

加成性雜訊環境下運用特徵參數統計補償法於強健性語音辨識  
Feature Statistics Compensation for Robust Speech Recognition in Additive Noise Environments

謝宗學 Tsung-hsueh Hsieh  
國立暨南國際大學電機工程學系  
Department of Electrical Engineering  
National Chi Nan University  
[s94323532@ncnu.edu.tw](mailto:s94323532@ncnu.edu.tw)

洪志偉 Jeih-weih Hung  
國立暨南國際大學電機工程學系  
Department of Electrical Engineering  
National Chi Nan University  
[jwhung@ncnu.edu.tw](mailto:jwhung@ncnu.edu.tw)

摘要

在自動語音辨識的研究上，如何有效地降低背景雜訊的影響，以增加語音辨識系統的強健性，一直是一大研究重點，其中語音特徵參數正規化法是廣為人用的強健技術之一。然而，對於多變的語音訊號該如何準確的估算出其不同時段的統計值，是影響語音特徵參數正規化法效果的一個重要因素。本論文主要是針對在加成性雜訊環境下，對不同的特徵參數提出更有效率且準確的統計值補償法，以降低加成性雜訊對語音特徵參數的影響。我們提出了運用虛擬雙通道碼簿為基礎之特徵參數補償技術，其中包含三種方法：倒頻譜統計補償法、線性最小平方回歸法與二次最小平方回歸法。我們將這些方法運用在四種不同的語音特徵參數的補償上，發現都能有效降低加成性雜訊對語音特徵的影響，進而大幅提升辨識率，同時，在與傳統以段落為基礎的特徵參數正規化技術比較下，我們所提出的方法可達到更佳的強健效果。

Abstract

In this paper, we propose several compensation approaches to alleviate the effect of additive noise on speech features for speech recognition. These approaches are simple yet efficient noise reduction techniques that use online constructed pseudo stereo codebooks to evaluate the statistics in both clean and noisy environments. The process yields transforms for noise-corrupted speech features to make them closer to their clean counterparts. We apply these compensation approaches on various well-known speech features, including mel-frequency cepstral coefficients (MFCC), autocorrelation mel-frequency cepstral coefficients (AMFCC), linear prediction cepstral coefficients (LPCC) and perceptual linear prediction cepstral coefficients (PLPCC). Experimental results conducted on the Aurora-2 database show that the proposed approaches provide all types of the features with a significant performance gain when compared to the baseline results and those obtained by using the conventional utterance-based cepstral mean and variance normalization (CMVN).

關鍵詞：自動語音辨識、虛擬雙通道碼簿、倒頻譜統計補償法、線性最小平方回歸法、二次最小平方回歸法

Keywords: automatic speech recognition、pseudo stereo codebooks、cepstral statistics compensation,

linear least squares regression, quadratic least squares regression

## 一、緒論

本論文主要重點是在加成性雜訊環境下，對語音特徵參數補償法的探討，目的是使測試語音的統計特性在經過補償後能更接近訓練語音的統計特性。

我們運用四種語音特徵參數擷取技術結合兩大類特徵參數補償法，並觀察兩類特徵參數補償法之間的差異與優缺點。本論文中所討論的兩大類特徵參數補償法分別為：

### (1)以段落為基礎之特徵參數正規化法

即傳統的整段式倒頻譜平均與變異數正規化法[1](utterance-based cepstral mean and variance normalization, U-CMVN)與分段式倒頻譜平均與變異數正規化法[2](segmental cepstral mean and variance normalization, S-CMVN)。前者是以一整段語句為基準去估算該維特徵參數的統計值，並執行特徵參數正規化法；後者則是將每段語句以一小段的片段為基準，去估算該片段的統計值，然後執行特徵參數正規化。

### (2)以碼簿為基礎之特徵參數補償法

為了更精確地估測語音特徵參數統計值，以執行特徵參數補償與正規化法進而消除雜訊影響，我們提出透過虛擬雙通道碼簿，來幫助我們更準確地估算出代表訓練語音與測試語音的統計值，並藉由較準確的統計值來執行特徵參數補償，以提升辨識效率。其中包含三種方法：倒頻譜統計補償法(cepstral statistics compensation, CSC)、線性最小平方回歸法(linear least squares regression, LLS)與二次最小平方回歸法(quadratic least squares regression, QLS)。

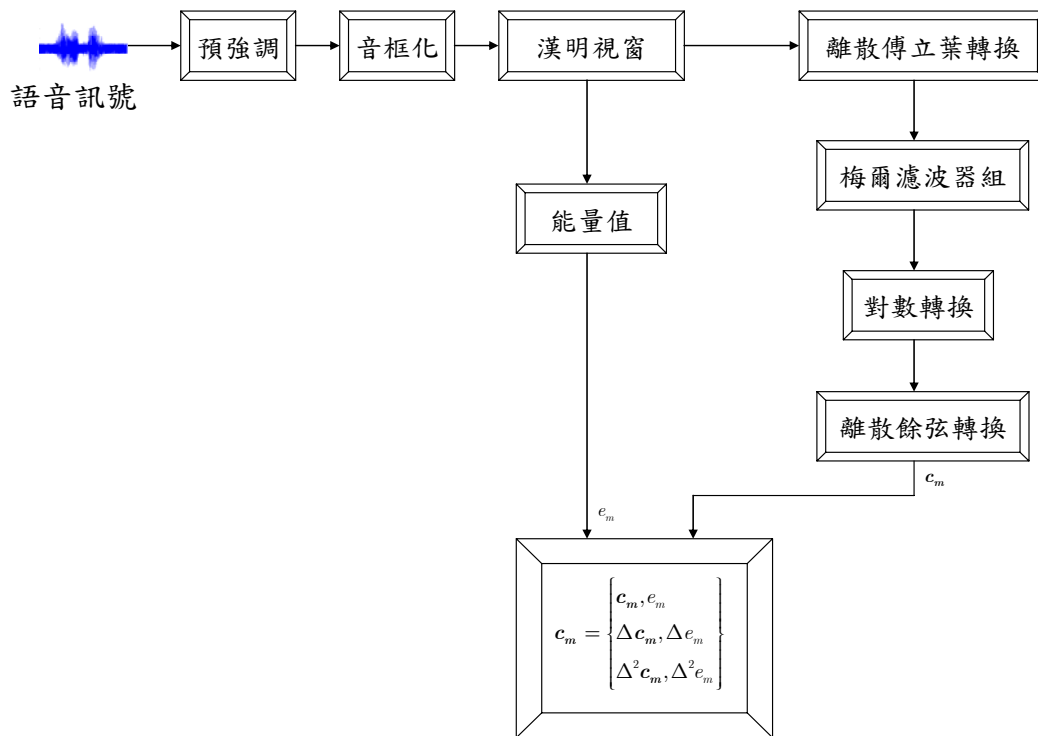
在之後的第二章裡，我們簡單介紹本論文所使用的四種語音特徵參數抽取流程。第三章介紹傳統之以段落為基礎特徵參數正規化法，第四章之討論即為本論文之重點：包括虛擬雙通道碼簿的建立方法，及三種以碼簿為基礎之特徵參數補償法。第五章與第六章分別為實驗環境介紹與實驗結果及討論。最後，第六章包含了簡要的結論。

