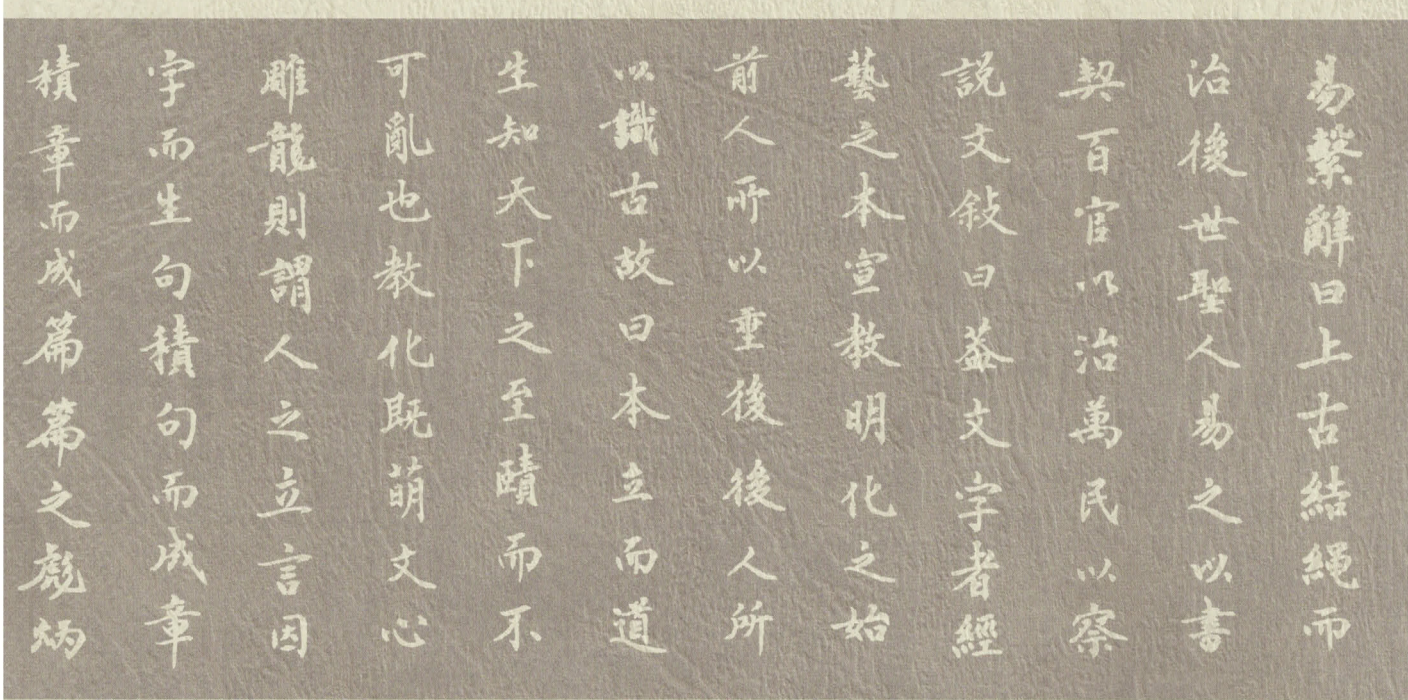
International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI, Linguistics Abstracts, and ACL Anthology.



Vol.21 No.2 December 2016 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

The Association for Computational Linguistics and Chinese Language Processing, Taipei

International Journal of

Computational Linguistics & Chinese Language Processing

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 21, no. 2, December 2016

Contents

Papers

基於詞語分布均勻度的核心詞彙選擇 [A Study on Dispersion Measures for Core Vocabulary Compilation]…….…. 1
白明弘(Ming-Hong Bai),吳鑑城(Jian-Cheng Wu),簡盈妮(Ying-Ni Chien),黃淑齡(Shu-Ling Huang),林慶隆(Ching-Lung Lin)

N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings...……... 19
Yu-Ming Hsieh and Wei-Yun Ma

使用字典學習法於強健性語音辨識 [The Use of Dictionary Learning Approach for Robustness Speech Recognition]……... 35
顏必成(Bi-Cheng Yan),石敬弘(Chin-Hong Shih),劉士弘(Shih-Hung Liu),陳柏琳(Berlin Chen)

評估尺度相關最佳化方法於華語錯誤發音檢測之研究 [Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection]..... 55
許曜麒(Yao-Chi Hsu),楊明翰(Ming-Han Yang),洪孝宗(Hsiao-Tsung Hung),林奕儒(Yi-Ju Lin), 陳冠宇(Kuan-Yu Chen),陳柏琳(Berlin Chen)

基於字元階層之語音合成用文脈訊息擷取 [Character-Level Linguistic Features Extraction for Text-to-Speech System]…… 71
陳冠宏(Kuan-Hung Chen),廖書漢(Shu-Han Liao),廖元甫(Yuan-Fu Liao),王逸如(Yih-Ru Wang)

融合多任務學習類神經網路聲學模型訓練於會議語音辨識之研究 [Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition]……. 85
楊明翰(Ming-Han Yang),許曜麒(Yao-Chi Hsu),洪孝宗(Hsiao-Tsung Hung),陳映文(Ying-Wen Chen), 陳冠宇(Kuan-Yu Chen),陳柏琳(Berlin Chen)

Reviewers List & 2016 Index…….. 105

Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 1-18

基於詞語分布均勻度的核心詞彙選擇

A Study on Dispersion Measures for Core Vocabulary Compilation

白明弘*、吳鑑城*、簡盈妮*、黃淑齡*、林慶隆*

Ming-Hong Bai, Jian-Cheng Wu, Ying-Ni Chien,

Shu-Ling Huang and Ching-Lung Lin

摘要

核心詞彙是不受文本類型、主題、應用情境等影響，穩定使用的詞彙。在自然語言中，核心詞彙的數量相對稀少，卻構成溝通內容的主要部份，因此是語言學習中重要的一環。傳統的核心詞彙選擇方法主要依據專家知識與經驗法則，在語料庫語言學興起後，詞頻與詞彙分布均勻度統計提供了客觀的統計數據協助核心詞彙的選取。在本論文中，我們提出一個多面向均勻度整合公式，使詞語均勻度的計算能夠同時考慮到不同的分類面向。其次，我們也針對傳統公式統計結果偏差的問題，提出詞頻正規化的方法。對於實驗的評估，我們提出了一個以異源語料庫評估核心詞彙的方法，可以比較各種統計公式的優缺點與特性。在實驗結果的部份，我們實際比較了多種不同的核心詞彙表選擇公式，分析不同公式的特質，並驗證了詞頻正規化的確能夠修正傳統公式的缺點。最後，我們也驗證了整合多面向均勻度的計算方法，確實可以選擇到更具核心特質的詞彙。

* 國家教育研究院編譯發展中心
Development Center for Compilation and Translation, National Academy for Educational Research
E-mail: wujc@mail.naer.edu.tw
The author for correspondence is Jian-Cheng Wu.

Abstract

Core vocabulary is a set of words that are stable used across different text types, theme, and application scenario. In natural language, the number of core vocabulary is relatively small, the core vocabulary, however, plays an important part in language learning because it constitutes a major part of communication content. The traditional core vocabulary selection method is mainly based on the expert knowledge and rule of experience. With the rise of corpus linguistics, word frequency and dispersion uniformity provide objective statistical data to assist the selection of core vocabulary. In this paper, we propose a formula that integrates multi-dimensional uniformity , so that the estimation of word uniformity can take different classification dimensions into account. Secondly, we also propose a method of word frequency normalization for the problem of deviation of the traditional method. For evaluation, a method of evaluating the core vocabulary with a heterogeneous corpus is proposed and it can compare the advantages, disadvantages, and characteristics of various statistical formulas. In the results, we actually compare the different core vocabulary selection formulas, analyzed the characteristics of different formulas, and verified the word frequency normalization can correct the shortcomings of the traditional formula. Finally, we also verified that the proposed method which integrates multi-dimensional uniformity can pick out the vocabulary with more core characteristics.

關鍵詞：語料庫語言學、核心詞彙、邊緣詞彙、分布均勻度。

Keywords: Corpus Linguistics, Core Vocabulary, Fringe Vocabulary, Dispersion Uniformity.

1. 緒論

核心詞彙(core vocabulary)是指一組不受文本類型、主題、應用情境等影響，穩定使用的詞彙(Huang, Zhang & Yu, 2005)。這些詞彙的穩定性是多方的，除了跨類別、主題、應用情境之外，還包括跨年齡層、性別等性質(Stuart, 1991)。這些詞彙相對於非核心詞彙(邊緣詞彙, fringe vocabulary)來說數量較稀少，卻構成溝通內容的主要部份(Vanderheiden & Kelso, 1987)。在語言的使用上，當一個句子缺乏邊緣詞彙時，雖難以確切指稱物品，但仍足以傳達說話者的主要意涵(Liu, 2012)，因此核心詞彙是語言學習中重要的一環。核心詞彙除了被應用在語言教學之外也應用在詞典編輯、輔助溝通系統、比較語言學等領域(Juilland & Chang-Rodríguez, 1964; Carroll, 1970; Juilland, Brodin & Davidovitch, 1970; Rosengren, 1971; Huang *et al*., 2005; Liu, 2012)。

傳統的核心詞彙選擇方法主要依據專家知識與經驗法則(Huang *et al*., 2005)，語料庫語言學興起後，統計式的方法逐漸取代經驗法則。然而從語料庫的觀察中可以發現，單

純使用詞頻無法分離核心詞彙與邊緣詞彙，許多高頻的詞彙只在特定的情境下高頻出現。我們在中研院平衡語料庫(後簡稱平衡語料庫)中觀察四個詞頻接近的詞在不同主題中的分布情況（如圖 1），「*網路*」在大部份主題中都屬低頻詞，只有在科學主題之下才大量出現。「*企業*」則在社會及科學主題中大量出現。相對而言「*今天*」和「*一定*」在各類主題中出現的次數較為平均。在此例中，前兩個詞語屬邊緣詞彙，後兩個詞語屬核心詞彙。由此例的觀察可以發現詞語的核心程度與分布均勻度有高度的相關，因此許多研究者提出以分布均勻度來衡量詞語的核心程度(Juilland & Chang-Rodríguez, 1964; Carroll, 1970; Juilland *et al.*, 1970; Rosengren, 1970; Huang *et al.*, 2005; Liu, 2012)，當分布程度越均勻時，該詞語的核心程度就越高。

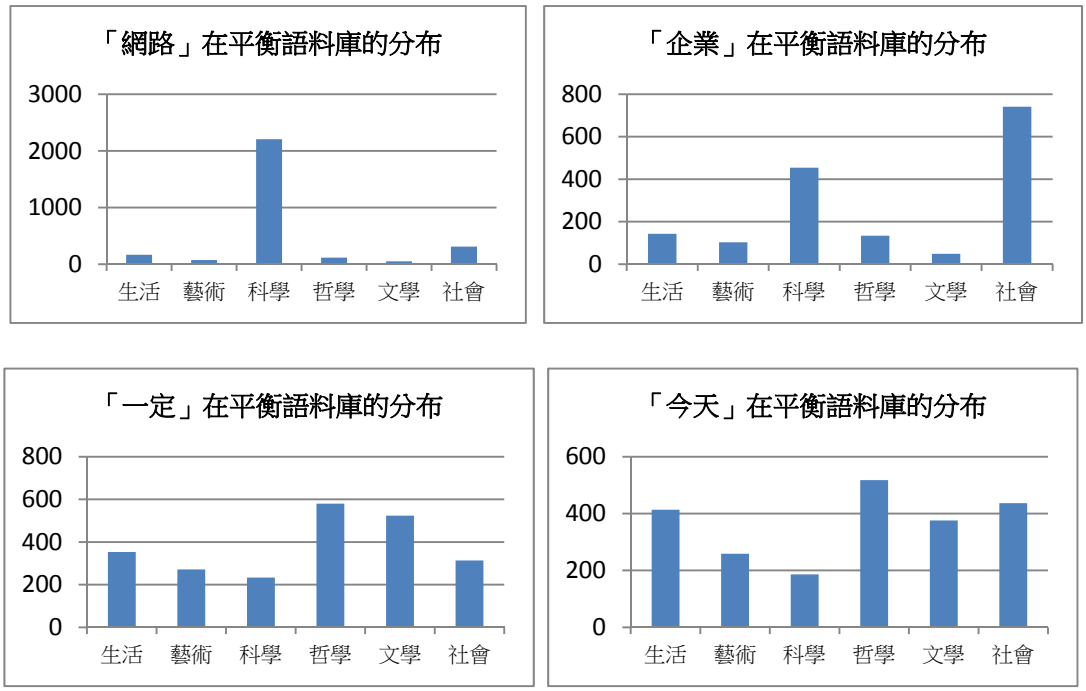


圖 1. 觀察四個詞頻接近的詞語在平衡語料庫不同主題中的分布（單位：次數/每百萬詞)
[Figure 1. The Observation of the distribution of four words with similar frequency in different topics in Sinica corpus (unit: words/per million words)]

詞語分布均勻度確實是衡量詞語核心程度的有效指標，然而，均勻度的計算具有多重分類面向。以上述的例子（圖 1）而言，均勻度的計算面向為主題，統計所得到的均勻度，也僅限於在主題中分布的均勻程度，無法保證在其他面向(如時間、語式、文體、媒體等)的均勻度。Huang *et al.* (2005)曾提到從不同分類面向計算詞語的均勻度的概念，不過該文並未提出整合不同面向均勻度的具體方法。

在本論文中，我們首先提出一個整合多面向均勻度的方法，使詞語均勻度的衡量能夠同時考慮不同的分類面向。其次，我們提出詞頻正規化的方法來修正傳統公式遇到分類區塊大小不一造成統計結果偏差的問題。最後，我們提出了一個以異源語料庫評估核心詞彙庫的方法，以比較各種均勻度公式的優缺點與特性。

以下為本論文的結構說明：我們將在第 2 節中介紹詞語分布均勻度的計算方法；在第 3 節中提出核心詞彙庫的驗證方法；第 4 節將說明實驗的設計、實驗的結果與數據分析；在第 5 節中我們將討論使用均勻度公式在選擇詞彙的一些特質；在第 6 節我們將為本文做結論。

2. 詞語分布均勻度的計算

2.1 語料區塊的切分

計算詞語分布均勻度前必須先將語料庫切分成幾個區塊，然後再計算詞語在區塊中的分布是否均勻。在這樣的計算程序中，每個區塊代表一個語言使用情境的實例，當分布均勻度越高時，即表示詞語受情境的影響越小。因此語料區塊的切分方法將關係到核心詞彙選擇的結果。

語料區塊切分的方法大致可分成隨機切分、依篇章、時間或主題類別切分等。隨機切分的優點是方法簡單、不需要倚賴額外的訊息，且仍有不錯的效果。不過有幾個問題必須注意：首先，當語料本身收錄的文章主題不平衡時，收錄篇數較多的主題可能分布在較多區塊中，因而造成此主題的高頻詞均勻度被高估的問題。其次，在切分區塊時必須選擇適當的區塊大小，因為過小的區塊及過大的區塊都會造成均勻度計算的結果偏差。以極端的例子來說，當區塊極小到以句子為單位時，則均勻度序列將趨近於頻率序列，亦即，以均勻度選擇的詞彙庫將近似於以詞頻選擇的詞彙庫。此問題的原因在於，單一詞語通常在單一句子中很少重覆出現，因此，當一個詞的頻率越高時，必然出現在越多句子之中，因而計算出來的均勻度也越高。另一個極端例子是當區塊極大時，例如只切割成 5 個區塊，假設 *a* 詞在 5 個區塊中均勻地出現數萬次，而 *b* 詞在各區塊中恰巧都低頻出現過。依均勻度來看 *a* 詞和 *b* 詞可能有相近的均勻度，這種現象在區塊數極少時發生的機率很高。

如果依篇章、時間或主題分類來切分區塊，則必需考慮到幾個問題。首先，語料中每個區塊的大小差異可能很大。以平衡語料庫為例，社會主題所收錄的文章數量大約是藝術主題的 4 倍之多。然而，過去研究所提出的均勻度公式大多只考慮詞語在區塊間的詞頻差異，而沒有考慮每個區塊的相對大小。忽略區塊的大小差異的問題將造成詞頻分布均勻度的偏差，亦即原本分布均勻的詞語，卻因為區塊大小不一，而造成均勻度低估的現象。我們將在 2.3 節提出詞頻正規化的方法來解決這個問題。其次，語料的分類方式非常多種，以平衡語料庫來說，目前的分類方式就有主題、文類、媒體、語式、文體、子主題、時間等不同的切割區塊方法，以不同分類切分語料代表以不同面向來觀察詞語的均勻度。我們將在 2.4 節中提出整合多重均勻度的計算方式。

2.2 均勻度公式

為了建立語言學習的參考詞表，過去有很多種均勻度公式被提出來。Juilland & Chang-Rodríguez(1964) 以及 Juilland *et al.*(1970)都以標準差的概念為基礎，提出分布均勻度(Juilland's coefficient of dispersion, JD) 定義如下：

$$JD = 1 - \frac{V}{\sqrt{n-1}} \tag{1}$$

其中的變數說明如下：

n: 語料庫切分區塊數

V: 詞語在區塊中分布的變異數：$V = \frac{\sigma}{\bar{f}}$

σ: 詞語在區塊中分布的標準差：$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(f_i - \bar{f})^2}{n-1}}$

f_i: 詞語在第 *i* 區的詞頻

$\bar{f}$: 詞語在區塊中的平均詞頻，即$\bar{f} = F/n$

F: 詞語在語料中的總詞頻

Juilland 提出分布均勻度的目的是為了適度調整從語料庫中統計的詞頻，因為高頻的詞語未必都是重要的詞語，尤其是只出現在特定主題的詞彙。所以 Juilland 認為選擇學習參考詞表，除了考量詞頻之外，同時也要考量詞語被廣泛使用在不同的主題類別。Carroll (1970) 認同 Juilland 以分布均勻度調整詞頻的觀點，但認為以標準差為基礎所設計的均勻度並不是一個好的評估方法。他提出以訊息熵(entropy)為基礎的分布均勻度公式(Carroll's coefficient of dispersion, CD) 定義如下：

$$CD = \frac{H}{log_2(n)} \tag{2}$$

其中的變數說明如下：

H： 訊息熵 $H = logP - \frac{\sum_{i=1}^{n} p_i \, log_2 p_i}{P}$

p_i：詞語在第 i 區塊中的分布比例，$p_i = f_i/s_i$

s_i：第 i 區塊總詞數

P：$\sum_{i=1}^{n} p_i$

Rosengren (1971) 則調整方均根公式，提出新的均勻度公式(Rosengren's coefficient of dispersion, RD)如下：

$$RD = \frac{\left(\sum_{i=1}^{n}\sqrt{f_i}\right)^2}{n} \cdot \frac{1}{F} \tag{3}$$

其中的變數說明如下：

n: 語料庫切分區塊數

f_i: 詞語在第 *i* 區的詞頻

F: 詞語在語料中的總詞頻

Lyne(1985) 則依據 chi-square 為基礎提出均勻度公式(Lyne's coefficient of dispersion, LD)如下：

$$LD = \frac{1-\chi^2}{4F} \tag{4}$$

Huang *et al.*(2005)為了預測與驗證 Swadesh(1952) 所提出的基本概念表(basic concepts)，提出一個均勻度為基礎的方法。他們所提出的均勻度公式 (Distributional Consistency, DC)如下：

$$DC = \left(\frac{\sum_{i=1}^{n}\sqrt{f_i}}{n}\right)^2 \cdot \frac{1}{f} \tag{5}$$

基本上 Huang 等人所提出的均勻度公式 DC 與 Rosengren 所提出的 RD 相同，不過兩個研究者的目的不同，Rosengren 的 RD 公式是作為調整詞頻的參數之一，為讓調整後的詞頻能夠同時具備高使用率與跨主題均勻分布特性。而 Huang *et al.* (2005)則直接使用 DC 值做為 Swadesh 基本概念表的預測指標。

2.3 公式的使用限制與正規化

在 2.1 節中曾提到，依篇章、主題、時間、語式等分類將語料切分成區塊，則區塊所包含的詞數差異可能很大，進而造成分布均勻度的偏差。例如：「開始」在平衡語料庫中的分布大約每百萬詞出現 600 次，而且在不同文類中，分布十分穩定。但因為平衡語料庫六個主題所收錄的文章數量差異很大，而使得「*開始*」在各主題中出現的頻率差異也很大（見表 1）。

如果我們以每百萬詞出現頻率來看就會發現，「*開始*」在每個文類中都以非常穩定的分布出現。因此，在使用均勻度公式之前，我們可以先將頻率正規化。頻率正規化(frequency normalization)的公式如下：

$$u_i = \frac{f_i \cdot 10^6}{N_i} \tag{6}$$

其中 N_i 表示區塊 i 收錄的總詞數。在計算均勻度時，可將原公式中的詞頻數 f_i 改成正規化後的詞頻 u_i ，但需注意的是，與詞頻有關的參數亦需要調整。我們以 DC 公式的正規化為例來說明，正規化後的公式(Normalized Distributional Consistency, NDC)定義如下：

$$NDC = \left(\frac{\sum_{i=1}^{n}\sqrt{u_i}}{n}\right)^2 \cdot \frac{1}{\bar{u}} \quad (7)$$

其中 $\bar{u}$ 為詞語在所有類別中的平均詞頻，也必需依正規化的詞頻重新計算，即 $\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n}$。

表 1. 「開始」在平衡語料庫各主題出現的頻率與每百萬詞頻率
[Table 1. The original frequency and the frequency in per million words of "Kāi shǐ" in topics of Sinica corpus]

主題	文類大小	頻率	每百萬詞頻率
藝術	846,593	624	737.07
生活	2,243,362	1,471	655.71
文學	2,234,564	1,481	662.77
科學	1,128,083	632	560.24
哲學	1,126,081	698	619.85
社會	3,624,244	2,113	583.02
總計	1,202,927	7,019	626.53

2.4 整合不同面向均勻度

在前面曾提到，將語料庫依不同主題類別切分後計算均勻度，即代表以不同面向來觀察詞語的核心程度。但是在過去的研究中，並沒有提出一套整合不同面向均勻度的方法。所以，在這一節中，我們嘗試使用模糊集合論(Zadeh, 1965; Kwakernaak, 1978)的角度來解釋核心詞彙。

假設 U 表示中文詞彙的模糊宇集合，$S_j \subset U$ 是在某一分類面向的分布均勻模糊集合。則某一詞語 x 屬於 S_j 的均勻度可以表示成 $\mu_{S_j}(x)$ 。我們以 NDC 均勻度公式當作詞語分布均勻度的評估方法，我們將 x 屬於 S_j 的程度定義如下：

$$\mu_{S_j}(x) = NDC_{S_j}(x) \quad (8)$$

其中，$NDC_{S_j}(x)$ 表示以 NDC 均勻度公式計算在 S_j 主題類別面向的均勻度。

我們要計算詞語 x 整合多個分類面向的均勻分布率，亦即是要計算 x 同時屬於多個模糊集合的程度即 $\mu_{\{\cap_{j=1}^{m} S_j\}}(x)$ ， 其中 m 表示有 m 種模糊集合，亦即有 m 種語料庫切分的方法，我們假設均勻模糊集合之間為獨立事件，則依據模糊集合論，可以表示為：

$$\mu_{\cap_{j=1}^{m} S_j}(x) = \prod_{j=1}^{m} \mu_{S_j}(x) = \prod_{j=1}^{m} NDC_{S_j}(x) \quad (9)$$

這個結果代表當要計算多面向均勻度時，在模糊集合獨立性的假設下，我們可以將多面向的均勻度(Multi-dimensional Normalized Distributional Consistency, MNDC) 定義為每個面向均勻度的乘積，亦即：

$$MNDC(x) = \mu_{\cap_{j=1}^{m} S_j}(x) = \prod_{j=1}^{m} NDC_{S_j}(x) \quad (10)$$

以實際的例子來說明，假設要計算詞語在篇章、主題、時間三個面向的均勻度，計算詞語 x 在這三個面向的均勻度可表示為：

$$MNDC(x) = NDC_{篇章}(x) \cdot NDC_{主題}(x) \cdot NDC_{時間}(x) \quad (11)$$

3. 核心詞彙表驗證

為了驗證不同核心詞彙抽取公式的效果，我們使用兩個獨立建置的語料庫來分別進行核心詞彙的抽取與驗證，一、來源語料庫：用來抽取核心詞彙的語料庫。二、驗證語料庫：用來驗證核心詞彙的核心程度。

在驗證的方法上，依照過去研究對核心詞彙的定義，大致可歸納出兩種特性，第一、跨情境特性，第二、詞彙重要性。在跨情境特性的驗證上，我們提出核心詞彙在驗證語料庫中不同主題下的再利用率定義如下：

$$Reuse\ Rate(C) = \frac{1}{n}\sum_{i=1}^{n} \frac{|C \cap L(T_i)|}{|C|} \quad (12)$$

其中 C 表示從來源語料抽取的核心詞彙表，T_i 表示驗證語料庫 T 中的主題 i 語料，$L(T_i)$ 表示構成 T_i 的詞表。公式的意義是檢驗核心詞彙表在驗證語料庫的不同情境中，是否都能保持很高的使用率。

再利用率可用來評估詞表的跨主題特性，但不保證這些詞在使用上是重要的。所以，我們以語料庫覆蓋率來驗證詞彙的重要性，定義如下：

$$Coverage(C) = \frac{\#\ of\ tokens\ in\ T\ which\ was\ covered\ by\ C}{\#\ of\ tokens\ in\ T} \quad (13)$$

當詞表在語料庫中的覆蓋率越高時，表示詞表中的詞彙涵蓋了語料庫中較為重要的詞語和概念。

4. 實驗設計與結果

4.1 核心詞彙來源語料庫

在實驗的設計上，我們使用國家教育研究院建置的華語文語料庫(Corpus of Contemporary Taiwanese Mandarin , 簡稱 COCT)中的書面語語料作為核心詞彙抽取的來源語料庫(柯華葳等人，2016)。COCT 書面語語料庫蒐集來源以圖書為主，目前共包含 1 億 1,220 萬字。這些文章在分類上包含出版年份(1986-2015 年)、書籍冊別，書籍主題(10 類)，我們利用篇章、書籍冊別、書籍主題、與出版年份四個分類面向來計算詞彙的分布均勻度。

4.2 驗證語料庫

在核心詞彙的驗證上，我們採用中央研究院平衡語料庫 4.0 版，該語料庫包含了約 1,000 萬詞，每一份文本都標示主題、文類、媒體、語式及文體等不同的後設資料。(Chinese Knowledge Information Processing Group [CKIP], 1995)

4.3 不同均勻度公式比較

在第一個實驗中，我們驗證了不同均勻度公式在核心詞彙的抽取效果。在來源資料庫切分上，我們採用隨機區塊切分法，將 COCT 書面語語料庫隨機切分成包含 3,000 字的小區塊，然後使用不同公式計算每個詞語的分布均勻度、詞頻或調整詞頻，公式說明如下：

Freq: 詞語在語料庫中的頻率。

Entropy: 以訊息熵公式計算均勻度值。

DC: 以 Distributional Consistency 計算詞語的均勻度值。

JD: Juilland 的分布均勻度。

JU: Juilland 的詞頻調整法。

CD: Carroll 的分布均勻度。

CU: Carroll 的詞頻調整法。

LD: Lyne 的詞頻調整法。

RAF: Rosengren 的詞頻調整法。

在實驗中，我們分別使用上列的公式計算來源語料庫中詞彙的分布均勻度、詞頻或調整詞頻，然後將每個詞表依照公式評估值由高而低排序，最後，在每個詞表中取出前 10,000 個詞做為核心詞彙表抽取的結果。

在核心詞彙抽取效果的比較上，我們使用每個公式統計出來的結果各取前 10,000 詞當作詞彙表，並以平衡語料庫的六個主題來驗證詞彙表的再利用率。表 2 為 9 個公式抽取的詞彙表的再利用率驗證結果。從結果來看，我們可以發現 Entropy , DC, CD 三個方法所抽取的詞彙表，在平衡語料庫中都有較高的再利用率，這表示這三個方法所抽出來的詞彙，具有高度跨主題性質。

為進一步觀察這些詞的重要性，我們評估這些詞彙表在平衡語料庫中的覆蓋率(如表 3)。在此表中我們發現 Freq, DC, JU, CU, RAF 這五個方法的覆蓋率都很高。其中 RAF 和 CU 能夠最有效覆蓋驗證語料庫，由此可知使用均勻度來調整詞頻的確可以更精確地找出重要性高的詞彙。不過，使用 RAF 及 CU 所抽出詞彙表的再利用率則不及 DC，這表示 RAF 及 CU 的結果包含了較多的非核心詞彙。綜合來說，從再利用率的角度觀察，Entropy、DC、CD 三種公式在核心詞彙的抽取上有最佳的核心性質，而從覆蓋率的

角度觀察，這三個公式的語料庫覆蓋率相較於詞頻及調整詞頻的公式並無太大的差異，亦即這三個公式在詞彙的選擇上能夠兼顧到詞彙的核心性與重要性。

表 2. 各核心詞彙表在平衡語料庫的不同主題中的再利用率
[Table 2. The re-utilization rate of each core vocabulary in different topics in Sinica corpus]

	生活	藝術	科學	哲學	文學	社會	平均
Freq	0.969	0.920	0.897	0.927	0.972	0.973	0.943
Entropy	**0.992**	0.946	**0.924**	**0.967**	**0.999**	**0.996**	0.971
DC	**0.992**	**0.947**	**0.924**	0.966	0.998	**0.996**	0.970
JD	0.991	0.940	0.916	0.964	0.998	0.995	0.967
JU	0.972	0.924	0.900	0.930	0.975	0.975	0.946
CD	**0.992**	0.946	**0.924**	**0.967**	**0.999**	**0.996**	0.971
CU	0.978	0.931	0.906	0.939	0.980	0.981	0.952
LD	0.991	0.939	0.915	0.964	0.998	0.994	0.967
RAF	0.988	0.943	0.918	0.956	0.993	0.992	0.965

表 3. 不同均勻度公式在驗證語料庫中的覆蓋率
[Table 3. The coverage of different uniformity formulas in verification corpus]

	生活	藝術	科學	哲學	文學	社會	整體覆蓋率
Freq	0.809	0.828	**0.833**	0.893	0.853	0.814	**0.832**
Entropy	0.806	0.826	0.826	0.893	0.855	0.809	0.829
DC	0.807	0.828	0.829	0.894	**0.856**	0.811	**0.831**
SD	0.788	0.806	0.818	0.872	0.829	0.792	0.810
JD	0.798	0.819	0.816	0.887	0.852	0.799	0.821
JU	0.809	0.829	**0.833**	0.894	0.854	0.814	**0.832**
CD	0.806	0.826	0.826	0.893	0.855	0.809	0.829
CU	**0.810**	0.830	**0.833**	0.894	0.855	**0.815**	**0.833**
LD	0.798	0.818	0.815	0.887	0.852	0.797	0.820
RAF	**0.810**	**0.831**	**0.833**	**0.895**	**0.856**	**0.815**	**0.833**

4.4 正規化的實驗

在第二個實驗中，我們以 DC 及 Entropy 兩個公式為例，比較詞頻正規化前計算均勻度及正規化後計算均勻度的效果。在來源資料庫的切分上，我們採用書籍與主題分類兩種切分方式，因為這兩種切分方式的區塊中包含的詞數大小較不一致。書籍切分法即是以每一冊書籍視為一個區塊將 COCT 書面語語料庫切分成區塊；而主題分類切分法依 10 個主題類別[1] 將 COCT 書面語語料庫切分成 10 個區塊，再使用不同公式計算每個詞語跨區塊的分布均勻度。在均勻度的計算上，DC 及 Entropy 兩個公式都分別計算詞頻正規化前的均勻度與詞頻正規化後的均勻度(Normalized DC 與 Normalized Entropy)。實驗結果同樣於排序後取前 10,000 個詞作為詞彙表，並驗證詞彙表的再利用率。

表 4 為以書籍切分法抽取詞彙表的結果，為了觀察，我們將驗證語料庫切分成生活、藝術、科學、哲學、文學、社會六個類別。由表 4 我們可以發現四個詞彙表在生活、文學及社會三類文本中的再利用率都很高，所以正規化前後沒有差別，而藝術、科學及哲學類再利用率較低，正規化後有微幅的改善。整體來說，正規化之後有微幅的改善，但差距不甚明顯。這代表以單冊書籍為切分區塊不致產生太大的偏差。

表 4. 詞頻正規化前後的再利用率比較，以書籍為語料庫切分區塊
[Table 4. The comparison of re-utilization rate before and after word frequency normalization(using book as a block unit)]

	生活	藝術	科學	哲學	文學	社會	平均
Entropy	0.991	0.928	0.901	0.962	0.999	0.994	0.962
Normalized Entropy	**0.992**	**0.933**	**0.906**	**0.965**	0.999	0.994	**0.965**
DC	0.994	0.943	0.917	0.970	0.999	0.996	0.970
Normalized DC	0.994	**0.945**	**0.919**	0.970	0.999	0.996	**0.971**

表 5 為以主題類別切分法抽取詞彙表的結果，由結果可以發現以主題分類當作切分區塊來計算均勻度，所抽取的詞彙再利用率較書籍區塊切割來得差。經過正規後，Normalized Entropy 在六個類別中的再利用率都比未正規化的 Entropy 有明顯的提高，平均改進達到 5%。而 Normalized DC 的再利用率也比未正規化的 DC 提高 3-4%。這結果一方面顯示了切分區塊過少時，較難選出具核心特性的詞彙；另一方面，以主題分類作為區塊切割，將造成區塊大小的落差較大，所以不論是 Entropy 或是 DC 公式，先將詞頻正規化，結果都能得到大幅的改善。

由此結果發現以下情形： 一、Entropy 比 DC 更容易受到語料庫切分不均的影響，所以 Entropy 正規化後效果提昇比 DC 來得明顯。二、切分區塊數極少的時候，區塊大小不均的落差較大，所造成的影響程度較為嚴重，所以此時正規化的效果在兩個方法中都極為明顯。三、生活、文學及社會類型文本對核心詞彙的再利用率較藝術、科學及哲

[1]參考中文圖書分類法並經調整後分成 10 類：總類，商業及金融類，哲學及宗教類，史地類，藝術類，語言文學類，社會科學類，應用科學類，科學類，休閒類。

學來得高。

表 5. 詞頻正規化前後的再利用率比較，以主題分類為語料庫切割區塊
[Table 5. The comparison of re-utilization rate before and after word frequency normalization(using topic as a block unit)]

	生活	藝術	科學	哲學	文學	社會	平均
Entropy	0.919	0.812	0.810	0.798	0.905	0.936	0.863
Normalized Entropy	**0.967**	**0.873**	**0.839**	**0.893**	**0.975**	**0.973**	**0.920**
DC	0.945	0.858	0.848	0.841	0.937	0.957	0.898
Normalized DC	**0.974**	**0.894**	**0.862**	**0.909**	**0.980**	**0.980**	**0.933**

