併合式倒頻譜統計正規化技術於強健性語音辨識之研究
A Study of Hybrid-based Cepstral Statistics Normalization Techniques for Robust Speech Recognition

何冠旻 Guan-min He
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
s96323528@ncnu.edu.tw

杜文祥 Wen-Hsiang Tu
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
aero3016@ms45.hinet.net

洪志偉 Jeih-weih Hung
國立暨南國際大學電機系
Dept of Electrical Engineering, National Chi Nan University
Taiwan, Republic of China
jwhung@ncnu.edu.tw

摘要

一語音辨識系統，在雜訊干擾的環境下，其辨識效能通常會明顯下降，如何改善此問題，是歷年來許多語音處理領域之學者所研究的重點。本論文也是針對此問題，提出了幾種新的語音強健性技術，來降低雜訊的干擾，以提升語音辨識的效能。

在本論文中，我們提出了新的語音特徵統計估測資訊演算法，藉此改進五種有名的強健性語音特徵正規化技術的效能，這些正規化技術包括了倒頻譜平均消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)、高階倒頻譜動差正規化法(HOCMN)、倒頻譜增益正規化法(CGN)以及倒頻譜統計圖等化法(HEQ)等，這些技術皆被證明有效提升語音特徵之強健性。這些方法中的關鍵步驟之一，爲特徵統計資訊的估測。在傳統上，有三種統計估測的演算法，分別爲整句式、分段式與碼簿式演算法。在此論文中，我們討論這三種估測方式可能的優缺點，進而提出新的估測方式，稱作併合式統計估測演算法，其適當地組合碼簿式與整句式(或分段式)統計值估測法所求得的特徵統計資訊。在一系列之雜訊環境下的語音辨識實驗中，我們驗證了新提出的併合式統計估測法相對於傳統三種估測法而言，能夠更有效地改進上述五種語音特徵正規化技術的效能，而能得到更明顯的辨識精確率提昇。此外，我們所提出的併合碼簿與分段式的統計估測法具有近似線上運算的功能，因此更具有實際應用之價值。

Abstract

. Cepstral statistics normalization techniques have been shown to be very successful at improving the noise robustness of speech features. In this paper, we propose a hybrid-based scheme to achieve a more accurate estimate of the statistical information of features in these techniques. By properly integrating codebook and utterance/segment knowledge, the

resulting hybrid-based normalization methods significantly outperform conventional utterance-based, segment-based and codebook-based ones in recognition accuracy.

For the Aurora-2 clean-condition training task, the proposed hybrid codebook/segment-based histogram equalization (CS-HEQ) achieves an average recognition accuracy of 90.66%, which is better than utterance-based HEQ (87.62%), segment-based HEQ (85.92%) and codebook-based HEQ (85.29%). Furthermore, the high-performance CS-HEQ can be implemented with a short delay and can thus be applied in real-time online systems. A similar performance promotion can be also found in the methods of hybrid-based cepstral mean subtraction (CMS), cepstral mean and variance normalization (CMVN), cepstral gain normalization (CGN) and higher-order cepstral moment normalization (HOCMN).

關鍵詞：語音辨識、碼簿、特徵統計值估測法、強健性語音特徵參數

Keywords: speech recognition, codebook, feature statistics estimate, robust speech features

一、緒論

一語音辨識系統，當其應用於真實環境時，常因環境中諸多無法預期的變異性(variation)，而使其辨識效能受到明顯影響，爲了降低諸多的變異性所發展的各種技術，一般而言統稱爲強健性技術(robustness techniques)，而本論文中，我們則是主要著重於發展降低環境之雜訊干擾或通道效應的強健性技術。

在諸多降低錄音環境之雜訊干擾的強健性演算法中，有一大類的方法是將訓練與測試環境下的語音特徵其時間序列統計特性加以正規化(normalization)，以降低訓練與測試環境之間的不匹配，達到提昇辨識率的目的。在這些演算法中，首要步驟通常是估測語音特徵的統計值相關資訊，例如在 CMVN 法中所需估測的統計值爲平均值(mean)與變異數(variance)，而在 HEQ 法中必需估測出特徵時間序列的機率分佈(probability distribution)。這些統計估測值的精確度，直接影響到其對應之正規化演算法的效能。

在過去關於上述特徵統計正規化法的文獻中，根據不同的樣本來源，大致上有三種統計值估測法，分別爲整句式、片段式與碼簿式的估測法，顧名思義，第一種直接使用了整句的語音特徵來估測統計值，第二種則使用了部分(片段)的語音特徵，而第三種則間接透過語音特徵建立的碼簿[7]來作統計值之估測。我們發現這三種方法各有其優缺點，因此在本論文中，我們所提出的新統計估測技術，適當地併合碼簿與整句或片段的特徵資訊，希望得到更精準的語音特徵統計值，進而使各種特徵統計正規化法，在受雜訊干擾的環境中能夠更有效地提昇語音特徵的強健性，以改善辨認精確度。

本論文其餘的章節概要如下：在第二章，我們將簡要介紹過去三種特徵統計值估測法之步驟及其可能的優缺點。第三章則介紹我們新提出的兩種併合式(hybrid-based)的統計值估測法，及其如何運用於各種特徵統計正規化法中。在第四章中，我們介紹語音辨識實驗之語音資料庫、及新提出的兩種統計估測法在各種特徵統計正規化法的語音辨識結果及其相關討論。最後，第五章爲一簡要結論及未來研究之展望。

二、整句式、分段式與碼簿式特徵統計值估測法

我們在本論文中所討論的五種著名強健性語音特徵正規化技術，分別爲倒頻譜平均消去法(CMS)[1]、倒頻譜平均值與變異數正規化法(CMVN)[2,3]、高階倒頻譜動差正規

化法(HOCMN)[4]、倒頻譜增益正規化法(CGN)[5]以及倒頻譜統計圖等化法(HEQ)[6]等，這些技術所需使用的特徵統計相關資訊，例如：平均值、變異數、高階動差或是機率分佈等，可由不同的方法估測，而有不同的效果。在本章中，我們將介紹過去學者所提之主要三種特徵統計值估測法，包括了整句式(utterance-based)[8]、分段式(segment-based)[8]與碼簿式(codebook- based)[9]三類方法，及它們可能的優點與缺點。

