改良調變頻譜統計圖等化法於強健性語音辨識之研究

Improved Modulation Spectrum Histogram Equalization for Robust Speech Recognition

高予真 陳柏琳
Yu-Chen Kao and Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{80247004s, berlin}@ntnu.edu.tw

摘要

在自動語音辨識技術的發展上，語音強健性長久以來都是相當重要的研究領域。近年來以調變頻譜的處理和正規化進行強健性語音辨識，已然成為一項活躍的研究議題。調變頻譜統計圖等化法(SHE)是其中一種相當有效的技術，可用以補償調變頻譜因環境干擾而產生的非線性扭曲。在過去研究中，我們改善了調變頻譜統計圖等化法，使其運算複雜度和所需的儲存空間下降，並稱之為多項式擬合調變頻譜統計圖等化法(PSHE)；在此論文中，我們嘗試進一步改進此方法，結合前人的研究中將語音特徵在時域與空間域作分頻的概念，對於語音特徵的高低頻成份分別進行 PSHE 處理並將之結合，嘗試解除原本 SHE 和 PSHE 所依據的語音特徵維度必須獨立和相鄰音框語音特徵無關的兩個假設，將時域與空間域上的文脈資訊列入考慮。本論文的實驗採用 Aurora-2 語料庫進行自動語音辨識實驗；經一系列實驗結果顯示本論文所提出的方法是有實際成效的，能夠顯著地提升語音辨識率。

關鍵詞：調變頻譜，調變頻譜統計圖等化法，強健性語音辨識，多項式擬合，空間域文脈，時域文脈

Keywords: modulation spectrum, spectral histogram equalization, robust speech recognition, polynomial fitting, spatial context, temporal context

一、緒論

目前的自動語音辨識(automatic speech recognition, ASR)系統，在不受各種環境變因干擾的理想錄音環境下，可以得到相當優秀的辨識效果；但在實務應用上，語者的差異、錄音過程產生的噪音、其他環境聲響及通道效應(channel effect)等環境上的變因，會使訓練環境和測試環境間產生環境不匹配(environmental mismatch)的問題，在本論文中也稱為雜訊(noise)。雜訊可以粗略地分成加成性噪音(additive noise)及摺積性噪音(convolutional noise)：加成性噪音即除了實際所需的語音訊號外，系統所接收到的其他聲音，其在時域(time domain)及頻域(spectrum domain)上與原語音訊號是相加的關係，因而得名；摺積性噪音又稱為通道效應(channel effect)，是語音從發聲到接收的過程中經過的各種實體介質及電子設備所造成的扭曲，在時域上與原語音訊號為摺積(convolution)的關係，而在頻域上則與原語音訊號為相乘的關係。

人耳對雜訊有非常優良的強健性(robustness)，這些雜訊對人耳的影響並不大；但對於自動語音辨識系統而言，這樣的不匹配會使語音辨識的正確率(recognition accuracy)大幅降低，需要採用若干強健性語音辨識(robust speech recognition)技術減少環境不匹配

所造成的影響，使自動語音辨識在不同的環境下仍能保有一定的辨識正確率。強健性語音辨識技術依其特性可以大致分為三大類型[1,2]：

1. 以聲學模型為基礎之強健性技術(model-based techniques)：藉由修改已訓練之聲學模型(acoustic model)的模型參數，使聲學模型能夠適應與訓練時不同的環境，從而減少環境不匹配造成的問題。例如經典的最大相似度線性回歸法(maximum likelihood linear regression, MLLR)[3]、平行模型結合法(parallel model combination, PMC)[4]、基於向量泰勒展開式(vector Taylor series)的模型調適[5]等。此類方法通常能對強健性有相當不錯的改善，但所需要的調適語料較多，運算複雜度也較高[1]。
2. 語音強化(speech enhancement)：強化所接收到的語音訊號，使該語音訊號所受到的環境因素干擾減少或消失，從而模擬在理想錄音環境下所取得的語音訊號，藉以降低雜訊的影響。例如經典的頻譜消去法(spectral subtraction, SS)[6]、訊號子空間法(signal subspace approach)[7]、維納濾波器(Wiener filtering)[8]、或是基於統計估測子的語音強化技術[9]等。這一類的方法經常是針對人耳的特性設計，但其引入的非線性扭曲有時會對自動語音辨識系統有負面的影響[10]。
3. 強健性語音特徵擷取(robust speech feature extraction)：藉由改變語音特徵擷取的過程，找出較不會因環境不匹配而改變其特性的語音特徵參數。其中有一部份的方法希望找到一種通用的特徵表示法，使乾淨的語音和受雜訊干擾的語音能表現出類似的特性[11-13]；而另一些方法則是試著運用各種補償的方式，將語音特徵當中受到的干擾還原成未受干擾前的樣子[14,15]。本論文的主要的討論都集中在強健性語音特徵擷取中。

在強健性語音特徵擷取的研究中，其中一個重要的研究領域稱為語音特徵正規化(feature normalization)。這個領域的研究主張將語音特徵序列中的某些特性變為一致，使這種新的語音特徵表示法能較不受雜訊的影響。其中，本論文討論的主要為基於統計分佈的語音特徵正規化(distribution-based feature normalization)，亦即將同一維度的語音特徵序列視為隨機變數(random variable)的一組樣本(sample)，利用這些樣本估計該隨機變數的統計量，據此對特徵序列的分佈進行線性或非線性的轉換。例如基於動差正規化(moment normalization)的倒頻譜平均值減去法(cepstral mean subtraction, CMS)[12]、倒頻譜平均值變異數正規化法(cepstral mean and variance normalization, CMVN)[13]、高階倒頻譜動差正規化法(higher order cepstral moment normalization, HOCMN)[16]，以及可以消除更多非線性環境因素影響的統計圖等化法(histogram equalization, HEQ)[11]等都是此一研究方向的成員。此類的技術大多具有直觀、快速且有效的特性，是強健性語音特徵擷取的領域不可缺少的一環。

許多過去研究[17-19]都說明了統計圖等化法能夠有效地補償非線性的雜訊干擾，而對辨識的正確率有顯著的提升，但統計圖等化法仍然有一些不盡正確的假設。例如其假設語音特徵中各維度間彼此獨立，因而可以對個別維度分別進行正規化，但常見的運用利用離散餘弦轉換(discrete cosine transform, DCT)求取的語音特徵，各維度之間仍具有部份的相關性；而語音是隨時間緩慢變化的訊號，在統計圖等化法中將每一個音框(frame)個別看待的方式也無法有效抓住時域上與前後其他音框的相關性。針對這種比較嚴格的假設，有許多不同的方法被提出，如運用迴歸(regression)技術或時域平均(temporal average, TA)技術引入前後文資訊[20,21]，抑或是將空間(spatial)域及時域的高低頻成份進行正規化，以分頻帶的方式引入文脈資訊(context information)[22,23]。

