# 聯合語者、雜訊環境與說話內容因素分析之強健性語音辨認

國立台北科技大學電子工程系
Department of Electronic Engineering
National Taipei of Technology

吳聖堂　Sheng-Tang Wu
t8418093@ntut.edu.tw

方偉德　Wei-Te Fang
t9418025@ntut.edu.tw

廖元甫　Yuan-Fu Liao
yfliao@ntnt.edu.tw

## 摘要

摘要—本論文主要研究於強健性語音辨認上，我們提出聯合語者、雜訊環境與語音內容因素分析(Joint Speaker and Noisy Environment and Speech Content Factor Analysis；JSEC)，主要是透過聯合因素分析，在特徵空間做即時語音辨認模型補償(online recognition model compensation)，使得調適出來的模型與測試環境能夠盡量匹配，進而提升辨識效果。此外，我們先將 JSEC 分解成語音和非語音二個模型做模型調適、估算影響因素，接著每個模型再利用階層式的概念，語音特性考慮之因素分成雜訊環境特徵空間、語者特徵空間、說話內容特徵空間與獨特因素空間分別估算，非語音特性考慮之因素則分成雜訊特徵空間和獨特因素空間分別估算，最後再把語音和非語音組合回辨認用的模型，用此方式來降低我們的參數量。我們使用 Aurora2 語料庫做實驗，在複合情境的訓練模式下，我們得到最佳的辨識錯誤率為 4.37%，比傳統強健性參數求取方法 MVA (Mean subtraction，Variance normalization，and ARMA filtering)[1][2]的錯誤率 4.99%低了許多，也比我們先前提出的 JSE (Joint Speaker and Noisy Environment Factor Analysis)[11]方法的錯誤率相當甚至好一點。除了辨認率之外，我們提出的方法也能使得調適模型的參數量大幅下降，JSEC 參數量約為傳統 MVA 的 4 倍，也比 JSE 方法少了十分之一的參數量，因此為更有效率的調適方法。

關鍵詞：強健性語音辨認，因素分析，Aurora2

## 一、緒論

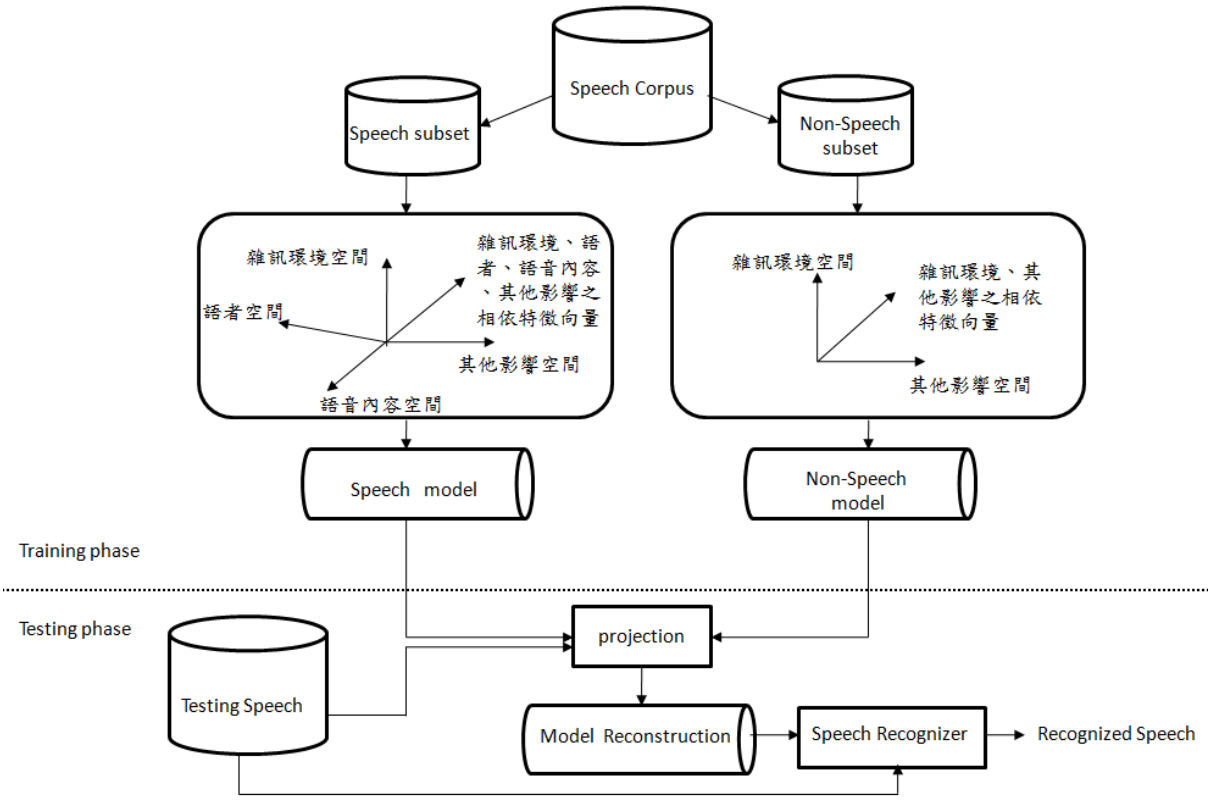
語音辨認系統受雜訊環境、語者特性與通道效應等影響，導致辨識率下降。通常處理這些影響因素或環境不匹配問題，有兩種較常見的方法：強健性語音參數求取(robust speech feature extraction)與語音模型調適(speech model adaptation)。

強健性參數求取之方法，我們可以舉幾個經典的例子：倒頻譜正規化 ARMA (Auto-regression and Moving Average)濾波技術(Mean subtraction，Variance normalization，and ARMA filtering；MVA)[1][2]、分布等化法(Histogram Equalization；HEQ)[3][4]，與兩階式維納濾波器(two-stage Wiener filter)[5]等，它們的特點是有效且容易實現。

至於模型調適方法，又可分為是否需要先驗知識，不需要先驗知識的方法，主要有：最大相似度線性回歸(Maximum Likelihood Linear Regression；MLLR)[6]、最大事後機率調適法(Maximum A Posteriori；MAP)[7]調適法、平行模型結合(Parallel Model Combination；PMC)[8]等，以上皆為經典且常見之語音模型調適法，經常被應用於語音和語者辨認系統。

而需先驗知識的方法，常見的方法如雙聲源為基礎之分段線性補償(Stereo-based Piecewise Linear Compensation, SPLICE)[9]。我們也曾利用事先收集大量語者與環境先驗知識，提出基於雜訊環境參考模型內插法(Reference Model Weighting；RMW)、雜訊特徵最大相似度線性迴歸(Eigen-Maximum Likelihood Linear Regression；EMLLR)[10]，基於聯合語者與雜訊環境因素分析 (Joint Speaker and Noisy Environment Factor Analysis；JSE)[11]等方法，效果皆相當不錯。

此論文我們提出了需先驗知識的語者、雜訊環境與語音內容因素分析之強健性語音辨認(Joint Speaker and Noisy Environment and Speech Content Factor Analysis；JSEC)， JSEC 主要運用在雜訊分析處理，所考慮的影響因素及估算順序如圖一的 JSEC 架構圖。




圖一、JSEC 架構圖

JSEC 在訓練端將訓練語料分成兩類，一類為左邊具有語音特性之語句做影響因素之分類，另一類為右邊非語音特性之語句做影響因素之分類，再利用階層式的概念，將語音特性分成了雜訊環境特徵空間、語者特徵空間、說話內容特徵空間與獨特因素空間分別估算，非語音特性為雜訊特徵空間與獨特因素空間分別估算，最後再把語音和非語音組合回辨認用的模型。當我們得到不同影響因素的空間後，最後在測試端，輸入測試語料後，測試語料對個別特徵空間做投影，即可對模型做即時的調適。

## 二、聯合語者、雜訊環境與語音內容因素分析

### 2.1 JSEC 模型表示

JSEC 主要考慮測試語料在辨識時，受到雜訊環境、語者、語音內容與其他因素的影響。而 JSEC 比 JSE 多考慮的語音內容影響，可分為具有語音特性的部分，以及非具有語音特性。

我們定義具有語音特性之 JSEC 模型是由古典 MAP(Classical MAP)、特徵雜訊環境、特徵語者、語音內容(zero~nine、oh、silence)四個模型結合而成：

$$M_{speech} = m_{sp} + u_{sp}x_{sp} + v_{sp}y_{sp} + g_{sp}r_{sp} + d_{sp}z_{sp} \quad (1)$$