4.5 多面向整合的效果

在第三個實驗中，本研究比較單一面向均勻度所抽出的詞表和整合多面向均勻度所抽出詞表的效果差異。在來源資料庫的切分上，採用了四種切分法，分別計算 NDC 均勻度：

Text NDC: 隨機切成 3,000 字的區塊，計算 NDC 均勻度。

Book NDC: 以書籍為單位，切分成 850 個區塊，計算 NDC 均勻度。

Class NDC: 以主題分類切分成 10 個區塊，計算 NDC 均勻度。

Year NDC: 以出版年份分成 29 個區塊，計算 NDC 均勻度。

MNDC: 將上列四種均勻度值以 MNDC 公式計算出整合均勻度。

在詞表的選取上，對於每一個方法，我們同樣依均勻度值將詞語排序，並抽取前 10,000 個詞作為詞彙表，同樣以平衡語料庫來驗證這 5 個詞彙表。

表 6 以再利用率評估詞表抽取的結果，從結果中我們可以發現，MNDC 公式所得到的詞表有最佳的再利用率，這表示整合多個面向的均勻度(MNDC 公式)的確可以選出核心程度最高的詞彙。

表 7 則是以語料庫覆蓋率來評估整合公式，從結果可以發現, MNDC 公式依然有很高的語料庫覆蓋率，但 Text NDC 所選出來的詞表，在語料庫覆蓋率上則高於 MNDC。這代表多面向均勻度(MNDC)在詞彙的選擇上，比較傾向於高核心程度的選擇，而放棄一些頻率高而核心程度較低的詞。

表 6. 多面向均勻度整合的詞表再利用率
[Table 6. The re-utilization rate of multi-dimensional uniformity integrated word list]

	生活	藝術	科學	哲學	文學	社會	平均
Text NDC	0.992	0.946	0.924	0.965	0.998	0.996	0.970
Book NDC	0.994	0.945	0.919	0.970	0.999	0.996	0.971
Class NDC	0.974	0.894	0.862	0.909	0.980	0.980	0.933
Year NDC	0.992	0.937	0.914	0.966	0.996	0.996	0.967
MNDC	**0.996**	**0.951**	**0.928**	**0.973**	**0.999**	**0.997**	**0.974**

表 7. 多面向均勻度整合的語料庫覆蓋率
[Table 7. The corpus coverage rate of multi-dimensional uniformity integrated word list]

	生活	藝術	科學	哲學	文學	社會	整體覆蓋率
Text NDC	**0.807**	**0.828**	**0.829**	**0.894**	**0.856**	**0.811**	**0.830**
Book NDC	0.795	0.818	0.808	0.888	0.853	0.794	0.818
Class NDC	0.734	0.752	0.735	0.836	0.795	0.731	0.757
Year NDC	0.782	0.799	0.799	0.881	0.836	0.786	0.807
MNDC	0.803	0.824	0.821	0.893	**0.856**	0.805	0.825

由於詞表再利用率與語料庫覆蓋率兩個評估值存在取捨的關係，所以從表 6 及表 7 中並無法明顯確認MNDC 的效果。為了觀察詞表再利用率與語料庫覆蓋率的變化關係，本研究觀察當詞表的取詞數從 1,000 逐漸增加到 40,000 詞，詞表再利用率與語料庫覆蓋率的變化關係。從圖 2 可以發現在相同的語料庫覆蓋率基準之下，MNDC 比 Text NDC 及其他單一面向均勻度有穩定的再利用率。

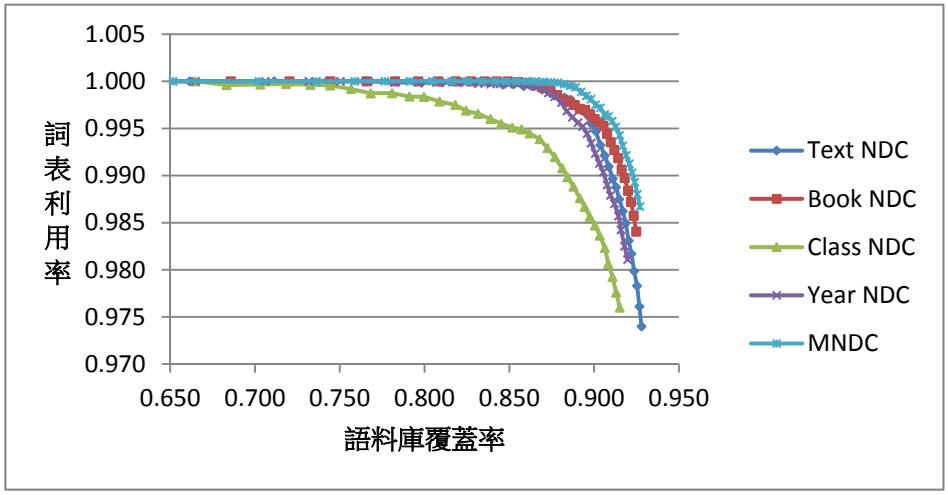


圖 2. 語料庫覆蓋率與詞表利用率關係圖
[Figure 2. The relationship between corpus coverage and word list utilization rate]

5. 分析與討論

過去的研究普遍認為以均勻度抽取詞表，會收錄較多的功能詞。本節中，我們將探討均勻度對詞彙選擇的實際影響。

首先，我們分別觀察以詞頻與均勻度公式收取 10,000 詞的詞表時，兩個詞表的詞類比例差異比較，結果如表 8。從表中我們可以發現，均勻度公式所收錄的名詞比詞頻法大量減少了約 1,000 詞，因為大部份的名詞多少都具有使用情境的特質，分布較不均勻。因為名詞收詞量的減少，所以其他詞類的收錄稍有增加，這是遞補的現象。只有副詞(ADV)、及物動詞(Vt)與不及物動詞(Vi)有較為明顯的增加。從這個結果中，並未顯示均勻度的選詞比詞頻選詞更偏好功能詞。

其次，我們把詞表的收詞數減少為 500 詞來觀察依均勻度收詞的詞類分布(表 9)，我們發現收詞數很少時，副詞(ADV)、 對等連接詞(C), 連接詞(POST)、時態標記(ASP)、介詞(P) 等詞類的收詞比例高很多。相對來說，名詞(N)、及物動詞(Vt)、不及物動詞(Vi)的收詞比例少很多。從表 9 的結果我們可以發現，當收詞量偏少時，均勻度對功能詞的偏好現象十分明顯。這一方面是因為功能詞在分布上的確是比較不受語境的影響，另一方面，功能詞的詞頻也比一般詞來得高。

在語言教學參考詞彙表的編輯上，以均勻度選擇詞彙表的確可以收錄比較穩定的跨領域核心詞彙。但是對於初階語言教學來說，因為初階所收錄的詞彙量較少，統計所得的詞彙表將包含較大比例的功能詞。對初階學習者來說，功能詞的語義模糊而抽象、歧義性高、語法規則複雜(Klammer, Schulz & Volpe, 2009)，所以不宜在初階詞表中收錄過多的功能詞。以華測會所編輯的華語八千詞表(Steering Committee for the Test of Proficiency - Huayu[SC-TOP], 2016) 為例，入門級所收錄的 500 詞中，名詞佔了極高的比例(約 60%)。而在基礎級中，名詞收錄比例降到約 50%左右。進階、高階及流利三級的名詞收錄則維持在約 40% ，可見初階語言教學的確需要包含較多的名詞。因此，不論是均勻度或是詞頻所選出來的詞表，在初階教學使用上都還需要經過收詞比例的調校以增加實詞的比例。

表 8. 前 10,000 詞收詞詞類比較
[Table 8. The comparison of Parts of Speech of the top 10,000 words]

	依詞頻收錄詞數	依均勻度收錄詞數	依詞頻收錄之百分比	依均勻度收錄之百分比	收錄差異百分比
ADV/副詞	634	755	6.34%	7.55%	19.09%
A/(非謂)形容詞	96	92	0.96%	0.92%	-4.17%
C/連接詞	111	119	1.11%	1.19%	7.21%
ASP/時態標記	10	11	0.10%	0.11%	10.00%
DET/定詞	191	222	1.91%	2.22%	16.23%

M/量詞	169	168	1.69%	1.68%	-0.59%
N/名詞	4,752	3,749	47.52%	37.49%	-21.11%
P/介詞	102	108	1.02%	1.08%	5.88%
Vt/及物動詞	2,355	2,811	23.55%	28.11%	19.36%
T/語助詞	54	48	0.54%	0.48%	-11.11%
Vi/不及物動詞	1,479	1,869	14.79%	18.69%	26.37%
POST/後置詞	47	48	0.47%	0.48%	2.13%

表 9. 依均勻度收錄前 500 詞與收錄 10,000 詞的詞類分布比較
[Table 9. The comparison of the distribution of the Parts of Speech of top 500 and 10,000 words according to uniformity]

	收錄 10,000 詞詞類分布	**收錄 500 詞詞類分布**
ADV/副詞	7.55%	21.60%
A/（非謂）形容詞	0.92%	0.60%
C/連接詞	1.19%	7.00%
ASP/時態標記	0.11%	1.00%
DET/定詞	2.22%	8.80%
M/量詞	1.68%	4.00%
N/名詞	37.49%	16.80%
P/介詞	1.08%	5.60%
Vt/及物動詞	28.11%	21.40%
T/語助詞	0.48%	2.20%
Vi/不及物動詞	18.69%	7.80%
POST/後置詞	0.48%	3.20%

6. 結論

在本研究中，我們提出一個整合多面向均勻度的計算方法，使詞語均勻度的衡量能夠同時考慮不同的分類面向，更全面地評估詞彙的核心程度。其次，我們提出詞頻正規化的方法來修正傳統均勻度公式遇到切分區塊大小不一致時，造成統計均勻度偏差的問題。最後，我們提出了一個以異源語料庫評估核心詞彙庫的方法，可以準確地比較及分析各種均勻度公式所選取詞表的的優缺點與特性。最後，我們以實驗證實，正規化後的均勻度公式的確可以有效改善分布均勻度的評估，而整合多面向均勻度的計算方法，確實可

以選擇到更具核心特質的詞彙。

在語言教學的應用上，過去許多研究者認為均勻度公式偏好選擇功能詞。所以我們在本文中也探討了以詞頻及分布均勻度作為詞彙選取方法的差異。結果發現，在初階詞彙表的選擇上，無論是頻率法或是均勻度法排序，序位最高的詞彙當中，功能詞所佔的比例都非常高，對學習者來說較不適宜。所以我們建議必須經過收詞比例的調校以增加實詞的比例。

參考文獻 References

Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour*, *3*(2), 61-65.

Chinese Knowledge Information Processing Group [CKIP]. (1995). *A Description to the Sinica Corpus*. Technical Report 95-02, Academia Sinica, Taipei.

Huang, C.-R., Zhang, H. & Yu, S.-W. (2005). On predicting and verifying a basic lexicon: proposals inspired by distributional consistency. In *POLA forever: festschrift in honor of Professor William S-Y. Wang on his 70th birthday*, Taipei: Language and Linguistics, Academia Sinica, 57-69.

Juilland, A. G. & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.

Juilland, A. G., Brodin, D. R. & Davidovitch, C. (1970). *Frequency dictionary of French words*. Paris, French: Mouton.

Klammer , T., Schulz, M. R. & Volpe A. D. (2009). *Analyzing English Grammar* (6th ed). Harlow, England: Longman.

Kwakernaak, H. (1978). Fuzzy random variables - I: Definitions and theorems. *Information Sciences*, 15, 1-29.

Liu, C-.P. (2012). *The effects of theme-narrative instruction with core vocabulary on oral narrative ability in elementary students with severe hearing impairment.* (Master's thesis, National University of Taiwan).

Lyne, A. A. (1985). Dispersion. *In The vocabulary of French business correspondence*(101-124). Paris, French: Slatkine-Champion.

Rosengren, I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)*, *1*, 103-127.

Steering Committee for the Test of Proficiency - Huayu[SC-TOP]. (2016). *8000 Chinese Words*. Steering Committee for the Test Of Proficiency-Huayu, http://www.sc-top.org.tw/english/download.php.

Stuart, S. L. (1991). Topic and vocabulary use patterns of elderly men and women of two age cohorts. *ETD collection for University of Nebraska - Lincoln.* Paper AAI9208116.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society 96*, 152-63.

Vanderheiden, G. C. & Kelso, D. (1987). Comparative analysis of fixed-vocabulary communication acceleration techniques. *AAC Augmentative and Alternative Communication*, *3*, 196-206.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353.

柯華葳、林慶隆、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍、吳鑑城、白明弘、陳茹玲、李詩敏、黃淑齡、劉寶琦、丁彥平、簡盈妮、張玳維、余昱瑩（2016）。*華語文八年計畫「建置應用語料庫及標準體系」105 年工作計畫期中報告*。教育部補助之研究計畫。臺北市：國家教育研究院。[Ko, Hwa-Wei, Ching-Lung Lin, Jason S. Chang, Hao-Jan Howard Chen, Zhao-Ming Gao, Ya-Hsun Tsai, Yu-Wen Chang, Po-Hsi Chen, Li-Ping Chang, Jian-Cheng Wu, Ming-Hong Bai, Ju-Ling Chen, Shih-Min Li, Shu-Ling Huang, Pao-Chi Liu, Yen-Ping Ting, Ying-Ni Chien, Tai-Wei Chang, Yu-Ying Yu. (2016). *『The 8-year project of construction and application of Mandarin Chinese corpus and standard systems 』105-year work plan interim report.* Taipei: National Academy for Educational Research.]

Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 19-34

N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings

Yu-Ming Hsieh* and Wei-Yun Ma*

Abstract

Rescoring approaches for parsing aims to re-rank and change the order of parse trees produced by a general parser for a given sentence. The re-ranking performance depends on whether or not the rescoring function is able to precisely estimate the quality of parse trees by using more complex features from the whole parse tree. However it is a challenge to design an appropriate rescoring function since complex features usually face the severe problem of data sparseness. And it is also difficult to obtain sufficient information requisite in re-estimatation of tree structures because existing annotated Treebanks are generally small-sized. To address the issue, in this paper, we utilize a large amount of auto-parsed trees to learn the syntactic and sememtic information. And we propose a simple but effective score function in order to integrate the scores provided by the baseline parser and dependency association scores based on dependency-based word embeddings, learned from auto-parsed trees. The dependency association scores can relieve the problem of data sparseness, since they can be still calculated by word embeddings even without occurrence of a dependency word pair in a corpus. Moreover, semantic role labels are also considered to distinct semantic relation of word pairs. Experimental results show that our proposed model improves the base Chinese parser significantly.

Keywords: Word Embedding, Parsing, Word Dependency, Rescoring.

1. Introduction

How to solve structural ambiguity is an important task in building a high-performance statistical parser, particularly for Chinese. Since Chinese is an analytic language, words play different grammatical functions without inflections. A great deal of ambiguous structures will be produced by a parser if no structure evaluation is applied. Therefore, the major task of a

* Institute of Information science, Academia Sinica, Taipei, Taiwan
E-mail: {morris, ma}@iis.sinica.edu.tw

parser is to determine the most plausible parse tree from these ambiguity structures. Re-ranking approaches are widely used in parsing natural language sentences for further advancing the performance of statistical parsers (Shen, Sarkar & Toshi, 2003; Hsieh, Yang & Chen, 2007; Johnson & Ural, 2010; Le, Zuidema & Scha, 2013; Zhu, Qiu, Chen & Huang, 2015). It is an intuitive and efficient strategy to determine the most plausible parse tree from a set of candidate parse trees of a sentence through a rescoring approach.

Treebanks are a widely used resource in parsing task, as it provides useful statistical distributions regarding grammar rules, words, part-of-speeches (PoS), and word-to-word association[1]. However it is difficult to obtain sufficient information requisite in re-estimation of tree structures from existing annotated Treebanks since sizes of treebanks are generally small and insufficient, resulting in a common problem of data sparseness, especially for more complex features in a re-scoring scenario, such as word-to-word dependency associations. So learning information and knowledge from analyzing large-scaled unlabeled data is a compulsory strategy, which is proved useful in the previous works (Wu, 2003; Chen, 2008; Yu *et al*., 2008).

In this paper, we utilize a large amount of auto-parsed trees to learn the syntactic and semantic information and present a simple but effective score function in order to integrate the scores provided by the base parser and word-to-word dependency association scores. The dependency association scores are based on dependency-based word embeddings, learned from a large amount of auto-parsed tress. The score function can relieve the problem of data sparseness, since the dependency association scores can still be calculated by word embeddings even without the occurrence of a dependency word pair in a corpus. In addition, Kim, Song, Park & Lee (2015) proves that the dependency labels (i.e., semantic role labels) in re-ranking parsed tree are important information. As a result, semantic role labels are also considered to distinct semantic relation of word pairs.

Word embeddings have become increasingly popular lately, proving to be valuable as a source of features in a broad range of NLP tasks (Turian, Ratinov & Bengio, 2010; Socher *et al*., 2013; Bansal, Gimpel & Livescu, 2014). The *word2vec* package (Mikolov, Chen, Corrado & Dean, 2013) is among the most widely used word embedding models today. Their success is largely due to an efficient and user-friendly implementation that learns high quality word embeddings from very large corpora. The *word2vec* package learns low dimensional continuous vector representations for words by considering window-based contexts, i.e., context words within some fixed distance of each side of the target words. Another different context type is dependency-based word embedding (Bansal *et al*., 2014; Levy & Goldberg,

[1] Word-to-word association is also called word dependency, a dependency implies its close association with other words in either syntactic or semantic perspective.

2014; Melamud, McClosky, Patwardhan & Bansal, 2016), which considers syntactic contexts rather than window contexts in *word2vec*. Dependency-based word embedding is able to capture functional similarity (as in *lion:cat*) rather than topical similarity or relatedness (as in *lion:zoo*) that *word2vec* would probably provide. Further, Melamud *et al*. (2016) prove that the approach should depend on the tasks to choose the right context type, windows size, and dimensionality in word embedding. From the experiments done by Bansal *et al*. (2014) and Melamud *et al*. (2016), results show benefits of such modified-context embeddings in dependency parsing task. Kim *et al*. (2015) proclaim similar arguments that semantic view should be taken into consideration in re-ranking parse trees because a dependency word pair implies both syntactic and semantic relations.

We propose a rescoring approach for parsing based on a combination of original parsing score and semantic plausibility of dependencies to assist the determination of the parse tree among the *n*-best parse trees. The original parsing score is produced from the Chinese parser (Hsieh, Bai, Chang & Chen, 2012), and the semantic plausibility of dependencies is calculated from dependency-based word embedding. There are three main steps in our rescoring approach. The first step is to have the parser produce *n*-best structures. Second, we extract word-to-word associations (word dependency) from a large amount of auto-parsed data and build dependency-based word embedding. The last step is to build a structural rescoring method to find the best tree structure from the *n*-best candidates. We conduct experiments on the standard data sets of the Chinese Treebank. The results indicate that our proposed approach improves the base Chinese parser significantly.

The remainder of this paper is organized as follows. In Section 2, we describe the rescoring approach and introduce a strategy to extract word dependency associations from a large-scale unlabeled corpus. In Section 3, we report the results of experiments conducted to evaluate the proposed rescoring approaches on different scores of dependency. Section 4 provides a discussion on the related work. Section 5 contains our concluding remarks.

2. Rescoring Syntactic Parse Trees with Dependency Embeddings

In this section, we will describe our rescoring approach. First, we need a parser to generate n-best parse trees with their structural scores, and then select the best parse tree through a score function which considers the structure and the dependency embeddings. Figure 1 shows a flow chart of our rescoring approach. Given an input sentence, the ‘parser’ is responsible for word segmentation, part-of-speech tagging, semantic role labeling, and generate *n*-best parse trees. And then the ‘rescoring’ is based on a combination function of the original parsing score and the semantic score of dependencies to determine from the quality of the *n*-best parse trees.

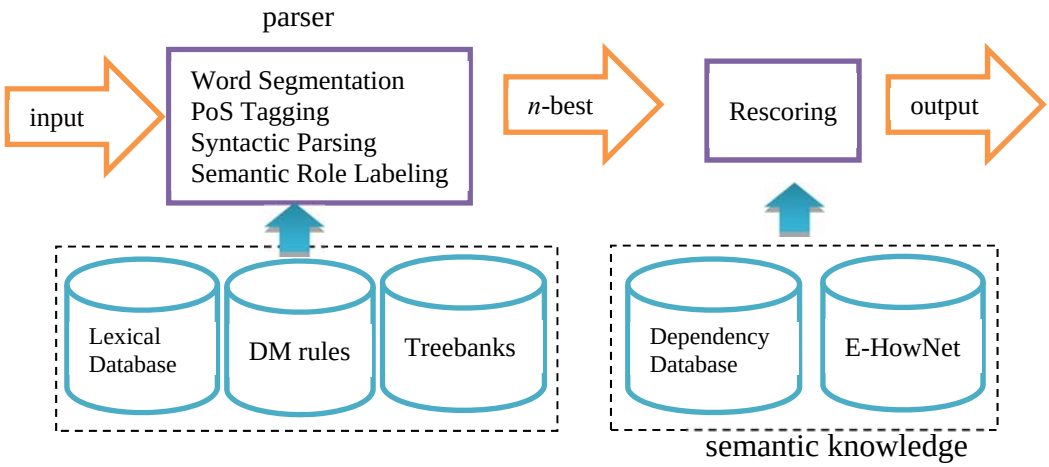


Figure 1. A flow chart of rescoring approach.

We adopt *word2vecf* (Levy & Goldberg, 2014) package to train dependency-based word embeddings from the parsed trees of our corpus. Similar to the approach of Levy & Goldberg (2014), the two steps are needed to achieve this training stage.

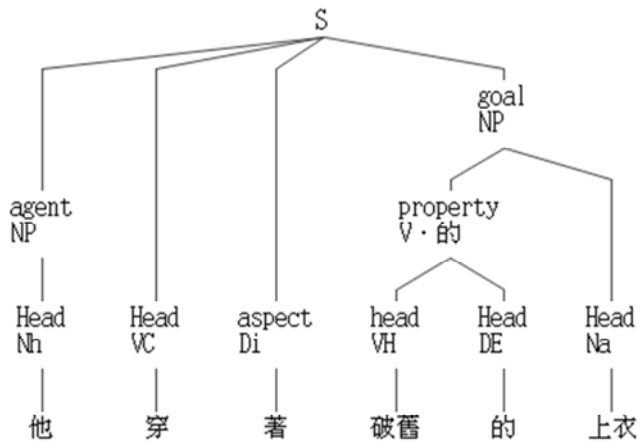
- Step 1: Extract word associations from each tree.
- Step 2: Train a dependency embeddings including target embeddings and context embeddings.

Figure 2 illustrates the first step of word association extraction. Based on the head word information (i.e. the semantic role of the word is 'Head' or 'head', called the 'head word'), we extract dependence word-pairs between head words and their arguments or modifiers. For example, if we have a sentence '*他穿著破舊的上衣 / he is wearing shabby clothes*', five word dependency pairs will be extracted from this tree structure including head words, modifier words, part-of-speeches, semantic role labels, frequencies and etc. The word dependency in (*agent 他/he Nh, Head[S] 穿/wear VC*) represents a head word '*穿/wear*' , its PoS VC (Active Transitive Verb), and its modifier (*他/he, Nh*) with the semantic role label '*agent*'[2].

[2] The 'agent' is a semantic role label. There are 60 semantic roles including thematic roles of events such as 'agent'; 'theme', 'instrument', and secondary roles of 'location', 'time', 'manner' and roles for nominal modifiers. Please refer to CKIP technical report (Chinese Knowledge Information Processing Group [CKIP], 2013) for detail information.

S(agent:NP(Head:Nh:他)|Head:VC:穿|aspect:Di:著|goal:NP(property:V‧的(head:VH:破舊|Head:DE:的)|Head:Na:上衣)):

He is wearing shabby clothes.



agent	他	Nh	**Head[S]**	**穿**	**VC**	1
Head[S]	**穿**	**VC**	aspect	著	Di	1
Head[S]	**穿**	**VC**	goal	上衣	Na	1
property	破舊..的	VH..DE	**Head[NP]**	**上衣**	**Na**	1
property	破舊	VH	**Head[V]**	**的**	**DE**	1

Figure 2. Extraction of word-to-word association (word dependency) from a parse tree.

The second step is to train dependency embeddings. We transform the information of word-to-word associations (in Figure 2) into the *word2vecf* format (in Figure 3). The specific format actually contains two types: the left column is the original word dependency, and right column is the inversion of these words.

word dependency	**Inverse dependency**
(穿, agent_他)	(他, agentI_穿)
(穿, aspect_著)	(著, aspectI_穿)
(穿, goal_上衣)	(上衣, goalI_穿)
(上衣, property_破舊)	(破舊, propertyI_上衣)
(的, property_破舊)	(破舊, propertyI_的)

Figure 3. The transformed word dependency knowledge and the dependency-based embedding format.

The final step is rescoring. We integrate the original parsing score with dependency embedding score of parsed *n*-best trees, and then select the best tree based on the rescoring scores. This step will be illustrated more clearly in the next section.

2.1 Measuring Dependency Plausibility of a Parse Tree

We use trained dependency-based word embeddings, including the target and the context embeddings, to design our score function. Figure 4 indicates an example of word/context embeddings. The symbol *u* is target word embedding, and the symbol *v* is context word embedding, where *n* and *m* represent size of *u* and *v* embeddings respectively. Both embeddings dimension is 300. For example, the embedding of the target word '***穿/wear***' is $u_{t=穿}$ = [0.35, -0.33, -0.01, ⋯, -0.17], and the embedding of the target word '***goal_旗袍***' is $v_{t=旗袍}$ = [0.05, -0.06, 0.01, ⋯, 0.04].

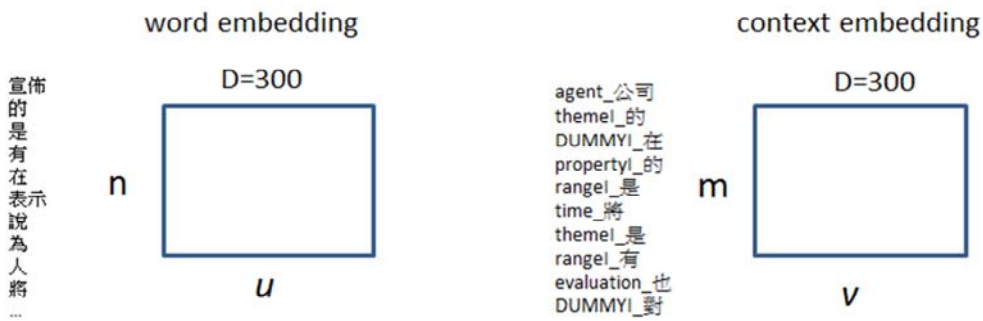


Figure 4. Dependency embeddings: Word and context embedding.

Table 1. The top-10 closest words to the target word '穿/wear' and its context activated.

Words similar to: '穿'		**Contexts activated by: '穿'**	
穿上	0.7544027174	goal_旗袍	0.6248433766
穿戴	0.7354367107	goal_泳裝	0.6048768995
穿著	0.7281331792	goal_母奶	0.5800340081
改穿	0.7054656651	agent_梁朝偉	0.5690256649
身穿	0.6858005123	goal_枕頭	0.5657152053
挑	0.6851677693	agent_模特兒	0.5602549151
換穿	0.6826460874	complement_昏倒	0.5553763580
穿出	0.6747521378	agent_樂器	0.5528911113
試穿	0.6619956193	Head_丟擲	0.5506488225
		goal_拳頭	0.5498251675

The word similarity list and context activated of the word '*穿/wear*' are shown in Table 1[3]. The *DepScore* is our dependency embedding score in Equation (1). The *DepEmb* of a parse *y(s)* tree with dependency embeddings is to represent the dependency associations of a parse tree as a set of target *t* and context *c*. Each dependency in a parse tree can be regarded as a *(t,c)* set. The semantic plausibility of a parse tree is then defined as the sum of the scores of all dependencies in the tree. That is, the semantic score of a parse tree y is defined as

$$DepScore\big(y(s)\big) = \sum_{(t,c)\epsilon y(s)} DepEmb(u_t, v_c) \tag{1}$$

$$DepEmb(u_t, v_c) = log \frac{\exp(u_t \cdot v_c)}{\sum_{x\epsilon Rel(t)} \exp(u_t \cdot v_x)}$$

$DepEmb(u_t,v_c)$ is calculated by taking exponential of dot product $u_t \cdot v_c$, following by a nomalizaion. $x=Rel(t)$ means a word *t* and its dependency word *x* from dependency database (see Equation 1). Finally, we obtain dependency embedding score and frequency for each pair shown in Table 2. With the dependency embedding and frequency information, our design of rescoring will be discussed in the next section.

Table 2. The target word '穿/wear' and its dependency word DepEmb score and frequency .

By dependency embedding score:		**By frequency:**	
TOTAL_穿	1.0000000000	TOTAL_穿	9754.000000
goal_高鞋	0.0008847827	goal_衣服	256.000000
goal_護士服	0.0007814124	aspect_了	197.000000
goal_麵粉袋	0.0007664880	goal_制服	186.000000
goal_官服	0.0007657732	quantity_都	160.000000
goal_衣裙	0.0007545836	quantity_所	146.000000
agent_靴子	0.0007487881	aspect_起	145.000000
reason_擊	0.0007484265	deontics_要	127.000000
goal_西褲	0.0007455655	negation_不	118.000000
goal_健體鞋	0.0007444242	evaluation_也	102.000000
theme_楊孟霖	0.0007376180	agent_他	100.000000

2.2 Rescoring Model for Parse Trees

A re-ranking model ranks a set of candidate dependency parse trees according to its criterion. The criterion of our re-ranking model is a combination of syntactic and semantic score. Given

[3] The interface of the provided information is revealed in the following website. http://irsrv2.cs.biu.ac.il:9998/?word=wear.

a sentence *s*, we define the rescoring model as follows, and the best parse tree $\hat{y}$ of s is obtained from *y*.

$$\hat{y} = \underset{y \epsilon gen(s)}{\operatorname{argmax}} \lambda * CDMScore(y(s)) + (1 - \lambda) * DepScore(y(s)) \quad (2)$$

where *gen(s)* is a set of *n*-best outputs of a baseline parser. In Equation (2), *CDMScore(y)* is the original parsing score generated by CDM Parser (a context-dependent PCFG Parser). The log probability of a parse tree *y* is used. *DepScore(y)* is the final semantic score of a parse tree *y*. CDMScore and DepScore are normalized, i.e. (*i*-min)/(max-min). The symbol λ is an weighting parameter between *CDMScore(y)* and *DepScore(y)*. We substitute λ for every interval of 0.1 from 0 to 1, and design the relating λ from development sets.

3. Experiments

We conducted experiments on our experimental data setting and the evaluation results. And we investigate different types of context in word dependency extraction process and analyze the test results.

3.1 Experimental Settings

Several parts are introduced below to illustrate our experimental design, including corpus, software, evaluation criteria.

Treebank: We employ Sinica Treebank[4] as our experimental corpus. It contains 61,087 syntactic tree structures and 361,834 words. The syntactic theory of Sinica Treebank is based on the Head-Driven Principle; that is, a sentence or phrase is composed of a phrasal head and its arguments or adjuncts. We use the same dataset in Hsieh *et al.* (2012), and divide the treebank into four parts: the training data (55,888 sentences), the development set (1,068 sentences), the test data T06 (867 sentences), and the test data T07 (689 sentences). The test datasets (T06, T07) are the datasets used in CoNLL06 and CoNLL07 dependent parsing evaluation individually. The only difference between Sinica Treebank data and CoNLL data is that the CoNLL is in dependency format. We use labeled information of gold-standard word segmentation and POS tags as our input data in all our experiments.

Large Corpus: The Gigaword corpus contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency (traditional characters), and 380 million characters from Xinhua News Agency (simplified characters). We used the Central News Agency (CNA) portion of Chinese Gigaword Version 2.0 (LDC2009T14). We need to perform word segmentation and part-of-speech tagging before parsing, and the baseline parser is used to

[4] Please refer to the Sinica Treebank webpage for further information: http://treebank.sinica.edu.tw/.

parse the sentences in the Gigaword corpus. In our experiments, word dependencies are extracted from CNA texts. Finally, we obtain 37,711,822 parse trees and extract 224,371,806 word-to-word associations (word dependency).

Chinese Parser: The parser includes some components: Chinese word segmentation, PoS tagging, syntactic parsing, and semantic role labeling. The Chinese word segmentation system and part-of-speech tagging system reaches a high performance of over 95% and 96% in accuracy respectively (Tsai & Chen, 2014). The extracted grammar rules of Sinica Treebank are used in the syntactic parser. We follow You and Chen's method (You & Chen, 2004) to assign semantic role automatically. The system adopts a probabilistic model of head-modifier relations and achieves 92.71% accuracy in labeling the semantic roles.

Estimating Parsing Performance: In evaluation, we use a structural evaluation system called PARSEVAL to compare the parsing results with the gold standard. Throughout the experiment, the bracketed *f*-score (BF) from PARSEVAL is used as the parsing performance metric.

Dependency-based Embedding: Following the illustration about word dependency extraction in section 2.1, we create word embedding from large-scale corpus by *word2vecf* tool with parameter *dimension = 300*, *negative sample=15*, *minimum frequency=5*, and *iterations=5* in our experiments. Finally, we have 568,180 words and 1,721,301 contexts in u and v embedding respectively. $\lambda = 0.7$ in Equation (1) is used in the all experiments below.

3.2 Results of Rescoring Approach

First, we use the *n*-best tree structures produced from F-PCFG parser and observe their bracketed *f*-scores (BF) variation. The oracle *n*-best BF of F-PCFG parser are listed in Table 3. In the data set T07, we find that for the 20-best result, the oracle BF score is 94.66%. In contrast, in the 1-best result, the oracle BF score is 83.91%.

Table 3. Oracle BF score as a function of number N of N-best parses.

	sent. #	1	3	5	10	20
T06	**867**	**88.56**	**92.77**	**94.11**	**95.69**	**96.35**
T07	**689**	**83.91**	**89.83**	**91.50**	**93.57**	**94.66**

Rescoring evaluation: The rescoring evaluation results of the proposed model 'Rescoring-emb' and its competitors are given in Table 4. The 'F-PCFG' parser adopts a linguistically-motivated grammar generalization method to obtain a binarized grammar from original CFG rules extracted from treebank (Hsieh, Yang & Chen, 2015). The 'CDM' Parser proposed by Hsieh *et al*. (2012) achieves the best score in Traditional Chinese Parsing task of SIGHAN Bake-offs 2012 (Tseng, Lee & Yu, 2012). Compared with the other two parsers of F-PCFG and CDM, the approach of 'Rescoring-emb' takes additional semantic score of

dependency parse trees into consideration and achieves high performance on BF scores.

