出國報告 （出國類別：參與國際會議）

第 14 屆國際語音通訊研討會

服務機關：國立暨南國際大學
姓名職稱：謝欣汝, 博士班研究生
派赴國家：法國里昂
出國期間：2013 年 8 月 25 日至 8 月 29 日
報告日期：2013 年 11 月 18 日

摘　要

本次出國前往參加於法國里昂所舉辦的第十四屆國際語音通訊研討會，其主要目的為吸取國際上鑽研語音學學門之學者交換其所得之知識與經驗，並藉由此次機會分享新穎創新之技術。此會議舉辦日期為 2013 年 8 月 25 日至 8 月 29 日為期五天，所探討的主題涵蓋所有和自然語言與語音處理相關的研究議題，例如:語音分析、自動語音辨識、聲音訊號源分離、語音合成等等，而本人所研究的主題偏重於語音訊號特徵擷取的部分，提出在語音訊號之調變頻譜上，運用知名的統計圖等化法以抵抗各式環境噪音源對於語音特徵擷取之影響，進而有效提升語音辨識系統的精確率。本次研討會安排 4 場主題演講和 50 個不同主題的口頭報告與 30 場不同主題的海報展示，內容十分豐富多元，藉此呈現全球卓越研究學者們最新的研究成果。經由參加此會議，瞭解語音學門最近的趨勢及走向，對於未來在研究題材上的選擇十分具有參考價值，同時也能提升自我的國際觀與外語能力。

目　　錄

頁數

一、 目的

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。而此國際會議提供一個平台，探討的主題涵蓋所有和自然語言與語音處理相關的研究議題，例如:語音分析、運用類神經網路之自動語音辨識、聲音訊號源分離、強健性語音特徵處理、語音合成、語言模型、聲學模型等等，藉此達到國際間學者相互交流與研究成果分享之目的，並致力開發語音相關之嶄新技術，祈望將語音學門相關之研究，發展至另一個高峰，且利用此次難得的機會充實自我對於語音學門相關領域的知識與增廣視野。

二、 參與會議之過程

『第 14 屆國際語音通訊研討會』為一年舉辦一次之國際性的語音處理研討會。此研討會在語音學門中，無論國內外皆擁有崇高的學術地位。因此國際間眾多的學者、研究人員皆會前往參加此會議。會議於 2013 年 8 月 25 日至 8 月 29 日於法國里昂之里昂會議中心(LYON CONVENTION CENTER)舉行(如照片 1)，由國際語音通訊學會(International Speech Communication Association, ISCA)所主辦。其主要目的為提供創新技術與具前瞻性研究一個學術交流平臺，透過國際間不同領域專長的學者及研究人員在知識與創新思維上的交流，以激發語音學門的蓬勃發展。本人很高興有這個難得的機會參與此會議並發表相關論文，主要行程如下：

表一 行程表

日期	工作事項
2013/08/21-25	個人行程
2013/08/26	啟程至法國里昂
2013/08/26-29	參與會議
2013/08/30	回程

會議第一天(8/25)是大會舉行 Tutorials 的時間，並無任何論文發表的安排，且正式的開幕儀式是在 8/26 舉行，因此這天並沒有參加。

會議第二天(8/26)於中午過後，趕搭巴黎至里昂的高速鐵路至開會地點，完成報到手續並領取大會議程、論文集與其他資料後，隨後聆聽幾場 Oral Section 和 Poster Section，分別有關『單通道之語音增強』、『聲音端點偵測和語音切割』相關的議題。其中在語音增強的 Section 中聽到幾篇論文是將『非負矩陣分解』技術應用至語音增強的領域，這個創新的點子非常令人耳目一新。另外在聲音端點偵測和語音切割的 Section 中，看到有考慮多通道長時間觀察語音的變異來達到良好的端點偵測，使用『有限狀態機』的技術來提高端點偵測的效果，這些研究題目真的非常的有趣。此外還稍微觀看了有關於『語

音的產生與語音知覺』相關的議題，而這方面的研究是非常適用於語音學習與教學的領域。而這一天聽完幾場 Oral Section 的感想是覺得自己的英文聽力能力還有待加強，尤其是英國人和印度人的口音最讓我感到頭痛，而相反的，我個人覺得從觀看 Poster Section 的經驗中我還可以學到更多較為深入的一些專業知識，並且將學者們所提出的方法或技術應用在我的研究之中，例如可以將『非負矩陣分解』技術運用於我研究的語音特徵抽取的過程中，而且當下若對 Poster 裡的內容有任何問題，我也可以主動的去請教該篇論文的作者們，透過這種交流方式，可以讓我在短時間之內吸取到許多知識。

會議第三天(8/27)於上午聆聽一場跟我的研究最密切相關的 Oral Section，即是『雜訊強建性用於自動語音辨識』方面的研究，在這個 Section 中，有幾篇論文讓我印象深刻，例如有新加波著名學者李海州教授的團隊，跟我一樣的也於本次會議中發表了對於統計圖等化法改良方面的研究，李教授的團隊提出了一套有考慮到語音特徵之屬性的統計圖等化法，並且利用 k-means 分群法和 EM 訓練的方式，將一些機器學習的方式應用於統計圖等化法之改良上，此創新的想法令人耳目一新。此外，印度 IBM 研究單位也發表了一篇關於『平均值消去法的改良』，此方法解決了以往在語音辨識中，較短的語句因其統計資訊量不足，而造成辨識系統效能下降的缺點，並且將此方法改良成類似可以 Real-time 執行的演算法。下午稍微聆聽了兩場關於語者辨識和特徵擷取運用於自動語音辨識的 Oral Section，這兩場的內容和我的研究比較相近因此也較容易進入狀況。

