出國報告　（出國類別：參與國際會議）

第八屆國際智能資訊處理研討會

服務機關：國立暨南國際大學
姓名職稱：范顥騰, 博士班研究生
派赴國家：南韓首爾
出國期間：2013 年 3 月 31 日至 4 月 5 日
報告日期：2013 年 5 月 28 日

摘　要

本次出國前往參加於首爾所舉辦的第八屆國際智能資訊處理會議，其會議舉辦主要目的為網羅世界各國鑽研此領域的學者交換其所得之知識與經驗，並分享和交流新型技術。此會議舉辦的時間為102年4月1日至4月3日為期三天，所探討的主題涵蓋資訊管理與應用、智慧系統之理論與應用、機器學習、圖像辨識、電腦視覺應用與自然語音和語言處理等多重領域，而我所研究的主題偏重於語音處理，投稿至此會議的論文為提出使用知名小波轉換法(wavelet transform)降低辨識環境因雜訊干擾所造成的不匹配情形，進而提升語音辨識系統的精確率。本大會以資訊處理為主軸方向，安排1個主題演講與26 個場次分別就不同主題以口頭報告的方式呈現全球各卓越研究者最近之相關研究成果，同時陳列資訊處理領域具代表性的研究成果，除了提供研究人員相互交流的平臺，也給予了未來方法走向。經由參加此會議，了解各國頂尖學術研究與技術人員的資訊處理趨勢，對於未來研究題材的選取頗具參考之價值，同時也提升自我國際觀與外語能力。

目　　　錄

頁數

一、　目的

『科技始終來自於人性』造就了世界日新月異的發展與開發，其中資訊處理的開發也不斷地蓬勃發展，到現今相關研究學者仍努力研發新型態技術。此領域囊括影像、電腦視覺、類神經網路、模糊系統、圖像辨識、自然語音和語言處理與資訊處理與應用等多元主題，而此國際會議提供一個訊息交流的場所，對於資訊處理相關課題的基本層面與進階發展為主要開發方向，達到經驗分享及相互交流之目的並致力開發嶄新技術；同時，此會議的使命也在於透過此平臺定期所討論的議題挑戰不同層面的研究。藉由參加此類型會議，不但能與世界各國創新研究接軌，也可吸收其研究經驗與交換彼此心得，經由發表論文提升國際能見度，對於未來研究的延伸或期刊的投稿有其莫大的幫助。，

二、 參與會議之過程

『第八屆國際智能資訊處理研討會』在2013年4月1日至4月3日於南韓首爾舉行，由混合資訊創新研究所(AICIT)所主辦。其主要任務為提供傳遞最新資訊之媒介，透過各國不同領域的學者相互知識與創新思維的交流，進階來開發新型應用技術，其會議安排二場主題演講與26個場次的主題報告，內容豐富並且多元，本人很榮幸能參與此會議並發表其論文。

本會議因屬國際型會議，參與人員從世界各國前來，舉辦地點位於奧林匹克花園酒店。會議第一天下午舉行開幕式並隨後進行主題演講，由學者 Dongsoo HAN 主講「Crowdsourcing Radio Map Construction for Wi-Fi Positioning Systems」，其主題有關於利用 Wi-Fi 定位系統(WPS)架構於室內的定位系統；談到定位，不外乎目前最知名的全球衛星定位系統(GPS)，應用層面廣泛，相對本身也有其缺點，容易受到地形與建築物之影響，如在室內使用，受到多重路徑或是訊號干擾的效應，所接收到之信號衰減相當嚴重。其報告內容主要針對上述現象提出有效率演算法，降低偵測所導致的誤判問題，提高定位的正確性，並且整合兩者系統達到無障礙系統的理念。目前語音辨識課題也相對應遇到類似的問題，如何在不同語言差異或實際環境影響之下提升辨識效能，這也是現今最嚴峻的課題之一。

在 Keynote Speech 演講後，接續為4個場次共為34個主題報告，其涵蓋範圍包含晶片設計、自動控制等，其中印象較深刻為使用獨立成份分析(ICA)理論應用在膚色偵測上，論文題目為「ICA based Skin Pigmentation Detection」。ICA 演算法發展近數十年頭，應用於訊號處理、影像處理與語音處理方面上都有極佳的效果，而此篇論文運用此方式辨別膚色的像素數值，進而分門別類偵測各個膚色的種類，對於醫學上具有相當的應用價值。

第二天早上安排 20 個場次共 51 個主題報告，而本人論文排定於第二場次第三位，論文內容(詳見於第三節)是針對生活環境因雜訊干擾造成不匹配導致辨識系統效能下降的主題研究：現今自動語音辨識系統(ASR)在無雜訊的環境下，能得到相當高的辨識

圖 1. 2013 國際智能資訊處理會議報到櫃臺

Outline
Introduction
The wavelet threshold de-noising (WD) process
Experimental results and analyses
Conclusion and future works
Introduction (1/3)
The environmental mismatch caused by two types of noise :

圖 2. 於 ICIIP 2013 口頭報告會場發表論文之過程

結果；但實際環境中往往有許多的干擾因子，使得此系統效能下降得非常嚴重。因此，探究語音之學者發展出一連串的強健性方法，進一步改善辨識系統的效能，同時也能達到不受外在干擾因素的影響。過程中，主持人也提問如何有效提升語音辨識系統的效能：這是一個極為嚴峻的課題，同時也是現今語音學者所要達到的目標。語音訊號除了外在的雜訊干擾外，同時也因各國語言上的腔調或是說話的速度有所影響，如何去補償這些所帶來的效應，從古至今已提出許多強健性演算法與模型調適法，雖然目前聲學模型在乾淨的環境下已可到相當高的辨識結果，一旦受到雜訊干擾或是其餘因素就會造成系統效能的下降，因此要達到此目標還有許多困難需克服。

接著，下一位輪到泰國的學者發表「Automatic Retrieval of Particular Oncology Documents from PubMed by Semantic-based Text Clustering」有關於提高醫學檢索資料庫的準確性。檢索意即從大量的資料中精確搜尋使用者所需的資訊，因此快速且易於使用

圖 3. 各國研究人員成果張貼處

圖 4. 口頭報告會場參與會議之人員

的工具成為發展主軸，作者提出方法經由測試結果可達到 92%準確率。

會後，在註冊櫃臺附近擺設不同領域研究成果(如圖 3)，休息之餘觀察各國先進研究人員目前所探討相關主題,從中也發現同一主題融合不同領域的內容形成創新的思維。此種研究整合方式是目前我所欠缺的研究能力其中之一,領域本身並無確切分界線且也無獨立的特性，需拋開自我的成見，多吸收不同領域之知識，累積一定的數量，同時不斷勇於嘗試,縱使實驗結果異於期望,也不代表是錯誤;相對地,而是要懂得嘗試錯誤，這是我此行最大的收穫。