而非語音特性之 JSEC 模型是由特徵雜訊與古典 MAP(Classical MAP)模型結合而成：

$$M_{nonspeech} = m_{non} + u_{non}x_{non} + d_{non}z_{non} \quad (2)$$

其中：

$m_{sp}$、$m_{non}$：由語音參數串接而成的超向量，模型參數串接而成的超向量。

$x_{sp}$、$x_{non}$：特徵雜訊環境空間的投影量，初始假設平均值為 0 變異數為 1 的標準高斯分佈。

$y_{sp}$：特徵語者特徵空間的投影量，初始假設平均值為 0 變異數為 1 的標準高斯分佈。

$r_{sp}$：語音內容特徵空間的投影量，初始假設平均值為 0 變異數為 1 的標準高斯分佈。

$z_{sp}$、$z_{non}$：獨特因素特徵空間的投影量，初始假設平均值為 0 變異數為 1 的標準高斯分佈。

$u_{sp}$、$u_{non}$：特徵雜訊環境特徵空間。

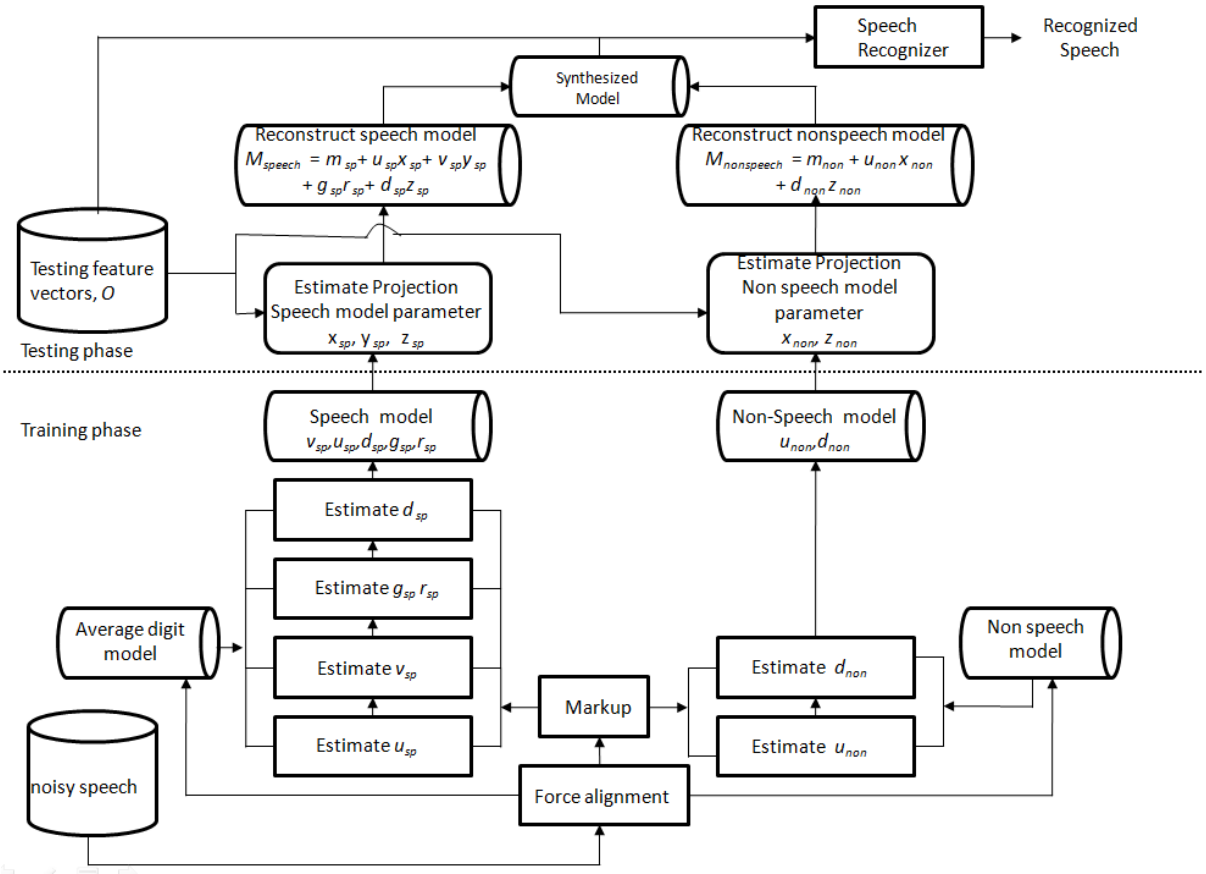
$v_{sp}$：特徵語者特徵空間。

$d_{sp}$、$d_{non}$：獨特因素特徵空間。

$g_{sp}$：語音內容特徵空間。

## 2.2 JSEC 系統架構

圖二是 JSEC 之系統流程圖，在訓練端，我們將訓練語料做 Force-alignment，變成不同語音內容的片段語句。具有語音特性的片段語句訓練一個名為 speech 的聲學模型，我們便是利用這種方式來降低最後重建模型之參數量。並且對具有語音特性的片段語句做標記分類，接著再依序估算雜訊、語者、說話內容，最後是獨特因素的特徵空間，分別以 $u_{sp}$、$v_{sp}$、$g_{sp}$、$d_{sp}$ 表示。非具有語音特性的片段語句，則訓練一個 non speech 的模型，並且僅對不同雜訊做標記，同於具有語音特性的部分。接著依序估算雜訊與獨特因素特徵空間，分別以 $u_{non}$ 與 $d_{non}$ 表示。



圖二、JSEC 之系統流程圖

在測試端，我們估算測試語料具有語音特性影響因素的投影量 $x_{sp}$、$y_{sp}$、$z_{sp}$，然後投影到建立好的 $u_{sp}$、$v_{sp}$、$d_{sp}$，得到偏移量 $u_{sp}x_{sp} + v_{sp}y_{sp} + d_{sp}z_{sp}$；與非具有語音特性影響因素的投影量 $x_{non}$、$z_{non}$，然後投影到建立好的 $u_{non}$、$d_{non}$，得到偏移量 $u_{non}x_{non} + d_{non}z_{non}$。得到兩者偏移量後，另外再估算訓練端具有語音特性之說話內容的偏移量 $g_{sp}r_{sp}$，用意是把單一的聲學模型，可以依照不同說話內容之影響，調適為不同語音內容特性的聲學模型，然後再加上以未切割語料訓練的聲學模型、靜音模型與停頓模型部分，即可重建出每一句測試語料獨有的模型，最後做辨識結果。

在得到所需要的轉換矩陣之後就可以進行第二步驟，也就是將原始參數向量對轉換矩陣做內積運算，就能轉換成新參數向量，然後送進模型訓練。接下來兩段將敘述主成分分析和線性鑑別分析的原理和實作的步驟。

2.3 特徵空間的估計

我們類似於參考文獻[12]之古典 MAP、特徵語者和特徵通道等方法，表示各種因素的關係。而由不同高斯混合分布(mixture)的共變異數串接而成的對角矩陣則可作為參數估測的初始值。

本文所提到的聯合因素分析是參考[13]的作法，將語音模型，利用擷取 average speech model 的平均值所構成的超向量作為基準，就像是傳統的 MAP 語者調適一樣，而由不同混合成分的共變異數$\sum_c$串接而成的對角矩陣$\sum$則可作為參數估測的初始值。在模型參數估測之前先定義系統的機率假設。

波氏統計

我們先使用波氏統計[14]主要是以average speech model的平均值、變異數以及權重來計算機率統計量。假設語者$s$以及語者特徵向量$\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{t}$，對於每一個混合成分$c$，我們定義波氏統計如下：

$$N_c(s) = \sum_t \gamma_t(c) \tag{3}$$

$$F_c(s) = \sum_t \gamma_t(c)\left(Y_t - m_c\right) \tag{4}$$

$$S_c(s) = diag\left(\sum_t \gamma_t(c)\left(Y_t - m_c\right)\left(Y_t - m_c\right)^*\right) \tag{5}$$

其中：

$\gamma_t(c)$代表語者特徵向量於時間時落於混合成分的事後機率，而$m_c$代表average speech model中混合成分的$c$平均值。接著設$N(s)$為$CF$ x $CF$的對角矩陣，其中的對角區塊是由

$N_c(s)$ ($c$=1,...,$C$)所構成。設$F(s)$為$CF$ x 1的超向量，是由每一個$F_c(s)$ ($c$=1,...,$C$)串接而成。設$S(s)$為$CF$ x $CF$的對角矩陣，其中的對角區塊是由$S_c(s)$ ($c$=1,...,$C$)所構成。

2.3.1 語者、雜訊環境、語音內容特徵空間

求得波式統計量之後，由參考文獻[13]我們可以計算出具有語音特性的語者、雜訊環境特徵空間，和非語音特性的雜訊環境特徵空間。

語者、雜訊環境、語音內容特徵空間估算方法相同，但是語音內容算出來的隱藏變數$r_{sp}$，必須儲存起來給測試端使用，因為在辨認的時候並不知道要說哪些語音內容,先假設每一個 model 平均值在哪裡，再利用 ML 法重估超參數取得 $g_{sp}$ 之後，即可將說話內容投影到對應位置。