另外，近年來亦有一些研究顯示，環境中的干擾因素不只會改變語音特徵的分佈特性，也會使語音特徵的時域結構(temporal structure)產生扭曲。調變頻譜(modulation spectrum)[24]為一有效描繪整個語句語音特徵之時域結構的媒介，相較於一般的語音特徵能呈現出更廣泛的語音變化特性。而調變頻譜正規化的研究，便試圖將上述語音特徵分佈特性正規化的概念，應用在語音特徵的調變頻譜上。不同於在時域上語音特徵正規化的技術，調變頻譜正規化技術考慮了語句的整體變化情形，與語音特徵正規化技術採用不同的角度切入環境干擾的問題。類似於語音特徵正規化的研究途徑，調變頻譜平均值正規化法(spectral mean normalization, SMN)及調變頻譜平均值變異數正規化法(spectral mean and variance normalization, SMVN)[25]、調變頻譜統計圖等化法(spectral histogram equalization, SHE)[26]等方法都屬於此一研究領域的成果。另外，也有一些研究根據調變頻譜的特性發展新的正規化方法，例如調變頻譜取代法(modulation spectrum replacement, MSR)[27]、基於濾波器設計的時域序列結構正規化法(temporal structure normalization, TSN)[28]、以及正規化高低頻比例的強度頻譜比例正規化法(magnitude ratio equalization, MRE)[26]等。其中 SHE 所採用的概念與作用於特徵上的 HEQ 類似，但 HEQ 是直接調整特徵的數值，SHE 調整的則是特徵變化的趨勢與規律，此兩種調整標的是不同的，因此具有高度的互補性[29,30]。

有鑑於此，本論文延續以分頻帶的方式引入文脈資訊之研究，提出將其概念應用在調變頻譜統計圖等化法中的「基於空間域－時域文脈統計資訊的調變頻譜統計圖等化法」(ST-PSHE)。利用簡單的高通(high-pass)及低通(low-pass)濾波器取得高頻及低頻的文脈資訊，針對這些文脈資訊進行調變頻譜統計圖等化法，再將正規化後的高低頻成份結合成為新的語音特徵，藉此改善傳統統計圖等化法中的限制，又同時能調整語句的時域結構資訊，也就是特徵變化的規律。在第二章及第三章中，我們將先簡要介紹語音特徵正規化的方法及基於調變頻譜的正規化方法；第四章則詳細說明本論文所提出之改良式架構；接著，實驗的設定、結果與分析將在第五章中呈現，而第六章則為結論與未來可能的研究方向。

二、語音特徵正規化技術

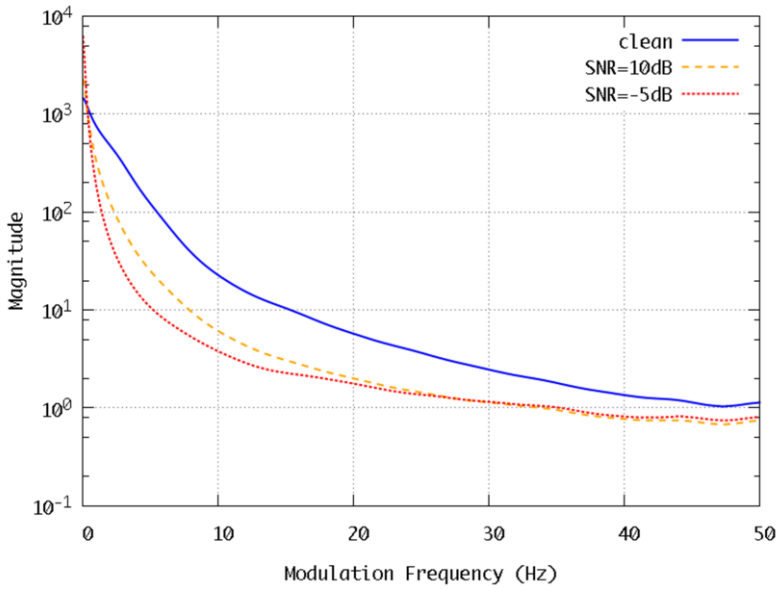
（一）動差正規化法

動差正規化(moment normalization)的技術，主要透過正規化每一個語句(utterance)中各維度特徵統計分佈的動差，來減少雜訊對語音特徵的影響。例如倒頻譜平均值減去法[12](下稱 CMS)希望藉由將每一個語句的第一階動差(first-order moment)，也就是期望值減去，來減少雜訊的影響；而倒頻譜平均值變異數正規化法[13](下稱 CMVN)則更進一步將正規化的範圍擴展至第二階動差，使不同語句間的變異數(variance)也變得一致。令一語句中，某一維度的語音特徵時間序列為$\{x[n]\}$，μ為$\{x[n]\}$的期望值，σ^2為其變異數，則經此兩個方法正規化過的特徵分別可以表示為：

$$\hat{x}_{\mathrm{CMS}}[n] = x[n] - \mu \tag{1}$$

$$\hat{x}_{\mathrm{CMVN}}[n] = \frac{x[n] - \mu}{\sigma} \tag{2}$$

由於通道效應在倒頻譜(cepstrum)上與原本的語音訊號為相加的關係，CMS 的正規化可以有效地消去一些穩定(stationary)的通道效應，而使得語音辨識的正確率有相當明顯的改善。另一方面，CMVN 對變異數的正規化，更進一步地補償了不同語句的語音



圖一、Aurora-2 語料庫中不同訊噪比語句 MFCC 特徵 c1 參數之調變頻譜的差異

特徵間因為雜訊而產生的動態範圍(dynamic range)差異，使得雜訊對語音特徵的影響更為縮小。在這些基礎之下，也有學者提出正規化語音特徵的第三階動差或任意階數的動差的技術[16]。