Table 4. Results on T06 and T07 data set.

	T06	T07
F-PCFG	88.56	83.91
CDM	89.91	85.86
Rescoring-emb	90.55	86.41

From Table 4, our rescoring method obtains improvement from 88.56% to 90.35% and the BF score is between the oracle score 1-best and 3-best. The result of the experiment is similar to Charniak & Johnson (2005) proposed re-ranking model. An example of the improved *n*-best parse tree after baseline parsing is presented below. The sequence of these results represents the scores of the original tree in order and the best result is in the 4th tree after rescoring approach.

S(NP(DM:第一天|DM:有５０００多個|Head:Na:人)|Head:VC:參觀)
VP(NP(DM:第一天|DM:有５０００多個|Head:Na:人)|Head:VC:參觀)
VP(DM:第一天|NP(DM:有５０００多個|Head:Na:人)|Head:VC:參觀)
S(DM:第一天|NP(DM:有５０００多個|Head:Na:人)|Head:VC:參觀)
…

In addition, we conducted another experiment without a semantic role label in word dependency pairs. The BF score decreased from 90.55% to 90.05% in T06 data set. The results show that using semantic role labels in word dependency is useful. Furthermore we attempt to compare the effect on using the traditional conditional probability method called 'Rescoring-freq'. Therefore, we modify the Equation (2) into Equation (3). The symbol *m* is the modifier word of the dependency word pair, *h* is the head word, *freq(m,h)* is frequency of the *(m,h)* dependency, and *freq(h)* is frequency of the *h*. If *freq(m,h)* is 0, we will replace it by *1/total word dependency*.

$$\hat{y} = \underset{y \epsilon gen(s)}{\operatorname{argmax}} \ \lambda * CDMScore(y(s)) + (1-\lambda) * WAScore(y(s)) \tag{3}$$

$$WAScore\big(y(s)\big) = \sum_{(m,h)\epsilon y(s)} \log P(m|h)$$

$$P(m|h) = \frac{freq(m,h)}{freq(h)}$$

We also compare the performance of using the traditional conditional probability method with our approach. From the experimental results, the *BF* fell by 0.3% in T07 dataset in Table 5, denoting that the embedding-based scoring has better results than the traditional approach since the embedding score can relax the data sparseness problem, i.e., since dependency scores can be still calculated by word embeddings even without the occurrence of a dependency word pair in a corpus. This verifies the finding of Melamud, Levy & Dagan (2015) in their lexical substitution research based on the word embedding model.

Table 5. Results of 'Rescoring-emb' and 'Rescoring-freq'

	T06	T07
F-PCFG	88.56	83.91
CDM	89.91	85.86
Rescoring-emb	90.55	86.41
Rescoring-freq	90.35	86.19

3.3 Effects of Word Sense Information

In addition to word information, we give a study about the effect on embedding with word sense information. Regarding to word sense information, we use the head senses of words expressed in E-HowNet[5] as words' semantic information. For example, the E-HowNet definition of 車輛(Na), is {LandVehicle|車:quantity={many|多}}, and its head sense is "LandVehicle|車". For detailed description about E-HowNet, readers may refer to Huang, Chung & Chen (2008).

Therefore, to obtain sense definition of lexicons, we convert the word dependency data in Figure 3 corresponding to the E-HowNet dependency (or concept-to-concept relation) as shown in Figure 5. We have two special cases to handle during the process: 1) unknown words, 2) sense ambiguity. The unknown words are skipped in the present experiment. As for the sense ambiguity, we retain the ambiguity of words in E-HowNet, since Zhao & Huang (1999) demonstrated that the retained ambiguity does not have an adverse impact on their identification system.

[5] The E-HowNet information please refer to webpage: http://ehownet.iis.sinica.edu.tw/

Concept-to-concept dependency	Inverse dependency
(PutOn|穿戴, agent_3rdPerson|他人)	(3rdPerson|他人, agentI_PutOn|穿戴)
(PutOn|穿戴, aspect_AspectValue|時貌值)	(AspectValue|時貌值, aspectI_PutOn|穿戴)
(PutOn|穿戴, goal_clothing|衣物)	(clothing|衣物, goalI_PutOn|穿戴)
(clothing|衣物, property_used|舊)	(used|舊, propertyI_clothing|衣物)
(relation|關聯, property_used|舊)	(used|舊, propertyI_relation|關聯)

Figure 5. The concept-to-concept relation and its inverse dependency.

After mapping, we obtain 1,858 target concepts and 93,058 context concepts. We train the concept level embedding and obtain *DepScore* in Equation (1). The experimental result does not seem to improve the overall performance. However some lexicons with the same PoS tag appearing in the similar context may improve through concept relation. For example the common noun "桌子/table (Na)" and "車子/car (Na)" can be as a patient of the verb to move "移動桌子 vs. 移動車子" and in the context denoting location, "在桌子上 vs. 在車子上." But if these words are tagged with sense information, table as {furniture|家具} and car as {LandVehicle|車}, the two nouns are distinguished more easily. Furthermore the unknown word influence may cause deficiency in sense information. In ten-thousand-word Sinica Corpus[6], there are around 8.04% unknown words which do not appear in E-HowNet. And the concept information deficiency is even more serious in CNA Corpus. We suspect that the *DepScore* may not be precise enough because of this factor. In the future, we aim to solve the problem by developing a sense predictor based on lexical analysis and word embeddings to predict the sense of unknown words.

4. Related Work

Our re-ranking estimation approach can be divided into two parts. The first one is rescoring model based on large scale corpus, and the second part is about designing a score function based on word dependency associations.

4.1 Rescoring Model Based on Large Scale Corpus

Most rescoring approaches rely on a post-processing to select the best structure from the n-best parse trees (Shen *et al*., 2003; Hsieh *et al*., 2007; Johnson & Ural, 2010; Hayashi, Kondo & Matsumoto, 2013; Le *et al*., 2013; Zhu *et al*., 2015) or a robust structural evaluation in their parsing models (Wang, Sagae & Mitamura, 2006; Hsieh *et al*., 2012). Treebank is a

[6] Sinica Corpus is the first Balanced Modern Chinese Corpus with part-of-speech tagging. Please refer to the website for detail information (http://asbc.iis.sinica.edu.tw/).

widely used resource, but it is generally small-sized. To overcome the data sparseness problem, some certain strategies of rule generalization and specialization are devised to improve the coverage and precision of the extracted grammar rules (Johnson, 1998; Sun & Jurafsky, 2003; Klein & Manning, 2003; Hsieh *et al*., 2015). However, these studies focused only on syntactic information of parse trees and no semantic information is used in their model. Kim *et al*. (2015) proves that the dependency labels (i.e., semantic role labels) in re-ranking parsed tree are important information. As a result we add semantic role label information in word dependency to distinct semantic relation of word pairs.

4.2 Word Dependency Associations

Common knowledge is needed in a robust parser. How to extract useful information from unannotated large scale corpus and represent the knowledge has been a research issue (Wu, 2003; Chen, 2008; Yu *et al*., 2008; Hsieh, Chang & Chen, 2014). Similarly, we train our word dependency associations from large-scale corpus. The representation of the dependency associations is like knowledge graph embedding (TransR) (Wang, Zhang, Feng & Chen, 2014) or dependency-based word embedding (*word2vecf*). TransR proposed by Lin *et al*. is adopted for representing semantic scores (Lin, Liu, Sun, Liu & Zhu, 2015). TransR models entities and relations in distinct spaces, and then translates the entities in an entity space into the space of a specific relation. Melamud *et al*. (2016) indicate *word2vecf* (Levy & Goldberg, 2014) in pre-trained embedding on unlabeled data in the Stanford Neural Network Dependency (NNDEP) parser (Chen & Manning, 2014) yields improved performance. In our approach, we use *word2vecf* to train dependency-based word embeddings (target, context) and calculate word dependency association scores in our task.

5. Conclusions

In this paper, we present a rescoring approach for parsing based on a combination of original parsing score and word dependency associations to assist the determination of the best parse tree among the *n*-best parse trees. To overcome the data sparseness problem, our word dependency associations are modeled through dependency-based word embeddings, learned from a large amount of auto-parsed trees, and semantic role labels are also considered to distinct semantic relation of word pairs. The experiment results indicate that our proposed approach improves the base Chinese parser significantly.

Reference

Bansal, M., Gimpel, K., & Livescu, K. (2014). Tailoring continuous word representations for dependency parsing. In *Proceedings of ACL 2014*, 809-815.

Charniak, E. & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of ACL 2005*, 173-180.

Chen, L. (2008). Autolabeling of VN Combination Based on Multi-classifier. *Journal of Computer Engineering*, *24*(5), 79-81.

Chen, D. & Manning, C.D. (2014). A fast and accurate dependency parser using neural networks. In *Proceedings of EMNLP 2014*, 740-750.

Chinese Knowledge Information Processing Group. (2013). *Introducing Semantic Roles in Sinica Treebank*. CKIP Technical Report 2013-01, Institute of Information Science, Academia Sinica, Taipei.

Hayashi, K., Kondo, S., & Matsumoto, Y. (2013). Efficient stacked dependency parsing by forest reranking. *Transactions of the ACL*, *1*(1), 139-150.

Hsieh, Y.M., Bai, M.H., Chang, J.S., & Chen, K.J. (2012). Improving PCFG Chinese parsing with Context-Dependent Probability Re-estimation. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 216-221.

Hsieh, Y.M., Chang, J.S., & Chen, K.J. (2014). Ambiguity Resolution for Vt-N Structures in Chinese. In *Proceedings of EMNLP 2014*, 928-937.

Hsieh, Y.M., Yang, D.C., & Chen, K.J. (2007). Improve Parsing Performance by Self-Learning. *Computational Linguistics and Chinese Language Processing (IJCLCLP)*, *12*(2), 195-216.

Hsieh, Y.M., Yang, D.C., & Chen, K.J. (2015). Linguistically-Motivated Grammar Extraction, Generalization and Adaptation. In *Proceedings of IJCNLP 2005*, 177-187.

Huang, S.L., Chung, Y.S. & Chen, K.J. (2008). E-HowNet: the Expansion of HowNet. In *Proceedings of the First National HowNet Workshop*, 10-22.

Johnson, M. (1998). PCFG Models of Linguistics Tree Representations. *Computational Linguistics*, *24*(4), 613-632.

Johnson, M. & Ural., A. E. (2010). Reranking the Berkeley and Brown Parsers. In *Proceedings of HLT-NAACL 2010*, 665-668.

Kim, A., Song, H., Park, S., & Lee, S. (2015). A Re-ranking Model for Dependency Parsing with Knowledge Graph Embeddings. In *Proceedings of IALP 2015*, 177-180.

Klein, D. & Manning, C. (2003). Accurate Unlexicalized Parsing. In *Proceedings of the ACL 2003*, 423-430.

Le, P., Zuidema, W., & Scha, R. (2013). Learning from errors: Using vector-based compositional semantics for parse reranking. In *Proceedings of CVSC*, 11-19.

Levy, O. & Goldberg, Y. (2014). Dependency-Based Word Embeddings. In *Proceedings of ACL 2014*, 302-308.

Lin, Y., Liu, Z., Sun, M., Liu, Y., & Zhu, X. (2015). Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI 2015*, 2181-2187.

Melamud, O., Levy, O., & Dagan, I. (2015). A Simple Word Embedding Model for Lexical Substitution. In *Proceedings of the Workshop on Vector Space Modeling for NLP 2015*.

Melamud, O., McClosky, D., Patwardhan, S., & Bansal, M. (2016). The Role of Context Types and Dimensionality in Learning Word Embeddings. In *Proceedings of NAACL-HLT 2016*, 1030-1040.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.

Shen, L., Sarkar, A., & Toshi, A. (2003). Using LTAG based features in parse reranking. In *Proceedings of EMNLP*, 89-96.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y. & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, 1631-1642.

Sun, H. & Jurafsky, D. (2003). The effect of rhythm on structural disambiguation in Chinese. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, *17*(1), 39-46.

Tsai, Y.F. & Chen, K.J. (2014). Reliable and Cost-Effective Pos-Tagging. *Journal of Computational Linguistics & Chinese Language Processing*, *9*(1), 83-96.

Tseng, Y.H., Lee, L.H., & Yu, L.C. (2012). Traditional Chinese Parsing Evaluation at SIGHAN Bake-offs 2012. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 199-205.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: A simple and general method for semi-supervised learning. In *Proceedings of ACL*, 384-394.

Wang, M., Sagae, K., & Mitamura, T. (2006). A Fast, Accurate Deterministic Parser for Chinese. In *Proceedings of COLING-ACL 2006*, 425-432.

Wang, Z., Zhang, J., Feng, J. & Chen, Z. (2014). Knowledge Graph Embedding by Translating on Hyperplanes. In *Proceeding AAAI-2014*, 1112-1119.

Wu, A. (2003). Learning verb-noun relations to improve parsing. In *Proceedings of the Second SIGHAN workshop on Chinese Language Processing*, 119-124.

You, J.M. & Chen, K.J. (2004). Automatic Semantic Role Assignment for a Tree Structure. In *Proceedings of the 3rd SigHAN Workshop on Chinese Language Processing (2004)*, 109-115.

Yu, K., Kawahara, D., & Kurohashi, S. (2008). Chinese Dependency Parsing with Large Scale Automatically Constructed Case Structures. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*, 1049-1056.

Zhao, J. & Huang, C. (1999). The Complex-feature-based Model for Acquisition of VN construction Structure Templates. *Journal of Software*, *10*(1), 92-99.

Zhu, C.X, Qiu, X.P., Chen, X.C. & Huang, X.J. (2015). A Re-ranking Model for Dependency Parser with Recursive Convolutional Neural Network. In *Proceedings of ACL 2015*, 11-19.

Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 35-54

使用字典學習法於強健性語音辨識

The Use of Dictionary Learning Approach for Robustness Speech Recognition

顏必成*、石敬弘*、劉士弘+、陳柏琳*

Bi-Cheng Yan, Chin-Hong Shih, Shih-Hung Liu and Berlin Chen

摘要

在有雜訊的環境下，自動語音辨識系統(Automatic Speech Recognition, ASR)的效能往往會有明顯衰退的現象。本論文旨在研究語音強健性技術，希望能夠透過語音特徵的調變頻譜(Modulation Spectrum)正規化以萃取出較具有強健性的語音特徵。為此，我們使用 K-奇異值分解(K-SVD)的字典學習法(Dictionary Learning)於分解調變頻譜的強度(Magnitude)成分，在最小化還原訊號誤差且在其權重矩陣稀疏性的限制下，希望能獲取較具強健性的語音特徵。此外，因調變頻譜強度成分皆為正值，所以我們提出非負K-SVD的方法來解決這個議題，希望能增進自動語音辨識系統在抗噪上的效能。本論文的所有實驗皆於國際通用的 Aurora-2 連續數字資料庫進行；實驗結果顯示相較於僅使用梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficient, MFCC)之基礎實驗和其它常見的調變頻譜分解方法，我們所提出的字典學習法與其改進方法皆能顯著地降低語音辨識錯誤率。最後，我們也嘗試將所提出的字典學習方法與一些經典的強健性技術結合，如：進階前端標準法(Advanced Front-End, AFE)、變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)、統計圖等化法(Histogram Equalization, HEQ)，以驗證其實用性。

關鍵字: 強健性、自動語音辨識、調變頻譜、稀疏編碼、字典學習法。

*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
E-mail: {60447055S, 60447003S, berlin}@ntnu.edu.tw
+中央研究院資訊科學研究所
Institute of Information Science, Academia Sinica
E-mail: journey@iis.sinica.edu.tw

Abstract

The performance of automatic speech recognition (ASR) often degrades dramatically in noisy environments. In this paper, we present a novel use of dictionary learning approach to normalizing the magnitude modulation spectra of speech features so as to retain more noise-resistant and important acoustic characteristics. To this end, we employ the K-SVD method to create sparse representations for a common set of basis vectors that span the intrinsic temporal structure inherent in the modulation spectra of clean training speech features. In addition, taking into account the non-negativity property of amplitude modulation spectrum, we utilize the nonnegative K-SVD method, paired with the nonnegative sparse coding method, to capture more noise-robust features. All experiments were conducted on the Aurora-2 corpus and task. The empirical evidence shows that our methods can offer substantial improvements over the baseline NMF method. Finally, we also integrate the proposed variants of the K-SVD method with other well-known robustness methods like Advanced Front-End (AFE), Cepstral Mean and Variance Normalization (CMVN) and Histogram Equalization (HEQ) to further confirm their utility.

Keywords: Robustness, Automatic Speech Recognition, Modulation Spectrum, Sparse Coding, Dictionary Learning.

1. 緒論

語音是人類最常使用的一種訊息表達方式。在日常生活中，語音往往帶有大量且重要的訊息，因此我們對語音訊號進行處理、分析，毫無疑慮的非常具有發展性，而語音辨識藉由將語音訊號轉換成文字為目標，無論是在語意情感分析、語言輔助學習、智慧機器語音識別上有著相當廣泛的應用。

然而大多數的自動語音辨識系統，在不被干擾的情況下，皆能獲得良好的語音辨識效果，但是在現實環境中，自動語音辨識卻往往因為測試環境與訓練環境不匹配(Mismatch) (Tabrikian, Fostck & Messer, 1999)，使得此系統之效能衰退之現象。上述所造成環境不匹配問題的種種因素包含了：語者腔調變異、加成性背景雜訊、摺積性通道雜訊及其他語者發音的干擾等。所謂的語音辨識之強健性技術(Li, Deng, Gong & Haeb-Umbach, 2014)，即是致力於降低上述因素所帶來之影響，進而使語音辨識系統在不匹配問題存在的環境下，仍能保有一定的辨識能力。

近年來，字典學習(Dictionary Learning)方法被廣泛地應用在圖像(Lu, Shi & Jia, 2013)、語音處理之領域(Gemmeke, Viratnen & Hurmalainen, 2011) (He, Sun & Han, 2015)，其核心概念是利用字典來線性地表示訊號並獲得其稀疏表示(Sparse Representation)，字典學習是從字典(Dictionary)中選取少量的原子(Atoms)來表示訊號，其中每一個原子都可以當作

是一個基礎訊號的表達，而所有原子組成的集合稱為字典。在字典學習方法中，我們可以直接挑選經過處理過後的訊號成為範本字典(Gemmeke *et al*., 2011)，或者是由其他自動學習字典的方法來求得字典，一般常見的字典學習方法有最優方向法(Method of Optimal Directions) (Engan, Aase & Husoy, 1999)、K-奇異值分解法(K-SVD) (Aharon, Elad & Bruckstein, 2006)、隨機梯度下降法(Stochastic Gradient Descent) (Bottou, 1998)及線上字典學習法(Online Dictionary Learning) (Mairal, Bach, Ponce & Sapiro, 2010)。另外在字典學習法中，其原子相對應的權重也須要一併更新，關於權重的更新方式可以由範數的限制做區分，0-式範數(0-Norm)常見的方法為匹配追蹤演算法(Matching Pursuit, MP) (Mallat & Zhang, 1993)、正交匹配追蹤演算法(Orthogonal Matching Pursuit, OMP) (Pati, Rezaiifar & Krishnaprased, 1993)，上述兩種方式都是透過計算殘差與原子的關聯程度來求取權重。1-式範數(1-Norm)常見的方法為基礎追求法(Basis Pursuit, BP) (Chen, Donoho & Saunders, 2001)以及最小絕對壓縮選擇法(LASSO) (Tibshirani, 1996)，此兩種方法將目標函數視為最佳畫圖函數，並透過迭代更新求得其解。

本論文旨在探究使用字典學習法以及一些改進方法來分解調變頻譜強度成分，以獲得較具強健性的語音特徵。字典重建是字典學習方法的主要問題，其目標在於如何從原始訊號中學習出具有代表性的原子來組成字典，且在字典學習方法裡通常被隨著使用稀疏編碼來求取原子的權重，而稀疏編碼的目的在於將訊號表示為各個原子的稀疏線性組合，期望能夠求取具調變頻譜局部性的重要資訊。在本論文中，我們分別使用了 K-SVD 演算法搭配匹配追蹤演算法以及正交匹配追蹤演算法，來得到乾淨的語音調變頻譜強度成分。另一方面，因調變頻譜強度成分皆為正值，所以我們提出使用非負 K-SVD 搭配非複數稀疏編碼(Hoyer, 2004)來解決這個議題，以增進自動語音辨識系統在抗噪上的效能。此外，我們嘗試將字典學習方法與一些經典的特徵強健性技術結合，如：進階前端標準法(Advanced Front-End, AFE)、變異數正規化法(Cepstral Mean and Variance Normalization, CMVN)、統計圖等化法(Histogram Equalization, HEQ)，以驗證這些改進方法之實用性。

2. 相關文獻

在語音辨識中，強健性語音特徵技術主要有兩個方法，第一是以模型為基礎的強健性技術(Model-based Technique)，第二是以語音特徵為基礎的強健性技術(Feature-based Technique)，分別介紹如下：

第一，以模型為基礎的強健性技術是使用少量的測試環境之調適語料來對聲學模型進行調整，使聲學模型可以去近似於輸入雜訊語音的機率分布參數，達到降低環境不匹配的情形。常見的技術有最大相似度線性回歸法(Maximum Likelihood Linear Regression, MLLR) (Leggetter & Woodland, 1995)、最大事後機率法則(Maximum a Posteriori, MAP) (Gauvain & Lee, 1994)、平行模型結合法(Parallel Model Combination, PMC) (Gales & Young, 1996)、向量泰勒級數(Vector Taylor Series, VTS) (Kim, Un & Kim, 1998)、遺失特徵理論(Missing Feature Theory, MFT) (Van Segbroeck & Van Hamme, 2011)。

第二，以語音特徵為基礎的強健性技術是在不更改聲學模型的情況下，利用乾淨的

語音特徵去訓練，期望將帶有雜訊的語音特徵還原成乾淨的語音特徵，本文使用語音參數正規化法(Feature Normalization)，此方法目的在正規化語音特徵本身的特徵值及統計分布，再利用測試語音特徵的特徵值來消除雜訊干擾所帶來的影響，此方法常見的技術有倒頻譜平均值減去法(Cepstral Mean Subtraction, CMS) (Viikki & Laurila, 1998)、變異數正規化法(Cepstral Mean and Variance Normalization, CMVN) (Viikki, Bye & Laurila, 1998)、統計圖等化法(Histogram Equalization, HEQ) (de la Torre *et al*., 2005)、倒頻譜平均值與變異數正規化結合自回歸動態平均濾波法(Cepstral Mean and Variance Normalization plus Auto-regressive-moving Average Filtering, MVA) (Chen & Bilmes, 2007)。

語音參數正規化法的特色是有效且容易實現於大部分的自動語音辨識系統，但是只適用於作用於噪音緩慢變化的情形，並且其改進效果往往是有限的。而語音模型調適法，在辨識上可以有較好的效果，由於此方法會根據噪音環境來更新聲學模型，所以此方法在實作上非常的耗時。

此外，與本論文最相關的研究是語音特徵的調變頻譜強健性技術(張庭豪，2015)，其論文實做了多種非負矩陣分解法處理於調變頻譜的強健性技術。並在稀疏非負矩陣分解法(Sparse Nonnegative Matrix Factorization, SNMF) (Hoyer, 2004)針對基底矩陣做稀疏化的限制，其結果顯示對基底矩陣稀疏化對語音強健性的辨識結果是有相當的助益。因此我們利用稀疏編碼的方法，並且延伸到字典學習法。利用帶有稀疏性的方法，更精確地表示語音訊號。我們希望在訓練階段時，讓字典學得不受噪音干擾的乾淨語音特徵，並在測試時讓帶有噪音的語料透過乾淨語音特徵字典在不超過誤差以及範數的限制下求取其對應的稀疏權重，以此還原噪音語料，期望能降低環境不匹配性並獲得較優良的辨識效果。

3. 調變頻譜正規化法

3.1 調變頻譜之簡介

對於一語音特徵時間序列$x[n]$而言，其調變頻譜定義如下：

$$X[k] = DFT(x[n]) = \sum_{t=0}^{N-1} x[t]e^{-j\frac{2\pi tk}{N}} \quad ,0 \le k \le \frac{N}{2} \tag{1}$$

其中，n 與 k 依序為音框索引與調變頻率索引，DFT 為離散傅立葉轉換(Discrete Fourier Transform, DFT)，$x[n]$代表某一維度語音特徵時間序列，$X[k]$代表語音特徵時間序列$x[n]$的調變頻譜。式(1)可看出調變頻譜可以廣泛的分析語句中語音特徵隨時間變化的資訊。而$X[k]$頻譜序列可視為一種對於原始語音訊號作降低取樣(Down-Sampled)後的調變訊號(由訊號取樣率轉至音框取樣率)，此序列即為所屬語音特徵時間序列之調變頻譜(Modulation Spectrum)。由式(1)可知，調變頻譜$X[k]$之最高頻率與特徵序列$x[n]$之取樣頻率(音框取樣率)相關。例如，在一般設定下，音框取樣率為 100 Hz，則最高調變頻率為 50 Hz。

過去已有不少學者研究語音特徵之調變頻譜的特性，發現了調變頻譜中的低頻成分是比高頻成分還要重要的特性(Chen & Bilmes, 2007)。而調變頻譜之低頻成分(約 1Hz 至 16Hz)對於語音辨識精確度也有密切的關係，潛藏了最重要的語意資訊。其中，最重要的是位於 4 Hz，有學者指出，4 Hz 是人耳聽覺最為敏感之調變頻率(Gales & Young, 1996)；也有學者認為，4 Hz 為人類大腦皮層感知之重要調變頻率(Kim *et al.*, 1998)。當語音訊號受到雜訊影響時，其語音特徵時間序列會受到影響而失真，及其調變頻譜也會跟著受到牽連。很多學者提出作用在調變頻譜的正規化法，以改善調變頻譜受到雜訊干擾的影響。因此，我們可將許多發展在語音特徵時間序列的正規化法應用在調變頻譜使其正規化，而正規化的對象是對其調變頻譜強度成分$|X[k]|$來進行處理，並保持其相位角不變$\theta[k] = \angle X[k]$的部分。接著處理更新後的強度成分會與原始相位成分結合，再經由反傅立葉轉換(Inverse Discrete Fourier Transform, IDFT)來求得新的語音特徵時間序列。若調變頻譜的強度能夠被有效的正規化，便能夠有效解決雜訊產生的環境不匹配之問題，使自動語音辨識系統使用新的語音特徵時能夠獲得較佳的辨識率。以下將會簡單回顧一些常見的調變頻譜正規化法。

3.2 調變頻譜平均正規化法(Spectral Mean Normalization, SMN)

假設當各種音素在理想環境中占的比例接近一致時，每一維度特徵的調變頻譜之平均值應該為一個定值(Huang, Tu & Hung, 2009)：

$$\left|\tilde{X}[k]\right| = |X[k]| - \mu_s + \mu_a \tag{2}$$

在式(2)中，$|X[k]|$為原始的調變頻譜強度成分，μ_s為單一語句的調變頻譜強度成分之平均值，μ_a為所有訓練語句的調變頻譜強度成分之平均值，而$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

3.3 調變頻譜平均與變異數正規化法(Spectral Mean and Variance Normalization, SMVN)

除了要正規調變頻譜強度成分之平均值，也要正規其變異數(張庭豪，2015)。假設特徵向量參數之平均值在理想環境中比例接近一致時，平均值應為零，且特徵向量參數之分布可以利用變異數來進行檢測：

$$\left|\tilde{X}[k]\right| = \frac{|X[k]| - \mu_s}{\sigma_s}\sigma_a + \mu_a \tag{3}$$

在式(3)中，μ_s與σ_s為單一語句的調變頻譜強度成分之平均值與變異數；μ_a與σ_a為所有訓練語句的調變頻譜強度成分之平均值與變異數，$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

3.4 調變頻譜統計圖等化法(Spectral Histogram Equalization, SHE)

利用非線性的轉換(Nonlinear Transform)，不只將調變頻譜強度成分之平均值與變異數作正規化，也使訓練語句與測試語句的調變頻譜強度成分趨於擁有同一個機率分布函數，

正規化全部階層的動差(Viikki & Laurila, 1998)：

$$\left|\tilde{X}[k]\right| = F_{ref}^{-1}\left(F_X(|X[k]|)\right) \quad (4)$$

在式(4)中，$F_X(\cdot)$為單一語句的調變頻譜強度的機率分布(Probability Distribution Function, PDF)，F_{ref}則是利用所有訓練語句之調變頻譜強度所求的參考機率分布，$\left|\tilde{X}[k]\right|$便是更新過後的調變頻譜強度成分。

3.5 分頻段調變頻譜統計正規化法

此方法的概念是想要改進調變頻譜統計正規化法；調變頻譜統計正規化法是將全部調變頻帶的頻譜強度值視為是同一隨機變數(Random Variable)的樣本(Samples)，且將之一併進行正規化的動作。但是前面提到在語音辨識中，不同調變頻率的成分有不同的重要性，低頻成分是比高頻成分還要相對重要的，因為語言的重要資訊較集中於低頻成分。因此有學者提出將調變頻帶分成許多子頻段，再分別對每一個子頻段的頻譜強度作上述所提的調變頻譜正規化的方法，而不是單純直接對整個全部調變頻帶做處理(Viikki *et al.*, 1998)。因為要強調低調變頻率的重要性，所以在低頻部分的子頻段擁有較細的頻寬，子頻段的數量也比較多，而高調變頻率便持有相反的特性。由於掌握住低頻成分的資訊，根據學者的實驗數據，顯示出將調變頻率分頻段且正規化的做法，能比全頻帶正規化的方式獲得較好的效能。

4. 使用字典學習法於調變頻譜分解

4.1 字典學習法介紹

字典學習法是一種利用字典來表示資料的方法。其精髓在於透過學習而來的字典(Dictionary)，配合稀疏編碼(Sparse Coding)挑選出字典中重要的原子(Atoms)；並且使得一些多餘的資訊(Redundancy)稀疏，最後以線性組合的方式近似還原出原始訊號(Tosic & Frossard, 2011)。相對於其他降維方法的技術如：主成分分析法(Principal Component Analysis)、線性鑑別分析(Linear Discriminant Analysis)，字典學習法沒有拘泥於減少資料維度的特性；反之，字典學習法追求的是如何習得資料中重要的特徵，以及潛在的資料意義。所以字典通常都是透過遠大於輸入資料的維度，鉅細靡遺表示資料。就計算複雜度而言，相比之下較花時間且並非每種資料都適合升維的字典學習法，我們也可以依照資料的屬性字典設計成降維的字典。一般而言，字典學習法可以依照資料選取的概念分成三種常見的方法，分別是：機率學習法(Tosic & Frossard, 2011) (Wipf & Rao, 2004) (Probabilistic Learning Methods)、向量編碼與分群法(Gersho & Gray, 1991) (Vector Quantization or Clustering Methods)、特殊結構法(Tosic & Frossard, 2011) (Particular Construction Methods)。關於機率字典學習法(Wipf & Rao, 2004)，採用稀疏貝氏學習法(Sparse Bayesian Learning)去解決字典的目標函數以及字典更新問題。而特殊結構字典學習法(Yaghoobi, Daudet & Davies, 2009)，他們探討著針對不同輸入信號時，可以在目標函數加入不同的參數矩陣函數(Parametric Function)使得字典最佳化。而向量編碼與分群

法除了著名的 K-SVD 以外(Aharon *et al*., 2006)，近年來有些研究學者基於 K-SVD 資料分布是高斯分布的假設而提出拉式分布的說法，改善 K-SVD 重建項(Reconstruction Term)2-式範數的限制為 1-式範數，其提出的方法為 l_1-K-SVD(Mukherjee, Basu & Seelamantula, 2016)，其方法能有效的改善 2-式範數過於平滑(Over-smoothing)的問題，其方法除了提升了辨識率，以及加快了收斂速度，並拉近了原始字典與習得字典的尤拉距離(Euclidean Distance)。

4.2 K-SVD字典學習法

K-SVD 字典學習法是由 Aharon 等提出(Aharon *et al*., 2006)，其根本想法源自於向量編碼問題(Vector Quantization)以及 K-means 演算法(Gersho & Gray, 1991)，屬於廣義的 K-means (Generalized K-means)演算法。Aharon 等認為編碼問題可以針對字典原子加入 0 式範數(0-Norm)的限制使得字典在近似還原輸入資料時，能刪減掉不必要的資訊，並重新表示為式(5)：

$$min_{D,X}\{\|Y - DX\|_F^2\} \quad subject\ to\ \forall i\ ,\ \|X_i\|_0 \le T_0 \tag{5}$$

其中 Y 為原始訊號、D 為欲學習的字典、X 為字典相對應的權重矩陣、T_0是一個非零的實數。迭代更新的關係式(5)也可以等價表示為式(6)：

$$min_{D,X} \sum_i \|X_i\|_0 \quad subject\ to\ \|Y - DX\|_F^2 \le \epsilon \tag{6}$$

其中ϵ為一個給定的錯誤容忍值。

在 K-SVD 字典學習法中，我們透過類似 K-means 迭代更新的方式求解算式(6)。在稀疏編碼階段中，我們可以使用任何匹配追蹤的方法找出權重矩陣 X，並且透過 0 式範數的限制，讓權重矩陣中的向量X_i內的元素(Element)個數不會超過T_0，藉此使得矩陣 X 稀疏。接著在字典更新階段，我們是以字典的行(Column)向量D_i為單位，亦記作原子(Atom)逐步迭代更新。透過殘差矩陣(Residual Matrix) E_k，如式(7)，其中x^k為權重矩陣(Weight Matrix)X 的第 k 列(Row)向量、d_k為字典的第 k 行。E_k代表著整個重建信號 DX 少了第 k 組原子(Atom)之權重向量後與輸入信號(Input Signal)Y 的差。

$$E_k = Y - \left(DX - d_k x^k\right) \tag{7}$$

接著我們定義了ω，ω是為了取得E_k中所對應的權重矩陣不為 0 的列向量而存在的數字集合，其定義式如下式(8)。其中x^k表示權重矩陣中第 k 列的向量(Row vector)；而x_i代表權重矩陣中的第 i 行的向量(Column vector)。

$$\omega = \{i \big| 1 \le i \le N,\ x^k(i) \ne 0\} = \{i | 1 \le i \le N, x_i(k) \ne 0\} \tag{8}$$

當E_k加入ω集合後，得到的非零的元素矩陣，我們重新記做E_k^w稱為限制殘差矩陣(Restricted Residual Matrix)。來對E_k^w使用奇異值矩陣分解法取得E_k^w中第一筆重要的資訊，分解後所得到的左奇異值向量矩陣我們記為 U；奇異值矩陣我們記作Δ；右奇異值向量矩陣寫成V^T。那麼新的原子(Atom) d_k等於第一筆左奇異值向量的值u_1；而對應的權重向量(Weight Vector) X^k可以透過第一個奇異值與第一筆右奇異值向量的乘積，$\Delta_{(1,1)} \cdot V_1$來更

演算法 1、K-SVD 字典學習法

初始階段:

$D^0 \in R^{n\times K}$，設 $J = 1$。

1、稀疏編碼階段(Sparse Coding Stage):本論文使用 MP、OMP 的演算法求得x_i：

for $i = 1,2, \ldots\ldots, N$

$$min_x\{\|y_i - Dx\|_2^2\} \quad subject\ to\ \ \|x\|_0 \leq T_0$$

2、字典更新階段(Dictionary Update Stage):

- 定義:字典 D 的行向量:d_k、權重矩陣的列向量:x^k、而有被原子使用到的對應權重行向量記錄成集合$\omega_k = \{i|1 \leq i \leq N, x^k(i) \neq 0\}$。
- 計算

$$E_k = Y - (DX - d_x x^k)，$$

- 殘差矩陣E_k透過ω_k取得E_k中不為零的行向量，得到新組成矩陣E_k^ω

  在利用 SVD 分解法$E_k^\omega = U\Delta V^T$。

  更新: $d_k = u_1$

  更新: $x^k = \Delta(1,1) \cdot v_1$

$J = J + 1$

反覆 1、2 直到收斂

新，更新式(9)如下：

$$\text{update: } d_k = u_1$$
$$X^k = \Delta_{(1,1)} \cdot V_1 \tag{9}$$

完整的 K-SVD 作法如演算法 1 所示。

4.3 非負K-SVD字典學習法

非負 K-SVD 字典學習法是和 K-SVD 字典學習法同時誕生的，源自於同一組研究團隊。為了使 K-SVD 更能夠完整的描述輸入資料為正數時的情況；進而在 K-SVD 求解權重以及更新字典時加入非負數的限制，而新生成的演算法稱為 NN-K-SVD。NN-K-SVD 為了使輸出的字典(Dictionary)以及權重矩陣(Weight)皆為正數，所以在學習的稀疏編碼階段(Sparse Coding Stage)時，求解權重的方法利用非負數稀疏編碼法(Non-negative Sparse Coding) (Hoyer, 2004)。

演算法 2、非負 K-SVD 字典更新規則

初始階段:

令 $d_i = \begin{cases} 0 & u_i<0 \\ u_i & \text{otherwise} \end{cases}$，$\mathrm{x(i)} = \begin{cases} 0 & v_1(i)<0 \\ v_1(i) & otherwise \end{cases}$，

其中u_1、v_1是$E_r^{\omega_k}$接過奇異值分解後，第一筆奇異值左向量與右奇異值向量。

重複 J 次:

(1) 令:$\mathrm{d} = \frac{E_r^{\omega_k} x}{x'x}$，　當 $\mathrm{d(i)} = \begin{cases} 0 & d(i)<0 \\ d(i) & otherwise \end{cases}$

(2) 令:$\mathrm{x} = \frac{d' E_r^{\omega_k}}{d'd}$，　當 $\mathrm{x(i)} = \begin{cases} 0 & x(i)<0 \\ x(i) & otherwise \end{cases}$

在非負 K-SVD 演算法中，學習的方法與 K-SVD 如出一轍。不過為了滿足非負數的限制，我們對目標函數式(5)了一些調整。我們在最小平方法裡加入非負數的限制。也就是說對於字典，我們只保留前 L 大的原子(Atom)來參與字典學習，如式子(10)。

$$min_x \|y - D^L x\| \quad \text{s.t.} \quad x \geq 0 \tag{10}$$

在稀疏階段(Sparse Coding Stage)，我們使用上述提及的非負稀疏編碼法(NNSC)中求取矩陣 S 的方式，求解權重 x。而在字典更新階段(Dictionary update stage)，為了使得原子d_k (Atom)更新後也為正數，則加入原子之值為正的限制，如式子(11)所示：

$$min_{d_k, x^k} \left\| E_r^{\omega_k} - d_k x^k \right\| \quad s.t. \quad d_k, x^k \geq 0 \tag{11}$$

而更新殘差矩陣$E_r^{\omega_k}$利用 SVD 分解時可能產生負數，我們採用上圖的演算法，如果第一次 SVD 分解後的左右奇異值向量皆負數；則同乘(-1)。接下來奇異值向量內的元素(Element)小於零則設為零，其他大於零的數字則保留不變。開始重複 J 次使得$E_r^{\omega_k}$越來越接近$\mathrm{d}x'$。完整非負 K-SVD 更新規則如演算法 2 所示。

4.4 權重的更新方法

4.4.1 匹配追蹤

匹配追蹤(Matching Pursuit, MP)，是字典學習法第一階段稀疏編碼(Sparse Coding)中，求得權重矩陣(Weight) X 的常見解法。演算法概念為貪婪法則。首先將輸入信號(Input signal) x 令成冗餘項量(Residual) r。再利用投影量的方式衡量出與冗餘向量 r 相關程度最大的原子(Atom) d_i後，則冗餘向量 r 與最相關的原子 d_i的內積值式(12)即為對應權重 α 在第 i 維度的分數。

$$\left(d_i^T \cdot r\right) \tag{12}$$

演算法 3: 匹配追蹤演算法(MP)

1. 輸入資料:輸入信號 x、字典 D。
2. 輸出資料:權重向量 α。
3. 目標函數:

$$\min_{\alpha \in R^n} \|x - D\alpha\|_2^2 \qquad s.t \quad \|\alpha\|_0 \leq L$$

4. 初始步驟:

$$\alpha \leftarrow 0 \text{、} r(\text{冗餘向量}) \leftarrow x$$

5. while $\|\alpha\|_0 < L$ do
   - 找尋與冗餘向量(Residual Vector)關聯程度最大的原子。

$$\hat{\imath} \leftarrow argmax_{\hat{\imath}=1,\ldots\ldots,p} |(d_{\hat{\imath}}^T \cdot r)|$$

   - 更新冗餘向量以及對應的權重(Weight)

$$\alpha[\hat{\imath}] \leftarrow \alpha[\hat{\imath}] + (d_{\hat{\imath}}^T \cdot r)$$

$$r \leftarrow r - (d_{\hat{\imath}}^T \cdot r) d_{\hat{\imath}}$$

6. end while

之後再將輸入信號 r 減去信號 r 在最相關的原子d_i上的投影量分量即為式(13)，所得到的冗餘向量 r 則為下一次迭代的輸入信號。

$$r = r - \left(d_i^T \cdot r\right) d_i \tag{13}$$

完整匹配追蹤的算法如演算法 3 如示。

4.4.2 正交匹配追蹤

正交匹配追蹤法(Orthogonal Matching Pursuit, OMP)是匹配追蹤法的改良方法。正交匹配追蹤法(OMP)在衡量冗餘向量與原子的關聯程度時考慮的是正交投影量，而不是投影量。而計算正交投影量的方法為 Gram-Schmidt 正交化法，其中更新冗餘向量 τ 及權重 α 如式(14)及(15)所示：

$$\tau \leftarrow (\mathrm{I} - D_\tau (D_\tau{}^T D_\tau)^{-1} D_\tau{}^T)\mathrm{x} \tag{14}$$

$$\alpha_\tau \leftarrow (D_\tau{}^T D_\tau)^{-1} D_\tau{}^T \mathrm{x} \tag{15}$$

4.4.3 非負數稀疏編碼法

非負數稀疏編碼法(Non-Negative Sparse Coding, NNSC)的更新演算法是參照非負矩陣分解的乘法式更新式(Multiplicative Update Rules) (張庭豪，2015)，其特性為藉著輸入矩陣皆為正數，然而透過乘法更新的關係使得更新後的矩陣也為正數。非負數稀疏編碼是為了求解式(10)而產生的方法。

演算法 4: 非負數稀疏編碼(NNSC)

目標函數:

$$min_{A,S}\frac{1}{2}\|X-AS\|^2+\lambda\sum_{ij}S_{ij}$$

限制條件如下: $\forall_{ij}: X_{ij}$ 、$A_{ij}\geq 0$ 、$S_{ij}\geq 0$、$\lambda\geq 0$ and

$\forall_i: \|a_i\|=1$，a_i *是* A*行向量*$(column)$.

(1) 初始階段:透過隨機變數並設變數為正數初始A^0與S^0，其 中 A 矩陣的行向量須為單位向量，並設時間參數 t=0。

(2) 反覆迭代直到收斂:

- 更新 A 矩陣:

  $A'=A^t-\mu(A^tS^t-X)(S^t)^T$。

  任何負數出現在A'中都設為 0，重新使得A'行向量成完單位向量，並設定$A^{t+1}=A'$。

- 更新 S 矩陣，運用乘法更新式(Multiplicative Update Rule):。

  $$S^{t+1}=S^t.*(A^TX)./(A^TAS^t+\lambda)$$

t=t+1

$$min_{A,S}\frac{1}{2}\|X-AS\|^2+\lambda\sum_{ij}S_{ij}$$

限制條件為: $\forall_{ij}: X_{ij}$ 、$A_{ij}\geq 0$ 、$S_{ij}\geq 0$ and $\forall_i: \|a_i\|=1$ (16)

其中 X 為輸入矩陣。X、S、A 為非負矩陣且 A 矩陣的行向量須為單位向量、λ 為平衡係數且$\lambda\geq 0$。

求解時我們先隨機初始非負矩陣 A 與矩陣 S 來滿足式(10)；其中 A 的行向量須為單位向量，並設時間參數 t=0。透過類似非負矩陣分解的乘法更新式反覆迭代更新 A 與 S 使得式(16)得到最佳解。NNSC 的求解如演算法 4 所示。

5. 實驗結果與分析

5.1 實驗語料庫

Aurora-2 是歐洲電信標準協會(ETSI) 所發行的語料庫(Hirsch & Pearce, 2000)，以美國成年人的聲音作為錄音來源，內容是連續的英文數字由 0(Zero)到 9(Nine)跟 Oh 等發音字詞。語料庫內有乾淨及附有雜訊的語音，雜訊中有八種不同的加成性雜訊與兩種不同的通道效應，而通道效應是使用國際電信聯合會(ITU)標準中的 G.712 和 MIRS。根據不同的雜訊干擾，分成三個測試集：Set A、Set B 及 Set C。Set A 的語音分別含有地下鐵(Subway)、

人聲(Babble)、汽車(Car)和展覽會館(Exhibition)等四種加成性雜訊與 G.712 通道效應；Set B 的語音則分別含有餐廳(Restaurant)、街道(Street)、機場(Airport)和火車站(Train Station)等四種加成性雜訊與 G.712 的通道效應；Set C 分別加入了地下鐵(Subway) 與街道(Street)兩種雜訊與 MIRS 通道效應。而其中的訊噪比(SNR)則有七種，為 clean、20dB、15dB、10dB、5dB、0dB 和-5dB，並且提供二種訓練模式：乾淨情境訓練模式(Clean-condition Training)與複合情境訓練模式(Multi-condition Training)。本研究的基礎實驗皆使用乾淨情境訓練模式，故測試集中所有加成性噪音是與訓練語料不同的語句，而只有測試集 C 的通道效應與訓練語料不同。

5.2 實驗設定

在本文中的基礎實驗是採用梅爾倒頻譜係數(MFCC)做為語音特徵參數，取樣頻率(Sampling Rate)為 8000Hz，預強調(Pre-Emphasis)參數設為 0.97，使用的窗函數為漢明窗(Hamming Window)，音框長度(Frame Length)是 25 毫秒，音框間距(Frame Shift)為 10 毫秒。每一個音框的特徵使用 13 維梅爾倒頻譜係數(第 1 維至第 12 維還有第 0 維)，加上其一階差量計算和二階差量計算，共 39 維之特徵參數。在特徵的強健性處理方法，本文在處理特徵時，只針對 13 維的靜態特徵參數(Static Feature)進行處理，處理完成後才額外將一階差量和二階差量加入。

5.3 辨識效能評估方式

辨識效能的評估方式是採用美國標準與科技組織(NIST)所訂立的評估標準，進行正確轉譯文句字串與辨識字串的比較。評估方式是以詞正確率(Word Accuracy Rate)為主，計算正確轉寫文句詞串與辨識詞串彼此間，詞的取代個數(Substitutions)、詞插入個數(Insertions)和詞刪除個數(Deletions)：

$$\text{詞正確率}(\%) = \frac{\text{詞正確辨識個數} - \text{詞插入個數}}{\text{輸入詞總數}} \times 100\% \tag{17}$$

本文參照國際學者之設定，在對每一種噪音的訊噪比的結果作加總的動作時，去掉極端的訊噪比 clean 跟-5 dB，只計算範圍 20dB 到 0dB 中的平均詞精確率或平均詞錯誤率的結果再取其平均值。本論文的全部實驗皆是利用平均詞精確率來評估計算辨識的結果。

5.4 基礎實驗結果

首先，最基本的實驗是在以梅爾倒頻譜係數(MFCC)於乾淨語料訓練下所得到的辨識結果。本論文也比較常見的時間序列域特徵正規化法，包含有倒頻譜平均值與變異數正規化法(CMVN)、統計圖等化法(HEQ)、ETSI 所提供的進階前端標準(Advanced Front-End Standard, AFE)，以及作用在調變頻譜上的調變頻譜平均值與變異數正規化法(SMVN)。另外，在調變頻譜上的矩陣分解方法中我們以常用的非負矩陣分解法(NMF)來當作是一

表 1. 基礎實驗數據結果

[Table 1. Recognition accuracy rates (%) averaged over different noise types and different SNRs for several representative acoustic feature normalization methods.]

更新特徵	Set A	Set B	Set C	Avg.
MFCC	54.87	48.87	63.95	54.29
SMVN	59.02	63.60	58.49	60.75
CMVN	75.93	76.76	76.82	76.44
HEQ	80.03	82.05	80.10	80.85
AFE	87.68	87.10	86.29	87.17
MFCC+NMF	67.09	70.98	68.22	68.87
CMVN+NMF	83.56	85.51	83.27	84.28
HEQ+NMF	83.84	85.88	83.70	84.63
AFE+NMF	87.74	87.65	86.32	87.42

個強基礎實驗(Strong Baseline)，此非負矩陣分方法用可在不同的時間序列域中，如 NMF 結合基本的 MFCC(記做 MFCC+NMF)、結合 CMVN(記做 CMVN+NMF)、結合 HEQ(記做 HEQ+NMF)以及結合 AFE(記做 AFE+NMF)，其實驗結果如表 1 所示。表 1 顯示出幾個實驗現象。第一，正規化法消去了語音特徵的平均值及其變異數，其原理就是消去了穩定的通道效應且減少語音特徵分布的差異，因此能有效地抗噪，由表 1 可看出常見的時間序列域特徵正規化法，如 CMVN，都能有效地大幅提升辨識正確率(比 MFCC 進步約 20%)，而在調變頻譜上的正規化法 SMVN 雖然相較於 MFCC 也能提升辨識率，但提昇的幅度卻不太大(比 MFCC 進步約 6%)，這有待更進一步的深入探究。第二，統計圖等化法利用正規化特徵參數的整體分布，對特徵參數的統計分布之所有動差進行正規化，比傳統正規化法多考量了更多的統計資訊,因此在抗噪效果上又比傳統正規化法來得好，如表 1 所示，HEQ 比 CMVN 多了約 4%的進步。第三，由知名 ETSI 單位所提供的 AFE 特徵跟其他正規化法相比，會有最好的辨識結果，相較於 MFCC 有 33%的進步，且相較於 HEQ 也有約 7%的進步。最後，不同的時間序列域之調變頻譜的非負矩陣分解方法都能有效抗噪進而提升辨識正確率，在 MFCC 的調變頻譜上做非負矩陣分解可以得到最多的進步(約 14%)，次之的是在 CMVN 的調變頻譜上有約 8%的進步，接著是 HEQ 的調變頻譜上約有 4%的進步，最少的進步是在 AFE 的調變頻譜上(約 0.3%的進步)。在本基礎實驗中，NMF 的基底個數設定是 5(最好的結果)(張庭豪，2015)。

5.5 基於K-SVD字典學習法於調變頻譜分解

在本小節中，我們將探討所提出使用的 K-SVD 字典學習法於調變頻譜分解之實驗。在調變頻譜上，考量到輸入資料的維度是 513 維，然而運算相當的耗費時間，再衡量到須與前人比較應用在調變頻譜的抗噪技術，所以我們的字典仍屬於降低資料維度的字典，關於此小節實驗設定部分，本論文所使用的字典維度分別設定為 5、10 及 30。

我們的實驗目的是希望透過學習而得到的乾淨語音字典近似噪音語音信號，藉此還原噪音訊號。字典學習法的實驗分成兩個階段，分別是訓練階段，以及測試階段。在訓練階段時先運用四個資料集的乾淨語句調變頻譜特徵，當作輸入資料後並運用 K-SVD 字典學習法搭配權重的解法求得乾淨的字典以及乾淨的權重矩陣。而在此階段時，我們捨棄乾淨的權重矩陣，只保留乾淨的語音字典特徵。接下來在測試階段時，輸入的資料為具有噪音調變頻譜，並且以一句話為單位地利用在訓練階段時所創造的乾淨字典配合權重的解法求得該句對應的權重。最後再將測試語句所對應的權重矩陣乘上訓練時的乾淨字典來還原調變頻譜的特徵。

表 2. 使用 K-SVD 搭配 MP 求解權重於 MFCC、CMVN、HEQ、AFE 等特徵
[Table 2. Recognition accuracy rates (%) for MFCC, CMVN, HEQ, and AFE features for use in K-SVD integrated with MP method.]

更新特徵	Set A	Set B	Set C	Avg.
MFCC+ Dict(5)	63.57	69.00	63.61	65.39
MFCC+ Dict (10)	62.82	68.61	61.50	64.31
MFCC+ Dict (30)	63.70	70.14	59.95	64.59
CMVN+ Dict (5)	82.40	84.23	83.10	83.24
CMVN+ Dict (10)	81.97	83.94	82.72	82.87
CMVN+ Dict (30)	80.83	82.31	81.38	81.51
HEQ+ Dict (5)	81.73	84.50	82.58	82.94
HEQ+ Dict (10)	79.96	82.82	80.85	81.21
HEQ+ Dict (30)	78.85	81.89	79.41	80.05
AFE+ Dict (5)	85.98	87.02	84.87	85.96
AFE+ Dict (10)	85.66	86.66	84.35	85.56
AFE+ Dict (30)	86.09	86.94	84.39	85.81

本論文所使用的 K-SVD 搭配兩種不同求解權重的方法，一為 MP，另一為 OMP，並且作用在不同的語音特徵上，如 MFCC、CMVN、HEQ 及 AFE，實驗結果分別列在表 2 及表 3。對於實驗的結果，我們有四點發現。第一，由表 2、表 3 的實驗結果我們可以得知辨識結果並非隨著字典的原子數增加而變好。會導致此現象的原因可能因為字典是

學習乾淨語音訊號，當字典維度升高時，代表著字典把乾淨特徵學得越詳細。而當噪音影響加劇時，訊號中乾淨的語音特徵相對地不明顯，此時字典維度越高反而越不能辨識出噪音訊號是屬於哪一種乾淨語音特徵，所以在噪音影響下反而無法提升辨識效果。第二，觀察表 3 中 AFE 特徵下三個維度(5、10 及 30)的辨識結果。可以發現字典維度 30 比起維度 5 時的辨識率提升了 0.3%。關於此現象的原因，我們覺得是由於 AFE 特徵已經對噪音影響乾淨訊號的不匹配性有了非常有效的處理。故將字典為度提高時，能更仔細的描述噪音語音特徵。第三，我們可以從表 2、表 3 比較ＯＭＰ以及ＭＰ求取權重對於語音辨識的效果。可以發現不同特徵且字典維度 5 的情況下使用 OMP 求取權重的辨識效果除了 HEQ 特徵降低 0.42%的辨識正確率外，其他特徵比起 MP 的辨識效果都比較好。由於 OMP 求取權重的方法是衡量殘差和原子們的垂直分量，此方法可以除了確保每次迭代求取的權重都是最佳解以外，還可以加快權重收斂的速度。

表 3. 使用 K-SVD 搭配 OMP 求解權重於 MFCC、CMVN、HEQ、AFE 等特徵
[Table 3. Recognition accuracy rates (%) for MFCC, CMVN, HEQ, and AFE features for use in K-SVD integrated with OMP method]

更新特徵	Set A	Set B	Set C	Avg.
MFCC+ Dict(5)	65.82	71.00	65.33	67.38
MFCC+ Dict (10)	62.11	68.77	61.32	64.07
MFCC+ Dict (30)	62.02	67.10	59.78	62.97
CMVN+ Dict (5)	83.65	85.74	84.06	84.48
CMVN+ Dict (10)	79.54	83.95	82.67	82.05
CMVN+ Dict (30)	83.28	85.34	83.59	84.07
HEQ+ Dict (5)	81.22	84.18	82.18	82.52
HEQ+ Dict (10)	79.11	82.29	79.99	80.47
HEQ+ Dict (30)	75.73	78.98	77.18	77.30
AFE+ Dict (5)	86.09	86.91	85.02	86.01
AFE+ Dict (10)	85.88	86.71	84.62	85.74
AFE+ Dict (30)	86.56	87.47	84.88	86.31

5.6 基於非負K-SVD字典學習法於調變頻譜分解

由於本實驗是針對調變頻譜強度部分做特徵增強，故經處理後的強度若出現負值將不符合強度的物理意義。然而 K-SVD 字典學習法在訓練階段時，所用到的 SVD 分解法，將會把輸入資料分解出負數元素(Element)連同後續的更新步驟中一併地加入實驗所創造的乾淨字典。且在實驗的最後階段，透過具有負數的字典乘上權重矩陣所還原出來的調變頻譜也可能會有負數。有鑑於前文所提及的狀況，我們透過非負 K-SVD 字典學習法有效

地改善 K-SVD 在更新時產生負數的問題。

非負 K-SVD 與 K-SVD 的差異除了求取權重時採用的方法為使用乘法更新式的非負稀疏編碼外，非負 K-SVD 在求解目標函數也加入非負的限制，並在更新字典以及權重時使用相當嚴格的演算法確保字典以及權重在每次迭代後的結果皆為正數。

表 4. 使用非負 K-SVD 與非負稀疏編碼求解權重於 MFCC、CMVN、HEQ、AFE 等特徵
[Table 4. Recognition accuracy rates (%) for MFCC, CMVN, HEQ, and AFE features for use in NN-KSVD integrated with NNSC method]

更新特徵	Set A	Set B	Set C	Avg.
MFCC + Dict(5)	65.59	71.22	64.56	67.12
CMVN + Dict(5)	83.80	85.83	84.24	84.62
HEQ + Dict(5)	82.24	85.16	83.40	83.60
AFE + Dict(5)	87.50	88.27	86.84	87.54

由於表 2 和表 3 的數據呈現出在維度為 5 時有最好的辨識結果，所以本小節的實驗非負 K-SVD 字典學習法維度設定就只有設定為 5。關於非負 K-SVD 字典的實驗結果，我們可以歸納出三個要點。首先，與表 1 的四種常見時間域序列特徵相比(如：MFCC、CMVN、HEQ 及 AFE)，我們引入的非負 K-SVD 於調變頻譜領域的辨識結果都比基礎實驗的四種常見特徵有效。MFCC 的辨識率提升了 12.83%、CMVN 的辨識率提升了 8.18%、HEQ 的辨識率提升了 2.75%、AFE 的辨識率提升了 0.37%。第二，比較表 4 與表 3 並針對字典維度 5 的情況下，我們比較 K-SVD+OMP 與 NNK-SVD+NSC 的辨識效果。發現除了 MFCC 降低了 0.26%以外，CMVN、HEQ 以及 AFE 三種不同特徵的辨識結果，皆為非負 K-SVD 較為優異，其中 CMVN 提升了 0.14%、HEQ 提升了 1.08%、AFE 提升了 1.53%。由此可知，當考量調變頻譜的特性並加入非負的限制後，對於辨識效果的提升是有幫助的。第三，我們與強基礎實驗相比，透過觀察表 4 以及表 1 的 MFCC+NMF、CMVN+NMF、HEQ+NMF、AFE+CNMF，可以得知 NMF 是一個很有用的抗噪技術，但本論文所提出使用的非負 K-SVD 在輸入特徵為 CMVN 以及 AFE 時，相較於 NMF 方法在 CMVN 與 AFE 分別有 0.34%以及 0.12%的辨識進步率。最後，由第一點分析延伸。我們覺得將字典學習法以及稀疏編碼加入非負的限制後，並實作在調變頻譜域上是相當具有前瞻性的，然而相關實驗仍持續進行中，期望在此非負的限制條件下其辨識結果仍是值得期待的。

6. 結論與未來展望

本論文探討了字典學習法應用在語音特徵的處理，並將之運用在調變頻譜上，希望能夠擷取出更強健性的基底向量，而達到降低訓練語料和測試語料的不匹配性進而達到語音強健性的目的。本論文使用了兩種字典學習的方法，第一種是採用 K-SVD 字典學習法，但我們可以從本論文的實驗觀察到，隨著字典的增大，辨識效果並沒有如期的提升。由

此可知實作在調變頻譜的實驗仍是適合降維的字典。第二種是使用非負的 K-SVD 字典學習方法，該方法在非負的限制下，我們更可以詮釋調變頻譜強度能量皆為正值的概念。其效果相較於使用 K-SVD 字典學習法也有所提昇，抗噪效果在部分特徵的實作上也勝過非負矩陣分解的方法。

在未來展望的方面，我們希望試著使用超完備的字典來表示原始語音訊號，但由於使用 K-SVD 演算法的計算複雜度過於龐大，所以我們期望使用線上字典學習方法來學習字典，希望能夠加快學習字典的收斂速度；再者，我們也希望能嘗試在訓練階段時使用範本字典選取的方法，此方法藉由事先選擇較為重要的語音特徵，來代替字典學習中字典的學習階段。另外，我們也希望能繼續探討在使用非負 K-SVD 演算法時，造成語音辨識效果不升反降的情形。最後在未來，我們希望能透過訊號分離(Source Separation) (Gemmeke *et al*., 2011)的方式解決環境不匹配的問題，如此一來辨識結果應該能得到提升。

參考文獻 Reference

Aharon, M., Elad, M. & Bruckstein, A. M. (2006). The KSVD: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Transactions on Signal Processing*, *54*, 4311-4322.

Bottou, L. (1998). Online algorithms and stochastic approximations. In D. Saad (Eds.), *Online Learning and Neural Networks*. Cambridge, UK: Cambridge University Press.

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

Chen, S. S., Donoho, D. L. & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, *43*(1), 129-159.

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

Engan, K., Aase, S. O. & Husoy, J. H. (1999). Method of optimal directions for frame design. In *Proc. of IEEE International Conference of Acoustic, Speech, and Signal Processing*, *5*, 2443-2446.

Gales, M. J. F. & Young, S. J. (1996). Robust continuous speech recognition using parallel model combination. *IEEE Transactions on Speech and Audio Processing*, *4*(5), 352-359.

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

Gemmeke, J. F., Viratnen, T. & Hurmalainen, A. (2011). Exemplar-based sparse representations for noise robust automatic speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(7), 2067-2080.

Gersho, A. & Gray, R. M. (1991). *Vector qusntization and signal compression*. Norwell, MA: Kluwer Academic.

He, Y., Sun, G. & Han, J. (2015). Spectrum enhancement with sparse coding for robust speech recognition. *Journal of Digital Signal Processing*, *43*, 59-70.

Hirsch, H. G. & Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proc. of ISCA ITRW ASR 2000*, 181-188.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, *5*, 1457-1469.

Huang, S. Y., Tu, W. H. & Hung, J. W. (2009). A study of sub-band modulation spectrum compensation for robust speech recognition. In *Proceeding of ROCLING XXI:Conference on Computational Linguistics and Speech Processing*, 39-52.

Kim, D. Y., Un, C. K. & Kim, N. S. (1998). Speech recognition in noisy environments using first-order vector Taylor series. *Speech Communication*, *24*(1), 39-49.

Leggetter , C.J. & Woodland, P.C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. *Computer Speech Language*, *9*(2), 171-185.

Li, J., Deng, L., Gong, Y. & Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(4), 745-777.

Lu, C., Shi, J. & Jia, J. (2013). Online robust dictionary learning. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2013)*, 415-422.

Mairal, J., Bach, F., Ponce, J. & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*, 19-60.

Mallat, S. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, *41*(12), 3397-3415.

Mukherjee, S., Basu, R. & Seelamantula, C. S. (2016). ℓ 1-K-SVD: A robust dictionary learning algorithm with simultaneous update. *Signal Processing*, *123*, 42-52.

Pati, Y. C., Rezaiifar, R. & Krishnaprasad, P. S. (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*.

Tabrikian, J., Fostck, G. S. &Messer, H. (1999). Detection of environmental mismatch in a shallow water waveguide. *IEEE Transactions on Signal Processing*, *47*(8), 2181-2190.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267-288.

Tosic, I. & Frossard, P. (2011). Dictionary learning. *IEEE Signal Processing Magazine*, *28*(2), 27-38.

Van Segbroeck, M. & Van Hamme, H. (2011). Advances in missing feature techniques for robust large-vocabulary continuous speech recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(1), 123-137.

Viikki, O., Bye, D. & Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. In *Proc. of ICASSP*, 733-736.

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

Wipf, D. P. & Rao, B. D. (2004). Sparse Bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, *52*(8), 2153-2164.

Yaghoobi, M., Daudet, L. & Davies, M. E. (2009). Parametric dictionary design for sparse coding. *IEEE Transactions on Signal Processing*, *57*(12), 4800-4810.

張庭豪(2015)。**調變頻譜分解之改良於強健性語音辨識**(碩士論文)。取自 http://etds.lib.ntnu.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=rs0dbQ/record?r1=1&h1=0
[Chang, T.-H. (2015). *Several Refinements of Modulation Spectrum Factorization for Robust Speech Recognition* (Master's thesis). Retrieved from http://etds.lib.ntnu.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=rs0dbQ/record?r1=1&h1=0]

Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 55-70

評估尺度相關最佳化方法於華語錯誤發音檢測之研究

Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection

許曜麒*、楊明翰*、洪孝宗*、林奕儒*、陳冠宇+、陳柏琳*

Yao-Chi Hsu, Ming-Han Yang, Hsiao-Tsung Hung,

Yi-Ju Lin, Kuan-Yu Chen and Berlin Chen

摘要

錯誤發音檢測(Mispronunciation Detection)與錯誤發音診斷(Mispronunciation Diagnosis)為電腦輔助發音訓練系統的一部分，它們能輔助第二外語學習者準確地找出語句中錯誤發音的部位以增進學習者的口說熟練度。本論文延續過去學者的研究，大致可將貢獻分為三點：1) 比較不同的發音分數做為錯誤發音檢測的評估依據，並探討對於錯誤發音檢測效能的影響；2) 我們透過最佳化評估尺度相關訓練法則估測深層類神經網路聲學模型的參數以及發音檢測決策函數之參數；3) 使用 F_1 度量作為目標函數時，若將二類的 F_1 度量線性組合並調整權重，可有效處理資料類別不平衡的問題。一系列的實驗將建立在華語錯誤發音檢測與診斷任務，從實驗中可以觀察到我們提出的方法之優點。

關鍵詞：電腦輔助發音訓練、錯誤發音檢測、自動語音辨識、鑑別式訓練與深層類神經網路。

*國立台灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {ychsu, mh_yang, alexhung, lin_yj, berlin}@ntnu.edu.tw
+中央研究院資訊科學所
Institute of Information Science, Academia Sinica
E-mail: kychen@iis.sinica.edu.tw

Abstract

Mispronunciation detection and diagnosis are part and parcel of a computer assisted pronunciation training (CAPT) system, collectively facilitating second-language (L2) learners to pinpoint erroneous pronunciations in a given utterance so as to improve their spoken proficiency. This thesis presents a continuation of such a general line of research and the major contributions are three-fold. First, we compared the performance of different pronunciation features in mispronunciation detection. Second, we propose an effective training approach that estimates the deep neural network based acoustic models involved in the mispronunciation detection process by optimizing an objective directly linked to the ultimate evaluation metric. Third, we can linearly combine two F_1-score when we consider F_1-score as final objective function. It can effectively deal with the label imbalance problem. A series of experiments on a Mandarin mispronunciation detection task seem to show the performance merits of the proposed methods.

Keywords: Computer Assisted Pronunciation Training, Mispronunciation Detection, Automatic Speech Recognition, Discrimetive Training, Deep Neural Networks.