由於 $u_{sp}$, $v_{sp}$, $g_{sp}$, $u_{non}$ 其估計方法相同，以下我們以具有語音特性的特徵語者(Eigen-voice)模型為例，求取 $v_{sp}$。

我們利用 Expectation Maximization(EM)演算法進行 10 次的疊代，反覆更新 $v_{sp}$ 使之趨於定值：

**隱藏變數$y_{sp}$事後分佈**

先假設隱藏變數$y_{sp}(s)$是平均值為0變異數為1的標準高斯分佈，當我們輸入語音資料後就像MAP語者調適一樣會有不同的分佈。根據[12]假設，令$L_{sp,y}(s) = I_{sp,y} + V^*{}_{sp,y}(s)\sum{}_{sp,y}{}^{-1}(s)N_{sp,y}(s)V_{sp,y}(s)$，其中$\sum{}_{sp,y}{}^{-1}$為variance之超向量，而隱藏變數$y_{sp}(s)$的分佈可由平均值$L^{-1}{}_{sp,y}(s)V^*{}_{sp,y}(s)\sum{}_{sp,y}{}^{-1}(s)F_{sp,y}(s,m)$與共變異數$L^{-1}{}_{sp,y}(s)$去描述該機率分佈。

**ML法重估超參數**

初始參數 $m$ 與$\sum$是擷取自 average speech model 的相關組合，參數 $v_{sp}$ 則是採用隨機的初始值，假設語者為 $s$，定義累積的統計量如下：

$$N_c = \sum_s N_c(s) \quad (c = 1, \dots, C) \tag{6}$$

$$U_c = \sum_s N_c(s)\, E\left[y_{sp}(s) y^*{}_{sp}(s)\right] \quad (c = 1, \dots, C) \tag{7}$$

$$C' = \sum_s F(s)\, E\left[y^*{}_{sp}(s)\right] \tag{8}$$

$$N = \sum_s N(s) \tag{9}$$

對每一個混和成分 $c$ =1,...,$C$ 、每一個混合成分的元素 $f$ =1,...,$F$ ，設 $i$ = ( $c$ -1)$F$ + $f$，令 $u_i$ 代表 $u$ 的第 $i$ 列，而 $C_i'$ 代表 $C'$ 的第 $i$ 列，因此特徵特徵空間 $v_{sp}$ 的更新公式可表示成：

$$v_i U_c = \mathrm{C}_i \quad (i=1,....,CF) \tag{10}$$

上述的表示式，可以從參考文獻[13]得到相關的表示。

2.3.2 獨特因素特徵空間

由於 $d_{sp}, d_{non}$ 其估計方法相同，以下我們以具有語音特性的獨特因素模型為例，求取 $d_{sp}$。

我們利用 Expectation Maximization(EM)演算法進行 10 次的疊代，反覆更新 $d_{sp}$ 使之趨於定值。

**隱藏變數$d_{sp}$事後分佈**

先假設隱藏變數 $z_{sp}(s)$是平均值為 0 變異數為 1 的標準高斯分佈，當我們輸入語音資料後就像 MAP 語者調適一樣會有不同的分佈。根據[12]假設，令$L_{sp,d}(s) = I_{sp,d} + d^2{}_{sp,d}(s)\sum_{sp,d}{}^{-1}(s)N_{sp,d}(s)$，其中$\sum_{sp,d}{}^{-1}$為 variance 之超向量，而隱藏變數 $z_{sp}(s)$的分佈可由平均值$L^{-1}{}_{sp,d}(s)d_{sp,d}(s)\sum_{sp,d}{}^{-1}(s)F_{sp,d}(s,m)$與共變異數$L^{-1}{}_{sp,d}(s)$去描述該機率分佈。

**ML法重估超參數**

初始參數 $m$ 與$\sum$是擷取自 average speech model 的相關組合，參數 $d_{sp}$ 則是採用隨機的初始值，假設語者為 $s$，定義累積的統計量如下：

$$N_c = \sum_s N_c(s) \quad (c = 1, \dots, C) \tag{11}$$

$$U_c = \sum_s diag\ \left(N(s)\ E\left[z_{sp}(s)z^*{}_{sp}(s)\right]\right)\ (c = 1, \dots, C) \tag{12}$$

$$\mathrm{C} = \sum_s diag\ \left(F(s)\ E\left[z^*{}_{sp}(s)\right]\right) \tag{13}$$

$$N = \sum_s N(s) \tag{14}$$

對每一個混和成分 $c$ =1,...,$C$ 、每一個混合成分的元素 $f$ =1,...,$F$ ，設 $i$ = ( $c$ -1)$F$ + $f$，令 $u_i$ 代表 $u$ 的第 $i$ 列，而 $C_i$ 代表 $C$ 的第 $i$ 列，因此特徵特徵空間 $d_{sp}$ 的更新公式可表示成：

$$v_i U_c = \mathrm{C}_i \quad (i=1,....,CF) \tag{15}$$

上述的表示式，可以從參考文獻[13]得到相關的表示。

2.3.3 投影量 *x, y, r, z* 的估計

當我們在訓練端得到求取出的具有語音特性的超參數 $u_{sp}$、$v_{sp}$、$g_{sp}$、$d_{sp}$，以及非語音特性的超參數 $u_{non}$、$d_{non}$ 後，測試端的參數再依照具有語音特性之影響因素，經過估算而得到個別雜訊影響之投影量 $x_{sp}$、語者影響之投影量 $y_{sp}$、說話內容之投影量 $r_{sp}$、獨特因素之投影量 $z_{sp}$，非語音特性之影響因素一樣經過估算而得到投影量 $x_{non}$、$z_{non}$。

語音特性和非語音特性之投影量算法一樣，我們以估算語音部分的投影量為例：

**雜訊影響之投影量$x_{sp}$**

$$\text{假設} L_{sp,x}(s) = I_{sp,x} + V_{sp,x}{}^{*}(s)\sum_{sp,x}{}^{-1}(s)N_{sp,x}(s)V_{sp,x}(s) \tag{16}$$

$$x_{sp}= \mathrm{E}[x(s)] = L_{sp,x}{}^{-1}(s)u_{sp,x}{}^{*}(s)\sum_{sp,x}{}^{-1}(s)F_{sp,x}(s,m) \tag{17}$$

**語者影響之投影量$y_{sp}$**

$$\text{假設} L_{sp,y}(s) = I_{sp,y} + V_{sp,y}{}^{*}(s)\sum_{sp,y}{}^{-1}(s)N_{sp,y}(s)V_{sp,y}(s) \tag{18}$$

$$y_{sp}= \mathrm{E}[y(s)] =L_{sp,y}{}^{-1}(s)v_{sp,y}{}^{*}(s)\sum_{sp,y}{}^{-1}(s)F_{sp,y}(s,m) \tag{19}$$

**語音內容影響之投影量$r_{sp}$**

$$\text{假設} L_{sp,r}(s) = I_{sp,r} + V_{sp,r}{}^{*}(s)\sum_{sp,r}{}^{-1}(s)N_{sp,r}(s)V_{sp,r}(s) \tag{20}$$

$$r_{sp}= \mathrm{E}[r(s)] =L_{sp,r}{}^{-1}(s)g_{sp,r}{}^{*}(s)\sum_{sp,r}{}^{-1}(s)F_{sp,r}(s,m) \tag{21}$$

**獨特因素之投影量$z_{sp}$**

$$\text{假設} L_{sp,z}(s) = I_{sp,z} + d_{sp,z}{}^{2}(s)\sum_{sp,z}{}^{-1}(s)N_{sp,z}(s) \tag{22}$$

$$z_{sp}= \mathrm{E}[z(s)] =L_{sp,z}{}^{-1}(s)d_{sp,z}{}^{*}(s)\sum_{sp,z}{}^{-1}(s)F_{sp,z}(s,m) \tag{23}$$

得到 $x_{sp}$、$y_{sp}$、$r_{sp}$、$z_{sp}$ 後，投影到 $u_{sp}$、$v_{sp}$、$g_{sp}$、$d_{sp}$ 特徵空間，即可對模型做即時的調適，重建出每句測試語料獨有的辨認模型，而變異數、轉移機率與權重之影響很小，故暫且假設不考慮變異數、轉移機率與權重等問題。

## 三、實驗結果與分析

### 3.1 實驗設定

本論文實驗是以國際上廣泛用在雜訊環境語音辨識技術強健性的標準語料庫 Aurora 2 為主。Aurora2 是以 TIDigits 為基礎，加上不同雜訊以及通過特定的通道效應製成。Aurora2 是一個連續數字串語料庫，每句音段包含一至七個連續數字，長度最多不超過三秒鐘。