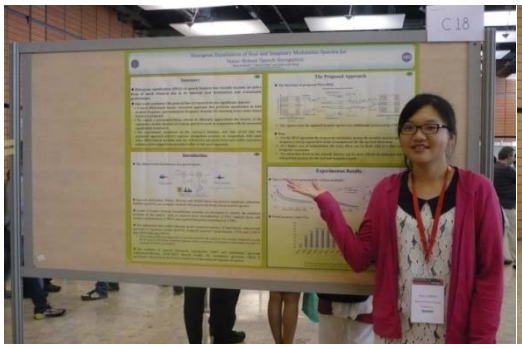
會議第四天(8/28)是大會排定我要報告的日期，由於我的報告時間是排在下午時段，因此早上還有以放鬆的心情稍微聆聽兩場 Oral Section，其內容主要是關於『語音分析』與『語者辨識』方面的議題，其中語者辨識這個題目是我非常想研究的題目之一，其主要目的為抽取一些和語者特性相關的一些特徵，例如:說話者聲帶振動的基本頻率(F0)和其音高的偵測及音高變化的曲線等等的資訊，來達到可以有效的自動辨認出說話者身份的技術，我覺得這方面的研究更為實用，例如:可以用於各行各業的保全系統中。下午到處看看其他學者所發表的論文後，整理一下心情，即將換我上場了，可能因為之前有出國報告的經驗，因此報告時心情不那麼緊張，反而還有點期待敢快有人來觀看我的

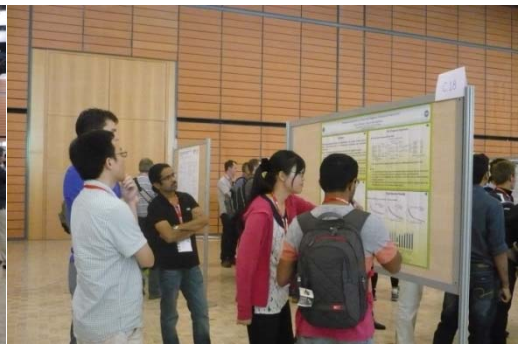
研究。在我報告的過程中，有許多位學者前來觀看我的研究並且給與我一些意見，也有學者給予我言語上的鼓勵，例如像前文所提到的新加波著名學者李海州教授團隊的蕭博士，表達對我的研究內容感到興趣，也對我的研究結果感到驚豔，還跟我一起討論統計圖等化法，其往後的方展方向可以朝向調整語音辨識器的聲學模型參數，以使得統計圖等化法的效果再往上精進。在討論的過程中我很投入，也很開心可以認識一位志同道合的研究學者，由於明年第 15 屆的國際語音通訊研討會，其舉辦的地點正好是在李海州教授所任教的大學舉辦，因此蕭博士也邀請我來參加明年的會議，並且歡迎我到新加坡玩，我感到非常的榮幸。此外，在我報告的同時，還見到幾位去年同樣在此會議中認識的日本籍研究生前來觀看我的研究並且跟我做學術上的意見交流，他也對於我將調變頻譜之實部和虛部分開處理的方式感到特別新奇，並且肯定我在研究上的發現與突破，我覺得參加國際研討會額外的好處就是，可以見到以往曾經在學術上一起交流過心得的研究生或學者們，雖然在研究這條路上，辛苦與孤單的感覺時常存在著，但看到別人都還在努力，自己也要加倍努力才行。

會議第五天(8/29)早上聆聽了語者辨識和語音增強的議題，其中在語音辨識的 Section 中有許多國際大型的研究單位都相繼提出新的語者辨識系統，這些系統其改主的地方就是加了許多不同的辨識環境與不同通道效應特性，來讓語者辨識相關的演算法及研究議題，在此系統上可以得到更全面性的衡量與評估方法之好壞。緊接著聆聽下午的強建性語音辨識的研究議題和幾場 poster 後則搭車回旅館休息。經由這幾天參加會議的過程中，可以感覺到，對於目前語音處理領域的研究中，會把許多其他領域會用到的技術整合在語音處理中，或者是將處理的範圍更加的擴大，例如像有些大型的研究，都是將前端的語音特徵處理技術再結合後端的聲學模型調整或是語言模型參數的調適來達到辨識器效能的提升。在回旅館的途中天色已暗，夜晚使里昂的風景更加迷人，索性搭上末班纜車前往具有『山上的大象』之稱的著名教堂參觀及觀賞里昂美麗的夜景。隔天整理好心情，下午時段驅車前往里昂機場，返程至桃園國際機場。其會議舉行地點及參與報告之過程如照片 1 所示。

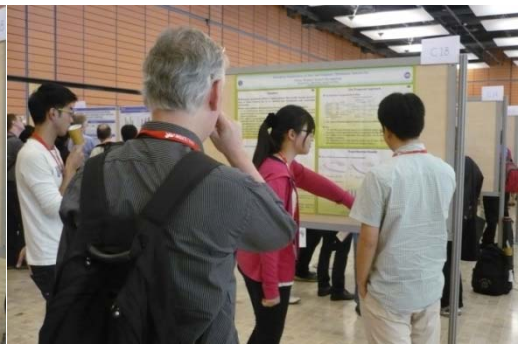
(a)