1. 緒論

全球化時代來臨，為提升個人的競爭力，外語能力已列為基本的技能之一。因此電腦輔助語言學習(Computer Assisted Language Learning, CALL)在現今已是非常具有潛力的研究；其目的是透過電腦自動判斷外語學習者的學習狀況並給予有幫助的回饋。近年來，由於中國市場的快速發展，全球華語學習熱潮席捲而來，學習華語的人數預估已經超過一億。在許多非華語語系的亞洲、歐洲以及美洲國家，華語已經逐漸成為一種必須學習的語言(Hu, Qian, Soong & Wang, 2014)。語言學習可分為聽(Listening)、說(Speaking)、讀(Reading)和寫(Writing)等四類學習面向，其中口說與書寫測驗的評量往往需要專業的語言教師來評斷，但語言教師的培養尚無法滿足遍佈全球的需求。本篇論文將專注於電腦輔助發音訓練(Computer Assisted Pronunciation Training, CAPT)，也就是「說」的技術進行討論。

電腦輔助發音訓練最主要目的就是要讓第二外語(Second-Language, L2)學習者有更多的機會練習發音；過去第二外語學習者要進行發音練習都需要配合語言教師的授課時間，若將電腦輔助發音訓練普及到現有的智慧型行動裝置，將會有更多的第二外語學習者因此受惠。電腦輔助發音訓練中的首要任務為自動錯誤發音檢測；檢測過程是請學習者讀誦口說教材，針對學習者念誦的錄音，標記學習者的發音是正確發音(Correct Pronunciation)或錯誤發音(Mispronunciation)，標記的目標可以是音素(Phone)層次(Witt & Young, 2000)、音節(Syllable)層次(Zhang, Huang, Soong, Chu & Wang, 2008)或詞(Word)層次(Chen & Jang, 2015)。當系統指出學習者的錯誤發音時，將可以針對該錯誤發音進行

偏誤回饋，該階段被稱為錯誤發音診斷(Harrison, Lau, Meng & Wang, 2008; Harrison, Lo, Qian & Meng, 2009; Lo, Zhang & Meng, 2010; Wang & Lee, 2012; Wang & Lee, 2015)。錯誤發音檢測為電腦輔助發音訓練中的第一步，當錯誤發音檢測可以精準的預測學習者的發音狀況時，才能有效的進行錯誤發音診斷。本研究主旨在探討如何提升錯誤發音檢測之效能？錯誤發音檢測可視為二類分類問題，對於發音檢測的結果在語言專家與系統的決策之間可以產生四種結局：若學習者的發音正確，系統卻判斷為發音錯誤稱為是錯誤的拒絕(False Rejections, FR)；而學習者發音錯誤，系統認定為發音正確則稱為錯誤的接受(False Acceptances, FA)；學習者發音正確，系統判斷為發音正確稱為正確的接受(True Acceptances, TA)；學習者發音錯誤，系統判定為發音錯誤稱為正確的拒絕(True Rejections, TR)。上述的四種指標可以計算出其它評估的標準，例如召回率(Recall)與精準度(Precision)，有許多發音檢測的研究皆以該評估方式作為評量系統優劣的準則(Hu *et al.*, 2015; Huang, Xu, Wang & Silamu, 2015)。我們可更進一步使用召回率與精準度的調和平均－F_1度量(F_1-Score)做為準則，F_1度量在自然語言處理(Natural Language Processing, NLP)與資訊檢索(Information Retrieval, IR)等研究中廣為使用，甚至有許多任務直接將該指標作為模型訓練的目標(Fujino, Isozaki & Suzuki, 2008; Dembczynski, Waegeman, Cheng & Hüllermeíer, 2011; Ye, Chai, Lee & Chieu, 2012)。在錯誤發音檢測任務中也有類似想法的研究已被探討(Huang *et al.*, 2015; Qian, Soong & Meng, 2010; Huang, Wang & Abudureyimu, 2012)。

近年來，在語音辨識系統中的聲學模型已由深層類神經網路(Deep Neural Network, DNN)取代傳統的高斯混合模型(Gaussian Mixture Model, GMM)，並在語音辨識任務上取得巨大的進步(Hinton *et al.*, 2012)。在錯誤發音檢測的相關研究中也因為深層類神經網路聲學模型的使用而在效能上有顯著的提升(Hu *et al.*, 2015; Qian, Meng & Soong, 2012; Hu *et al.*, 2014)。基於上述研究的啟發，我們延續過去學者以最大化錯誤發音檢測任務的效能(Huang *et al.*, 2015; Hsu, Yang, Hung & Chen, 2016)為目標函數對模型進行調整的想法，並實作於深層類神經網路聲學模型的架構上探討對於錯誤發音檢測任務的影響。

本篇論文在第二節將介紹錯誤發音檢測相關研究的發展近況；第三節簡單的回顧錯誤發音檢測任務中較常被使用的方法；第四節則是討論基於第三節的發音檢測方法如何實現最大化錯誤發音檢測F_1度量之訓練；第五節則是從實驗中探討最大化錯誤發音檢測F_1度量之訓練對於發音檢測任務的影響；最後，在第六節，我們提出結論與一些未來可能的研究方向。

2. 文獻探討

錯誤發音檢測大致可分為基於門檻值(Thresholding-Based)與基於分類器(Classification-Based)等兩種做法。兩者差別在於是否使用明確的門檻值來判斷發音為正確或錯誤；基於分類器則是整合多種特徵並訓練二元分類器來決定發音是否合格。基於門檻值等方法早期由(Hsu *et al.*, 2016)提出三種發音檢測特徵：對數相似度值(Log-Likelihood)、對數事後機率(Log Posterior Probability)、段落區間長度(Segment

Duration)對於發音檢測效果的影響。學者 Kim 在實驗中指出對數事後機率為表現較好的發音檢測分數 (Kim, Franco & Neumeyer, 1997)。之後則有學者簡化事後機率的計算方式並將其稱作 GOP (Goodness of Pronunciation) (Witt & Young, 2000)，之後也有研究針對 GOP 等方法進行改良(Zhang *et al*., 2008)。因為基於門檻值之方法展現了簡潔有效的優點，所以學者們提出以最大化錯誤發音檢測之F_1度量作為目標對聲學模型進行鑑別式訓練 (Huang *et al*., 2012)。

而基於分類器的發音檢測方法，較早是由(Wei, Hu, Hu & Wang, 2009)所提出的，陸續也有許多不同的發音特徵(Lee & Glass, 2012; Laborde *et al*., 2016)或是不同分類模型 (Hu *et al*., 2015)。事實上，在錯誤發音診斷任務中，早已開始整合多種特徵，例如引入韻律特徵(Strik, Truong, De Wet & Cucchiarini, 2007)。但這類的做法都只有在特定的發音才能使用(例如荷蘭語的/x/或/k/)，且韻律特徵容易因為不同語者而產生無法預期的變化。然而，也有學者基於語音辨識模組來進行發音診斷(Hu *et al*., 2015)，但從實驗數據看來距離理想的準確率還有一段差距。有些學者認為錯誤發音檢測與診斷應該要視為語音辨識的任務(Harrison *et al*., 2008; Harrison *et al*., 2009; Qian *et al*., 2012)，將訓練資料的錯誤型態(Error Pattern)都記錄在模型中；倘若測試資料出現訓練時從未的錯誤型態，辨識結果將會無法預期，且該情況會因為外語學習者的母語不同使得更容易發生。

3. 錯誤發音檢測

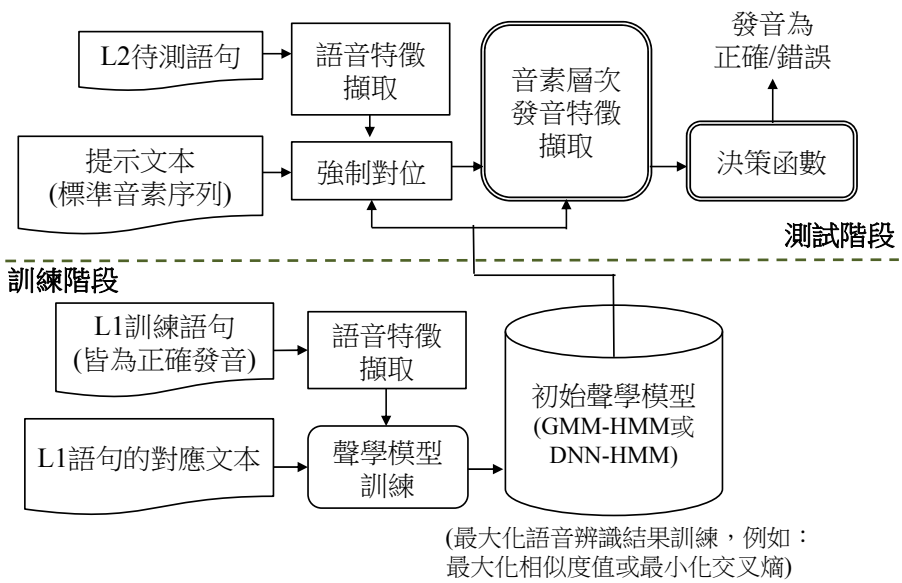


圖 1. 基礎錯誤發音檢測流程圖
[Figure 1. The flowchart of the mispronunciation detection process.]

錯誤發音檢測的基本流程如圖 1 所示。我們首先使用母語者的語料庫訓練語音辨識所需的聲學模型，在將外語學習者的發音語句與正確的文本做強制對位；接著將聲學模型算出的事後機率作為發音檢測特徵進行錯誤發音檢測。錯誤發音檢測的早期研究中，有學者延續(Kim *et al*., 1997)的觀察並將事後機率改良並稱作 GOP (Witt & Young, 2000)，也是最常被使用的發音檢測方法。GOP 的計算方式如下：

$$\mathrm{GOP}(u,n) \equiv \frac{1}{T_{u,n}} \log P\left(q_{u,n}|\mathbf{O}_{u,n}\right) \tag{1}$$

$$= \frac{1}{T_{u,n}} \log \frac{p(\mathbf{O}_{u,n}|q_{u,n})P(q_{u,n})}{\sum_{\tilde{q}\in Q_{u,n}} p(\mathbf{O}_{u,n}|\tilde{q})P(\tilde{q})} \tag{2}$$

$$\approx \frac{1}{T_{u,n}} \log \frac{p(\mathbf{O}_{u,n}|q_{u,n})}{\max_{\tilde{q}\in Q_{u,n}} p(\mathbf{O}_{u,n}|\tilde{q})} \tag{3}$$

其中 GOP 是音素段落$\mathbf{O}_{u,n}$對應目標音素$q_{u,n}$的事後機率，其中 u 與 n 表示第 u 個語句的第 n 個音素，根據貝氏定理將式(1)轉換成式(2)；$Q_{u,n}$是該段落對應的音素集合，可以是全部音素或部分較混淆的音素，$T_{u,n}$則是音素段落的經歷時間(Duration)。我們假設每個音素的事前機率相同，且只使用最大相似度值的音素，即最混淆音素做為分母項，如式(3)。其中$p\left(\mathbf{O}_{u,n}|q_{u,n}\right)$是已知音素$q_{u,n}$要取得音素段落$\mathbf{O}_{u,n}$的相似度值，計算$p\left(\mathbf{O}_{u,n}|q_{u,n}\right)$可以透過已知的文本內容對語句進行強制對位取得對應音素$q_{u,n}$的狀態序列$\mathbf{s}^* = \{s_{t_s}, s_{t_s+1}, \dots, s_{t_e}\}$，同時也可以得到音素段落區間對應的起始時間$t_s$與結束時間$t_e$。式(3)所計算的 GOP 分數作為決策發音錯誤與否的評估依據，並經過式(3)決定發音程度的分數。我們定義函數 D(·)表示發音的決策函數：

$$\mathrm{D}(u,n) = \frac{1}{1+\exp(\alpha\cdot\mathrm{GOP}(u,n)+\beta)} \tag{4}$$

而 D(·)接近 1 表示發音可能錯誤，接近 0 則表示發音正確，β表示決策用的門檻值，而參數α用來將 GOP 分數放大或縮小。上述兩個參數皆可以設計為音素相依，若為音素相依則用$\boldsymbol{\alpha}$與$\boldsymbol{\beta}$表示。接著我們利用指示函數判定發音是否錯誤：

$$\mathbb{I}(\mathrm{D}(u,n)) = \begin{cases} 1 & \text{if } \mathrm{D}(u,n) \geq \tau \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

τ為全域的固定門檻值，大部分都是透過發展集調整至一個較合適的值。然而 GOP 是錯誤發音檢測的方法中較普遍的作法，且不需依賴人工標記好的錯誤發音，屬於非監督式學習(Unsupervised Learning)的方法。

此外，已有學者提出利用深層類神經網路聲學模型的輸出為事後機率$P(s_t|\mathbf{o}_t)$的方法作為發音檢測的分數，稱作對數音素事後機率(Log Phone Posterior, LPP) (Hu *et al*., 2015)。其計算方式為音素段落$\mathbf{O}_{u,n}$對應的狀態事後機率之幾何平均。與 GOP 的算法類似，透過已知的文本內容對語句進行強制對位取得對應目標音素$q_{u,n}$的狀態序列$\mathbf{s}^{(q_{u,n})} =$

$\{s_{t_s}, s_{t_s+1}, \ldots, s_{t_e}\}$，而計算 LPP 的公式可以寫成：

$$\mathrm{LPP}(u,n) = \log P(q_{u,n}|\mathbf{O}_{u,n}; t_s, t_e) \tag{6}$$

$$\approx \frac{1}{t_e - t_s + 1} \sum_{t=t_s}^{t_e} \log P\left(s_t^{(q_{u,n})} \middle| \mathbf{o}_t\right) \tag{7}$$

透過式(7)算出目標音素$q_{u,n}$的 LPP，$\mathbf{s}^{(q_{u,n})}$為音素$q_{u,n}$在音素段落$\mathbf{O}_{u,n}$的最佳路徑所對應的狀態序列。從我們實驗中可以發現使用 LPP 產生發音分數在發音檢測任務的效果與 GOP 相近，但 LPP 的計算複雜度遠低於 GOP。如式(3)所見，GOP 在分母項需要將所有音素的相似度值算出；而 LPP 只需要計算目標音素$q_{u,n}$的狀態事後機率之幾何平均，符合深層類神經網路架構的輸出狀態事後機率。當我們取得以 LPP 表示的發音分數後，寫成決策函數的形式則為：

$$\mathrm{D}(u,n) = \frac{1}{1+\exp(\alpha \cdot \mathrm{LPP}(u,n) + \beta)} \tag{8}$$

4. 最大化錯誤發音檢測$\mathbf{F_1}$度量之訓練

在發音檢測任務中有許多研究都以改良 GOP 為主軸提升錯誤發音檢測的效能，近期有學者將鑑別式訓練應用在 GOP 估測，以最大化 F_1 度量為目標作鑑別式訓練(Huang *et al.*, 2015)，學者 Huang 使用高斯混合模型-隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)建構聲學模型，並使用 GOP 進行錯誤發音檢測，透過調整聲學模型中的參數來提升錯誤發音檢測的表現。而在本論文，我們將聲學模型改用深層類神經網路-隱藏式馬可夫模型(Deep Neural Networks-Hidden Markov Model, DNN-HMM)，在錯誤發音檢測的部分則用 LPP 做為發音分數，透過決策函數決定發音是否錯誤；並以最大化$\mathrm{F_1}$度量為目標做鑑別式訓練更新深層類神經網路聲學模型的參數以及決策函數的參數。首先，$\mathrm{F_1}$度量的計算方式如下：

$$F_1 = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{9}$$

其中F_1為召回率與準確率兩種評估標準的調合平均，而召回率與準確率可以表示為：

$$\mathrm{Precision} = \frac{C_{H \cap D}}{C_D} \tag{10}$$

$$\mathrm{Recall} = \frac{C_{H \cap D}}{C_H^{(\mathcal{M})}} \tag{11}$$

C_D表示訓練資料中被系統標記為錯誤發音的音素數量；$C_H^{(\mathcal{M})}$則是訓練資料中被語言專家標記為錯誤發音的音素數量，因此該值是一個固定的常數；而$C_{H \cap D}$就是系統與語言專家同時認為該音素段落為錯誤發音的數量。將式(10)與式(11)代入式(9)並簡化後得到：

$$F_1 = \frac{2C_{H\cap D}}{C_D + C_H^{(\mathcal{M})}} \tag{12}$$

接著將我們在第 3 節定義的錯誤發音決策函數 D(·)透過指示函數𝟙(·)轉成非 1 即 0 的數值，訓練資料的所有音素段落經過決策函數 D(·)與指示函數𝟙(·)的總和為C_D；每個音素段落的決策與專家評斷之結果 H(·)相乘的總和則為$C_{H\cap D}$，如式(13)：

$$F_1 = \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathbb{1}(\mathrm{D}(u,n))\cdot \mathrm{H}(u,n)}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathbb{1}(\mathrm{D}(u,n)) + C_H^{(\mathcal{M})}} \tag{13}$$

然而上述定義的F_1度量並不是可微分的函數，因為在計算$C_{H\cap D}$與C_D時使用到的指示函數𝟙(·)在基於梯度法(Gradient Based Method)的參數更新方式中較難處理。因此我們定義一個平滑(Smooth)的F_1度量，如式(14)：

$$\Xi(\boldsymbol{\theta}) = \frac{2\cdot C_{H\cap D}^{S}}{C_D^{S} + C_H^{(\mathcal{M})}} \tag{14}$$

$$= \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)\cdot \mathrm{H}(u,n)}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n) + C_H^{(\mathcal{M})}} \approx F_1 \tag{15}$$

由於錯誤發音決策函數 D(·)已將發音檢測分數經過激發函數轉為 0 到 1 之間的值，因此計算$C_{H\cap D}^{S}$與C_D^{S}時不使用指示函數 𝟙(·)也可以近似F_1度量的算法，如式(15)。定義了目標函數後，我們使用隨機梯度上升法(Stochastic Gradient Ascent Algorithm)來更新參數。最後對於最大化F_1度量鑑別式訓練的流程列出摘要：

(1) 首先透過華語母語者(L1)訓練資料使用於 DNN-HMM 聲學模型的訓練，且訓練資料皆為正確發音，並以最小化交叉熵為目標函數更新聲學模型。

(2) 基於步驟(1)訓練的聲學模型，透過第 3 節提及的 LPP 算法(式(7))得出每筆訓練資料的發音分數，接著透過決策函數(式(8))將發音分數轉成決策值(值域 0 到 1 之間)。

(3) 接續步驟(2)算出的決策值透過式(15)算出近似的F_1度量作為目標函數並迭代的訓練決策函數的參數$\boldsymbol{\alpha}, \boldsymbol{\beta}$以及 DNN-HMM 聲學模型的參數，而決策函數的參數可為音素相依。

相較於原本的流程在訓練資料中加入了二語(L2)的資料(包含正確與錯誤發音)；且決策函數與聲學模型的參數也以發音檢測任務的目標函數進行調適。

5. 實驗結果

5.1 華語學習者口語語料庫

本論文使用臺灣師範大學邁向頂尖大學計畫所提供的華語學習者口語語料庫(Hsiung & Sung, 2014)，語者部分包含華語母語者(L1)與華語非母語者(L2)，錄音內容有單音節、雙音節、多音節與短文等情境；其中華語非母語者語料庫為音素層次的發音標記，每筆資

料皆是由 1 至 4 人進行審視，並採用多數決判斷發音為正確或錯誤。我們將語料庫分成訓練集、發展集與測試集，如表 1。

表 1. 華語學習者口語語料庫
[Table 1. Statistics of the Mandarin Annotated Spoken Corpus.]

		時間(小時)	語者(個)	音素數量(個)	發音錯誤之音素數量(個)
訓練集	L1	6.68	44	72,846	NA
	L2	14.04	74	107,202	24,150
發展集	L1	1.4	10	14,186	NA
	L2	3.39	18	25,900	5,227
測試集	L1	3.21	25	32,568	NA
	L2	7.49	44	55,190	14,247

5.2 聲學模型訓練

本研究的語音辨識模組的建立是使用美國約翰霍普金斯大學學者所發展的大詞彙連續語音辨識工具－"Kaldi"(Povey *et al*., 2011)；其它的實驗以 Python 程式語言為主，並參考各種函式庫像是"Scikit-learn"(Pedregosa *et al*., 2011)和"Theano"(Bergstra *et al*., 2010)等，其提供機器學習或深層學習與 GPGPU 運算結合的開發環境。在聲學模型的設定我們採每個音素基於 GMM-HMM 的聲學模型皆由 3 個狀態所組成，而每個狀態至少 16 個高斯混合而成，並以的 L1 訓練集與 L1 發展集作為訓練資料調整聲學模型參數。輸入特徵為梅爾倒頻譜係數，每個音框由 12 維梅爾倒頻譜係數、1 維能量特徵和 3 維度音調(pitch)特徵所組成；並對 16 維語音特徵取相對的一階差量係數(Delta Coefficient)和二階差量係數(Acceleration Coefficient)合併成 48 維的特徵向量；其中取一階和二階差量係數是為了獲得語音特徵在時間的相關資訊。

在 DNN-HMM 聲學模型的部分，每個隱藏層皆使用邏輯函數(sigmoid function)作為激發函數，到輸出層則使用軟式最大化函數(Softmax)轉換成機率。輸入特徵是梅爾頻譜係數(Mel-Scale Frequency Spectral Coefficients, MFSC)取得的對數能量特徵並透過濾波器組(Filter Banks)所產生的 40 維輸出；鄰近音窗我們採用前後各 5 個音框，共含 11 個音框，每個音框皆為 40 維的濾波器組產生加上 3 維度音調(Pitch)特徵；並對 43 維語音特徵取相對的一皆差量係數和二皆差量係數，則輸入的語音特徵就會得到 11 個 129 維的特徵向量串成一個 1,419 維度的特徵向量。

在自動語音辨識的結果我們以音節錯誤率(Syllable Error Rate, SER)與音素錯誤率(Phone Error Rate, PER)來表示，如表 2；解碼過程為自由音節解碼(Free-Syllable Decoding)且無任何語言模型限制，辨識錯誤率是使用表 1 的 L1 測試集所計算的。從表 2 的辨識結果可以觀察到無論是音節錯誤率或是音素錯誤率皆由 DNN-HMM 聲學模型大幅度勝

過 GMM-HMM 聲學模型。

表 2. 自動語音辨識實驗結果
[Table 2. ASR experimental results.]

	音節錯誤率(%) (syllable error rate, SER)	音素錯誤率(%) (phone error rate, PER)
GMM-HMM	50.87	34.30
DNN-HMM	41.71	28.14

5.3 評估方法

表 3. ROC 分析的四項指標在發音檢測任務中的定義
[Table 3. The definition of the confusion matrix used in the mispronunciation detection task.]

	描述
錯誤的接受 (false acceptances, FA)	實際上學習者的發音錯誤，系統卻認定為發音正確。
錯誤的拒絕 (false rejections, FR)	實際上學習者的發音正確，系統卻判斷為發音錯誤。
正確的接受 (true acceptances, TA)	實際上學習者的發音正確，系統也判斷為發音正確。
正確的拒絕 (true rejections, TR)	實際上學習者的發音錯誤，系統也認定為發音錯誤。

如同本論文在第 1 節提及的，二分類問題會有四種結局(如表 3)，基於這四項指標可以延伸出非常多變的評估方式；例如召回率與精準度是分類問題中經常被使用的評估方式，而召回率與精準度的調和平均，也就是F_1度量更是廣為使用。無論是正確或錯誤發音的檢測結果都是重要的指標，因此我們先定義正確發音檢測的召回率($\mathrm{Recall}_{\mathcal{C}}$)、精準度($\mathrm{Precision}_{\mathcal{C}}$)與$\mathrm{F}_1$度量($\mathrm{F1}_{\mathcal{C}}$)的計算方式：

$$\mathrm{Recall}_{\mathcal{C}} = \frac{\text{正確接受(TA)的個數}}{\text{實際為正確發音的個數}} = \frac{\#\mathrm{TA}}{\#\mathrm{TA}+\#\mathrm{FR}} \tag{16}$$

$$\mathrm{Precision}_{\mathcal{C}} = \frac{\text{正確接受(TA)的個數}}{\text{系統判斷為正確發音的個數}} = \frac{\#\mathrm{TA}}{\#\mathrm{TA}+\#\mathrm{FA}} \tag{17}$$

$$\mathrm{F1}_{\mathcal{C}} = \frac{2\cdot\mathrm{Recall}_{\mathcal{C}}\cdot\mathrm{Precision}_{\mathcal{C}}}{\mathrm{Recall}_{\mathcal{C}}+\mathrm{Precision}_{\mathcal{C}}} \tag{18}$$

而錯誤發音檢測的召回率($\mathrm{Recall}_{\mathcal{M}}$)、精準度($\mathrm{Precision}_{\mathcal{M}}$)與$\mathrm{F}_1$度量($\mathrm{F1}_{\mathcal{M}}$)的計算方式為：

$$\text{Recall}_{\mathcal{M}} = \frac{\text{正確拒絕(TR)的個數}}{\text{實際為錯誤發音的個數}} = \frac{\#\text{TR}}{\#\text{TR}+\#\text{FA}} \tag{19}$$

$$\text{Precision}_{\mathcal{M}} = \frac{\text{正確拒絕(TR)的個數}}{\text{系統判斷為錯誤發音的個數}} = \frac{\#\text{TR}}{\#\text{TR}+\#\text{FR}} \tag{20}$$

$$\text{F1}_{\mathcal{M}} = \frac{2\cdot\text{Recall}_{\mathcal{M}}\cdot\text{Precision}_{\mathcal{M}}}{\text{Recall}_{\mathcal{M}}+\text{Precision}_{\mathcal{M}}} \tag{21}$$

精準度的資訊在其它常見的評估方式(準確率(Accuracy)或ROC曲線等)中不易觀察，但對於錯誤發音檢測任務而言精準度也是非常重要的指標之一，因此同時考慮召回率與精準度的F_1度量指標為本論文後續實驗討論最常使用的評估標準。

5.4 錯誤發音檢測實驗

延續 5.2 小節設定的初始聲學模型(GMM-HMM 與 DNN-HMM)，並使用第 3 節提到的 GOP 分數(式(3))作為評估發音品質的特徵；並代入決策函數(式(4))與指示函數(式(5))檢測學習者的發音為正確或錯誤，決策函數與指示函數的參數皆使用全域的數值(未調整為音素相依或音素狀態相依)，其結果如表 4 所示。由表 4 可以得知基於 DNN-HMM 作為聲學模型產生 GOP 分數並應用在發音檢測任務效果更勝 GMM-HMM 聲學模型的效果發音檢測的F_1度量皆有約 3%的絕對進步(正確發音檢測的F_1度量由 0.836 提升至 0.863；錯誤發音檢測的F_1度量由 0.546 提升至 0.579)。已有許多學者在實驗中證明了深層學習在發音檢測任務的突破(Hu *et al*., 2015; Qian *et al*., 2012; Hu *et al*., 2014)。

表 4. 不同聲學模型在發音檢測任務的實驗結果
[Table 4. Mispronunciation detection results achieved by using different acoustic models.]

GOP	Correct pronunciation detection			Mispronunciation Detection		
	Recall	Precision	F1	Recall	Precision	F1
GMM-HMM	0.828	0.844	0.836	0.562	0.532	0.546
DNN-HMM	0.877	0.849	0.863	0.552	0.609	0.579

接著我們探討發音檢測任務使用第 3 節所提到的對數音素事後機率(LPP)作為發音分數，如表 5。GOP 與 LPP 的方法在F_1度量的表現相近(正確發音檢測的F_1度量由 0.863 降低至 0.854；錯誤發音檢測的F_1度量由 0.579 提升至 0.587)，其變化皆在 0.01 之間。如第 3 節提到的 LPP 的計算複雜度遠低於 GOP，因此接下來的實驗將以 LPP 作為主要的發音分數。

表 5. 基於 DNN-HMM 聲學模型使用不同發音分數的發音檢測基礎實驗
[Table 5. Mispronunciation detection results achieved by incorporating the DNN-HMM acoustic model with different decision features.]

	Correct pronunciation detection			Mispronunciation Detection		
	Recall	Precision	F1	Recall	Precision	F1
GOP	0.877	0.849	0.863	0.552	0.609	0.579
LPP	0.850	0.857	0.854	0.594	0.580	0.587

本論文探討的主題為最大化發音檢測效能之鑑別式訓練。首先我們定義欲進行發音檢測的聲學模型與決策函數，延續 5.2 小節的語音辨識基礎實驗中表現最好的聲學模型 DNN-HMM，以及在發音檢測任務上效果不遜色於 GOP 所提供的發音分數，也就是 LPP 發音分數作為發音分數；並透過非線性的邏輯函數轉換為決策值，供我們計算評估該模型的表現。然而，在第 4 節討論的F_1度量之目標函數為錯誤發音檢測的F_1度量，但是在電腦輔助發音訓練等任務中正確發音檢測也是非常重要的部分，我們不希望系統對於學習者本身正確的發音造成誤判。延續第 4 節的定義的F_1度量目標函數，並再次定義錯誤/正確發音檢測的F_1度量近似算法：

$$\Xi_{\mathcal{M}}(\boldsymbol{\theta}) = \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)\cdot\mathrm{H}(u,n)}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}\mathrm{D}(u,n)+C_H^{(\mathcal{M})}} \tag{22}$$

$$\Xi_{\mathcal{C}}(\boldsymbol{\theta}) = \frac{2\sum_{u=1}^{U}\sum_{n=1}^{N_u}(1-\mathrm{D}(u,n))\cdot(1-\mathrm{H}(u,n))}{\sum_{u=1}^{U}\sum_{n=1}^{N_u}(1-\mathrm{D}(u,n))+C_H^{(\mathcal{C})}} \tag{23}$$

最後我們使用參數 φ 作為正確發音與錯誤發音的F_1度量之線性組合作為最終的目標函數：

$$\tilde{\Xi}(\boldsymbol{\theta}) = \varphi\cdot\Xi_{\mathcal{M}}(\boldsymbol{\theta}) + (1-\varphi)\cdot\Xi_{\mathcal{C}}(\boldsymbol{\theta}) \tag{24}$$

從我們的實驗中可以觀察到參數 φ 對於結果發音檢測發展集的影響，如圖 2，以最大化F_1度量調整決策函數的參數(未更新聲學模型)。從圖中可以發現參數 φ 對於錯誤發音檢測結果的影響十分顯著(圖 2 下半部)，當 φ=0.8 時在錯誤發音檢測有最好的效果(F_1度量為 0.566，基礎實驗的F_1度量為 0.527)。但是在正確發音中並非 φ=0.8 為效果最佳，但參數 φ 的調整對於正確發音檢測效能的影響較小，因此我們挑選參數 φ 則以錯誤發音檢測之效能為優先考量。有趣的是，在訓練資料中正確發音與錯誤發音的比例正好接近 0.8，這表示透過調整參數 φ 的方式巧妙的處理了資料類別不平衡的問題。

基於圖 2 的實驗結果，我們將固定 φ=0.8 並依續探討最大化發音檢測效能之鑑別式訓練對不同階段的參數進行調整所帶來的影響。在表 6 中我們基於 LPP 所算出的發音分數透過決策函數並算出整體的F_1度量，以最大化F_1度量更新決策函數(+MFC (DF))或聲學模型(+MFC (AM))的參數，甚至是決策函數與聲學模型的參數同時調整(+MFC (Both))。從表 6 可以發現無論是更新任何階段的參數在發音檢測任務上都可以得到顯著的提升。首先討論更新決策函數的參數(+MFC (DF))，與基礎實驗相比(LPP)則有明顯的提升；而

只更新聲學模型參數(+MFC (AM))可以得到更好的效果，若同時更新兩階段的參數(+MFC (Both))效果最佳。從實驗結果中可以發現更新聲學模型參數有著最大幅度的進步，由此可見初始的聲學模型是為語音辨識任務所設計，若經過調適則可以得到更好的效果。

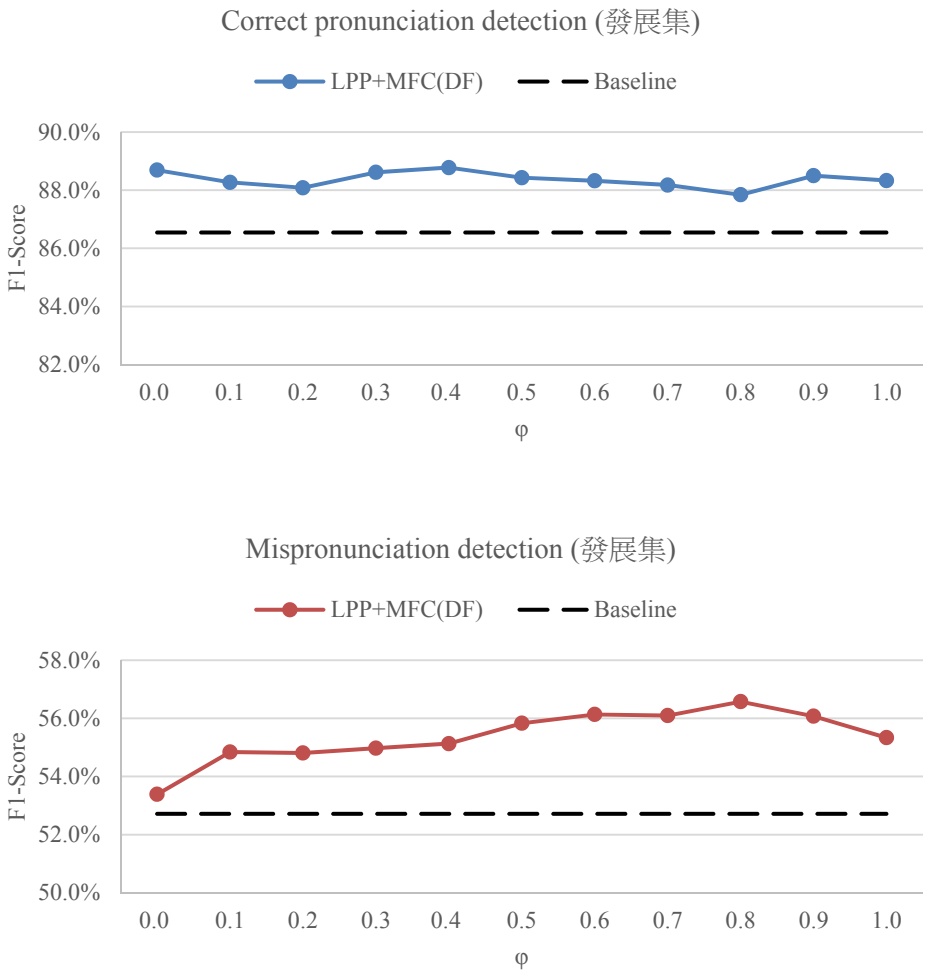


圖 2. 不同 φ 在發展集的發音檢測效能

[Figure 2. Mispronunciation detection results on the development set with different threshold values φ.]

表 6. 基於 LPP 最大化 F1 度量鑑別式訓練於不同設定的發音檢測效能
[Table 6. Mispronunciation detection results achieved by using LPP features with/without MFC training.]

	Correct pronunciation detection			Mispronunciation detection		
	Recall	Precision	F1	Recall	Precision	F1
LPP	0.850	0.857	0.854	0.594	0.580	0.587
+MFC (DF)	0.863	0.866	0.865	0.617	0.611	0.614
+MFC (AM)	0.906	0.870	0.888	0.612	0.694	0.650
+MFC (Both)	0.907	0.871	0.889	0.613	0.697	0.652

6. 結論與未來展望

本論文著重在電腦輔助發音訓練的錯誤發音檢測任務，並以最佳化錯誤發音檢測效能為主軸進行一系列的實驗。基於過去學者的研究，我們認為以最大化發音檢測之F_1度量為目標函數進行模型訓練是非常有潛力的。因此我們延伸該作法至現今語音辨識模組十分熱門的部份－深層類神經網路聲學模型，取代傳統的高斯混合聲學模型。從實驗結果可以發現以最大化F_1度量為目標對決策函數或聲學模型的參數進行調整，甚至是同時調整，都可以在效果上得到提升；尤其對於聲學模型參數進行調整的進步幅度令人印象深刻。且以F_1度量作為目標進行訓練在不同的評估方式也可以得到進步。在未來我們希望從特徵與模型等兩個面向來探討對於電腦輔助發音訓練任務的影響。在特徵的部分，我們期望從不同角度來獲取跟發音狀況高相關性的特徵，其中韻律特徵非常具有潛力；在模型的部分除了持續探討更新穎的聲學模型外，我們也預期將語音辨識所使用的調適技術移轉到該任務，例如一些非監督式的語者調適或是針對不同語言進行模型調適等。

致謝

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(104-2911-I-003-301)與行政院科技部研究計畫(MOST 104-2221-E-003-018-MY3 和 MOST 105-2221-E-003-018-MY3)之經費支持，謹此致謝。

參考文獻 References

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D. & Bengio, Y. (2010). Theano: A CPU and GPU math compiler in Python. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 1-7.

Chen, L. Y. & Jang, J. S. R. (2015). Automatic pronunciation scoring with score combination by learning to rank and class-normalized DP-based quantization. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(11), 1737-1749.

Dembczynski, K. J., Waegeman, W., Cheng, W. & Hüllermeier, E. (2011). An exact algorithm for F-measure maximization. In *Advances in Neural Information Processing Systems*, 1404-1412.

Fujino, A., Isozaki, H. & Suzuki, J. (2008). Multi-label Text Categorization with Model Combination based on F1-score Maximization. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 823-828.