（二）統計圖等化法

統計圖等化法為影像處理領域常用的演算法，用以調整如明度、色彩平衡等影像參數[31]；而在自動語音辨識的領域，也有學者提出利用統計圖等化法來補償雜訊在語音特徵上造成的失真，許多研究也證明了它的有效性[18,32-35]。前一節所介紹的 CMS 與 CMVN，乃至於更高階動差的正規化方法，均是以線性(linear)的方式補償雜訊對語音特徵的干擾，但對於非線性的扭曲補償效果有限，統計圖等化法則彌補了動差正規化法的此一缺失。相較於動差正規化法，統計圖等化法不對動差進行正規化，而是利用一非線性(non-linear)的轉換，將所有語音特徵的統計分佈直接變得與未受雜訊干擾時的統計分佈一致，並且無需對該統計分佈擁有先驗知識(prior knowledge)，即可有效地改善雜訊語音的辨識正確率。

統計圖等化法主要的做法，是將目前語句中特徵分佈的累積密度函數(cumulative distribution function, CDF)，對應至由訓練語料所統計出來的參考分布，藉此將整句話的特徵還原至與訓練語料相同的統計分佈。令$F(\cdot)$為目前語句語音特徵時間序列$\{x[n]\}$的機率分佈(以一個將值對應到 CDF 的函數表示)，而$G(\cdot)$為根據所有訓練語料統計出的參考分佈，統計圖等化法正規化後的語音特徵可以表示為：

$$\hat{x}_{\mathrm{HEQ}}[n] = G^{-1}\big(F(x[n])\big) \tag{3}$$

傳統的統計圖等化法通常以查表法(table lookup)描述$G(\cdot)$函數的對應關係，但這樣的方法不僅較費時，也需要花費許多空間來記錄表格。在[33]中，我們提出利用一多項式函數來逼近$G^{-1}(\cdot)$，可以降低計算時間與儲存空間，同時獲得比原始的 HEQ 相似或較佳的辨識正確率。此方法稱為多項式擬合統計圖等化法(polynomial-fit histogram equalization, PHEQ)，本論文中之統計圖等化法皆以此方式實作，如下式所示：

$$\hat{x}_{\mathrm{PHEQ}}[n] = G^{-1}\big(F(x[n])\big) = \sum_{m=0}^{M} a_m \big(F(x[n])\big)^m \tag{4}$$

（三）基於濾波器的正規化技術

除了在統計分佈上進行處理外，也有一些語音特徵正規化的方法試圖從濾波器的設計出發。例如相對頻譜法(relative spectra, RASTA)[36]便是利用人類語音主要資訊集中在特定調變頻譜頻帶的原理，設計一帶通濾波器(band-pass filter)，藉以移除語音特徵中與語音較不相關的成份；而在[37]中，則是使用低通濾波器(low-pass filter)對特徵進行平滑化(smoothing)，以降低語音特徵中不穩定或突發的雜訊對語音特徵造成的干擾。值得一提的是，式(1)也可以視為是一個高通濾波器(high-pass filter)的脈衝響應(impulse response)，因此從另一個角度來解讀，CMS 亦是利用濾波的概念來移除穩定通道效應的一種技術。

三、調變頻譜於強健性語音辨識之研究

（一）調變頻譜之定義與特性

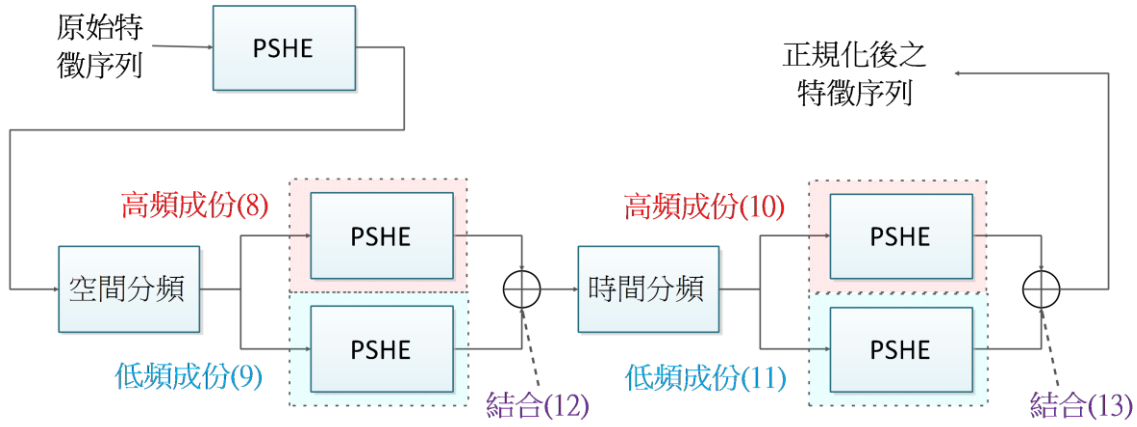
令一語句中，某一特定維度之語音特徵時間序列為$\{x[n]\}$，其中 n 為音框(frame)的索引值，該語音特徵序列的調變頻譜可以定義為：

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{\frac{-kn2\pi i}{N}} \tag{5}$$

其中$i = \sqrt{-1}$為虛數單位，其中k為調變頻率的索引，N 為語句中音框的總數，所得之序列$\{X[k]\}$即為$\{x[n]\}$的調變頻譜。式(5)可以視為一離散傅立葉轉換(discrete Fourier transform, DFT)，調變頻譜中的頻率範圍與語音特徵時間序列之取樣率有關：在本論文的基礎語音特徵設定中，每兩個相鄰音框之間隔為 10ms，亦即語音特徵時間序列之取樣率為 100Hz，根據奈奎斯特定理(Nyquist–Shannon sampling theorem)[38]，調變頻譜之最高頻率為 50Hz。

調變頻譜在分析語音特徵之時域結構上，是很有用的工具；過去有研究[39]指出，調變頻率大約 1Hz 到 16Hz 間的低頻成份，與語音辨識的正確率有明顯的關聯，而其中以 4Hz 附近所包含的資訊最為重要。關於人類聽覺的研究[40]也不約而同地發現：4Hz 的調變頻率在人類的聽覺感知中佔有很重要的地位。

當語音訊號受到雜訊干擾時，不只其語音特徵時間序列的分佈特性會改變，其時域結構也會有一定程度的扭曲，亦即使其調變頻譜產生失真。一些過去針對調變頻譜的研究[25,30]發現，語音訊號受到環境干擾的影響越劇烈，亦即訊噪比(signal-to-noise ratio, SNR)越低的時候，調變頻譜中對語音辨識最重要的 1Hz 到 16Hz 成份強度越受到壓抑，而偏離乾淨狀況的調變頻譜越遠。舉例來說，圖一是 Aurora-2 語料庫所有測試集梅爾倒頻譜系數(Mel-frequency cepstral coefficients, MFCC)[41]中 c1 系數的調變頻譜。由於除了環境干擾外，尚有個別語者的差異等因素，因此此圖採用測試集中所有句語句調變頻譜之平均值，以突顯環境條件的不同，降低個別語句差異造成的影響。從此圖中可以觀察到，當訊噪比降低時，整個調變頻譜的所有頻帶都會產生失真，尤其以包含最多語音內容資訊的頻帶為甚。



圖二、ST-PSHE 流程示意圖

（二）調變頻譜之正規化

調變頻譜正規化的相關技術，旨在使受到環境干擾而扭曲的調變頻譜恢復為未受干擾的樣貌。針對強健性語音辨識正規化調變頻譜的過程大致上可以如下三個步驟說明：

1) 分析：將受到環境干擾的整句語句之語音特徵時間序列$\{x[n]\}$進行離散傅立葉轉換，得到該語句的調變頻譜$\{X[k]\}$。以離散傅立葉轉換取得之序列為一複數序列，可再分解成該調變頻譜的強度頻譜$\{|X[k]|\}$及相位頻譜$\{\angle X[k]\}$。

2) 正規化：針對前一步驟所得到的強度頻譜及相位頻譜進行處理。其中相位頻譜通常維持原狀，僅改變強度頻譜中的強度，並得到新的強度頻譜$\{|Y[k]|\}$。

3) 還原：依據原本的相位頻譜$\{\angle X[k]\}$和第二步驟中所得之新的強度頻譜$\{|Y[k]|\}$，進行反離散傅立葉轉換(inverse discrete Fourier transform, IDFT)，取得還原後的語音特徵時間序列。

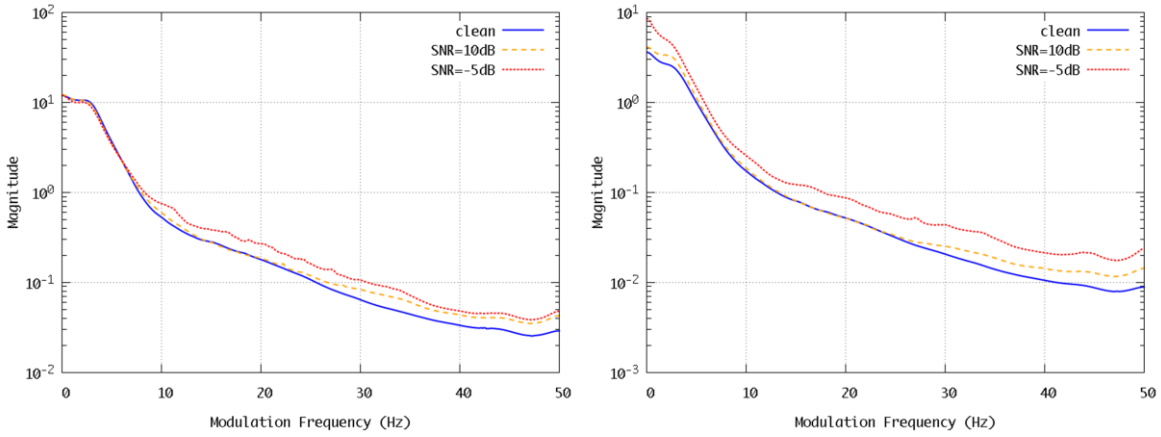
若上述第二步驟中的強度頻譜能夠被適當地正規化，則可以有效降低環境干擾對調變頻譜的失真，進而使還原後的語音特徵參數，在自動語音辨識系統中得到較好的辨識精確率。以下將簡述數種調變頻譜正規化的方法：

1. 強度頻譜比例正規化法(magnitude ratio equalization, MRE)

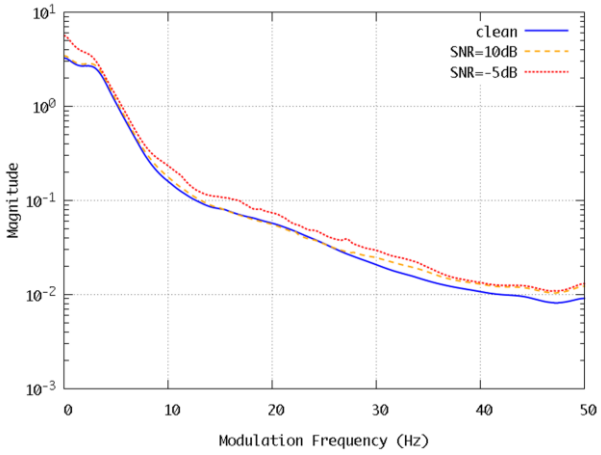
此技術[26]計算調變頻譜中低頻成份強度和高頻成份強度的比例，在語句受到環境干擾時，將此比例調整回未受干擾情況下的比例。由於調變頻譜受環境干擾時「低頻下降，高頻抬升」的現象十分明顯，若能找到高頻成份和低頻成份間適當的界線，此方法能有不錯的成效，且運算十分快速。

2. 調變頻譜統計圖等化法(spectral histogram equalization, SHE)

在第二節中所介紹的統計圖等化法為影像處理領域常用的演算法，亦有在在語音特徵時間序列分佈正規化之應用[11]。此技術[26]將統計圖等化法應用在調變頻譜強度的正規化上，利用一非線性的轉換(non-linear transform)，使得可能受環境干擾的測試語句調變頻譜分佈，趨向於乾淨訓練語句的調變頻譜分佈。令$F(\cdot)$為目前語句語音特徵時間序列



(a) (b)



(c)

圖三、c1 特徵的調變頻譜受雜訊干擾之情形：(a)僅經 PSHE 處理後的全頻帶特徵之調變頻譜 (b)僅經 PSHE 處理後的空間高頻成份之調變頻譜 (c)進一步經分頻處理後的空間高頻成份之調變頻譜

$\{x[n]\}$之調變頻譜強度$\{|X[k]|\}$的機率分佈，$G(\cdot)$為所有訓練語料的調變頻譜強度機率分佈，也就是參考分佈，此方法中正規化後的頻譜強度$|Y[k]|$與原始頻譜強度$|X[k]|$的關係為：

$$|Y[k]|_{\mathrm{SHE}} = G^{-1}\big(F(|X[k]|)\big) \tag{6}$$

由於此方法為非線性轉換，可以較完整地使測試語句的調變頻譜分佈趨近訓練語料的調變頻譜分佈。但其使用查表法(table-lookup)記錄訓練語料的調變頻譜分佈，需要記錄相當大量的資料方能較完整地逼近真實的分佈；且查表法相當於使用一系列的分段線性函數(piecewise linear function)來逼近真實分佈，但調變頻譜的分佈中，並非所有區域都適合使用線性函數進行逼近，如此勢必使所需的記錄點數大輅增加。在[29]中，我們提出多項式擬合調變頻譜統計圖等化法(polynomial-fit spectral histogram equalization, PSHE)利用一個多項式函數逼近$G^{-1}(\cdot)$，其能夠有與原本的 SHE 相近的辨識正確率，但同時大輅降低空間需求與計算時間。本論文中所使用之 SHE 皆為此多項式版本的 PSHE，此方法可以下式來表示：

$$|Y[k]|_{\mathrm{PSHE}} = G^{-1}\big(F(|X[k]|)\big) = \sum_{m=0}^{M} a_m \big(F(|X[k]|)\big)^m \tag{7}$$

四、基於空間域－時域文脈統計資訊的調變頻譜統計圖等化法

為了改進統計圖等化法中「獨立看待每一個音框」及「獨立處理各別維度」的嚴格假設，在[22]中，我們提出了使用語音特徵的空間域－時域文脈統計資訊的統計圖等化法，或稱為 ST-PHEQ。此技術將語音特徵在空間域及時域上分別分成高頻成份及低頻成份，分別進行統計圖等化法後，再將這些成份重新結合形成新的語音特徵，藉此取得時域上及空間域上的文脈資訊。

在本論文中，我們嘗試將這樣的概念運用在調變頻譜的統計圖等化法，稱為「基於空間域－時域文脈統計資訊的調變頻譜統計圖等化法」，或簡稱 ST-PSHE。其中 S 代表空間域(spatial)，T 代表時域(temporal)，而 PSHE 則為前一章所述的多項式擬合調變頻譜統計圖等化法。ST-PSHE 的流程如圖二所示。首先，為了得到高頻與低頻的成份，我們使用一組簡易的差分(differencing)及平均(averaging)濾波器，分別擷取語音特徵的高頻成份及低頻成份的特徵序列，如下表所示：

	高頻(差分)	低頻(平均)
空間域	$x_d^{s,\mathrm{hp}}[n]=\begin{cases} x_d[n] & ,\text{if } d=1 \\ \dfrac{x_d[n]-x_{d-1}[n]}{2} & ,\text{otherwise} \end{cases}$ (8)	$x_d^{s,\mathrm{lp}}[n]=\begin{cases} 0 & ,\text{if } d=1 \\ \dfrac{x_d[n]+x_{d-1}[n]}{2} & ,\text{otherwise} \end{cases}$ (9)
時域	$x_d^{\mathrm{t,hp}}[n]=\begin{cases} x_d[n] & ,\text{if } n=1 \\ \dfrac{x_d[n]-x_d[n-1]}{2} & ,\text{otherwise} \end{cases}$ (10)	$x_d^{\mathrm{t,lp}}[n]=\begin{cases} 0 & ,\text{if } n=1 \\ \dfrac{x_d[n]+x_d[n-1]}{2} & ,\text{otherwise} \end{cases}$ (11)

其中$x_d[n]$為該語句中第n個音框第d維度的語音特徵值，$n=1$及$d=1$代表第一個音框及第一個維度，依此類推；$x_d^{s,\mathrm{hp}}[n]$、$x_d^{s,\mathrm{lp}}[n]$、$x_d^{\mathrm{t,hp}}[n]$及$x_d^{\mathrm{t,lp}}[n]$則分別代表空間域高頻、空間域低頻、時域高頻、時域低頻的子頻帶成份特徵。

對於每一個語句，在進行了一次 PSHE 之後，其全頻帶(full-band)的調變頻譜已經具有和訓練語料的調變頻譜相同的分佈，但時域或空間域上的高低頻成份卻還是有一部份的不匹配現象。因此在進行 PSHE 以後，要將處理後的特徵依式(8)及式(9)在空間域上分為高頻特徵與低頻特徵，將此兩個頻帶的特徵分別求取其調變頻譜並以 PSHE 正規化並由調變頻譜還原回特徵域之後，再依下式將空間域高低頻成份結合：

$$\hat{x}_d[n]=\hat{x}_d^{\mathrm{s,hp}}[n]+\hat{x}_d^{\mathrm{s,lp}}[n] \tag{12}$$

其中$\hat{x}_d^{\mathrm{s,hp}}[n]$為空間域高頻成份經 PSHE 正規化後之特徵，$\hat{x}_d^{\mathrm{s,lp}}[n]$則為空間域低頻成份經 PSHE 正規化後之特徵。由於式(8)與式(9)的設計使得此兩個頻帶具有互補關係，故將兩個頻帶的特徵直接相加即可還原回原本全頻帶的特徵。進行完空間域上的分頻正規化以後，將結合後的全頻帶特徵再次依據式(10)及式(11)在時域上分為高頻特徵與低頻特徵，同樣將此二頻帶分別進行 PSHE 後，利用與空間域高低頻結合相同的方式，依下式所示將時域之高低頻成份結合：

$$\tilde{x}_d[n]=\tilde{x}_d^{\mathrm{t,hp}}[n]+\tilde{x}_d^{\mathrm{t,lp}}[n] \tag{13}$$

其中$\tilde{x}_d^{\mathrm{t,hp}}[n]$為時域高頻成份經 PSHE 正規化後之特徵，$\tilde{x}_d^{\mathrm{t,lp}}[n]$則為時域低頻成份經 PSHE 正規化後之特徵，經過此一過程產生最終經 ST-PSHE 處理後的特徵。其中，亦可以選擇跳過時域分頻的部份(稱為 S-PSHE)、跳過空間域分頻的部份(稱為 T-PSHE)、或是將時域分頻與空間域分頻兩部份調換順序(稱為 TS-PSHE)，此部份的差異將於第五章中探討。

表一、各種基礎特徵及強健性技術的辨識正確率(%)

特徵	訊噪比							平均值
	乾淨	20dB	15dB	10dB	5dB	0dB	-5dB	
MFCC	99.71	92.44	80.56	58.61	30.04	9.31	3.39	54.19
CMS	99.72	98.13	94.27	80.45	50.64	23.81	13.04	69.46
CMVN	99.69	97.97	94.98	87.25	67.52	34.87	13.73	76.52
MVA	99.66	97.96	95.98	90.27	76.46	50.70	22.86	82.27
PHEQ	99.65	98.52	96.56	91.19	75.78	45.39	18.14	81.49
ST-PHEQ	99.58	98.59	96.99	92.26	78.95	50.36	20.04	83.43
PSHE	99.47	97.55	94.29	86.54	68.54	37.58	16.09	76.90
CMVN+PSHE	99.56	98.38	96.59	92.26	80.63	56.24	26.93	84.82
PHEQ+PSHE	99.45	98.39	96.61	92.71	82.05	58.75	28.34	85.70

表二、PSHE 結合空間域或時域文脈資訊的辨識正確率(%)

特徵	訊噪比							平均值
	乾淨	20dB	15dB	10dB	5dB	0dB	-5dB	
S-PSHE	99.39	97.29	93.69	85.73	68.77	40.17	17.75	77.19
T-PSHE	99.41	97.31	93.78	85.90	68.89	40.18	17.68	77.21
TS-PSHE	99.45	97.10	93.44	85.50	68.65	39.96	17.13	76.93
ST-PSHE	99.28	97.28	94.21	86.70	69.48	40.06	17.71	77.55

在傳統的 HEQ 或是 SHE 中，都假設雜訊對於語音只具有單調(monotonic)的的干擾，亦即會改變特徵或調變頻譜中所有數值的大小，但各數值之間的相對排序(ordering)是維持不變的。ST-PSHE 除了打破時域及空間域上的獨立假設以外，此種將高低頻分別正規化再結合的方式也可能會改變調變頻譜不同頻率強度的大小順序，而使得非單調的干擾能夠一併被考慮進來。有鑑於此，在訓練階段統計時域分頻部份的參考分佈時，需要使用空間域分頻部份已經正規化過的語音特徵進行統計，而非原始未經正規化的語音特徵。

值得注意的是，本論文中時域分頻的方法，其概念與前人針對 SHE 所提出的分頻處理類似，並具有相仿的成效：在[25]中，調變頻譜被依等比音程(octave)的比例分為若干個頻帶，越低頻的成份越加細分，並針對每一個頻帶進行獨立的 SHE 處理；而在[30]中，調變頻譜被畫分為兩個頻帶獨立進行 SHE 處理，而劃分的頻率則為可調整之參數。在此兩種技術中，對頻帶的畫分都是直接將某個特定頻率以下及以上的成份畫分為不同的頻帶；然而本論文中進行分頻的濾波器在高頻帶與低頻帶之間有重疊，在高低頻之間沒有一個確切的分割點，將高低頻結合後也不會產生明顯的不連續現象。另外，本論文中分頻的濾波器為有限脈衝響應(finite impulse response, FIR)濾波器，分頻的過程不需轉換至調變頻譜，可直接在特徵上快速並穩定(numerical stability)地進行實作。

表三、ST-PSHE 與其他強健性技術結合之辨識正確率(%)

特徵	訊噪比							平均值
	乾淨	20dB	15dB	10dB	5dB	0dB	-5dB	
CMVN+ST-PSHE	99.45	98.44	96.82	92.8	82.01	58.44	29.39	85.70
PHEQ+ST-PSHE	99.41	98.28	96.59	92.44	82.03	59.13	29.32	85.69
ST-PHEQ+ST-PSHE	99.37	98.12	96.42	92.28	82.16	60.08	30.98	85.81

表四、ST-PSHE 與 AFE 比較及結合的辨識正確率(%)

特徵	訊噪比							平均值
	乾淨	20dB	15dB	10dB	5dB	0dB	-5dB	
AFE	99.74	98.89	97.68	94.27	85.47	62.54	30.26	87.77
AFE+ST-PSHE	99.70	98.82	97.64	94.28	85.89	63.86	32.22	88.10

五、實驗與分析

（一）實驗語料庫

本論文的實驗所使用的語料庫為 Aurora-2 英文連續數字語料庫[42]，此語料庫由歐洲電信標準協會(European Telecommunications Standards Institute, ESTI)所發行，內容皆是由美國成年人錄製的連續數字。此語料庫包含 G.712 和 MIRS 兩種不同的通道效應，及機場、人聲、汽車、展覽會館、餐廳、地下鐵、街道、火車站等八種加成性噪音，加成性噪音分別以乾淨、20dB、15dB、10dB、5dB、0dB、-5dB 等七種不同的訊噪比混入語音中。此語料庫含有兩組不同的訓練語料，分別有 8,440 句的訓練語句。在乾淨訓練(clean-condition training)語料中，所有語句皆乾淨不含任何噪音；而在複合情境(multi-condition training)訓練語料中，含有及地下鐵、人聲、汽車、展覽會館等四種噪音，其訊噪比由 5dB 到 20dB 外加乾淨語音，兩組訓練語料皆含 G.712 通道效應。本論文中的實驗一律使用乾淨訓練語料進行訓練。

在測試語料部份，訊噪比範圍皆是由-5dB 到 20dB 外加乾淨語音。測試集 A 有 28,028 句，分為四個子集，含有和複合情境訓練語料中相同的噪音和通道效應；測試集 B 有 28,028 句，分為四個子集，含有餐廳、機場、街道、火車站等四種噪音，以及和訓練語料相同的通道效應；測試集 C 有 14,014 句，分為兩個子集，含有地下鐵和街道兩種噪音，通道效應為 MIRS。由於本論文使用乾淨訓練語料，所有加成性噪音皆是訓練語料中未曾見過，而只有測試集 C 的通道效應與訓練語料不同。

（二）基礎實驗設定

本論文的基礎實驗是採用梅爾倒頻譜係數[41]做為語音特徵參數，其中預強調(pre-emphasis)參數設為 0.97，窗函數(window function)為漢明窗(Hamming window)，其參數設為 0.46，取樣音框長度為 25 毫秒，音框間距(frame shift)為 10 毫秒。每個音框內的資訊，在完成特徵擷取以後由 39 維的語音特徵向量表示。其中前 13 維為梅爾倒頻譜係數的前 12 項(c1~c12)及第零倒頻譜係數(c0)，14 維到 26 維為前 13 維的一階差量係數

(delta coefficient)，最後 13 維則為前 13 維的二階差量係數(acceleration coefficient)。本論文的實驗中，擷取特徵的過程共使用 23 組梅爾濾波器(Mel filter)。

評估語音特徵所使用的聲學模型訓練及辨識，皆使用 HTK 套件[43]完成。其中每個數字皆由一個由左到右形式的連續密度隱藏式馬可夫模型(continuous density hidden Markov model, CDHMM)表示，每個模型扣除前後之銜接用狀態(state)共有 16 個狀態，每個狀態以含 20 個高斯混合(Gaussian mixture)的高斯混合模型(Gaussian mixture model, GMM)表示。靜音(silence)模型則為 3 個狀態和 36 個高斯混合。

（三）辨識效能評估方式

本論文辨識效能評估的方法採用美國標準與科技組織(The National Institute of Standards and Technology, NIST)所訂定之用以評估轉譯文句與正確文句比較的標準。評估的指標為詞正確率(word accuracy)，計算方式如下：

$$\text{詞正確率} = \frac{\text{詞正確辨識個數} - \text{詞插入個數} - \text{詞刪除個數}}{\text{此句中詞的總數}} \tag{14}$$

另外，本論文中靜音詞(silence 和 short pause)將不列入詞正確率的計算。而在 Aurora-2 語料庫的設定中，每一個測試子集的平均辨識率，只以 0dB(含)到 20dB(含)間的辨識精確率計算平均。本論文亦以此計算方式評估辨識效能。

（四）實驗結果與討論

首先，作為比較的基準，我們在表一中列出了 MFCC 特徵及一些基礎強健性語音辨識技術的辨識正確率。其中 PHEQ 及 PSHE 的多項式階數均是根據 Aurora-2 語料庫進行挑選之最佳設定值，本論文後續實驗皆依循此組設定，而不另行最佳化多項式階數。而由表一中也可以發現：由於 PHEQ 非線性轉換的特性，比起使用線性轉換的 CMS 及 CMVN 能夠補償更多雜訊造成的干擾，在辨識正確率上有較好的表現，而同樣引入時域及空間域文脈資訊進行分頻的 ST-PHEQ，相較於原本的 PHEQ 亦有大幅的改進，顯示這些文脈資訊對於語音辨識的強健性有巨大的幫助。

而在調變頻譜的正規化方面，雖然單獨使用 PSHE 沒有太突出的表現，但由於 PSHE 正規化的是整個語句中特徵變化的趨勢與規律，與其他直接調整語音特徵數值的方法(如 CMVN 與 PHEQ)具有良好的互補性[29]。進一步將 PSHE 運用在經 CMVN 或 HEQ 正規化後的特徵上，可以獲得相當突出的成果，其效能甚至高於 ST-PHEQ。依這樣的結果來看，顯然使用調變頻譜這種描述語句整體變化資訊的表示法是有其重要性的。另外，在雜訊的干擾相當嚴重的環境下(如訊噪比為-5dB 的情況)，應用 PSHE 後，其改善的幅度多於在所有環境下的平均情況，甚至在同時應用 HEQ+PSHE 的情況下，訊噪比 -5dB 的辨識正確率高達原始 MFCC 特徵的 8 倍以上。此結果說明了調變頻譜確實能捕捉到一些無法直接透過正規化語音特徵改善的問題，尤以在雜訊較強時為甚。

本論文所提出的方法，其實驗結果則列在表二中。與原本的 PSHE 相較，針對其正規化後的特徵進行分頻帶的正規化，無論以何種順序組合時域與空間域兩個元素，都能取得更好的結果，這顯示了 PSHE 雖然能夠使調變頻譜上的分佈變得一致，但在時域與空間域高低頻成份的調變頻譜中仍然存在著一些未被消除的干擾，藉由將這些成份也納入正規化的範圍，可以補足 PSHE 這一點不足之處。在圖三中，我們以空間域高頻成份

為例，顯示了即使 PSHE 已將全頻帶特徵的調變頻譜變得較為一致，在子頻帶特徵的調變頻譜中，仍然存在著因為雜訊而產生的失真；而這個失真在經過 ST-PSHE 的處理以後，則有顯著的改善，並達到跟全頻帶的調變頻譜相近的一致程度。另外，單獨在空間域上或是時域上進行分頻的正規化，都能夠相對地減少大約 1.3%的字錯誤率(word error rate)，而依照空間域－時域的順序進行分頻正規化，更能夠相對減少 2.8%的錯誤。但若將順序反過來，依照時域－空間域的順序進行，則改進的幅度反而變得非常有限。

前文中提到在調變頻譜上的正規化方法，若與在特徵時域上的正規化方法結合，會產生很明顯的互補效應，而使辨識率大幅上升。因此在表三當中，我們也嘗試將 ST-PSHE 與 CMVN、HEQ 以及同樣應用時域及空間域文脈資訊進行分頻的 ST-PHEQ 進行結合，探索與這些方法結合的效果。由於調變頻譜雖然抓住了整個語句的特徵變化模式，但對於比較區域性的雜訊干擾及個別音框的扭曲則較難詳盡地描述，因此若能在進行 ST-PSHE 前先利用特徵上的正規化方法 CMVN 及 HEQ 處理過，則能同時正規化整體變化模式及個別音框的數值，與單純處理調變頻譜相較，可以取得超過 36%的相對字錯誤率減少。而若在進行 ST-PSHE 之前先使用 ST-PHEQ 處理過一次，雖然同樣是運用分頻取得文脈的概念進行，但由於處理的面向不同，因此仍然有很大的互補成份存在，其結果較單獨使用 ST-PSHE 相對減少了 36.8%的辨識錯誤，與 ST-PHEQ 比較也相對降低了 14.4%的字錯誤率。

最後，我們也將本論文所提出的方法與歐洲電信協會(European telecommunications standards institute, ETSI)發展的 AFE (advanced front end)[44]進行比較。如表四所示，由於 AFE 包含了較複雜的語音活動偵測(voice-activity detection, VAD)及噪音抑制(noise reduction)的技術，AFE 的辨識正確率相較於 ST-PSHE 明顯是較好的；但進一步將 AFE 的特徵施以 ST-PSHE 的處理，並將之與原本的 AFE 特徵線性結合之後，仍然能夠相對地減少大約 2.7%的辨識錯誤，顯示這兩樣技術彼此仍然有能夠互補的層面存在。值得注意的是，以 ST-PSHE 處理後的 MFCC 特徵雖然平均的辨識正確率不如 AFE，但在極端的噪音環境下(訊噪比-5dB)反而能取得較好的效果，再次顯示調變頻譜的正規化對於嚴重的雜訊干擾是很有效的。

六、結論

在本論文中，我們探討了使用將語音特徵在時域與空間域進行分頻的方式以取得文脈資訊，進而減緩傳統 SHE 以及 PSHE 的嚴格限制。ST-PSHE 和傳統的方法相較，不僅全頻帶的調變頻譜具有一致的分佈，高頻成份與低頻成份的調變頻譜分佈也納入正規化的範圍，進一步地減少了雜訊對調變頻譜的干擾。實驗的結果也說明了本論文所提出的方法確實能夠達成較高的辨識正確率表現，並能夠與其他特徵正規化的方法互補。

展望未來研究，我們提出兩點可能的方向。第一是將此技術應用到更複雜的語音辨識任務上，如屬於大詞彙連續語音辨識(large vocabulary continuous speech recognition, LVCSR)的 Aurora-4 語料庫[45]和 MATBN 語料庫[46]上，以更進一步驗證我們所提出之方法是否在較複雜的語音辨識任務上也能夠有相同的表現。第二是在整個語句的調變頻譜之外，更深入地探討運用不同的分析單位處理調變頻譜，以期能捕捉更多層面的資訊而進一步提升語音辨識的強健性，並使此方法能夠應用在實時(real-time)的系統中。

七、誌謝

本論文之研究承蒙教育部－國立臺灣師範大學邁向頂尖大學計畫（102J1A0800）與行政院國家科學委員會研究計畫 (NSC 101-2221-E-003-024-MY3, NSC 101-2511-S-003-057-MY3, NSC 101-2511-S-003-047-MY3 和 NSC 102-2221-E-003-014-）之經費支持，謹此致謝。

參考文獻

[1] J. Droppo and A. Acero, "Environmental robustness," in *Springer handbook of speech processing*, 1st ed., J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer, 2008, ch. 33, pp. 653–679.

[2] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[3] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[4] M. J. Gales, "Model based techniques for noise robust speech recognition," Ph.D. dissertation, Cambridge University, 1995.

[5] P. Moreno, B. Raj, and R. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 2, 1996, pp. 733 –736.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[7] Y. Ephraim and H. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[8] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[10] I. Soon and S. Koh, "Low distortion speech enhancement," *IEEE Proceedings of Vision, Image and Signal Processing*, vol. 147, no. 3, pp. 247–253, 2000.

[11] A. de la Torre, J. C. Segura, C. Benitez, A. M. Peinado, and A. J. Rubio, "Non-linear transformations of the feature space for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 2002, pp. 401–404.

[12] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[13] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communucation*, vol. 25, no. 1-3, pp. 133–147, 1998.

[14] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. on Spoken Language Processing*, 2000.

[15] J. Wu, Q. Huo, and D. Zhu, "An environment compensated maximum likelihood

training approach based on stochastic vector mapping," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, vol. 1, 2005, pp. 429–432.

[16] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 205–220, 2009.

[17] B. Chen and S.-H. Lin, "Distribution-based feature compensation for robust speech recognition," in *Recent Advances in Robust Speech Recognition Technology*. Bentham Science Publishers, 2011, ch. 10, pp. 155–168.

[18] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[19] D. P. Ibm, S. Dharanipragada, and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proc. Int. Conf. on Spoken Language Processing*, 2000, pp. 556–559.

[20] B. Chen, W.-H. Chen, S.-H. Lin, and W.-Y. Chu, "Robust speech recognition using spatial-temporal feature distribution characteristics," *Pattern Recognition Letter*, vol. 32, no. 7, pp. 919–926, 2011.

[21] S.-S. Wang, Y. Tsao, and J.-W. Hung, "Filtering on the temporal probability sequence in histogram equalization for robust speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 2013.

[22] H.-J. Hsieh, J.-W. Hung, and B. Chen, "Exploring joint equalization of spatial-temporal contextual statistics of speech features for robust speech recognition," in *Proc. Annu. Conf. of the Int. Speech Communication Association*, 2012.

[23] V. Joshi, R. Biligi, U. S., L. Garcia, and C. Benitez, "Sub-band level histogram equalization for robust speech recognition," in *Proc. Annu. Conf. of the Int. Speech Communication Association*, 2011.

[24] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *The Journal of the Acoustical Society of America*, vol. 95, no. 3, pp. 1593–1602, 1994.

[25] W.-H. Tu, S.-Y. Huang, and J.-W. Hung, "Sub-band modulation spectrum compensation for robust speech recognition," in *Proc. IEEE Workshop on Automatic Speech Recognition Understanding*, 2009, pp. 261–265.

[26] L.-C. Sun, C.-W. Hsu, and L.-S. Lee, "Modulation spectrum equalization for robust speech recognition," in *IEEE Workshop on Automatic Speech Recognition Understanding*, 2007, pp. 81–86.

[27] J.-W. Hung, W.-H. Tu, and C.-C. Lai, "Improved modulation spectrum enhancement methods for robust speech recognition," *Signal Processing*, vol. 92, no. 11, pp. 2791–2814, 2012.

[28] X. Xiao, E. S. Chng, and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1662–1674, 2008.

[29] Y.-C. Kao and B. Chen, "Leveraging distributional characteristics of modulation spectra for robust speech recognition," in *Proc. Int. Conf. on Information Science, Signal Processing and their Applications*, 2012, pp. 120–125.

[30] L.-C. Sun and L.-S. Lee, "Modulation spectrum equalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*,

vol. 20, no. 3, pp. 828–843, 2012.

[31] T. Acharya and A. Ray, *Image Processing: Principles and Applications*. Wiley, 2005.

[32] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 845–854, 2006.

[33] S.-H. Lin, B. Chen, and Y.-M. Yeh, "Exploring the use of speech features and their corresponding distribution characteristics for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 84–94, 2009.

[34] S. Prasad and S. A. Zahorian, "Nonlinear and linear transformations of speech features to compensate for channel and noise effects," in *Proc. European Conf. on Speech Communication and Technology*, 2005, pp. 969–972.

[35] J. C. Segura, C. Benitez, A. de la Torre, A. J. Rubio, and J. Ramirez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Processing Letters*, vol. 11, no. 5, pp. 517–520, 2004.

[36] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, 1994.

[37] C.-P. Chen, K. Filali, and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *Proc. Annu. Conf. of the Int. Speech Communication Association*, 2002.

[38] C. E. Shannon, "Communication in the presence of noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[39] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the importance of various modulation frequencies for speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, 1997.

[40] S. Greenberg, "On the origins of speech intelligibility in the real world," in *Proc. ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, 1997.

[41] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[42] D. Pearce, H. G. Hirsch, and D. Gmbh, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA Workshop on ASR*, 2000.

[43] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[44] D. Macho, L. Mauuary, B. Noe, Y. M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust dsr front-end on aurora databases," in *Proc. Annu. Conf. of the Int. Speech Communication Association*, 2002.

[45] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," ETSI STQ-Aurora DSR Working Group, Tech. Rep. AU/384/02, 2002.

[46] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese Broadcast News Corpus," *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 10, no. 2, pp. 219–236, 2005.