(b)

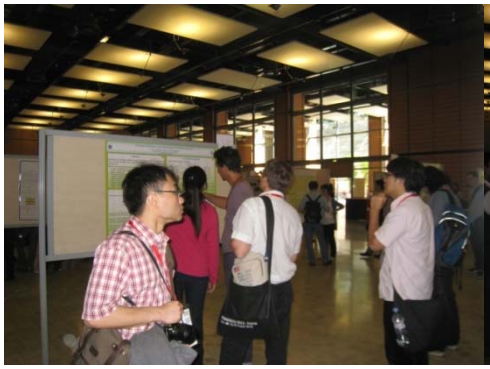




(c) (d)



(e) (f)



(g) (h)

照片 1、 (a)研討會舉行地點(b)會議立牌前留影(c)~(h)發表論文之過程

三、 心得

很榮幸有這次出國參加國際研討會的機會，這也是我第一次來到法國，法國一直以來是我非常想去的國家之一，不僅可以很開心的與世界各國的頂尖學者分享本人最新的研究成果，還可以與其他研究學者們互相做學術上的意見交流，內心真的感到十分充實。在進行論文發表的過程中，有許多位學者拿著照像機將我的研究成果拍照帶回去，其中還夾雜幾位印度籍的學者，心裡非常開心，一直以來我都很崇拜印度人有很好的數理能力。在與他人做學術成果分享的同時，也有許多位知名學者給予我言語上的鼓勵，對於我所提出的方法感到非常有興趣，經由這些寶貴的經驗讓我對研究的信心度又往上提升了。 這次去一趟法國，感覺法國是個非常先進且浪漫的國度，無論街道上或地底下，大眾運輸工具的高度發展，不僅讓當地人感到方便且讓我們這些大老遠來到的旅客也能享受這便利的生活環境。此外法國到處都有 Wi-Fi 的熱點可以免費供人使用，真的是非常便利。提到法國里昂最著名的就是到處充滿著非常真實的壁畫，例如:名人牆、圖書館壁畫等等，都真實到不注意看的話不會發現他是用畫上去的。參加國際研討會，對我們這些研究生來說真是一大福氣，除了可以見到以往只能在論文上面看到的作者本尊，還可以跟那些大師做面對面的學術交流。此外，還可以欣賞當地風土民情，受到異國文化的洗禮，讓自我的國際觀也更加充實，並且提升自我的語文能力。

四、 建議

對於本次參與國際會議之經驗，有以下建議提供參考：

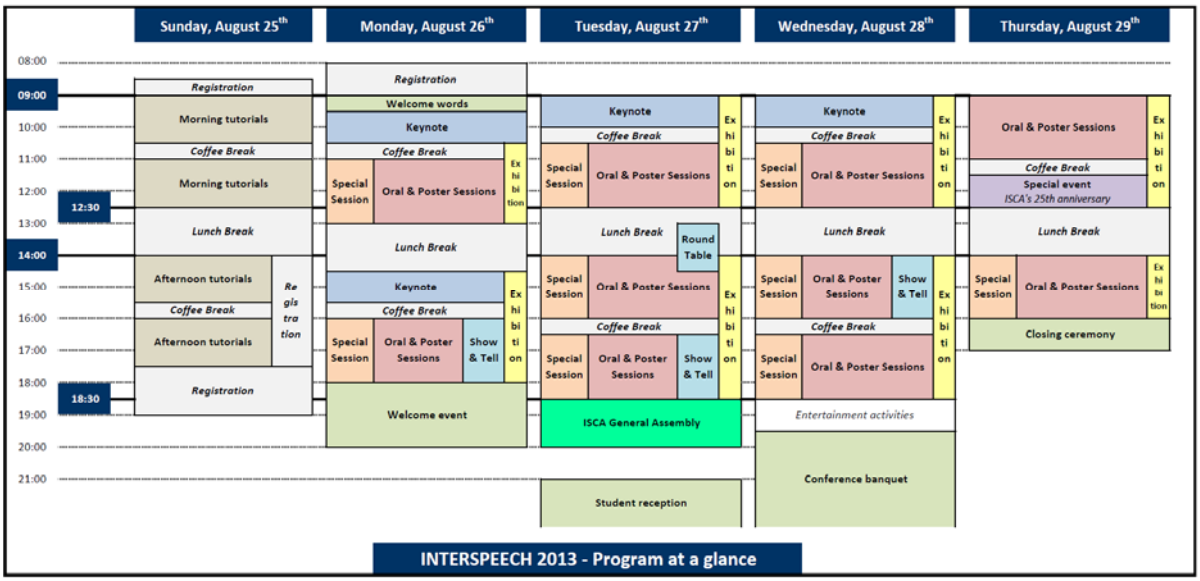
1. 增加補助出國經費之管道

   以博士班學生而言，現行之申請出國經費補助，分別有教育部和國科會這兩種管道。如這兩者申請皆未通過，等同於是博士生需自費前往，這對於出國發表論文的學生來說在無形間卻變成一種負擔。因此如果可以增加申請管道，相信一定可以達到鼓勵學生出國參與國際會議之效果。

2. 視情況增加補助額度

   申請出國經費補助通常其補助的費用主要為機票的部分費用，但幾天下來旅館住宿費用很有可能也會花到上萬塊錢，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。