Harrison, A. M., Lau, W. Y., Meng, H. M. & Wang, L. (2008). Improving mispronunciation detection and diagnosis of learners' speech with context-sensitive phonological rules based on language transfer. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 2787-2790.

Harrison, A. M., Lo, W. K., Qian, X. & Meng, H. (2009). Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training. In *Proceedings of the International Symposium on Languages, Applications and Technologies (SLaTE)*, 45-48.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

Hsiung, Y. & Sung, Y. (2014). Development of Mandarin Annotated Spoken Corpus (MAS Corpus) and the Learner Corpus Analysis. *1st Workshop on the Analysis of Linguistic Features (WoALF)*.

Hsu, Y. C., Yang, M. H., Hung, H. T. & Chen, B. (2016). Mispronunciation detection leveraging maximum performance criterion training of acoustic models and decision functions. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 2646-2650.

Hu, W., Qian, Y. & Soong, F. K. (2014). A DNN-based acoustic modeling of tonal language and its application to Mandarin pronunciation training. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 3206-3210.

Hu, W., Qian, Y., Soong, F. K. & Wang, Y. (2015). Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers. *Speech Communication*, *67*, 154-166.

Huang, H., Wang, J. & Abudureyimu, H. (2012). Maximum F1-score discriminative training for automatic mispronunciation detection in computer-assisted language learning. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 815-818.

Huang, H., Xu, H., Wang, X. & Silamu, W. (2015). Maximum F1-score discriminative training criterion for automatic mispronunciation detection. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(4), 787-797.

Kim, Y., Franco, H. & Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 649-652.

Laborde, V., Pellegrini, T., Fontan, L., Mauclair, J., Sahraoui, H. & Farinas, J. (2016). Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 2686-2690.

Lee, A. & Glass, J. (2012). A comparison-based approach to mispronunciation detection. In *Proceedings of the International Conference on Spoken Language Technology Workshop (SLT)*, 382-387.

Lo, W. K., Zhang, S. & Meng, H. M. (2010). Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 765-768.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2825-2830.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer G. & Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Qian, X., Meng, H. M. & Soong, F. K. (2012). The Use of DBN-HMMs for Mispronunciation Detection and Diagnosis in L2 English to Support Computer-Aided Pronunciation Training. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 775-778.

Qian, X., Soong, F. K. & Meng, H. M. (2010). Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT). In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 757-760.

Strik, H., Truong, K. P., De Wet, F. & Cucchiarini, C. (2007). Comparing classifiers for pronunciation error detection. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 1837-1840.

Wang, Y. B. & Lee, L. S. (2012). Improved approaches of modeling and detecting error patterns with empirical analysis for computer-aided pronunciation training. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 5049-5052.

Wang, Y. B. & Lee, L. S. (2015). Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(3), 564-579.

Wei, S., Hu, G., Hu, Y. & Wang, R. H. (2009). A new method for mispronunciation detection using support vector machine based on pronunciation space models. *Speech Communication*, *51*(10), 896-905.

Witt, S. M. & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication*, *30*(2), 95-108.

Ye, N., Chai, K., Lee, W. & Chieu, H. (2012). Optimizing Fmeasures: a tale of two approaches. In *Proceedings of the International Conference on Machine Learning (ICML)*, 289-296.

Zhang, F., Huang, C., Soong, F. K., Chu, M. & Wang, R. (2008). Automatic mispronunciation detection for Mandarin. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 5077-5080.

Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 71-84

基於字元階層之語音合成用文脈訊息擷取

Character-Level Linguistic Features Extraction for Text-to-Speech System

陳冠宏*、廖書漢*、廖元甫*、王逸如+

Kuan-Hung Chen, Shu-Han Liao, Yuan-Fu Liao and Yih-Ru Wang

摘要

優良的語言文脈訊息是語音合成的關鍵部分，傳統的文脈訊息都是依賴於自然語言處理(Natural Language Processing，NLP)，即使用 parser 分析文字。但是 parser 設計困難無法專門為語音合成設計；所以我們想直接以字元為處理單元建立一個 end-to-end 的語音合成系統，在這想法下我們改用字元層級(character-level)的 word2vec 與遞迴類神經網路，直接將輸入字元序列轉換成隱藏特徵向量當做語言合成的文脈訊息。最後我們利用一中英夾雜語音合成系統測試此想法，語音合成的實驗的結果表明，我們提出的方式的確比傳統使用 parser 的方式有更好的性能。

Abstract

High quality linguistic features is the key to the success of speech synthesis. Traditional linguistic feature extraction methods are usually relied on a word-level natural language processing (NLP) parser. Since, a good parser requires a lot of feature engineering to build, it is usually a genral-purpose one and often not specially designed for speech synthesis. To avoid these difficulties, we propose to replace the conventional NLP parser by a character embedding and a chacter-level

* 國立台北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
E-mail: { s970428, sam8105111}@gmail.com; yfliao@mail.ntut.edu.tw
+ 國立交通大學電機工程系
College of Electricl and Computer Engineering, National Chiao-Tung University
E-mail: yrwang@mail.nctu.edu.tw

recurrent neural network language model (RNNLM) module to directly convert input character sequences, character-by-character, into latent linguistic feature vectors. Experimental results on Chinese-English speech synthesis system showed that the proposed approach achieved comparable performance with transitional NLP parser-based methods.

關鍵詞：語音合成、文脈訊息、文字向量、遞迴類神經網路語言模型

Keywords: Speech Synthesis、Linguistic Features、Word2vec、RNNLM

1. 簡介

在語音合成系統中分為兩大模組(如圖 1)，分別是文本分析(自然語言處理)與聲音合成(語音訊號處理)。其中前端的文本分析通常會做文字正規化、斷詞(word segmentation)、part of speech(POS)標註與相關文法分析，甚至是藉由韻律預測從文本提取文脈訊息特徵。例如在中文語音合成中經常採用條件隨機場(Conditional Random Fields，CRF)做斷詞和 POS 標註。另一方面在後端聲音合成一般會透過決策樹依據文脈訊息選擇最適合的聲音或韻律特徵，將選出來的聲學參數給語音編碼器合成語音波形產出聲音。因此若是我們想在語音合成系統中獲得自然、流暢的合成聲音，需要能萃取有效且有用的文脈訊息。

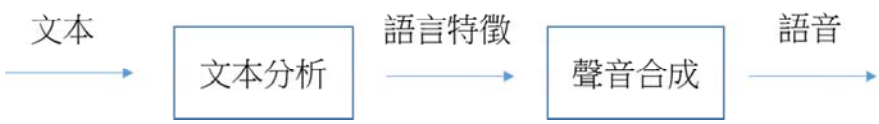


圖 1. 傳統 TTS 系統架構圖
[Figure 1. The tranditional two-stage TTS approach]

在傳統方式上，文本分析主要使用 NLP 的 parser(The Stanford Natural Language Processing Group, 2015)。使用 parser 的好處是經過語言學專家所設計的斷詞能夠比較完整的表現出每段語句內的詞性狀態。但是要建立傳統斷詞很難做，需要大量專業人士標註的資料庫以及專家設計好的斷詞特徵參數，所以我們通常只能使用現成的，無法專門根據合成的需求來設計。再加上標註的過程中不同的標註人員可能會對同一句話產生不一致的標註方式，這導致機器不好學習。所以為了解決這些問題，我們想要讓機器能夠自動去學習每一個字元的屬性與字元串間的關係。就像我們人在閱讀文章時不會先對文章做斷詞及 POS 標註，我們也是一個字一個字讀過去，然後透過前後文的拆解去理解字義、詞性。所以我們能不能想一個方法跟人在閱讀的時候一樣，利用以字為單元這種比較簡單的方式，避免使用 parser 做斷詞和 POS 標註的問題，以改善原有系統求取語言特徵的缺點。

近年來，類神經網路(DNNs) (Licstar, 2013)(在 NLP 方面有越來越多的研究，提出很多字元層級的語言模型，因為使用 DNNs 可以建立更好的模型，並且學習大量無須標註的文本資料，所以基於使用類神經網路的方法能夠有效的減少 POS 標註這類的語言特徵參數設計工程(feature engineering)；例如(Greff, Srivastava, Koutník, Steunebrink, &

Schmidhuber, 2015)使用以字元為層級的語言模型，在相同的性能下，會比 n-gram 模型小。而且對於中文的語音合成來說，以字元層級為處理單元的方式更具有意義，因為中文是沒有空格的連續字元串，也沒有詞的分隔符號，所以中文的詞在定義上是很模糊的，在這情況下許多以字元層級的處理方法方法被提出，像是(Zheng, Chen & Xu, 2013)中文斷詞與 POS 標註的深層學習網路以及(Ding, Xie, Yan, & Zhang, 2015)採用雙向長短期記憶遞迴類神經網路(BLSTM)直接從中文預測韻律邊界的標註，這些研究都證明了使用 DNNs 是能夠實現比傳統 CRF 有類似甚至更優良的效能。

在這麼多使用 DNNs 獲得不錯成果的研究下，我們想說看可不可以使用大量沒有經過標註的語料來訓練 DNN，這樣不僅可以免去標註工作也無需專家設計的斷詞或 POS 特徵參數，而要達成使用無標註語料的目的，我們需要靠字元層級(character-level)的文字處理的幫助，因此我們捨棄以詞為單元，將輸入簡化成以字元為單位，不經過 parser 做斷詞、求詞性等前處理，而是一個字元一個字元逐次輸入網路中；希望能將字元轉到更高維的向量空間，在向量空間進行分析，能有效的從語料中，自動學習字元之間隱藏的相對關係。所以我們將透過建立語料的字向量空間並進行分析，將文本字元分類，產生字元語意、文法腳色資訊等文脈訊息；另一方面則使用遞迴神經網路進行字元串訓練，分析目前輸入字元在整句話中的狀態，並猜測下一狀態可能為何，最後並擷取隱藏層在各個神經元的狀態輸出做為字元的時間順序資訊當作字元時序狀態的文脈訊息。如此一來我們就能利用大量未標註的文字語料，自動擷取文脈訊息，並探討更多種文脈訊息的可能性。

2. 傳統文脈訊息擷取方法

一般文本分析的文脈訊息是輸入文本經由 parser 做斷詞、抓 POS、位置再加上切割資訊、聲調資訊等合起來的，所以要得到好的文脈訊息這些參數需要精確和有用，而 parser 在文本分析中扮演舉足輕重的角色。

在語言學上，詞是能夠獨立運用而且含有語義內容的最小語言單位。在英文文本中，每個單字(word)即是一個詞，具有完整意義，而且每個單字間都以空白做區隔，但是在中文文本裡，詞與詞之間是不會有空白做為區隔的。因此在中文的 NLP 中，為了讓電腦能夠分辨文本中的詞義，就必須先正確的將詞區隔開來，才能進一步發展出相關演算法。例如機器翻譯、資訊檢索與擷取、語言分析、語音辨識和合成，為此發展出斷詞器來使用，而斷詞器主流使用 CRF 處理輸入文本，透過訓練 CRF 對輸入句子做猜測判斷兩個字中間是否為斷點，訓練方法要依照專家設計斷詞參數去學習怎麼預測詞斷點。但詞斷點要參照已經標註好的資料庫來學習，所以對於未知詞的錯誤率無法有效降低。如何改善中文斷詞的條件機率模型相關資料可以參照(黃昭銘，2010)，當中有提到使用一個線性的 CRF 來達成更準確的中文斷詞，如果分割出的詞與相對應標準語料庫的詞不同時，透過該詞和前後斷詞的重組，可求出更適當的斷詞。

詞性標註(Part Of Speech Tagging)在 NLP 中也是一大課題。一般 parser 就包含了斷詞與 POS Tagging 兩部分，順序上是先斷詞再做詞性標註，而詞性標註就是透過適當的方式對經過斷詞處理後的每個詞給予一個合適的詞性，也就是要確定這個詞是名詞、動詞或是副詞等等，關於詞性標註可以參閱(Brill, 1992)基於 POS 標註的簡單規則。但是詞性標註的困難點在於詞性不定的問題上，這種現象是自然語言中有很多詞語本身包含很多詞性，擺在不同地方就會變換詞性。對於人在閱讀而言，這種詞性歧義現象比較容易排除，但是對於機器而言則不容易區分。傳統使用的詞性標註規則是由語言學家根據語言規律進行人工標註完成的，有了標註完的資料庫，在使用上利用 CRF 從已知的序列求對應序列的方式，來猜測輸入語料的詞性。詳細內容可以參閱相關研究，例如(唐大任，2002)中文 parser 之研究，內文探討了設計 parser 時利用的斷詞與構詞規則，和標記詞類方式的一些問題。

在 NLP 中使用 parser 是非常普遍的，對英文而言使用 parser 能得到不錯的結果，但是在中文使用起來卻不盡理想，因為中文結構在歧義性上有太多變化，需要搭配上下文來觀察，無法單純靠人工標註及專家設計所有可能，所以我們需要一套新方法來克服這些問題。

3. 字元階層文脈訊息擷取方法與語音合成系統

在前一章中我們已經知道傳統架構擷取語言特徵會遇到的問題點，所以我們希望能使用大量無標註資料庫與非監督式學習(unsupervised learning)來訓練深層類神經網路，達到無需人工介入讓機器自動學習文本取代 parser。在這基礎下我們使用以字元為層級的訓練方式來達到我們零標註的目的。在字元屬性求取上我們採用 word2vec 來對文本進行字元的語意與文法角色分類，字元之間的時序順序關係我們利用遞迴式神經網路，擷取字元在句子中的狀態來當作時序關係，最後得到字元屬性與字元時序狀態的文脈訊息。

3.1 新系統架構

為了達成以字元層級為輸入的目的，我們設計另一個新的架構，將 parser 部分做替換，主要是對字元語意、文法角色等資訊利用 word2vec 來產生。字元時間前後資訊則使用 RNNLM 擷取，最後形成一個新的文脈訊息全部是由機器自行學習未標註語料所產生出來的，這樣能達成我們想使用字元層級的目的也能避開傳統 parser 的問題。以下為新系統架構以及新方法的詳細說明。

首先，圖 2 為我們的新語音合成系統架構，將語料經過正規化，文本轉拼音方面，由於我們的中文拼音字典並沒有歸納出哪些字應該何時讀破音字的資訊，在此選用第一組拼音來對應，再將拼音轉音素。我們更動的地方是將 parser 拿掉，以 word2vec 與 RNNLM 做替換，主要是用 word2vec 與 RNNLM 求取字元語意、文法角色資訊與字元時間前後資訊當作文脈訊息，以合成出新的語音。

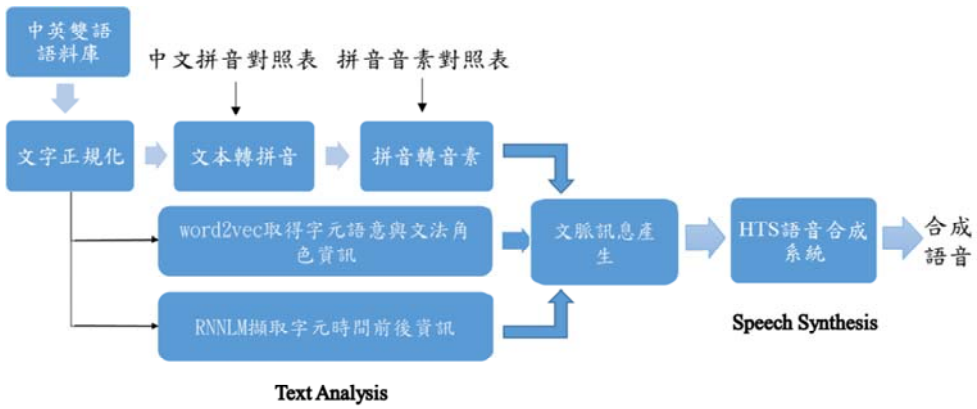


圖 2. 新語音合成系統架構圖
[Figure 2. The proposed character-based TTS apporach]

3.2 字元語意與文法屬性之文脈訊息擷取

文字探勘以及 NLP 在數據分析世界中一直是非常重要的一部分。其中 word2vec 被頻繁的討論以及使用，因為使用它能夠將輸入的詞轉到向量空間上並進行演算，分析後可以發現在向量空間中，相聚在一起的詞向量轉換回文字後會是相同屬性的詞彙。也就是說它有能將字詞語意或文法角色做分類的能力，而且它無需給定標註過的資料庫就能對語料直接進行訓練，這非常符合我們想要避開人工標註的初衷。

Word2vec 是 Google 公司在 2013 年開放的一款用於訓練詞向量的軟體工具。它根據給定的語料庫，通過優化後的訓練模型快速有效的將一個詞語表達成向量形式，其核心架構包括 CBOW(Continuous Bag-Of-Words Model)和 Skip-gram。圖 3 為 word2vec 的核心架構圖。

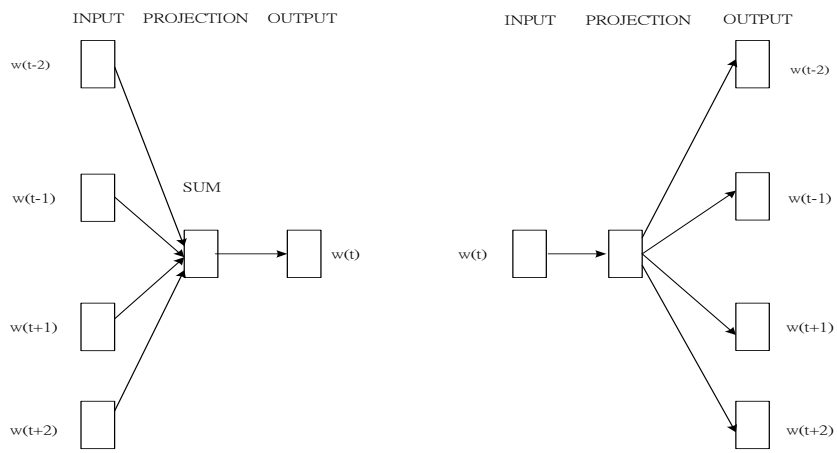


圖 3. word2vec 架構圖
[Figure 3. The continuous bag of word (CBOW) and Skip-gram word2vec models]

CBOW 模型可以理解為以上下文決定當前詞出現的機率，在 CBOW 模型中，上下文所有的詞對當前詞出現機率的影響的權重是一樣的，因此叫 CBOW 模型。如在袋子中取詞，取出數量足夠的詞就可以了，至於取出的先後順序是無關緊要的。Skip-gram 模型與 CBOW 模型正好相反，Skip-gram 模型使用當前的詞向量來預測該詞之前和之後各 K 個詞的機率，距離當前詞越遠的詞和當前詞的相關詞的相關性越小，所以可以給予這些詞以較小的權重。詳細內容可以參閱(Mikolov, Sutskever, Chen, Corrado & Dean, 2013)詞語分佈表示與組合性，內文中提到 word2vec 提供一個更準確以及非常簡單的訓練方式，能夠大大的改善學習單詞和短語的相關性。

因此我們利用 word2vec 的特性，將所有的字轉到向量空間，並進行分類，將性質相近的字群聚在一起。表 1 為我們在一前置實驗中，利用 word2vec 分析 Chinese Gigaword 語料庫的結果，我們可以看見 class1 是魚部首被分為一類，裡面都為魚類生物居多，class2 把木部首的植物類分在一起，class3 則是把親戚稱謂歸類在一起，class5 是價值衡量單位，class7 是將語助詞分在一起，而 class4 與 class6 分別是英語和日語，將不同語言各自為一類，由此可知將字元轉換到向量空間處理分析，可以有效的將字元語意、字元文法角色等屬性擷取出來。

表 1. 藉由 character embedding 產生的字分類
[Table 1. Experimental results of character embedding-based character clustering]

Class	Characters
1	顱、鯛、鰭、鰐、哼、塭、犮、蛉、蝳、蟾、躠、魟、魡、魣、魨、魴、鮎、鮜、鮪、鮫、鮭、鯁、鯇、鯉、鯊、鯏、鯡、鯧、鯰、鯷、鰆、鰕、鰱、鰷、鰹、鰾、鱅、鱒、鱓、鱘、鱮、鱸、鸅
2	杏、杞、杷、枇、枴、枸、柑、柚、柿、株
3	丈、么、兄、兒、叔、夫、女、妻、姨、姪、娶、婦、婿、媳、嫁、嫂、子、孺、弟、戚、歲、母、父、璞、甥、男、舅、親、齡
4	Peanuts、Pentavision、Peranakan、percussion、Perfectv、Persona、phantasy、Phantom、 Phoenix、 Pick
5	億、元、兆、千、圜、幣、新、百、美、胘、萬、鈔、銖、鎊、餘
6	ムック、ひろみ、ベィン、トヨタ、あやか、タケシ、ヘリオス、ブルース、タイトロープ、チュートリアル
7	哎、哼、唉、唔、唬、啥、喂、喲、嗐、嗯、嘻、嘿、噢、嬤、欸、砥

也因為 word2vec 它所訓練出來的字向量空間能夠表達字的屬性，並且它能將訓練出來的字向量進行排列，讓屬性接近的字分類在一起，所以使用它來代替傳統 parser 可能是行得通的。

3.3 字元時序狀態之文脈訊息擷取

在我們提出的字元概念下，我們希望能從文本中獲得字元在當前語句中的狀態，讓我們學習文章的脈絡，可以從這句預測下句可能為何。為了完成這功能，遞迴神經網路(RNN)可能是個不錯的選擇，例如(Mikolov & Zweig, 2012)基於 RNNLM 上下文相關性的研究也指出使用遞迴類神經網路模型進行訓練能從隱藏層中連續的輸出向量獲得字詞在句子中的狀態。

本文是以 Mikolov 改良的 RNNLM 來進行，遞迴式類神經網路包含輸入層（input layer）、隱藏層(hidden layer)、輸出層(output layer)和類別層(class layer)。而 U、V、W 和 C 為各層的權重。w(t)為輸入，t 依時間排序為為 1 到 N，也是 RNN 的權重，s(t)為隱藏層輸出也就是神經元(neurons)的值也是它的 state，y(t)為輸出須與輸入同維度。而 c(t)為類別層，Mikolov 提出輸出層分解可以降低語言模型中的運算複雜度，使訓練效率提高。圖 4 為 Mikolov 改良的 RNNLM 架構。

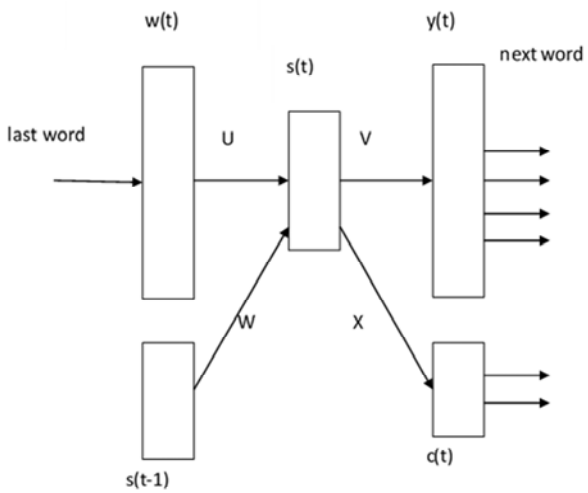


圖 4. Mikolov 改良的 RNNLM 架構
[Figure 4. The block diagram of the RNNLM model.]

遞迴神經網路最大的優勢在於，可以真正充分的利用所有上文訊息來預測下一個詞，不像其它神經網路只能一次看 n 個字，只能從前 n 個字來預測下一個詞，簡單來說 RNN 就是一個有隱藏層的自我相連網路，隱藏層同時接收來自 t 時刻的輸入和 t-1 時刻的隱藏層輸出做為輸入，這使得 RNN 具有短期記憶能力，能夠學習到較長時間的文章脈絡，建立起語言模型。

表 2 為我們在一前置實驗中，利用 Chinese Gigaword 語料庫與 RNNLM 學習字元時序關係的結果，我們可以發現 RNNLM 模型產生出來的句子不管在語意，時間順序，或是流暢度而言都蠻貼近一般新聞的文字。

表 2. 使用 RNNLM 產生之文句
[Table 2. Experimental results on RNNLM-based sentences generation.]

時序方向	範例
Forward	在民主黨的表現，他們1致認為，1切不可能會對我們的壓力．
	警方今天在高雄縣警察局長劉松藩訪問時表示，這次的民進黨立法院黨團將在明天召開記者會，會中表示，民進黨決定將全力支持．
Backward	．遇待國惠最的國美開離定決已，判談的國美與國合聯在國美兼理總副國美，示表統總李
	收作點3818以，點4十7百1漲大數指價股權加量行發，點95.591以貨期月期10.959金基太亞銷沖天今元美兌幣台新

因此應用 RNNLM 的特性來訓練文本，可以學習前面以及當前看過的句子來猜測下一的字或是下一句會是甚麼，也因為遞迴神經網路能夠這樣分析句子前後時間關係，所以能夠知道整個語句的脈絡，而這剛好與傳統文本分析的時間順序資訊相似，因此我們將其用在我們的新架構上看看能否真的有用。

4. 實驗結果與分析

本實驗中，目的是為了使用新語言特徵取代現有 parser 的斷詞、POS 資訊，然後產生新的文脈訊息，代入語音合成系統合成語音，並與傳統方法 parser 做比較；新系統與舊系統的差別只有文脈訊息的不同，舊系統使用 parser 做文本的分析，而新系統使用字元層級的 word2vec 與 RNNLM 來分析文本，為求其公平性，新舊系統皆使用中英夾雜語料，訓練語句長度為段落訓練。而我們要比較的是新舊系統合成音檔之偏好，選用的合成語料是訓練語料中沒有被訓練到的部分，最後新舊系統比較的音檔皆為同一句話以示公平，而測試者不知道哪個音檔為新系統所合成避免分數灌水。

4.1 實驗設定

4.1.1 語料

我們使用的訓練語料、合成語料皆為我們與台灣數位有聲書協會合作錄製的”NTUT Audiobook Corpus Vol.2”，在此語料庫中我們請專業錄音員為我們錄製男聲語料，所有音檔皆在專業錄音室錄音，錄製時是以段落為單位唸完，保留語句之間的連結性；合成語料則從其中各別抽取中文 280 句及中英夾雜 160 句來做合成，抽出的句子皆不在訓練語料當中。表 3 為訓練語料資料表。

表 3. 訓練語料資料表
[Table 3. Statistics of the speech corpus for speech synthesis experiments.]

	中文語料	中英夾雜語料-CE	純英文
文本內容出處	生命科學大師：遺傳學之父孟德爾的故事(張文亮著)	線上文本 (工研院提供)	CMU
訓練語料總句數	約 4800 句	約 3500 句	約 990 句
訓練語料每句詞數	20-35 詞	10-30 詞	5-15 單字
訓練語料時間長度	約 172 分鐘	約 201 分鐘	約 79 分鐘

4.1.2 文脈訊息求取方法與設定

新方法與舊方法中只有前級文本分析不同，後級的語音合成系統(HMM-based Speech Synthesis System，HTS)的設定，兩者完全相同；舊系統文脈訊息依然採用 parser 來分析文本，新系統的文脈訊息則是去除 parser 產生的資訊，改使用字元層級的 word2vec、RNNLM 來分析，我們用 word2vec 將字元歸類成 64 類，RNNLM 隱藏層設 256 維，並為隱藏層中每個 neurons 的狀態設定門檻(threshold)把輸出量化成 0 或 1。在問題集中我們去掉有使用斷詞和 POS 的項目，然後添加使用 word2vec 和 RNNLM 產生的新項目，然後建立新的決策樹。表 4 為舊系統與新架構所使用的語言特徵，其中舊系統中紅色斜體字部分在我們新架構中將被廢除，新加入的藍色斜體字部分為 word2vec 與 RNNLM 的輸出參數。

表 4. 傳統 TTS 系統和新架構的文脈訊息差異
[Table 4. Comparison of linguistic features used by the traditional and the proposed speech synthesis approaches.]

	傳統語言特徵	新架構語言特徵
音素(PHONE)	音素在音節中的位置	音素在音節中的位置
音節(SYLLABLE)	*音素數量*，*在詞中的位置*	X
詞(WORD)	*音節數量*，*在短語中的位置*	X
短語(CLAUSE)	*詞數量*，在句子中的位置	在句子中的位置
句子(UTTERANCE)	短語數量，在段落中的位置	短語數量，在段落中的位置
段落(PARAGRAPH)	句子的數量	句子的數量
WORD2VEC 類別	X	*字元是屬於哪一類*
RNNLM	X	*字元間前後時間順序*

4.1.3 語音合成設定

本研究的中英夾雜語音合成系統使用”NTUT Audiobook Corpus Vol.2”語料和 HTS 合成；首先我們中文和英文聲音編碼統一使用 X-SAMPA 編碼為標準。在訓練語音合成模型時，所有錄音皆為 48KHz，聲學特性我們取 34 維梅爾倒頻譜係數(MFCCs)，音調(pitch)輪廓每 5 毫秒至 25 毫秒為一音框(frame)長度，最後每個音素(phone)我們使用 5 個狀態(state)的 HMMs 來訓練。

4.2 評估方法

系統偏好的評估方式是傳統使用 parser 的語音合成系統與我們提出使用字元階層提取特徵的新語音合成系統來做比較。我們以聲音的相似度、自然度和可理解度的評估方式，將測試音檔給 10 位以國語為母語的人士進行評分，新舊系統偏好度測試為 2 選 1 方式，為標準 A/B/X 測試，不存在兩者皆好；而新舊系統評分採平均主觀值分數(mean opinion score,MOS)來評估，其評分方式為 1~5 分。表 5 為測試音檔設定。

表 5. 測試音檔設定
[Table 5. Statistics of the synthesized speech database for all evaluation experiments]

	NTUT Audiobook Corpus Vol.2	
測試類型	中文	中英夾雜
總句數	60 句	40 句
音檔數	10 個	20 個
每句字數	10-20 字	10-20 字

4.3 實驗結果

圖 5 與圖 6 分別為新舊系統在純中文與中英夾雜的相似度、自然度與可理解度偏好的比較，表 6 與表 7 則分別為新舊系統在純中文與中英夾雜的主觀 MOS 分數比較。評比偏好部分，從測試結果中我們可以發現，在圖 5 與圖 6 中不管是在相似度、自然度與可理解度來看，大部分測試者偏好新系統。而由表 6 來看新系統的聲音比傳統架構所合成出來的聲音稍微自然與相似原語者的聲音，但在表 7 中英夾雜測試分數看起來新舊系統分數差別不大，所以只能說兩者大概相當。不過概觀來看可以發現使用字元層級(character-level)的 word2vec 與 RNNLM 來取代傳統 parser 進行文脈訊息擷取，能合成出相當貼近人聲的聲音。

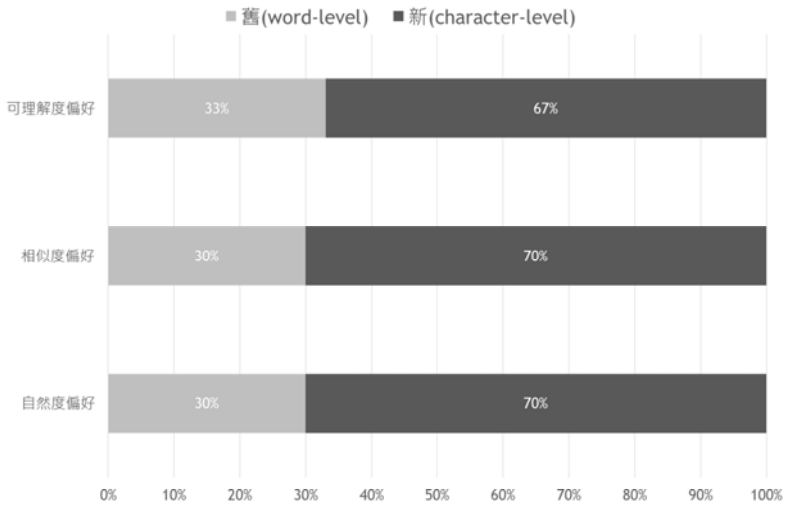


圖 5. 新舊系統純中文偏好比較

[Figure 5. Experimental results of the A/B/X preference test on pure Chinese sentence synthesis]

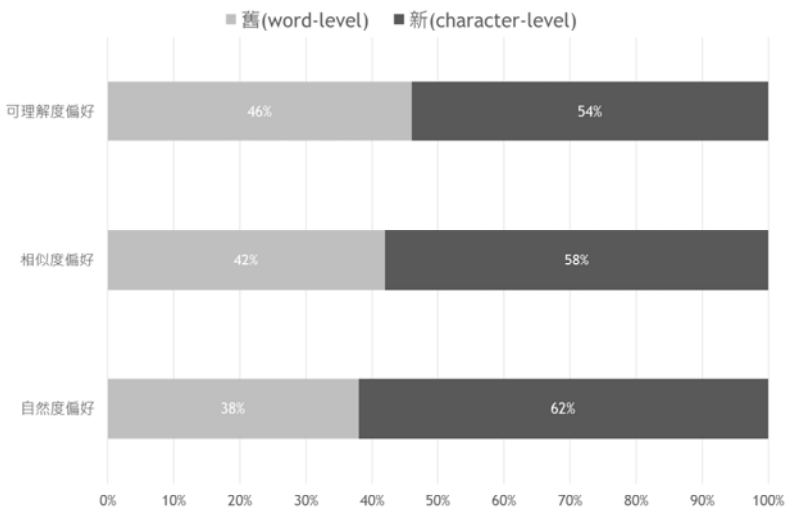


圖 6. 新舊系統中英夾雜偏好比較

[Figure 6. Experimental results of the A/B/X preference test on mixed Chinese-English sentence synthesis]

表 6. 新舊系統純中文聲音 MOS 主觀分數比較
[Table 6. Comparison of the MOS scores of the conventional and the proposed approaches on pure Chinese sentence synthesis]

純中文	傳統語音合成系統	新語音合成系統
自然度評分	3.09	3.44
相似度評分	3.25	3.38
可理解度評分	4.27	4.27

表 7. 新舊系統中英夾雜聲音 MOS 主觀分數比較
[Table 7. Comparison of the MOS scores of the conventional and the proposed approaches on mixed Chinese-English sentence synthesis]

中英夾雜	傳統語音合成系統	新語音合成系統
自然度評分	3.26	3.22
相似度評分	3.26	3.36
可理解度評分	3.24	3.24

4.4 實驗討論

在本次實驗中新系統大概獲得 3 點多分，在市面上的 TTS 合成系統多為 4 分以上，雖比不上大公司的合成系統，不過這分數還算是合理，證明使用字元層級的文脈訊息在語音合成上是可行的。對此我們猜測新架構會更好的原因，是因為傳統 parser、POS 是人所設計的，但是真正口語上我們不一定會這樣唸，而我們新系統則是不靠人工設計，讓機器自己從語料當中學習句子關係，因此可能合成出來的聲音會比較流暢。不過在中英夾雜測試下兩者分數差距不大，可能是 RNNLM 在中英夾雜的英文句子部分少，所以無法有效得知字元在句子狀態，不過整體看來新架構確實有比較好的成績。但是本研究只是初步實驗新架構方法以實驗數據來證明真的會比傳統好，真正比較好的詳細原因需要未來繼續深入探討。

5. 結論

在本研究中，我們將一般語音合成中的文本分析做替換，將以前以詞為單位求取文脈訊息的方式，替換成以字元為處理單位。用 word2vec 求取字元的語意屬性與文法角色分類和利用 RNNLM 猜測字元在句子中的狀態，以這種方式能夠避開斷詞、POS 需要大量人工標註資料庫的缺點；而新舊系統在各項評比中都是新系統合成出來的聲音較為優良，所以我們提出的字元層級的文脈訊息擷取方法確實能達到相當甚至超越傳統方式的成績。

致謝

本研究感謝教育部『大學以社教機構為基地之數位人文計畫』（A36 號）與科技部專題計畫（MOST 104-2221-E-027-079, 105-2221-E-027-119 and 103-2218-E-027-006-MY3）支持。

參考文獻 References

Brill, B. (1992). A SIMPLE RULE-BASED PART OF SPEECH TAGGER . In *ANLC '92 Proceedings of the third conference on Applied natural language processing*, 152-155.

