discriminant analysis (HDA), were investigated as well. Finally, we considered the influence of noises on the above data-driven linear feature transformation techniques. All experiments were carried out on the Mandarin broadcast news speech. The speech recognition tasks included not only Free Syllable Decoding but mandarin Large Vocabulary Continuous Speech Recognition.

關鍵詞：語音辨識、線性鑑別分析、異質性線性鑑別分析、異質性鑑別分析、最大相似度線性轉換。

一、簡介

語音長久以來是人與人之間最自然且最方便的溝通方式[1]。隨著數位電子科技的蓬勃發展以及無線通訊與網際網路的創新普及，傳統的桌上型電腦不再是人們唯一主要的資訊存取平台，取而代之的是各式各樣的家電產品，這些設備將變成是可以計算、通訊與上網的智慧型設備，而且朝輕薄短小的趨勢演進發展。同時，不是每種設備都具有螢幕、鍵盤和滑鼠等這些人們習以為常的輸出入裝置，取而代之的將是語音控制，扮演著未來人類與各式智慧型設備間最主要的人機介面。另一方面，日常生活中可以存取與使用的多媒體影音資訊愈來愈多，例如廣播電視節目、語音信件、演講錄影和數位典藏等。這些多媒體資訊可以從網路上大量地取得，已經成為傳統文字資訊外社會大眾廣泛使用的資訊來源。顯而易見的是，在上述的絕大部分多媒體資訊中，語音可以說是最具語意的主要內涵之一，當播放出多媒體的語音資訊或是顯示出對應的正確轉寫文字時，我們就可以大概地瞭解其中所要傳達的主題或概念。因此，語音辨識技術對多媒體資訊處理也扮演著相當重要的角色，近年來在國際上有相當多從事多媒體語音內涵自動轉寫的研究被發表，其中常以廣播新聞、電話交談式語音、演講與口述歷史典藏的語音辨識研究為主。

語音辨識系統基本上可看作一種圖樣辨識(Pattern Recognition)系統，如圖 1 所示。所有這類系統都是屬於分類(Classification)問題，分類問題包含了兩個成分，第一是特徵擷取(Feature Extraction)，第二是分類器的設計。如果特徵擷取出的特徵向量可以保留重要的成份或者是可以帶有很高的鑑別力(Discriminability)，如此分類器就可以使用較簡單的方法來作分類，自然地分類的結果也會比較精準。當今以人耳聽覺感知為語音特徵的梅爾倒頻譜係數(Mel-frequency Cepstral Coefficients, MFCC) [2]或是感知線性預測係數(Perceptual Linear Prediction Coefficients, PLPC) [3]已成為主流的語音特徵擷取方法之一，配合它們的一階與二階時間軸導數(Time Derivatives)、以及特徵平均值與變異數正規化(Mean and Variance Normalization)強健性(Robustness)處理後，可以在大詞彙連續語音辨識上得到不錯的效果。近幾年來

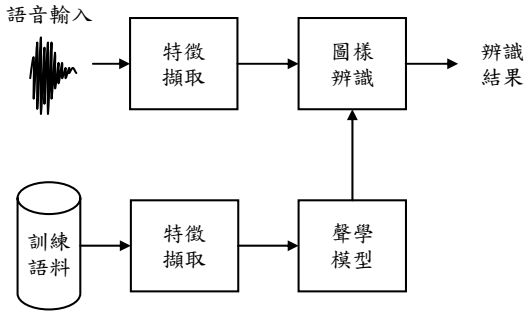

圖 1 基於圖樣辨識的語音辨識系統。

陸續有研究嘗試針對這些語音特徵作進一步處理，最常見的是對語音特徵向量作線性轉換並降低維度只保留重要或具有鑑別力的特徵成分，例如使用線性鑑別分析(Linear Discriminant Analysis, LDA) [4]、異質性線性鑑別分析(Heteroscedastic Linear Discriminant Analysis, HLDA) [5, 6]、異質性鑑別分析(Heteroscedastic Discriminant Analysis, HDA) [7]等。其中線性鑑別分析是假設所有類別特徵向量的分佈變異是相同的，而異質性線性鑑別分析與異質性鑑別分析則是打破這樣的假設。同時，也有許多的研究嘗試以核函數線性鑑別分析(Kernel Linear Discriminant Analysis, Kernel LDA) [8]對語音特徵向量做進一步處理，希望藉由核函數將特徵向量投射到高維度特徵空間作鑑別分析，解決在原特徵空間可能存在的非線性鑑別問題。

另一方面，由於在聲學模型(例如隱藏式馬可夫模型狀態觀測機率分佈)常使用具對角化共變異矩陣(也就是假設特徵向量維度間彼此為無關的)的高斯分佈，但是上述的語音特徵向量或是鑑別分析並不保證此一特性，因而有學者提出以最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT)，嘗試讓轉換過後的共變異矩陣的值集中在對角線上，在不影響聲學模型相似度估測的條件下，儘量滿足對角化共變異矩陣的要求。因此，目前在大詞彙連續語音辨識的語音特徵擷取上常見到線性鑑別分析與最大相似度線性轉換結合(LDA-MLLT) [9]或是異質性線性鑑別分析與最大相似度線性轉換結合(HLDA-MLLT) [10, 11]等的一些作法。本論文主旨在於探討各種資料導向線性特徵轉換和各種結合時域-頻域資訊方法於中文大詞彙連續語音辨識之應用，並結合其他強健性技術，研究噪音對於上述各種資料導向線性特徵轉換技術的影響。我們以中文廣播新聞為實驗語料，初步地以外場採訪記者(Field Reporter)語料部分作為語音辨識實驗題材。

本論文接下來的安排如下：在第二節我們將介紹各種資料導向線性轉換的技術，包含了主成份分析、線性鑑別分析、異質性線性鑑別分析、最大相似度線性轉換；第三節介紹大詞彙連續語音辨識系統；第四節將描述實驗結果；第五節為結論與未來展望。