五、 附錄


Sunday, August 25th
Monday, August 26th
Tuesday, August 27th
Wednesday, August 28th
Thursday, August 29th
08:00
09:00
10:00
11:00
12:00
12:30
13:00
14:00
15:00
16:00
17:00
18:00
18:30
19:00
20:00
21:00
Registration
Morning tutorials
Coffee Break
Morning tutorials
Lunch Break
Afternoon tutorials
Coffee Break
Afternoon tutorials
Registration
Registration
Welcome words
Keynote
Coffee Break
Special Session
Oral & Poster Sessions
Exhibition
Lunch Break
Keynote
Coffee Break
Special Session
Oral & Poster Sessions
Show & Tell
Exhibition
Welcome event
Keynote
Coffee Break
Special Session
Oral & Poster Sessions
Exhibition
Lunch Break
Round Table
Special Session
Oral & Poster Sessions
Exhibition
Coffee Break
Special Session
Oral & Poster Sessions
Show & Tell
Exhibition
ISCA General Assembly
Student reception
Keynote
Coffee Break
Special Session
Oral & Poster Sessions
Exhibition
Lunch Break
Special Session
Oral & Poster Sessions
Show & Tell
Exhibition
Coffee Break
Special Session
Oral & Poster Sessions
Exhibition
Entertainment activities
Conference banquet
Oral & Poster Sessions
Exhibition
Coffee Break
Special event
ISCA's 25th anniversary
Lunch Break
Special Session
Oral & Poster Sessions
Exhibition
Closing ceremony
INTERSPEECH 2013 - Program at a glance

Histogram Equalization of Real and Imaginary Modulation Spectra for Noise-Robust Speech Recognition

Hsin-Ju Hsieh[1,2], *Berlin Chen*[2] *and Jeih-weih Hung*[1]

[1]National Chi Nan University, Taiwan
[2]National Taiwan Normal University, Taiwan
s101323902@ncnu.edu.tw, berlin@ntnu.edu.tw, jwhung@ncnu.edu.tw

Abstract

Histogram equalization (HEQ) of acoustic features has received considerable attention in the area of robust speech recognition because of its relative simplicity and good empirical performance. This paper presents a novel HEQ-based feature extraction approach that performs equalization in both acoustic frequency and modulation frequency domains for obtaining better noise-robust features. In particular, the real and imaginary acoustic spectra are first individually transformed to the modulation domain via discrete Fourier transform (DFT). The HEQ process is then carried on the corresponding magnitude modulation spectra so as to compensate for the noise distortions. Finally, the equalized modulation spectra are converted back to form the real and imaginary acoustic spectra, respectively. By doing so, we can enhance not only the magnitude but also the phase components of the acoustic spectra, and thereby create more noise-robust cepstral features. The experiments conducted on the Aurora-2 clean-condition database and task reveal that the presented approach delivers superior recognition accuracy in comparison with some other HEQ-related methods and the well-known advanced front-end (AFE) extraction scheme, which supports the potential utility of this novel approach.

Index Terms: noise robustness, feature extraction, modulation spectrum, histogram equalization, automatic speech recognition

1. Introduction

Varying environmental effects, such as ambient noise and interferences caused by the recording devices and transmission channels, often lead to severe mismatch between the acoustic environments for the training and testing speech data in automatic speech recognition (ASR), and this environmental mismatch inevitably degrades the performance of ASR dramatically [1]. Substantial efforts have been made and also a number of techniques have been developed to address this issue for improving the ASR performance in the past decades. Broadly speaking, these noise/interference processing techniques fall into three main categories: enhancement, normalization and adaptation [2], while these techniques can be conducted either in the speech feature domain or in the acoustic model domain.

Regarding the popular speech feature representation, Mel-frequency cepstral coefficient (MFCC), which reflects the spectral characteristics within a short period of time, has exhibited high discriminating capability for acoustic units and thus gives excellent recognition accuracy in nearly noise-free laboratory environments. However, MFCC is vulnerable to noise/interference and often requires compensation prior to being used in real-world scenarios. The compensation can be carried out in the various intermediate states during the extraction of the MFCC feature stream for a speech signal. Roughly speaking, according to the MFCC extraction procedure, a time-signal is segmented to a series of overlapping frames, and then each frame signal is transformed into the acoustic spectrum, next into the (linear) critical-band spectrum and the logarithmic critical-band spectrum, and eventually into the cepstrum. First of all, spectral subtraction [3,4], Wiener filtering [5] and MMSE-based log-spectral amplitude estimation (MMSE log-STSA) [6] are exemplary methods that process the frame-based acoustic spectra. Second, a suite of feature moment normalization methods are developed to regulate the statistical moments of the cepstra, such as cepstral mean normalization (CMN) [7], cepstral mean and variance normalization (CMVN) [8] and cepstral histogram normalization (CHN) [2, 9, 10], to name but a few. Since the statistical moments are directly evaluated by the temporal series of cepstra, these moment normalization methods implicitly enhance the cepstra in temporal characteristics. On the other hand, the approaches that employ filtering on the temporal sequence of logarithmic critical-band spectrum or cepstrum include, but are not limited to, RASTA [11], temporal structure normalization (TSN) [12] and CMVN plus ARMA filtering (MVA) [13]. These temporal-filtering approaches in general emphasize the relatively low varying components (except the DC part) of the feature temporal sequence, which encapsulate rich linguistic information cues that are conducive for speech recognition. Additionally, the methods of spectral histogram equalization (SHE) [14] and modulation spectrum replacement/filtering (MSR/MSF) [15] directly modify the modulation spectrum, which is specifically referred to as the Fourier transform of the temporal sequence of cepstra.