Ding, C., Xie, L., Yan, J., Zhang, W. & Liu, Y. (2015). Automatic prosody prediction for Chinese speech synthesis using BLSTM-RNN and embedding features. In *proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 98-102.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. (in press). LSTM: A Search Space Odyssey. *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS*, Retrieved from https://arxiv.org/abs/1503.04069

Licstar, (2013 年 7 月 29 日)。Deep Learning in NLP (一)詞向量和語言模型。【部落格文字資料】。取自 http://licstar.net/archives/328。[Licstar. (2013, July 29). Deep Learning in NLP (1) Word embedding and Language model [Web blog message]. Retrieved from http://licstar.net/archives/328]

Mikolov, T. & Zweig, G. (2012). Context dependent recurrent neural network language model. In *proceedings of 2012 IEEE Spoken Language Technology Workshop (SLT)*. doi: 10.1109/SLT.2012.6424228

Mikolov, T., Sutskever, I., Chen,K., Corrado, G. & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In *proceedings of Advances in neural information processing systems 26 (NIPS 2013)*, 3111-3119.

The Stanford Natural Language Processing Group. (2015). Stanford-Parser Version3.6.0 Release in 2015/12/09: http://nlp.stanford.edu/software/lex-parser.shtml

Zheng, X., Chen, H. & Xu, T. (2013). Deep Learning for Chinese Word Segmentation and POS Tagging. In Proceedings s of the 2013 Conference on EMNLP, 647-657.

唐大任 (2002)。*中文斷詞之研究*（碩士論文）。取自 http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22NT900435069%22.&searchmode=basic [Tang, D.-R. (2002). *A Study of Chinese Parser*(Master's thesis). Retrieved from http://140.113.39.130/cgi-bin/gs32/tugsweb.cgi?o=dnctucdr&s=id=%22NT900435069%22.&searchmode=basic]

黃昭銘 (2010)。*改善條件隨機域模型於中文斷詞*（碩士論文）。取自 http://etd.lib.nsysu.edu.tw/ETD-db/ETD-search-c/view_etd?URN=etd-0203110-093833 [Huang, J.-m. (2010). *An Enhanced Conditional Random Field Model for Chinese Word*

Segmentation (Master's thesis). Retrieved from http://etd.lib.nsysu.edu.tw/ETD-db/ETD-search-c/view_etd?URN=etd-0203110-093833]

Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 85-104

融合多任務學習類神經網路聲學模型訓練於會議語音辨識之研究

Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition

楊明翰*、許曜麒*、洪孝宗*、陳映文*、陳冠宇+、陳柏琳*

Ming-Han Yang, Yao-Chi Hsu, Hsiao-Tsung Hung, Ying-Wen Chen, Kuan-Yu Chen, and Berlin Chen

摘要

本論文旨在研究如何融合多任務學習(Multi-Task Learning, MTL)技術於聲學模型之參數估測，藉以改善會議語音辨識(Meeting Speech Recognition)之準確性。我們的貢獻主要有兩點：1)我們進行了實證研究以充分利用各種輔助任務來加強多任務學習在會議語音辨識的表現。此外，我們還研究多任務與不同聲學模型像是深層類神經網路(Deep Neural Networks, DNN)聲學模型及摺積神經網路(Convolutional Neural Networks, CNN)結合的協同效應，期望增加聲學模型建模之一般化能力(Generalization Capability)；2)由於訓練多任務聲學模型的過程中，調整不同輔助任務之貢獻(權重)的方式並不是最佳的，因此我們提出了重新調適法，以減輕這個問題。我們基於在台灣所收錄的中文會議語料庫

*國立台灣師範大學資訊工程學系
Department of Computer Science and Information Engineering, National Taiwan Normal University
E-mail: {mh_yang, ychsu, alexhung, cliffchen, berlin}@ntnu.edu.tw
+中央研究院資訊科學所
Institute of Information science, Academia Sinica
E-mail: kychen@iis.sinica.edu.tw

(Mandarin Meeting Recording Corpus, MMRC)建立了一系列的實驗。與數種現有的基礎實驗相比，實驗結果揭示了我們所提出的方法之有效性。

關鍵詞：多任務學習，深層學習，類神經網路，會議語音辨識。

Abstract

This paper sets out to explore the use of multi-task learning (MTL) techniques for more accurate estimation of the parameters involved in neural network based acoustic models, so as to improve the accuracy of meeting speech recognition. Our main contributions are two-fold. First, we conduct an empirical study to leverage various auxiliary tasks to enhance the performance of multi-task learning on meeting speech recognition. Furthermore, we also study the synergy effect of combing multi-task learning with disparate acoustic models, such as deep neural network (DNN) and convolutional neural network (CNN) based acoustic models, with the expectation to increase the generalization ability of acoustic modeling. Second, since the way to modulate the contribution (weights) of different auxiliary tasks during acoustic model training is far from optimal and actually a matter of heuristic judgment, we thus propose a simple model adaptation method to alleviate such a problem. A series of experiments have been carried out on the Mandarin meeting recording (MMRC) corpora, which seem to reveal the effectiveness of our proposed methods in relation to several existing baselines.

Keywords: Multi-Task Learning, Deep Learning, Neural Network, Meeting Speech Recognition.

1. 緒論

口語對話是人與人之間最自然的溝通方式，可以預期它也是人們與人工智慧助理等機器間最重要的互動方式。近六十年來，自動語音辨識的研究活動十分活躍，並且已取得了巨大的成功。在研究初期，語音辨識器只能在安靜的環境中識別一個單獨的詞彙。1980年代，以高斯混合模型-隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)做為聲學模型使得語音辨識有能力進行大詞彙量連續語音識別。由於GMM-HMM 的架構易於訓練模型和進行聲學解碼，因此近二十年來 GMM-HMM 是自動語音辨識系統的主流聲學模型，聲學模型的研究主要集中在以更好的模型結構與訓練演算法改良 GMM-HMM。顯著的成果包含狀態聯繫(State Tying) (Young & Woodland, 1993)、鑑別式訓練(Discriminative Training) (Povey, 2004)與最大相似度線性轉換(Maximum Likelihood Linear Transformation, MLLT) (Gales, 1998)。在 GMM-HMM 模型主導語音界的時期內，研究學者們也探索了許多不同的聲學模型方法，然而卻沒有一種方法可以像 GMM-HMM 滿足建置成本和辨識效能的平衡。過去的五年內我們看見了深層學習架構和

技術在電腦視覺、語言及語言學習領域的巨大成功。深層類神經網路與其變體最終取代了 GMM，混合深層類神經網路-隱藏式馬可夫模型(Hybrid Deep Neural Networks-Hidden Markov Model, DNN-HMM)已成為大多數自動語音辨識系統的聲學模型。DNN 的竄起可歸功於以下六種因素：1)深層學習架構及演算法；2)通用計算圖形處理器(General Purpose Graphical Processing Units, GPGPU)的發展；3)數千小時的已轉寫語音訓練資料及更多的未標記資料；4)行動式網際網路和雲端計算；5)從生活到工作環境都廣泛地出現語音辨識技術需求。

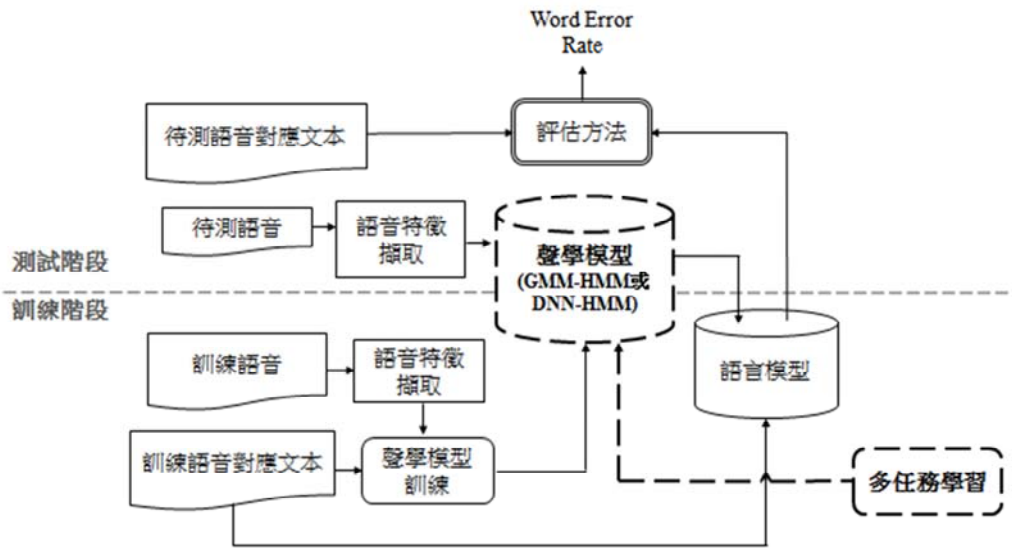


圖 1. 語音辨識流程與本論文嘗試改良之處(虛線標記部分)
[Figure 1. Illustration of the Enhanced Speech Recognition System]

雖然自動語音辨識技術已經是一項成熟的技術，但是在實際應用上仍有許多問題需要被解決。例如使用智慧型手機錄音時往往離手機麥克風較遠，錄音品質容易受環境影響。此外，現今語音辨識領域也面臨著海量詞彙、自由不受限的任務、吵雜的遠距離語音、自發性的口語及語言混雜情景的挑戰(Yu & Deng, 2014)。而會議語音辨識[1]正涵蓋了上述大部分的困境與挑戰，是一個相當困難的語音辨識任務。本論文將問題視為訓練與測試環境不匹配；除了語音和文字的使用不同，也包含了多語言的混用。簡而言之，上述問題是在考驗語音辨識之一般化能力。為克服此問題，我們首先比較各種輔助任務來加強多任務學習在會議語音辨識的表現。其次，藉由多樣異質模型的整合也是常見的方法，本論文透過實驗來驗證異質模型結合的效果。最後，我們提出了重新調適法以減輕傳統調整輔助任務權重方法的缺陷。所有結果是驗證於一在台灣所收錄的會議語料庫，內容多為中文和中英文混用語句。圖 1 所示為語音辨識流程與本論文嘗試改進的部分(以

[1]會議語音的內容如表 1 所示。不難發現語料中除了中英文交互使用外，也有自然對話所包含的語助詞、口頭禪、口吃與贅詞。專有名詞與時髦新詞也可能出現在語料中。

虛線標記)。

表 1. 會議語音範例
[Table 1. Meeting transcription examples extracted from MMRC]

語句編號	語句內容
U0006-002932	教材的 scenario 可能 target audience 如果是這群人可能是應該怎麼怎麼讓他 aggregate 出一套理論讓他再發揮更大
U0000-001623	呃考法律的法學院的碩士吼 然後呃就醫學院的碩士還有那個
U0000-001624	那個牙科的碩士吼那個那個補習班補得很兇啊那生意都很好
U0002-000118	呃 cover 到然後也講到這個 我們上次談的這個 megatrend 吼
U0002-000122	就是 knowledge formulation 呃 diversification 那個 is great

本論文後續章節安排如下：第二小節將簡介類神經網路的相關文獻探討，第三小節介紹多任務學習的發展演進以及我們想要探討的輔助任務，第四小節介紹我們提出的重新調整法，第五小節則解析基礎實驗及多任務學習的實驗結果，最後在第六小節進行結論與探討未來可能的研究方向。

2. 類神經網路之相關文獻探討

在機器學習的領域中，類神經網路的起源可以追溯到 1943 年的數學家 McCulloch，他設計了一套數學方法模擬神經元運作的模式，是開啟了類神經網路研究大門的先驅。接著在 1957 年，Rosenblatt 是第一個將類神經網路概念付諸實行的學者，提出了感知器(Perceptron)模型。1975 年，Werbos 提出倒傳導演算法(Backpropagation Algorithm) (Werbos, 1974)改善類神經網路參數更新的方式。終於在 1988 年，Rumelhart 等人發明了多層感知器(Multilayer Perceptron, MLP) (Rumelhart, Hinton & Ronald, 1988)，因為多層感知器適用於更多元的問題，使得類神經網路的研究熱潮再度熱絡起來。

在語音辨識領域中，從 1992 年起，就陸續有許多將類神經網路與隱藏式馬可夫模型結合(Hidden Markov Model, HMM)的研究。例如在 1998 年，Cook 等人(Cook *et al*., 1999)使用廣播新聞的語料，訓練多個遞迴神經網路(Recurrent Neural Network, RNN)與 MLP 的聲學模型，並透過 ROVER (Recognition Output Voting Error Reduction, ROVER) (Fiscus, 1997)的方法統整這些模型的辨識結果。2000 年時，學者們指出類神經網路也是合適的特徵擷取工具，例如 Bottleneck 特徵或 Tandem 特徵(Hermansky, Ellis & Sharma, 2000)。

早期的類神經網路研究受限於硬體計算資源的不足，且不易進行平行化處理，使得相關研究沒有顯著地突破。直到 2006 年開始，學者們在訓練演算法與架構上提出了一系列改進(Hinton, Osindero & Teh, 2006; Poultney, Chopra & Cun, 2006, Bengio, Lamblin, Popovici & Larochelle, 2007)，而後幾年的 GPGPU 運算設備發展迅速，使得深層類神經

網路模型計算成本問題大幅降低，也讓學者們願意投入此研究。現今主流的聲學模型設計即是利用深層類神經網路取代高斯混合模型(Hinton *et al*., 2012)。除了深層類神經網路之外，學者們也嘗試引入類神經網路變體，例如 CNN (Abdel-Hamid *et al*., 2014)與 RNN (Graves, Mohamed & Hinton, 2013)。這些新穎的深層模型在語音辨識領域也有顯著的成功(Sercu, Puhrsch, Kingsbury & LeCun, 2016)。

3. 多任務學習探討

多任務學習(Caruana, 1997)或者學會學習(Learning To Learn) (Thrun & Pratt, 1988)是一種機器學習的技術，其目的是希望藉由共同學習數個相關的輔助任務，以提升主任務的一般化能力。多任務學習大約在二十年前開始成為一項熱門的研究，有許多論文以理論的角度分析多任務學習的行為與一般化能力界限(Generalization Bound)，並提出了一系列有關多任務學習的統計理論，也進一步得知，透過相關輔助任務所產生的參數假設空間(Parameter Hypothesis Space)作為基礎，能夠提供更好的初始參數假設空間給其它新的輔助任務。近年來，多任務學習的研究開始探索自動地學習任務之間的關係，Zhang 等人將學習任務之間的關係視為求解凸函數(Convex Function)的過程(Zhang & Yeung, 2014)；假設數個線性回歸任務的參數具有相同的矩陣常態分佈事前機率(Matrix-Variate Normal Distribution Prior)，並由共變異數矩陣定義任務與任務之間的關係，模型訓練時能間接學習如何替正例任務關係(Positive Task Correlation)與負例任務關係(Negative Task Correlation) 的相關性建立模型。後來 Zhang 等人更融入了多任務特徵選取(Multi-Task Feature Selection) 與相關性學習(Relationship Learning)對高維度的輸入資料進行處理。據學者的研究證明，假設多個任務之間彼此相關，通過一起學習的方式來共享內在的表示資訊，就能夠達到知識轉移的效果。其實驗結果也證實此方法在模型遇到沒看過的資料(Unseen Data)時也能有不錯的成效。

3.1 語音辨識中的多任務學習

多任務學習結合深層類神經網路的架構(Multi-Task Deep Neural Network, MTL-DNN)如圖 2 所示。語音辨識領域中也有許多研究先進嘗試融合語音領域及多任務學習技術。例如 Parveen 等人(Parveen & Green, 2003)探討了 11 種不同的分類任務與語音增強任務的影響，例如語者的性別或情緒等。實驗結果發現在分類任務中，多任務訓練優於單任務訓練。Chen 等人(Chen & Mak, 2015)則是透過多任務學習的特性，使得資源豐富(Resource-Rich)的語言能夠在模型訓練的過程中，輔助資源貧乏(Resource-Poor)語言，提升它的辨識效果。Seltzer 等人的研究(Seltzer & Droppo, 2013)則是探討了以目前音框的音素標記、鄰近音素狀態標記(State Contexts)及鄰近音素的音素標記 (Phone Contexts) 做為輔助任務訓練聲學模型的效果，其文獻指出在英語音素語料(TIMIT) (Garofolo, Lamel, Fisher, Fiscus & Pallett, 1993)的語音辨識任務中有顯著地進步。然而，Seltzer 等人的研究只停留在單連音素(Monophone)，沒有使用到三連音素(Triphone)的資訊，也沒有探討若使用三連音素狀態標記取代單連音素狀態標記做為輔助任務的成效，這正是本論文想探

究的其中一項問題。另外，學者們嘗試將多種語言的語料混合在一起，共同訓練一個跨語言的聲學模型(Ghoshal,Swietojanski & Renals, 2013; Huang, Li, Yu, Deng & Gong, 2013)，證實這種做法確實能夠提升準確率。多語言的資料主要使用於訓練階段，所有語言的訓練資料皆會調整底層共享的隱藏層。每種語言有各自對應的輸出層，各個輸入語言的聲學特徵除了調整底層的隱藏層外，也會更新其所對應語言的輸出層，其它語言的輸出層將不會被 更新。如此一來，隱藏層就可被視為一層一層的特徵萃取器。總而言之，多任務學習允許學習多個任務時，以建設性(Constructive)或破壞性(Destructive)錯誤訊號梯度更新隱藏層。在多任務學習的框架下，模型將會同時學習：(1)主任務；(2)一個或數個相關的輔助任務；(3)任務間共享的隱藏層。

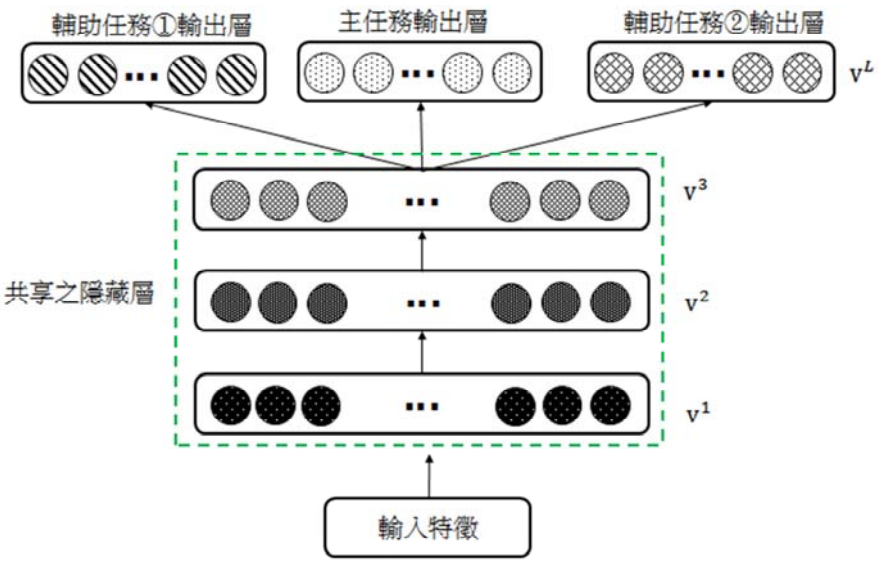


圖 2. 多任務深層類神經網路示意圖
[Figure 2. Illustration of the MTL-DNN system]

3.2 輔助任務探討

有學者的研究指出，多任務學習並不保證效能會提升，訓練的演算法及任務是否相關同樣也是重要的關鍵(Caruana, 1997)。有鑒於此，本論文從兩大類研究面向，篩選出 10 種輔助任務進行探討，如圖 3 所示。其中一個面向是語言與音韻學資訊，此類型的資訊主要分為 3 類：音框對應狀態標記、音框對應音素標記與多語言及跨語言資訊。另一個面向則是自動語音辨識回饋，我們採用的是模型壓縮技術(Model Compression) (Buciluǎ, Caruana & Niculescu-Mizil, 2006; Hinton, Vinyals & Dean, 2015)，從已訓練完成的強健模型中，將知識在訓練過程中轉移到待訓練模型。接下來將詳細介紹我們使用的輔助任務：

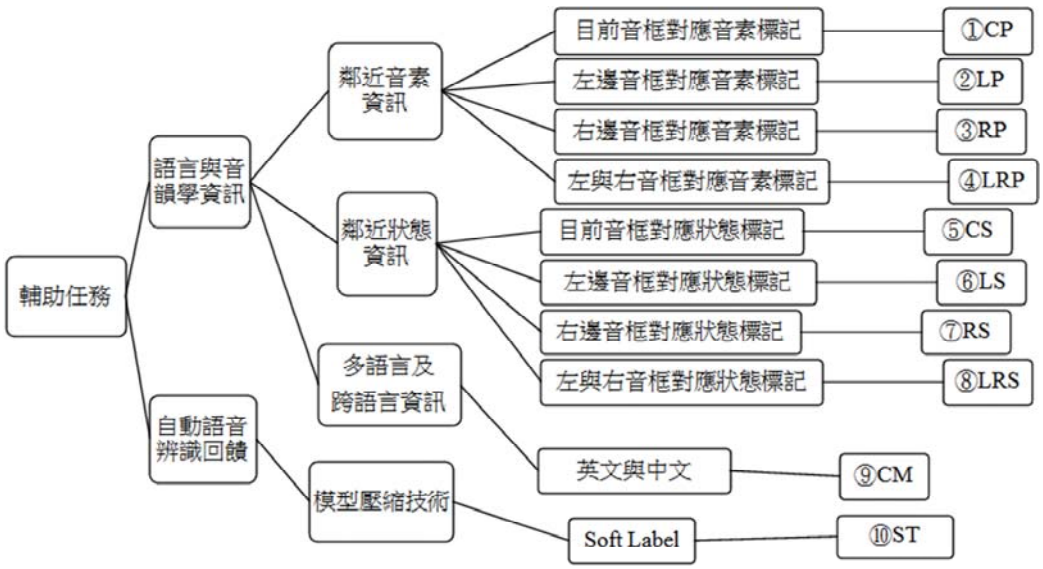


圖 3. 本論文使用的輔助任務一覽
[Figure 3. Auxiliry tasks used in this paper]

3.2.1 音框對應狀態標記：

音框對應狀態標記是以預測目前音框的前一個或後一個音框的 HMM 狀態標記做為輔助任務。由於以往類神經網路聲學模型的訓練方式，通常是以預測目前音框的 HMM 狀態標記為目標，這使得主任務的訓練目標並沒有鄰近音框的狀態資訊。這類輔助任務則是期望能夠提供模型訓練的過程中能夠加入這些額外資訊。以預測鄰近音框 t+1 的狀態標記為例，假設目前音框表示為$\mathbf{o}_t$，下一個時間點的狀態標記表示為s_{t+1}，則右邊音框狀態標記之目標函數表示為：

$$\mathcal{F}_{\text{Right-State}} = \sum_t \ln P_{\text{RS}}(s_{t+1}|\mathbf{o}_t) \tag{1}$$

3.2.2 音框對應音素標記：

這類輔助任務初始的設計理念跟 Triphone 模型類似，不同的地方在於以往產生 Triphone 模型所使用的 Decision Tree State Tying 技術(Young & Woodland, 1993)是靜態的二元表示，且與 DNN 的隱藏層無關。而融入多任務學習能夠在訓練的過程中，提供鄰近音素的資訊，達到動態更新，自動影響隱藏層的效果。我們除了遵循原有的設計理念外，也進一步延伸出不同的輔助任務。以預測目前音框所屬音素為例，假設時間 t 的音框所對應的音素標記表示為q_t，音框對應語音特徵向量表示為$\mathbf{o}_t$，則此輔助任務的目標函數可以表示為：

$$\mathcal{F}_{\text{Current-Phone}} = \sum_t \ln P_{\text{CP}}(q_t|\mathbf{o}_t) \tag{2}$$

3.2.3 多語言及跨語言資訊：

很自然地我們可以猜想，不同語言之間應該具有共同的發音模式。舉例來說，許多的子音和母音是跨語言共享的，運用語言之間共享的特性，來建立統計模型更優於僅使用單一語言建立的模型，這項優勢已經被許多研究報告證明。近年來，這類的研究透過深層類神經網路做為多語言及跨語言資訊的傳遞媒介也越來越火熱，其主要思路是認為低層靠近特徵的隱藏層，傾向於學習語言獨立(Language-Independent)的資訊；而較高層的隱藏層學習較多語言相關(Language-Dependent)的知識(Schultz & Waibel, 2001)。Swietojanski 等人提出以非監督式(Unsupervised)的方法，以多語言的資料目標語言的類神經網路模型進行初始化(Swietojanski, Ghoshal & Renals, 2012)。多語言訓練資料用於訓練一個多語言的 DNN 模型時，對每種語言來說，僅訓練特定語言的輸出層與共享底層隱藏層，也比每種語言重新訓練各自的 DNN 模型要容易得多。直到最近，仍然有許多研究先進們，追隨這樣的想法來進行改良。而在我們的任務中，由於會議語料具有中英文夾雜的特性，我們也嘗試希望透過與不同語言的語料一起訓練，使聲學模型更具一般化能力。

3.2.4 自動語音辨識回饋：

機器學習中，想要改進模型預測的準確率，最簡單且有效的方式，就是用同一組訓練資料訓練多個不同的模型，並且平均它們的預測結果。但是想要訓練多個模型在預測時結合它們預測的結果卻十分耗費計算與時間成本，尤其是當這些模型都屬於大規模的類神經網路時，所耗費的成本更是無法想像。因此，Buciluǎ 等人的研究顯示，把知識從這些已訓練的模型中擷取出來是可能的(Buciluǎ *et al.*, 2006)。

一般來說，我們都會認為用於訓練的目標函數應該盡可能地反映使用者的實際目標。儘管如此，模型的目標函數常常設計成要在訓練資料集上有最佳的辨識效能為準則。這樣的訓練方式使得模型盡可能的讓訓練資料所屬的類別機率越大越好，反而忽略了錯誤答案之間可能隱含的關係。以影像辨識為例，雖然高級跑車的圖片可能會被預測成不同的物體。但是以經驗來看，高級跑車被誤認為垃圾車的機率，應該比被誤認為胡蘿蔔的機率高。假設模型本身對於不同類別的輸出機率已經偷偷告訴我們這些知識，那麼在訓練時若能加入這些資訊，應該有助於提升模型的一般化能力。因此，我們嘗試融合前人的方法，從訓練有素(Well-Trained)的模型中蒸餾出有用的知識，這些知識又被稱為柔性標記(Soft Label) (Hinton *et al.*, 2015)。以 Soft Label 取代傳統非 1 及 0 表示法(Hard Label)，做為訓練模型的標記，這種做法的優點是將不同類別之間的排序資訊也融入訓練的過程中。假設現有已訓練完成的模型，則知識的蒸餾可透過加高輸出層 Softmax 函數的溫度 T，產生 Soft Label，而訓練新的模型時，將 Soft Label 做為輔助任務來進行訓練。Softmax 加上溫度 T 如下式所示：

$$v_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{3}$$

隨著溫度 T 上升，會使得函數輸出值較平緩(Smooth)。此外，Soft Label 在訓練過程中得到的錯誤訊號較小，也較容易滿足目標函數設定的目標。假設目前音框表示為$\mathbf{o}_t$，目前時間點的狀態標記表示為s_t，則 Soft Label 之目標函數表示為：

$$\mathcal{F}_{\text{Soft-Label}} = \sum_t \ln P_{\text{Soft}}(s_t|\mathbf{o}_t) \quad (4)$$

4. 模型重新調適法

主任務與輔助任務的影響力界限該如何拿捏，一直是個重要的研究議題，假使所選擇的輔助任務與主任務不相關，或沒有適當地調整任務的權重，辨識的效果便會大打折扣。為了減輕這個問題，我們受到的學者的研究啟發(Huang *et al*., 2013)，提出了重新調適法。這個方法的核心概念為：在訓練模型時，模型參數較容易擬合最後幾次迭代所送入的訓練資料。因此我們可以將 MTL-DNN 訓練所萃取出的共享隱藏層(Shared Hidden Layers, SHLs)，視為是一個含有額外知識的特徵擷取模組，以此模組為基礎，對主任務重新進行調適訓練，進一步突顯主任務的影響力。重新調適法的流程十分簡單：首先，先以多任務的方式訓練類神經網路聲學模型。待訓練結束後，保留底層 SHLs 的部分，在 SHLs 上層加上新的 Softmax 層，使用主任務的訓練資料與標記進行調適訓練。這種做法將多任務學習視為監督式預訓練(Supervised Pre-Training)，輔助任務在預訓練中扮演正則項(Regularizer)的角色，將模型定位在參數空間中較好的位置，使得後續的微調(Fine Tuning)容易找到良好的局部最小值。實驗也證實了我們提出的重新調適法，確實能夠提升會議語料的辨識率。

5. 實驗

5.1 實驗環境簡介

本論文主要使用的語料庫為國內某大公司錄製的中文會議語料庫(MMRC)，其中收錄了約 43.18 小時的會議語料。語料庫文本經由專家進行轉寫與標記。會議參與人員有 23 位語者。共有 40,022 句。本實驗將會議語料庫分為訓練集、發展集與測試集，如表 2 所示。其中訓練集有 36.02 小時，35,769 句；發展集有 3.52 小時，3,269 句；測試集有 3.64 小時，984 句。

表 2. 中文會議語料庫
[Table 2. Statistics of the Mandarin Meeting Recording Corpus]

MMRC	訓練集	發展集	測試集	總計
小時數	36.02	3.52	3.64	36.02
語句數(句)	35,769	3,269	984	35,769

中文會議語料庫中，所有會議的談話內容及參與人員的對話方式並沒有經過設計，較貼近一般科技公司實際開會的流程。例如聊到專業技術時，通常會出現中英文夾雜的

對話；發表談話時可能有贅詞或口吃的現象，甚至根據開會氣氛的變化，語速和音量也可能產生差異；會議進行時也可能受到外部出現不可預期的噪音干擾；對話過程中，主題可能斷斷續續不連貫；加上不同會議可能位於不同的地點，錄音品質、所使用的麥克風都可能不同。例如有些會議室只有近距離麥克風，有些則是只有遠距離麥克風，抑或是會議室可能有回音干擾等等。因此會議語音是一種十分具備挑戰性的語料。

本論文的實驗使用美國約翰霍普金斯大學學者所發展出的一套大詞彙連續語音辨識的發展工具軟體 Kaldi (Povey *et al*., 2011)，以及 Python 程式語言上的函式庫等提供機器學習或是深層學習與 GPGPU 運算結合的開發環境。此外，多語言及跨語言資訊的輔助任務中，我們用於訓練聲學模型的語料庫如表 3 及表 4 所示。英文的語料為英文音素語料語料庫(TIMIT)，其中訓練集有 3.14 小時，3,696 句；發展集有 0.34 小時，400 句。中文的語料為中文廣播新聞語料庫(Mandarin Chinese Broadcast News Corpus, MATBN) (Wang, Chen, Kuo & Cheng, 2005)，訓練集有 25.60 小時，34,672 句；發展集有 1.3 小時，292 句。實驗所採用的語言模型為以 MMRC 語料所訓練的 *N*-連詞(*N*-gram)語言模型。詞彙的數量有 33,814 詞。音素方面，語料庫含有中文與英文的音素。中文音素切分為聲母及韻母，所含總音素數量為 247 個。

表 3. TIMIT 語料庫
[Table 3. Statistics of the TIMIT corpus]

TIMIT	訓練集	發展集	測試集	總計
小時數	3.14	0.34	(未使用)	3.48
語句數(句)	3,696	400	(未使用)	4,096

表 4. MATBN 語料庫
[Table 4. Statistics of the MATBN corpus]

MATBN	訓練集	發展集	測試集	總計
小時數	25.60	1.36	(未使用)	26.96
語句數(句)	34,672	292	(未使用)	34,964

5.2 基礎實驗

本小節主要想比較不同的語音特徵以及不同的聲學模型對中文會議語料辨識字錯誤率的影響，所使用的聲學模型高斯混合模型(Gaussian Mixture Model, GMM)與深層類神經網路模型。GMM 使用的語音特徵有三種，包含梅爾倒頻譜特徵(MFCC)、線性鑑別分析加上最大化相似度線性轉換(LDA+MLLT)與線性鑑別分析加上最大化相似度線性轉換與語者調適訓練(LDA+MLLT+SAT)，所有的特徵皆使用倒頻譜正規化法，為了表示方便，特

徵與代號的對應表示如下：*tri2* 表示使用梅爾倒頻譜特徵(MFCC)、*tri3* 表示為線性鑑別分析加上最大化相似度線性轉換(LDA+MLLT)的特徵，線性鑑別分析加上最大化相似度線性轉換與語者調適(LDA+MLLT+SAT)則表示為 *tri4*。

在會議的情景下，如果能將不同語者的說話方式與習慣，也就是語者的資訊，加入聲學模型的訓練，應該會有不錯的成效。從實驗結果也可以驗證我們的想法，在 GMM 的基礎實驗中，效果最好的特徵為線性鑑別分析加上最大化相似度線性轉換與語者調適(*tri4*)，所以接下來以類神經網路為主的實驗，都是以 *tri4* 特徵訓練 GMM 以產生類神經網路訓練資料的標記。類神經網路的語音特徵則是梅爾濾波器組特徵(Mel-Frequency Filterbanks, FBANK) 40 維加上 3 維的聲調特徵(Pitch) 共 43 維。鄰近音窗的大小設定為 11 個音框(取目前音框前後各 5 個音框)。並對 43 維語音特徵取相對的一階差量係數(Delta Coefficient)和二階差量係數(Acceleration Coefficient)，輸入特徵的維度為 1419 維。DNN 實驗設定使用 6 層隱藏層，每層 2048 個神經元，隱藏層的活化函數(Activation Function)使用 Sigmoid。CNN 實驗使用的設定是沿著頻率軸掃描的 CNN，架構以摺積層(Convolution Layer)-池化層(Pooling Layer)-摺積層-全連接隱藏層 2 層的順序排列。第一層摺積層的摺積核(Filter)大小為 8 維，共 128 個。第二層摺積層的摺積核大小為 4 維，共 256 個。池化層採用最大池化法(Max Pooling)運算，池化窗的大小為 3 維，池化步伐(Pooling Step)為 3。為了聚焦在輔助任務與重新調適法的效果，類神經網路的權重我們沒有透過預訓練調整。