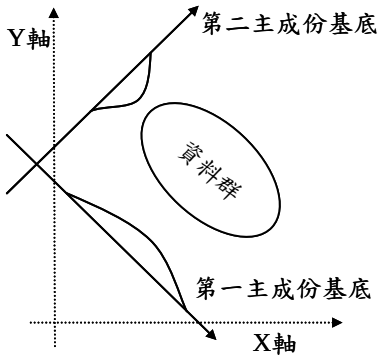

圖 2 主成份分析示意圖。

二、資料導向線性特徵轉換技術

近幾年來資料導向(Data-Driven)線性特徵轉換在語音特徵擷取的研究上佔有相當重要的地位。因為資料導向線性特徵轉換可以藉由語音訓練資料的統計資訊來自動地找出語音特徵空間中重要的基底向量，使得經轉換後的語音特徵能保有重要的成份或具有較高的鑑別力，且可以進一步去除多餘的維度；由於基底向量是根據訓練資料而來，所以找出的基底向量將較能代表語音訊號的特徵。以下將針對實驗組的資料導向線性特徵轉換方法做介紹。

2.1 主成份分析

主成份分析(Principal Component Analysis, PCA)在圖樣辨識中為很常見的技術。它的精神在於將維度間為相關(Correlated)的一群特徵向量用較少維度來表示，且使得維度間變成彼此無關(Uncorrelated)，同時能保有特徵向量的變異量(Variation) [12]，所以主成份分析可以找出特徵主要成份所在的基底向量。如圖 2 所示，投影在第一主成份基底向量的資料擁有最大的變異量，投影在第二主成份基底擁有次大的變異量，且基底向量間各自為單位正交(Orthonormal)。主成份分析的作法為使用所有訓練資料 $X=[x_1,x_2,x_3\cdots,x_N]$，$N$ 為資料的總數，每個 x_i 為 n 維向量，來統計訓練資料的整體共變異矩陣(Covariance Matrix) T ：

$$T=\frac{1}{N}\sum_{i=1}^{N}\left(x_i-\bar{X}\right)\left(x_i-\bar{X}\right)^T \tag{1}$$

其中 $\bar{X}$ 為整體的平均向量，所以 T 為 $n\times n$ 維矩陣。

$$\bar{X}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{2}$$

對 T 求特徵向量分解(Eigenvectors Decomposition)，以特徵值(Eigenvalues)最大的前 p 個特徵向量(Eigenvectors)當成基底矩陣(亦稱轉換矩陣) θ_p 的行向量，最後所有資料可以利用求得的轉換矩陣投影到新特徵空間 $Y=[y_1,y_2,y_3\cdots,y_N]$ ：

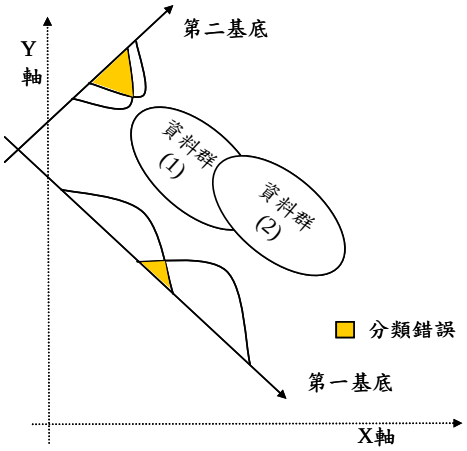

圖 3 線性鑑別分析示意圖。

$$y_i = \theta_p^T x_i \ , i = 1,2,\cdots,N \tag{3}$$

故本文以下所提及不同的資料導向線性鑑別分析法的差別只在於轉換矩陣的求取不同，最終都是藉由式(3)求得新的資料向量表示方式。

2.2 線性鑑別分析

線性鑑別分析(Linear Discriminant Analysis, LDA) [4]與主成份分析在作法上相似，都是求取一個轉換矩陣再藉此作線性轉換與特徵降維。但線性鑑別分析的目的在於使得轉換後特徵之間可以保有最大的分類鑑別資訊。如圖 3 所示，資料群投影到第一基底後比投影到第二基底後可以有較大的鑑別力。線性鑑別分析則有幾個假設：第一，假設投影後並不是所有的維度都包含鑑別資訊，所有鑑別資訊都包含在前 p 維子空間，而後(*n-p*)維子空間即可省略；第二，假設每個類別都是高斯分佈；第三，所有類別分佈的變異量都相同。線性鑑別分析的作法必須使用到訓練資料的類別資訊來統計各類別的分佈，所以算是一種監督式學習(Supervised Learning)。當資料以向量方式呈現時，希望類別間共變異矩陣 B (Between-class Covariance Matrix)轉換後的行列式值越大越好，且類別內共變異矩陣 W (Within-class Covariance Matrix)轉換後的行列式值越小越好。也就是要求取一個基底矩陣 θ_p 使得經線性轉換過後訓練資料的兩行列式(Determininit)比值最大：

$$\hat{\theta} = \arg\max_{\theta_p} \frac{\left|\theta_p^T B \theta_p\right|}{\left|\theta_p^T W \theta_p\right|} \tag{4}$$

在此，假設資料共分成 J 個類別，N_j 為屬於第 j 個類別的訓練語料個數，N 為所有訓練語料的個數，W_j 為第 j 個類別的共變異矩陣。$\bar{X}_j$ 為訓練語料第 j 個類別的平均向量，$\bar{X}$ 為所有訓練語料的平均向量，$g(i)$ 表示向量 x_i 所屬的類別，而 W 與 B 的定義如下：

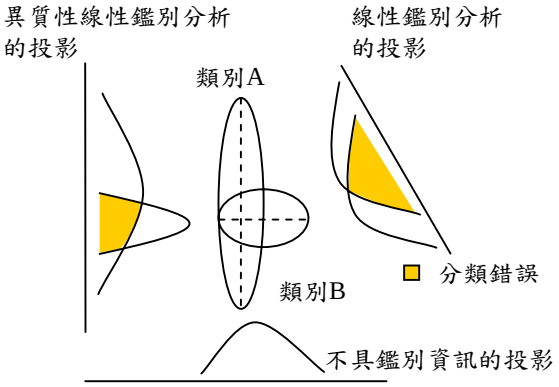

圖 4 線性鑑別分析與異質性線性鑑別分析的比較。

$$W = \frac{1}{N}\sum_{j=1}^{J} N_j W_j \tag{5}$$

$$B = \frac{1}{N}\sum_{i=1}^{J} N_j \left(\bar{X}_j - \bar{X}\right)\left(\bar{X}_j - \bar{X}\right)^T \tag{6}$$

其中

$$W_j = \frac{1}{N_j}\sum_{x_i \text{ st. } g(i)=j} \left(x_i - \bar{X}_j\right)\left(x_i - \bar{X}_j\right)^T \tag{7}$$

$$\bar{X}_j = \frac{1}{N_j}\sum_{x_i \text{ st. } g(i)=j} x_i \ , j = 1 \ldots J \tag{8}$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{9}$$

從文獻可以得知[12]，要求轉換矩陣 θ_p 使得式(4)有極大值等同於對 $W^{-1}B$ 做特徵向量分解，以特徵值最大的前 p 維特徵向量當成基底向量。

2.3 異質性線性鑑別分析

由於大多的訊號特徵分佈的變異量皆為異質性，所以異質性線性鑑別分析免除了前一小節提到的線性鑑別分析所作假設，也就是各類別分佈變異量皆相同的假設。如圖 4 所示，經異質性線性鑑別分析的投影過後的訓練資料將會較經線性鑑別分析投影過後的訓練資料有較少的分類錯誤。當我們假設語音特徵各類別分佈有不同變異量時，線性轉換基底的求取則有以數值方法[5, 13]和固定解方法[14]兩種解法。此兩種解法將使用不同的最佳化方法取得相同的目標函式。

(a) 數值方法(Numerical Method)

Kumar 在 1997 年的博士論文提出現實中特徵的分佈之變異可為異質性(Heteroscedastic)，針對此假設來一般化線性鑑別分析，也就是去除各類別分佈變異量為相同的要求，同樣再以最大相似度(Maximum Likelihood, ML)估測為目標函式[5]，進一步推導出異質性線性鑑別分析。異質性線性鑑別分析如同線性鑑別分析一般，都藉由選取原 n 維特

徵空間中的 p 維子空間，並捨棄另外(n-p)維子空間，來達到降維的目的。因為最後的目標是在於提高分類的正確率，所以暗示被捨棄的(n-p)維不帶有任何的分類資訊。對於高斯分佈來說，捨棄的(n-p)維被假設不帶有分類資訊即同等於假設此(n-p)維在類別分佈的中心值和變異量是相同的。因此可以假設特徵向量的前 p 維和後(n-p)維彼此獨立，所以線性轉換後的平均向量 μ_j 及共變異矩陣 $\sum_j$ 可以拆開成帶有鑑別資訊和不帶有鑑別資訊：

$$\mu_j = \begin{bmatrix} \mu_{j,1} \\ \vdots \\ \mu_{j,p} \\ \mu_{0,p+1} \\ \vdots \\ \mu_{0,n} \end{bmatrix} = \begin{bmatrix} \mu_j^p \\ \mu_0 \end{bmatrix} \tag{10}$$

μ_j^p 為投影後第 j 個類別平均向量的前 p 維，μ_0 為投影後第 j 個類別平均向量的後(n-p)維，同樣地，$\sum_j^p$ 、$\sum^{n-p}$ 分別為前 p 維與後(n-p)維的共變異矩陣；換言之，所有類別的 μ_0 、$\sum^{n-p}$ 都是相同的，但 μ_j^p 、$\sum_j^p$ 都是不同的。

$$\sum_j = \begin{bmatrix} \sum_j^p & 0 \\ 0 & \sum^{(n-p)} \end{bmatrix} \tag{11}$$

如果原始資料分佈為高斯分佈，則資料經線性轉換後亦為高斯分佈，且轉換前後機率密度函式(Probability Density)的關係如下：

$$P(x_i) = \frac{|\theta|}{\sqrt{(2\pi)^n \left|\sum_{g(i)}\right|}} e^{-\frac{\left(y_i-\mu_{g(i)}\right)^T \sum_{g(i)}^{-1}\left(y_i-\mu_{g(i)}\right)}{2}} \tag{12}$$

其中 θ 為 $n\times n$ 維線性轉換矩陣，每一行(Column)為一基底向量；$y_i = \theta^T x_i$ ，x_i 為第 i 個資料向量，下標 $g(i)$ 為 x_i 對應的類別。最大相似度估測期望達到訓練語料落在所屬類別的機率是最大，也就是希望全體資料向量(訓練資料)在其所屬類別的對數相似度能最大，所以對式(13)求極大值。

$$\begin{aligned} \log L(\theta,\{x_i\}) &= \sum_{i=1}^{N} \log p(x_i) \\ &= -\frac{1}{2}\sum_{i=1}^{N}\left(\theta^T x_i - \mu_{g(i)}\right)^T \sum_{g(i)}^{-1}\left(\theta^T x_i - \mu_{g(i)}\right) \\ &\quad -\frac{1}{2}\sum_{i=1}^{N}\log\left((2\pi)^n\left|\sum_{g(i)}\right|\right) + N\log|\theta| \end{aligned} \tag{13}$$

其中 N 為訓練語料的資料向量總數。首先固定 θ ，對式(13)各別轉換後平均向量 μ_j 和共變異矩陣 $\sum_j$ 做微分，可求得 μ_j 和 $\sum_j$ 。μ_j^p 、μ_0 、$\sum_j^p$ 與 $\sum^{n-p}$ 分別表示如下：

$$\mu_j^p = \theta_p^T \bar{X}_j \tag{14}$$

$$\mu_0 = \theta_{n-p}^T \bar{X} \tag{15}$$

$$\sum_j^p = \theta_p^T W_j \theta_p \tag{16}$$

$$\sum^{n-p} = \theta_{n-p}^T T \theta_{n-p} \tag{17}$$

經由化簡並去掉與 θ 無關的變數後，最大化類別相似度的線性轉換矩陣 $\hat{\theta}$ 可表示成：

$$\begin{aligned} \hat{\theta} = \arg\max_{\theta}\Bigg\{ & N\log|\theta| - \frac{N}{2}\log\left|\theta_{n-p}^T W_j \theta_{n-p}\right| \\ & -\sum_{j=1}^{J}\frac{N_j}{2}\log\left|\theta_p^T W_j \theta_p\right|\Bigg\} \end{aligned} \tag{18}$$

由式(18)可以看出轉換後的共變異矩陣並不是對角化的，可以在式子推導的過程中假設轉換後的共變異矩陣僅對角線有值，稱為對角化異質性線性鑑別分析(Diagonal HLDA, DHLDA)，式子為：

$$\begin{aligned} \hat{\theta} = \arg\max_{\theta}\Bigg\{ & N\log|\theta| - \frac{N}{2}\log\left|Diag\left(\theta_{n-p}^T T \theta_{n-p}\right)\right| \\ & -\sum_{j=1}^{J}\frac{N_j}{2}\log\left|Diag\left(\theta_p^T W_j \theta_p\right)\right|\Bigg\} \end{aligned} \tag{19}$$

期間 θ 的求取並沒有固定解(Close-form Solution)，只能利用數值方法的最佳化技術來求解，對 θ 微分然後迭代更新 θ 的值，直到收斂。

(b) 固定解方法(Close-form-solution Method)

由於上述的數值方法沒有固定解，造成需要非常多次的迭代來更新 θ 值。每作一次實驗，都需花費相當多的時間來求取異質性線性鑑別分析的基底矩陣，所以我們亦參考英國劍橋大學 Gales 教授在 1999 年提出的方法對異質性線性鑑別分析求解[14, 15]。這個方法會產生固定解，並且保證整體相似度會隨著每次迭代更新 θ 而增加，但只適用在對角化異質性線性鑑別分析的求解。

$$\begin{aligned} \log L(\theta;\{x_i\}) &= \sum_{i=1}^{N}\log P(x_i) \\ &= -\frac{1}{2}\sum_{i=1}^{N}\left(x_i - \bar{X}_{g(i)}\right)^T \theta \sum_{g(i)}^{-1} \theta^T \left(x_i - \bar{X}_{g(i)}\right) \\ &\quad -\frac{1}{2}\sum_{i=1}^{N}\log\left((2\pi)^n\left|\sum_{g(i)}\right|\right) + N\log|\theta| \end{aligned} \tag{20}$$

其中 θ 為 $n\times n$ 維矩陣，每一行(Column)為一基底向量，$|\theta|$ 可分解成 $a_i^T c_i^T$。a_i 為 θ 的第 i 行，為 $n\times 1$ 維向量。c_i 為 θ 第 i 行的餘因子(Cofactors)，為 $1\times n$ 維向量。由於假設轉換後類別共變異矩陣為對角化，如此可以把 $\left(x_i - \bar{X}_{g(i)}\right)^T \theta \sum_{g(i)}^{-1} \theta^T \left(x_i - \bar{X}_{g(i)}\right)$ 用一行一行(row by row)的方式表示，則式(20)可以代換成如下：

$$\log L\left(\theta;\{x_i\}\right)$$
$$=-\frac{1}{2}\sum_{i=1}^{N}\sum_{k}\frac{\left(a_k^T x^{g(i)}(i)\right)^2}{\left(\sigma_{diag_k}^{g(i)}\right)^2}-\frac{1}{2}\sum_{i=1}^{N}\log\left((2\pi)^n\left|\Sigma_{g(i)}\right|\right)$$
$$+N\log\left(a_i^T c_i^T\right) \tag{21}$$

其中 $x^{g(i)}(i)=x_i-\overline{X}_{g(i)}$，$g(i)$ 為 x_i 對應的類別，$\left(\hat{\sigma}_{diag_k}^{g(i)}\right)^2$ 為經線性轉換後共變異矩陣對角線上第 k 個元素。把式(21)與 a 無關的變數集中成 K，並做代數轉換：

$$\log L\left(\theta;\{x_i\}\right)$$
$$=-\frac{1}{2}\sum_{k}\left(a_k^T G^k a_k\right)+N\log\left(a_i^T c_i^T\right)+K \tag{22}$$

其中

$$G^k=\sum_{j=1}^{J}\frac{N_j}{\sigma_{diag_k}^{j\ 2}}W_j \tag{23}$$

式(22)對 a_i 做微分並使其等於零，最後可求得：

$$a_i=G^{i^{-1}}c_i^T\sqrt{\frac{N}{c_i G^{i^{-1}}c_i^T}} \tag{24}$$

由於假設只有前 p 維子空間帶有分類資訊，所以 G^i 和 $\left(\sigma_{diag_k}^{j}\right)^2$ 可以拆成如下。

$$G^k=\begin{cases}\sum_{j=1}^{J}\frac{N_j}{\left(\sigma_{diag_k}^{j}\right)^2}W_j & (k\le p)\\ \sum_{j=1}^{J}\frac{N_j}{\left(\sigma_{diag_k}^{j}\right)^2}T & (k>p)\end{cases} \tag{25}$$

$$\left(\sigma_{diag_k}^{j}\right)^2=\begin{cases}a_k^T W_j a_k & (j\le p)\\ a_k^T T a_k & (j>p)\end{cases} \tag{26}$$

2.4 最大相似度線性轉換

最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT) [7, 16]與上述資料導向線性特徵轉換的作法是一樣的，都是求取一個轉換基底矩陣來轉換特徵空間。但在用途跟精神有所不同，首先，最大相似度線性轉換並不執行降維動作，而是保留訓練資料特徵的所有維度。其次，它的精神在於使轉換後類別的共變異矩陣對角化，同時希望分別使用對角化或全秩(Full)共變異矩陣所計算出的特徵向量的整體類別相似度差要最小；也就是說雖然經線性轉換後類別的共變異矩陣只保留對角線的值，但希望整體類別相似度與保留共變異矩陣所有元素是相同的。如此可以用來解決上述主成份分析、線性鑑別分析和異質性線性鑑別分析的重大缺陷，就是經線性轉換後類別的共變異矩陣不是對角化的現象，而造成後端隱藏式馬可夫模型估測失真。最大相似度線性轉換假設每個類別為高斯分佈且為全秩共變異矩陣，訓練資料整體對數相似度可以簡化成：

$$-\sum_{j=1}^{J}\frac{N_j}{2}\log\left|\theta^T W_j\theta\right|+C \tag{27}$$

由於希望轉換後類別共變異矩陣只保留矩陣對角線上的值，且整體類別相似度受影響越小越好。所以等同求解：

$$\hat{\theta}=\arg\max_{\theta\in R^{n\times n}}N\log\left|\theta\right|-\sum_{j=1}^{J}\frac{N_j}{2}\log\left|diag\left(\theta^T W_j\theta\right)\right| \tag{28}$$

比較式(19)與式(28)可以發現，當式(28)的線性轉換保留所有的維度時，兩式是相等的，所以得知最大相似度線性轉換為對角化異質性線性鑑別分析的特例。

三、大詞彙連續語音辨識系統

本論文實驗環境採用建立於台灣師範大學資工所發展的一套大詞彙連續語音辨識系統[17, 18]，主要包括前端處理、聲學模型訓練、詞典的建立(Lexicon Construction)、語言模型訓練和詞彙樹複製搜尋(Tree-Copy Search)等部分。

3.1 前端處理

前端處理部份即是本篇論文的討論重點，主要將著眼於擷取特徵值的各種資料導向線性特徵轉換方式，其中以梅爾頻譜濾波器(Mel-frequency Filter Banks)的輸出作為語音訊號的基礎資料向量，進一步在資料向量時間點的前後各結合 4 個資料向量來擴大資訊，此後結合時域與頻域做轉換，取得更精準的特徵資訊；如第二章節所述的線性轉換方式，使得經轉換後的特徵值能保有重要的語音成份或具有較高的鑑別力。除不同的線性轉換處理方法之外，並與梅爾倒頻譜特徵參數(Mel-frequency Cepstral Coefficients,MFCC)及感知線性預測係數(Perceptual Linear Prediction Coefficients, PLPC)比較，且加入強健性語音特徵處理方式，倒頻譜平均消去法(Cepstral Mean Subtraction,CMS) [19]或倒頻譜正規化法(Cepstral Normalization,CN) [20]，降低噪音對特徵值的影響。

3.2 聲學模型

聲學模型是採用傳統的連續密度隱藏式馬可夫模型（Continuous Density Hidden Markov Model, CDHMM）。模型的總數量有 151 個，其中包含了 1 個靜音模型（Silence），112 個聲母模型（Initials），以及 38 個韻母模型（Finals）。每個模型的狀態數分別為 3 至 6 個不等，每個狀態皆為高斯混合分

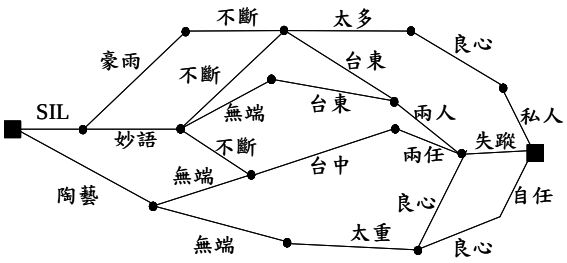

圖 5 詞圖為語音辨識所有可能候選詞與詞句的簡潔表示。

佈，其中每個高斯混合分佈的分佈個數分別為 1 至 128 個不等。此外，聲母和韻母共有 403 種不同的音節組合。

3.3 詞典建立及語言模型訓練

本系統使用詞雙連(Word Bigram)以及詞三連(Word Trigram)語言模型，並以中央通訊社(Central News Agency, CNA) 2001 與 2002 年所收集到的約一億七千萬個中文字語料作為背景語言模型訓練時的訓練資料 。在本論文中的語言模型使用了 Katz 語言模型平滑技術[21]，並採用 SRL Language Modeling Toolkit (SRILM)的研究工具軟體[22]來訓練語言模型。

3.4 詞彙樹複製搜尋

本系統的大詞彙連續語音辨識方法是採用由左至右(Left-to-right)、音框同步(Frame-synchronous)的詞彙樹複製搜尋方式[23]。在詞彙樹中每個分枝(Arc)代表一個 INITIAL 或 FINAL 的隱藏式馬可夫模型，由樹根(Root)到任一個樹梢(Leaf)的路徑代表一個詞或一些發音相同的詞，路徑上的分枝就是代表這個詞或這些詞會使用到的隱藏式馬可夫模型。具體來說，所採用的詞彙樹複製搜尋演算法，搜尋時每個音框會同時存在數棵詞彙樹複製(Tree Copies)，每個詞彙樹代表不同的語言模型歷史或限制(Language Model History or Constraint)。實際上，搜尋時產生的不完全路徑(Partial Paths)如果擁有相同的語言模型歷史，則會被歸類在同一棵詞彙樹複製裡，進行隱藏式馬可夫模型狀態層次(State-level)維特比動態規劃搜尋。在每個音框中，若有不完全路徑已抵達樹梢時，代表一個完整詞已可被產生；同時，不同棵詞彙樹複製間已抵達樹梢的不完全路徑，若具有相同的語言模型歷史，則會進行再結合(Recombination)，保留最大分數者，並以它們的語言模型歷史為標註，產生新的一棵詞彙樹複製，或加入到一棵已存在且具有相同語言模型歷史的詞彙樹複製中。值得注意的是，在實作時並不需要真的建立如此多的詞彙樹複製，僅需建立一棵詞彙樹作為搜尋時路徑展開參考之用即可，並分別記錄搜尋時存活下來的隱藏式馬可夫模型狀態節點(也就是不完全路徑目前拜訪到的節點)的相關資訊。另一方面，由於存活的隱藏式馬可夫模型狀態節點可能會隨音框數呈指數倍增加，因此必須以光束剪裁(Beam Pruning) 技術適當地剪裁分數較低的狀態節點或不完全路徑。在執行剪裁動作時會同時考量每一個詞彙樹複製內部狀態節點(Internal Node)下涵蓋的可能拜訪樹梢節點代表之所有詞對應的語言模型機率，並以其中最大者當做每一個詞彙樹複製內部狀態節點的語言模型前看(Look-ahead)分數，再加上內部狀態節點本身搜尋時所累積的解碼分數(Decoding Score)及聲學前看分數當成剪裁比較的依據。在本系統採用詞單連(Unigram)語言模型前看技術，對每一個詞彙樹複製內部狀態節點，會以其所在分枝(或隱藏式馬可夫模型)之可能拜訪樹梢節點中具最大詞單連語言模型機率，做為該內部狀態節點的語言模型前看分數。此外，在每個音框，會記錄存活的詞彙樹複製樹梢節點中分數較高者的相關資訊(這些樹梢節點本身代表著可能的候選詞)，諸如它們的語言模型歷史、對應候選詞開始與結束的音框以及搜尋時聲學解碼的分數，然後再依此資訊建立起一個詞圖，如圖 5。並在詞圖上使用更高階的語言模型，如詞三連(Trigram)、詞四連(Fourgram)語言模型等，重新進行一次詞圖動態規劃搜尋(Word Graph Rescoring)，找出最佳的文句。在本系統中，詞彙樹複製搜尋階段是使用詞雙連語言模型，而在詞圖搜尋階段則是使用詞三連語言模型。

四、實驗

實驗語料與資料導向線性特徵轉換技術，以及這些技術與最大相似度線性轉換結合的相關實驗結果，將詳細敘述如下。

4.1 實驗語料

本論文主要使用的語料庫為 MATBN 電視新聞語料[24]，為中央研究院資訊所口語小組耗時三年與公共電視台合作錄製完成。其中包含 2001 年的新聞 30 小時、2002 年 146 小時及 2003 年 24 小時。本論文初步地選擇採訪記者語料作為實驗語材，其中包含 25.5 小時的訓練集(5,774 句)，供聲學模型訓練之用；1.5 小時的評估集(292 句)，供辨識評估之用，其中男女語料各半。另外，為了探討不同語音特徵參數對於語音辨識的強健性，本論文對公視新聞的辨識語料加入 Aurora 2.0 不同訊噪比(SNR)下的各種噪音[25]。Aurora 2.0 是由歐洲電信標準協會(European Telecommunications Standards Institute, ETSI)所發行的語料，提供了八種不同噪音，有地下鐵(Subway)、人聲(Babble)、汽車(Car)、展覽會館(Exhibition)、餐廳(Restaurant)、街道(Street)、機場(Airport)、火車站(Train Station)。本論文使用 Aurora 2.0 提供的不同噪音搭配不同程度的訊噪比，分別是 -5dB、0dB、5dB、10dB、15dB20dB。強健性實驗的方法為：使用 Aurora 2.0 內的不同噪音搭配不同的 SNR 比例加入我們的測試語料中，使用原始訓練語料所訓練出的模型來辨

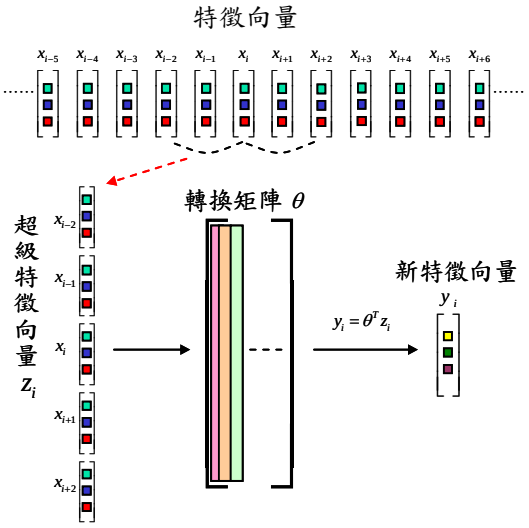

圖 6 頻域-時域特徵擷取示意圖。

識加入噪音的測試語料，藉此來驗證不同線性特徵轉換與強健性技術所求得語音特徵值的強健性。

4.2 資料導向線性特徵轉換技術的應用

本論文把資料導向線性特徵轉換應用在頻域-時域特徵擷取，希望藉由資料導向線性特徵轉換來同時擷取頻域上與時域上重要或具鑑別力的特徵向量。如圖 6 所示，首先由資料向量 x_i 本身加上前後各取 k 個資料向量形成超級資料向量 z_i (Supervector)，此處的 k 為 2。超級資料向量 z_i 經由基底矩陣 θ 線性轉換後可得新資料向量 y_i。其中的 θ 就是由資料導向線性特徵轉換求得。本論文中 x_i 為 18 組梅爾濾波器(Mel Filter Banks)的輸出，前後各取 4 個資料向量，所以 z_i 為 162 維向量，最後新資料向量 y_i 為 39 維。

4.3 中文自由音節辨識

本小節主要在探討各種資料導向線性特徵轉換效能的比較，主要有三個觀察。首先從表 1 我們可看到 DHLDA-I (數值方法求解)與 DHLDA-II (固定解方法求解)的效能在伯仲之間，這是因為兩個方法都是從同一個目標函式所推導而成，自然地兩者會有相似的辨識率。不過數值方法由於迭代次數太多，要找到跟固定解方法一樣好的轉換矩陣是有困難的。其次，表 1 可以看出，LDA 的辨識率比 DHLDA-I 和 DHLDA-II 都差，本論文推測這是因為線性鑑別分析應用在頻域-時域特徵擷取時，擷取出的特徵向量並不滿足類別共變異矩陣為對角化，造成隱藏式馬可夫模型參數估測失真，進而影響辨識率。所以一但線性鑑別分析結合最大相似度線性轉換後，辨識率就與 DHLDA-I 和 DHLDA-II 相差不遠，如表 2 所示。但這並不代表異質性的線性鑑別分析沒有功效，觀察表 1 可以發現 HLDA 的辨識率是最低的，但在表 2 所示 HLDA+MLLT 的辨識率與 HLDA 相比，大幅度的減少 15.71%相

表 1 各種資料導向線性特徵轉換與強健性技術結合，數據皆為音節錯誤率。(+:表示加上倒頻譜平均消去法處理(CMS)或加上倒頻譜正規化法處理(CN))

Method	Baseline	+CMS	+CN
MFCC	44.97	41.68	41.06
PLP	46.46	42.50	41.82
PCA	45.40	39.82	38.89
LDA	43.17	38.80	38.30
HLDA	47.10	40.08	39.22
DHLDA-I	40.90	37.41	36.80
DHLDA-II	40.50	37.05	36.45

表 2 各種資料導向線性特徵轉換與最大相似度線性轉換及強健性技術結合，數據皆為音節錯誤率。

Method	Baseline	+CMS	+CN
MFCC	44.67	40.67	40.10
PLP	46.92	42.36	41.94
PCA	41.98	37.53	37.03
LDA	40.78	37.06	36.47
HLDA	39.70	36.57	36.12

對錯誤率。HLDA+MLLT 的辨識率與 LDA+MLLT 的辨識率比較，減少 2.65%相對錯誤率。而 HLDA+MLLT 的辨識率與 DHLDA-II 的辨識率相比，也減少 1.98%相對錯誤率。

再來討論音節強健性實驗，比較主成份分析與線性鑑別分析，從表 3 可以看出在 20dB 至 5dB 的訊噪比下，LDA+MLLT+CN 的辨識率都比 PCA+MLLT+CN 來的高；只有在 0dB 與-5dB 的訊噪比下會相反。比較 LDA 與 DHLDA-II 的辨識率，從表 3 可以看出在 20dB 至 5dB 的訊噪比下，LDA+MLLT+CN 的辨識率都比 DHLDA-II+CN 來的高；只有在 0dB 與-5dB 的訊噪比下會相反，其中-5dB 時，DHLDA-II+CN 的辨識率表現十分出色，也因為如此使得 DHLDA-II+CN 在所有噪音下表現都比 LDA+MLLT+CN 好。此外特別的是，在乾淨環境下，DHLDA-II+CN 與 LDA+MLLT+CN 的辨識率比較，只減少了 0.05%相對錯誤率；但在噪音環境下，DHLDA-II+CN 與 LDA+MLLT+CN 的辨識率比較，減少 0.98%相對錯誤率。HLDA+MLLT+CN 在所有噪音所有訊噪比環境下都比 LDA+MLLT+CN 和 DHLDA-II+CN 優秀。但在-5dB 時還是表現比 DHLDA-II+CN 差。此外異質性線性鑑別分析不僅在乾淨環境下有最好的辨

表 3 資料導向線性特徵轉換技術的強健性實驗，數據皆為音節錯誤率。

+CN	With MLLT					Without MLLT
	MFCC	PLP	PCA	LDA	HLDA	DHLDA-II
Clean	40.10	41.94	37.03	36.47	36.12	36.45
20dB	41.43	42.96	38.31	37.78	37.34	38.08
15dB	44.11	45.59	40.83	40.53	39.88	40.87
10dB	50.01	51.50	46.63	46.56	45.89	46.96
5dB	62.26	63.45	58.67	58.60	58.35	59.47
0dB	82.20	83.09	78.26	78.64	77.73	77.79
-5dB	103.94	104.33	98.41	98.79	96.66	94.18
Average	63.99	65.15	60.19	60.15	59.31	59.56

表 4 特徵擷取在大詞彙連續語音辨識的音節(S)、字(C)、詞(W)錯誤率(%)；TC 為詞彙樹複製搜尋，WG 為詞圖搜尋。

+CN	Method	TC (S)	TC (C)	TC (W)	WG (S)	WG (C)	WG (W)
With MLLT	MFCC	19.64	27.95	37.78	19.52	26.76	35.55
	PLP	23.80	32.40	42.36	23.63	31.19	40.05
	PCA	18.32	26.85	36.87	18.03	25.20	34.06
	LDA	18.00	26.47	36.45	17.52	24.59	33.27
	HLDA	17.40	25.79	35.66	17.04	24.21	32.80
Without MLLT	DHLDA-I	18.09	26.56	36.61	17.67	24.88	33.79
	DHLDA-II	17.97	26.40	36.04	17.49	24.77	33.53

識率，在噪音環境下也是最好，這代表異質性線性鑑別分析除了可以擷取有鑑別力的語音特徵，同時也可以增進強健性。

4.4 中文大詞彙連續語音辨識

依上述 4.3 自由音節辨識所顯示的結果，各自選取最好的一組資料轉換技術，實作在大詞彙連續語音辨識系統中，實驗人耳聽覺感知相關特徵擷取技術與資料導向線性特徵轉換技術，以及這些轉換技術與最大相似度線性轉換、並結合其他強健性技術的組合，以探討這些技術在大詞彙連續語音辨識的效能。從表 4 可以看到不管在何種評估方法，HLDA+MLLT+CN 都有最高的辨識率。以詞圖搜尋的字正確率來說，相比於 LDA+MLLT+CN，減少了 1.41%相對錯誤率。另外，雖然 DHLDA-II+CN 的辨識率比不上 HLDA+MLLT+CN，但仍然可以得到比 LDA+MLLT+CN 高的辨識率，且其在轉換矩陣最佳化的求取比 HLDA 容易、計算量也較小，算是不錯的一種組合選擇。

五、結論與未來展望

在實驗中可以發現，應用在頻域-時域(Spatial-Temporal)特徵擷取的主成份分析與線性鑑別分析結合最大相似度線性轉換(MLLT)後，辨識率都大幅度的提昇。不論是在乾淨環境下或噪音環境下，HLDA-MLLT 的辨識率都比 LDA-MLLT 高上許多。這表示 HLDA 不只比 LDA 更能取出有鑑別力的特徵向量，而且對於噪音干擾也較具有強健性的幫助。

此外，近來對於資料導向線性特徵轉換的研究改進，尚有最小分類錯誤(Minimum Classification Error, MCE) [26]和(Maximum Mutual Information, MMI) 估測來取代最大相似度(Maximum Likelihood, ML) 估測外。BBN 與 IBM 等研究單位也提出以最小音素錯誤率(Mimmun Phoneme Error, MPE)估測，稱為最小音素錯誤異質性線性鑑別分析(Minimum Phoneme Error Based Heteroscedastic Linear Discriminant Analysis, MPE-HLDA) [27]。在未來，資料導向線性特徵轉換可以用來結合各種不同的特徵擷取技術，自動地找出具有重要性或鑑別力的特徵。例如，結合各種不同特徵擷取技術所產生的特徵向量再經由資料導向線性特徵轉換後，可以產生更具有代表性的語音特徵。

六、參考文獻

[1] B. H. Juang, S. Furui, “Automatic recognition and understanding of spoken language—a first step toward natural human-machine communication,” in Proc. IEEE, vol. 88, no. 8, pp. 1142-1165, 2000.

[2] S. B. Davis, P. Mermelstein, “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” IEEE Trans. on Acoustic, Speech, and Signal Processing, vol. 28, no. 4, 1980, pp. 357-366.

[3] H. Hermansky, “Perceptual Linear Predictive (PLP) Analysis of Speech,” Journal of the Acoustical Society of America, vol. 87, 1999, pp. 1738-1752.

[4] R.O. Duda, P.E. Hart, “Pattern Classification and Scene Analysis,” John Wiley and Sons, New York, 1973.

[5] N. Kumar, “Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition,” Ph.D. thesis, John Hopkins University, Baltimore, 1997.

[6] M.J.F. Gales, “Maximum Likelihood Multiple Subspace Projections for Hidden Markov Models,” IEEE Transactions on Speech and

Audio Processing, vol. 10, no. 2, 2002, pp. 37-47.

[7] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, “Maximum Likelihood Discriminant Feature Spaces,” in Proc. IEEE International Conference on Acoustics, Speech, Signal processing, vol. II, 2000, pp. 1129-1132.

[8] S. Mika, “Fisher Discriminant Analysis With Kernels,” in Proc. IEEE International Workshop on Neural Networks for Signal Processing, 1999, pp. 41-48.

[9] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, A. Sixtus, “Large Vocabulary Continuous Speech Recognition of Broadcast News – The Philips/RWTH Approach,” Speech Communication, vol. 37, 2002, pp. 109-131.

[10] P.C. Woodland, “The Development of the HTK Broadcast News Transcription System: An Overview,” Speech Communication, vol. 37, 2002, pp. 47-67.

[11] T. Hain, P.C. Woodland, G. Evermann, M.J.F. Gales, D. Povey, G. Moore, L. Wang, X. Liu, “Automatic Transcription of Conversational Telephone Speech,” to appear in IEEE Transactions on Speech and Audio Processing.

[12] K. Fukunaga, “Introduction to statistical pattern recognition,” E.2nd, Academic Press, 1990.

[13] N. Kumar, A. G. Andreou, “Heteroscedastic Discriminant Analysis and Reduced Rank HMMs for Improved Speech Recognition,” Speech Communication, vol. 26, no. 4, p.283-297, Dec. 1998.

[14] M.J.F. Gales, “Semi-tied Covariance Matrices for Hidden Markov Models,” IEEE Transactions on Speech and Audio Processing, vol. 7, no. 3, pages 272–281, 1999.

[15] M.J.F. Gales, “Maximum Likelihood Multiple Projection Schemes for Hidden Markov Models,” Cambridge University Technical Report RT-365, 2001.

[16] R.A. Gopinath, “Maximum likelihood modeling with Gaussian distributions,” in Proc. of ICASSP Seattle, 1998.

[17] Berlin Chen, Jen-Wei Kuo, Wen-Hung Tsai (2004). “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” in Proc. ICASSP 2004.

[18] Berlin Chen, Jen-Wei Kuo, Wen-Huang Tsai (2005). “Lightly Supervised and Data-Driven Approaches to Mandarin Broadcast News Transcription,” International Journal of Computational Linguistics and Chinese Language Processing, Vol. 10, No. 1, pp.1-18, March 2005.

[19] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” IEEE Trans. Acoust. Speech Signal Process. 1981.

[20] O. Viikki and K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” Speech Communication, Vol. 25, pp. 133-147, August 1998.

[21] S.M. Katz, “Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer,” IEEE Trans. Acoustics, Speech and Signal Processing, Vol. 35, No.3, pp. 400-401, 1987.

[22] A. Stolcke, “SRI language Modeling Toolkit,” version 1.3.3, http://www.speech.sri.com/projects/srilm/.

[23] X.L. Aubert, “An Overview of Decoding Techniques for Large Vocabulary Continuous Speech Recognition,” Computer Speech and Language, January 2002.

[24] H.M. Wang, B. Chen, J.W. Kuo, S.S. Cheng, “MATBN: A Mandarin Chinese Broadcast News Corpus,” accepted to appear in International Journal of Computational Linguistics and Chinese Language Processing, vol. 10, no. 2, June 2005, pp. 219-236.

[25] ETSI Website: http://www.elda.org/article20.html.

[26] Wu Chou and Biing Hwang Juang,, “Pattern Recognition in Speech and Language Processing ,” September 2002, pp. 1-40.

[27] B. Zhang and S. Matsoukas, “Minimum Phoneme Error Based Heteroscedastic Linear Discriminant Analysis for Speech Recognition,” in Proc. ICASSP 2005.