語料首先通過理想濾波器將 20 kHz 降頻為 8 kHz，此為定義的乾淨(Clean)語料，每個乾淨音段先經特定的通道效應，再依各種訊雜比(SNR20、SNR15、SNR 10、SNR 5、SNR 0 和 SNR -5dB)加上不同的加成性雜訊。

訓練語料混合各種訊雜比及不同環境雜訊的複合情境訓練訓練模式。測試語料部分則是依照原本 Aurora2 自行建立的不同通道效應與加成性雜訊，共分成 A、B、C 三種測試組合。

本論文採用梅爾倒頻譜係數，及聲學模型為連續性密度的隱藏式馬可夫模型，模型的狀態轉移只停留在原始狀態，及由左至右轉移到下一個相鄰的狀態。

數字聲學模型的單位為全詞模型，十一個英文數字聲學模型(0～9 和 oh)。每個聲學模型有 16 個狀態，每個狀態含 3 個高斯分布模型。除數字聲學模型外，還有靜音(silence)模型和停頓(short pause)模型。辨識效能評估上，採取辨識詞錯誤率，這考慮了刪除型錯誤、插入型錯誤和取代型錯誤。我們實驗分為二類：simple backend (3 mixture)和 complex backend (20 mixture)，如表一所示。

| Backend | Speech model | Silence model | Short pause model |
|---|---|---|---|
| Simple | 16 state, each state 3 mixture | 3 state, each state 6 mixture | 1 state, each state 6 mixture |
| Complex | 16 state, each state 20 mixture | 3 state, each state 64 mixture | 1 state, each state 64 mixture |

表一、複合情境訓練模式各種參數組合辨識結果

另外我們實驗對照需要用到我們先前提出的 JSE 方法，其模型可表示為：

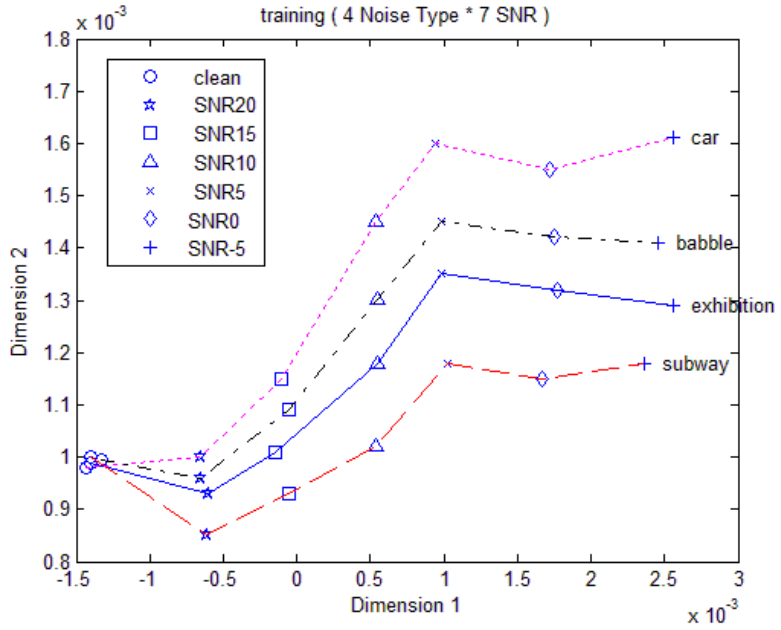
$$M = m + ux(s) + vy(s) + dz(s) \tag{24}$$

其中 $m$ 為初始模型中所有平均值串成的超向量(super-vector)；$v$、$u$、$d$ 分別為特徵聲音、特徵雜訊、獨特因素之特徵空間；$vy(s)$、$ux_h(s)$、 $dz(s)$分別為人聲、雜訊、獨特因素在各自特徵空間的平均偏移量。和 JSEC 最大不同在於少考慮了講話內容因素，模型的特徵向量比 JSEC 龐大。

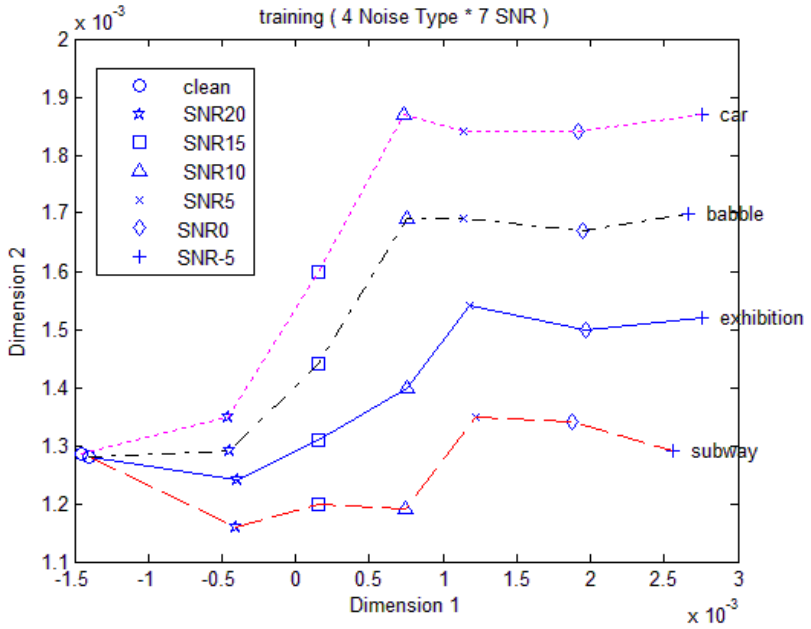
### 3.2 特徵空間分析

我們想要得知此方法是否正確，能不能有效地將影響因素個別分開，所以先使用 simple

backend 分析特徵空間並畫出特徵空間投影圖，目的是讓各種不同雜訊的測試語料，能投影到正確的特徵空間上。在建構特徵空間時，我們先估算已經做好分類資訊的統計量，接著再依照 u、v、g、d 之順序，逐一估算個別之特徵空間。爲了方便分析，我們取前兩維的特徵向量作 x 軸和 y 軸，建構一個二維空間，首先以雜訊類型(地下鐵、人聲、汽車、展覽會)做分析，我們採用七種 SNR(clean、SNR20、SNR15、SNR10、SNR5、SNR0、SNR-5)做特徵空間分析，其結果如圖三、圖四所示。




