

Digital Speech Processing, Final Exam

June. 26, 2007, 10:10-12:10

- OPEN EVERYTHING
- 除專有名詞可用英文以外，所有文字說明一律以中文為限，未用中文者不計分
- Total points: 160
- Note that you don't need to be able to answer all the questions.

1(+2)

Ch 9

1. (20) In eigenvoice approach, explain how the eigenvoice space is constructed, what that means, and why rapid speaker adaptation can be achieved with very limited quantity of data?

ch 10

2. (20) In Latent Semantic Analysis the elements w_{ij} of the word-document matrix \bar{W} is

$$w_{ij} = (1 - \varepsilon_i) \frac{c_{ij}}{n_j},$$

小 \rightarrow 複合

Where c_{ij} is the number of times the word w_i occurs in the document d_j , n_j is the total number of words in d_j , and

$$\varepsilon_i = -\frac{1}{\log N} \sum_{j=1}^N \left(\frac{c_{ij}}{t_i} \right) \log \left(\frac{c_{ij}}{t_i} \right), t_i = \sum_{j=1}^N c_{ij},$$

where N is the total number of documents. Explain the meaning of all these parameters.

ch 11

3. (20) For Linear Discriminative Analysis (LDA), explain the meaning of within/between class scatter matrices S_W/S_B , and the meaning of the optimization criterion.

2

ch 13

4. (20) Explain why and how vector space model based on subword units are useful in retrieving speech information using speech queries.
4. (20) Explain why and how vector space model based on subword units are useful in retrieving speech information using speech queries.

ch 14

5. (20) Write down the two steps in each iteration of EM algorithm and explain how they operate and what they mean.

2

ch 15

6. (20) Write down what you know about likelihood ratio test and its applications in digital speech processing.

ch 16

7. (20) Explain the three key elements in a spoken dialogue system: what they are, how they operate, and how they are linked together.

2

ch 17

8. (20) Compare the client-server model and server-only model for Distributed Speech Recognition (DSR), and discuss the different considerations.

第九份投影片

聲音的一些特質 相關的參數併成向量

把 random vector 投影 => 把聲音特性區分出來

把聲音的一堆參數 用較少量的參數表示

本來需要很多 dimension 來表示 現在可以用比較少的 dimension 去表示

投影之後要分的越開越好

找投影 variance 最大的

為什麼他可以馬上做 adaptation?

我講話他還是會認得，很快就可以調整，比較少的資料就可以分辨出相對特質。資料量需要比較少

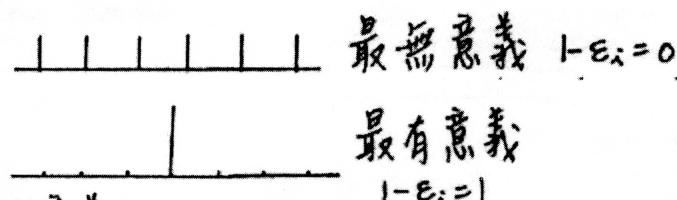
2.

最後一句：解釋一下，所以看一下講義就好了，

如果第二題講一下問參數什麼意思的話就 **10.0** 打開就是了

講義第十份第一面 entropy 大 → 越亂越不重要 小比較重要

取($P_i \log P_i$) 就是看她的 entropy,



前面乘的那個東西是 normalize

3. CH11

找個維度 讓兩個 data 分的最開

S_w (希望聚集一些)越小越好 一群的聚集程度

s_b (希望鬆散一些)越大越好 各群的分散程度

4. 第 13 份

像是說"紐約受到恐怖攻擊的新聞"

就要列出"美國總統布希..."

"賓拉登..."