(一) 整句式特徵統計值估測法

假設某單一語句之某一維特徵序列表示爲

$$\{x[n];1 \le n \le N\} \qquad \text{(式 2-1)}$$

其中 N 爲特徵序列之特徵總個數(即音框總數)。在整句式特徵統計值估測法裡，我們利用(式 2-1)所列之單句所有特徵，共同估測第 m 項特徵 $x[m]$ 的統計值。換言之，我們假設 $x[m]$ 對應至一隨機變數 $X[m]$，進而假設整句特徵序列 $\{x[n];1 \le n \le N\}$ 爲此隨機變數之樣本(sample)，根據這些樣本，我們可估測出 $X[m]$ 此隨機變數的各種統計值，例如：

1. $X[m]$ 的期望值(平均值)爲

$$\mu_{X[m],(u)}[m] = \frac{1}{N}\sum_{n=1}^{N} x[n], \qquad \text{(式 2-2)}$$

2. $X[m]$ 的變異數(variance)爲

$$\sigma^2_{X[m],(u)}[m] = \frac{1}{N}\sum_{n=1}^{N}\left(x[n] - \mu_{X[m],(u)}[n]\right)^2, \qquad \text{(式 2-3)}$$

3. $X[m]$ 的第 J 階中央動差(central moment)爲

$$\xi^{(J)}_{X[m],(u)}[m] = \frac{1}{N}\sum_{n=1}^{N}\left(x[n] - \mu_{X[m],(u)}[n]\right)^J, \text{ 其中 } J \text{ 爲任意之正偶數} \qquad \text{(式 2-4)}$$

4. $X[m]$ 的動態範圍(dynamic range)爲

$$d_{X[m],(u)}[m] = \max_{1\le n\le N}\{x[n]\} - \min_{1\le n\le N}\{x[n]\}, \qquad \text{(式 2-5)}$$

5、 $X[m]$ 的機率分佈函數(probability distribution function)爲

$$F_{X[m],(u)}(z) = \frac{1}{N}\sum_{n=1}^{N} u(z - x[n]). \qquad \text{(式 2-6)}$$

其中， $u(\bullet)$ 爲單位步階函數(unit step function)，定義爲：

$$u(z) = \begin{cases} 1, & \text{if } z \ge 0 \\ 0, & \text{if } z < 0 \end{cases}$$

在以上五個式子中的各種統計值的代號中，我們以下標 $_{(u)}$ 來代表這些統計值是由整句(utterance)的特徵估測而得，值得注意的是，以上所算之針對某一項特徵 $x[m]$ 所得的各種統計值，事實上跟 $x[m]$ 於序列之順序 m 無關，意即在整句式估測法而言，我們只須計算一次統計值，就可將此統計值供整句裡每項特徵 $x[m]$ 作正規化使用。換言之，不同項特徵 $x[m]$ 共用同一組統計值。接下來，我們將討論整句式特徵統計值估測法運用在語音特徵正規化技術之可能的優缺點。

- **整句式特徵統計值估測法運用在語音特徵正規化技術之優缺點**

在以前的文獻裡，大多數強健性語音特徵正規化技術所用的統計值，皆是藉由前述的整段語句之語音特徵所求得，雖然執行上簡單有效率，而且確實對語音特徵有明顯提升強健性的效果，但還是有一些潛在的缺點，例如，整句式語音特徵正規化技術無法達到即時處理(real-time processing)的要求，因爲對一連串的語音特徵序列而言，必須等到最後一個語音特徵得到之後，才能求取統計值。除此之外，隨著語句的不同，而產生的

語音特徵序列的長度(音框數)也不一樣，不同語句所包含的音素數目或種類，及其長度的變化可能相差很大，導致影響到所估測之統計資訊的準確性。

(二) 分段式特徵統計值估測法

首先，假設某單一語句特徵$\{x[n];1 \le n \le N\}$其中某一片段特徵序列表示爲

$$\{x[k];m-L \le k \le m+L\} \quad \text{(式 2-7)}$$

其中$2L+1$爲片段特徵序列之特徵總數，(式 2-7)所表示的即爲以單項特徵$x[m]$爲中心點，前後各延展L項特徵所得的動態特徵片段。在分段式特徵統計值估測法裡，第m項特徵$x[m]$的統計值，是藉由(式 2-7)所列之片段語句所有特徵中求得。換言之，我們假設$x[m]$對應至一隨機變數$X[m]$，進而假設片段特徵序列$\{x[k];m-L \le k \le m+L\}$爲此隨機變數之樣本，然後根據此樣本，我們可估測出隨機變數$X[m]$的各種統計值，如：

1. $X[m]$的期望值(mean)爲

$$\mu_{X[m],(s)}[m] = \frac{1}{2L+1}\sum_{k=m-L}^{m+L} x[k], \quad \text{(式 2-8)}$$

2. $X[m]$的變異數(variance)爲

$$\sigma^2_{X[m],(s)}[m] = \frac{1}{2L+1}\sum_{k=m-L}^{m+L}\left(x[k]-\mu_{X[m],(s)}[m]\right)^2, \quad \text{(式 2-9)}$$

3. $X[m]$的第J階中央動差(central moment)爲

$$\xi^{(J)}_{X[m],(s)}[m] = \frac{1}{2L+1}\sum_{k=m-L}^{m+L}\left(x[k]-\mu_{X[m],(s)}[m]\right)^J, J\text{爲任意之正偶數} \quad \text{(式 2-10)}$$

4. $X[m]$的動態範圍(dynamic range)爲

$$d_{X[m],(s)}[m] = \max_{m-L\le k\le m+L}\{x[k]\} - \min_{m-L\le k\le m+L}\{x[k]\}, \quad \text{(式 2-11)}$$

5. $X[m]$的機率分佈函數(probability distribution function)爲

$$F_{X[m],(s)}(z) = \frac{1}{2L+1}\sum_{k=m-L}^{m+L} u\left(z-x[k]\right). \quad \text{(式 2-12)}$$

在以上五個式子中的各種統計值的代號中，我們以下標$_{(s)}$來代表這些統計值是由片段(segment)的特徵估測而得，從以上數式之各種統計值求取法得知，本節所用的估測法不同於上一節的整句式統計值估測法，它所針對某一項特徵$x[m]$所得的各種統計值，事實上與$x[m]$中的序列順序m有關，也就是說在分段式統計值估測法中，我們必須要個別計算整段語句中每一項特徵$x[m]$的統計值，接著將每項特徵$x[m]$的統計值供當下的特徵$x[m]$作正規化處理。換言之，不同項特徵$x[m]$所用的統計值會不一樣。以下，我們將討論分段式特徵統計值估測法運用在語音特徵正規化技術之可能的優缺點。

- **分段式特徵統計值估測法運用在語音特徵正規化技術之優缺點**

分段式語音特徵正規化技術可以彌補整段式技術的缺點，使其可以達到近似即時處理的效果。假如片段(即動態的視窗)長度越短，即時處理的優點越明顯，且因片段長度設爲固定常數，其包含的音素數目相對而言較少，不同片段所包含的音素數目較爲一致，所估測之統計值的準確性受到一片段特徵中的音素數目影響較小，因此降低了相同音素在不同語句之間的變異性。然而其缺點爲，若片段長度不夠長，代表能用以估測的樣本數較少，則估測到的統計值可能會較不精確，導致特徵統計正規化的效果變差，這意味著在這分段式技術中，可能無法同時達成即時處理的效果與大幅正規化的辨識準確性，因此通常必須在即時處理與強健性效能這兩者優點之間作取捨(trade-off)。

(三) 碼簿式特徵統計值估測法

在這一節中，我們簡要介紹如何利用碼簿資訊來估測特徵的各種統計值，而碼簿構成的詳細程序請參照文獻[7,9]。首先，假設某單一語句之某一維特徵序列$\{x[n];1 \le n \le N\}$所構成的一組碼簿，表示爲

$$\{y[r], w_r; 1 \le r \le M\} \quad \text{(式 2-13)}$$

其中w_r爲每一碼字$y[r]$所對應的權重值，而M爲碼字總數。在碼簿式特徵統計值估測法方面，我們利用(式 2-13)所示之整組碼字，去求得每一項特徵$x[m]$所對應之隨機變數$X[m]$其統計值如下：

1. $X[m]$的期望值(mean)爲

$$\mu_{X[m],(c)}[m] = \sum_{r=1}^{M} w_r y[r], \quad \text{(式 2-14)}$$

2. $X[m]$的變異數(variance)爲

$$\sigma^2_{X[m],(c)}[m] = \sum_{r=1}^{M} w_r \left(y[r] - \mu_{X[m],(c)}[m]\right)^2, \quad \text{(式 2-15)}$$

3. $X[m]$的第J階中央動差(central moment)爲

$$\xi^{(J)}_{X[m],(c)}[m] = \sum_{r=1}^{M} w_r \left(y[r] - \mu_{X[m],(c)}[m]\right)^J, J \text{爲任意之正偶數} \quad \text{(式 2-16)}$$

4. $X[m]$的動態範圍(dynamic range)爲

$$d_{X[m],(c)}[m] = \max_{1 \le r \le M} \{y[r]\} - \min_{1 \le r \le M} \{y[r]\}, \quad \text{(式 2-17)}$$

5. $X[m]$的機率分佈函數(probability distribution function)爲

$$F_{X[m],(c)}(z) = \sum_{r=1}^{M} w_r u(z - y[r]). \quad \text{(式 2-18)}$$

從上述各式所示，我們得知某一項特徵$x[m]$所對應的各種統計值，其實與$x[m]$中的序列順序m無關，也就是說在碼簿式統計值估測法方面，我們從一組碼字$\{y[r], w_r; 1 \le r \le M\}$中，只計算一次統計值，就可供整段語句中的每項特徵$x[m]$作正規化處理。換言之，不同項特徵$x[m]$將共用同一組統計值，所以本節估測法類似於整句式統計值估測法，然而主要的差別在於，整句式統計值估測法所求得的統計值是從整段語句之特徵序列求得的；碼簿式統計值估測法所求得的統計值是間接從一組碼字求得的，而不是直接從自身語句之特徵序列求得。以下，我們將討論碼簿式特徵統計值估測法運用在語音特徵正規化技術之優缺點。

- **碼簿式特徵統計值估測法運用在語音特徵正規化技術之優缺點**

碼簿式語音特徵統計估測法不同於前兩小節所提的整段式與分段式統計估測法，是由碼簿來幫助我們估算出代表訓練語音特徵與測試語音特徵的統計值，藉此有效執行各種語音特徵正規化演算法，而且也有近似即時處理的優點。在過去本實驗室的研究中[8]，我們提出了兩種碼簿式特徵統計正規化法，包括了碼簿式倒頻譜平均消去法(C-CMS)與碼簿式倒頻譜平均值與變異數正規化法(C-CMVN)，其發現 C-CMS 與 C-CMVN 的辨識結果都比前一類之整段式或分段式的方法來的好，但將其延伸至其他種類的特徵正規化法（如 HOCMN、 CGN 與 HEQ 等）時，發現其效果並沒有比整段式或分段式方法來的好，因此碼簿的統計值估測法可能不是都適用於每一種特徵正規化技術。此可能肇因於碼簿式估測法的一些缺點，例如，在雜訊語音碼簿求取的過程中，

只利用每句語音前幾個音框作爲純雜訊的代表，這會造成雜訊語音碼簿的估測不精確，而且當雜訊環境爲非穩定性(non-stationary)時，所得到的雜訊語音碼簿可能更不精準，因此其改善辨識率結果就可能較不理想。

三、併合式倒頻譜統計正規化技術

在本章中，我們參照上一章所述的整句式、分段式與碼簿式三種特徵統計估測法，提出兩種新的特徵統計估測法，稱之爲併合式(hybrid-based)統計估測法，第一種併合式統計估測法是整合了語音特徵碼簿與**整句**語音特徵的統計資訊，第二種併合式統計估測法則是整合了語音特徵碼簿與**片段**語音特徵的統計資訊。在以下各節中，我們將介紹如何將它們運用於第二章所提到之五種著名的特徵參數統計正規化技術(CMS, CMVN, HOCMN, CGN與HEQ)中，以期得到更準確的特徵強健化結果。

(一) 併合式倒頻譜平均消去法與併合式倒頻譜平均值與變異數正規化法

在這裡我們將一同介紹併合式倒頻譜平均消去法(hybrid-based CMS)與併合式倒頻譜平均值與變異數正規化法(hybrid-based CMVN)。假設某一維原始之輸入特徵序列爲 $\{x[n];1 \le n \le N\}$，則經過 CMS 處理後的輸出特徵參數表示式如下：

$$\tilde{x}[n] = x[n] - \mu[n], \qquad 1 \le n \le N, \tag{式 3-1}$$

而經過 CMVN 處理後的特徵參數表示式如下：

$$\tilde{x}[n] = \big(x[n] - \mu[n]\big) / \sigma[n], \qquad 1 \le n \le N, \tag{式 3-2}$$