基礎實驗的步驟流程如下：我們先透過訓練集的語音特徵訓練單連音素的 GMM 聲學模型。根據單連音素模型，訓練三連音素的 GMM 聲學模型。基於三連音素模型再使用不同的特徵(例如 *tri2*、*tri3* 及 *tri4*)分別訓練出三組不同的 GMM。接著我們利用上述三組 GMM，對訓練集的語音資料進行強制對齊(Forced Alignment)，取得每個音框對應的機率密度函數編號(不同的 GMM 所求取的機率密度函數之編號會有差異)，做為訓練資料的標記。最後，我們分別保留這三組 GMM 計算出來的初始機率、轉移機率與強制對齊的資訊(標記)，以最小化交叉熵(Minimum Cross-Entropy, MCE)的目標函數，重新訓練三組 DNN，取代原有的 GMM 來產生每個音框所對應 HMM 狀態的機率。訓練 DNN 與 MTL-DNN 時我們使用小批次隨機梯度下降法(Mini-Batch Stochastic Gradient Descent)，每次 mini-batch 抽樣 256 筆訓練語料特徵輸入類神經網路，微調(Fine-Tuning)使用倒傳導演算法進行網路參數的調整。實驗結果如表 5 所示，可以發現聲學模型改成使用類神經網路後，字錯誤率從 GMM 的 51.88%降低到 38.30%。使用摺積神經網路效果更加顯著，能從 51.88%降低到 38.16%。

表 5. 不同類神經網路模型用於 MMRC 的字錯誤率(%)
[Table 5. Recognition Results achieved by various systems (in Character Error Rate(%))]

模型	特徵	測試集
GMM_*tri4*	MFCC	51.88
DNN6※	FBANK	38.30
CNN-DNN2※	FBANK	38.16
CNN-DNN4	FBANK	35.60
LSTM	FBANK	36.48

※未使用預訓練的類神經網路模型

5.3 多任務學習之實驗結果

本論文的輔助任務可分成 2 大類，一類是語言與音韻學資訊，此類型的資訊主要分為 3 種：音框對應音素標記(代號①到④)、音框對應狀態標記(代號⑤到⑧) 與多語言及跨語言資訊(代號⑨)。另一類則是自動語音辨識回饋(代號⑩)，自動語音辨識回饋有許多研究面向，二 我們所採用的是模型壓縮技術，從已訓練完成的強健模型中產生 Soft Label，提供待訓練模型進行訓練。接下來我們將詳細介紹實驗中輔助任務的設定：

1)音框對應音素標記：音框對應音素標記又可分為 4 種：前一個時間點(左邊)之音框對應的音素標記、目前音框對應的音素標記、下一個時間點(右邊)之音框對應的音素標記與同時預測前一個時間點(左邊)。

2)音框對應狀態標記：概念與前項輔助任務相同，音框對應狀態標記也可分為 3 種：前一個時間點(左邊)之音框對應的狀態標記、目前音框對應的狀態標記與下一個時間點(右邊)之音框對應的狀態標記。值得注意的是，目前音框對應的狀態標記想要預測的目標是不同的特徵所訓練之高斯混合模型產生的三連音素模型的狀態，舉例來說，若主任務以 *tri1* 的狀態標記為目標進行訓練時，輔助任務的目標則是以 *tri2* 的狀態標記為目標進行訓練。

3)多語言及跨語言資訊：我們分別使用 TIMIT 語料庫與 MATBN 語料庫，做為英文和中文的輔助資訊，語料庫統計資訊如表 3 及表 4 所示。我們的實驗可分為兩種，一種是輔助任務只使用 TIMIT 語料庫進行訓練。另一種則是使用 TIMIT 語料庫與 MATBN 語料庫。

4)自動語音辨識回饋：Soft Label 的實驗我們針對不同的溫度、已訓練模型與待訓練模型是否同質、訓練標記精確與否及不同的任務權重比例進行實驗。

上述輔助任務的權重除了特別標註外，皆設定為 1。我們先探討預測目前的音框屬於哪一種音素之輔助任務是否對辨識有幫助，可以從表 6 中觀察到，輔助任務預測音素 *tri3* 有較好的效果，反而預測音素 *tri4* 效果卻不如 *tri3* 明顯。可能是因為經過語者調適訓練的音素 *tri4*，與主任務的 *tri4* 標記性質過於相近，導致效果不明顯。因此，我們可

以發現選擇輔助任務時，除了任務需與主任務相關之外，選擇異質性的任務會有較佳的辨識效果。

表 7 為其它輔助任務在中文會議語料庫的字錯誤率。首先，我們可以先從音素對應標記的輔助任務(②與③)觀察到：預測下一個(右邊)音框所屬的音素標記的辨識效果(37.90%)較預測上一個(左邊)音框的音素標記的效果(38.58%)明顯。而從狀態對應標記的輔助任務(⑥與⑦)觀察到：預測上一個(左邊)音框所屬的狀態標記的辨識效果(36.79%)較預測上一個 (左邊)音框的狀態標記的效果(39.19%)明顯，但同時預測左邊及右邊的音素標記或狀態標 記的辨識率並不如預期，分別為 38.33%與 38.83%，原因應該是由於左右音框標記的輔助任務所佔的權重十分難選擇，需要仰賴經驗來調整。

表 6. 不同音素標記在 MMRC 會議的字錯誤率(%)
[Table 6. Recognition results achieved by MTL-DNN6 with different phoneme labels (in Character Error Rate(%))]

模型	任務編號	輔助任務預測音素	測試集 1
DNN6	Baseline	-	38.30
MTL-DNN6	①	*mono*	37.76
MTL-DNN6	①	*tri2*	37.60
MTL-DNN6	①	*tri3*	36.98
MTL-DNN6	①	*tri4*	37.14

表 7. 不同輔助任務在 MMRC 的字錯誤率(%)
[Table 7. Recognition results achieved by MTL-DNN6 with different auxiliary tasks (in Character Error Rate(%))]

模型	任務代號	備註		測試集
DNN6	Baseline	-		38.30
MTL-DNN6	②	Phone	Left	38.58
MTL-DNN6	③		Right	**37.90**
MTL-DNN6	④		Both	38.33
MTL-DNN6	⑥	State	Left	**36.79**
MTL-DNN6	⑦		Right	39.19
MTL-DNN6	⑧		Both	38.83

表 8 為多語言及跨語言資訊於中文會議語料庫之辨識結果，可以發現 MMRC 語料與 TIMIT 語料一起進行訓練的字錯誤率為 38.06%，可以使得聲學模型在訓練時能夠額外獲取英文音素的知識，在中英文轉換頻繁的語料確實有幫助。而 MMRC、TIMIT 與 MATBN 一起訓練的模型辨識率卻些微上升到 38.23%。我們認為原因可能有二：其一，輔助任務與主任務皆以中文為主，因此幫助並不明顯。其二，輔助任務貢獻(權重)的選擇十分關鍵，也需要經過不斷地嘗試才能找到適合此任務的權重。

表 9 為 Soft Label 的實驗，我們先探討不同的輸出層溫度的影響。以產生 Soft Label 的模型為以狀態標記 *tri4* 所訓練的模型(DNN6)為例，從實驗中可以發現，溫度較高的效果較佳：溫度設定為 5 的辨識率(35.91%)優於溫度設定為 2(36.58%)。而使用較精確的狀態標記(*cnn_ali*)訓練的狀況，則是溫度 2 的辨識效果 (36.20%)優於溫度 5 的辨識效果(36.53%)。因為輸出層溫度較高，表示類別間的排序資訊較豐富。模型訓練時如果以較精確的標記表示(*cnn_ali*)時，不需要過多的排序資訊就能夠訓練得不錯。而當使用較模糊的狀態標記(*tri4*) 時，則需要更多的排序資訊才有較佳的效果。最後是產生 Soft Label 的模型與待訓練模型屬於同質模型的情景。從數據可以發現，當兩者都是摺積神經網路時，字錯誤率為 37.30%，進步的幅度並不大，這表示同質的模型產生的 Soft Label 幫助有限。

表 8. 多語言及跨語言資訊在 MMRC 的字錯誤率(%)
[Table 8. Recognition results obtained by using multi/cross-lingual corpora (in Character Error Rate(%))]

模型	任務代號	共同訓練之語料庫	測試集
DNN6	Baseline	-	38.30
MTL-DNN6	⑨	+TIMIT	38.06
MTL-DNN6	⑨	+TIMIT+MATBN	38.23
MTL-CNN2-DNN2	⑨	+TIMIT	38.17
MTL-CNN2-DNN2	⑨	+TIMIT+MATBN	37.64
MTL-CNN2-DNN2※	⑨	+TIMIT+MATBN	**37.31**

※輔助任務權重為 0.7

表 9. Soft Label 在 MMRC 的字錯誤率(%)
[Table 9. Recognition results obtained by integrating the soft label technique with various DNN models (in Character Error Rate(%))]

模型	溫度	目標標記	任務權重比例 (主:輔)	測試集
DNN6	-	*tri4*	-	38.30
MTL-DNN6 [1]	2	*tri4*	0.5:1	36.58
MTL-DNN6 [1]	5	*tri4*	0.5:1	**35.91**
MTL-CNN2-DNN4 [2]	5	*tri4*	0.5:1	37.30
MTL-DNN6 [2]	2	*cnn_ali*	0.5:1	36.20
MTL-DNN6 [2]	5	*cnn_ali*	0.5:1	36.53

[1]產生 Soft Label 的模型為 DNN6;
[2]產生 Soft Label 的模型為 CNN2-DNN4

最後，重新調適法的實驗結果如表 10 所示。在我們的任務中，調整所有網路的參數之辨識效果優於僅調整輸出層的參數之辨識效果，因此表 10 的數據列出的是重新調整所有網路參數的辨識錯誤率。從實驗中可以發現重新調適的方法對 Soft Label 的模型較無效果，可以推斷 Soft Label 訓練的模型效果較穩定。在多語言及跨語言的任務中，即使模型並沒有調整到最佳的權重(38.23%)，但是在經過調適後，辨識效果提升十分明顯(36.33%)。進步幅度較大的主因可能是因為用於預訓練時，所使用到的訓練語料較多的緣故。另外，預測鄰近音框的音素標記與狀態標記的任務經過重新調適法調整訓練後，預測左邊音框音素標記的辨識率進步到 37.73%，而預測右邊音框音素標記的辨識率進步到 37.13%。而預測左邊音框狀態標記的辨識率進步到 37.94%，預測右邊音框狀態標記的辨識率則可以進步到 37.97%。總結來說，重新調適法並不直接受限輔助任務的優劣，可以嘗試在更多元的設定。

表 10. 重新調適法在 MMRC 的字錯誤率(%
[Table 10. Recognition results achieved by the proposed method (in Character Error Rate(%))]

重新調適之模型	任務代號	備註	測試集
DNN6	baseline	-	38.3
MTL-DNN6	②	Left Phone	37.76
MTL-DNN6	③	Right Phone	**37.13**
MTL-DNN6	⑥	Left State	37.94
MTL-DNN6	⑦	Right State	37.97
MTL-DNN6	⑨	+TIMIT	37.13
MTL-DNN6	⑨	+TIMIT+MATBN	**36.33**
MTL-DNN6	⑩	溫度 2， $tri4$標記	36.96
MTL-DNN6	⑩	溫度 5，cnn_ali標記	37.14

6. 結論與未來展望

聲學模型在會議語音辨識的研究上扮演著十分重要的角色，本論文旨在研究如何融合多任務學習技術於聲學模型之參數估測，藉以改善會議語音辨識之準確性。研究成果與貢獻可分成兩點：1)在多任務學習技術中，我們探究 10 種不同輔助任務在類神經網路聲學模型的成效。其中以使用 Soft Label 及多語言及跨語言資訊做為輔助任務的進步幅度最大。以 Soft Label 做為輔助任務可以使得字錯誤率從 38.30%降低到 35.91%。而以多語言及跨語言資訊做為輔助任務也能使得字錯誤率從 38.30%下降到 38.23%。另外，我們除了結合多任務學習與類神經網路聲學模型之外，也嘗試融合多任務學習於新穎的摺積神經網路上。以跨語言與多語言的輔助任務為例，多任務學習摺積神經網路可以使字錯誤率從 38.23%進一步下降到 37.64%；2) 我們提出重新調適法，使得未調整到最佳輔助任務貢獻(權重)的模型，經過重新調整後，其辨識準確率能夠提升。在多語言及跨語言的任務中，有著最佳的進步幅度，辨識錯誤率從 38.23%降低到 36.33%。

未來可近一步探討五大面向：1)輔助任務選擇：多任務學習的研究大多還是停留在使用語音學及音韻學的資訊(音素，音框狀態)做為輔助任務，未來我們希望能夠探究更有效的輔助任務(例如透過詞、句子或更高層次的特徵)；2)輔助任務間的關係：目前在語音辨識中，多任務學習的研究尚未探討任務之間彼此的關係，例如不同粗細程度的輔助任務之間可能具有階層式的關係，若能將這些關係融入類神經網路訓練中，或許也是一個研究方向；3)輔助任務的影響深度：由於現在多任務學習中，輔助任務只放在類神經網路輸出層，或許可以融入課程學習(Curriculum Learning) (Bengio, Louradour, Collobert

& Weston, 2009)的概念，針對類神經網路不同深度的隱藏層設計不同難度的任務；4)融合更新穎的模型：現今遞迴式神經網路與其改良之長短期記憶網路(Long Short-Term Memory) (Li, Mohamed, Zweig & Gong, 2016)在語音辨識中也取得了不錯的成果，我們也希望善用這些新穎的聲學模型來改進辨識的正確率；5)最後，由於大多數的辨識錯誤常發生於中文詞與英文詞發音很類似的狀況，例如『size』被辨識為『才是』或 bottleneck 被辨識成『把那個』。辨識錯誤之詞在句子中的位置，又會間接影響到句子後面其它詞的辨識結果，導致一連串的辨識錯誤連鎖發生。如果想要使得會議語音辨識更加地符合實際需求，其中一項重要的目標，應該是提升關鍵詞的辨識準確率。如果能夠依據關鍵詞辨識正確與否，以最佳化辨識關鍵詞效能的訓練準則調整聲學模型也是一個值得研究的方向。

致謝

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(104-2911-I-003-301)與行政院科技部研究計畫(MOST 104-2221-E-003-018-MY3 和 MOST 105-2221-E-003-018-MY3)之經費支持，謹此致謝。

Reference

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533-1545.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*, 153.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 41-48.

Buciluǎ, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, 535-541.

Caruana, R. (1997). *Multitask learning* (Doctoral dissertation, University of Carnegie Mellon).

Chen, D., & Mak, B. K. W. (2015). Multitask learning of deep neural networks for low-resource speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *23*(7), 1172-1183.

Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B., Morgan, N., Renals, S., Robinson, T., & Williams, G. (1999). An overview of the SPRACH system for the transcription of broadcast news. In *Proceedings of the DARPA Broadcast News Workshop*.

Fiscus, J. (1997). A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 347-354.

Gales, M. J. F. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, *12*(2), 75-98.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM*. NIST speech disc 1-1.1. NASA STI/Recon technical report n, 93.

Ghoshal, A., Swietojanski, P., & Renals, S. (2013). Multilingual training of deep neural networks. In P*roceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 7319-7323.

Graves, A., Mohamed, A. R., & Hinton, G. E. (2013). Speech recognition with deep recurrent neural networks. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 6645-6649.

Hermansky, H., Ellis, D. P., & Sharma, S. (2000). Tandem connectionist feature extraction for conventional HMM systems. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 1635-1638.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82-97.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. Retrived from https://arXiv preprint arXiv:1503.02531.

Huang, J. T., Li, J., Yu, D., Deng, L., & Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proceedings of the International Conference on Speech Communication and Technology (INTERSPEECH)*, 7304-7308.

Li, J., Mohamed, A. R., Zweig, G., & Gong, Y. (2016). Exploring multidimensional LSTMs for large vocabulary ASR. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 4940-4944.

Parveen, S., & Green P. D. (2003). Multitask learning in connectionist ASR using recurrent neural networks. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 1813-1816.

Poultney, C., Chopra, S., & Cun, Y. L. (2006). Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems*, 1137-1144.

Povey, D. (2004). *Discriminative training for large vocabulary speech recognition* (Doctoral dissertation, University of Cambridge).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer G., & Vesely, K. (2011). The

Kaldi speech recognition toolkit. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*.

Rumelhart, D. E., Hinton, G. E., & Ronald, J. W. (1988). Learning representations by back-propagating errors. *Cognitive Modeling*, *5*(3).

Schultz, T., & Waibel, A. (2001). Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, *35*(1), 31-51.

Seltzer, M. L., & Droppo, J. (2013). Multi-task learning in deep neural networks for improved phoneme recognition. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 6965-6969.

Sercu, T., Puhrsch, C., Kingsbury, B., & LeCun, Y. (2016). Very deep multilingual convolutional neural networks for LVCSR. In *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 4955-4959.

Swietojanski, P., Ghoshal, A., & Renals, S. (2012). Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proceedings of the International Conference on Spoken Language Technology Workshop (SLT)*, 246-251.

Thrun, J. S., & Pratt, L. (1988). *Learing to learn*. Norwell, MA : Kluwer Academic Publishers.

Wang, H. M., Chen, B., Kuo, J. W., & Cheng, S. S. (2005). MATBN: a Mandarin Chinese broadcast news corpus. *Journal of Computational Linguistics and Chinese Language Processing*, *10*(2), 219-236.

Werbos, P. J. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences* (Doctoral dissertation, University of Harvard).

Young, S. J., & Woodland, P. C. (1993). The use of state tying in continuous speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2207-2219.

Yu, D., & Deng, L. (2014). *Automatic speech recognition: a deep learning approach*. London, England: Springer-Verlag.

Zhang, Y., & Yeung, D. Y. (2014). A regularization approach to learning task relationships in multitask learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *8*(3), 12.

The individuals listed below are reviewers of this journal during the year of 2016. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Guo-Wei Bian
Jing-Shin Chang
Tao-Hsing Chang
Yu-Yun Chang
Yi-Hsiang Chao
Chien Chin Chen
Kuang-Hua Chen
Tai-Shih Chi
Chih-Yi Chiu
Hong-Jie Dai
Hung-Yan Gu
Wei-Tyng Hong
Jia-Fei Hong
Shu-Kai Hsieh
Jen-Wei Huang
Jeih-Weih Hung
Chih-Chung Kuo
June-Jei Kuo
Wen-Hsing Lai
Chi-Chun Lee
Hong-Yi Lee
Tan Lee
Bor-Shen Lin
Chuan-Jie Lin
Shu-Yen Lin
Chu-Cheng Lin
Chao-Hong Liu
Jyi-Shane Liu
Chiarung Lu
Ming-Feng Tsai
Wei-Ho Tsai
Chin-Chin Tseng
Hsu Wang
Jenq-Haur Wang
Jiun-Shiung Wu
Cheng-Zen Yang
Huiling Yang
Jui-Feng Yeh
Ming-Shing Yu

2016 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 21

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2016.

The Author Index contains the primary entry for each item, listed under the first author's name. The primary entry includes the coauthors' names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author's name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

B

Bai, Ming-Hong
Jian-Cheng Wu, Ying-Ni Chien, Shu-Ling Huang and Ching-Lung Lin. A Study on Dispersion Measures for Core Vocabulary Compilation; 21(2): 1-18

C

Chang, Yung-Chun
Chun-Han Chu, Chien Chin Chen and Wen-Lian Hsu. Linguistic Template Extraction for Recognizing Reader-Emotion; 21(1): 29-50

Chen, Berlin
see Hsu, Yao-Chi, 21(2): 55-70
see Yan, Bi-Cheng, 21(2): 35-54
see Yang, Ming-Han, 21(2): 85-104

Chen, Chien Chin
see Chang, Yung-Chun, 21(1): 29-50

Chen, Kuan-Hung
Shu-Han Liao, Yuan-Fu Liao and Yih-Ru Wang. Character-Level Linguistic Features Extraction for Text-to-Speech System; 21(2): 71-84

Chen, Kuan-Yu
see Hsu, Yao-Chi, 21(2): 55-70
see Yang, Ming-Han, 21(2): 85-104

Chen, Yanping
Qinghua Zheng, Feng Tian and Deli Zheng. A Segmentation Matrix Method for Chinese Segmentation Ambiguity Analysis; 21(1): 1-28

Chen, Ying-Wen
see Yang, Ming-Han, 21(2): 85-104

Chien, Ying-Ni
see Bai, Ming-Hong, 21(2): 1-18

Chu, Chun-Han
see Chang, Yung-Chun, 21(1): 29-50

H

Hsieh, Yu-Ming
and Wei-Yun Ma. N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings; 21(2): 19-34

Hsu, Wen-Lian
see Chang, Yung-Chun, 21(1): 29-50

Hsu, Yao-Chi
Ming-Han Yang, Hsiao-Tsung Hung, Yi-Ju Lin, Kuan-Yu Chen and Berlin. Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection; 21(2): 55-70
see Yang, Ming-Han, 21(2): 85-104

Huang, Shu-Ling
see Bai, Ming-Hong, 21(2): 1-18

Huang, Yu-Yang
Rui Yan, Tsung-Ting Kuo and Shou-De Lin. Enriching Cold Start Personalized Language Model Using Social Network Information; 21(1): 51-68

Hung, Hsiao-Tsung
see Hsu, Yao-Chi, 21(2): 55-70
see Yang, Ming-Han, 21(2): 85-104

K

Kuo, Tsung-Ting
see Huang, Yu-Yang, 21(1): 51-68

L

Liao, Shu-Han
see Chen, Kuan-Hung, 21(2): 71-84

Liao, Yuan-Fu
see Chen, Kuan-Hung, 21(2): 71-84

Lin, Ching-Lung
see Bai, Ming-Hong, 21(2): 1-18

Lin, Shou-De
see Huang, Yu-Yang, 21(1): 51-68

Lin, Yi-Ju
see Hsu, Yao-Chi, 21(2): 55-70

Liu, Shih-Hung
see Yan, Bi-Cheng, 21(2): 35-54

Lu, Wen-Hsiang
see Wang, Ting-Xuan, 21(1): 69-90

M

Ma, Wei-Yun
see Hsieh, Yu-Ming, 21(2): 19-34

S

Shih, Chin-Hong
see Yan, Bi-Cheng, 21(2): 35-54

T

Tian, Feng
see Chen, Yanping, 21(1): 1-28

W

Wang, Ting-Xuan
and Wen-Hsiang Lu. Identifying the Names of Complex Search Tasks with Task-Related Entities; 21(1): 69-90

Wang, Yih-Ru
see Chen, Kuan-Hung, 21(2): 71-84

Wu, Jian-Cheng
see Bai, Ming-Hong, 21(2): 1-18

Y

Yan, Bi-Cheng
Chin-Hong Shih, Berlin Chen and Shih-Hung Liu. The Use of Dictionary Learning Approach for Robustness Speech Recognition; 21(2): 35-54

Yan, Rui
see Huang, Yu-Yang, 21(1): 51-68

Yang, Ming-Han
Yao-Chi Hsu, Hsiao-Tsung Hung, Ying-Wen Chen, Kuan-Yu Chen and Berlin Chen. Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition; 21(2): 85-104

Z

Zheng, Deli
see Chen, Yanping, 21(1): 1-28

Zheng, Qinghua
see Chen, Yanping, 21(1): 1-28

SUBJECT INDEX

A

Automatic Speech Recognition
Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection; Hsu, Y.-C., 21(2): 55-70

The Use of Dictionary Learning Approach for Robustness Speech Recognition; Yan, B.-C., 21(2): 35-54

C

Cold-Start Problem
Enriching Cold Start Personalized Language Model Using Social Network Information; Huang, Y.-Y., 21(1): 51-68

Complex Search Task
Identifying the Names of Complex Search Tasks with Task-Related Entities; Wang, T.-X., 21(1): 69-90

Computer Assisted Pronunciation Training
Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection; Hsu, Y.-C., 21(2): 55-70

Core Vocabulary
A Study on Dispersion Measures for Core Vocabulary Compilation; Bai, M.-H., 21(2): 1-18

Corpus Linguistics
A Study on Dispersion Measures for Core Vocabulary Compilation; Bai, M.-H., 21(2): 1-18

D

Deep Learning
Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition; Yang, M.-H., 21(2): 85-104

Deep Neural Networks
Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection; Hsu, Y.-C., 21(2): 55-70

Dictionary Learning
The Use of Dictionary Learning Approach for Robustness Speech Recognition; Yan, B.-C., 21(2): 35-54

Discrimetive Training
Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection; Hsu, Y.-C., 21(2): 55-70

Dispersion Uniformity
A Study on Dispersion Measures for Core Vocabulary Compilation; Bai, M.-H., 21(2): 1-18

E

Emotion Template
Linguistic Template Extraction for Recognizing Reader-Emotion; Chang, Y.-C., 21(1): 29-50

F

Factor Graph
Enriching Cold Start Personalized Language Model Using Social Network Information; Huang, Y.-Y., 21(1): 51-68

Fringe Vocabulary
A Study on Dispersion Measures for Core Vocabulary Compilation; Bai, M.-H., 21(2): 1-18

L

Language Model
Enriching Cold Start Personalized Language Model Using Social Network Information; Huang, Y.-Y., 21(1): 51-68

Linguistic Features
Character-Level Linguistic Features Extraction for Text-to-Speech System; Chen, K.-H., 21(2): 71-84

M

Meeting Speech Recognition
Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition; Yang, M.-H., 21(2): 85-104

Mispronunciation Detection
Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection; Hsu, Y.-C., 21(2): 55-70

Modulation Spectrum
The Use of Dictionary Learning Approach for Robustness Speech Recognition; Yan, B.-C., 21(2): 35-54

Multi-Task Learning
Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition; Yang, M.-H., 21(2): 85-104

N

Neural Networks
Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition; Yang, M.-H., 21(2): 85-104

P

Parsing
N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings; Hsieh, Y.-M., 21(2): 19-34

R

Reader-Emotion Detection
Linguistic Template Extraction for Recognizing Reader-Emotion; Chang, Y.-C., 21(1): 29-50

Rescoring
N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings; Hsieh, Y.-M., 21(2): 19-34

RNNLM
Character-Level Linguistic Features Extraction for Text-to-Speech System; Chen, K.-H., 21(2): 71-84

Robustness
The Use of Dictionary Learning Approach for Robustness Speech Recognition; Yan, B.-C., 21(2): 35-54

S

Segmentation Ambiguity
A Segmentation Matrix Method for Chinese Segmentation Ambiguity Analysis; Chen, Y., 21(1): 1-28

Segmentation Matrix
A Segmentation Matrix Method for Chinese Segmentation Ambiguity Analysis; Chen, Y., 21(1): 1-28

Sentiment Analysis
Linguistic Template Extraction for Recognizing Reader-Emotion; Chang, Y.-C., 21(1): 29-50

Smoothing
Enriching Cold Start Personalized Language Model Using Social Network Information; Huang, Y.-Y., 21(1): 51-68

Social Network Analysis
Enriching Cold Start Personalized Language Model Using Social Network Information; Huang, Y.-Y., 21(1): 51-68

Sparse Coding
The Use of Dictionary Learning Approach for Robustness Speech Recognition; Yan, B.-C., 21(2): 35-54

Speech Synthesis
Character-Level Linguistic Features Extraction for Text-to-Speech System; Chen, K.-H., 21(2): 71-84

T

Task Name Identification
Identifying the Names of Complex Search Tasks with Task-Related Entities; Wang, T.-X., 21(1): 69-90

Task-related Entity
Identifying the Names of Complex Search Tasks with Task-Related Entities; Wang, T.-X., 21(1): 69-90

Template-based Approach
Linguistic Template Extraction for Recognizing Reader-Emotion; Chang, Y.-C., 21(1): 29-50

Text Classification
Linguistic Template Extraction for Recognizing Reader-Emotion; Chang, Y.-C., 21(1): 29-50

W

Word Dependency
N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings; Hsieh, Y.-M., 21(2): 19-34

Word Embedding
N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings; Hsieh, Y.-M., 21(2): 19-34

Word2vec
Character-Level Linguistic Features Extraction
for Text-to-Speech System; Chen, K.-H.,
21(2): 71-84

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member： NT$ 1,000 (US$50.-)
group membership： NT$20,000 (US$1,000.-)
life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502 Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： _____

Name： _____ Date of Birth： _____

Country of Residence： _____ Province/State： _____

Passport No.： _____ Sex: _____

Education(highest degree obtained)： _____

Work Experience： _____

_____

Present Occupation： _____

Address： _____

_____

Email Add： _____

Tel. No： _____ Fax No： _____

Membership Category： ☐ Regular Member ☐ Life Member

Date： ____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）

Life Member ： US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）
（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目
（三）收集國內外有關計算語言學知識之圖書及最新發展之資料
（四）發行有關之學術刊物，論文集及通訊
（五）研定有關計算語言學專用名稱術語及符號
（六）與國際計算語言學學術機構聯繫交流
（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會
2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
   信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

<table>
<tr><td>會員類別</td><td colspan="2">□終身 □個人 □學生</td><td>會員編號</td><td colspan="2">（由本會填寫）</td></tr>
<tr><td rowspan="2">姓　　名</td><td rowspan="2"></td><td rowspan="2">性別</td><td rowspan="2"></td><td>出生日期</td><td>年　月　日</td></tr>
<tr><td>身分證號碼</td><td></td></tr>
<tr><td>現　　職</td><td></td><td>學　　歷</td><td colspan="2"></td></tr>
<tr><td>通訊地址</td><td colspan="5">□□□</td></tr>
<tr><td>戶籍地址</td><td colspan="5">□□□</td></tr>
<tr><td>電　　話</td><td colspan="2"></td><td>E-Mail</td><td colspan="2"></td></tr>
<tr><td colspan="6">申請人:　　（簽章）

中　華　民　國　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

   終身會員：　10,000.-
   個人會員：　1,000.-
   學生會員：　500.-（限國內學生）
   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)
   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638
   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw
   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)
PAYMENT FORM

Name: _____(Please print) Date: _____

Please debit my credit card as follows: US$ _____

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:_____

Card No.: _____ -_____-_____ -_____ Exp. Date:_____(M/Y)

3-digit code: _____ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE: _____

Phone No.: _____E-mail: _____

Address: _____

PAYMENT FOR

US$ ____ ❑ Computational Linguistics & Chinese Languages Processing (IJCLCLP)

Quantity Wanted: _____

US$ ____ ❑ Journal of Information Science and Engineering (JISE)

Quantity Wanted: _____

US$ ____ ❑ Publications:_____

US$ ____ ❑ Text Corpora: _____

US$ ____ ❑ Speech Corpora:_____

US$ ____ ❑ Others: _____

US$ _____ ❑ Membership Fees ❑ Life Membership ❑ New Membership ❑Renew

US$ _____ = Total

Fax 886-2-2788-1638 or Mail this form to:

ACLCLP
% IIS, Academia Sinica
Rm502, No.128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: _____(請以正楷書寫)　　日期:：_____

卡別：❑ VISA CARD　❑ MASTER CARD ❑ JCB CARD　發卡銀行：_____

信用卡號：_____-_____-_____-_____　有效日期：_____(m/y)

卡片後三碼：_____（卡片背面簽名欄上數字後三碼）

持卡人簽名：_____(簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$_____ ❑ 中文計算語言學期刊(IJCLCLP) _____

NT$_____ ❑ Journal of Information Science and Engineering (JISE)

NT$_____ ❑ 中研院詞庫小組技術報告_____

NT$_____ ❑ 文字語料庫 _____

NT$_____ ❑ 語音資料庫 _____

NT$_____ ❑ 光華雜誌語料庫1976~2010

NT$_____ ❑ 中文資訊檢索標竿測試集/文件集

NT$_____ ❑ 會員年費：❑續會　❑新會員　❑終身會員

NT$_____ ❑ 其他: _____

NT$_____ = 合計

填妥後請傳真至 02-27881638 或郵寄至:

11529台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

		Surface	AIR (US&EURP)	AIR (ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: _____Total Due:_____

- **OVERSEAS USE ONLY**
- PAYMENT： ☐ Credit Card (Preferred)
  ☐ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address：_____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息爲本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期）　年份：_____（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※ 此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人： 黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：_____ 收據抬頭：_____

地　　址：_____

電　　話：_____ E-mail: _____

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright : It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.

1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.

2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.

3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.

4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript

6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.

7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).

Please visit the following websites for details.

(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.21 no.2 December 2016

Contents

Papers

基於詞語分布均勻度的核心詞彙選擇 [A Study on Dispersion Measures for Core Vocabulary Compilation]..... 1
白明弘(Ming-Hong Bai), 吳鑑城(Jian-Cheng Wu), 簡盈妮(Ying-Ni Chien), 黃淑齡(Shu-Ling Huang), 林慶隆(Ching-Lung Lin)

N-best Rescoring for Parsing Based on Dependency-Based Word Embeddings..... 19
Yu-Ming Hsieh and Wei-Yun Ma

使用字典學習法於強健性語音辨識 [The Use of Dictionary Learning Approach for Robustness Speech Recognition]..... 35
顏必成(Bi-Cheng Yan), 石敬弘(Chin-Hong Shih) 劉士弘(Shih-Hung Liu), 陳柏琳(Berlin Chen)

評估尺度相關最佳化方法於華語錯誤發音檢測之研究 [Evaluation Metric-related Optimization Methods for Mandarin Mispronunciation Detection]..... 55
許曜麒(Yao-Chi Hsu), 楊明翰(Ming-Han Yang), 洪孝宗(Hsiao-Tsung Hung), 林奕儒(Yi-Ju Lin), 陳冠宇(Kuan-Yu Chen), 陳柏琳(Berlin Chen)

基於字元階層之語音合成用文脈訊息擷取 [Character-Level Linguistic Features Extraction for Text-to-Speech System]..... 71
陳冠宏(Kuan-Hung Chen), 廖書漢(Shu-Han Liao), 廖元甫(Yuan-Fu Liao), 王逸如(Yih-Ru Wang)

融合多任務學習類神經網路聲學模型訓練於會議語音辨識之研究 [Leveraging Multi-Task Learning with Neural Network Based Acoustic Modeling for Improved Meeting Speech Recognition]..... 85
楊明翰(Ming-Han Yang), 許曜麒(Yao-Chi Hsu), 洪孝宗(Hsiao-Tsung Hung), 陳映文(Ying-Wen Chen), 陳冠宇(Kuan-Yu Chen), 陳柏琳(Berlin Chen)

Reviewers List & 2016 Index..... 105

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing