Computational Linguistics and Chinese Language Processing
Vol. 17, No. 4, December 2012, pp. 69-84

語音辨識使用統計圖等化方法

Speech Recognition Leveraging Histogram Equalization Methods

謝欣汝*+、洪志偉+、陳柏琳*

Hsin-Ju Hsieh, Jeih-weih Hung, and Berlin Chen

摘要

統計圖等化法(Histogram Equalization, HEQ)是一種概念簡單且有效的語音特徵處理技術，近年來被廣泛地研究與應用於強健性語音辨識的領域。在本論文中，我們延續統計圖等化法的研究，提出一系列使用語音特徵的空間－時間之文脈統計資訊 (Spatial-Temporal Contextual Statistics)的語音特徵強健方法；其作法是在語音之倒頻譜特徵上，利用一個簡易的差分(Differencing)和平均(Averaging)的處理方式，來得到語音特徵之文脈統計資訊後予以正規化並結合。這些新方法的作法有別於傳統之個別維度獨立正規化(Dimension-Wise)的統計圖等化法，進一步地正規化不同空間與時間之間的特徵分布資訊，因此可以降低不同聲學環境所產生的偏差，並且嘗試消除傳統之統計圖等化法無法補償的問題，亦即隨機性雜訊(Random Noise)對語音所產生的影響。本論文所有的語音辨識實驗皆是作用於國際通用的連續語音語料庫 Aurora-2 上；實驗結果顯示，我們所提出之方法相較於許多著名的特徵強化法，皆有不錯的效果。

關鍵詞：語音辨識，雜訊強健性，統計圖等化法，特徵文脈的統計

* 國立臺灣師範大學資訊工程學系 Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: hsinju@ntnu.edu.tw; berlin@ntnu.edu.tw
+ 國立暨南國際大學電機工程學系 Department of Electrical Engineering, National Chi Nan University
E-mail: jwhung@ncnu.edu.tw

Abstract

Histogram equalization (HEQ) of speech features has received considerable attention in the field of robust speech recognition due to its simplicity and excellent performance. This paper is a continuation of this general line of research, presenting a novel HEQ-based feature normalization framework which takes advantage of joint equalization of spatial-temporal contextual statistics of speech features. In doing so, we explore the use of simple differencing and averaging operations to capture the contextual statistics of feature vector components for speech feature normalization. All experiments are conducted on the Aurora-2 database and task. Experimental results show that for clean-condition training, the methods instantiated from this framework achieve considerable word error rate reductions over the baseline system, which are indeed quite comparable to other conventional methods.

Keywords: Speech Recognition, Noise Robustness, Histogram Equalization, Feature Contextual Statistics.

1. 研究動機

『科技始終來自於人性』，這是一家手機廠商的廣告用語；隨著科技不斷的進步，電腦功能不斷地提升、相關資訊設備也日漸普及並且深入到你我的日常生活中，不僅爲人類生活帶來許多的便利性，更是大大地提升工作效率與生活品質。現今我們可以藉由電腦或其它資訊設備來完成大部分的工作，如此一來便使得人類與電腦間有著密不可分的關係。但目前人類與電腦的溝通方式，仍須仰賴鍵盤、滑鼠等工具，因此對於某些特定族群的使用者而言，這種不友善的操作介面無疑是一個障礙。我們相信以最自然且簡便的方式來操作這些科技產品，能將科技帶給人們的效益提升到最高。

由於語音是人們最自然且最普遍使用的溝通媒介，因此在不久的將來，語音必然會扮演著人類與智慧型電子設備間，最重要的互動媒介，而自動語音辨識(Automatic Speech Recognition, ASR)技術將會是一個關鍵的角色。然而在現實生活中，已有許多和自動語音辨識技術相關的應用，其中最廣爲人知的應用爲航空公司的語音訂位系統及銀行帳戶的語音查詢系統等；而這一類的系統能成功運作的原因主要是因爲限制系統辨識的詞彙個數。此外還有許多的自動語音辨識相關的應用，如語音轉譯文字軟體、互動式聲音問答系統和語音文件檢索等；然而要實現這類技術，將會面臨許多困難與障礙。

對於一套自動語音辨識系統而言，在語音訊號不受雜訊干擾的理想實驗室環境下，一般皆可獲得良好的辨識結果，但若要應用至日常生活的環境中，常會受到環境中諸多雜訊的干擾，例如:具有加成性的背景雜訊(Background Noise)或是錄音設備本身所產生的摺積性的通道效應(Channel Effect)等，皆會造成系統之訓練環境與測試環境之間存在不匹配(Mismatch)的情況，而嚴重地影響系統的辨識效能。因此，在自動語音辨識技術的

發展上，雜訊強健性(Noise Robustness)一直是一門重要的研究議題。並且，如何能以更有效的方式來處理雜訊所造成的影響，將是一個既複雜又頗具挑戰性的任務。

如前所述，對於語音訊號而言，環境中雜訊的干擾大致可分爲兩種類型: (1)加成性雜訊(Additive Noise)和(2)摺積性(Convolutional Noise)雜訊。其中加成性雜訊爲錄製語音時，原始語音與背景雜訊呈線性加成的關係一同被收錄進去，例如汽車乎嘯而過或周遭人們聊天所產生的噪音等；另一方面，摺積性雜訊則是指語音訊號經由不同傳輸通道所造成的通道效應，例如麥克風通道效應、電話線路或手持式電話所產生的通道效應等。圖 1 爲乾淨語音訊號受加成性雜訊與摺積性雜訊干擾的示意圖。

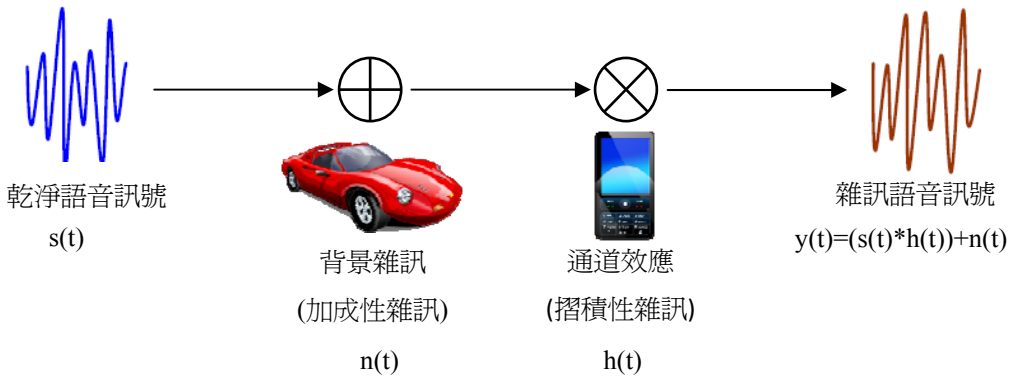


圖 1. 雜訊對語音訊號干擾示意圖

在強健性語音辨識的研究領域裡，過去已有許多學者已成功的發展出許多相關的演算法，其主要目的爲降低雜訊對語音訊號的影響，進而使得辨識結果能夠有效的提升。依據所發展出方法的特性，約略可分爲以下三大研究方向(Gong, 1995):

(1) 語音訊號增益法(Speech Enhancement):

考量人耳聽覺的特性，以增加語音訊號在感知上的品質。其主要的目的爲將語音訊號從受雜訊干擾之空間轉換至乾淨語音空間，期望轉換後的語音訊號能與對應的乾淨語音訊號相似。但此方法不保證一定可使自動語音辨識之效能提高。原因爲大多數的語音增益方法都會導致訊號失真的情形，雖然人耳對於些許訊號的失真有很好的容忍力，但是這些干擾對自動語音辨識器而言則相當敏感。常見的技術有頻譜消去法(Spectral Subtraction, SS) (Boll, 1979)、端點偵測(Voice Activity Detection, VAD) (ITU, 1996)等。

(2) 強健性語音特徵(Robust Speech Feature):

主要作法是希望從語音訊號中擷取較不易受到雜訊干擾而失真的強健性語音特徵參數，進而降低訓練語料和測試語料間存在的不匹配情況，因此可以有效的提升自動語音辨識的效能。其著名的方法有倒頻譜平均值消去法(Cepstrum Mean Subtraction, CMS) (Furui, 1981)、倒頻譜平均值與變異數正規化法(Cepstrum Mean

and Variance Normalization, CMVN) (Viikki & Laurila, 1998)與倒頻譜平均與變異數正規化法結合自動回歸動態平均濾波器法(Cepstral Mean and Variance Normalization plus Auto-Regressive-Moving Averaging Filtering, MVA) (Chen & Bilmes, 2006)等。

(3) 聲學模型調適法(Acoustic Model Adaptation):

藉由辨識器的學習，以轉換聲學模型內的分佈，進而獲得與輸入的雜訊語音向量近似的分佈。常見的技術有最大事後機率法則(Maximum a Posteriori, MAP) (Gauvain & Lee, 1994)、最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR) (Leggetter & Woodland, 1995)、平行模型結合法(Parallel Model Combination, PMC) (Hung *et al*., 2001)等。

本論文所提出之新方法是基於上述第二類的強健性語音特徵所發展出來的。而目前最廣泛被使用的語音特徵參數包含以人耳之聽覺特性爲考量依據而發展出的梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCCs) (Davis & Mermelstein, 1980)、線性預估倒頻譜係數(Linear Prediction Cepstral Coefficients, LPCC) (Atal, 1974)及感知線性預估倒頻譜係數(Perceptual Linear Prediction Cepstral Coefficients, PLPCC) (Hermansky, 1991)等。然而，透過這些特徵參數擷取方法所抽取出來的特徵，往往卻極爲容易受到雜訊的干擾而有所影響，而本論文所提出之方法皆是作用於梅爾倒頻譜係數的架構上。

對於以特徵爲基礎的強健性技術而言，由於和其他兩類別的強健性方法比較起來，無論是在做法上或者對於演算法之運算複雜度上，都相對比較簡易且效果十分顯著，因此目前已成功發展出一系列相關之演算法，例如倒頻譜平均值消去法(CMS)、倒頻譜平均值與變異數正規化法(CMVN)與統計圖等化法(HEQ)。基於這三種方法，可消除雜訊所造成的線性失真方法爲倒頻譜平均值消去法和倒頻譜平均值與變異數正規化法，而統計圖等化法則能補償雜訊所造成的非線性失真。本論文將此類方法歸納爲動差正規化法，將與其他種類的特徵正規化法在第二章節中給予詳盡的介紹。

本論文延續統計圖等化法的研究，提出一套新穎的語音特徵正規化技術，其作法是在語音之倒頻譜特徵上，利用一個簡易的差分和平均的處理方式，來得到原始語音特徵之相對應的文脈統計資訊後加以正規化並結合。此新方法的作法有別於傳統之個別維度獨立正規化的統計圖等化法，而是正規化不同空間與時間之間的特徵分布資訊，因此可以更進一步的降低不同聲學環境所產生的偏差，並且嘗試消除傳統之統計圖等化法無法補償的問題，即隨機性雜訊對語音所產生的影響。本論文後續安排如下：第二章節介紹一些著名的運用於時間序列之特徵正規化法的相關研究介紹；第三章則詳細介紹本論文所提出的改良式統計圖等化法，其對應之實驗結果與討論則在第四章節中呈現；最後，第五章節爲結論與未來展望。

2. 運用於時間序列之特徵正規化法的相關研究介紹

2.1 相對頻譜法(Relative Spectral, RASTA) (Hermansky & Morgan, 1994)

觀察人類發音的特性，發現其語音訊號之調變頻譜在低於 1Hz 或高於 12Hz 的範圍是屬於非語音的訊號(Non-Speech)，因此可以使用一個帶通濾波器(Band-Pass Filter)來移除非語音的成分，針對數個音框的特徵參數進行平滑的動作。此濾波器的轉移函數(Transfer Function)如下所示:

$$H_{RASTA}(z) = \frac{0.1\sum_{\theta=1}^{2}\theta(z^{\theta} - z^{-\theta})}{1-\alpha z^{-1}} \tag{1}$$

由式(1)可知此濾波器是由一差量濾波器和一無限長度脈衝響應(Infinite Impulse Response, IIR)之低通濾波器串接而成，當 $z=\alpha$ 時則產生一極點，因此可用參數 α 來控制其頻率響應之峰值所對應的頻率，且當 α 值愈大時，峰值所對應的頻率則變得更小，所以高頻部分的響應則會被壓得更低，而在本論文的辨識實驗中設為 0.94。此外，還有一個位於 0 的零點，可以有效的去除極低頻之慢速變化通道失真效應。

2.2 動差正規化法(Moment Normalization)

如前所述之倒頻譜平均值消去法、倒頻譜平均值與變異數正規化法，通常只需很少量的運算時間即可明顯提升語音辨識的效果因此被廣泛的應用，分別正規化語音特徵參數之第一階動差與第一、二階動差，其公式如下所示:

$$\overline{X}^{d} = \frac{1}{T}\sum_{t=1}^{T} x_t^d, \quad \hat{x}_t^d = x_t^d - \overline{X}^{d}, \quad 1 \le d \le D \tag{2}$$

$$\overline{X}^{d} = \frac{1}{T}\sum_{t=1}^{T} x_t^d, \quad \sigma^{d} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t^d - \overline{X}^{d})^2}, \quad \hat{x}_t^d = \frac{x_t^d - \overline{X}^{d}}{\sigma^{d}}, \quad 1 \le d \le D \tag{3}$$

其中，x_t^d 表示第 d 維的第 t 個音框的語音特徵參數，T 為總音框個數，$\overline{X}^{d}$ 和 σ^{d} 則分別代表第 d 維語音特徵的平均值(Mean)與變異數(Variance)，$\hat{x}_t^d$ 則為其分別正規化後所得之新的特徵。式(2)隱含著經由扣除平均值的處理方式，即可消除通道雜訊所造成的干擾，式(3)除了能消除通道雜訊所造成的影響，還可藉由正規化變異數的動作來降低不同維度間語音特徵分佈的差異程度，因此更能進一步的減少雜訊對語音特徵所造成的干擾。但因方法本身線性關係的限制，因此只能補償雜訊所造成之線性失真的情況，相反的此類方法對於雜訊所造成之非線性失真的情況，其補償效果則是非常有限的。此外另有學者提出正規化語音特徵的第三階動差或更高階的動差(Suk *et al.*, 1999)。

2.3 統計圖等化法(HEQ)

此方法原爲應用於影像處理的領域，以解決數位影像之色彩分佈不均、對比度不平衡等問題 (Acharya & Ray, 2005)。由於統計圖等化法是一種概念簡單且效果顯著的演算法，因此近年來已被廣泛的研究與應用於語音處理的領域中 (De la Torre *et al*., 2005; Hilger & Ney, 2006; Lin *et al*., 2009; Chen *et al*., 2011)。此方法不僅對語音特徵之平均值與變異數做正規化外，還可正規化更高階的動差。目的爲使訓練語料及測試語料的統計分佈特性能夠趨於一致，因此其效果一般而言是優於線性補償技術之倒頻譜平均值消去法及倒頻譜平均值與變異數正規化法。除此之外，容易與大部分之語音特徵表示法(Speech Feature Representation)或強健性方法結合且無需事先對雜訊做任何的假設則爲其附加的優點。主要的作法是將測試語料的機率分佈函數(Probability Distribution Function, PDF)對應至由訓練語料所統計出來的參考分布的機率密度函數，藉由此匹配轉換過程，以降低環境雜訊所造成之測試語料與訓練語料其統計特性不同的現象。

$$\hat{x}_t^d = F_{ref}^{-1}(F(x^d)), \quad 1 \le d \le D \tag{4}$$

式(4)爲統計圖等化法的轉換公式，其中 x^d 爲原始第 d 維的語音特徵參數， $F(x^d)$ 爲 x 之第 d 維的機率分佈且 $F_{ref}(\cdot)$ 爲此相同維度之參考的機率分佈，由此可知此方法是經過 D 次的獨立轉換，所得之 $\hat{x}_t^d$ 爲轉換後之新的語音特徵。

2.4 倒頻譜增益正規化法(Cepstral Gain Normalization, CGN)(Yoshizawa *et al*., 2001)

當乾淨語音受到雜訊影響之後，其雜訊語音特徵之平均值會與原始未受干擾的乾淨語音特徵值之平均值之間產生一個偏移量，同時兩者之間的動態範圍也會因爲受到雜訊的影響而產生不一致的情況，而使得辨識效果變差。因此使用倒頻譜增益正規化法即可消除上述之直流偏移量且正規化特徵參數的動態範圍，其公式如下所示:

$$\overline{X}^d = \frac{1}{T}\sum_{t=1}^{T} x_t^d, \quad \hat{x}_t^d = \frac{x_t^d - \overline{X}^d}{\max(x^d) - \min(x^d)}, \quad 1 \le d \le D \tag{5}$$

其中 max(·) 與 min(·) 分別是求出每一維原始倒頻譜特徵 x^d 之最大值與最小值的函數，而 $\overline{X}^d$ 爲原始每一維之倒頻譜特徵的平均值。

2.5 倒頻譜平均與變異數正規化法結合自動回歸動態平均濾波器法(MVA)

此方法的作法爲上述所提及之倒頻譜平均值與變異數正規化法再結合自動回歸動態平均濾波器(Chen *et al*., 2002)的處理，除了保有倒頻譜平均值與變異數正規化法的優點，利用一個 ARMA 低通濾波器(Low-Pass Filter)可消除因非穩定性雜訊(Non-Stationary Noise)所造成的異常尖峰(Sharp Peak)或波谷(Valley)並達到語音特徵的平滑化(Smoothing)、減緩音框間過度劇烈的快速變化。此外，這樣的結合方式還可進一步的改善調變頻譜平坦化的問題。其公式如下所示:

$$\hat{x}_t^d = \frac{1}{2M+1}\left(\sum_{m=0}^{M} x_{CMVN,t+m}^d + \sum_{m=1}^{M} x_{CMVN,t-m}^d\right), \quad 1 \le t \le T \tag{6}$$

其中 M 爲 ARMA 濾波器的階數，而在本論文的辨識實驗中設爲 2， x_{CMVN} 爲經過倒頻譜平均值與變異數正規化法處理後之特徵。

3. 統計圖等化法使用語音特徵的空間－時間之文脈統計資訊(ST-HEQ)

上述所提及之運用於時間序列上的強健性演算法，雖然已可達到不錯的辨識結果，但由於其並未考慮到對一語音訊號而言，不同頻率成分所造成的影響。如我們所知，對於語音辨識而言，不同頻率成分所占的重要性不盡相同且大部分的語音辨識資訊主要集中於 1Hz 至 16Hz 之間(Kanedera *et al.*, 1997)。因此我們認爲，若能進一步的對一語音訊號之不同頻率成分加以分析、處理，對語音辨識而言將會帶來更大的效益。

在語音訊號處理上，利用離散餘弦轉換(Discrete Cosine Transform, DCT)可得到接近不相關的語音特徵。例如：著名的梅爾倒頻譜特徵就是經由離散餘弦轉換後所得到的。在語音辨識的應用上，爲了要減少運算的複雜度，通常會假設特徵參數彼此間是互相無關(Uncorrelated) 的，但對數頻譜(Logarithmic Spectrum)並不符合此要求，所以將對數頻譜經由離散餘弦轉換後所得的倒頻譜特徵，彼此間的相關性大幅降低，進而較吻合特徵彼此無關的要求。因此大多數傳統的統計圖等化法，當其作用於倒頻譜域(Cepstrm Domain)時，皆是沿著語音特徵之時間序列，進行個別維度之語音特徵獨立正規化的動作，但當不同維度的語音特徵向量不是完全不相關的情形下，此一假設則是無效的。此外從短時間之語音訊號分析的觀點出發，其語音訊號的特性是隨時間緩慢變化的，因此我們假設鄰近串接的語音特徵向量能提供額外有助於正規化的資訊。另一方面，由於傳統的統計圖等化法，皆是假設雜訊對於乾淨語音的干擾是呈現單調(Monotonic)的轉換形式，但隨機性雜訊極可能會使得雜訊對於語音特徵的干擾變成非單調的轉換，此轉變將會導致無法復原的資料損失。

基於上述的觀點，本論文延續傳統統計圖等化法的研究，提出一個新穎的統計圖等化法使用語音特徵的空間－時間之文脈統計資訊；此方法不僅正規化語音特徵之整體(Overall)的統計資訊，更進一步將語音特徵之時間域(Temporal Domain)與空間域(Spatial Domain)上的局部(Local)統計資訊加以正規化。此外本方法的特點爲:突破以往傳統之統計圖等化法只考慮個別維度之統計資訊正規化的問題，並試圖藉由鄰近語音特徵向量所串接而成的文脈統計資訊來改善隨機性雜訊所造成的干擾。所提出之方法的整體概念與作法將在以下做詳細的介紹。

爲了能進一步的得到語音特徵在空間域上之不同頻率成分的文脈統計資訊。本論文將一個簡易的差分和平均的濾波器處理方式，作用於同一個音框 t 之任意兩個相鄰且不同維度的語音特徵 x_t^d 、 x_t^{d-1} ，進而將全頻帶(Full-Band)之統計圖等化法處理後的特徵一分爲二，其公式分別如下所示:

$$x_{s-diff,t}^{d}=\begin{cases}\dfrac{x_t^d-x_t^{d-1}}{2}, & 2\le d\le D\\ x_t^d, & d=1\end{cases} \tag{7}$$

$$x_{s-avg,t}^{d}=\begin{cases}\dfrac{x_t^d+x_t^{d-1}}{2}, & 2\le d\le D\\ 0, & d=1\end{cases} \tag{8}$$

其中 $x_{s-diff,t}^{d}$ 與 $x_{s-avg,t}^{d}$ 分別表示從原始語音特徵 x_t^d 之空間域上所擷取出的高頻(High-Frequency)和低頻(Low-Frequency)的語音成分。隨後將兩個分離出來的頻帶沿著其時間序列的方向，再次利用統計圖等化法來補償雜訊所造成的失真。最後，將這兩個正規化後的頻帶做線性相加的動作成新的語音特徵。相同的，將此處理方式作用於同一個維度之任意兩個相鄰的音框，亦可得到原始語音特徵 x_t^d 在時間域上之高頻 $x_{t-diff,t}^{d}$ 和低頻 $x_{t-avg,t}^{d}$ 的文脈統計資訊，且將這兩個頻帶予以正規化處理也可達到降低雜訊對語音特徵所造成的影響，其所提出之方法流程圖爲圖 2 所示。

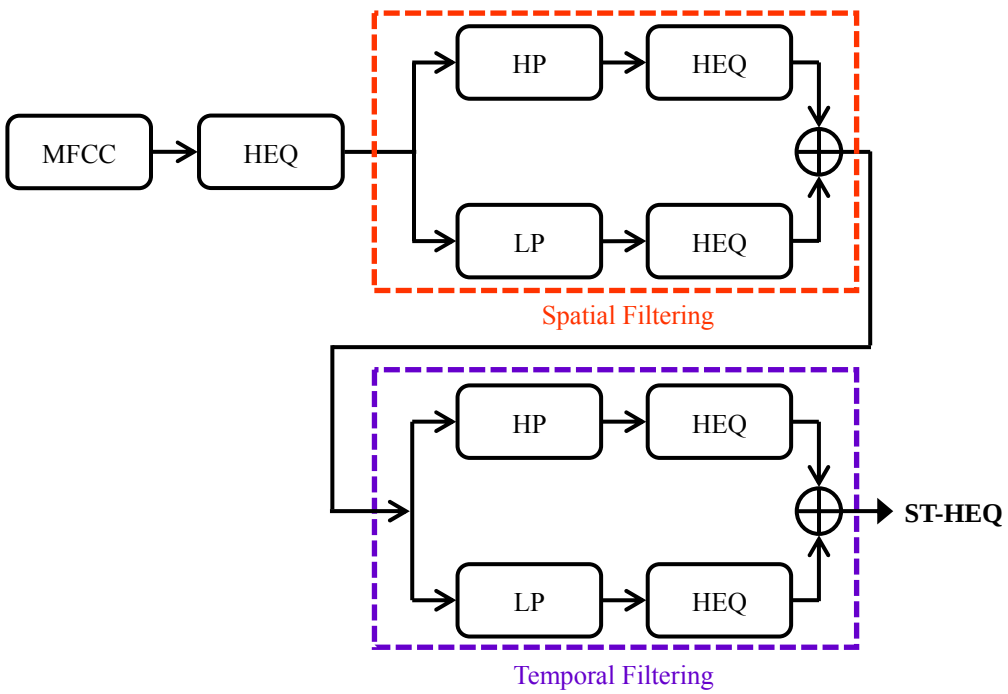


圖2. 結合式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊的流程圖

如前所述，對於語音辨識而言，不同頻率成分所占的重要性不盡相同。因此本論文另外提出了一個變型的 ST-HEQ 法，稱之爲加權式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊(Weighted Spatial-Temporal Contextual Statistics Histogram

Equalization, WST-HEQ)，此方法和 ST-HEQ 方法相比其主要的差別在於，ST-HEQ 於最後結合已正規化的高頻和低頻部分的語音特徵時，是直接將這兩個頻帶的特徵做線性加總的作法，而 WST-HEQ 於重建新的語音特徵時，則是利用一個變數 α 來控制語音特徵之高、低頻成分所占的比重，其重建方法如下所示:

$$\hat{x}_{all} = \hat{x}_{diff} \cdot \alpha + \hat{x}_{avg} \cdot (1-\alpha) \tag{9}$$

其中 $\hat{x}_{diff}$ 與 $\hat{x}_{avg}$ 分別代表已正規化之高、低頻語音特徵，$\hat{x}_{all}$ 爲重建之新的語音特徵且 α 爲一個小於 1 的變數。透過此作法，可進一步探討語音特徵在空間域和/或時間域上之不同頻率成分的重要性。

值得注意的是，對於語音特徵之時間域或空間域上的正規化處理方式，過去已有學者提出概念類似的語音特徵之單一域的正規化處理技術(Hung & Fan, 2009; Joshi *et al*., 2011)，而本論文所提出之結合式統計圖等化法使用語音特徵的空間－時間之文脈統計資訊的技術，於目前爲止則是相對較少被研究與探討的議題。

4. 各種強健性技術之辨識結果與討論

在本章節的一開始，會先介紹本論文在語音辨識實驗上所使用的語音語料庫，並且說明相關的實驗設定與辨識效能評估的方式，隨後會探討本論文所提出之新的語音特徵正規化方法的實驗結果，並與其他常見的特徵正規化方法作比較。此外，我們更進一步的將本論文所提出之方法與著名的 ETSI 進階前端標準(Advanced Front-End Standard, AFE) (Macho *et al*., 2002)做結合，來觀察此結合是否能更進一步地提升系統辨識的精確度。

4.1 實驗語音語料庫

爲了驗證本論文所提出的方法是否爲有效且可行的，因而在語音辨識實驗方面，我們將其作用在國際通用的連續語音語料庫 Aurora-2(ETSI, 2005)上。Aurora-2 本身爲一套含有雜訊的連續英文數字語音語料庫，其內容皆是由美國成年男女所錄製的。爲了評估各種雜訊對於語音的影響，用於測試的語料則分別夾雜了八種不同來源的加成性雜訊和兩種不同特性的通道效應。根據不同雜訊種類的干擾，進一步地將其分成三個測試集：Set A、Set B 與 Set C。其中 Set A 的語料分別含有地下鐵(Subway)、人聲(Babble)、汽車(car)和展覽會館(Exhibition)等四種加成性雜訊與 G.712 通道效應，Set B 的語料則分別含有餐廳(Restaurant)、街道(Street)、機場(Airport)和火車站(Train Station)等四種加成性雜訊與 G.712 的通道效應，此外 Set C 則分別加入了地下鐵(Subway)與街道(Street)兩種雜訊與 MIRS 通道效應。並且依語料中所含之雜訊成分的多寡，而有七種不同的訊雜比(Signal-to-Noise Ratios, SNRs)，分別爲 Clean (∞dB)、20dB、15dB、10dB、5dB、0dB 和 -5dB，其訊雜比的計算公式如下所示:

$$SNR(dB) = 10 \times \log\left(\frac{E_s}{E_n}\right) \tag{10}$$

其中 E_s 爲訊號能量而 E_n 指的是雜訊能量。Aurora-2 語音語料庫提供兩種訓練聲學模型的模式：乾淨情境訓練模式 (Clean-Condition Training) 與複合情境訓練模式 (Multi-Condition Training)，本論文統一使用乾淨語料訓練模式來進行實驗，訓練集的乾淨語句共有 8,440 句，其中並無加成性雜訊，卻包含了 G.712 的通道效應，因此在三個測試集中，訓練集只與測試集的 Set C 有通道上的不匹配。

表1. 語音特徵參數抽取設定

取樣頻率	8000 Hz
音框長度	25 ms
音框位移	10 ms
預強調濾波器	$1-(0.97)z^{-1}$
視窗類型	漢明窗
離散傅立葉點數	256點
濾波器組	共23個梅爾刻度三角濾波器
使用的特徵參數	13維MFCC(C0~C12) +13維ΔMFCC(ΔC0~ΔC12) +13維Δ^2MFCC(Δ^2C0~Δ^2C12)，共39維

4.2 實驗設定

本論文所使用的特徵參數是由 13 維(第 0 維至第 12 維)梅爾倒頻譜係數，加上其一階差量計算(Delta)及二階差量計算(Delta-Delta)，所形成總共 39 維之特徵參數，其詳細的參數設定如表 1 所示。而在訓練聲學模型的部分，則是使用劍橋大學所開發的隱藏式馬可夫模型工具(Hidden Markov Model Tool Kit, HTK)(CUED, n.d.)完成的，包含 11 個數字模型(zero, one, two, …, nine 和 oh)以及靜音模型，其中每個數字模型包含 16 個狀態且每個狀態包含 20 個高斯混合。

4.3 辨識效能的評估方式

本論文對於所有的辨識實驗而言，都是以詞正確率(Word Accuracy)來做爲評估一個演算法是否有效的依據。在此所指的詞(單字詞)則對應到每一個數字，且實驗數據皆是以百分比的方式來呈現，其計算公式爲：

$$\text{詞正確率} = \frac{\text{輸入詞總數} - (\text{取代型錯誤}+\text{插入型錯誤}+\text{刪除型錯誤})}{\text{輸入詞總數}} \tag{11}$$

值得注意的是，根據 Aurora-2 語音資料庫的設定，每一種雜訊的平均詞正確率計算方式是對於 20 dB 至 0 dB 的五種訊雜比詞正確率取平均，而排除乾淨情況與-5 dB 兩種極端的訊雜比的詞正確率，本論文後續的實驗結果之辨識率皆是依循此種呈現方式。

表2. 各種時間序列語音特徵正規化技術與改良式統計圖等化法的辨識率(%)

Method	Set A	Set B	Set C	Avg.
MFCC	54.87	48.87	63.95	54.29
ARMA	60.00	55.94	69.87	60.35
RASTA	67.44	71.90	68.45	69.43
CMS	66.81	71.79	67.64	68.97
CMVN	75.93	76.76	76.82	76.44
HEQ	80.03	82.05	80.10	80.85
CGN	80.08	81.48	80.20	80.66
MVA	80.89	82.00	81.49	81.45
S-HEQ	82.16	84.44	81.12	82.87
T-HEQ	80.42	82.53	80.73	81.33
ST-HEQ	82.52	84.90	81.81	83.33
TS-HEQ	81.85	84.41	81.11	82.72

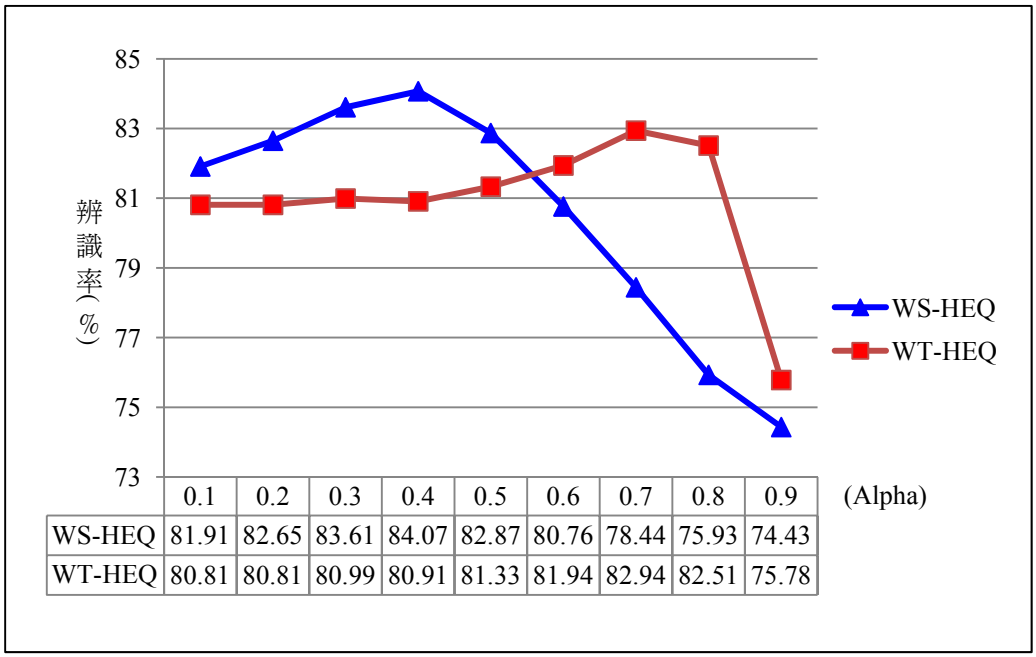
4.4 實驗結果與討論

在這一小節中，首先我們將比較本論文所提出之改良式語音特徵正規化技術(ST-HEQ)與前述所介紹之傳統的時間序列語音特徵正規化法的實驗結果，其結果如表 2 所示。隨後更進一步探討所提出之加權式統計圖等化法使用語音特徵之空間和/或時間之文脈統計資訊，對於語音辨識而言不同頻帶的重要性爲何，其結果如圖 3、4 所示。最後將 ST-HEQ 進一步結合 AFE，以便觀察這樣的結合是否有助於語音辨識率的提升，其結果如圖 5 所示。

根據表 2，我們觀察到:

1. 由於 HEQ 試圖將訓練語料和測試語料之統計分佈特性趨於一致，意謂此方法可正規化語音特徵的所有動差(All Moments)，因此可想而知的，其效果一般而言是優於 CMS、CMVN、CGN 等正規化之動差階數較少的方法。
2. 對統計圖等化法之語音特徵額外再擷取其空間和/或時間之文脈統計資訊，加以正規化後並加總，都有助於辨識效果之提升，其中 S-HEQ 的效果是優於 T-HEQ 且 ST-HEQ 優於 TS-HEQ，而最佳的正規化組合方式爲 ST-HEQ，足足可將原始辨識率從 54.29% 大幅提升至 83.33%，相對錯誤降低率約爲 64%，這顯示了此新方法對於強化語音特徵上有十分顯著的效果。
3. 對於在統計圖等化法處理後的特徵，只額外的正規化空間域上的特徵統計資訊(S-HEQ)會比額外正規化時間域上之特徵統計資訊(T-HEQ)來得好，這結果意謂由於原先已在語音特徵之時間序列域上做過一次 HEQ 處理，若額外的再做一次語音特徵

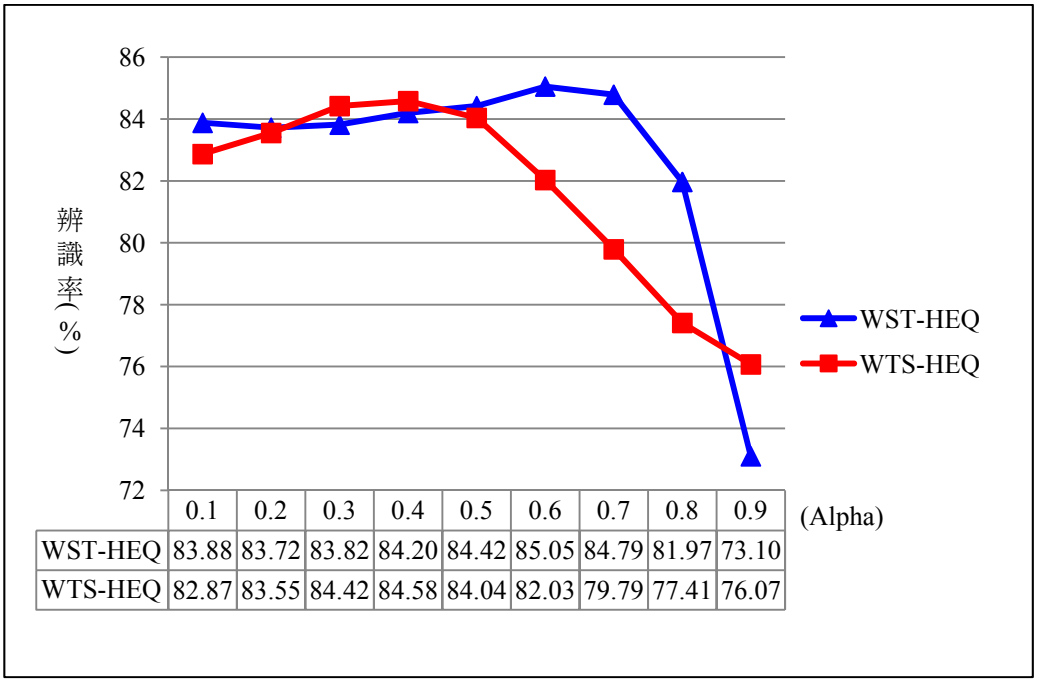
之時間域上的正規化，其效果是有限的。此時反而正規化語音特徵之空間域上的統計資訊會對語音辨識帶來更大的效益。



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	(Alpha)
WS-HEQ	81.91	82.65	83.61	84.07	82.87	80.76	78.44	75.93	74.43	
WT-HEQ	80.81	80.81	80.99	80.91	81.33	81.94	82.94	82.51	75.78	

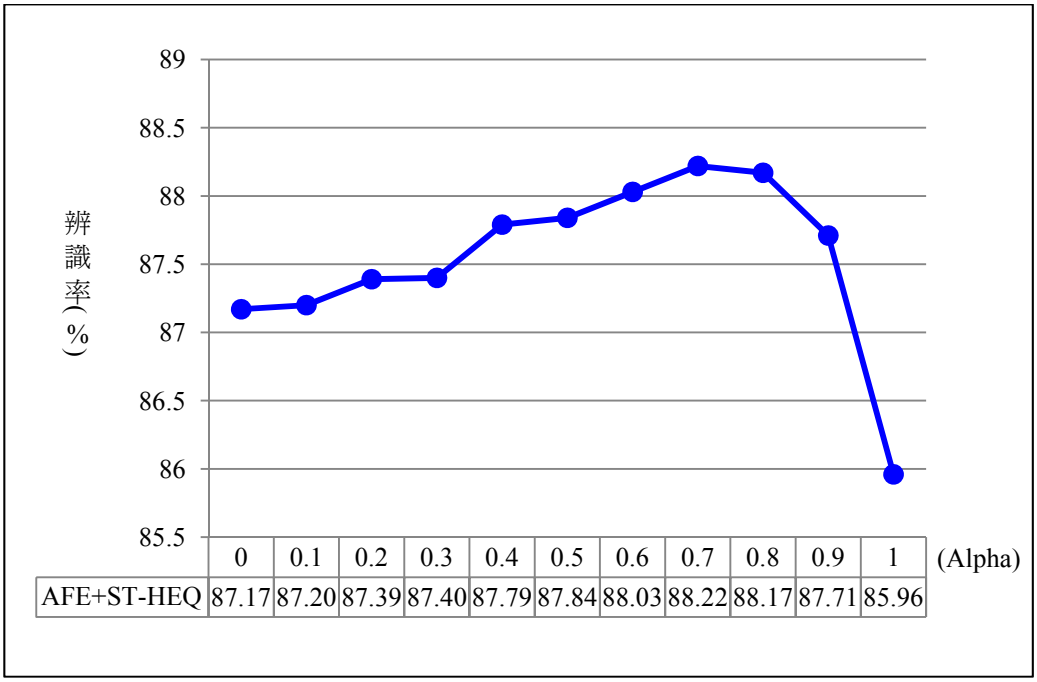
圖 3. 加權式統計圖等化法使用語音特徵的時間或空間之文脈統計資訊的辨識率(%)

圖 3 探討的是加權式統計圖等化法使用語音特徵之空間或時間之文脈統計資訊，根據圖 3 我們發現，對於 WS-HEQ 而言，當其參數 α 設爲 0.3 與 0.4 皆可使辨識率有進一步上升的效果，進步最大的幅度爲當 $\alpha = 0.4$ ，使辨識率從原始未加權的 82.87%上升至 84.07%，其絕對錯誤降低率爲 1.20%。此現象意味著進一步在統計圖等化法之語音特徵空間域上不同頻帶予以正規化，其低頻部分所包含的辨識資訊則比高頻部分所包含的還來得多。此外，對於 WT-HEQ 而言，當其參數 α 設爲 0.6 至 0.8 還可使辨識率有進一步上升的效果，進步最大的幅度爲當 $\alpha = 0.7$ 時，可使辨識率從原始未加權的 81.33%上升至 82.94%。對於此現象，我們所給予的解釋爲:對一個短時段的語音訊號而言，基於其語音能量主要集中於低頻部分的特性，在一般的寬帶雜訊環境下，理所當然的低頻區域的訊雜比會比高頻區域的訊雜比來的高，換句話說，與高頻相比，低頻區域較能提供辨識所需之資訊，相反的高頻則包含較多對辨識而言無用的雜訊成分。但在此由於已將原始語音特徵經過一次全頻帶之 HEQ 的處理即大部分的雜訊成分已被適當的補償，故此時，我們反而需要加重高頻正規化所占的比例，以藉此提高語音辨識的精確度。



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
WST-HEQ	83.88	83.72	83.82	84.20	84.42	85.05	84.79	81.97	73.10
WTS-HEQ	82.87	83.55	84.42	84.58	84.04	82.03	79.79	77.41	76.07

圖 4. 加權式統計圖等化法使用語音特徵的時間和空間之文脈統計資訊的辨識率(%)



	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AFE+ST-HEQ	87.17	87.20	87.39	87.40	87.79	87.84	88.03	88.22	88.17	87.71	85.96

圖 5. 進階前端(AFE)結合改良式語音特徵正規化法(ST-HEQ)之辨識率(%)

圖 4 所呈現的結果是將圖 3 所得到之最佳辨識結果的語音特徵再做一次不同域的語音特徵正規化，例如就 WST-HEQ 而言，其作法是先得到 α 設爲 0.4 的 S-HEQ 的特徵後再做一次 T-HEQ 後並嘗試找出結合後之最佳的 α 值。從圖 4 中我們發現 WST-HEQ 當其參數 α 設爲 0.4 至 0.7 時皆可提升語音辨識的精確度，其中當 $\alpha = 0.6$ 時可得到最佳的辨識結果，此外若以 WTS-HEQ 而言，當 α 設爲 0.2 至 0.5 時也可更進一步的使辨識率有所提升，且最好的參數設定爲當 $\alpha = 0.4$ 時，此外 WST-HEQ 的結果是優於 WTS-HEQ。以上實驗結果皆證明，額外的增加時間域和/或空間域上之文脈統計資訊且適度的調整已正規化之語音特徵的高、低頻成分所占的比例，將有助於語音辨識精確度的進一步提升。

AFE 是一個著名且成效非常好的一種前端之語音特徵擷取技術，從前面的實驗結果得知，本論文所提出之方法的辨識率，沒有優於 AFE，其主要的原因是 ST-HEQ 僅僅是作用於原始 MFCC 上，並沒有額外的做雜訊估測或語音增強的程序。ST-HEQ 作用在原始 MFCC 上只額外的正規化語音特徵之時間域和空間域之文脈統計資訊，如此簡易的作法即可有效的消除雜訊所造成的影響。最後我們試圖將本論文所提出之 ST-HEQ 與 AFE 做結合，來觀察所提出之新方法是否能對 AFE 帶來額外的對語音辨識有用的辨識資訊，其結果如圖五所示。由於在 AFE 的處理程序上有丟棄部分非語音音框的動作(Frames Dropping)，因此我們的結合方式是將本論文所提出之 ST-HEQ 方法直接作用於 AFE 的語音特徵上，並且和 AFE 特徵做線性加權的組合，其結果如圖 5 所示。其中當 α 設爲 1 時則爲 ST-HEQ 直接作用於 AFE 上之辨識結果，且當 α 設爲 0 時則代表 AFE 之辨識結果，從圖 5 我們發現，當 α 設爲 0.1 至 0.9 時相對於原始 AFE 的結果，都能使辨識率有進一步提升的空間，且最好的情況是發生於當 $\alpha = 0.7$ 時，其相對錯誤降低率約有 8%，此結果再次證明了我們所提出之方法的有效性。

5. 結論與未來展望

在本論文中，我們延續統計圖等化法的研究，提出了一套新穎的語音特徵正規化技術，此方法不僅能正規化語音特徵整體的統計資訊，利用一簡易的濾波器處理技術，能更進一步的對語音特徵的空間－時間之文脈統計資訊加以正規化，使得雜訊對語音訊號所造成的影響能夠大幅的降低。此外本方法的特點爲：突破以往傳統之統計圖等化法只考慮個別維度之統計資訊正規化的問題，並試圖藉由正規化鄰近語音特徵向量所串接而成的文脈統計資訊來改善傳統之統計圖等化法無法補償隨機性雜訊所造成的干擾。在國際通用的語音語料庫 Aurora-2 上，我們驗證了所提出之 ST-HEQ 法能夠大幅提升各種雜訊環境下之語音辨識的精確度。此外，我們所提出之 ST-HEQ 其辨識率都明顯高於原始 HEQ、S-HEQ、T-HEQ 和 TS-HEQ。最終更進一步的結合進階式前端標準，實驗結果顯示這樣的結合是有助於辨識效能的提升。在未來的研究中，我們嘗試將本論文所提出之演算法和其他著名的特徵正規化法做結合，來觀察辨識率是否有進一步上升的空間，並且將我們所提出之方法擴展到具有更大詞彙量的語音辨識語料庫上，以便觀察我們所提出之方法在不同複雜度之辨識系統上的效能。

致謝

本論文之研究承蒙教育部－國立台灣師範大學邁向頂尖大學計畫（101J1A0900 和 101J1A0901）與行政院國家科學委員會研究計畫(NSC 101-2221-E-003 -024 -MY3 和 NSC 99 -2221-E-003 -017 -MY3)之經費支持，謹此致謝。

參考文獻

Acharya, T. & Ray, A. K. (2005). *Image processing: principles and applications*, Wiley-Interscience.

Atal, B. S. (1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America*, 55, 1304-1312.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), 133-120.

Chen, B., Chen, W. H., Lin, S. H. & Chu, W. Y. (2011). Robust speech recognition using spatial-temporal feature distribution characteristics. *Pattern Recognition Letters*, 32(7), 919-926.

Chen, C. P., Bilmes, J. & Kirchhoff, K. (2002). Low-resource noise-robust feature post-processing on Aurora 2.0. In *7th International Conference on Spoken Language Processing (ICSLP)*.

Chen, C. & Bilmes, J. (2006). MVA processing of speech features. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), 257-270.

CUED. (n.d.). The hidden Markov model toolkit. Available from: http://htk.eng.cam.ac.uk.

Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366.

De la Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Benitez, M. C. & Rubio, A. J. (2005). Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(3), 355-366.

ETSI. (2005). ETSI standard documentation, “Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithms; back-end speech reconstruction algorithm”, ETSI ES 202 212 ver.1.1.2, 2005.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2), 254-272.

Gauvain, J. L. & Lee, C. H. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2), 291-298.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech communication*, 16(3), 261-291.

Hermansky, H. (1991). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4), 1738-1752.

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4), 578-589.

Hilger, F. & Ney, H. (2006). Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Transactions on Speech and Audio Processing*, 14(3), 845-854.

Hung, J. W. & Fan, H. T. (2009). Subband feature statistics normalization techniques based on a discrete wavelet transform for robust speech recognition. *Signal Processing Letters, IEEE*, 16(9), 806-809.

Hung, J. W., Shen, J. L. & Lee, L. S. (2001). New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination (PMC) techniques. *IEEE Transactions on Speech and Audio Processing*, 9(8), 842-855.

ITU. (1996). ITU-T Recommendation G.729-Annex B: A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.

Joshi, V., Bilgi, R., Umesh, S., Garcia, L. & Benitez, C. (2011). Sub-band level histogram equalization for robust speech recognition. In *12th Annual Conference of the International Speech Communication Association (ICSLP)*.

Kanedera, N., Arai, T., Hermansky, H. & Pavel, M. (1997). On the importance of various modulation frequencies for speech recognition. In *European Conference on Speech Communication and Technology (Eurospeech)*.

Leggetter, C. J. & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9, 171-185.

Lin, S. H., Chen, B. & Yeh, Y. M. (2009). Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1), 84-94.

Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., & Saadoun, F. (2002). Evaluation of a noise-robust DSR front-end on Aurora databases. In *7th International Conference on Spoken Language Processing (ICSLP)*.

Suk, Y. H., Choi, S. H. & Lee, H. S. (1999). Cepstrum third-order normalization method for noisy speech recognition. *Electronics Letters*, 35(7), 527-528.

Viikki, O. & Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(1-3), 133-147.

Yoshizawa, S., Hayasaka, N., Wada, N., & Miyanaga, Y. (2004). Cpestral gain normalization for noise robust speech recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1, 209-212.