Our work in this paper presents a novel application of histogram equalization (HEQ) [9,10] to reduce the distortion of acoustic spectral features in modulation domain for speech recognition. Unlike the conventional HEQ approaches that often operate on the temporal stream of the cepstra (which is denoted by CHN earlier) or the Mel-filter smoothed logarithmic spectra [9, 10], the presented method performs HEQ on the DFT for the temporal series of acoustic spectra (i.e., the modulation spectra of the acoustic spectra) with respect to each acoustic frequency bin. Furthermore, the real and imaginary parts of the acoustic spectra are treated individually in the presented framework, and it is different from most well-known acoustic spectral-domain robustness techniques, such as spectral subtraction and Wiener filtering, that process the magnitude acoustic spectra directly. By and large, our presented approach has the following three advantages. First, via the HEQ operation the long-term

correlation among the acoustic spectra (at the same frequency) can be captured for the compensation of spectral distortion. Second, at a higher cost of computation, the noise effect can be dealt with in a finer (acoustic) frequency resolution. Third, the distortion dwelt in the acoustic spectra can be more extensively mitigated due to the independent process for the real and imaginary parts. All of the aforementioned advantages will be confirmed via empirical evaluation.

The rest of this paper is organized as follows. Section 2 provides the essential fundamentals for HEQ and briefly describes how it can be crystallized for robust ASR. Section 3 elucidates our proposed normalization framework. Then, the experimental settings and a series of ASR experiments conducted are presented in Sections 4 and 5, respectively. Finally, Section 6 concludes this paper and suggests avenues for future work.

2. Brief Introduction of HEQ

Histogram equalization (HEQ) that can effectively reduce the statistical mismatch between the training and testing data has been well studied and practiced in the field of pattern recognition. In the HEQ algorithm, an arbitrary data series, denoted by $\{x_1, x_2, \cdots, x_N\}$, is viewed as the sample set of a random variable X with a cumulative distribution function (CDF) $F_X(x)$. Then, via the mapping procedure:

$$y_i = F_T^{-1}(F_X(x_i)), \quad 1 \le i \le N \tag{1}$$

the CDF of another random variable Y with the obtained new data series $\{y_1, y_2, \cdots, y_N\}$ as samples can approximate a predefined target CDF $F_T(y)$ as long as the number of data, *N*, is sufficiently large. The target CDF $F_T(y)$ is usually set to be simply a standard Gaussian distribution with zero mean and unity variance, or approximated by the histogram of the training data.

More recently, HEQ has been adopted to compensate for speech features for noise-robust ASR. The CHN [2, 9, 10] and SHE [14] methods mentioned in the previous section are two good instantiations developed along this line of thought, which operate HEQ on the temporal domain and modulation domain of MFCC features, respectively.

3. Proposed Approach

This section describes a novel HEQ-based feature extraction framework in an attempt to improve the noise robustness of speech features. First, in the preprocessed stage, any utterance $x[\ell]$ in the training and testing sets is shaped by a high-pass pre-emphasis filter, and framing as well as windowing operations are performed in turn. Then, each windowed frame signal is transformed to the acoustic frequency domain via short-time Fourier transform (STFT), and the resulting *complex-valued* acoustic spectrum is denoted by

$$X[n,k] = X_r[n,k] + jX_i[n,k], \quad 0 \le n \le N-1,\ 0 \le k \le K-1 \tag{2}$$

where $X_r[n,k]$ and $X_i[n,k]$ denote the acoustic real and imaginary spectra, respectively, *n* and *k* respectively refer to the indices of frame and discrete frequency, and *N* and *K* are respectively the numbers of frames and acoustic frequency bins. By the way, $\{X[n,k]\}$ in Eq. (2) is sometimes called the spectrogram of the utterance. Next, the acoustic real and imaginary spectra, $X_r[n,k]$ and $X_i[n,k]$ in Eq. (2), *with respect to a fixed frequency bin k* are updated via the subsequent steps.

Step I: Compute the modulation spectrum separately for $X_r[n,k]$ and $X_i[n,k]$ along the n-axis by discrete Fourier transform (DFT) as follows

$$\mathcal{X}_r[k,m] = \sum_{n=0}^{N-1} X_r[n,k] e^{-j\frac{2\pi nk}{N}} \quad \text{and} \quad \mathcal{X}_i[k,m] = \sum_{n=0}^{N-1} X_i[n,k] e^{-j\frac{2\pi nk}{N}}, \quad 0 \le m \le N-1,\ 0 \le k \le K-1. \tag{3}$$

where m refers to the index of the discrete modulation frequency. The resulting spectra can be expressed in polar form as

$$\mathcal{X}_r[k,m] = \mathcal{A}_r[k,m] e^{j\theta_r[k,m]} \quad \text{and} \quad \mathcal{X}_i[k,m] = \mathcal{A}_i[k,m] e^{j\theta_i[k,m]}, \tag{4}$$

where $\mathcal{A}_r[k,m]$ and $\mathcal{A}_i[k,m]$ are, respectively, the magnitude component of $\mathcal{X}_r[k,m]$ and $\mathcal{X}_i[k,m]$, and $\theta_r[k,m]$ and $\theta_i[k,m]$ are, respectively, the phase component of $\mathcal{X}_r[k,m]$ and $\mathcal{X}_i[k,m]$.