短暫的休息後，參與關於 VoIP 的議題，其題目為「A Detection Method of Subliminal Channel based on VoIP Communication」。VoIP 全名為 **V**oice **o**ver **I**nternet **P**rotocol 是一種透過網際網路進行電話通訊的技術，應用上包含 Skype 與 Viber 通訊軟體等，最近許多資訊隱藏也同時用於及時(real time)通訊技術上，此篇論文主要探討透過一演算法將隱藏後的訊息提高正確解讀的比率,結果顯示正確可提升 50%並且達到即時處理的效果。

下午的場次偏重於商業管理的議題，雖然跟本人研究並無太大的相關性，但也可了解此領域的最新趨勢。第三天安排為會議主持人討論，因並無擔任此職務，故沒有參加此活動，所以會議三天告一段落。

三、 發表論文之內容

本人提出小波消噪法(wavelet de-noising)用於語音特徵序列上，來提升其強健性，進一步改善語音辨識系統的效能。方法流程上，先分別對每一維語音特徵時間序列做統計值正規化處理，如平均值與變異數正規化法(mean and variance normalization, MVN)或增益正規化法(cepstral gain normalization, CGN)，緊接著運用小波消噪法做後續處理。在國際通用的語料庫(Aurora-2)實驗結果上得出，上述方法可顯著提升雜訊環境下語音辨識精確度約 20%的相對改善率。以下就動機、方法與實驗結果分述如下：

1. 動機

現今語音辨識系統裡，常因環境的各種干擾因素，如雜訊與通道干擾等，使其接收到的語音產生嚴重失真，進而明顯降低其辨識能力，因此，許多相關學者致力於發展改善上述問題之強健性語音辨識技術。這些方法中，有一大類是著重於對於語音特徵其時間序列統計值的正規化處理，如平均值與變異數正規化法或增益正規化法等。另外，離散小波轉換(discrete wavelet transform, DWT)是近年來在數位信號處理領域上的新型技術，相對於傳統的離散傅立葉轉換(discrete Fourier transform, DFT)而言，DWT 可以呈現訊號在頻譜以外的資訊，例如時頻域變化特性。許多基於 DWT 的訊號分析或處理技術也日益增多，例如各類的小波消噪(wavelet de-noising)技術。

在一般的小波消噪演算法中，假設是雜訊干擾集中在訊號的中高頻成分，因此只對於中高頻率區域做消噪處理，但當我們初步將小波消噪演算法使用於時域上的語音訊號時，發現其對語音辨識的提升度並不如預期理想，此原因在於，語音訊號受雜訊干擾的情形未必符合小波消噪之前提假設，亦即雜訊未必主要對時域上的語音之中的高頻成分產生干擾。

根據以上的觀察，在本論文中，我們提出了一個語音強健性的方法，主要步驟是將小波消噪法作用於經統計正規化後的語音特徵時間序列上，發現此時小波消噪法即可帶來明顯的語音辨識率提升。

2. 方法流程

主要流程架構如圖 5 所示：

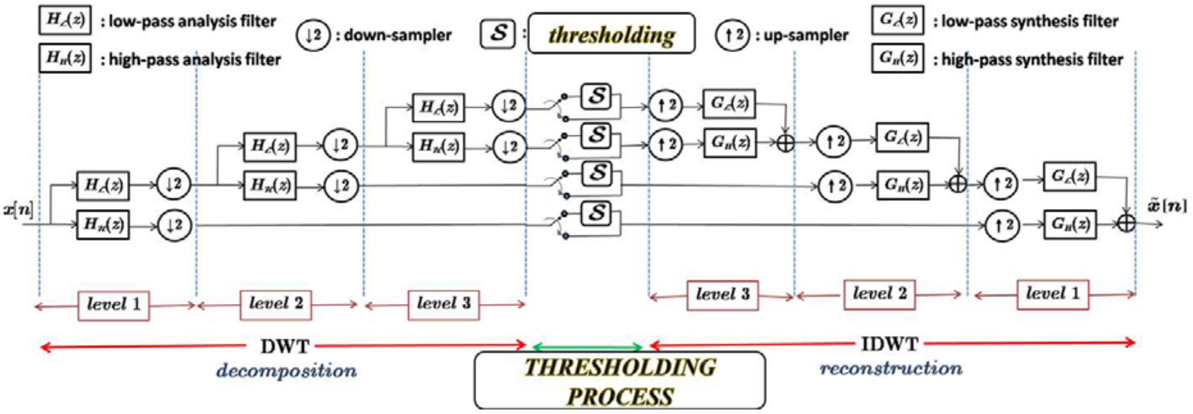


圖 5. 小波消噪法流程

對語音特徵序列$\{x[n]\}$執行$L-1$階的離散小波轉換，得到L段由低頻（近似成分）到高頻（細節成分）的子訊號，在此，以$\{x_1[n]\}$、$\{x_2[n]\}$,..., $\{x_L[n]\}$表示之。保留最低頻的子訊號$\{x_1[n]\}$不動，而更新其他子訊號，更新的方式主要分為兩種：

a. 硬式門檻(hard thresholding)決策法：

$$\tilde{x}_\ell[n] = T_h(y_\ell[n],\theta) = \begin{cases} y_\ell[n] & \text{if } |y_\ell[n]| \ge \theta, \\ 0 & \text{elsewhere} \end{cases} \qquad (1)$$

若$x_\ell[n]$其強度在某個門檻值θ以上，則視其為較不受干擾的訊號點而保留原值，反之，則將其值設為 0。

b. 軟式門檻(soft thresholding)決策法：

$$\tilde{x}_\ell[n] = T_h(y_\ell[n],\theta) = \begin{cases} \operatorname{sgn}(|y_\ell[n]|)(|y_\ell[n]|-\theta) & \text{if } |y_\ell[n]| \ge \theta, \\ 0 & \text{elsewhere} \end{cases} \qquad (2)$$

若$x_\ell[n]$其強度在某個門檻值θ以上，則將其強度扣除此門檻值，反之，則將其值設為 0。

最後，將將最低頻的原始子訊號$\{x_1[n]\}$結合其他更新後較高子頻帶$\{\tilde{x}_2[n]\}$,..., $\{\tilde{x}_L[n]\}$，使用$L-1$階反離散小波轉換(inverse DWT)，得到新語音特徵序列$\{\tilde{x}[n]\}$。

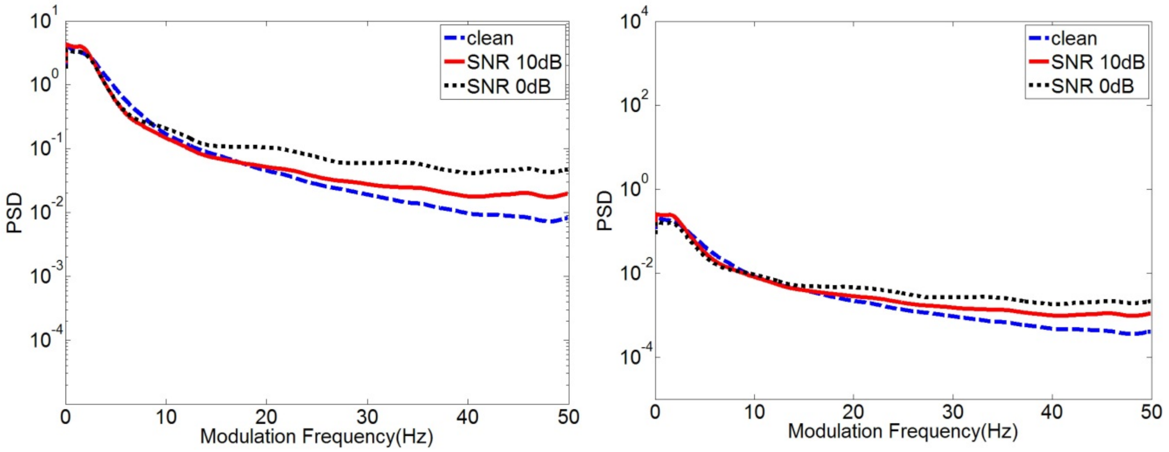
3. 實驗結果

所採用語音資料庫為歐洲電信標準協會(European Telecommunication Standard Institute, ETSI)所發行之語料庫 AURORA 2.0，其測試語料包含三個集合，Sets A 與 B 為加成性雜訊環境，Set C 則同時包含加成性雜訊與摺積性雜訊。原始語音特徵為 39 維（13 維 MFCC 加上其一階與二階差分值）。聲學模型為隱藏式馬可夫模型(hidden Markov models, HMM)，以 HTK 軟體訓練而得。

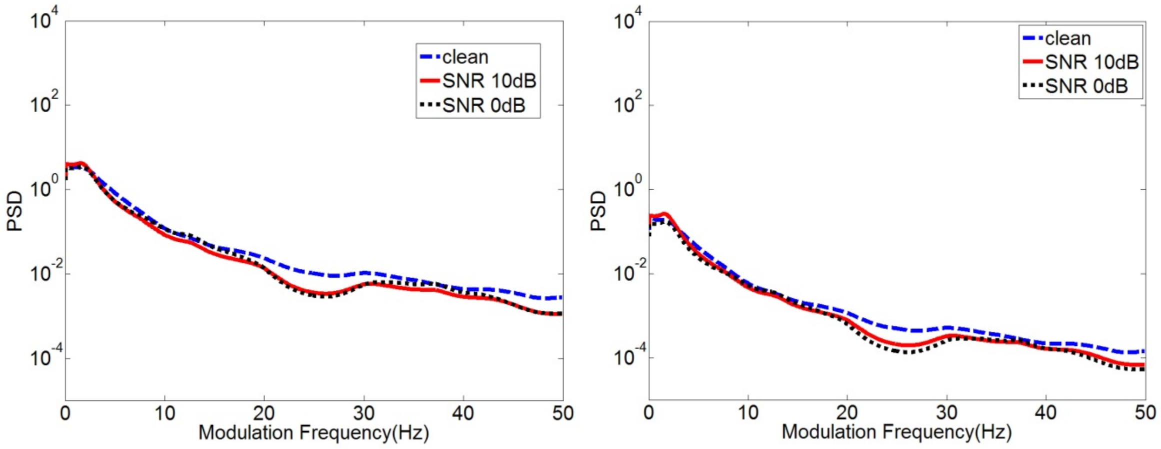
表 1 所提出的方法與特徵正規化結合所得之辨識率

Method	Set A	Set B	Set C	Avg	RR
MFCC	71.89	68.24	77.57	71.57	—
MVN	85.05	85.62	85.70	85.41	48.68
WD with MVN	**88.29**	**89.07**	**88.62**	**88.67**	**60.15**
HEQ	86.91	88.32	87.50	87.59	56.36
MVA	88.12	88.81	88.50	88.47	59.46
TSN	89.42	90.03	89.03	89.59	63.37
CGN	87.64	88.55	87.73	88.02	57.87
WD with CGN	**89.81**	**90.75**	**89.89**	**90.20**	**65.54**

從表 1 可知當小波消噪法（WD）與 MVN 或 CGN 結合時，相對於單一正規化法而言，前述之三種子頻帶組合之消噪處理，幾乎都可使辨識率更佳，對 MVN 而言，WD 帶來的進步率最高 3.32%，而對 CGN 而言，WD 帶來的進步率最高為 2.18%。我們同時可看出，WD 的加入可使整體辨識率超過 90%，相對於原始 MFCC 而言，有高達約 19%的進步。另外，從功率頻譜密度(PSD)的角度上觀察



(a) (b)



(c) (d)

圖 6. 經由(a)MVN (b)CGN (c)WD with MVN (d)WD with CGN 處理所得之功率頻譜密度圖

圖 6(a)、圖 6(b)顯示 MVN 與 CGN 處理之特徵序列做小波門檻消噪法後的調變頻譜強度圖，將這兩圖與圖 6(c)、圖 6(d)相較，可看出中高頻以上的失真可明顯被降低，代表雜訊效應進一步被消弭。

四、　心得

由於南韓是第一次前往的國家，之前透過報章雜誌或是電視傳播有些許印象，但參與會議後，經過短期的語言與文化交流，也有了新的認識與感受。很榮幸能從眾多投稿文件中脫穎而出，進一步發表目前的研究成果，過程裡吸收到不同領域研究員的先進知識與經驗，從而開拓自我的研究觀與國際觀，此行對我而言給予了對於研究不同的見解與認知。

在會議兩天中，觀察不同國家最近的研究成果，每個主題所提出創新的思考方式，增添日後我所要研究的方法走向，之前思考模式都只是前往單一方向，未往整體發展為主軸；因先用俯觀的方式找尋所要探討的主題，延伸出的思維也不會容易被侷限住，相對於方法也就會多樣化，較不易陷入思想泥沼裡。

在英文口頭發表過程裡，深刻體會到外語能力對於國際會議的重要性，無論是應答或是基本對話。雖之前已參加過幾次，但掌握度並不熟練且也不流暢；這部分無法一蹴可幾，需持之以恆每天練習。每次參加此類型會議回來，都對於自我的語文能力深刻反省，期許自己有朝一日能達成目標。

綜合以上所述，此韓國會議行，在發表過程或是探討研究成果上收穫良多，期望自己把握每次出國發表的機會，因為不論在國際觀或是語文能力都有顯著的提升，同時也可以反觀自我的研究方向，是否具備宏觀與微觀。最後，希望國家機關能摒除學術城鄉的差距鼓勵支持研究生多積極參與研討會，不僅可提升國際競爭力也可增加自我的研究價值。

五、　建議

對於這次參與國際會議之經驗，有以下建議提供參考：

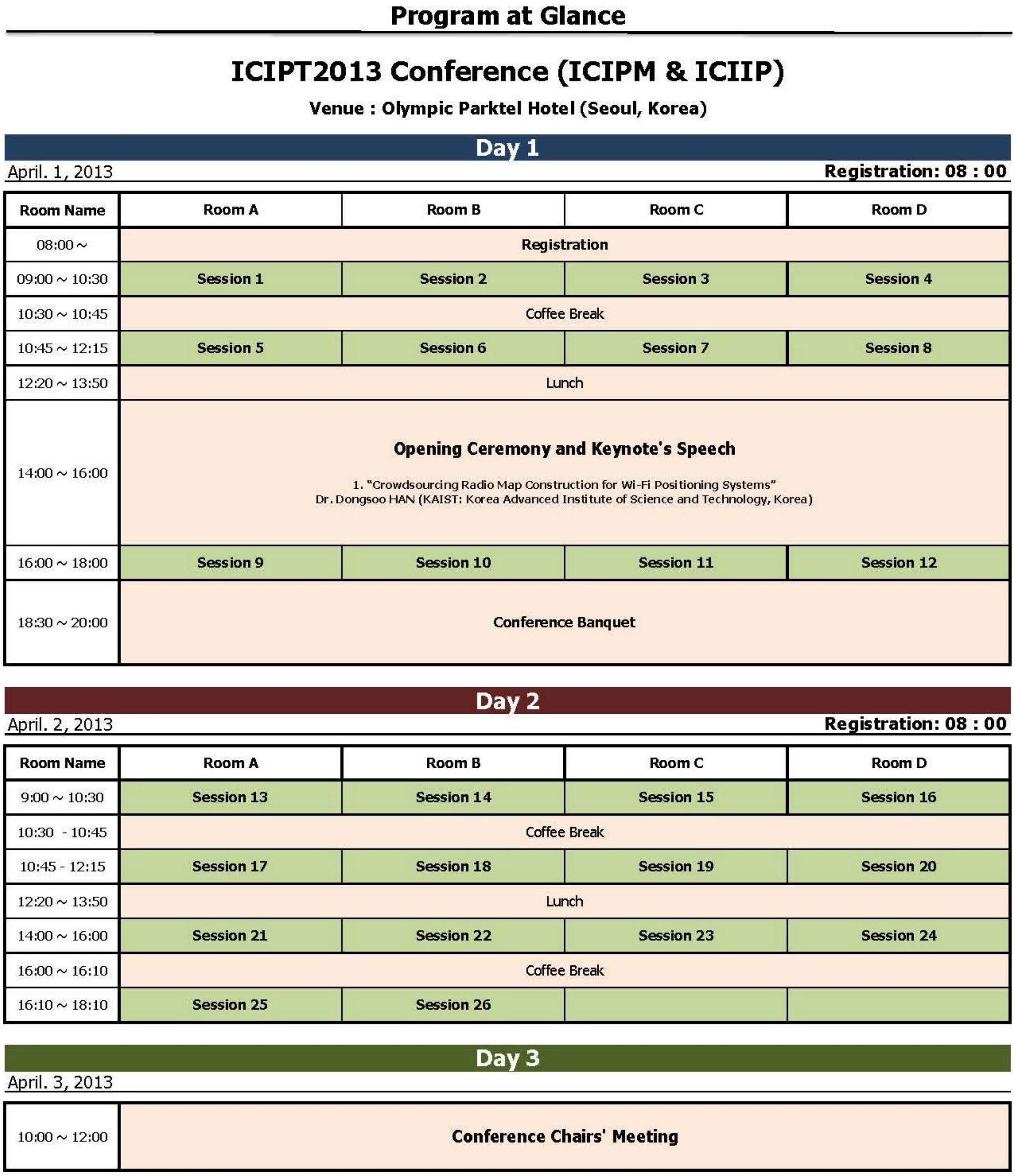
1. 多增加補助出國經費的管道。

   就目前的出國經費的補助，以博士班而言分別可向教育部與國科會申請。這兩者如申請審核未通過，等同於是博士生需自費前往，對於此種情況而言，出國發表論文對報告者來說無非是一種負擔，機票加旅館住宿費用亞洲國家動輒 2 萬左右，歐美國家更是需要 4 萬以上。如可增加申請管道，可鼓勵研究生出國參與國際會議，增加國際觀與自我的英文能力，無非可提升研究者之自我成長能力。

2. 視情況增加補助的額度

   以申請經驗來說，通常補助的額度為機票的部分費用，但有時的情況是旅館費用都可能花上萬塊錢，譬如歐美國家。如能依照情況補助機票加旅館的部分費用，這對於研究生而言是一種莫大的幫助。

六、 附錄



Program at Glance

ICIPT2013 Conference (ICIPM & ICIIP)

Venue : Olympic Parktel Hotel (Seoul, Korea)

Day 1

April. 1, 2013 **Registration: 08 : 00**

Room Name	Room A	Room B	Room C	Room D
08:00 ~	Registration			
09:00 ~ 10:30	Session 1	Session 2	Session 3	Session 4
10:30 ~ 10:45	Coffee Break			
10:45 ~ 12:15	Session 5	Session 6	Session 7	Session 8
12:20 ~ 13:50	Lunch			
14:00 ~ 16:00	**Opening Ceremony and Keynote's Speech** 1. “Crowdsourcing Radio Map Construction for Wi-Fi Positioning Systems” Dr. Dongsoo HAN (KAIST: Korea Advanced Institute of Science and Technology, Korea)			
16:00 ~ 18:00	Session 9	Session 10	Session 11	Session 12
18:30 ~ 20:00	Conference Banquet			

Day 2

April. 2, 2013 **Registration: 08 : 00**

Room Name	Room A	Room B	Room C	Room D
9:00 ~ 10:30	Session 13	Session 14	Session 15	Session 16
10:30 - 10:45	Coffee Break			
10:45 - 12:15	Session 17	Session 18	Session 19	Session 20
12:20 ~ 13:50	Lunch			
14:00 ~ 16:00	Session 21	Session 22	Session 23	Session 24
16:00 ~ 16:10	Coffee Break			
16:10 ~ 18:10	Session 25	Session 26		

Day 3

April. 3, 2013

10:00 ~ 12:00	Conference Chairs' Meeting

Leveraging wavelet de-noising in temporal sequences of speech features for noise-robust speech recognition

Hao-teng Fan
National Chi Nan University
Puli Township
Nantou County, Taiwan
886-49-2910960
s99323904@mail1.ncnu.edu.tw

Jan-Yee Lee
Kun-shan University
Yongkang District
Tainan City, Taiwan
886-6-2727175
tpejohnny@gmail.com

Jeih-weih Hung
National Chi Nan University
Puli Township
Nantou County, Taiwan
886-49-2910960
jwhung@ncnu.edu.tw

I-Chia Lu
National Chi Nan University
Puli Township
Nantou County, Taiwan
886-49-2910960
s98323530@ncnu.edu.tw

ABSTRACT

In this paper, we propose to employ the wavelet de-noising (WD) techniques in temporal-domain feature sequences for enhancing the noise robustness in order to improve the accuracy of noisy speech recognition. In the proposed method, the temporal domain feature sequence is first processed by some specific statistic normalization scheme, such as mean and variance normalization (MVN) and cepstral gain normalization (CGN), and then dealt with the wavelet de-noising algorithm. With this process, we find that the wavelet de-noising procedure can effectively reduce the middle and high modulation frequency distortion remaining in the statistics-normalized speech features. On the Aurora-2 digit database and task, experimental results show that the above process can significantly improve the accuracy of speech recognition under noise environments. The pairing of WD and CMVN/CGN provides about 20% relative error reduction associated with the MFCC baseline, outperforms the individual CMVN/CGN, and makes the overall recognition rate beyond 90%.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *Speech recognition and synthesis.*

General Terms

Languages, Human Factors, Experimentation

Keywords

wavelet transform, speech recognition, noise robustness

1. INTRODUCTION

The current automatic speech recognition systems perform well when tested on data similar to those used for training. However, the lack of robustness of recognition systems seems to be a serious obstacle in noisy environments to practical speech recognition. As a result, a lot of methods have been proposed to improve the robustness of speech recognition systems, and one category of them is to compensate speech features to make them less distorted by noise. The Mel-Frequency Cepstral Coefficients (MFCCs) are a popular speech feature type due to their low correlation and their ability to arrive at a compact and computationally efficient representation of a speech signal. However, MFCCs are not very noise-robust, and thus most approaches of this category are designed for MFCCs in order to improve the noise robustness while retain their inherent merits. One direction of this category of methods is to normalize the statistics of MFCC temporal streams in both training and testing conditions. These feature statistics normalization techniques include cepstral mean subtraction (CMS) [1], mean and variance normalization (MVN) [2], cepstral gain normalization (CGN) [3] and histogram equalization (HEQ) [4]. In these feature statistics normalization methods, the MFCCs in the temporal sequence are often viewed as the samples of a random variable, and thus the required statistical information of this random variable is estimated directly via these samples.

Conventional signal analysis primarily relies on the technique of Fourier transform. However, wavelet transform gets more popular primarily due to some of its particular properties that Fourier transform lacks. The main difference is that wavelet transform considers both the time and frequency aspects of a signal while the standard Fourier transform takes cares of the frequency parts only. The short-time Fourier transform (STFT) is somewhat similar to the wavelet transform since it is also time- and frequency-localized, but it has the issues with the frequency/time resolution trade-off. In contrast, wavelet transform often gives a better signal representation in multi-resolution analysis with balanced resolution at any time and frequency [5].

Many de-noising algorithms [6,7] have been developed through wavelet transform, and present very good results in alleviating noise. However, our preliminary experimental results reveal that the performance of the wavelet denoising (WD) algorithms in dealing with noisy speech signals is sensitive to parameter settings. One possible explanation is, when performing these WD algorithms to reduce the middle/high-frequency noise components, the speech components in the same frequency ranges are impaired and/or weakened at the same time.

In this paper, we present a novel application for the WD algorithms. They are performed on the temporal domain of speech features rather than the time domain of speech signals. We use the WD algorithms to remove or reduce the middle/high "modulation frequency" distortions of speech features. Our experimental

ICIPT '13, April 01 - 03 2013, Seoul, Republic of Korea
Copyright 2013 ACM 978-1-4503-1783-2/13/04…$15.00.

results indicate that such a use of the WD algorithms effectively improves the noise-robustness of speech features and brings significant accuracy promotion. Compared to the conventional application in the time domain, using WD algorithms in the temporal domain does not distort the speech portions very much because the primary speech information for recognition is located in the low-modulation frequency region (about below 15 Hz) [8]. This is a probable reason of the success for WD algorithms in processing temporal feature streams. Furthermore, when applying in the temporal domain, the performance of the WD algorithms is relatively insensitive to the values of tuning parameters, implying that these WD algorithms are robust, and there is no need to set the parameters in WD meticulously in order to obtain the nearly optimal performance.

The remainder of the paper is organized as follows: Section 2 gives brief discussions of DWT and wavelet denoising algorithms. The novel WD method operating in temporal feature stream is presented in Section 3. Section 4 contains experimental results and discussions about the presented WD. Finally, in Section 5 we give conclusions and future works.

2. DWT AND WAVELET DENOISING

Relative to discrete Fourier transform (DFT), DWT can show the additional information of a signal in the spectrum, such as the properties in a short-time range. The number of DWT-related signal analysis technology is also increasing, For example, a lot of research about wavelet threshold de-noising has been proposed [6,7]. In the following section, we briefly review the discrete wavelet transform (DWT) and the concept of the wavelet threshold de-noising method.

2.1 Discrete wavelet transform (DWT)

Mathematically, the DWT of a signal $f[n]$ is the outcome of passing $f[n]$ through a series of filters. The resulting signals represent different frequency components of $f[n]$. First, the signal $f[n]$ is passed through two filters (called analysis filters) with impulse responses $g[n]$ and $h[n]$ in parallel to obtain two signals. The two filters are low-pass and high-pass, respectively, and are related to each other by

$$g[n] = (-1)^n g[N-1-n], \tag{1}$$

where N is the filter length, and eq. (1) shows $g[n]$ and $h[n]$ are a quadrature mirror filter. Since half the frequencies of the signal $f[n]$ have now been removed in the output signals of two filters, we can discard half the points according to Nyquist's rule. The filter outputs are then down-sampled by 2:

$$f_{low}[n] = \sum_k f[k]g[2n-k], \tag{2}$$

and

$$f_{high}[n] = \sum_k f[k]h[2n-k]. \tag{3}$$

Compared with the original signal $f[n]$, the two sub-band signals $f_{low}[n]$ and $f_{high}[n]$ are halved in time resolution but doubled in frequency resolution. Besides, $f_{low}[n]$ and $f_{high}[n]$ are called the approximation and detail parts of $f[n]$, respectively.

The standard form of an *L*-level DWT of $f[n]$ is to perform the one-level DWT as shown in Figure 1 on the approximate (low-frequency) part obtained from the (*L*-1)-level DWT of $f[n]$, where $L > 1$. In the multi-level DWT, the decomposition (low-pass and high-pass filtering together with down-sampling) is repeated to further increase the frequency resolution. The resulting sub-band signals reveal a different time-frequency localization.

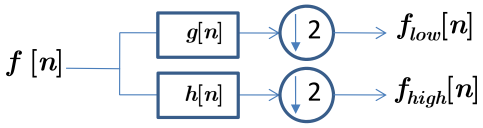


Figure 1. a one-level discrete wavelet transform (DWT)

On the contrary, we can reconstruct the original signal from summing up the up-sampling sub-band signals through the reconstruction high-pass and low-pass filters. The process is called the inverse discrete wavelet transform (IDWT). The one-level IDWT, which is the inverse process of the DWT shown in Figure1, is depicted in Figure 2. The two sub-band signals $f_{low}[n]$ and $f_{high}[n]$ are first up-sampled by 2 and then passed through the low-pass and high-pass synthesis filters. Note that in IDWT the two synthesis filters are with impulse responses $g[-n]$ and $h[-n]$, respectively, which are related to the analysis filters with equal magnitude and reversed phase in frequency response. It is easy to obtain the L-level IDWT by extending the one-level IDWT structure.

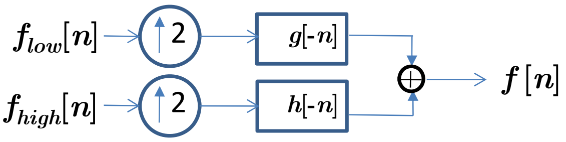


Figure 2. a one-level inverse discrete wavelet transform (IDWT)

2.2 Wavelet threshold de-noising

Among many DWT-based methods in handling noise, the algorithm of "wavelet threshold de-noising" (WD) is widely used due to its high effective performance in reducing noise and its various parameters which can be tuned to cope with different noise conditions. The basic idea of WD is to use the de-correlation characteristic of the output signals from the DWT, which helps to separate clean signals and noise. DWT enables the signal energy to be concentrated at some larger wavelet coefficients while makes the noise energy distribute throughout the wavelet domain. As a result, the signal amplitude of the wavelet coefficients are greater than the noise amplitude, and the thresholding method is adopted to retain the signal coefficient and to attenuate.

The WD algorithm presented in [7] can be summarized as follows:

Let $x[n]$ represent a clean signal with a finite length L and $y[n]$ stand for the signal corrupted by additive white Gaussian noise $d[n]$ with a variance σ^2. We have

$$y[n] = x[n] + d[n]. \qquad (4)$$

Performing DWT on $y[n]$, we have

$$y_\ell[n] = x_\ell[n] + d_\ell[n], \qquad (5)$$

where $y_\ell[n]$, $x_\ell[n]$ and $d_\ell[n]$ represent the ℓ^{th} sub-band DWT coefficients of $y[n]$, $x[n]$ and $d[n]$, respectively.

Then, based on the wavelet threshold de-noising algorithm in [ref], the estimated clean signal of $x_\ell[n]$ in wavelet domain, denoted as $\tilde{x}_\ell[n]$, has two expressions as follows according to different threholding strategies:

- hard-thresholding:

$$\tilde{x}_\ell[n] = T_h(y_\ell[n],\theta) = \begin{cases} y_\ell[n] & \text{if } \left|y_\ell[n]\right| \geq \theta \\ 0 & \text{elsewhere} \end{cases}, \qquad (6)$$

and

- soft-thresholding:

$$\begin{aligned} \tilde{x}_\ell[n] &= T_s(y_\ell[n],\theta) \\ &= \begin{cases} \text{sgn}(\left|y_\ell[n]\right|)(\left|y_\ell[n]\right| - \theta) & \text{if } \left|y_\ell[n]\right| \geq \theta, \\ 0 & \text{elsewhere} \end{cases} \end{aligned} \qquad (7)$$

where θ is a threshold. That is, if the coefficient magnitude is less than the specified threshold vale, it is identified as noise and we specify the new coefficient to be zero. On the other hand, if the coefficient magnitude is greater than or equal to the threshold θ, then it is identified as a signal coefficient and the new coefficient is set to be the original coefficient as eq. (6) or the original coefficient minus the threshold as eq. (7).

For the threshold θ, several selection rules are suggested in [7]:

- the "**rigrsure**" rule: It is based on Stein's unbiased risk estimation (SURE) [9]. The risk for a particular threshold value θ is estimated. The θ that minimizes the risk is then the used threshold.
- the "**sqtwolog**" rule: It uses a fixed-form threshold that yield the minimax performance multiplied by a small factor proportional to $\log(L_\ell)$, where L_ℓ is the length of $y_\ell[n]$.
- the "**heursure**" rule: It stands for heuristic SURE and is a mixture of the above two rules. If the signal-to-noise ratio (SNR) is low within this sub-band, it uses a fixed threshold (following the "sqtwolog" rule), otherwise the threshold from the "rigrsure" is used.

The notations "rigrsure", "sqtwolog" and "heursure" are from the naming of the argument for the MATLAB function "**thselect**" [10].

All the above threshold selection rules are related to the noise variance. In [10], three alternatives to determine the noise variance are provided:

- "**one**": the noise is assumed to follow a standard Gaussian distribution, and thus the variance is set to one.
- "**sln**": A common value, which is the estimated noise variance of the first sub-band signal $y_1[n]$, is used for every sub-band.
- "**mln**": The noise variance for each sub-band is estimated and then applied to the associated sub-band.

Again, here the three notations "one", "sln" and "mln" are from the naming of the argument for the MATLAB function "wden" [10].

3. WAVELET THRESHOLD DENOISING FOR NORMALIZED SPEECH FEATURES

Here, we present a novel application of wavelet de-noising techniques, which perform in temporal-domain feature sequences to enhance the noise robustness. In the proposed scheme, the original MFCC features are first compensated by either of two normalization methods, mean and variance normalization (MVN) [2] and cepstral gain normalization (CGN) [3], and then processed by wavelet de-noising (WD). The pairing of MVN/CGN and WD helps to further enhance the noise robustness compared with the individual component method. In the following sub-sections, we describe the ideas and detailed structures of the presented novel scheme, followed by some preliminary performance demonstrations.

3.1 Basic ideas

In most wavelet de-noising (WD) algorithms, it is assumed that the noise interference primarily concentrates in the relatively high frequency components of signals. Based on this assumption, these algorithms deal with the detail parts (high frequency portions) and leave the approximation parts (low frequency portions) unchanged. However, when speech signals are first updated by either of these WD algorithms and then undergo the training and testing processes for speech recognition, the various parameters (for example, the level of DWT, the selecting threshold, and the hard/soft thresholding strategy) in the WD algorithm usually require a careful and heuristic tuning in order to achieve better recognition accuracy. In other words, the effect of WD in improving recognition performance is quite sensitive to the used parameters. This is very probably due to two reasons:

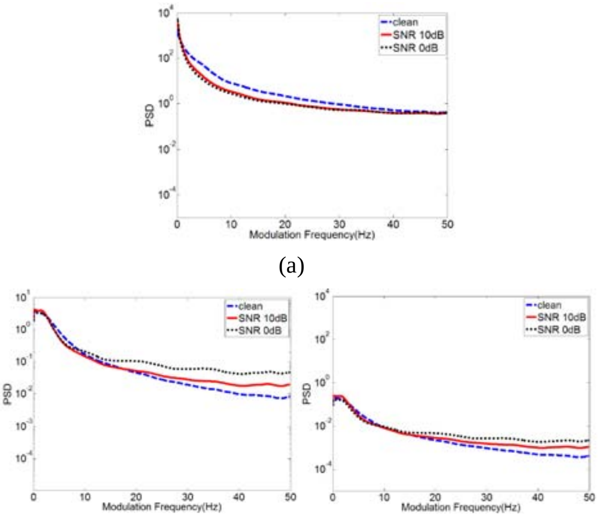
1. The noise and/or interference existing in the speech utterance is not always high-pass in spectrum, which somewhat contradicts the assumption of many WD algorithms.

2. Some important speech information with median/high frequencies is eliminated or undermined by WD algorithms if the parameters in WD are not well defined. Roughly speaking, the lower bound of the sampling rate for speech signals in recognition is around 8 kHz, implying the speech information helpful in recognition is within the range [0, 4 kHz]. Inappropriately dealing with the high frequency components, which contain a wealth of recognition information, may damage the speech signal and result in low recognition accuracy.

In contrast to the time-domain speech signals, the temporal-domain speech feature streams reveal a significant band-limited characteristic. A lot of research has confirmed that a temporal feature stream corresponding to a clean speech utterance possesses most of its energy and useful information for recognition in low modulation frequency. More precise speaking, [8] has shown that most of the useful linguistic information is in the modulation frequency components between 1 Hz and 16 Hz, with the dominant component at around 4 Hz. From this viewpoint, the WD algorithms are appropriate to process the temporal speech stream in order to alleviate the high modulation frequency distortion without damaging the speech contents that are mainly located in low modulation frequencies.

The feature normalization techniques like MVN and CGN can enhance the noise robustness and bring higher recognition accuracy relative to the baseline MFCC features. However, some of these algorithms can be further improved by inspecting the possible inferior effects they cause or the residual distortion they care less. For example, [11] shows that MVN often results in a relatively "flat" modulation spectrum, which can be adjusted with an ARMA filter to emphasize the low (modulation) frequency portion as in the proposed algorithm. The technique of TSN enhances MVN by designing an utterance-based temporal filter.

Figures 3(a), (b) and (c) shows the averaged power spectral density (PSD) curves for the unprocessed, MVN- and CGN-processed MFCC c1 streams for a set of 1001 utterances at three signal-to-noise ratios (SNRs) in The Aurora-2 database [12]. Noise results in significant PSD distortion in the original MFCC stream, as shown in Figure 3(a). However, Figures 3(b) and 3(c) reveal that both MVN and CGN reduce the PSD mismatch especially at low modulation frequencies (roughly below 5 Hz), while the distortion at higher frequency components still remains.



(a)

(b)

(c)

Figure 3: the averaged PSD curves of feature streams at three SNRs, clean, 10 dB and 0 dB, corresponding to the MFCC c1 features (a) unchanged (b) processed by MVN (c) processed by CGN

Inspired by the above observations, in this thesis we propose to conduct the WD algorithm on the MVN- or CGN-processed temporal feature streams with the hope to improve the noise robustness further. It is expected that WD helps to reduce the high modulation frequency distortions left behind by MVN/CGN to further alleviate the mismatch caused by noise, and thereby to promote the recognition accuracy. We also hope that, in the proposed scheme the WD algorithms performing in the temporal domain of speech features do not require meticulous parameter settings to present good results.

3.2 Proposed method

We consider the mel-scaled filter-bank cepstral coefficients (MFCC) for speech recognition. Let $x^{(m)}[n]$ be the m^{th} cepstral coefficient of the n^{th} frame of an utterance. As a result, we have M feature streams:

$$\left\{x^{(m)}[n]; 0 \le n \le N-1\right\}, \qquad 0 \le m \le M-1$$

where M is the number of cepstral coefficients within a frame and N is the number of frames in the utterance. For the sake of compact notation, we omit the superscript "(*m*)" in $x^{(m)}[n]$ in the discussions hereafter, unless otherwise mentioned.

First, the original feature stream $\{x[n]\}$ of each utterance in both the training and testing sets is processed by either MVN [2] and CGN [3] to produce new streams, denoted as $\tilde{x}[n]$. Next, we operate the wavelet denoising algorithm on the stream $\{\tilde{x}[n]\}$:

Step I: Split the stream $\{\tilde{x}[n]\}$ by DWT.

The stream $\{\tilde{x}[n]\}$ is decomposed into L sub-streams $\{\tilde{x}_\ell[n]; 1 \le \ell \le L\}$ by performing an $(L-1)$-level discrete wavelet transform (DWT). Given that the frame rate of $\tilde{x}[n]$ is F_s in Hz, $\tilde{x}[n]$ is thus within the modulation spectral band $[0, F_s/2 \text{ Hz}]$, and the band range of the ℓ^{th} sub-stream $\{\tilde{x}_\ell[n]\}$ can be approximately represented as

$$\begin{cases}[0,\frac{1}{2^{L-1}}(\frac{F_s}{2})], & \text{if } \ell = 1\\ [\frac{2^{l-2}}{2^{L-1}}(\frac{F_s}{2}),\frac{2^{l-1}}{2^{L-1}}(\frac{F_s}{2})], & \text{if } \ell = 2,3,..,L\end{cases}. \qquad (8)$$

Step II: perform the thresholding operation on the sub-stream $\{\tilde{x}_\ell[n]\}$

For the sub-stream $\{\tilde{x}_\ell[n]\}$ to be updated, the resulting new sub-steam is either soft thresholding:

$$\hat{x}_\ell[n] = T_s(\tilde{x}_\ell[n],\theta) = \begin{cases} sgn(\tilde{x}_\ell[n])(\left|\tilde{x}_\ell[n]\right| - \theta), & \text{if } \left|\tilde{x}_\ell[n]\right| \ge \theta\,,\\ 0, & \text{if } \left|\tilde{x}_\ell[n]\right| < \theta\end{cases} \qquad (9)$$

or hard thresholding:

$$\hat{x}_\ell[n] = T_h(\tilde{x}_\ell[n],\theta) = \begin{cases}\tilde{x}_\ell[n], & \text{if } \left|\tilde{x}_\ell[n]\right| \ge \theta\\ 0, & \text{if } \left|\tilde{x}_\ell[n]\right| < \theta\end{cases}, \qquad (10)$$

where θ is the threshold, and T_s in eq. (9) and T_h in eq. (10) correspond to the soft- thresholding and hard-thresholding strategy, respectively. According to Section II, the threshold θ has three selections.

Step III: Use IDWT to obtain the final new stream

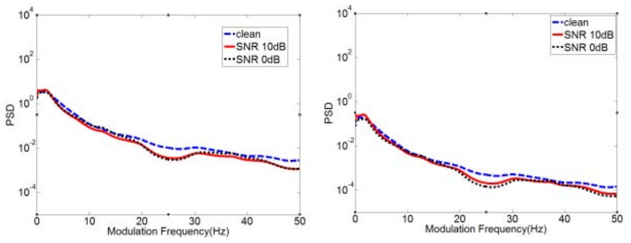
The final new feature stream is obtained by performing an (*L*-1)-level inverse DWT (IDWT) on the set of L sub-streams, including the new sub-streams $\{\hat{x}_\ell[n]\}$ from Step II and the other unaltered sub-streams.

3.3 Preliminary performance analysis

We give an initial evaluation on the proposed method that performs WD on MVN/CGN-processed features in terms of the effect of reducing the PSD distortion as in Figures 4(a)(b). The parameter settings for the WD algorithm here is as follows:

1. 2-level discrete wavelet transform
2. A soft threholding strategy
3. The "*heursure*" (heuristic Stein's unbiased risk) threshold selection rule.
4. The "*mln*" (multi-rescaling with level-dependent estimation) mode of noise variance estimation

Comparing Figures 4(a)(b) with Figures 3(b)(c), the PSD mismatch roughly above 10 Hz for MVN/CGN-processed c1 streams is apparently lowered by WD, partially supporting our statement that WD helps to deal with the residual distortion left behind by MVN/CGN. However, Figures 4(a)(b) show that in the range around the frequency of 25 Hz, the PSD mismatch gets relatively large, indicating WD may not deal with the distortions dwelled in the "border regions" of the sub-bands very well.



(a) (b)

Figure 4. the PSD curves of feature streams at three SNRs, clean, 10 dB and 0 dB, corresponding to the MFCCs (a) processed by MVN and then WD (b) processed by CGN and then WD

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Our recognition experiments use the AURORA 2.0 continuous-digit string database, which was published by European Telecommunication Standard Institute (ETSI) [12]. The contents of each string contain 11 digits: zero, one, two, three, four, five, six, seven, eight, nine and oh, and all the strings are uttered by American adult males and females. For evaluating the impact of noise on speech, additive noise and channel effects are used to corrupt the clean utterances. Eight types of additive noise are used, which are recorded in the environments of subway, babble, car, exhibition, restaurant, street, airport and train station, respectively. The noise is added to each clean speech signal at seven different signal-to-noise ratio (SNR) levels: clean, 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and -5 dB. As for the channel distortion, either of the two impulse responses, G712 and MIRS, is convolved with each clean speech signal.

For the baseline experiments for the "clean-condition training task" defined in Aurora 2 database, each utterance in the clean training and noisy testing sets is converted to a sequence of 39-dimensional feature vectors, each consisting of 13 MFCC (c1-c12, c0) and their first- and second-order derivatives. We operate either MVN or CGN first, and then the presented WD on these baseline features. With the features in the training set, we perform acoustic model training with the Hidden Markov Model Tool kit (HTK) [13]. The resulting acoustic models include 11 digit models (zero, one, two, three, four, five, six, seven, eight, nine and oh) and a silence model. Each digit model contains 16 states and 20 Gaussian mixtures per state.

The settings for the WD algorithm are as follows:

1. A 3-level discrete wavelet transform (DWT) is applied on the feature stream, resulting four octave sub-band streams with the ranges approximately [0, 6.25 Hz], [6.25 Hz, 12.5 Hz], [12.5 Hz, 25 Hz] and [25 Hz, 50 Hz].
2. The two higher sub-band streams (above 12.5 Hz) are processed by thresholding, while the two lower sub-band streams (below 12.5 Hz) are kept unchanged.
3. The wavelet function type is Daubechies 2 (db2).
4. The soft thresholding function with a heuristic SURE threshold value is selected.

The experimental results shown in this section are the outcome of the above parameter settings. We will vary these parameters and present the corresponding results in section V.

Tables 1 give the recognition accuracy rates for WD performing on the MVN/CGN-processed MFCC features, respectively. For comparison, the achieved accuracy rates for MVN and CGN, together with some other well-known methods, HEQ [4], MVA [11] and TSN [14], are listed in this table. From this table, we have some findings below:

1. Compared with the MFCC baseline, all the methods listed here improve the recognition accuracy by more than 13%, corresponding to a high relative error rate reduction of around 50%.
2. Compared with MVN alone, the pairing of WD and MVN gives better recognition accuracy rates for all the three Sets. Similar to the case of MVN, integrating WD and CGN achieves higher recognition performance than CGN alone. The overall averaged recognition accuracy is around 90%.
3. The three methods, HEQ, MVA and TSN, provide MVN with further accuracy promotion. TSN behaves the best, followed by MVA and then HEQ. The presented WD for MVN features performs better than HEQ and MVA and worse than TSN.
4. The WD for CGN features gives the optimal recognition accuracy among all the methods shown here. It outperforms TSN by 0.61% in averaged recognition accuracy. In summary, the presented WD for MVN and CGN features has similar effectiveness as the popular methods, HEQ, MVA and TSN.

Therefore, these results confirm the success of performing WD on MVN/CGN-processed features and support our proposal that WD helps alleviate the median and high (modulation) frequency distortions left behind by MVN and CGN.

	Set A	Set B	Set C	Avg	RR
MFCC baseline	71.89	68.24	77.57	71.57	-
MVN	85.05	85.62	85.70	85.41	48.68
WD with MVN	**88.29**	**89.07**	**88.62**	**88.67**	**60.15**
HEQ	86.91	88.32	87.50	87.59	56.36
MVA	88.12	88.81	88.50	88.47	59.46
TSN	89.42	90.03	89.03	89.59	63.37
CGN	87.64	88.55	87.73	88.02	57.87
WD with CGN	**89.81**	**90.75**	**89.89**	**90.20**	**65.54**

Table 1. The recognition accuracy rates (%) achieved by different methods. RR (%) stands for the relative error rate reduction.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose to apply the wavelet threshold denoising (WD) algorithm on the MVN and CGN features in order to improve noise robustness. The resulting recognition accuracy rates are significantly promoted relative to those achieved by MVN and CGN features. In future works, we will investigate if the presented approach can benefit the recognition in a large vocabulary task. Besides, we will perform WD on the feature streams processed by other normalization methods, such as higher-order cepstral moment normalization (HOCMN) [15] and cepstral shape normalization (CSN) [16], to see if the noise robustness can be further enhanced.

6. REFERENCES

[1] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 29, Apr 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," International Conference on Spoken Language Processing, 2002.

[3] Y. Shingo, H. Noboru, W. Naoya and M. Yoshikazu, "Cepstral gain normalization for noise robust speech recognition," International Conference on Acoustics, Speech and Signal Processing, 2004.

[4] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," IEEE Trans. on Audio, Speech and Language Processing, pp.845-854, 2006

[5] http://en.wikipedia.org/wiki/Wavelet

[6] S. Mallat and W. L. Hwang, "Singularity detection and processing with wavelets," IEEE Transactions on Information Theory, 1992.

[7] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," Biometrika, vol. 81, pp. 425-455, 1994.

[8] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, “On the importance of variousmodulation frequencies for speech recognition”, Eurospeech, 1997.

[9] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," The Annals of Statistics, 1981.

[10] http://www.mathworks.com/help/toolbox/wavelet/ref/wden.html

[11] C. Chen and J. Bilmes, "MVA processing of speech features," IEEE Transactions on Audio, Speech, and Language Processing, 2006.

[12] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," ISCA 2000.

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

[14] 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, 2008.

[15] C.W. Hsu and L. S. Lee, "Higher order cepstral moment normalization (HOCMN) for robust speech recognition," ICASSP 2004

[16] J. Du and R.Wang, "Cepstral shape normalization (CSN) for robust speech recognition," ICASSP 2008.