圖三、simple backend JSEC 語音特性之雜訊特徵空間投影圖

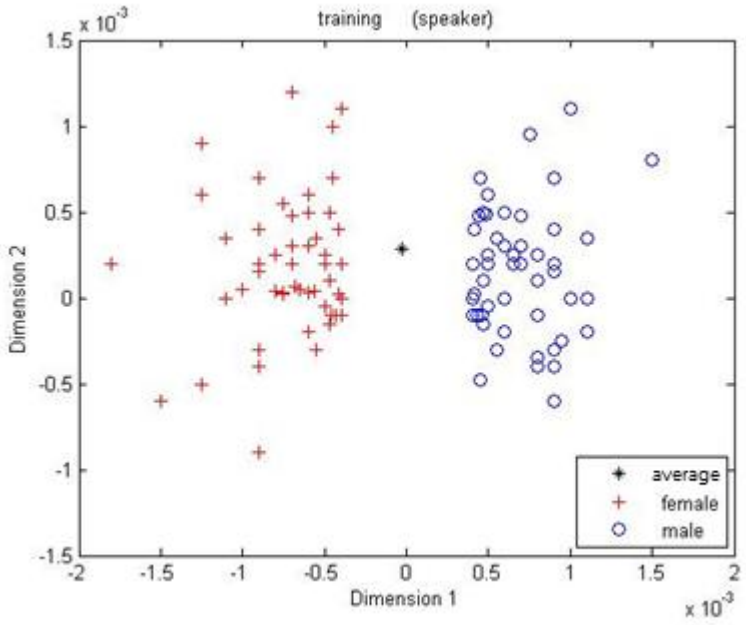



圖四、simple backend JSEC 非語音特性之雜訊特徵空間投影圖

我們可以看到圖三、圖四，在 clean 端，雜訊特性並不明顯，隨著 SNR 增加，雜訊特性

越來越明顯，而末端的線條便跟著逐漸分開。由以上之特徵空間投影圖，我們得知求出來之特徵空間能夠有效地將這些干擾因素個別分開，提升辨識效能。

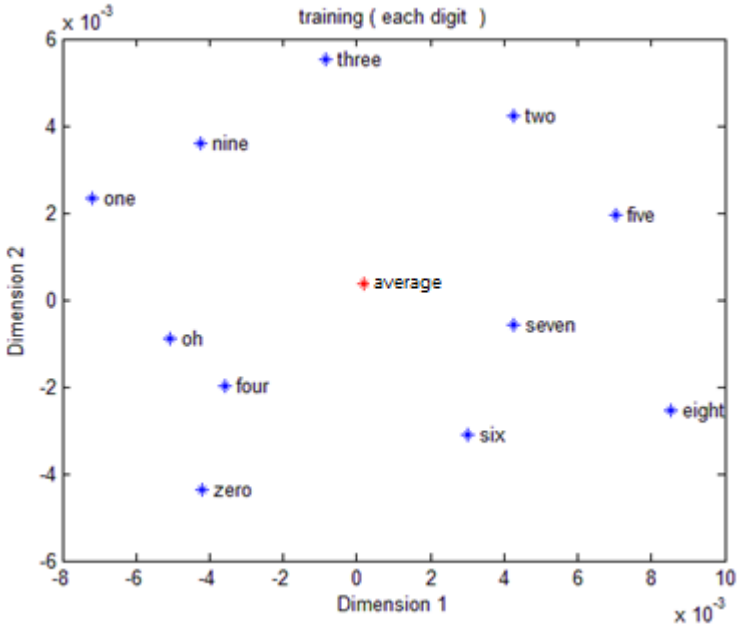
接著我們要做語者特性分析，如圖五：




圖五、simple backend JSEC 語者特徵空間投影圖

在圖五中，我們可以看到其投影結果，很明顯依照語者的不同被分成兩邊，我們以「o」與「+」的符號分別表示男生與女生的特性。

最後是語音內容之特徵空間投影分析，其結果如圖六所示：



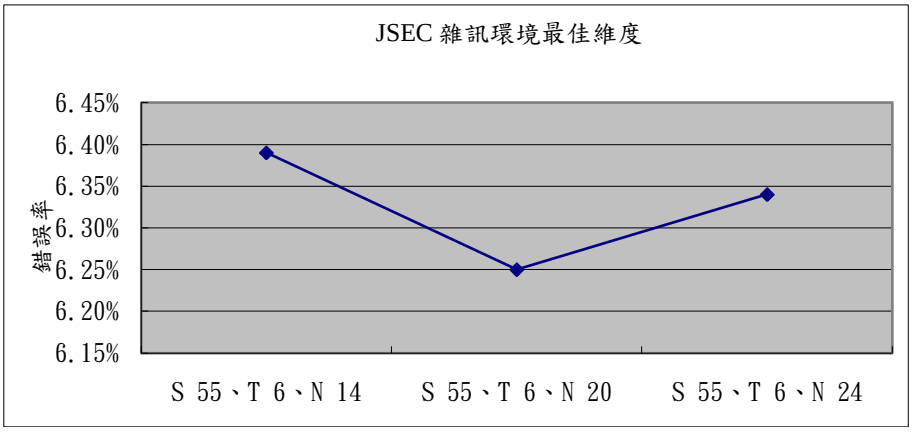
圖六、simple backend JSEC 語音內容之特徵空間投影圖

我們可以看到圖六其投影結果，依照語音內容被分開，而比較類似的音，例如 oh、four，似乎會比較靠近，而複合情境的點(digit)大約是在所有點的平均位置。由以上三種不同影響因素之特徵空間投影圖，我們預測辨識效果應當不錯。

## 3.3 simple backend

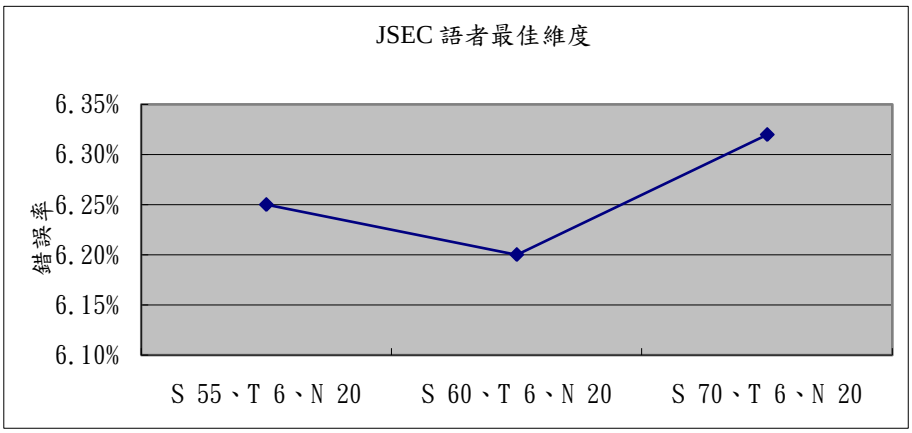
我們的實驗為了要有效率的找出特徵空間的最佳維度，首先固定語者(S) 55 維，語音內容(T) 6 維。雜訊(N)維度共 40 維，所以我們從 20 維開始找最佳效果，並且一次往上或往下增加 6 維(14 維、20 維和 24 維)尋找最佳維度。另外，由於調適模型中的變異數、轉移機率與權重影響很小，因此先假設與比較的 MVA、JSE 相同，並且把實驗分成 simple backend 和 complex backend 二組做維度組合分析。

辨識結果如圖七所示：



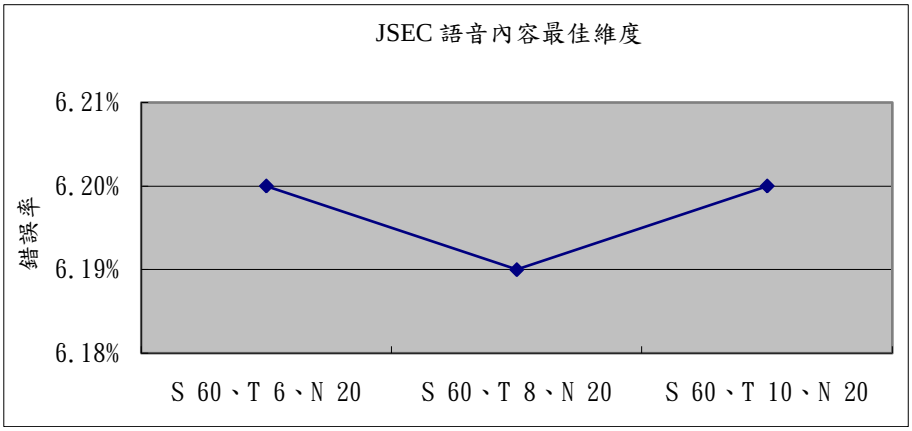

圖七、simple backend JSEC 雜訊環境最佳維度比較圖

圖七我們可以看到雜訊的最佳維度是 20 維，因此我們接著固定雜訊 20 維，語音內容一樣取 6 維，再取語者 55 維、60 維和 70 維，做測試可以得到以下結果：



圖八、simple backend JSEC 語者最佳維度比較圖

圖八我們可以看到語者的最佳維度是 60 維，因此我們接著固定雜訊 20 維，語者取 60 維，再取語音內容 6 維、8 維和 10 維，做測試可以得到以下結果：

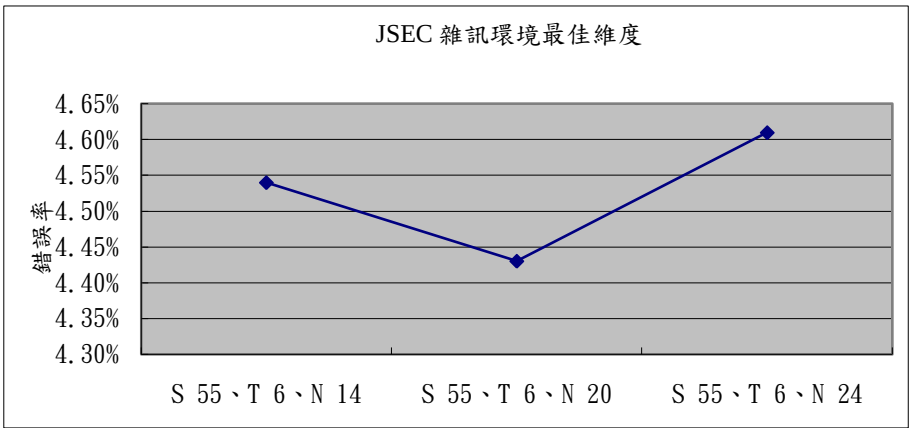



圖九、simple backend JSEC 語者最佳維度比較圖

圖九我們可以看到語音內容的最佳維度是 8 維。

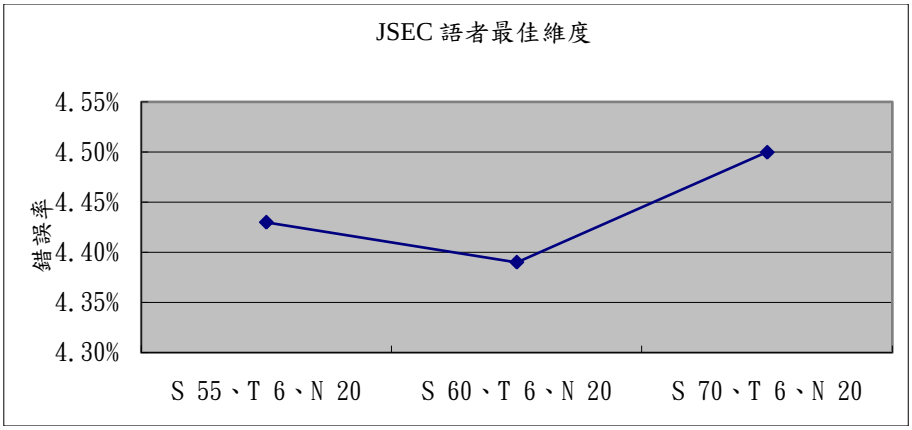
3.3.2 Complex backend

由於調適模型中的變異數、轉移機率與權重影響很小，因此先假設與比較的 MVA、JSE 相同。維度測試首先固定語者(S) 55 維，語音內容(T) 6 維，雜訊(N)分別以 14 維、20 維和 24 維做測試後可得以下結果：



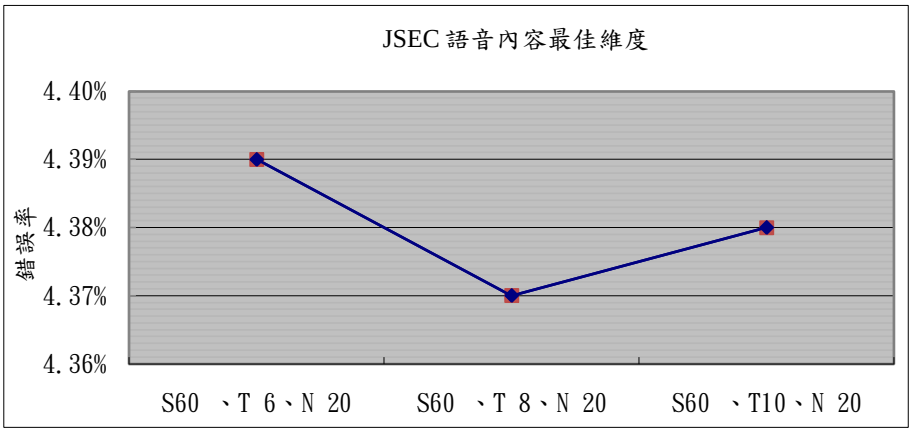

圖十、complex backend JSEC 雜訊環境最佳維度比較圖

由圖十，我們可以看到雜訊的最佳維度是 20 維，因此我們接著固定雜訊 20 維，語音內容一樣取 6 維，再取語者 55 維、60 維和 70 維，做測試可以得到以下結果：




圖十一、complex backend JSEC 語者最佳維度比較圖

由圖十一我們可以看到語者的最佳維度是 60 維，因此我們接著固定雜訊 20 維，語者取 60 維，再取語音內容 6 維、8 維和 10 維，做測試可以得到以下結果：




圖十二、complex backend JSEC 語音內容最佳維度比較圖

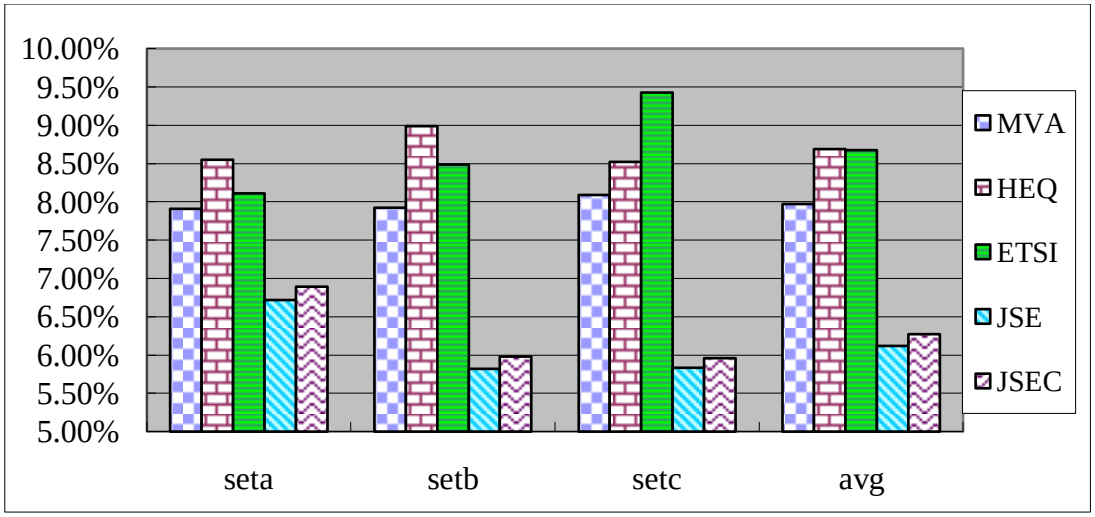
由圖十二，我們可以看到語音內容的最佳維度是 8 維。

### 3.4 實驗結果與討論

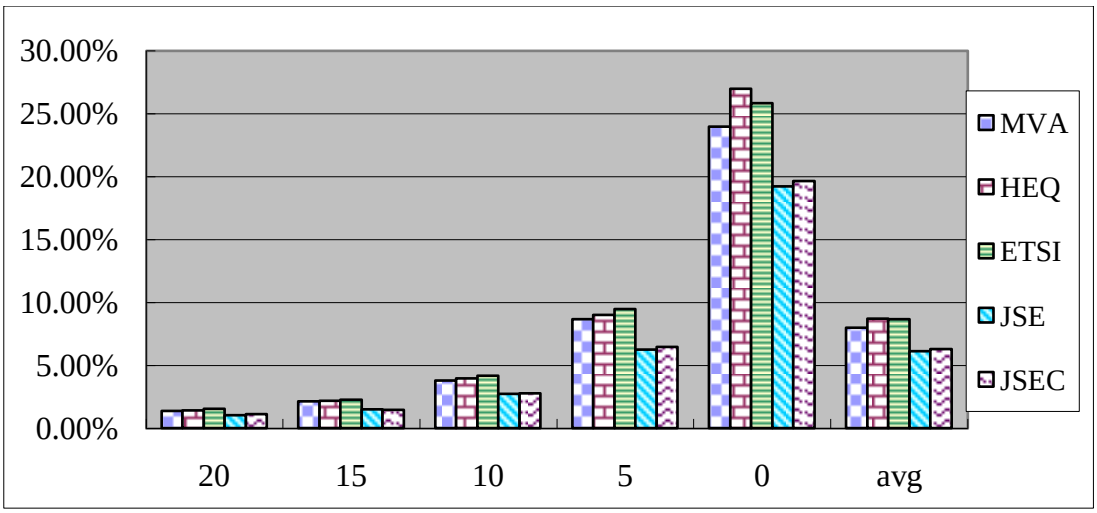
最後我們參數設定使用最佳的雜訊 20 維、語者 60 維、語音內容 8 維，和其他系統方法做實驗對照，並分成 simple backend 和 complex backend 二組實驗討論。

#### 3.4.1 Simple backend

從圖十三和圖十四我們發現，JSE 與 JSEC 平均錯誤率遠優於 MVA 的 7.97%，但是 JSEC 卻略差 JSE 0.15%，我們認為 JSEC 由於聲學模型變成只有一個時，做 simple backend 的實驗，可能會導致模型不夠複雜，因此造成辨識率下降。




圖十三、simple backend 各系統方法不同環境之比較圖




圖十四、simple backend 各系統方法不同 SNR 之比較圖

但在調適模型所需的參數量的方面，如表二和圖十五。我們可以從圖十五發現，調適模型所需的參數量明顯比原本 JSE 的方法降低很多，JSE 比 JSEC 多 9 – 10 倍的參數量，而 JSEC 只比 MVA 多了 4 倍的參數量，效能更好、運算量更小。

| simple backend | | | |
|---|---|---|---|
| 模型所需參數 | MVA | JSE | JSEC |
| mean | 21528 | 21528 | 2808 |
| variance | 21528 | 21528 | 2808 |
| weight | 552 | 552 | 72 |
| Transition | 3598 | 3598 | 358 |

| | | | |
|---|---|---|---|
| $u$ | – | 430560 | – |
| $v$ | – | 1291680 | – |
| $d$ | – | 21528 | – |
| $u_{non}$ | – | – | 14040 |
| $d_{non}$ | – | – | 702 |
| $u_{sp}$ | – | – | 37440 |
| $v_{sp}$ | – | – | 112320 |
| $d_{sp}$ | – | – | 1872 |
| $g_{sp}$ | – | – | 14976 |
| $r_{sp}$ | – | – | 88 |
| 總共的參數量 | 47206 | 1790974 | 187484 |
| 參數量的比例 | 1 | 37.94 | 3.97 |

表二、simple backend MVA 與改變參數量的 JSEC 比較表

其中 JSEC 的 mean 與 variance=(39*3*16=1872)+(39*6*3=702)+(39*6*1=234)

=2808 (25)

Transition =18x18 +5x5+3*3 =358 (26)

weight =3*16+6*3+6*1 =72 (27)

$u_{non}$ =20x702 =14040 (28)

$d_{non}$ =1x702 =702 (29)

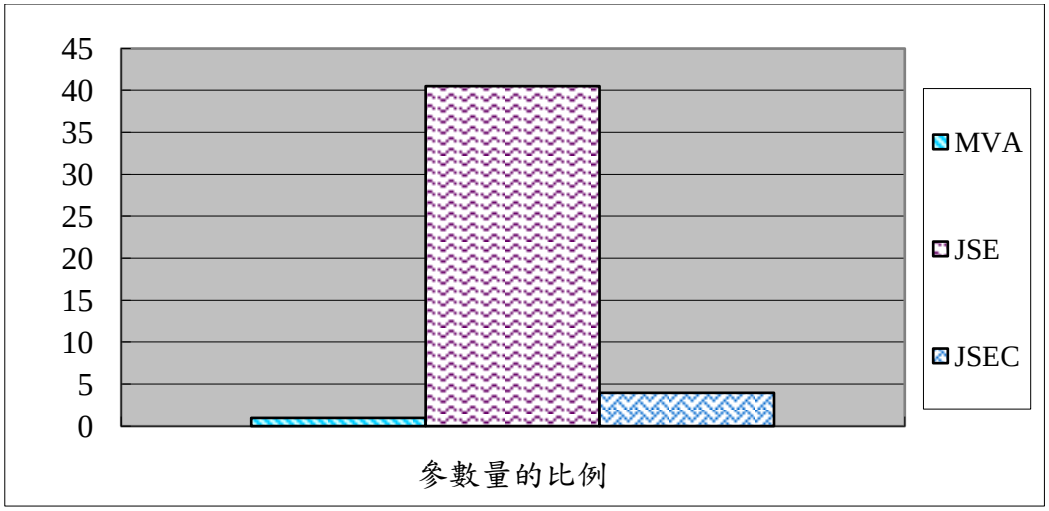
$u_{sp}$ =20x1872 =37440 (30)

$v_{sp}$ =60x1872 =112320 (31)

$d_{sp}$ =1x1872 =1872 (32)

$g_{sp}$ =8x1872 =14976 (33)

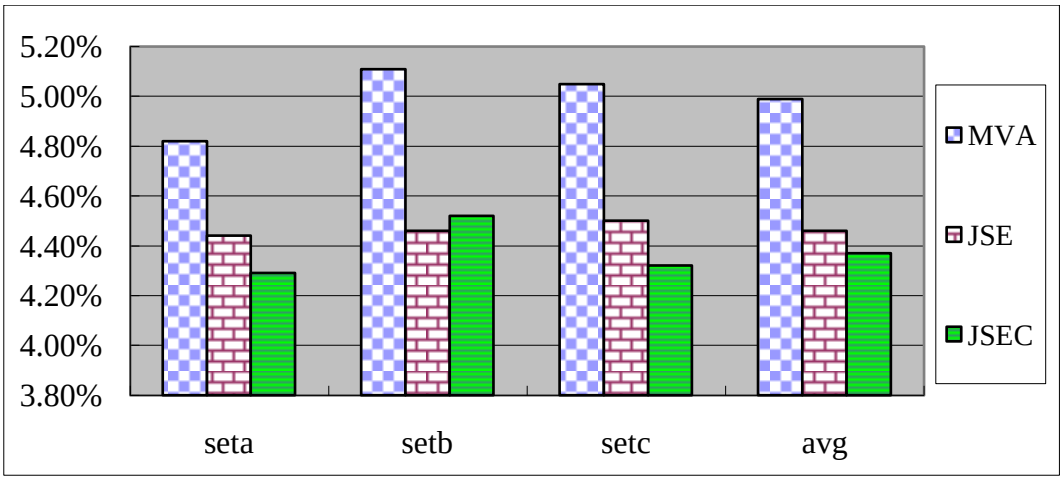
$r_{sp}$ =11x8 =88 (34)



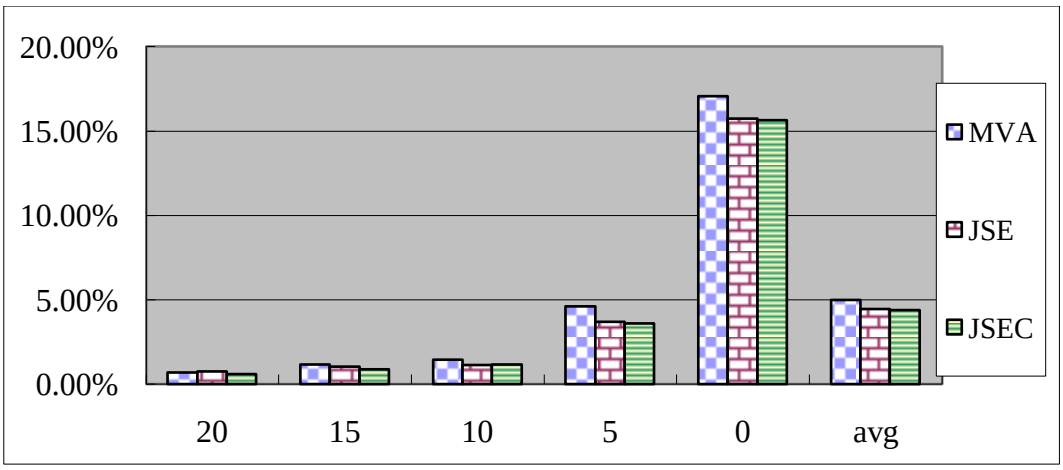

圖十五、simple backend MVA 與改變參數量的 JSEC 比例圖

3.4.1 Complex backend

我們另外做了一組 complex backend 實驗，把 mixture 數從 3 拉到 20，我們一樣使用最佳的雜訊 20 維、語者 60 維、語音內容 8 維，最後實驗數據如圖十六和圖十七。




圖十六、complex backend 各系統方法不同環境之比較圖



圖十七、complex backend 各系統方法不同 SNR 之比較圖

圖十六和圖十七可發現，mixture 從 3 拉到 20 之後，JSEC 錯誤率 4.37%，可以和 JSE 錯誤率 4.46%相當，甚至更低一點。

最後我們一樣統計了調適模型所需的參數量，如表三和圖十五所示：

| complex backend | | | |
|---|---|---|---|
| 模型所需參數 | MVA | JSE | JSEC |
| mean | 147264 | 147264 | 22464 |
| variance | 147264 | 147264 | 22464 |
| weight | 3776 | 3776 | 576 |
| Transition | 3598 | 3598 | 358 |
| $u$ | – | 2945280 | - |
| $v$ | – | 8835840 | - |
| $d$ | – | 147264 | - |
| $u_{non}$ | – | - | 149760 |
| $d_{non}$ | – | - | 7488 |
| $u_{sp}$ | – | - | 249600 |
| $v_{sp}$ | – | - | 748800 |
| $d_{sp}$ | – | - | 12480 |
| $g_{sp}$ | – | - | 99840 |
| $r_{sp}$ | – | - | 88 |
| 總共的參數量 | 47206 | 12230286 | 1313918 |
| 參數量的比例 | 1 | 40.51 | 4.35 |

表三、complex backend MVA 與改變參數量的 JSEC 比較表

其中，JSEC 的 mean 與 variance=(39*20*16=12480)+(39*64*3=7488) +(39*64*1=2496)

=22464 (35)

weight =20*16*11+64*3+64*1=576 (36)

$u_{non}$ =20x7488 =149760 (37)