## 二、各種語音訊號特徵參數抽取流程的介紹

本章節介紹在語音訊號處理中四種常用的語音特徵參數及其抽取流程，分別為梅爾倒頻譜係數(mel-frequency cepstral coefficients, MFCC)、自相關梅爾倒頻譜係數[3](autocorrelation mel-frequency cepstral coefficients, AMFCC)、線性預測倒頻譜係數[4][5](linear prediction cepstral coefficients, LPCC)以及感知線性預測倒頻譜係數[6](perceptual linear prediction cepstral coefficients, PLPCC)。我們將使用這四種語音特徵參數來驗證本論文所提出的強健性語音特徵參數技術，並且與其他強健性方法運用在這四種特徵參數上做比較。

### (一) 梅爾倒頻譜係數 (mel-frequency cepstral coefficients, MFCC)

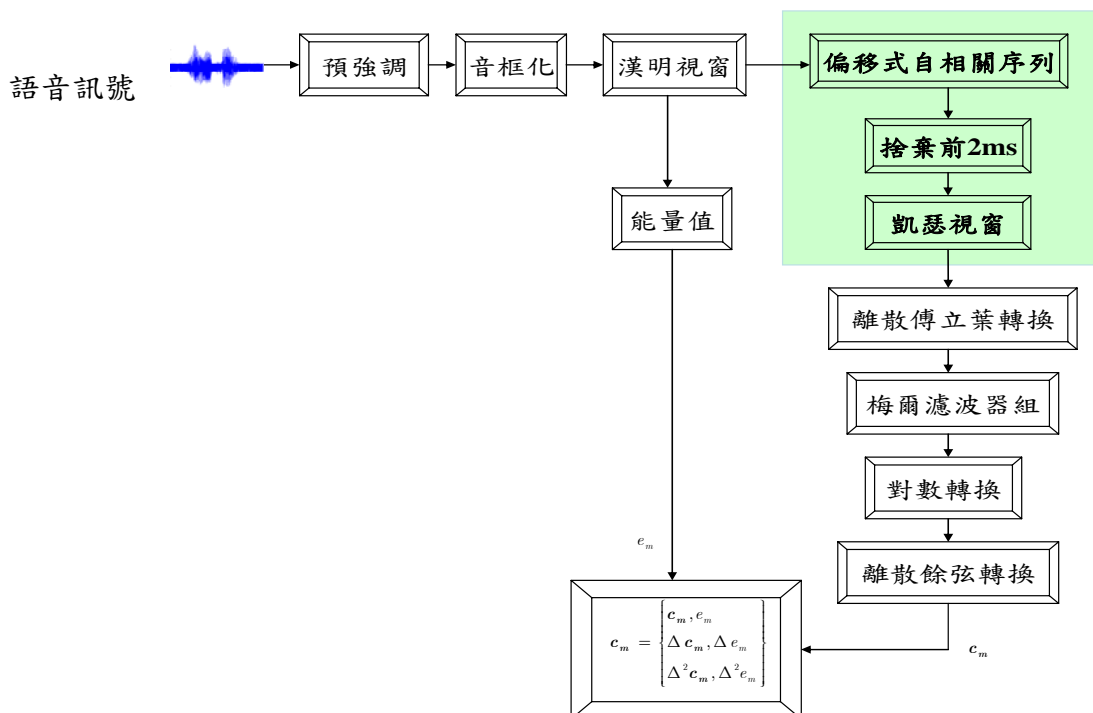
圖一為梅爾倒頻譜係數擷取流程圖，梅爾倒頻譜係數結合了人在發音上與聽覺上的諸多性質，是目前語音研究上，最常被使用的特徵參數。



圖一、梅爾倒頻譜特徵擷取流程圖

(二) 自相關梅爾倒頻譜係數 (autocorrelation mel-frequency cepstral coefficients, AMFCC)

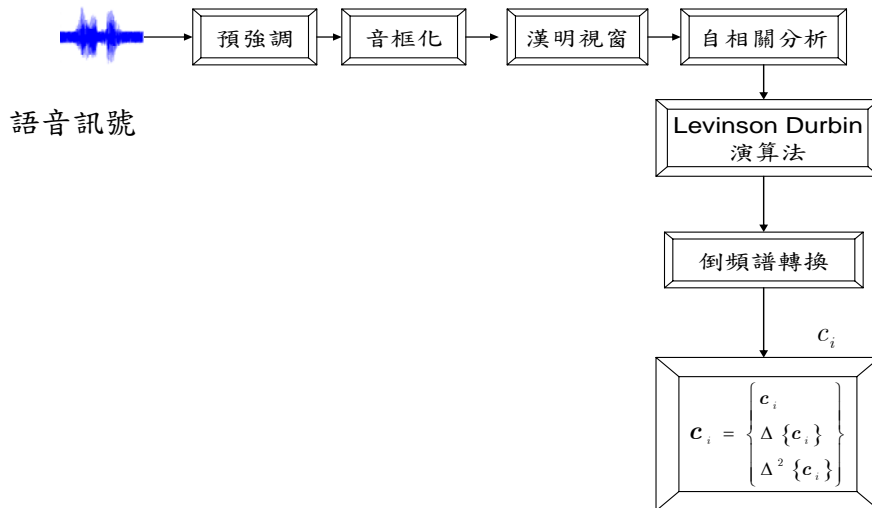
自相關梅爾倒頻譜係數是利用自相關序列結合梅爾倒頻譜係數求取步驟來做特徵參數抽取[1]，其流程即在一音框內的語音訊號經過漢明視窗處理後，取其偏移式自相關係數(biased autocorrelation coefficients)，並捨棄其前端約 2ms 係數後，再經過一凱瑟視窗以降低高頻效應。除了上述步驟外，其餘取頻譜、對數轉換及離散餘弦轉換等流程，皆與梅爾倒頻譜係數抽取流程相同。圖二為自相關梅爾倒頻譜係數之抽取流程。



圖二、自相關梅爾倒頻譜特徵擷取流程圖

(三) 線性預測倒頻譜係數(linear prediction cepstral coefficients, LPCC)

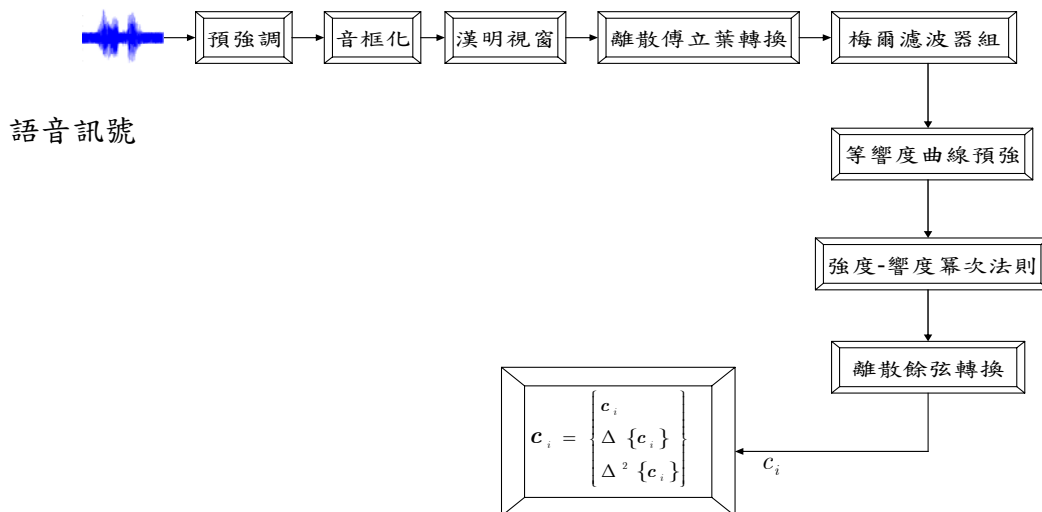
線性預測(linear prediction)的基本原理是假設目前的聲音取樣值可由在前面的  $p$  個取樣值，以線性組合來預測。圖三即為線性預測倒頻譜係數之擷取流程圖。如同前兩種特徵參數擷取技術，我們將語音訊號經過預強調後，切割成許多一小段的音框與漢明視窗的處理後取其自相關係數，透過 Levinson Durbin 演算法求得線性預測係數，最後將線性預測係數轉換成倒頻譜，便得到線性預測倒頻譜係數(linear prediction cepstral coefficients, LPCC)。



圖三、線性預測倒頻譜特徵擷取流程圖

(四) 感知線性預測倒頻譜係數(perceptual linear prediction cepstral coefficients, PLPCC)

感知線性預測倒頻譜係數的擷取流程圖如圖四，與線性預測係數倒頻譜擷取流程不同之處在於：(1)它經過模擬人耳的梅爾濾波器組，對於頻譜作頻率校準(frequency warping)的處理。(2)它利用等響度曲線(equal loudness curve)對強度頻譜做預強調。(3)對於經過預強調後的強度頻譜取三次方根(cubic root)，相當於對強度與響度之間做校準(intensity-loudness warping)的動作。這些改變都是針對人的聽覺特性而做的。



圖四、感知線性預測倒頻譜係數的擷取流程圖

### 三、整段式與分段式之強健性語音特徵等化技術

倒頻譜平均值與變異數正規化法(CMVN)是常被使用來強健語音特徵參數的方法之一，其作法是將一連串語音資料中的每一維倒頻譜特徵參數做統計量的調整，以便消除雜訊對語音的影響；在作法上，整段式的倒頻譜平均值與變異數正規化法[1](U-CMVN)是利用一整段語音特徵求取平均值與變異數，因此所用的音框長度隨特徵係數序列長短而異，而分段式倒頻譜平均值與變異數正規化法[2](S-CMVN)的作法，是將每一維語音特徵參數，以當時的音框為中心，對其前後數十個音框做分段統計量的計算，然後對當下的音框作正規化處理，以下將分別對這兩種方法詳加介紹。

#### (一) 整段式倒頻譜平均值與變異數正規化法(utterance-based cepstral mean and variance normalization, U-CMVN)

乾淨的語音訊號在經過加成性雜訊干擾後，其倒頻譜之平均值會和原本的乾淨語音倒頻譜的平均值之間會存在一個偏移量，而其變異數相對於乾淨語音特徵參數而言則會有壓縮性，因此會造成訓練與測試特徵的不匹配而降低辨識效果。而使用倒頻譜平均值與變異數等化法[1](CMVN)可將每一維倒頻譜特徵參數之平均值化為零，並將其變異數正規化為 1，這樣就能降低上述所謂的偏移量與壓縮性，進而提升倒頻譜參數的強健性。

整段式倒頻譜平均值與變異數等化法的作法如(式 3-1)，假設  $\{Y[n], n = 1, 2, \dots, N\}$  為一由語音資料擷取所得到的某一維倒頻譜特徵參數序列，而經過整段式倒頻譜平均值與變異數等化法處理後，得到新的特徵參數  $\{Y_{U-CMVN}[n], n = 1, 2, \dots, N\}$ ，其中的  $\{Y_{U-CMVN}[n], n = 1, 2, \dots, N\}$  平均值與標準差是經由整段語音的音框求取而得，如式(3-2)與式(3-3)。

$$Y_{U-CMVN}[n] = \frac{Y[n] - \mu_Y}{\sigma_Y}, \quad n = 1, 2, \dots, N \quad (\text{式 3-1})$$

其中

$$\mu_Y = \frac{1}{N} \sum_{n=1}^N Y[n] \quad (\text{式 3-2})$$

$$\sigma_Y = \sqrt{\frac{1}{N} \sum_{n=1}^N (Y[n] - \mu_Y)^2} \quad (\text{式 3-3})$$

#### (二) 分段式倒頻譜平均值與變異數正規化法(segmental cepstral mean and variance normalization, SCMVN)

如同整段式倒頻譜平均值與變異數正規化法，分段式倒頻譜平均值與變異數正規化法[2]目的亦是降低雜訊對語音的干擾，不同的是正規化時，其平均值與變異數是分段求得而非整段，如(式 3-4)、(式 3-5)與(式 3-6)。假設當第  $n$  個音框為當時的正規化之特徵參數，則其前  $P/2$  個音框與其後  $P/2$  個音框的參數皆用以求取統計值。也就是說，長度為  $P+1$  的視窗在一段語音特徵的時間軸上作橫移，視窗的中心點代表當時的音框，其前後的  $P/2$  個音框為其求取統計值之片段，執行平均值與變異數正規化。至於語句中前段的特徵向量中，由於特徵向量數目少於  $P/2$ ，所以求取統計值之長度為特徵參數的起始音框至該音框的後  $P/2$  音框；語句中後段之求

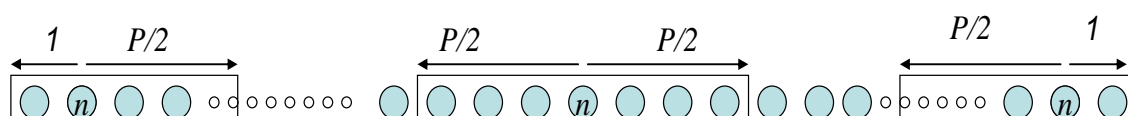
取統計值之長度亦同理可得，其表示法如圖五。

$$Y_{S-CMVN}[n] = \frac{Y[n] - \mu[n]}{\sigma[n]}, \quad n = 1, 2, \dots, N \quad (\text{式 3-4})$$

$$\text{其中} \quad \mu[n] = \frac{1}{P+1} \sum_{i=n-\frac{P}{2}}^{n+\frac{P}{2}} Y[i] \quad (\text{式 3-5})$$

$$\sigma[n] = \sqrt{\frac{1}{P+1} \sum_{i=n-\frac{P}{2}}^{n+\frac{P}{2}} (Y[i] - \mu[n])^2} \quad (\text{式 3-6})$$

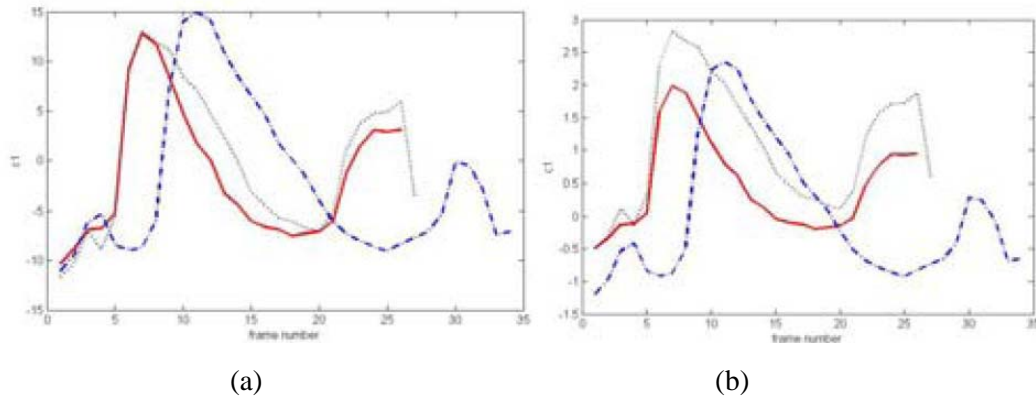
其中  $P+1$  為正規化片段的長度。



圖五、特徵參數經分段正規化視窗處理的示意圖

### (三) 討論

本章所介紹兩種語音特徵參數正規化法目的都是希望藉由正規化倒頻譜統計值，來降低訓練語音與測試語音之間的不匹配。其中，整段式倒頻譜平均值與變異數正規化法(U-CMVN)是以一整段語音的所有音框，去計算特徵參數的統計值，執行上比較簡單，不過此方法存在以下幾個缺點：(1)在語音特徵參數擷取流程中，在得到最後一個音框前，我們無法執行 U-CMVN。因為它是以一整段語調中所有特徵參數的音框，當做統計值的計算與進行正規化的基礎，所以 U-CMVN 是無法以線上方式(on-line manner)執行的。所謂線上方式(on-line manner)，即我們在擷取每一段倒頻譜特徵參數後，以即時(real-time)的方式計算出屬於該特徵參數之統計值，並執行正規化的動作。(2)一段語音的音框數通常是影響統計值正確性之因素，但我們無法控制原始一整段語音的長短。(3)因為不同聲學單位(acoustic units)的長度或總數量，在不同段語音之間會有變化，所以一段語音中同一個聲學單位被正規化後的特徵參數，在另一段語音中可能會不同。圖六(a)是從 AURORA2 中乾淨的訓練語料庫中，三段不同語音 "FAC\_1911446"，"FAC\_1473533A"與"FAC\_1O1"擷取出聲學單位"one"之原始第一維倒頻譜參數 c1 之輪廓；圖六(b)則是圖六(a)經過 U-CMVN 處理後的版本。從圖六(a)可發現未經 CMVN 處理的 c1，其輪廓在([-10.3, 12.8], [-11.6, 13.0]及[-11.0, 15.0])這三個範圍內有相似的分佈狀況；反觀圖六(b)，因為不同語音中的特徵參數是被不同的平均值與變異數作正規化，使正規化後三個 c1 的輪廓變得不太相同，它們的分佈狀況在([-0.5, 2.0], [-0.5, 2.8]及[-1.1, 2.3])，這三個範圍跟之前相比之下是比較不同的。



圖六：AURORA2 中乾淨的訓練語料庫中，三段不同語音"fac\_1911446"，"fac\_1473533A"與"fac\_101"擷取出聲學單位"one"之(a)原始第一維梅爾倒頻譜特徵 c1 輪廓(b)經 U-CMVN 處理後第一維梅爾倒頻譜特徵 c1 輪廓

另一方面，分段式倒頻譜平均值與變異數正規化法，它是以移動一固定長的分段視窗來計算正規化每個音框所要用的統計值，並作統計值正規化法。假如我們使得該分段視窗夠短的話，在執行上它是可以較接近於線上方式(on-line manner)的。再者，因為在一個短分段中的聲學單位總數目相對而言較少，所以使用分段式倒頻譜平均值與變異數正規化法，可降低相同聲學單位在不同段語音間的特徵參數變異性。而根據之後的實驗結果顯示，使用分段式倒頻譜平均值與變異數正規化法(S-CMVN)在本論文中所用之四種語音特徵參數上(MFCC、AMFCC、LPCC、PLPCC)，所得到的辨識效果，的確可比整段式倒頻譜平均值與變異數正規化法(U-CMVN)效果來得好。這也間接證明了，以分段式進行特徵參數統計值的估算，相對於整段式來的準確。而以前者估算出的統計值，進行特徵參數統計值正規化法所得之特徵參數也能更加降低雜訊對語音的影響。

#### 四、運用虛擬雙通道碼簿為基礎之雜訊強健技術的介紹

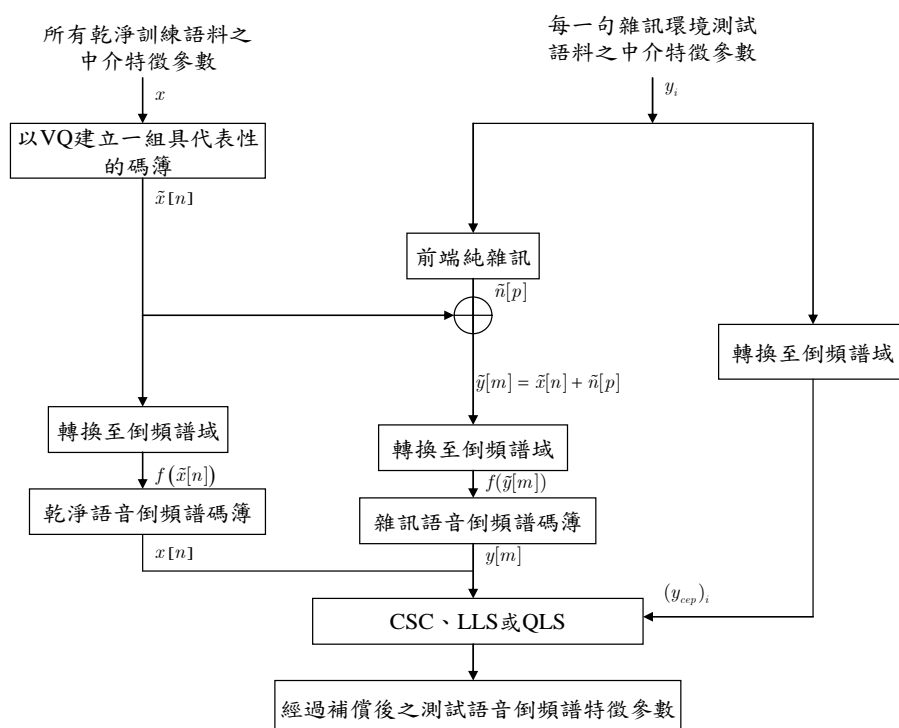
本論文所提出三種特徵參數補償法語音強健技術，依序為倒頻譜統計補償法[7](cepstral statistics compensation, CSC)、線性最小平方回歸法[7](linear least squares regression, LLS)以及二次方最小平方回歸法[7](quadratic least squares regression, QLS)。在執行這幾個特徵參數補償法前，我們先訓練兩組分別代表乾淨語音與雜訊語音的碼簿(codebooks)，我們稱之為虛擬雙通道碼簿(pseudo stereo codebooks)。藉由這兩組碼簿的使用，我們得以發展上述三種特徵參數補償法。

運用所謂的虛擬雙通道碼簿來計算乾淨語音與含雜訊語音之統計值，進而執行三種減低雜訊的技巧是簡單而又有效率的。倒頻譜統計補償法(CSC)、線性最小平方回歸法(LLS)以及二次方最小平方回歸法(QLS)，這三種特徵參數統計量補償法的概念就是對含雜訊之語音倒頻譜係數做轉換(transformation)，使得經過轉換後的語音倒頻譜其統計值更相似於乾淨訓練語音倒頻譜的統計值，進而提升語音辨識強健性。

##### (一) 虛擬雙通道碼簿之建立方法

本論文中四種語音特徵參數擷取流程的虛擬雙通道碼簿之建立方法，其步驟如圖七所示，首先

必須在擷取所有用以訓練之乾淨語音訊號的 MFCC、AMFCC、LPCC、PLPCC 不同的倒頻譜特徵參數中，保留下具有語音與雜訊為線性相加特性的中介特徵參數(intermediate feature)，將這些乾淨語音的中介特徵參數訓練成一組碼簿，接著將此乾淨語音碼簿中所有的碼字與測試語音中所選取的一段純雜訊之中介特徵參數作線性相加，如此便可得到兩組分別代表乾淨語音與測試語音在該中介特徵參數域中的碼簿，最後將這兩組中介特徵參數的碼簿轉到倒頻譜域中，以利之後特徵補償演算法的進行，此兩組分別代表乾淨語音與雜訊語音的倒頻譜特徵參數碼簿，我們稱之為虛擬雙通道碼簿 (pseudo stereo codebooks)。



圖七：虛擬雙通道碼簿之建立架構及以虛擬雙通道碼簿執行特徵參數補償法之流程圖

虛擬雙通道碼簿之建立過程詳述如下：

首先將語料庫中所有乾淨語料的每一段語音，透過特徵參數擷取流程轉換成一序列的中介特徵向量，如表一所述。這些由所有乾淨語料的語調所得到的中介特徵向量，透過向量量化 (vector quantization, VQ) 後可建立成一組包含  $N$  個碼字(codewords)的集合，以  $\{\tilde{x}[n], 1 \leq n \leq N\}$  表示。這組在中介特徵參數域上的乾淨語音碼簿中所有碼字，都可經過剩下的特徵參數擷取步驟轉換至倒頻譜域，如(式 4-1)所示：

$$\mathbf{x}[n] = f(\tilde{\mathbf{x}}[n]) \quad , \quad (式 4-1)$$



表一：本論文中所使用之四種語音特徵參數，及其具備語音與雜訊線性相加特性之中介特徵參數。

倒頻譜特徵參數型態	具備語音與雜訊線性相加之中介特徵參數
梅爾倒頻譜係數(MFCC)	梅爾頻譜(mel-spectrum)
自相關梅爾倒頻譜係數 (AMFCC)	梅爾頻譜(mel-spectrum)
線性預測倒頻譜係數 (LPCC)	自相關係數(autocorrelation coefficients)、強度頻譜 (magnitude spectrum)
感知線性預測倒頻譜係數 (PLPCC)	自相關係數(autocorrelation coefficients)、強度頻譜 (magnitude spectrum)

其中  $f(\cdot)$  為轉換函數，它是隨著我們所選擇的特徵參數型態而不同。因此  $\{\mathbf{x}[n], 1 \leq n \leq N\}$  這組經轉換至倒頻譜域的碼簿，即稱之為乾淨語音的倒頻譜碼簿。

至於含雜訊的測試語音方面，因為要完全以每段簡短的測試語音為基礎，去建立一組可靠的碼字是很困難的，所以我們試著藉由乾淨語音在中介特徵參數域上的碼字，來建立對應至該段的含雜訊之測試語音的碼簿。步驟如下：

對於一段測試語音，我們假設估計到的純雜訊在中介特徵參數域上可用一組向量來代表，以  $\{\tilde{\mathbf{n}}[p], 1 \leq p \leq P\}$  表示。因為乾淨語音與雜訊在中介特徵參數域上是近似線性相加的，因此含雜訊語音的碼字可表示成(式 4-2)：

$$\tilde{\mathbf{y}}[m] |_{m=(n-1)P+p} = \tilde{\mathbf{x}}[n] + \tilde{\mathbf{n}}[p] \quad (\text{式 4-2})$$

接著我們將  $\tilde{\mathbf{y}}[m]$  經過剩下的特徵參數擷取步驟轉換至倒頻譜域，如(式 4-3)所示：

$$\mathbf{y}[m] = f(\tilde{\mathbf{y}}[m]) \quad (\text{式 4-3})$$

因此， $\{\mathbf{y}[m], 1 \leq m \leq NP\}$  這組碼字即代表雜訊語音在倒頻譜域上的碼簿。 $\{\mathbf{x}[n]\}$  與  $\{\mathbf{y}[m]\}$  這兩組碼字可分別代表乾淨訓練語音與雜訊測試語音，我們稱之為虛擬雙通道碼簿(pseudo stereo codebooks)。所謂"虛擬"，即因為雜訊語音的碼簿並非直接由雜訊語音得到的，而是透過結合乾淨語音碼簿與雜訊估算值所得到的。值得注意的是，我們在建立乾淨語音碼簿  $\{\mathbf{x}[n]\}$  時是一次將語料庫中所有乾淨訓練語音做處理，這是屬於非線上方式(off-line manner)的處理。不過，當輸入每一段不同的測試語音，或雜訊環境改變時，雜訊語音碼簿  $\{\mathbf{y}[m]\}$  必須隨之更新。因為雜訊估算值  $\tilde{\mathbf{n}}[p]$  可粗略地以每一段測試語音的前幾個音框得到，因此雜訊語音的碼簿  $\{\mathbf{y}[m]\}$  可以在一個幾乎為線上運算方式(on-line manner)，即在不會有太長的延遲時間的運算情況下建立。

在本論文中，我們以虛擬雙通道碼簿為基礎來執行三種特徵參數補償法，以降低加成性雜訊的影響。以下，我們對三種特徵參數補償法做完整的介紹。

## (二) 倒頻譜統計補償法(cepstral statistics compensation, CSC)

我們利用虛擬雙通道碼簿，可以算出分別代表乾淨語音與雜訊語音的統計值，如(式 4-4)、(式 4-5)所示：

$$\mu_{x,i} \approx \frac{1}{N} \sum_{n=1}^N (\mathbf{x}[n])_i, \sigma_{x,i}^2 \approx \frac{1}{N} \sum_{n=1}^N [(\mathbf{x}[n])_i - \mu_{x,i}]^2 \quad (\text{式 4-4})$$

$$\mu_{y,i} \approx \frac{1}{NP} \sum_{m=1}^{NP} (\mathbf{y}[m])_i, \sigma_{y,i}^2 \approx \frac{1}{NP} \sum_{m=1}^{NP} [(\mathbf{y}[m])_i - \mu_{y,i}]^2 \quad (\text{式 4-5})$$

其中  $(\mathbf{v})_i$  代表一個任意向量  $\mathbf{v}$  第  $i$  維成份， $\mu_{x,i}$  與  $\sigma_{x,i}^2$  分別代表乾淨語音特徵向量  $\mathbf{x}$  第  $i$  維的平均值與變異數； $\mu_{y,i}$  與  $\sigma_{y,i}^2$  分別代表雜訊語音特徵向量  $\mathbf{y}$  第  $i$  維的平均值與變異數。以這些統計值來執行倒頻譜統計值補償法，我們轉換每一段雜訊語音之倒頻譜向量，如(式 4-6)：

$$(\mathbf{z})_i = \frac{\sigma_{x,i}}{\sigma_{y,i}} \times [(\mathbf{y})_i - \mu_{y,i}] + \mu_{x,i} \quad (\text{式 4-6})$$

在理想的情況下， $(\mathbf{z})_i$  與乾淨語音特徵向量  $(\mathbf{x})_i$  會有相同的平均值與變異量，由於雜訊語音倒頻譜的某些統計值被補償，使得補償過後的雜訊語音倒頻譜其統計值是近似於乾淨語音倒頻譜的統計值，因此我們將此方法稱為倒頻譜統計補償法(cepstral statistics compensation, CSC)。我們可以用矩陣的形式改寫(式 4-6)的倒頻譜統計補償演算法，如(式 4-7)：

$$\mathbf{z} = \Psi(\mathbf{y} - \boldsymbol{\mu}_y) + \boldsymbol{\mu}_x \quad (\text{式 4-7})$$

其中  $\boldsymbol{\mu}_x = [\mu_{x,1}, \mu_{x,2}, \dots]^T$ ,  $\boldsymbol{\mu}_y = [\mu_{y,1}, \mu_{y,2}, \dots]^T$ ,  $\Psi$  是一對角線為  $\{\sigma_{x,i} / \sigma_{y,i}\}$  之對角矩陣(diagonal matrix)。

事實上，CSC 的概念是類似傳統的倒頻譜平均與變異數正規化法(CMVN)，因為這兩種演算法的目的都是希望訓練語音與測試語音能得到相似的統計值。不過 CSC 擁有下列幾項優點：  
(1) CSC 可以一個幾乎為線上方式(on-line manner)的處理程序來執行，因為乾淨語音的碼簿是事先建立好的，而在建立雜訊語音碼簿時所需的雜訊估算值，通常可以在每段測試語音的前幾個音框來得到。

(2) 在 CSC 中，統計量是利用訓練語料庫中所有訓練語音所建立的碼簿所得；但在 CMVN 中，只利用單一語句去決定平均值與變異量。因此，我們可以預期碼簿幫助我們求得更準確的特徵參數統計值。

(3) 在 CSC 中，相同雜訊環境下不同語句的特徵參數接受相同的轉換，這使得不同語句之間的特徵參數，在對應至相同的聲學單位時，能保持特徵相似度。這是 CMVN 無法做到的，我們在上一章的圖六已做了說明。

(三) 線性最小平方回歸法(linear least squares regression, LLS)與 二次最小平方回歸法(quadratic least squares regression, QLS)

在這裡，線性最小平方回歸法與二次最小平方回歸法都是屬於多項式回歸法(polynomial regression approaches)，其概念就是希望雜訊語音的碼簿，在透過一個轉換函數的運算後能和乾淨語音碼簿的整體距離是最小的，如此我們便可預期，當雜訊語音倒頻譜經過相同轉換後，會更接近乾淨語音倒頻譜。以下我們做詳細的介紹。

在前面的介紹中，我們知道每個雜訊語音碼字  $\mathbf{y}[m]$  對應的乾淨語音碼字為  $\mathbf{x}[n]$ ，其中  $n = \lceil m/P \rceil$  ( $\lceil \cdot \rceil$  表示無條件進位運算， $P$  為純雜訊的向量數目)， $\{\mathbf{x}[n]\}$  與  $\{\mathbf{y}[m]\}$  這兩組碼字分別代表乾淨語音與雜訊語音倒頻譜  $\mathbf{x}$  與  $\mathbf{y}$ 。若我們能對每一個雜訊語音碼字  $\mathbf{y}[m]$  找到一個轉換函數  $\mathcal{T}(\cdot)$ ，使得  $\mathcal{T}(\mathbf{y}[m])$  與  $\mathbf{x}[n]$  之間的整體距離是最小的，那我們可以合理的預期雜訊語音倒頻譜  $\mathbf{y}$  經轉換後  $\mathcal{T}(\mathbf{y})$ ，會更接近乾淨語音倒頻譜  $\mathbf{x}$ 。為了簡單起見，我們假設轉移函數是執行在  $\mathbf{y}$  的每一維上。假設  $\mathcal{T}_i(\bullet)$  是  $\mathbf{y}$  的第  $i$  維成份的轉移函數，則定義一目標函數  $J_i$  將使得  $\mathcal{T}_i((\mathbf{y}[m])_i)$  與

$(\mathbf{x}[n])_i$  的整體平方距離最小，如(式 4-8)：

$$J_i = \sum_{m=1}^{NP} [\mathcal{T}_i((\mathbf{y}[m])_i) - (\mathbf{x}[n])_i]^2 \quad (\text{式 4-8})$$

其中  $n = \lfloor m/P \rfloor$ ，假設  $\mathcal{T}_i(\bullet)$  是一個  $K$  次多項式，則(式 4-8)中以處理  $\mathcal{T}_i(\bullet)$  來最小化  $J_i$ ，就變成一個典型的最小化平方(least squares)的問題，如(式 4-9)：

$$\mathcal{T}_i(u) = a_K^{(i)}u^K + a_{K-1}^{(i)}u^{K-1} + \dots + a_0^{(i)} \quad (\text{式 4-9})$$

(式 4-8)中的目標函數可以改寫成向量矩陣的形式，如(式 4-10)：

$$J_i = \|\mathbf{Y}_i \mathbf{a}_i - \mathbf{b}_i\|^2 \quad (\text{式 4-10})$$

其中矩陣  $\mathbf{Y}_i$  的第  $(m,n)$  項如(式 4-11)所示：

$$(\mathbf{Y}_i)_{mn} = [(\mathbf{y}[m])_i]^{K-n+1}, 1 \leq m \leq NP, 1 \leq n \leq K+1 \quad (\text{式 4-11})$$

且  $\mathbf{a}_i = [a_K^{(i)} \quad a_{K-1}^{(i)} \dots a_0^{(i)}]^T$ ，

$$\mathbf{b}_i = \left[ \left( \mathbf{x}[\lfloor 1/P \rfloor] \right)_i \left( \mathbf{x}[\lfloor 2/P \rfloor] \right)_i \dots \left( \mathbf{x}[\lfloor NP/P \rfloor] \right)_i \right]^T。$$

多項式  $\mathcal{T}_i(\bullet)$  中最小化  $J_i$  的係數向量  $\mathbf{a}_i$  即為最小平方解，如下(式 4-12)：

$$\hat{\mathbf{a}}_i = (\mathbf{Y}_i^T \mathbf{Y}_i)^{-1} \mathbf{Y}_i^T \mathbf{b}_i \quad (\text{式 4-12})$$

值得注意的是，多項式  $\mathcal{T}_i(\bullet)$  的次數  $K$  不可以設太大，以避免有過度擬合(over-fitting)情況或不良狀況的矩陣(ill-conditional matrix)  $\mathbf{Y}_i^T \mathbf{Y}_i$  產生。因此，我們只考慮  $K=1$  與  $K=2$  兩種情況：當  $K=1$  時，轉移函數  $\mathcal{T}_i(\bullet)$  是一個線性函數，我們稱之為線性最小平方回歸法(linear least squares regression, LLS)。當  $K=2$  時轉移函數  $\mathcal{T}_i(\bullet)$  即為一個二次函數，我們稱之為二次最小平方回歸法(quadratic least squares regression, QLS)。

如同本節一開始所提到，用這兩種多項式回歸法的概念就是希望雜訊語音的碼簿，在透過一個轉換函數的運算後能和乾淨語音碼簿的整體距離是最小的，當雜訊語音倒頻譜經過相同轉換後會更接近乾淨語音倒頻譜，如此便可提升辨識效果。

## 五、實驗設定

### (一) 語音資料庫簡介

本論文所使用的語音資料庫為歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)發行的 AURORA2 語音資料庫，它是一套連續的英文數字字串，內容是以美國成年男女所錄製的乾淨環境連續數字，再加上雜訊與通道效應。加性雜訊共有八種，分別為地下鐵、人聲、汽車、展覽館、餐廳、街道、機場、火車站等，前四種歸類為 Set A，後四種歸類為 Set B。

訊雜比(signal-to-noise ratio, SNR)則有七種，分別為 20dB, 15dB, 10dB, 5dB, 0dB, -5dB 與完全乾淨狀態。

### (三) 特徵參數的設定與辨識系統的訓練

本論文共使用四種特徵參數分別為梅爾倒頻譜係數(MFCC)、自相關梅爾倒頻譜係數

(AMFCC)、線性預測倒頻譜係數(LPCC)及感知線性預測倒頻譜係數(PLPCC)，其相關設定與個別對應之中介特徵參數，如表二所示。對於每個欲辨識的數字模型而言，本論文使用隱藏式馬可夫模型工具(hidden Markov model toolkit, HTK)來訓練，包含 11 個數字模型(0~9 以及 oh 11 個數字模型)以及靜音模型，每個數字模型包含 10 個狀態，各狀態包含 4 個高斯密度混合。隱藏式馬可夫模型是一種運用統計理論推導出來的模型，用來描述語音產生的過程，相當適合用在連續語音的辨認。HMM 有很多種類型，本論文採用由左到右的形式，也就是每個狀態在下一個時間只能跳到此刻狀態或下一個鄰近的狀態，隨著時間的增加，狀態由左至右依序轉移。另外，模型中的狀態觀測機率函數是選用連續式的高斯混合機率密度函數(Gaussian Mixture probability density function, 簡稱 GM)，因此我們也稱此模型為連續密度隱藏式馬可夫模型(continuous density HMM, 簡稱 CDHMM)。

表二、實驗中所用的特徵參數詳細資料

特徵參數種類	特徵參數維度	中介特徵參數維度
MFCC	12 維倒頻譜加上 1 對數能量維，並取其一階和二階差量，總共 39 維特徵參數。	23 維梅爾頻譜加上 1 對數能量維。
AMFCC	12 維倒頻譜加上 1 對數能量維，並取其一階和二階差量，總共 39 維特徵參數。	23 維梅爾頻譜加上 1 對數能量維。
LPCC	13 維倒頻譜，並取其一階和二階差量，總共 39 維特徵參數。	23 維強度頻譜，或是 24 維自相關係數。
PLPCC	13 維倒頻譜，並取其一階和二階差量，總共 39 維特徵參數。	23 維強度頻譜，或是 23 維自相關係數。

#### (四) 強健性特徵參數技術實驗設定

在分段式平均值與變異數正規化法(S-CMVN)，我們令式 3-5 與式 3-6 中使用的分段長度為  $P+1 = 101$  個音框，即大約為 1 秒的長度。

虛擬雙通道碼簿的建立方法當中，乾淨語音碼簿為  $\{\tilde{x}[n], 1 \leq n \leq N\}$ ，其中  $N$  值我們分別設為 32、64、128、256、512、1024。在實驗結果中，我們將只呈現得到最佳辨識率時的  $N$  值之整體實驗數據。

對於純雜訊的估測值  $\{\tilde{n}[p]\}$ ，我們是以在中介特徵參數域上，每一段測試語音的前 5 個音框當作該段語音的純雜訊音框。

在 LPCC 與 PLPCC 兩種特徵參數擷取過程裡，因為其具備語音與雜訊為線性相加的中介特徵參數有兩種，分別為強度頻譜(magnitude spectrum)與自相關係數(autocorrelation coefficients)，因此在實驗結果中我們以 MS-與 AC-分別代表之。

## 六、實驗結果與分析

首先，表三與表四分別為整段式倒頻譜平均與變異數正規化法(U-CMVN)與分段式倒頻譜

平均與變異數正規化法(S-CMVN)的辨識精確度，相對於原始未處理的各種倒頻譜特徵而言，U-CMVN 與 S-CMVN 皆能有效提升各種雜訊環境下的辨識率，這意謂這兩種方法的確具有提升特徵參數強健性的效能，而當我們將表四與表三的數據比較，可明顯看出 S-CMVN 相對於 U-CMVN 在辨識率能有更明顯的提升。這與我們之前分析的結果相吻合，即利用分段的方式估測特徵參數的統計值能比利用整段的方式更精確。

接下來，我們探討本論文所提出的三種以碼簿為基礎的特徵參數補償法的效果，表五、表六與表七分別為倒頻譜統計補償法(CSC)、線性最小平方回歸法(LLS)與二次最小平方回歸法(QLS)的辨識精確度。為了比較起見，我們將表四之 S-CMVN 的結果亦列於各表中。從這三個表的數據可知：

表三、整段式倒頻譜平均與變異數正規化法之辨識精確度(%)

Set A	subway	babble	car	exhibition	average	baseline
MFCC	72.91	69.71	68.71	69.22	70.14	61.99
AMFCC	68.60	72.02	68.94	65.58	68.78	65.52
LPCC	72.25	71.32	71.50	70.18	71.31	51.26
PLPCC	75.24	74.34	73.72	74.60	74.48	57.38
Set B	restaurant	street	airport	train station	average	baseline
MFCC	71.60	72.16	71.00	68.28	70.76	55.78
AMFCC	72.89	71.23	73.35	70.23	71.93	59.43
LPCC	73.44	72.82	74.68	70.69	72.91	49.58
PLPCC	76.48	75.79	77.01	73.23	75.63	54.51

表四、分段式倒頻譜平均與變異數正規化法之辨識精確度(%)

Set A	subway	babble	car	exhibition	average	U-CMVN
MFCC	75.71	73.42	72.36	72.63	73.53	70.14
AMFCC	72.12	74.82	73.49	70.21	72.66	68.78
LPCC	74.53	74.20	75.30	74.38	74.60	71.31
PLPCC	76.07	75.14	75.72	76.06	75.75	74.48
Set B	restaurant	street	airport	train station	average	U-CMVN
MFCC	75.65	75.23	74.97	71.93	74.45	70.76
AMFCC	75.58	76.11	75.02	72.14	74.71	71.93
LPCC	76.08	76.63	76.87	73.84	75.86	72.91
PLPCC	77.69	77.48	78.21	75.13	77.13	75.63

1. 相對於原始未處理的倒頻譜參數（數據列於表三）而言，這三種新的特徵參數補償法都能夠大幅提昇辨識精確度，意謂各種不同的特徵參數都能藉由這三種方法而提升其強健性。
2. 在大部分的情形下，這三種新的特徵參數補償法的表現都優於 S-CMVN 與 U-CMVN，這呼應了我們之前的推論：利用碼簿來估測特徵參數的統計值相較於利用整段或分段的方式估測更來的精確。
3. 雖然這三種特徵參數補償法作用於四種特徵參數上的實驗結果，所得到最佳辨識率時的 N 值都不相同，不過若個別觀察個特徵參數的實驗結果，可發現其具有規則性。如在特徵參數為 MFCC 時，三種特徵參數補償法之最佳實驗結果都在 N 值為 512 或 256 這些比較中段

的值；而特徵參數為 LPCC 時則在 N 值為較大值 1024 時，可得到最佳辨識率。

4. 一般而言，倒頻譜統計補償法的效果優於線性最小平方回歸法與二次最小平方回歸法，然而，其表現的差異並沒有十分明顯。

表五、倒頻譜統計補償法(CSC)之辨識精確度(%)，其中 N 表示乾淨碼簿的碼字數

Set A	subway	babble	car	exhibition	average	S-CMVN
MFCC (N=512)	78.71	75.84	80.54	77.40	78.12	73.53
AMFCC (N=512)	79.11	74.26	82.80	75.91	78.02	72.66
MS-LPCC (N=1024)	75.81	72.31	82.38	74.62	76.28	74.60
AC-LPCC (N=1024)	74.94	75.71	81.19	74.54	76.60	
MS-PLPCC (N=128)	77.79	76.14	81.25	78.71	78.47	75.75
AC-PLPCC (N=32)	78.64	76.57	78.55	77.70	77.87	
Set B	restaurant	street	airport	train station	average	S-CMVN
MFCC (N=512)	75.08	77.82	77.15	77.77	76.95	74.45
AMFCC (N=512)	73.15	79.36	77.28	79.10	77.22	74.71
MS-LPCC (N=1024)	73.57	76.55	78.48	79.16	76.93	75.86
AC-LPCC (N=1024)	76.35	75.95	80.24	79.61	78.04	
MS-PLPCC (N=128)	77.24	77.47	80.40	79.55	78.66	77.13
AC-PLPCC (N=32)	76.76	77.06	79.90	78.07	77.95	

表六、線性最小平方回歸法(LLS)之辨識精確度(%)，其中 N 表示乾淨碼簿的碼字數

Set A	subway	babble	car	exhibition	average	S-CMVN
MFCC (N=512)	78.92	76.09	80.01	76.57	77.90	73.53
AMFCC (N=512)	76.97	75.06	80.35	73.46	76.46	72.66
MS-LPCC (N=1024)	72.85	75.23	79.79	71.18	74.77	74.60
AC-LPCC (N=1024)	71.70	77.48	78.22	70.91	74.58	
MS-PLPCC (N=64)	79.74	77.70	81.37	79.22	79.51	75.75
AC-PLPCC (N=64)	76.63	77.15	76.02	75.93	76.43	
Set B	restaurant	street	airport	train station	average	S-CMVN
MFCC (N=512)	78.89	78.19	77.72	77.62	77.35	74.45
AMFCC (N=512)	74.45	77.55	77.29	76.95	76.56	74.71
MS-LPCC (N=1024)	75.29	74.91	79.42	77.97	76.90	75.86
AC-LPCC (N=1024)	77.82	74.32	80.66	78.12	77.73	
MS-PLPCC (N=64)	78.71	78.91	81.17	80.44	79.81	77.13
AC-PLPCC (N=64)	77.79	75.56	79.40	76.69	77.36	

## 七、結論與未來展望

本論文提出三種以虛擬雙通道碼簿為基礎的特徵參數補償法，分別為倒頻譜統計補償法(CSC)、線性最小平方回歸法(LLS)與二次最小平方回歸法(QLS)，個別作用於四種語音特徵參數：梅爾倒頻譜係數(MFCC)、自相關梅爾倒頻譜係數(AMFCC)、線性預測倒頻譜係數(LPCC) 與感知線性預測倒頻譜係數(PLPCC)上。我們發現，以虛擬雙通道碼簿為基礎之特徵參數補償法，

表七、二次最小平方回歸法(QLS)之辨識精確度(%), 其中 N 表示乾淨碼簿的碼字數

<b>Set A</b>	subway	babble	car	exhibition	average	S-CMVN
<b>MFCC (N=256)</b>	77.71	76.44	82.60	77.38	<b>78.53</b>	<b>73.53</b>
<b>AMFCC (N=512)</b>	72.61	76.94	81.08	71.02	<b>75.41</b>	<b>72.66</b>
<b>MS-LPCC (N=1024)</b>	69.82	74.99	80.80	71.04	<b>74.16</b>	<b>74.60</b>
<b>AC-LPCC (N=1024)</b>	67.35	75.45	77.81	69.54	<b>72.54*</b>	
<b>MS-PLPCC (N=64)</b>	76.83	76.74	82.80	77.94	<b>78.58</b>	<b>75.75</b>
<b>AC-PLPCC (N=512)</b>	71.48	73.05	74.29	72.48	<b>72.82*</b>	
<b>Set B</b>	restaurant	street	airport	train station	average	S-CMVN
<b>MFCC (N=256)</b>	73.88	77.19	77.70	79.18	<b>76.99</b>	<b>74.45</b>
<b>AMFCC(N=512)</b>	74.47	75.50	78.95	78.72	<b>76.91</b>	<b>74.71</b>
<b>MS-LPCC (N=1024)</b>	74.75	72.06	79.65	78.55	<b>76.25</b>	<b>75.86</b>
<b>AC-LPCC (N=1024)</b>	75.28	70.17	78.92	76.97	<b>75.34*</b>	
<b>MS-PLPCC (N=64)</b>	76.53	76.60	81.22	81.19	<b>78.89</b>	<b>77.13</b>
<b>AC-PLPCC (N=512)</b>	72.61	70.62	76.19	74.24	<b>73.41*</b>	

以線上方式即時地估算出雜訊語音特徵參數統計值，所估算出的統計值較為準確，也使得執行特徵參數補償法後語音特徵參數更為強健。相對於傳統特徵參數正規化法是以整段或分段語句為基礎去估算語音特徵參數的統計值後，而執行特徵參數正規化，以虛擬雙通道碼簿為基礎的特徵參數補償法，更能降低雜訊對語音的影響。

本論文只著重於加成性雜訊環境下的研究，因此在未來，我們期望能以虛擬碼簿為基礎的強健性語音技術，藉由結合一些通道補償技巧如：相對頻譜法(RASTA) [8]，使這些虛擬碼簿為基礎的強健性語音技術能延伸於消除通道失真的效應上。

## 八、參考文獻

- [1] S. Tiberwala and H. Hermansky, "Multiband and adaptation approaches to robust speech recognition", Eurospeech97, 1997, pp. 107-110
- [2] O. Viikki and K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization", in ESCA NATO Workshop Robust Speech Recognition Unknown Communication Channels, Pont-a-Mousson, France, 1997, pp.107-110.
- [3] Benjamin J. Shannon, Kuldip K. Paliwal, "Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition", Speech Communication 2006.
- [4] Atal, B.S. "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", Journal of the Acoustical Society of America, 1974.
- [5] J. Makhoul, "Spectral linear prediction: properties and applications," IEEE Transactions on Acoustics, Speech and Signal Processing, 1975.
- [6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J. Acoust. Soc. Am, vol. 87, no. 4, pp. 1738-1752, Apr. 1990.
- [7] Jieh-weih Hung, "Cepstral statistics compensation using online pseudo stereo codebooks for robust speech recognition in additive noise environments", ICASSP 2006.
- [8] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Transactions on Speech and Audio Processing, 2, pp.578-589, 1994