Step II: Update the magnitude components of the modulation spectra via HEQ, while keeping the phase components unchanged. The resulting new magnitude modulation spectra are expressed by

$$\tilde{\mathcal{A}}_r[k,m] = F_{T_r}^{-1}\left(F_{\mathcal{A}_r}(\mathcal{A}_r[k,m])\right) \quad \text{and} \quad \tilde{\mathcal{A}}_i[k,m] = F_{T_i}^{-1}\left(F_{\mathcal{A}_i}(\mathcal{A}_i[k,m])\right), \tag{5}$$

where the cumulative distribution functions (CDFs) $F_{\mathcal{A}_r}$ and $F_{\mathcal{A}_i}$ are estimated from those $\mathcal{A}_r[k,m]$ and $\mathcal{A}_i[k,m]$ of the utterance being processed, and the inverse CDFs $F_{T_r}^{-1}$ and $F_{T_i}^{-1}$ are from those $\mathcal{A}_r[k,m]$ and $\mathcal{A}_i[k,m]$ of the utterances in the clean training set. As such, combining the updated magnitude components with the original phase components results in the new modulation spectra

$$\tilde{\mathcal{X}}_r[k,m] = \tilde{\mathcal{A}}_r[k,m] e^{j\theta_r[k,m]} \quad \text{and} \quad \tilde{\mathcal{X}}_i[k,m] = \tilde{\mathcal{A}}_i[k,m] e^{j\theta_i[k,m]}. \tag{6}$$

Step III: Construct the new acoustic real and imaginary spectra, denoted by $\tilde{X}_r[n,k]$ and $\tilde{X}_i[n,k]$, respectively, by taking the inverse DFT of $\tilde{\mathcal{X}}_r[k,m]$ and $\tilde{\mathcal{X}}_i[k,m]$, in Eq. (6). Accordingly, we obtain the modified *complex-valued* acoustic spectrum as

$$\tilde{X}[n,k] = \tilde{X}_r[n,k] + j\tilde{X}_i[n,k]. \tag{7}$$

At the final stage, the processing is the same as in the case of MFCC extraction: the magnitude of the modified acoustic spectrum $\{\tilde{X}[n,k]\}$ in Eq. (7) associated with each frame is weighted by a Mel-frequency filter bank, and the nonlinear compression is achieved by using the logarithmic operation. Lastly, the less correlated MFCC features are derived after the application of the discrete cosine transform (DCT).

Because the main idea of the aforementioned framework is to perform HEQ on the ***m***odulation domain of the ***a***coustic ***s***pectrum, we will use the short-hand notation "MAS-HEQ" to denote it hereafter.

The MAS-HEQ framework has two remarkable characteristics:

1. In MAS-HEQ, the real and imaginary acoustic spectra are processed individually, which helps to enhance the magnitude and phase parts simultaneously. Note that the modulation-domain HEQ process cannot operate on the magnitude acoustic spectra directly (that is, to perform HEQ on the *magnitude* modulation spectrum of $|X[n,k]|$ in Eq. (2))

because the resulting new magnitude acoustic spectra are real-valued, but not necessarily nonnegative.

2. As for the comparison between MAS-HEQ and the well-practiced CHN (HEQ performing on the cepstral time series), MAS-HEQ focuses on equalizing the distribution of the data *at different modulation frequencies*, while CHN equalizes the distribution of the data *at different time indices*. Furthermore, MAS-HEQ bears some resemblance to the SHE technique [14] since both of them are operated on the modulation domain. However, the modulation spectrum processed by SHE is the DFT of the cepstra rather than the DFT of the acoustic spectra.

In this paper, we also leverage a polynomial-fitting scheme (denoted by PHEQ) [10] to efficiently approximate the inverse CDFs, $F_{T_r}^{-1}$ and $F_{T_i}^{-1}$ in Eq. (5), to work in concert with the presented MAS-HEQ. PHEQ provides the advantages of lower storage and time consumption when compared with the existing HEQ methods. It makes effective use of data fitting (or so-called least squares error regression) to estimate the inverse CDFs of the training data.

The notion of processing real and imaginary components of acoustic spectra in the modulation domain for speech enhancement has been investigated recently [16]. However, to our knowledge, there is still a dearth of work investigating the effectiveness of normalizing the real and imaginary components of acoustic spectra in the modulation domain for speech recognition. As will be shown in Section 5, such a joint normalization paradigm shows promise and performs quite well.

4. Experimental Setup

The speech recognition experiments were conducted under various noise conditions using the Aurora-2 database and task [17]. The Aurora-2 database is a subset of the TI-DIGITS, which contains a set of connected digit utterances spoken in English; while the task consists of the recognition of the connected digit utterances interfered with various noise sources at different signal-to-noise ratios (SNRs), in which the Test Sets A and B are artificially contaminated with eight different types of real world noises (e.g., the subway noise, street noise, etc.) in a wide range of SNRs (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and Clean) and the Test Set C additionally includes the channel distortion.

As for the baseline experiment, each utterance of the training and testing sets were converted to a series of 39-dim MFCC feature vectors (*c*0, *c*1-*c*12 plus their delta and delta-delta). The frame length and shift were set to 25 ms and 10 ms, respectively. In particular, each of the robustness algorithms to be evaluated is to produce the 13 static cepstra (*c*0, *c*1-*c*12) only, and then the 26 dynamic cepstra are computed accordingly.

More specifically, the acoustic model for each digit was a left-to-right continuous density HMM with 16 states, and each state has a 20-mixture diagonal GMM. The training and recognition tests used the HTK recognition toolkit [18], which followed the setup originally defined for the ETSI evaluations. All the experimental results reported below are based on clean-condition training, i.e., the acoustic models were trained only with the clean (uncontaminated) training utterances.

5. Experimental Results

At the outset, we evaluate the utility of MAS-HEQ in terms of recognition accuracy. For the purpose of comparison, the results of some well-known feature robustness methods are also reported here. These methods are roughly divided into two categories depending on the feature type to be adjusted directly:

1. Acoustic spectrum processing methods: ETSI advanced front-end (AFE) [19], MMSE-based log-spectral amplitude estimation (MMSE log-STSA) [6], Wiener filtering (WF) based on a priori signal-to-noise-ratio estimation [5] and two versions of spectral subtraction (SS) [3,4], denoted by SS_{Boll} and $SS_{Berouti}$ for short, respectively, in which the author names are represented by the subscripts.
2. Cepstrum processing methods: cepstral mean normalization (CMN) [7], cepstral mean and variance normalization (CMVN) [8], cepstral histogram normalization (CHN) [2], cepstral gain normalization (CGN) [20], CMVN plus ARMA filtering (MVA) [13], spectral histogram equalization (SHE) [14] and temporal structure normalization (TSN) [12].

In particular, we additionally perform CMN on the cepstral features produced by any of the acoustic spectrum processing methods, including the presented MAS-HEQ. Note that the CMN procedure has been inherently embedded in all of the cepstrum processing methods.

Table 1 shows the recognition accuracy rates for the various methods, from which we notice several particularities:

1. It comes as no surprise that every method can give rise to significant improvement in recognition rates for all the three test sets as compared to the MFCC baseline. The simple CMN process can achieve a relative error rate reduction of 32.12%, and all the other methods that integrate the CMN process produce even better results relative to CMN alone.
2. As for the cepstrum processing methods, SHE behaves the best, followed by TSN, MVA, CHN, CGN and then CMVN. There are several noteworthy points. First, CHN outperforms CMVN due to its further normalization on the statistical moments higher than the second order. Second, the constraint of unity dynamic range for CGN eliminates the outliers in the resulting data and makes it behaves as well as CHN. Next, MVA explicitly enhances the low time-varying components of CMVN features with a fixed ARMA filter and performs very well, while TSN, which employs a data-driven temporal filter, produces better results than MVA. Finally, the better outcome of SHE compared with CHN implies histogram equalization (HEQ) conducted in the modulation domain of cepstra provides superior robustness than in the temporal domain.
3. Regarding the acoustic spectrum processing methods, WF and two variants of SS behave less effective than the other spectral-domain methods possibly because they are initially designed for speech enhancement. MMSE log-STSA performs specifically well for Set C and the corresponding overall results are close to the best possible ones achieved by cepstrum processing methods. The cepstra derived from the well-known AFE without further CMN processing achieves an accuracy rate of 87.17%, higher than those obtained by any other methods discussed before. Nevertheless, CMN is not well additive to AFE probably due to the effect of over-normalization on the AFE-derived features. Finally, the presented MAS-HEQ turns out to be the best-performing one

among all of the tested methods in terms of the overall averaged recognition accuracy. Compared with AFE, MAS-HEQ is better for Test Set B and worse for Test Sets A and C. In brief, MAS-HEQ shows excellent performance in creating noise-robust speech features.

4. MAS-HEQ outperforms SHE consistently over different Test Sets, and on average, the respective accuracy improvement is around 3%. These results indicate that when considering the effectiveness of processing speech features in modulation domain via HEQ, the acoustic spectra seem to be a better choice than the cepstra. However, MAS-HEQ is less efficient than SHE in implementation since the number of the (discrete) acoustic spectra in MAS-HEQ is larger than that of the cepstra in SHE.

Apart from recognition performance, we also examine the presented MAS-HEQ with regard to its capability of reducing the mismatch in the power spectral density (PSD) of the cepstral sequence caused by noise. Figs. 1(a) to 1(d) depict the averaged PSD curves of the first MFCC feature *c*1 for the 1001 utterances in the Test Set B of the Aurora-2 database for three SNR levels, clean, 10 dB and 0 dB (with airport noise) before and after various processes (CMN, AFE and MAS-HEQ), respectively. First, for the unprocessed case as in Fig. 1(a), it shows that the noise causes a significant PSD mismatch over the entire modulation frequency band [0, 50 Hz]. Second, by comparing Fig. 1(b) with Fig. 1(a) we find that CMN just eliminates the distortion at the DC component and provides nearly no benefit for the PSD mismatch at any other frequency (even so, CMN can bring about significant accuracy improvement, as evident in Table 1). Finally, Figs. 1(c) and 1(d) show that both AFE and the presented MAS-HEQ can considerably reduce the PSD distortion, while MAS-HEQ appears more effective than AFE to mitigate the PSD mismatch at higher frequencies. These results may partly explain why MAS-HEQ outperforms AFE for processing Test Set B, as shown in Table 1, and they also reveal that MAS-HEQ can provide a more noise-robust feature representation.

6. Conclusions

In this study, we have proposed a novel noise-robustness framework, termed MAS-HEQ, for equalization of the acoustic spectra in modulation domain. Applying histogram equalization on the magnitude parts of the DFTs for the real and imaginary acoustic spectra separately enables MAS-HEQ to reduce the noise effect effectively and refine the resulting features elaborately. The experimental results conducted on Aurora-2 demonstrate that MAS-HEQ can provide superior performance over many state-of-the-art robustness methods, including the ETSI advanced front-end (AFE). As to future work, we envisage several directions, including extending the idea of our work to process the Mel-filter smoothed (complex-valued) spectra, analyzing the possible addition of our work with more other robustness methods and further confirming our observations on larger-scale ASR experiments.

Table 1. *Recognition accuracy rates (%) averaged over different noise types and different SNRs for the baseline MFCC and various robustness methods. RR (%) is the relative error rate reduction over the MFCC baseline.*

	Set A	Set B	Set C	Avg	RR
MFCC baseline	54.87	48.87	63.95	54.29	-
Cepstrum processing methods					
CMN	66.81	71.79	67.64	68.97	32.12
CMVN	75.93	76.76	76.82	76.44	48.46
CHN	80.03	82.05	80.10	80.85	58.11
CGN	80.08	81.48	80.20	80.66	57.69
MVA	80.89	82.00	81.49	81.45	59.42
TSN	83.26	84.50	82.83	83.67	64.27
SHE	83.37	85.08	83.47	84.08	65.17
Acoustic spectrum processing methods					
SS_{Boll}	73.03	76.84	73.00	74.55	44.32
$SS_{Berouti}$	78.70	82.81	79.69	80.54	57.43
WF	79.64	81.39	80.29	80.47	57.27
MMSE log-STSA	82.96	83.95	84.60	83.68	64.30
$^{\dagger}AFE_{(1)}$	87.68	87.10	86.27	87.17	71.93
$^{\dagger}AFE_{(2)}$	85.53	86.59	85.47	85.94	69.24
MAS-HEQ	86.89	88.66	85.33	87.29	72.19

$\dagger AFE_{(1)}$ denotes the original AFE, and $AFE_{(2)}$ denotes the pairing of AFE and CMN. Note: The CMN process is integrated with all of the methods except for $AFE_{(1)}$.

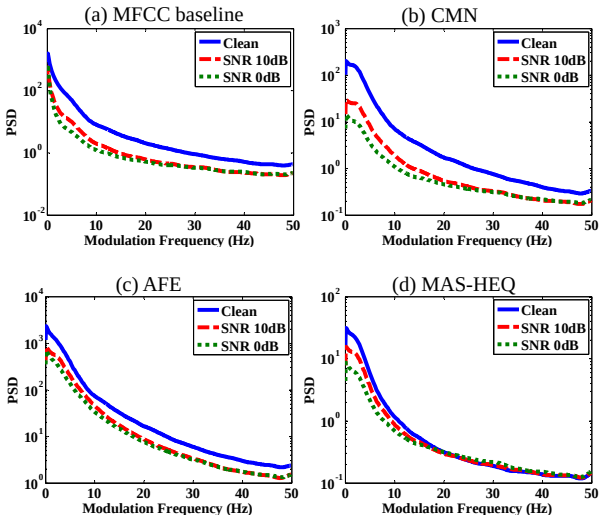


Figure 1. *The MFCC c1 PSD curves processed by various compensation methods: (a) the MFCC baseline (no compensation), (b) CMN, (c) AFE and (d) MAS-HEQ*

7. Acknowledgement

This work was sponsored in part by "Aim for the Top University Plan" of National Taiwan Normal University and Ministry of Education, Taiwan, and the National Science Council, Taiwan, under Grants NSC 101-2221-E-003 -024 -MY3 and NSC 99 -2221-E-003 -017 -MY3.

8. References

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communications*, 16, pp. 261-291, 1995.

[2] J. Droppo and A. Acero, "Environmental robustness," in *Springer Handbook of Speech Processing*, Chapter 33, pp. 653-679, 2008.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2), pp. 113-120, 1979.

[4] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 208-211, 1979.

[5] C. Plapous, C. Marro and P. Scalart, "Improved signal-to-noise ratio estimation for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, 14(6), pp. 2098-2108, 2006.

[6] 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*, 33(2), pp. 443-445, 1985.

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

[8] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communications*, 25(1-3), pp. 133-147, 1998.

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

[10] 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*, 17 (1), pp. 84-94, 2009.

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

[12] X. Xiao, E. S. Chng and H. Z. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, 16(8), pp. 1662-1674, 2008.

[13] C. P. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), pp. 257-270, 2007.

[14] L. C. Sun and L. S. Lee, "Modulation spectrum equalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3), pp. 828-843, 2012.

[15] J. W. Hung, W. H. Tu and C. C. Lai, "Improved modulation spectrum enhancement methods for robust speech recognition," *Signal Processing*, 92(11), pp. 2791-2814, 2012.

[16] Y. Zhang and Y. Zhao, "Spectral subtraction on real and imaginary modulation spectra," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4744-4747, 2011.

[17] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 2000 Automatic Speech Recognition: Challenges for the new Millenium*, pp. 181-188, 2000.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book* (for HTK Version 3.4), Cambridge University Engineering Department, Cambridge, UK, 2006.

[19] D. Macho, L. Mauuary, B. Noé, 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 *Proceedings of the Annual Conference of the International Speech Communication Association*, pp. 17-20, 2002.

[20] S. Yoshizawa, N. Hayasaka, N. Wada and Y. Miyanaga, "Cepstral gain normalization for noise robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 209-212, 2004.