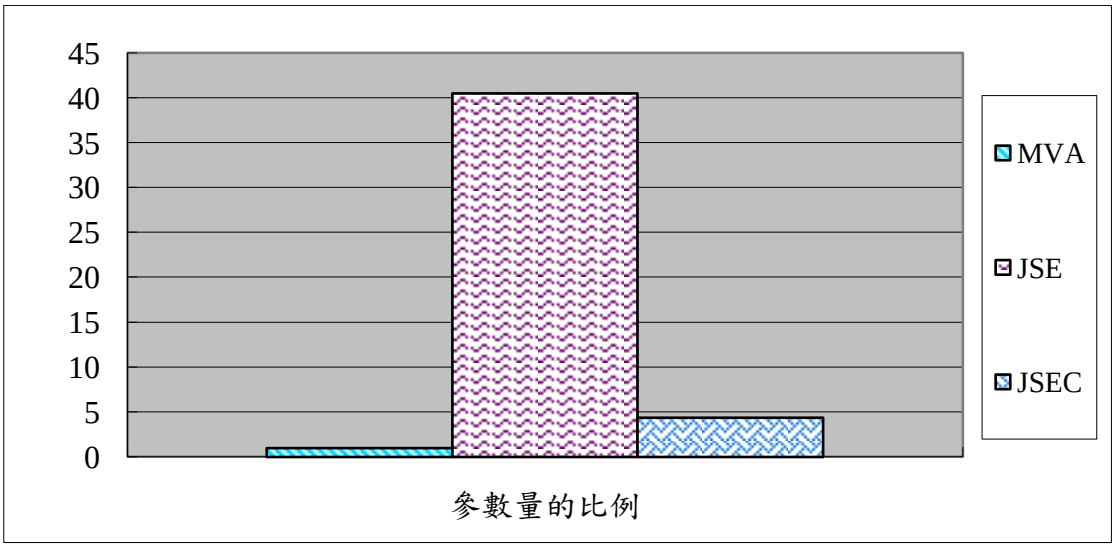
$d_{non}$ =1x7488=7488 (38)

$u_{sp}$ =20x12480=249600 (39)

$v_{sp}$ =60x12480=748800 (40)

$d_{sp}$ =1x12480=14800 (41)

$g_{sp}$ =8x12480=99840 (42)

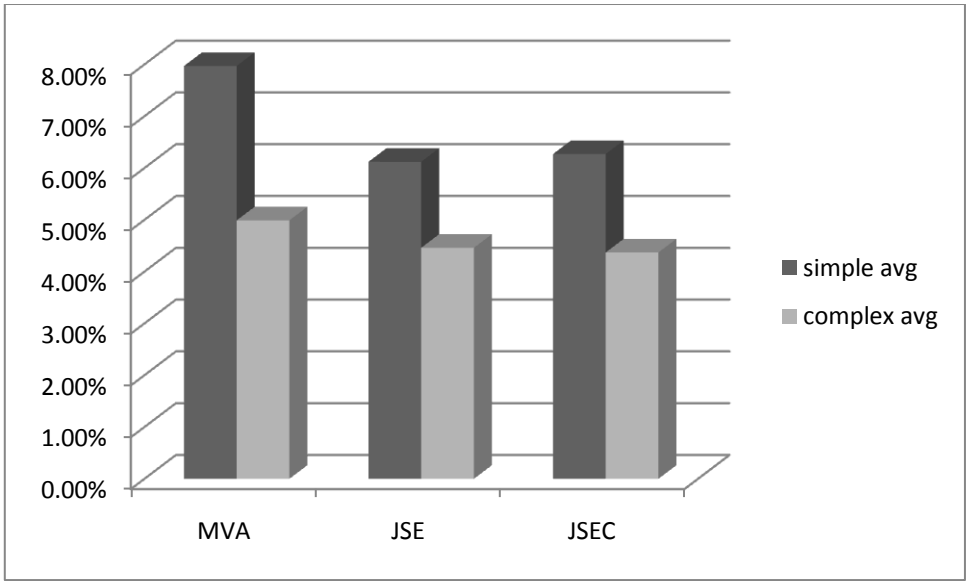



圖十八、complex backend MVA 與改變參數量的 JSEC 比例圖

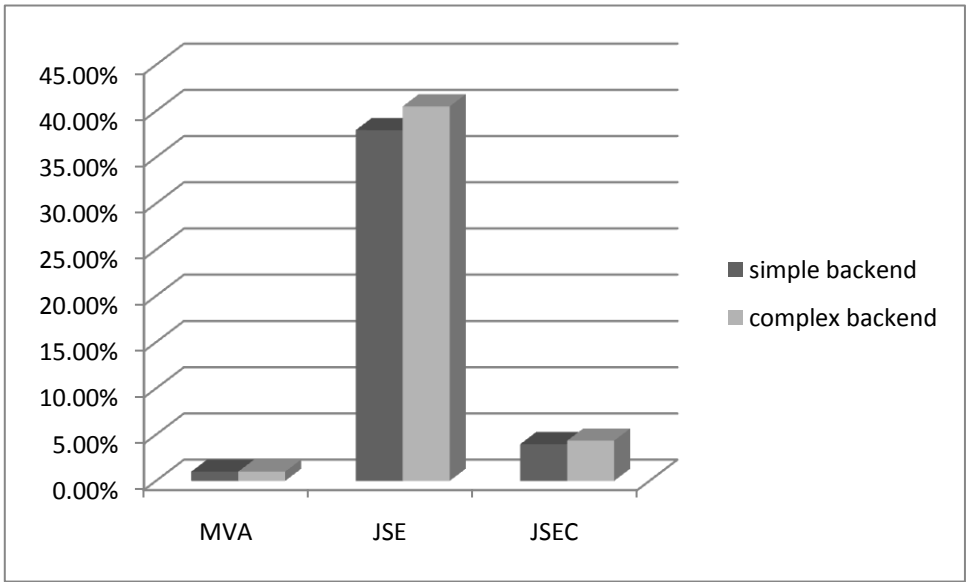
從圖十八我們可以觀察出參數量依然遠小於先前的 JSE，JSE 比 JSEC 多 9－10 倍的參數量，而 JSEC 只比 MVA 多了 4 倍的參數量，運算量低很多。

## 四、結論

本論文的主要研究目標是提出新的 JSEC 方法，並且使用 Aurora2 做實驗，實驗數據最後做了總整理如圖十九和圖二十。由圖中我們提出的 JSEC 在 complex backend 的實驗當中，我們發現可以和原來 JSE 系統的錯誤率差不多，甚至可以更低一些，而且也比傳統方法 MVA 錯誤率 4.99%低了許多；除了辨認率之外，我們提出的方法也能使調適模型的參數量比原來的 JSE 降低了十分之一，JSE 的參數量是 MVA 的 40 倍，但 JSEC 只比 MVA 多了 4 倍的參數量而已，因此為更有效率的調適方法。.




圖十九、平均錯誤率比較圖




圖二十、參數量比較之比例圖

## 五、致謝

本論文所進行工作之成果，部分在國科會專題計畫編號 98-2221-E-027-081-MY3 和 97-2628-E-027-003-MY3 的經費補助之下順利完成，特此致謝。

# 參考文獻


[1] C.P. Chen, K. Filali and J. Bilmes, “Frontend Post-Processing and Backend Model Enhancement on the Aurora 2.0/3.0 Databases,” in *Proc. ICSLP*, 2002.

[2] C.P. Chen, J. Bilmes and K Kirchhoff, “*Low-resource Noise-Robust Feature Post-Processing on Aurora 2.0*,” in *Proc. ICSLP*, 2002.

[3] A. de la Torre, J. C. Segura, M. C. Benitez, A. M. Peinado and A. J. Rubio, “*Non-linear transformation of the feature space for robust speech recognition*,” in *Proc. ICASSP*, vol. I, 2002.

[4] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. P. Cordoba, M. C. Benitez and A. J. Rubio, “*Histogram equalization of speech recognition for robust speech recognition*,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, 2005.

[5] ETSI standard document, “*Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm; back-end reconstruction algorithm*,” ETSI Standard ES 202 212, 2003.

[6] C.J. Leggetter and P.C. Woodland, “*Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models*,” *Computer Speech Lang.*, vol. 9, 1995.

[7] J.L. Gauvain and C.H. Lee, “*Maximum a Posteriori estimation for multivariate Gaussian mixture observations of Markov chains*,” *IEEE Trans.on Speech Audio Processing*, vol. 2, 1994.

[8] M. Gales and S. Young, “*Robust continuous speech recognition using parallel model combination*,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, September 1996.

[9] L. Deng, A. Acero, M. Plumpe and X. Huang. “*Large-Vocabulary Speech Recognition under Adverse Acoustic Environments*,” in Proc. ICSLP 2000.

[10] M.J.F. Gales and P.C. Woodland, “*Mean and variance adaptation within the MLLR framework*,” Computer Speech Lang., vol. 10, no. 3, pp. 249–264, 1996.

[11] 王瑞璟，基於聯合語者與雜訊環境因素分析之強健性語音辨認，國立台北科技大學電腦與通訊研究所碩士論文，2010 年，98 頁。

[12] P. Kenny, “*Joint Factor Analysis of Speaker and Session Variability : Theory and AlgorithmsMontreal*”, Technical report CRIM-06/08-13 Montreal, CRIM, 2005

[13] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P., “A Study of Inter-Speaker Variability in Speaker Verification,” IEEE Transactions on Audio Speech and Language Processing, July 2008.

[14] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", Ann. Math. Statist., vol. 41, no. 1, pp. 164–171, 1970.