其中 N 爲整段序列之特徵總數，而 $\mu[n]$ 與 $\sigma[n]$ 分別爲特徵 $x[n]$ 的平均值與標準差。

在第一種併合式特徵統計估測法中，$\mu[n]$ 與 $\sigma[n]$ 可由下列兩公式估測而得：

CU-CMS/CU-CMVN：

$$\mu_{(c,u)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(u)}[n], \tag{式 3-3}$$

$$\sigma^2_{(c,u)}[n] = \alpha\Big[\sigma^2_{(c)}[n] + \mu^2_{(c)}[n]\Big] + (1-\alpha)\Big[\sigma^2_{(u)}[n] + \mu^2_{(u)}[n]\Big] - \mu^2_{(c,u)}[n], \tag{式 3-4}$$

其中下標"(c)"、"(u)"與"(c,u)"分別代表使用碼簿式、整句式與併合碼簿/整句式統計值估測法，而 $\mu_{(c)}[n]$、$\mu_{(u)}[n]$、$\sigma^2_{(c)}[n]$ 與 $\sigma^2_{(u)}[n]$ 分別定義於前一章的(式 2-14)、(式 2-2)、(式 2-15)與(式 2-3)，α 爲權重值，介於 0 到 1 之間，被用來調整碼簿式統計資訊與整段式統計資訊之間的比例。藉由(式 3-3)與(式 3-4)所估測之平均值與變異數而成的 CMS 與 CMVN，我們分別稱爲併合碼簿/整句式 CMS(hybrid codebook/utterance-based CMS, CU-CMS) 與併合碼簿/整句式 CMVN(hybrid codebook/utterance-based CMVN, CU-CMVN)，由(式 3-3)與(式 3-4)可明顯看出，CU-CMS 與 CU-CMVN 所使用的平均值與變異數是將前一章所述之語音特徵碼簿與整段語音特徵的平均值與變異數作一線性的組合。如果權重值 $\alpha = 1$ 時，CU-CMS 和 CU-CMVN 將分別等同於碼簿式 CMS (C-CMS) 和碼簿式 CMVN (C-CMVN)，另一方面，如果 $\alpha = 0$ 時，CU-CMS 和 CU-CMVN 將分別等同於整句式 CMS (U-CMS)和整句式 CMVN (U-CMVN)。

在第二種併合式特徵統計估測法中，$\mu[n]$ 與 $\sigma[n]$ 可由下列兩公式估測而得：

CS-CMS/CS-CMVN：

$$\mu_{(c,s)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(s)}[n], \tag{式 3-5}$$

$$\sigma^2_{(c,s)}[n] = \alpha\Big[\sigma^2_{(c)}[n] + \mu^2_{(c)}[n]\Big] + (1-\alpha)\Big[\sigma^2_{(s)}[n] + \mu^2_{(s)}[n]\Big] - \mu^2_{(c,s)}[n] \tag{式 3-6}$$

其中下標"(c)"、"(s)"與"(c,s)"分別代表使用碼簿式、分段式與併合碼簿/分段式統計值

估測法，而$\mu_{(c)}[n]$、$\mu_{(s)}[n]$、$\sigma^2_{(c)}[n]$與$\sigma^2_{(s)}[n]$分別定義於前一章的(式 2-14)、(式 2-8)、 (式 2-15)與(式 2-9)，α爲權重值，介於 0 到 1 之間，被用來調整碼簿式統計資訊與分段式統計資訊之間的比例。藉由(式 3-5)與(式 3-6)所估測之平均值與變異數而成的 CMS 與 CMVN 法，我們分別稱爲併合碼簿/分段式 CMS(hybrid codebook/segment-based CMS, CS-CMS) 與併合碼簿/分段式 CMVN(hybrid codebook/segment-based CMVN, CS-CMVN)，類似前面所提之 CU-CMS 與 CU-CMVN，從(式 3-5)與(式 3-6)可看出，CS-CMS 與 CS-CMVN 所使用的平均值與變異數是將前一章所述之語音特徵碼簿與片段語音特徵的平均值與變異數作一線性的組合。如果權重$\alpha=1$時，CS-CMS 和 CS-CMVN 將分別等同於碼簿式 CMS (C-CMS)和碼簿式 CMVN (C-CMVN)，然而，若$\alpha=0$時，CS-CMS 和 CS-CMVN 將分別等同於分段式 CMS (S-CMS)和分段式 CMVN (S-CMVN)。

(二) 併合式高階倒頻譜動差正規化法

對一特徵時間序列$\{x[n];1\le n\le N\}$而言，經高階倒頻譜動差正規化法(HOCMN)處理後所得的新特徵時間序列如下式：

$$\tilde{x}[n]=\big(x[n]-\mu[n]\big)/\Big(\xi^{(J)}[n]\Big)^{1/J},\qquad 1\le n\le N \tag{式 3-7}$$

其中N爲整段序列之特徵數，$\mu[n]$與$\xi^{(J)}[n]$分爲特徵$x[n]$的平均值與第J階中央動差。

類似上一節所述，這裡我們有兩種方式來估測$\mu[n]$與$\xi^{(J)}[n]$，所對應的 HOCMN 法我們分別稱爲併合碼簿/整句式 HOCMN(CU-HOCMN) 與併合碼簿/分段式 HOCMN(CS-HOCMN)，它們在$\mu[n]$與$\xi^{(J)}[n]$的估測運算如下列數式：

CU-HOCMN (hybrid codebook/utterance-based HOCMN)：

$$\mu_{(c,u)}[n]=\alpha\mu_{(c)}[n]+(1-\alpha)\,\mu_{(u)}[n], \tag{式 3-8}$$

$$\xi^{(J)}_{(c,u)}[n]=\alpha\left(\sum_{r=1}^{M}w_r\Big(y[r]-\mu_{(c,u)}[n]\Big)^J\right)+(1-\alpha)\left(\frac{1}{N}\sum_{n=1}^{N}\Big(x[n]-\mu_{(c,u)}[n]\Big)^J\right). \tag{式 3-9}$$

CS-HOCMN (hybrid codebook/segment-based HOCMN)：

$$\mu_{i,(c,s)}[n]=\alpha\mu_{i,(c)}[n]+(1-\alpha)\,\mu_{i,(s)}[n], \tag{式 3-10}$$

$$\xi^{(J)}_{(c,s)}[n]=\alpha\left(\sum_{r=1}^{M}w_r\Big(y[r]-\mu_{(c,s)}[n]\Big)^J\right)+(1-\alpha)\left(\frac{1}{2L+1}\sum_{k=n-L}^{n+L}\Big(x[k]-\mu_{(c,s)}[n]\Big)^J\right). \tag{式 3-11}$$

其中$\mu_{(c)}[n]$、$\mu_{(u)}[n]$、$\mu_{(s)}[n]$分別定義於前一章的(式 2-14)、(式 2-2)與(式 2-8)，α爲權重值，介於 0 到 1 之間，用來調整碼簿統計資訊與整段或片段特徵統計資訊之間的比例。

(三) 併合式倒頻譜增益正規化法

對一特徵時間序列$\{x[n];1\le n\le N\}$而言，經倒頻譜增益正規法(CGN)處理後所得的新特徵時間序列如下式：

$$\tilde{x}[n]=\big(x[n]-\mu[n]\big)/d[n],\qquad 1\le n\le N, \tag{式 3-12}$$

其中N爲整段序列之特徵總數，$\mu[n]$與$d[n]$分別爲特徵$x[n]$的平均值與動態範圍。

類似上兩節的方法，這裡我們有兩種方式來估測$\mu[n]$與$d[n]$，所對應的 CGN 法我們分別稱爲併合碼簿/整句式 CGN(CU-CGN) 與併合碼簿/分段式 CGN(CS-CGN)，它們對$\mu[n]$與$d[n]$的估測運算分別如下數式：

CU-CGN(hybrid codebook/utterance-based CGN)：

$$\mu_{(c,u)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(u)}[n], \qquad \text{(式 3-13)}$$

$$d_{(c,u)}[n] = \max\left\{Y_{(c)} \cup X_{(u)}\right\} - \min\left\{Y_{(c)} \cup X_{(u)}\right\} \qquad \text{(式 3-14)}$$

CS-CGN(hybrid codebook/segment-based CGN)：

$$\mu_{(c,s)}[n] = \alpha\mu_{(c)}[n] + (1-\alpha)\,\mu_{(s)}[n], \qquad \text{(式 3-15)}$$

$$d_{(c,s)}[n] = \max\left\{Y_{(c)} \cup X_{(s)}\right\} - \min\left\{Y_{(c)} \cup X_{(s)}\right\} \qquad \text{(式 3-16)}$$

其中$Y_{(c)}$、$X_{(u)}$與$X_{(s)}$代表了(式 2-13)、(式 2-1)與(式 2-7)。$\mu_{(c)}[n]$、$\mu_{(u)}[n]$與$\mu_{(s)}[n]$分別定義於前一章的(式 2-14)、(式 2-2)與(式 2-8)，其中α爲一個介於 0 到 1 之間的權重值，代表了碼簿式統計資訊與整句式或分段式統計資訊這兩者之間所使用的比例，$\max(\cdot)$與$\min(\cdot)$分別爲取最大值與最小值的函數，而『∪』爲聯集符號，意指將碼簿與整句(或片段)語句的特徵串在一起。

(四) 併合式倒頻譜統計圖等化法

對一特徵時間序列$\{x[n];1 \le n \le N\}$而言，經倒頻譜統計圖等化法(HEQ)處理後所得的新特徵時間序列如下式：

$$\tilde{x}[n] = F_{ref}^{-1}\left(F_X(x[n])\right), \qquad 1 \le n \le N, \qquad \text{(式 3-17)}$$

其中N爲整段序列之特徵總數，$F_{ref}(\cdot)$爲預先定義的參考機率分布函數，而$F_X(\cdot)$則爲特徵$x[n]$的機率分布函數。

類似前面幾節所述，這裡我們有兩種方式來估測機率分布函數$F_X(\cdot)$，分別使用在 HEQ 上，因此我們分別稱爲併合碼簿/整句式 HEQ(CU-HEQ)與併合碼簿/分段式 HEQ(CS-HEQ)，它們對$F_X(\cdot)$的估測表示式如下所示：

CU-HEQ(hybrid codebook/utterance-based HEQ)：

$$F_{X,(c,u)}(z) = \alpha F_{X,(c)}(z) + (1-\alpha)\,F_{X,(u)}(z), \qquad \text{(式 3-18)}$$

CS-HEQ(hybrid codebook/segment-based HEQ)：

$$F_{X,(c,s)}(z) = \alpha F_{X,(c)}(z) + (1-\alpha)\,F_{X,(s)}(z), \qquad \text{(式 3-19)}$$

其中α爲一個介於 0 到 1 之間的權重值，代表了碼簿式統計資訊與整段式或分段式統計資訊這兩者之間所使用的比例，而$F_{X,(c)}(\cdot)$、$F_{X,(u)}(\cdot)$與$F_{X,(s)}(\cdot)$分別定義於前一章的(式 2-18)、(式 2-6) 與(式 2-12)。

四、實驗環境設定與各種強健性語音特徵正規化技術之實驗結果與討論

(一) 實驗環境設定

本論文採用歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行的 AURORA2 語音資料庫[10]，其內容是由連續的英文數字字串所構成。此語音資料庫有兩種不同的訓練環境：乾淨環境(clean-condition)與多重環境(multi-condition)以及三種不同的測試集合：A 組(地下鐵、人聲、汽車和展覽館雜訊)、B 組(餐廳、街道、機場和火車站雜訊)與 C 組(地下鐵、街道雜訊外加 MIRS 通道效應)雜訊語音集合。乾淨環境代表沒有任何雜訊的語音環境，而多重環境則代表適當加入各種附加雜訊的語音環境。本論文的實驗只採用乾淨環境的語音特徵作聲學模型的訓練，並對三組雜訊語音集合加以辨識。

在這裡，基礎實驗(baseline experiment)將採用未處理的梅爾倒頻譜特徵係數(MFCC)作爲訓練跟測試，所使用的MFCC特徵參數爲13維(c_0~c_{12})，再加上其一階和二階差量，總共有39維特徵參數作爲最終使用之特徵參數向量。

聲學模型爲由左向右(left-to-right)之隱藏式馬可夫模型(hidden Markov model, HMM)

的形式，是使用隱藏式馬可夫模型訓練軟體 HTK[11]訓練所得，其中包含 11 個數字模型(zero, one, two, …, nine 及 oh)以及靜音(silence)模型，每個數字模型包含 16 個狀態，而每個狀態則包含 20 個高斯密度混合。

(二) 各種強健性語音特徵正規化技術之實驗結果與討論

本章將介紹我們所提的五種強健性語音特徵正規化技術之辨識實驗結果(20dB、15dB、10dB、5dB 與 0dB 五種訊雜比下的辨識率平均)，分別爲 CMS、CMVN、HOCMN、CGN 以及 HEQ，而這些方法的特徵統計相關資訊，分別使用三種特徵統計值估測法(整句式、分段式與碼簿式)以及兩種併合式(hybrid-based)特徵統計值估測法(碼簿/整句式與碼簿/分段式)來求得，然後進一步比較、討論與分析。

在本論文中，所有片段式統計估測法實驗中的片段長度 $2L+1$ 都設爲 101(除了 HOCMN 之外，它的片段長度 $2L+1$ 設爲 87)；所有碼簿式與併合式實驗中的碼字數目 R 統一設定爲 16 或 256，而併合式實驗中的 α，我們固定設爲 0.5，使併合之雙方統計資訊所佔的比例相等。

1、各種倒頻譜平均消去法之實驗結果

從表一中，我們得知各種 CMS 的總平均辨識率情形。我們先探討三種傳統的特徵統計值估測法作用在 CMS 上的效果，發現 U-CMS、S-CMS 與 C-CMS 中，以 C-CMS (R=256)所得的整體平均辨識率最大，比基礎實驗結果提升了 10.75%，而相對錯誤率衰減率(RR)爲 37.83%。

接著，我們可看出本論文新提出的兩種併合式特徵統計值估測法運用在 CMS 上之效果，如以下幾點所述：

(1) CU-CMS 與 CS-CMS 皆明顯優於傳統之 U-CMS、S-CMS 與 C-CMS。

(2) CS-CMS(R=16)表現優於 CU-CMS(R=16)，相對於基礎實驗結果，在辨識率上提升了 14.61%，而相對錯誤率衰減率高達 51.41%。

(3)在各種不同型態的 CMS 中，以 CS-CMS (R=16)的總平均辨識結果最佳，而其它 CMS 的總平均辨識結果的優劣順序，依序爲：CS-CMS (R=256)、CU-CMS (R=16)、CU-CMS (R=256)、C-CMS (R=256)、C-CMS(R=16)、U-CMS 以及 S-CMS。因此，我們驗證了所新提出之兩種併合式 CMS 在提昇語音特徵強健性上，比 U-CMS、S-CMS 與 C-CMS 還要來的優越。

method	Set A	Set B	Set C	Average	RR
baseline	71.92	68.22	77.61	71.58	—
C-CMS (R=16)	80.83	79.29	86.13	81.27	34.10
C-CMS (R=256)	81.62	81.58	85.25	82.33	37.83
U-CMS	79.35	82.46	79.91	80.71	32.13
CU-CMS (R=16)	83.28	84.92	84.28	**84.14**	44.19
CU-CMS (R=256)	82.13	84.29	83.06	83.18	40.82
S-CMS	77.28	80.66	77.63	78.70	25.05
CS-CMS (R=16)	85.38	87.43	85.31	**86.19**	51.41
CS-CMS (R=256)	84.63	86.98	84.42	85.53	49.09

表一、各種 CMS 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

2、各種倒頻譜平均值與變異數正規化法之實驗結果

表二呈現了各種倒頻譜平均值與變異數正規化法(CMVN)的辨識率，由表二中，首先，我們可以看出三種傳統的特徵統計值估測法運用於 CMVN 時，其辨識結果相較於基礎實驗而言，在 U-CMVN、S-CMVN 與 C-CMVN 中，以 C-CMVN (R=256)所得的總平均辨識率最高，比基礎實驗結果提升了 15.14%，而相對錯誤率降低率達到了 53.27%。

接下來，本論文提出兩種併合式特徵統計值估測法運用在 CMVN 上，同樣其辨識結果相較於基礎實驗也有明顯的進步。由表二所示，在 CU-CMVN 與 CS-CMVN 中，以 CS-CMVN (R=16)的總平均辨識率最高，比基礎實驗結果提升了 17.38%，而相對錯誤率降低率高達 61.15%。因此兩種併合式 CMVN 在語音特徵強健性方面，跟前一節的併合式 CMS 一樣，皆優於 U-CMVN、S-CMVN 與 C-CMVN。

Method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-CMVN (R=16)	85.75	85.5	83.78	85.26	48.14
C-CMVN (R=256)	87.16	87.44	84.39	86.72	53.27
U-CMVN	85.03	85.56	85.61	85.36	48.49
CU-CMVN (R=16)	87.87	88.67	86.14	**87.84**	57.21
CU-CMVN (R=256)	87.25	88.06	85.78	87.28	55.24
S-CMVN	83.99	84.85	84.78	84.49	45.43
CS-CMVN (R=16)	88.98	89.82	87.19	**88.96**	61.15
CS-CMVN (R=256)	88.18	89.09	86.73	88.25	58.66

表二、各種 CMVN 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

3、各種高階倒頻譜動差正規化法之實驗結果

表三呈現了各種高階倒頻譜動差正規化法(HOCMN)的辨識率，在這裡，中央動差的階數 J 皆設爲 100。從表三可看出以下幾點現象：

(1) 在傳統之 U-HOCMN、S-HOCMN 與 C-HOCMN 中，以 U-HOCMN 所得的總平均辨識率最大，與基礎實驗相比，提升了 16.31%，而相對錯誤降低率達到了 57.39%。此現象跟前兩節所呈現的結果並不相同，因爲碼簿式的 HOCMN(C-HOCMN)表現並不理想，其可能原因爲，碼簿對於較低階的動差值（例如平均值與變異數）之估測較爲準確，但無法有效估測較高階的動差值。

(2) 本論文提出兩種併合式特徵統計值估測法，當其運用在 HOCMN 上時，其辨識結果相較於基礎實驗結果仍有明顯改善：CU-HOCMN(R=16)與 CS-HOCMN(R=16)分別比基礎實驗結果提升了 16.09%與 17.78%，相對錯誤降低率衰減分別高達 56.62%與 62.56%。

(3) 在各種型態的 HOCMN 中，以 CS-HOCMN (R=16)的總平均辨識率最佳，其它併合式 HOCMN 皆低於 U-HOCMN，不同於前面幾節所述的併合式 CMS 與 CMVN 所呈現的結果。更進一步觀察，可看出 CU-HOCMN 在 Set C 的辨識率相對較低，造成總平均辨識率不及 U-HOCMN，此現象可歸因於 Set C 中的語音包含了摺積性雜訊(convolutional noise)，而碼簿(codebook)中只考慮到加成性雜訊(additive noise)，所以

造成 Set C 之辨識結果不盡理想。

Method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-HOCMN (R=16)	84.86	83.40	85.68	84.44	45.25
C-HOCMN (R=256)	86.30	86.34	83.53	85.76	49.89
U-HOCMN	87.43	88.54	87.52	87.89	57.39
CU-HOCMN (R=16)	87.55	88.85	85.57	**87.67***	56.62
CU-HOCMN (R=256)	86.66	88.15	85.12	86.95	54.08
S-HOCMN	85.60	86.63	86.16	86.12	51.16
CS-HOCMN (R=16)	89.17	90.26	87.96	**89.36**	62.56
CS-HOCMN (R=256)	87.44	88.78	86.22	87.73	56.83

表三、各種HOCMN的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

4、各種倒頻譜增益正規化法之實驗結果

表四列出了各種倒頻譜增益正規化法(CGN)的辨識率，若與表三相比較，我們發現它們的結果十分類似，三種傳統估測法所對應的U-CGN、S-CGN與C-CGN中，以U-CGN所得的總平均辨識率最高，與基礎實驗結果相比，提升了 16.33%，而相對錯誤降低率爲 57.46%，相對而言，C-CGN 表現較差，此可能原因跟上一節所述類似，即碼簿可能無法較精確地估測 CGN 所需用到的動態範圍值，以及其未考慮到 Set C 的摺積性雜訊干擾。

然而，當兩種併合式特徵統計值估測法，分別使用在 CGN 時，其辨識結果都能有十分顯著的提昇，其中以 CS-CGN(R=16)的總平均辨識率最高，比基礎實驗提升了 17.88%，而相對錯誤降低率高達了 62.91%，我們證實了在併合式 CGN 中，除了 CU-CGN (R=256)略低於 U-CGN 之外，皆優於 U-CGN、S-CGN 與 C-CGN。

method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-CGN (R=16)	85.60	84.29	84.82	84.92	46.94
C-CGN (R=256)	86.65	87.07	84.41	86.37	52.04
U-CGN	87.62	88.49	87.32	87.91	57.46
CU-CGN (R=16)	88.11	89.14	86.07	**88.11**	58.16
CU-CGN (R=256)	86.97	88.46	85.30	87.23	55.07
S-CGN	86.36	87.31	86.99	86.86	53.76
CS-CGN (R=16)	89.31	90.40	87.89	**89.46**	62.91
CS-CGN (R=256)	88.42	89.73	86.92	88.64	60.03

表四、各種 CGN 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

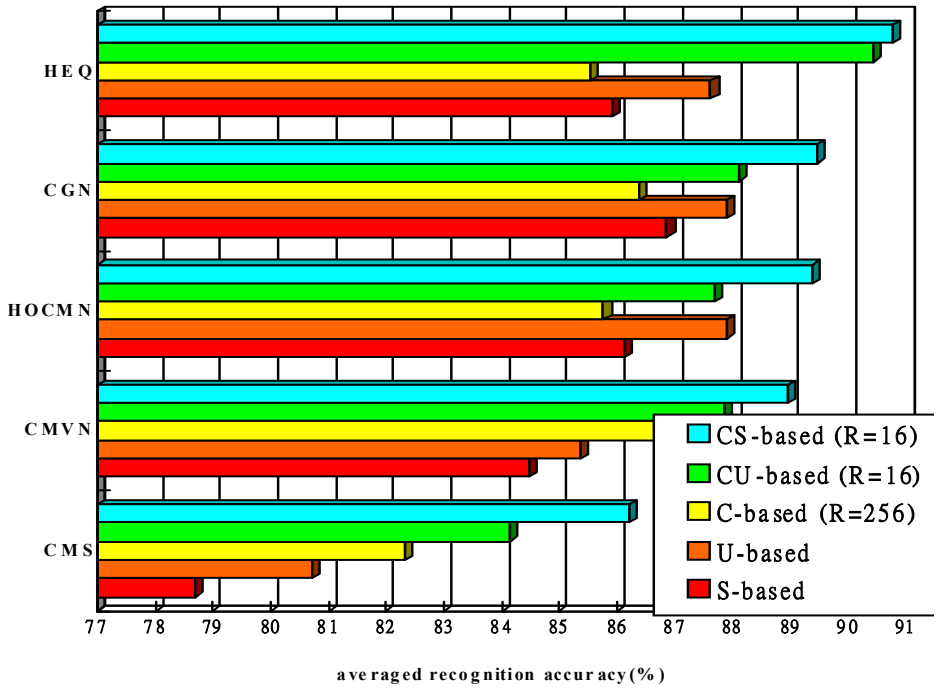
5、各種統計圖等化法之實驗結果

表五列出了各種不同型態的統計圖等化法(HEQ)之辨識率，其結果與表三與表四類似，我們明顯看出，雖然碼簿式的 HEQ(C-HEQ)效果不盡理想，然而當我們把碼簿與整句特徵合併，或把碼簿與片段特徵合併，所分別對應的 CU-HEQ 與 CS-HEQ，其帶來的辨識率提昇程度皆十分顯著，明顯超越了整句式 HEQ(U-HEQ)與片段式 HEQ(S-HEQ)，而其中又以 CS-HEQ 表現最佳。平均辨識率高達 90.76%，相對錯誤降低率爲 67.49%。

method	Set A	Set B	Set C	Average	RR
Baseline	71.92	68.22	77.61	71.58	—
C-HEQ (R=16)	80.15	80.41	76.03	79.43	27.62
C-HEQ (R=256)	86.23	85.77	83.71	85.54	49.12
U-HEQ	86.95	88.39	87.40	87.62	56.44
CU-HEQ (R=16)	90.21	91.16	89.37	**90.42**	66.29
CU-HEQ (R=256)	88.76	89.68	87.85	88.95	61.12
S-HEQ	85.10	86.82	85.64	85.90	50.39
CS-HEQ (R=16)	90.57	91.54	89.57	**90.76**	67.49
CS-HEQ (R=256)	89.11	90.21	88.35	89.40	62.70

表五、各種 HEQ 的整體平均辨識率與相對錯誤降低率(relative error rate reduction, RR)之比較

(三) 綜合討論



圖一、各種語音特徵正規化法在不同的特徵統計估測法下的總平均辨識率之比較

圖一是本章所有方法之總平均辨識率比較圖，從此圖與前面的五個表中，我們大致觀察到以下幾種情形：

(1) 整句式方法皆比分段式方法還要好，其可能原因爲，分段式方法是把整段語句分割成許多相鄰的片段語句，雖然可達到近似即時處理的效果，但也因片段資料量的不足導致統計估測值較不準確，因此造成辨識率的下降。在碼簿式方法方面，除了 C-CMS 與 C-CMVN 之外，效果皆低於整句式與分段式方法，其可能原因爲，虛擬雙通道中的雜訊估測只以整段語句的前幾個音框當雜訊的代表，所得之雜訊特性可能較不精準，造成建立的的碼簿比較不精確。

(2) 併合式方法幾乎皆比整句式、分段式與碼簿式方法來的好，例如：CU-HEQ (90.42%) 與 CS-HEQ (90.76%)優於 U-HEQ (87.62%)、S-HEQ (85.9%)與 C-HEQ (85.54%)。這結果證實了整句式、分段式與碼簿式方法，獨自所提升的辨識率較小，但結合了碼簿式與整句式(或分段式)的併合式方法，將使辨識率大幅地提升。

(3) 在併合式方法中，我們將α都設爲 0.5，這表示著碼簿與整段語句的統計資訊之使用比例相等，而沒有任何偏差。雖然$\alpha = 0.5$未必是最佳的設定參數，但至少代表了我們無須精微地挑選此參數值，便能得到明顯的併合效益。

(4) 在兩種併合式方法中，碼字數目 *R*=16 所對應的辨識結果明顯比 *R*=256 所對應的辨識結果有明顯的改善，此現象在 CMS、CMVN、HOCMN、CGN 與 HEQ 皆是如此，此可能原因在於碼簿資訊與整句語音資訊(或片段語音資訊)之間的不一致性。由於碼簿只呈現純語音部分的資訊，而整句或片段特徵可能同時包含了語音與非語音部分的資訊，因此增加碼字數目，使碼簿所對應的語音專屬資訊越多且越詳細，這將會使雙方的不一致性越來越明顯。然而此結果卻反而成爲我們所提出之併合式方法的優點，因爲這代表了我們在各種併合式倒頻譜統計正規化法中，只要使用較小的碼字數目，就可得到較佳的辨識結果，而且大幅降低演算法本身的運算複雜度。

(5) 在併合式方法中，以碼簿/分段式方法的辨識效能最高，它補足了分段式方法之辨識率較差的缺點，而仍保有分段式方法之近似及時處理的優點，因此極具實用價值。

五、結論

在本論文中，我們提出了兩種併合式的特徵統計估測法，它們是將語音特徵之碼簿與整句或分段的語音特徵適當地結合，進而估測出特徵的各項統計值，我們將此新方法分別使用在五種強健性語音特徵統計正規化技術上：倒頻譜平均消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)、高階倒頻譜動差正規化法(HOCMN)、倒頻譜增益正規化法(CGN)以及倒頻譜統計圖等化法(HEQ)，我們發現，跟整句式、分段式與碼簿式統計估測法相較之下，這兩種併合式的估測法皆能更明顯地提昇語音特徵正規化技術的效能，更有效地改善雜訊環境下的語音辨識率，因此可推論，我們所提出的新方法得以得到更精確的語音特徵的統計特性。

在未來的相關研究上，我們有以下幾個方向發展：

(1) 我們希望利用併合式特徵統計估測法，運用在其他的倒頻譜統計正規化技術上，例如：倒頻譜形狀正規化法(CSN)或其他階層的 HOCMN 等技術，以觀察其效能。

(2) 在實驗結果中，我們發現了併合式的倒頻譜統計正規化技術在摺積性雜訊環境下的辨識改善程度，較差於加成性雜訊環境，因此我們期望能結合消除通道效應的方法，例如相對頻譜法(RASTA)等，使其提升辨識結果。
(3) 除了本論文所用的數字語音資料庫外，我們將嘗試把所提的新方法運用在其它較大字彙的語音資料庫上，進一步驗證這些方法的實用價值。

參考文獻

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Transactions on Acoustics, Speech and Signal Processing*, Volume 29, Issue 2, pp. 254-272, 1981

[2] C.-P. Chen, K. Filaliy and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the Aurora 2.0/3.0 databases", *in Proc. International Conference on Spoken Language Processing* (*ICSLP*), pp. 241-244, 2002

[3] R. Haeb-Umbach, "Investigations on inter-speaker variability in the feature space", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 397-400, 1999

[4] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 197-200, 2004

[5] S. Yoshizawa, N. Hayasaka, W. Naoya, and Y. Miyanaga. "Cepstral gain normalization for noise robust speech recognition", *in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp. 209-212, 2004

[6] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust speech recognition", *European Conference on Speech Communication and Technology* (*Eurospeech*), pp. 1135-1138, 2001

[7] J.-W. Hung, "Cepstral statistics compensation and normalization using online pseudo stereo codebooks for robust speech recognition in additive noise environments", *IEICE Trans. Information and Systems*, pp. 296-311, 2008

[8] T.-H. Hsieh, “Feature statistics compensation for robust speech recognition in additive noise environments”, *M.S. thesis*, National Chi Nan University, Taiwan, 2007

[9] K.-C. Wu, “Study of cepstral statistics normalization techniques for robust speech recognition in additive noise environments”, *M.S. thesis*, National Chi Nan University, Taiwan, 2008

[10] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *in Proc. of ISCA IIWR ASR2000*, Paris, France, 2000

[11] http://htk.eng.cam.ac.uk/