等等有相關意義的 document 出來

performance measures 講的是

如果我要求字出現的次數的最低標準很高的話(像是出現 100 次以上才列入相關)

則自然我 recall(對應到的)的 relevant document 就少，反之亦然

其中又分成以 word(字)為單位、subword(音節)為單位、關鍵字為單位做 retrieve(檢索)等等。

answer:

而這題問題就是在問說

如果我採用 subword 為單位，利用 vector space model 做檢索搜尋，為什麼很有用？如何實作呢？

好處是我們將每個 document(像是剛剛說的"賓拉登...")跟 query(像是要查"紐約受到恐怖攻擊的新聞"都用一個 vector 去呈現他，

很普遍使用、簡單清楚(我們可以想像利用 class[100]來代表 100 個 class 會比 class1,class2,...來的省力，如以下假設：

將每個 query 跟 document 都設為一個 vector j，

則我們可以找到這個 vector 對個別 term(類似字串)的關係如下：

$$Z_{jt} = (1 + \ln[C_t]) * \ln(N/N_t)$$

C_t 代表這個 term 在 query 或 document 的出現次數

N 代表 document 總數

N_t 代表有這個 term 的 document 總數

其中 $\ln(N/N_t)$ 又稱為 IDF，表示此 term 的重要性。

而至於為什麼用在 subword 而非 word? 有什麼好處？

因為 subword 可以避免掉 word 會產生的 OOV 問題，並且可以支援多國語言，也較不佔空間。

然後我們就可以利用此頁最下方那個關係式描述某個 document 跟某個 query 的相似度(懶的打了xD)，然後下頁就告訴我們範例，

像是 Blind Relevance Feedback 就是利用 vector space model 先找出第一批相關 documents，然後再從裡頭找出更相關的 documents、...

query expansion by term association 先建一個 matrix，把所有 term 之間的關係都描述出來，在同一個 document 出現的 term 相關度就較高，

我們可以試著把一個 query "紐約受到恐怖攻擊的新聞" 拆成多個 term，

1.紐約 2.恐怖攻擊 3.新聞

然後重新修正 query 使得 relevance(1,2) relevance(2,3) relevance(1,3) 為最高

代表我問的問題中的字串 彼此相關程度比較高

像是 relevance(吃飯,恐怖攻擊) = 0.8 relevance(牛肉麵,恐怖攻擊) = 0.01

則就不建議 query 問 "當我吃牛肉麵被恐怖攻擊" 的新聞，而找 "當我吃飯被恐怖攻擊" 的新聞，即使你是比較關心牛肉麵。

5. CH14

EM

目標：根據某種標準(可以自己定)，估計出機率模型的參數。

two steps:

(1)E-step(expectation):

給定一個目標函式，例如可能性函式(likelihood function)、 $P(x|\theta)$ ，和觀察資料(observation data)，機率模型的參數(θ)，藉由一些潛在性資料(latent data)，例如 HMM problem 3 裡的 state sequence，去求出此目標函式的值。

給兩個 model， θ 和 θ_k ， θ_k 是第 k 個 iteration 的參數， θ 是我們要估計的參數，將兩個 model 帶入目標函式，算出估計值。

(2)M-step(maximization):

用 E-step 算出一系列的估計值，比較出最大的，此 θ 即為的 k+1 個 iteration 的參數 θ_{k+1} ，我們要保證 $P(x|\theta_{k+1}) \geq P(x|\theta_k)$ (單項遞增)。

6.

參考答案：15.0 page 2

```

likelihood ratio test
if P(X|H0)/P(X|H1) > P(H1)/P(H0) = th then
    choose H0
else
    choose H1

```

其中 H_0, H_1 為二個假設，根據這二個假設和已知可以求出事前機率 $P(H_0), P(H_1)$ ，當有一個現象產生時，將此現象套用這二個假設，即 $P(X|H_0), P(X|H_1)$ ，並將此二機率相除，若大於 th ，則表示 H_0 假設較吻合，反之表示 H_1 假設較吻合。

在 DSP 應用上，有 utterance verification、frame-level、confidence score(可延伸至 phone、word、multi level)

在 utterance verification 部分，可把二個假設換成是二個 model，一個是 word 的 model w_i ，一個是 anti-model(主要是背景、雜音) w'_i ，當有 observation 時，就放進兩個 model 去算機率，然後相除，當 $P(X|w_i)/P(X|w'_i) > th$ 時，表示這個 observation 是一個 word 的機率較高，反之這個 observation 是雜音或其他干擾聲音的機率較高，如此一來就可辨認某時刻的 observation 是人所發出的聲音還是雜音。

7. CH16

what&how

X_n :使用者第 n 次的對話輸入

F_n :將 X_n 翻譯出的語意 (類似達叔說的機器的 input?)

S_n :到對話目前為止(第 n 個部分)的語意 (應該是機器到第 n 步驟的理解狀況)

A_n :第 n 次對話後，系統的回應 (像是重複使用者說的話，或是進行確認等等)

three key elements:

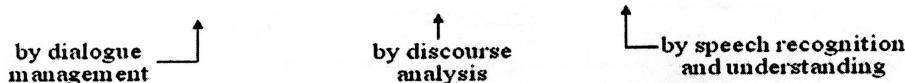
1. 語音辨識和理解-將 X_n 轉換成 F_n
2. 前文累積知識的分析-根據所有的 F_n ，將 S_{n-1} 轉換成 S_n
3. 對話處理-根據 S_n 選擇當時最佳的 A_n

link

(感覺是把講義上那兩行公式打上去?)

$$A_n^* = \arg \max_{A_n} \text{Prob}(A_n | X_n, S_{n-1})$$

$$A_n^* \approx \arg \max_{A_n, S_n} P(A_n | S_n) \sum_{F_n} P(S_n | F_n, S_{n-1}) P(F_n | X_n, S_{n-1})$$



F_n : semantic interpretation of the input speech X_n

Discourse: 語段

我覺得

S_n 改成像是 FSM 的 state 解釋比較恰當， X_n 是使用者講出的一串連續的話
(第 n 次應該就是用講話的停頓來分開的)

X_n 經由系統轉換成系統看的懂得 F_n ， F_n 可能有許多個語意包含在裡面
(ex: 我想要訂七點到巴黎的飛機 => 1.七點 2.巴黎 兩個 information)

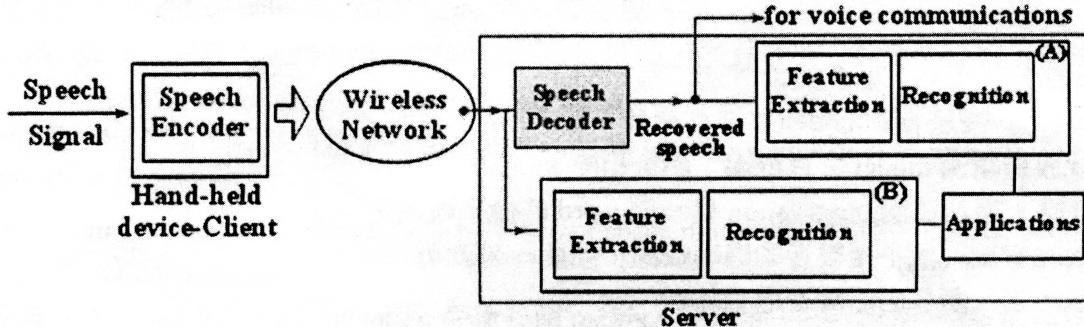
所以公式裡面有一個 sum，把第 n 次的所有 Fn 合起來
 這樣的 Fn 集合當作一個 FSM 的 input 形式，來產生下一個 state

8. CH17

這節的重點就像第 2 頁下面那張圖那樣，可以從 Client(這裡比較偏向討論 hand-held device，因為還要考慮 Wireless Network)送出服務要求，經由無線網路傳 Server。

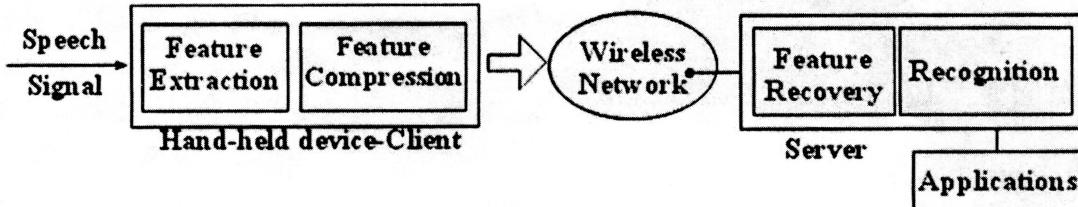
接收與解析語音分成三個部份：

Client-Only Model :



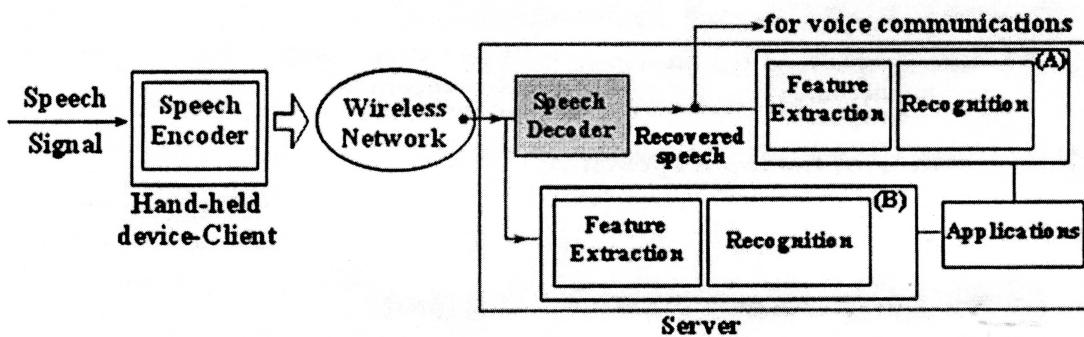
接收語音在 Client(廢話，說話的人是 Client)，解析的工作也在 Client 執行(就是手機內建整個語音處理系統)，再把最後解析出來的結果送到 Server。這個模式會受限於 hand-held device 的等級

Client-Server Model :



先在 Client 做好 Feature Extraction & Feature Compression，簡單說就是解析語音的各種特色(MFCC ?)，再經由無線網路送到 Server，在 Server 做剩下解析語音的工作。此舉動可以節省頻寬但是聲音會變奇怪，不好聽，影響講電話的品質(not compatible to existing wireless voice communications)在 Server，沒辦法從 MFCC 轉回原始的聲音

Server-Only Model :



送原始的聲音到 Server，再做處理。因為傳輸的都是原始的聲音，所以會比較好聽
 可是因為一般電話傳輸的不是全部的聲音，比較傾向人耳會聽的部份，所以會失真。

need to find recognition efficient feature parameters out of perceptually efficient feature parameters:

recognition efficient feature parameters(轉 MFCC)

perceptually efficient feature parameters(CELP)

找兩者參數的對應關係，但兩者所需參數因為不同需求，已經有所差異，因此參數之間無法完全對應

因為轉換時會刪掉相對不重要的部份，但刪去的部份可能是另一個方法的參數所需要的部分

server-only 應該不是送原始聲音？用 CELP encode 之後不會失真，但送到 server 轉 MFCC 去解析會使正確率下降

Principal Component Analysis (PCA)
研究人員的分析中往往涉及許多變數，要了解這麼多變數的相關型態是非常困難的。這些變數可能有高度相關，使很多資訊是重複的，也可能導致嚴重的多元共線性問題。

PCA 是一個可以將資料重新表達的方法，也就是 PCA 結果以新的相互不相關的變數取代原有相關的變數，此新的變數稱為「主成分」或「線性組合」，亦其為 principal components。(將變數視為多維空間的座標軸，PCA 的概念是對這些座標軸進行旋轉(線性組合)而已)

如果有過多重複資訊，PCA 能夠將原始資訊濃縮，使分析問題更明確且簡易。研究者必須去決定到底要取幾個主成份，取得數量越少越容易管理分析，而取得數量越多，就可獲得較多的資訊。PCA 有一個很大的優點，就是它不會有多元共線性的問題。

1. 聲音特質、相關參數、區分。

特性向量，挑取出來

eigenvoice：挑出 k 個音軸。

FIRK 太多

eigenvector, k 個 dimension

即最重要特徵

以特質分參數、特質

不想 distribution 太混亂，分開。

一個聲音一堆參數，究竟哪個？

eigenvector 反映聲音。

原 dimension \rightarrow 這是模擬之

傳統投影垂直很直，於是……

Principal Component Analysis (PCA)

研究人員的分析中往往涉及許多變數，要了解這麼多變數的相關型態是非常困難的。這些變數可能有高度相關，使很多資訊是重複的，也可能導致嚴重的多元共線性問題。

PCA是一個可以將資料重新表達的方法，也就是PCA結果以新的相互不相關的變數取代原有相關的變數，此新的變數稱為原有變數之線性組合，稱其為 principal components。(將多維視聽多度空間的座標軸，PCA的觀念是對這些座標軸進行旋轉(線性組合)而已)

如果有過多重複資訊，PCA能夠將原始資訊濃縮，使分析問題更明確且簡易。研究者須去決定到底要取幾個主成分，取得數量越少越容易管理分析，而取得數量越多，就可獲得較多的資訊。PCA有一個很大的優點，就是它不會有多元共線性的問題。

1. 聲音特質：相關參數區分

特性主成向量，擷取出來

eigenvalue：找出 k 個聲音軸。

FIRK太多

eigenvector：k 個 dimension

或最重要的一個

以聲音質為參照特質

不穩 & variation 太混亂，公母

-1個聲音 -1個聲軸 少量噪音

eigenvector 是次序

原子 dimension → 適切 模擬

後續投影重五很多對，請見 11.11.11.11

2007 6/26

1. (9.0 p.6~7)

a. how the eigenvoice space is constructed

對於每個 training speaker 我們可以 train 出其 speaker dependent phone model，然後將這個 speaker 的 phone model 的 Gaussian mean vectors 串接起來成一個 super vector。那對於 L 個 training speaker 來講，我們就有 L 個 super vector。那就可以把 L 個 super vector 當成 random vector x ，則每一個 training speaker 就是在 N (super vector 的大小) 維空間中的一個 vector。

再來利用 PCA 來找到這個 random vector x 的 k 個 orthonormal basis vectors set 而 $k \leq N$ ，這個找出來的 orthonormal basis vectors set 就是 L 個 speaker 的最重要 phone model 的特色，而不是所有的特色。這個 orthonormal basis vectors set 建立了這個 Eigenvoice space。

b. what that mean

Eigenvoice space 就是利用 speakers 最重要的特色，來調一個有不可見的資料的 model。

c. why rapid....

因為在利用 PCA 找到 random vector x 的 k 個 orthonormal basis vectors set 時，就一起減少了 phone model 特色的數量，所以在新的 speaker 加入需要調 model 時，就可以使用新 speaker 的少量資料來對照到已經在原有的 eigenvoice space 的已知點，藉由此已知點來得到全部的資料，而這個動作會比整個重頭調 model 來得快得很多，而且不需要這個 speaker 的整個 phone model。

6. (15.0 p1)

一種 statistical test，根據 likelihood ratio 決定要採用哪一個假設。

2 hypotheses: H_0, H_1 . 事前機率: $P(H_0), P(H_1)$

observation: $X : P(X|H_0), P(X|H_1)$

Likelihood ratio test: $P(H_i|X) = P(X|H_i) \times \frac{P(H_i)}{P(X)}, i = 0, 1$

根據 MAP principle:

當 $P(H_0|X) > P(H_1|X)$ 時選 H_0 $P(X|H_0) \times \frac{P(H_0)}{P(X)} > P(X|H_1) \times \frac{P(H_1)}{P(X)}$

當 $P(H_1|X) > P(H_0|X)$ 時選 H_1

則：

當選 H_0 時

$$\underbrace{\frac{P(X|H_0)}{P(X|H_1)}}_{>} > \underbrace{\frac{P(H_1)}{P(H_0)}}_{<}$$

當選 H_1 時 =>

$$\frac{P(X|H_0)}{P(X|H_1)} < \frac{P(H_1)}{P(H_0)}$$

CMS:

透過在 MFCC 中的 Fourier transform 將 convolution 變成相乘，取 log 之後相乘就會變相加，假設原本的 speech 的 mean 是 0，那只要對 input signal 取 mean 就能得到 convolution noise 的 mean。只要減掉 convolution noise 之後就能回到原本的 speech。

CMVN:

除了平均值之外變異係數通常也會有縮小的現象，因此就會造成訓練與測試特徵的不匹配，因此除了將 mean 變成 0 之外，將變異係數也調成 1

HEQ:

希望測試特徵和訓練特徵能夠有相同的統計分佈特性，因此將測試特徵和訓練特徵的 CDF (cumulative distribution function) 同時必進一個機率分佈。

7.

Client-server model

將計算分成兩部份在 Client 端抽 feature 後做 MFCC，壓縮後傳到 server 端後再做 recognition 以及 application。其中部份小的放在 client，部份大的放在 server，而且因為壓縮所以可省頻寬。

問題在於跟現有手機不相容，現在手機抽的 feature 是所謂的 perceptually efficient，這些 feature 可以 recover 到聽起來跟原本很像的聲音，但其實已經差很多了，跟 MFCC 裡的不一樣，MFCC 無法直接 recover 到跟原本很像的聲音，如果要在手機裡實做兩種抽 feature 的方法又跟現有手機不相容。雖然 MFCC 再加 pitch 可以 recover 到跟原來很像的聲音，但是也跟現在的手機會有不相容的問題。

Server-only model

將 perceptually efficient 的 feature 抽出來後傳到 server 端再解回來，跟現有的手機不會有不相容的問題，這樣的聲音聽起來可以跟原因很像，但是用這個解回來的聲音抽 feature 再做 recognition 正確率會變差，因為兩種 feature 是不同的，但跟現在的手機是相容的，只要改 server 就可以了。

Client-server model

在 client 端抽 Feature 以及 MFCC

在 Server 端做 Recognition 以及 Application

跟現在的手機不相容，但 recognition 正確率較高，且 MFCC 壓縮。

Server-only model

在 client 端抽 Feature。

在 Server 端做 MFCC、Recognition 以及 Application

跟現在的手機相容，但 recognition 正確率較差。

application:

Utterance Verification

語音辨識技術

最大期望演算法

在統計計算中，最大期望（EM）演算法是在機率（probabilistic）模型中尋找參數最大似然估計的演算法，其中機率模型依賴於無法觀測的隱藏變數（Latent Variable）。最大期望經常用在機器學習和計算機視覺的數據集聚（Data Clustering）領域。最大期望演算法經過兩個步驟交替進行計算，第一步是計算期望（E），也就是將隱藏變數象能夠觀測到的一樣包含在內從而計算最大似然的期望值；另外一步是最大化（M），也就是最大化在 E 步上找到的最大似然的期望值從而計算參數的最大似然估計。M 步上找到的參數然後用於另外一個 E 步計算，這個過程不斷交替進行。

最大期望過程說明

我們用 \mathbf{y} 表示能夠觀察到的不完整的變數值，用 \mathbf{x} 表示無法觀察到的變數值，這樣 \mathbf{x} 和 \mathbf{y} 一起組成了完整的數據。 \mathbf{x} 可能是實際測量丟失的數據，也可能是能夠簡化問題的隱藏變數，如果它的值能夠知道的話。例如，在混合模型（Mixture Model）中，如果「產生」樣本的混合元素成分已知的話最大似然公式將變得更加便利（參見下面的例子）。

估計無法觀察到的數據

讓 p 代表向量 θ : $p(\mathbf{y}, \mathbf{x}|\theta)$ 定義的參數的全部數據的機率分佈（連續情況下）

或者機率集聚函數（離散情況下），那麼從這個函數就可以得到全部數據的最大似然值，另外，在給定的觀察到的數據條件下未知數據的條件分佈可以表示為：

$$p(\mathbf{x}|\mathbf{y}, \theta) = \frac{p(\mathbf{y}, \mathbf{x}|\theta)}{p(\mathbf{y}|\theta)} = \frac{p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)}{\int p(\mathbf{y}|\mathbf{x}, \theta)p(\mathbf{x}|\theta)d\mathbf{x}}$$

主成分分析(PCA)

在統計學中，主成分分析（principal components analysis (PCA)）是一種簡化數據集的技術。它是一個線性變換。這個變換把數據變換到一個新的坐標系統中，使得任何數據投影的第一大方差在第一個坐標(稱為第一主成分)上，第二大方差在第二個坐標(第二主成分)上，依次類推。主成分分析經常用減少數據集的維數，同時保持數據集的對方差貢獻最大的特徵。這是通過保留低階主成分，忽略高階主成分做到的。這樣低階成分往往能夠保留下數據的最重要方面。但是，這也不是一定的，要視具體應用而定。

接著下面我們試著再闡述一些 **eigenvoice** 與 MAP 及 MLLR 之間的比較與不同之處

簡論而言，MAP 是結合了事前機率及由要 approach 的語料來決定要 approach 哪些 model 來產生給 new speaker 的 model，也因此需要很多的參數量。也就是要很多語料才行。

而 MLLR 的主要概念是，使用迴歸類別的理論，所謂的迴歸類別，理論上是由一群高斯分佈所組成。即假設說不同模型間共享一個線性轉換，藉此減低許多參數需要量。但仍需一定程度的量才行。

而本篇 report 所討論的 **eigenvoice** approach，一如前面所說，是以 PCA 來建構出一個最少的向量空間，藉此在極少的調適語料時，辨識率就能高點。而事實上，基底使用越多，就會越慢達到高點，

PCA 主成分分析，簡單的來說，一方面能保有原來變數的資訊(代表性)，而且主成份間也不能重疊(獨立性)，它是能以"少數"幾個主成份來代替原來"多個"解釋變數(精簡性)，有這三個主要特性。

打個比分，有 10 個學生，皆要考國文、英文、數學三科，而如何來代表一個學生的總成績呢，在主成分分析，就是先算出這三科的加權，然後每個學生的三個成績去乘以這個加權，所得到的一個'總成績'

之後要比較時，所拿這個每個學生總成績去做比較即可

要算加權須要經過二個步驟

- 1.共變異數矩陣
- 2.特徵值、特徵向量

取最大的特徵值的特徵向量(即:加權) 乘以 原資料 即 第一主成分值(即:總成績)

你要算第二主成分值也可以(就拿次大的特徵值)，但第一主分值通常就有很大的代表性了。

另外它能解決複迴歸分析時常遇到的困擾，就是所謂共線性問題。

=====

PCA

擁有Euclidean space vector norm 最小的特性。簡單來說，就是當我們用PCA 把資料降維後，用傳統的vector norm 下來看，誤差會是最小的。

LDA

擁有降維後能將各群data 分開的特性，但是以vector norm 來看，誤差不保證是最小的。

LDA 的推導

我們的目的是要找到一個vector w ，把資料投影到 w 上面去，得到新的coordinate y 。

$$y = w_t x$$

以LDA 的精神來看，是希望能將同一類的資料投影得越近越好，不同類的資料 map 得越遠越好。為了描述這個概念，我們需要用一些量化的值去表示，首先是每個 class data 的平均值(mean)。

$$m_i = \frac{1}{n_i} \sum_{x \in D_i} x$$

而投影過後的平均值是：

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in Y_i} y = \frac{1}{n_i} \sum_{x \in D_i} w^t x = \frac{1}{n_i} w^t \sum_{x \in D_i} x = w^t m_i$$

n_i 是第*i* 類的資料個數。 D_i 是第*i* 類資料的集合。 Y_i 是投影後第*i* 類資料的集合。

所以以上可以看出投影過後的每個class data 的平均值是原來在高維度空間的平均值投影。

再來我們可以定出投影後兩類資料的平均距離了

$$|\tilde{m}_1 - \tilde{m}_2| = |w^t (m_1 - m_2)|$$

我們也可以定出兩類資料在投影過後，分散的度量(scatter)

$$\tilde{s}_i^2 = \sum_{y \in Y_i} (y - \tilde{m}_i)^2$$

再來就依照LDA 的精神，投影過後的兩類資料越分開越好，就代表他們投影過後的平均值差越多越好。投影過後的同類資料越集中越好，就是投影過後的分散程度越小越好。我們可以因此得到以下的函式：

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

當決定一個投影 basis w ，我們可以求出一個 $J(w)$ 的值，我們希望分母越小越好，分子越大越好，我們很直覺地可以聯想到，求極值的方式不就是lagrange method 嗎？沒錯，當找出一個 w 可使 $J(w)$ 出現最大值，那個 w 就是我們要的basis。但是以上的 $J(w)$ 右邊的形式是間接跟 w 有關係，我們再做一下推導使得右邊的式子能跑出 w 這個 term 出來。

看

Digital Speech Processing, Final Exam

June. 17, 2008, 10:10-12:10

- OPEN Lecture Notes, Power Point(Printed Version) and Personal Notes.
- You have to use CHINESE sentences to answer all of the problems, but you can use English terminology
- Total points: 120
- Note that you don't need to be able to answer all the questions.

1. (20) In Maximum Likelihood Linear Regression (MLLR) approach, explain how the mean vectors of Gaussian mixtures in acoustic models are adjusted, why speaker adaptation can be achieved with relatively smaller quantity of data, and what are the limitations?

2. (20) In Latent Semantic Analysis the elements w_{ij} of the word-document matrix \bar{W} is
LSA

$$w_{ij} = (1 - \varepsilon_i) \frac{c_{ij}}{n_j}$$

Where c_{ij} is the number of times the word w_i occurs in the document d_j , n_j is the total number of words in d_j , and

$$\varepsilon_i = -\frac{1}{\log N} \sum_{j=1}^N \left(\frac{c_{ij}}{t_i} \right) \log \left(\frac{c_{ij}}{t_i} \right), t_i = \sum_{j=1}^N c_{ij},$$

where N is the total number of documents. Explain the meaning of the parameters w_{ij} , and the meaning of each row and column of this matrix.

3. (20) Explain why subword units are useful in retrieving speech information using speech queries, and if they introduce extra problems?

4. (20) For Linear Discriminative Analysis (LDA), explain the meaning of within/between-class scatter matrices S_W/S_B , and the meaning of the optimization criterion.

5. (20) Write down the two steps in each iteration of EM algorithm and explain how they operate and what they mean.

6. (20) Explain the three key elements in a spoken dialogue system: what they are, how they operate, and how they are linked together.