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.

Special Issue on "Selected Papers from ROCLING XXIX"
Guest Editors: Chi-Chun (Jeremy) Lee and Cheng-Zen Yang

易繫辭曰上古結繩而
治後世聖人易之以書
契百官以治萬民以察
說文敘曰蓋文字者經
藝之本宣教明化之始
前人所以垂後後人所
以識古故曰本立而道
生知天下之至賾而不
可亂也教化既萌文心
雕龍則謂人之立言因
字而生句積句而成章
積章而成篇篇之彪炳

Vol.22 No.2 December 2017 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

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. 22, no. 2, December 2017

Contents

Special Issue Articles: Selected Papers from ROCLING XXIX

Papers

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. i-ii

Forewords

The 29th Conference on Computational Linguistics and Speech Processing (ROCLING 2017) was held at Nangang Exhibition Center, Taipei, Taiwan on Nov. 27-28, 2017. ROCLING, which sponsored by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP), is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, bringing together researchers, scientists and industry participants from fields of computational linguistics, information understanding, and speech processing, to present their work and discuss recent trends in the field. This special issue presents the extended and reviewed versions of seven papers meticulously selected from ROCLING 2017, including 3 natural language processing papers and 4 speech processing papers.

The first paper presents a neural relevance-aware model (NRM) for spoken document retrieval (SDR). The notion of query intent classification is incorporated into the proposed NRM modeling framework to obtain more sophisticated query representations. This paper is awarded as one of two best papers of ROCLING 2017. The second paper discusses the question retrieval problem for community-based question answering (CQA). This paper proposes a retrieval approach using word embedding learning and participant reputation ranking in the community. The third paper from National Taiwan Normal University focuses on the text readability classification problem. This paper proposes two novel readability models built on top of a convolutional neural network based representation and the so-called fast text representation, respectively. The fourth paper discusses the acoustic echo cancellation problem. This paper presents an improved vector-space-based adaptive filtering algorithm to achieve fast converge and a better echo return loss enhancement performance. This paper is also awarded as one of two best papers of ROCLING 2017. The fifth paper focuses on the replay spoofing detection problem to tell whether the given utterance comes directly from the mouth of a speaker or indirectly through a playback. This paper presents a discriminative autoencoder (DcAE) neural network model which achieves up to 42% relative improvement in the equal error rate (EER) over the official baseline. The sixth paper discusses the tone group parser issue for Taiwanese text-to-speech. This paper presents a hypothesis of tonal government arguing that if the allotone selection can be made for each word in a sentence then the tone groups will be separated within the sentence. The last paper from National Tsing Hua University focuses on the speech emotion recognition problem. The approach proposed in this paper achieves improvements based on a set of trajectory-based spatial-temporal spectral features.

The Guest Editors of this special issue would like to thank all of the authors and reviewers for sharing their knowledge and experience at the conference. We hope this issue provide for directing and inspiring new pathways of NLP and spoken language research within the research field.

Guest Editors
Chi-Chun (Jeremy) Lee
Department of Electrical Engineering, National Tsing Hua University, Taiwan
Cheng-Zen Yang
Department of Computer Science and Engineering, Yuan Ze University, Taiwan

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 1-16

語音文件檢索使用類神經網路技術

On the Use of Neural Network Modeling Techniques for Spoken Document Retrieval

羅天宏[*]、陳映文[*]、陳冠宇[+]、王新民[#]、陳柏琳[*]

Tien-Hong Lo, Ying-Wen Chen, Kuan-Yu Chen,

Hsin-Min Wang and Berlin Chen

摘要

近年來由於含有語音資訊的多媒體內涵不斷增長，語音文件檢索已成為一個相當熱門的議題並吸引許多學者與實務家的投入研究。除了發展強健的索引機制和有效的檢索模型外，如何正確地且有效率地對於查詢內容進行模型化對於增進語音文件檢索的表現也扮演著非常關鍵的角色。有鑒於此，在本論文，我們提出一個新穎的基於類神經網路之相關性感知模型來得到較佳的查詢表示方式，同時可以避免使用傳統較耗費時間的準相關回饋程序。再者，我們嘗試將查詢意向分類的概念融入我們所提出的模型架構中，以進一步獲取更精緻的查詢表示方式。在 TDT-2 語音文件及所進行的初步實驗顯示出本論文所提出方法的效用。

關鍵詞：語音文件檢索，查詢意向，類神經網路，準相關回饋

[*] 國立臺灣師範大學資訊工程學系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: {teinhonglo, cliffchen, berlin}@ntnu.edu.tw

[+] 國立臺灣科技大學資訊工程系
Department of Computer Science & Information Engineering, National Taiwan University of Science and Technology
E-mail: kychen @mail.ntust.edu.tw

[#] 中央研究院資訊科學研究所
Institute of Information Science, Academia Sinica
E-mail: whm@iis.sinica.edu.tw

Abstract

Due to ever-increasing amounts of publicly available multimedia associated with speech information, spoken document retrieval (SDR) has been an active area of research that captures significant interest from both academic and industrial communities. Beyond the continuing effort in the development of robust indexing and effective retrieval methods to quantify the relevance degree between a pair of query and spoken document, how to accurately and efficiently model the query content plays a vital role for improving SDR performance. In view of this, we present in this paper a novel neural relevance-aware model (NRM) to infer an enhanced query representation, extricating the conventional time-consuming pseudo-relevance feedback (PRF) process. In addition, we incorporate the notion of query intent classification into our proposed NRM modeling framework to obtain more sophisticated query representations. Preliminary experiments conducted on the TDT-2 collection confirm the utility of our methods in relation to a few state-of-the-art ones.

Keyword：Spoken Document Retrieval, Query Intent, Neural Network, Pseudo-Relevance Feedback

1. 緒論 (INTRODUCTION)

伴隨著網際網路的發展與多媒體資訊的大量增長，影音的瀏覽與傳遞也逐漸成為我們的日常生活的一部分。在這環境下，如何利用語音的資訊，快速檢索符合資訊需求的內容，變成了一項新興的需求。因此，在過去的二十年(Chelba, Hazen & Saraclar, 2008) (Lee & Chen, 2005) (Huang, Ma, Li & Wu, 2011) (Chen, Chen, Chen & Chen, 2012)，語音文件檢索成為一個十分有魅力的研究主題。在語音文件檢索的任務上，過往有許多顯著成功的方法，如向量空間模型(Vector Space Model) (Salton, Wong & Yang, 1975)、Okapi BM25 model (Jones, Walker & Robertson, 2000)，以及主題模型(Topic Model) (Blei, Ng & Jordan, 2003)等。另一方面，將統計式語言模型(Statistical Language Model)應用在文字檢索(Information Retrieval)和語音文件檢索，在檢索任務上取得了嶄新的突破(Ponte & Croft, 1998) (Song & Croft, 1999) (Croft & Lafferty, 2003)，因此吸引了不少研究者的目光。在這樣的概念下，查詢對每個文件計算似然機率後作排名，我們稱這樣的排序方法為查詢似然測量(Query Likelihood Model Measure, QLM) (Manning, Raghavan & Schutze, 2008)。另一個知名的評估方式為 KL 散度測量(Kullback-Leibler Divergence Measure, KLM) (Zhai & Lafferty, 2001)，將查詢與文件皆表示為單元語法的語言模型(Unigram Language Model)，查詢與文件的相似程度即為兩個機率分佈的散度距離(Divergence Distance)。

最近，隨著深層類神經網路架構的流行，這類的方法也被大量應用在檢索的任務上。主要的研究方向為利用不同網路架構與訓練方法，以此來學習查詢與文件間的相似關係(Guo, Fan, Ai & Croft, 2016) (Mitra, Diaz & Craswell, 2017)。值得注意的是，大部分方法

的輸入資料皆是如詞頻的簡單統計資訊(Surface Statistics)，且期望類神經網路能夠區分出相關與不相關的文件。基於以上的想法，有兩個較為知名的方法，分別是深層結構語義模型(Deep Structured Semantic Model, DSSM) (Huang *et al*., 2013)和局部特徵向量模型(Locality Preserving Essence Vector, LPEV) (Chen, Liu, Chen & Wang, 2017)。DSSM 和 LPEV 兩種方法也皆致力於在非監督的環境下，將查詢與文件分別以低維度的向量作表示，使得向量不是單純的文字序列或簡單的統計關係(如鄰近的詞、共同出現的詞等更豐富的資訊)。除了以上兩種檢索模型，還有兩種較為知名的嵌入方法，分別是詞嵌入(word embedding-based methods)和段嵌入(paragraph embedding methods)的方法，前者的方法有略詞模型(Skip-gram model) (Mikolov, Sutskever, Chen, Corrado & Dean, 2013)和連續詞袋(Continuous bag-of-words) (Mikolov, Chen, Corrado & Dean, 2013)。後者的方法有分佈式內存模型(the distributed memory model) (Le & Mikolov, 2014)、分佈式詞袋模型(distributed bag-of-words model) (Le & Mikolov, 2014) (Chen, Lee, Wang, Chen & Chen, 2014)和本質向量模型(essence vector model) (Chen, Liu, Chen & Wang, 2016)等。

為了提升檢索效能，過往有許多方法嘗試猜測使用者意向進行查詢分類 (Shen *et al*., 2006)，給予對不同類別有興趣的使用者提供更為精確的結果。然而，如同傳統文字文件檢索，語音文件檢索也面臨了查詢過於簡短語意不清，且會隨著時間推移改變語句的意思，難以表達出使用者的資訊需求，因此查詢分類往往比文件分類困難。針對查詢語意不足的問題，查詢分類被定義為對使用者行為的建模，其一便是豐富語意表示的任務，稱為擴增查詢(Query Expansion)，擴增查詢主要可以分為兩個類別，第一種為引用外部的資源(如 Wikipedia 或 WordNet)分析語意，進一步擴展原始的查詢，其中的語意關係包括同義詞、反義詞、多義詞等；第二種為分析查詢的回饋，給予一個查詢，追蹤使用者點擊的文件，並以此做分類，以求第二次更為精確的結果，與非監督式的準相關回饋(Pseudo-Relevance Feedback) (Zhai & Lafferty, 2001) (Lavrenko & Croft, 2001) (Chen, Liu, Chen, Wang & Chen, 2016)的精神相似。第一個方法需要較複雜的自然語言處理技術，如語意表徵和推論。第二種方法則較為簡單，只需要取得前幾篇文章做分析，再適當地與原始查詢做結合即可。且因為分析的資料僅為回饋的文件，準相關回饋不需要額外的語料庫來學習。以下為幾個知名的準相關回饋(Manning *et al*., 2008)，相關性模型(Relevance Model, RM) (Lavrenko & Croft, 2001) 、簡易混合模型(Simple Mixture Model, SMM) (Zhai & Lafferty, 2001)、顯著詞模型(Significant Word Model, SWM) (Dehghani, Azarbonyad, Kamps, Hiemstra & Marx, 2016) (Chen, Chen, Wang & Chen, 2017)。儘管準相關回饋已經在檢索的領域上證實有效性，但仍需要做二次查詢，造成缺乏即時性的問題，因此實務上難以採用。

我們的動機便是解決上述提到的重要問題，有效性與即時性。因此這篇論文致力提出一個基於查詢分類的擴增查詢架構，其中利用到以類神經網路架構為基礎的 NRM，並探討了 NRM 利用查詢意向探索的可能性。

2. 相關文獻回顧 (RELATED WORK)

由於在現實中，查詢所用到的詞並不多，因此常用於估測查詢Q的詞出現機率的語言模型$P(w|Q)$的最大似然估測(Maximum Likelihood Estimator)便很難發揮所長。為了減緩這樣的侷限，有許多專家學者提出不同的方法，其中已經證實有效且泛用的方法便是準相關回饋，目的是用於估測出更為精確的查詢表示式。準相關回饋假設，查詢 Q 與檢索結果的前 N 篇文章相關，因此分析第一次查詢結果的文件$\mathbf{D}_F=\{D_1,\cdots,D_r,\cdots,D_{|\mathbf{D}_F|}\}$，來達到更為精確的查詢語言模型。我們探討的方法為以下三個，RM、SMM、SWM。

2.1 相關性模型(Relevance Model)

RM 假設每一個查詢Q皆有一個相關類別R_Q，且每個與查詢Q相關的文件皆由相關類別R_Q中產生。不幸的是，我們不會知道查詢Q的相關類別R_Q，因此作為替代，我們會將前 N 篇回饋的文章$\mathbf{D}_F$當作相關文章，並且利用準相關回饋近似真實的相關類別R_Q。對應查詢Q的相關性模型可利用以下公式計算：

$$P_{\text{RM}}(w|Q)=\frac{\sum_{D_r\in\mathbf{D}_F}P(D_r)P(w|D_r)\prod_{w'\in Q}P(w'|D_r)}{\sum_{D'_r\in\mathbf{D}_F}P(D'_r)\prod_{w''\in Q}P(w''|D'_r)}, \tag{1}$$

其中$P(D_r)$為文件的產生機率。由於我們沒有對於文件的先驗知識，因此決定機率的方式是用均勻分佈(Uniform Distribution)來實現。另一個語言模型$P(w|D_r)$則是利用最大似然估計的方式，利用文件的詞頻和該文件與查詢相似程度，計算出每個詞出現在文件D_r的機率，其餘符號以此類推。

2.2 簡易混合模型(Simple Mixture Model)

另一個估測查詢語言模型的觀點是 SMM。SMM 假設在回饋文件$\mathbf{D}_F$裡詞出現機率不是由單一模型估測，而是來自兩部份混合而成的語言模型，第一部份是專屬於該查詢Q的特殊詞的主題模型$P_{\text{SMM}}(w|Q)$；第二部份是廣泛出現在各個文件中的背景語言模型$P_{\text{BG}}(w)$。如此一來，便可透過分析回饋文件$\mathbf{D}_F$，最大化 SMM 的似然機率，並求得$P_{\text{SMM}}(w|Q)$，以下為估測時使用的損失函數(loss function)：

$$L=\prod_{D\in\boldsymbol{D}_F}\prod_{w\in V}[\alpha\cdot P_{\text{SMM}}(w|Q)+(1-\alpha)\cdot P_{\text{BG}}(w)]^{c(w,D)}, \tag{2}$$

α為預先定義好的參數，用於決定$P_{\text{SMM}}(w|Q)$和$P_{\text{BG}}(w)$兩者的比重關係。這樣的估計方式讓針對查詢Q的特殊詞得到較高的機率，進而獲得一個更有效的查詢模型。SMM 的假設提供一個有益的訊息，便是在回饋文件$\mathbf{D}_F$出現的詞，不僅與查詢Q本身相關，也與不存在於回饋文件$\mathbf{D}_F$的外部文件有關。舉例來說，背景詞儘管在特定查詢的回饋文件出現機率很高，但在其他文件的出現機率一樣很高，這樣詞的特徵便會被背景語言模型$P_{\text{BG}}(w)$給捕捉。另一類詞只在特定查詢的回饋文件的出現機率高，那麼這樣的特徵便被特殊詞的主題模型$P_{\text{SMM}}(w|Q)$捕捉。

2.3 顯著詞模型(Significant Word Model)

這樣的想法啟發自 Luhn's theory (Luhn, 1958)和 SMM，SWM 發現更精確估測查詢語言模型的方法。在 SWM 裡面，假設實際對於查詢有幫助的詞，必須不能是每個文件皆可看到的背景詞，且也不能過於集中出現在少數的回饋文件$\mathbf{D}_F$。因此 SWM 假設回饋文件的語言模型由下列三個模型混合而成，第一個模型為背景語言模型$P_{\mathrm{BG}}(w)$；第二個模型為特殊詞的語言模型$P_{\mathrm{S}}(w|Q)$，以及第三個學習到的顯著詞模型$P_{\mathrm{SW}}(w|Q)$，利用上述三個模型估測回饋文件$\mathbf{D}_F$的公式如下：

$$P(w|D) = \alpha \cdot P_{\mathrm{BG}}(w) + \beta \cdot P_{\mathrm{S}}(w|Q) + (1-\alpha-\beta) \cdot P_{\mathrm{SW}}(w|Q), \tag{3}$$

α和β為可調整的參數，用來決定$P_{\mathrm{BG}}(w)$、$P_{\mathrm{S}}(w|Q)$，以及$P_{\mathrm{SW}}(w|Q)$三者對於語言模型$P(w|D)$的貢獻。$P_{\mathrm{BG}}(w)$為表示常見詞的模型，估測方式是在全部文件集合中，詞的出現次數；$P_{\mathrm{S}}(w|Q)$為表示太特殊的詞的模型，估測方式是在回饋文件$\mathbf{D}_F$中，僅集中出現在少數特定文件的特殊詞。$P_{\mathrm{SW}}(w|Q)$則是既不是常見詞，也不是太特定的詞，因此$P_{\mathrm{SW}}(w|Q)$是利用上述兩個模型，便可在回饋文件$\mathbf{D}_F$中估測出$P_{\mathrm{SW}}(w|Q)$的最大似然機率，並使用最大期望算法(Expectation-Maximum Algorithm) (Dempster, Laird & Rubin, 1977)調整參數。

3. 查詢意向與建模方法 (QUERY INTENT AND MODELING FRAMEWORK)

儘管先前的準相關回饋在資訊檢索和語音資訊檢索上，皆大幅提升原先不足的查詢效能，但卻無法應用在實際的檢索系統上。最大的原因是準相關回饋需要做第二次查詢，消耗過多的計算時間(Manning *et al.*, 2008)。因此針對耗時的問題，我們利用基於類神經網路技術的神經相關感知模型(NRM)的架構。在這樣的架構下，我們不僅能夠有效地重新建構一個更有效的查詢表示式，同時還能解決準相關回饋的耗時問題。

3.1 建立查詢語言模型的相關性 (MODELING RELEVANCE FOR QUERY)

在許多擴增查詢的方法定義了不同的查詢相關性。RM 用了系統性的方法去近似相關性模型；SMM 和 SWM 則是分別利用背景語言模型(Background Language Model)以及額外的特殊詞模型(Specific Word Model)來獲得回饋文件的相關性。在這裡的研究，我們是利用類神經網路的技術來學習上述的建模過程。

更詳細的說，給定一個查詢的集合$\mathbf{Q} = \{Q_1, \cdots, Q_t, \cdots, Q_T\}$，每一個在集合裡的查詢會分別對應到查詢與文件的相關資訊$\mathbf{R} = \{R_1, \cdots, R_t, \cdots, R_T\}$。為了解決查詢的長度不一的問題，我們首先會將每一個查詢用高維度的詞袋模型$P_{Q_t} \in \mathbb{R}^{|V|}$來表示，其中$P_{Q_t}$為出現在對應查詢$Q_t$裡的詞次數，$|V|$則是語料庫的詞典長度。首先將$P_{Q_t}$正規化，讓向量裡的值合計為一，接著再利用編碼器$f(\cdot)$將原始查詢降至低維度空間，如下所示：

$$f(P_{Q_t}) = v_{Q_t} \tag{4}$$

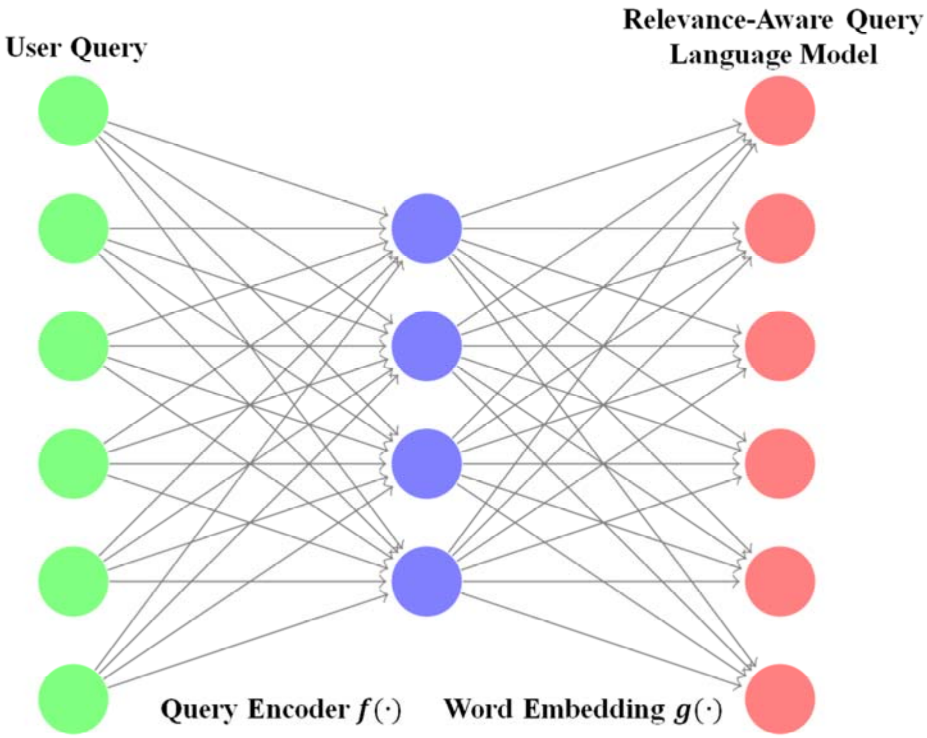

User Query
Relevance-Aware Query Language Model
Query Encoder $f(\cdot)$
Word Embedding $g(\cdot)$

圖 1. 前饋式神經網路的 NRM 模型架構
[Figure 1. NRM framework with feed-forward neural network]

在這篇論文裡面，$f(\cdot)$為全連接的前饋神經網路(feed-forward fully-connected neural network)。以最終要增加檢索效果的查詢表示式為目標，直覺的想法便是先推論一組查詢與詞嵌入(word embedding)的表示方式，再重新構建一個新的查詢語言模型。為了做到這點，我們在$f(\cdot)$之上再堆疊一個解碼器$g(\cdot)$，$g(\cdot)$是一個全連接的前饋式神經網路，權重矩陣可表示成$\mathbf{W} \in \mathbb{R}^{k\times|V|}$。其中$k$為的查詢的大小，$|V|$為詞典的大小。神經相關感知模型(NRM)可表示為以下的式子：

$$P_{\text{NRM}}(w|Q_t) = g\left(f\left(P_{Q_t}\right)\right) = \frac{\exp(v_{Q_t}\cdot v_w)}{\sum_{w'\in V}\exp(v_{Q_t}\cdot v_{w'})} \tag{5}$$

其中v_w是在矩陣$\mathbf{W}$的第w行，也是詞w的詞嵌入表示。最後，為了捕捉到特定查詢的相關性分佈，我們設計了一個訓練目標，最佳解便是在訓練資料上搜尋到一組最大化似然機率的模型參數，調整公式如下：

$$L = \prod_{t=1}^{\mathrm{T}}\prod_{w\in V} P(w|R_t) log P_{\text{NRM}}(w|Q_t) \tag{6}$$

其中$P(w|R_t)$為表示語查詢Q_t對應的相關性分佈。總結這個章節的內容， NRM 的架構主要可分為兩個部分，分別是推斷低維度表示的編碼器$f(\cdot)$和表示相關性的詞嵌入與任意單詞的對應矩陣$\mathbf{W}$，並以最大化似然機率的準則調整模型參數。

3.2 使用者查詢意向 (QUERY INTENT)

為了瞭解使用者的查詢意向(Query Intent)，我們使用基於高維度詞袋模型P_{Q_t}做為分群的依據，使用的演算法為 K-means。首先，我們將P_{Q_t}正規化，讓向量裡的每個值加總為一。接著以查詢詞的出現多寡做為分群的依據，迭代至收斂則停止，這時視為找出使用者意圖(Query Intent)，並當作特定的查詢意向訓練類神經網路$P_{\mathrm{NRM}}(w|Q_t)$。在測試階段便不再分群，而是將測試查詢的詞袋模型與各集群的中心點做相似度計算，以最像的 NRM 預測，其中i為最相似的集群，max相似度的計算以該查詢Q_t和集群中心c_i的 KL 散度距離。可以用下列的式子表示：

$$P_{NRM}(w|Q_t) = \max_{i\in c} P_{NRM_i}(P_{Q_t}) \quad (7)$$

利用使用者查詢意向預測出相關性模型後，再和原始查詢與模型線性組合得到更有鑑別力的查詢模型。詳細組合的式子如下：

$$P_{Q_{t'}} = \alpha P_{Q_t} + (1-\alpha) P_{NRM}(w|Q_t) \quad (8)$$

3.3 實作細節 (IMPLEMENTATION DETAILS)

在語音資訊檢索和資訊檢索的領域裡面，根據使用者的檢索行為，最直覺的定義相關性的方式便是與使用者的資訊需求有關，系統必須從使用者的查詢和語料庫的文件中找出相關性。在訓練階段，為了萃取出查詢與相關文件的相關性分佈(圖 2 的藍色部份)，並用這相關性的分佈來訓練 NRM，我們設計兩個不一樣的情境。第一個情境是監督式的環

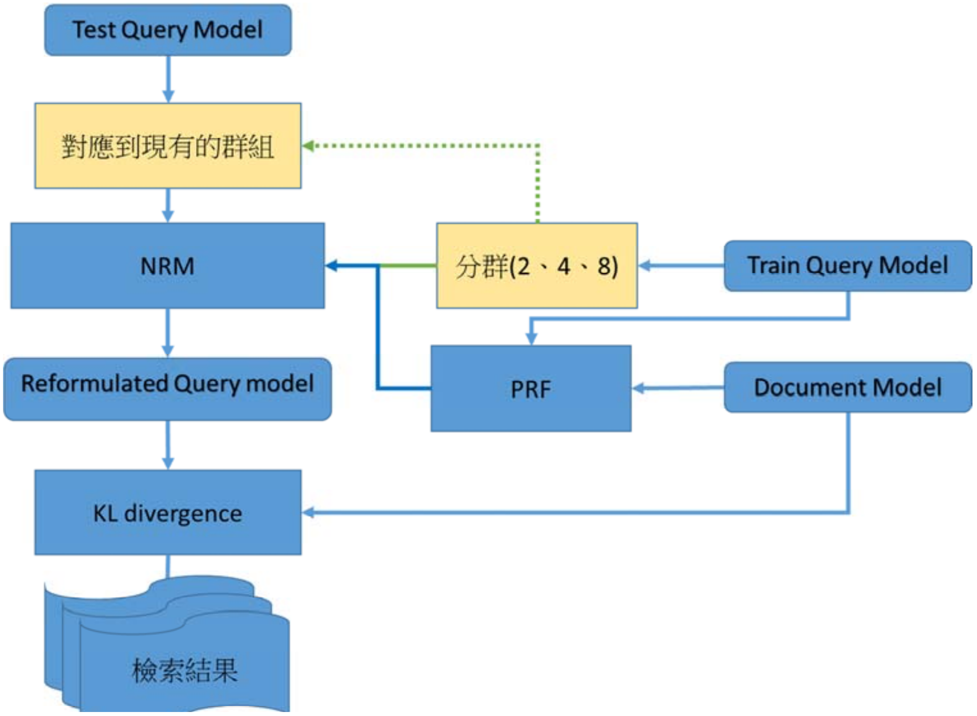


圖 2. 整體架構圖。藍色部份為檢索過程，黃色部分為查詢意向

[Figure 2. Model architecture. The modeling and retrieval is blue block, while yellow part is query intent.]

境，假設已知查詢與相關文件，如給定一個查詢，便有相關文件的正確答案。查詢與文件相關性可以用使用者的看到檢索結果後，點擊相關文件的資訊來代表。儘管點擊資訊可以反映出這篇文件與查詢是否相關，但要收集到這些正確資訊，本身就是一個浩大的工程。因此，我們提出第二種非監督式的情境，原先第一種情境的假設不存在，我們事前不知道相關文件與查詢的對應，只有一個多個查詢構成的集合，與不知道是否相關的文件集合。這樣的情境下，最自然的方法便經過一次查詢，得到相關性的排行後，取出前幾篇當作(準)相關的文件，這樣的策略就如同準相關回饋的方式。經過以上兩個情境的處理後，在訓練的階段，我們便有成對的查詢與相關文件，接著再利用準相關回饋處理這份資料，從中獲得每個查詢的相關性分佈$P(w|R)$，可利用擴展查詢的語言模型建模方法(如，RM、SMM、SWM)獲得。使用在 NRM 架構的神經網路配置如下，隱藏層的激勵函數為線性函數(linear function)，輸出層為 Softmax 函數(softmax function)，其中我們利用 Adam (Kingma & Ba, 2015)演算法尋找最佳解。在測試階段，每一個輸入的查詢皆會被表示為高維度的詞袋模型，經過一層編碼器$f(\cdot)$，將原先高維度表示的向量轉成低維度的詞嵌入向量，並保有原先資料點在高維度的關係，經過一層解碼器$g(\cdot)$，將原先被編碼成低維度表示的詞嵌入向量，依序解碼成 NRM 查詢語言模型，損失函數為交叉熵(Cross Entropy)。取得一個新的查詢語言模型後，我們再利用現有的 KLM 計算 NRM 查詢語言模型與文件語言模型的 KL 散度距離(divergence distance)。總結目前提到的好處，我們利用 NRM 的技術離線學習準相關回饋的建模方式，並在線直接推論新的表示方式。線上查詢時，系統可從原始缺乏使用者的資訊需求的簡短查詢中，重新建構出一個具有準相關回饋效能的的查詢語言模型，也因為耗時的處理過程皆在離線時做完，因此同時也解決耗時的缺點。

最後是建立在 NRM 的基礎之上，針對不同主題的訓練方式(圖 2 的黃色部分)，實作可分為兩個階段，首先是訓練階段，先利用 K-means 演算法將查詢分為 2、4、8 群，視為將查詢分成 2、4、8 個主題類別，接著將每一群當作訓練資料，分別訓練不同的 NRM。第二部分為測試階段，依據在訓練階段分好的 2、4、8 群，測試的查詢以 KL 散度距離計算該歸類在那群，並以該群訓練好的 NRM 模型預測新的查詢表示式，並進行檢索取得相關文件。

4. 實驗設定與結果 (EXPERIMENTS)

4.1 實驗設定 (EXPERIMENTAL SETUP)

我們使用 TDT2 (Topic Detection and Tracking collection)作為實驗數據 (Linguistic Data Consortium, 2000)。來自美國之聲的新聞廣播(Voice of America news broadcasts)的國語新聞(Mandarin news stories)被用來當作語音文件。所有的新聞故事都被標記上特定主題，以此作為評估效能時的相關與否。語音文件的平均詞錯誤率(Word Error Rate, WER)為 35% (Meng *et al.*, 2004)。測試時用來自新華社(Xinhua News Agency)的國語新聞當作檢索的查詢。更精確點來說，我們測試用的查詢可以分為兩個類別，使用新聞內容的長查詢

和使用新聞標題的短查詢。表 1 為 TDT2 一些基本的統計數據。評估檢索效能的方式，我們選用非內差的平均精確度 (non-interpolated mean average precision, MAP) (Manning *et al*., 2008) (Baeza-Yates & Ribeiro-Neto, 2011)作為評判尺度。

表 1. TDT-2 的統計資訊
[Table 1. Statistics of the TDT2 collection]

# 語音文件	2,265 新聞, 46.03 小時的語音錄音			
# 測試用查詢	16 新華社新聞 (Topics 20001~20096)			
	最短	最長	中位數	平均
文件長度	23	4,841	153	287.1
短查詢的長度	8	27	13	14.0
長查詢的長度	183	2,623	329	532.9
# 測試用查詢的相關文件	2	95	13	29.3

4.2 實驗結果 (EXPERIMENTAL RESULTS)

首先， 我們探討 NRM 在已知相關性和未知相關性兩種不同的情境的表現，實驗結果呈現於圖 3 與圖 4。比較的方法為向量空間式的深層結構語意模型(DSSM) (Huang *et al*., 2013)和局部保留本質向量模型(LPEV) (Chen *et al*., 2017)。DSSM 使用點擊資訊當作相關性，訓練網路的參數使相關文件和查詢的似然機率最大化。另一方面，LPEV 則意旨學習一個更好的低維度表示空間，同時保留原始語意結構。

由圖 3 中呈現的數據來看，有幾個比較明顯的結果。首先，DSSM 不論是在人工轉錄還是自動語音轉錄的文件，皆優於 LPEV。原因是在於 DSSM 是利用點擊資訊代表相關性，目的是訓練網路正確地分辨相關和不相關的文件。LPEV 則是意旨在學習一個文件和查詢的特殊的表示空間。目的性的不同，也使得 LPEV 在檢索結果上，比起 DSSM 差強人意。其次，因為在先前的實驗中，SWM 本身的表現普遍比 RM、SMM 較為優異。因此將這樣的方法使用在 NRM 的架構中，其中結合 SWM 和 NRM 的方法，我們將表示為 NRM(SWM)。評估結果正如我們的預期，一樣是 SWM 的表現較為穩定。我們發現在 NRM 的架構之下，不論傳統向量空間模型式的方法，或是經典的語言模型式的方法，NRM 的檢索結果效能普遍都能明顯勝出。有趣的是，NRM 方法甚至能勝過現有的重構查詢的方法，如 RM、SMM、SWM。這樣的好結果或許得歸功於離線的計算過程。相較於 RM、SMM、SWM 利用一次查詢的準回饋，線上做準相關回饋的計算，NRM 在離線時利用點擊資訊，並透過 RM、SMM、SWM 捕捉相關性分佈。我們認為這不同之處，可能讓 NRM 可以學習到有效的特徵,並利用這些特徵在線上即時取得更好的查詢結果，同時也解決耗時的問題。最後，我們可以明顯觀察出，在 NRM 架構之下，不論是那一

個方法，在所有的情況下皆大幅勝出 LPEV，以及在大部分的情況下贏過 DSSM。以上的結果進一步地證明 NRM 在語音文件檢索的有效性。

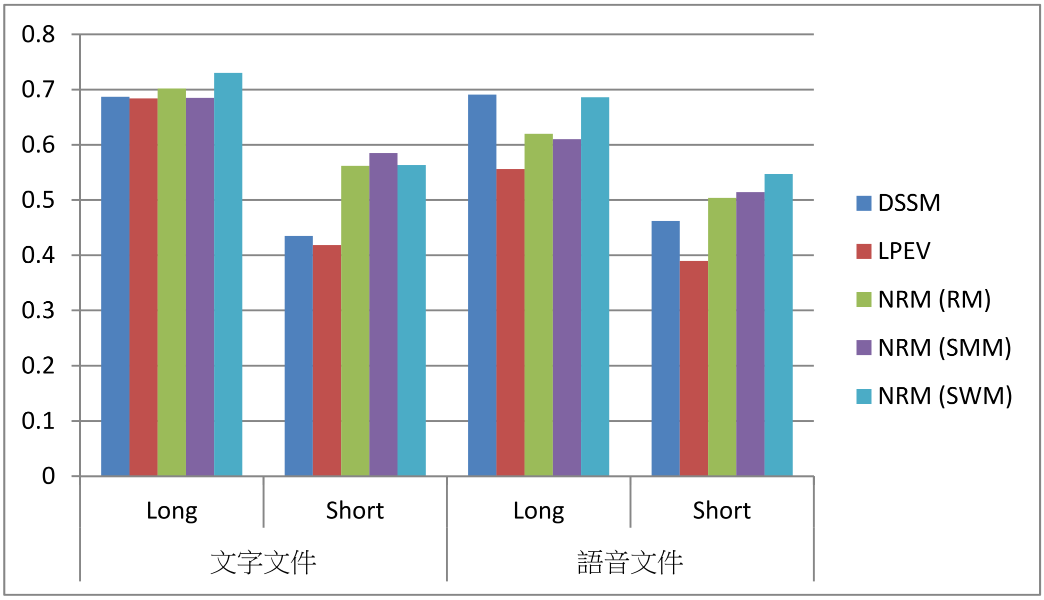


圖 3. 離線時利用點擊資訊訓練的 NRM 模型之檢索結果
[Figure 3. Retrieval results of the NRM offline trained on click-through information.]

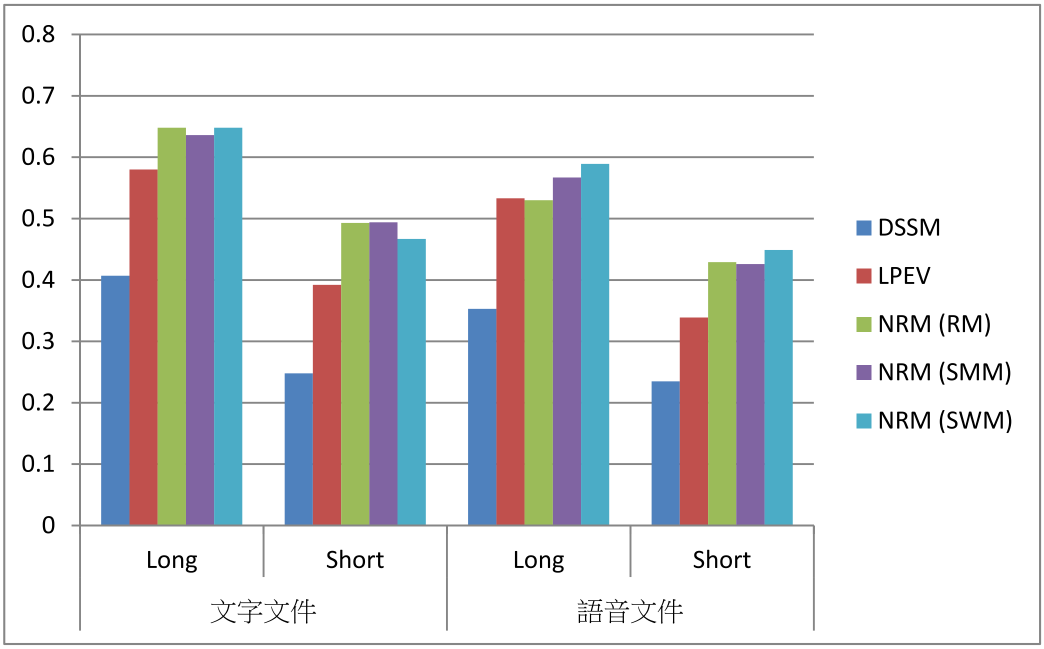


圖 4. 離線時利用準相關回饋訓練的 NRM 模型之檢索結果
[Figure 4. Retrieval results of the NRM trained with pseudo relevance feedback.]

接著是我們假設的第二個情境，查詢與相關文件的關係是未知。在這次的實驗中，我們查詢結果的前 10 篇當作相關文件，因此相關文件未必真的正確。我們可以從結果中觀察到幾個現象。首先，與表 2 不同，LPEV 在這次的實驗中，表現勝過 DSSM。因為 LPEV 的目標與 DSSM 不同，也因此辨識錯誤的相關文件的影響也較小，表現較為穩健。在語音文件的部分，SWM 的表現依舊可以勝過 SMM 和 RM，僅在手寫文本的部分略微下降。最後，整體來看，比較圖 3 與圖 4，這次的實驗普遍表現較差，可以觀察到正確答案對於以上幾個方法的影響，以 NRM 的結果來看，可以發現 NRM 非常依賴相關文件的信息是否正確。最後，即使是在非監督的環境下，NRM 仍然可在各種情境下，表現比 LPEV、DSSM 優異，再一次證明 NRM 的優秀之處。

表 2. 基於以上 NRM 之上，進一步利用使用者意向資訊
[Table 2. Based on NRM framework and further use query intent information.]

	點擊資訊				準相關回饋			
	文字文件		語音文件		文字文件		語音文件	
	Long	Short	Long	Short	Long	Short	Long	Short
NRM SWM	**0.730**	0.563	**0.686**	**0.547**	**0.648**	0.467	0.589	0.449
NRM SWM)-2	0.690	**0.571**	0.670	0.544	0.636	**0.475**	**0.593**	**0.470**
NRM(SWM)-4	0.694	0.562-	0.669	0.545	0.628	0.462	0.583	0.434
NRM(SWM)-8	0.712	0.564-	0.672	**0.547**	0.632	0.463	**0.593**	0.437

分群的實驗部分，我們假設查詢之間是有不同類別的關係，因此建立在原先 NRM 的基礎之上，利用簡單的分群演算法將查詢分群，不同的類別就訓練新的 NRM 模型，期望能達到 NRM 能學習到不同主題的特徵，進一步提升學習的效果。這裡的實驗，我們採用的是在第一個及第二個場景下皆表現較為亮眼的 SWM 方法。首先，我們將訓練資料中的查詢分群，並依據不同的群訓練新的 NRM，測試的長(短)查詢會依據不同歸屬的群，決定使用那個 NRM 預測新的查詢表示式，並以此新的模型結合舊有預測出來的模型(圖 3 與圖 4)線性組合，以此做為新的查詢，實驗結果呈現在表 2。從實驗結果可以看出，不論是那一種的情境，在分群後的訓練結果，大部分的效能都是下降或持平，少數幾個狀況下表現得比原先的結果較好。平均來看，分兩群的效果勝過四群與八群，四群的效能經常落在分兩群與分八群之間，偶而分八群的效果會是最佳的，但卻不會贏過太多。這樣的實驗結果，可以視為 SWM 是較為複雜的語言模型，會根據不同的查詢有不同的影響，所以當我們只將訓練資料分成兩群，對個別訓練的網路來說，可以學習到的訓練資料較為多樣，因此網路較能學習到 SWM 的準相關回饋的特性。反之，切割越多的資料後，讓網路的效能則變得較差。儘管分兩群的效果普遍勝於其他的設定， 但整體的效能比起舊有的模型，大部分仍是持平或退步。

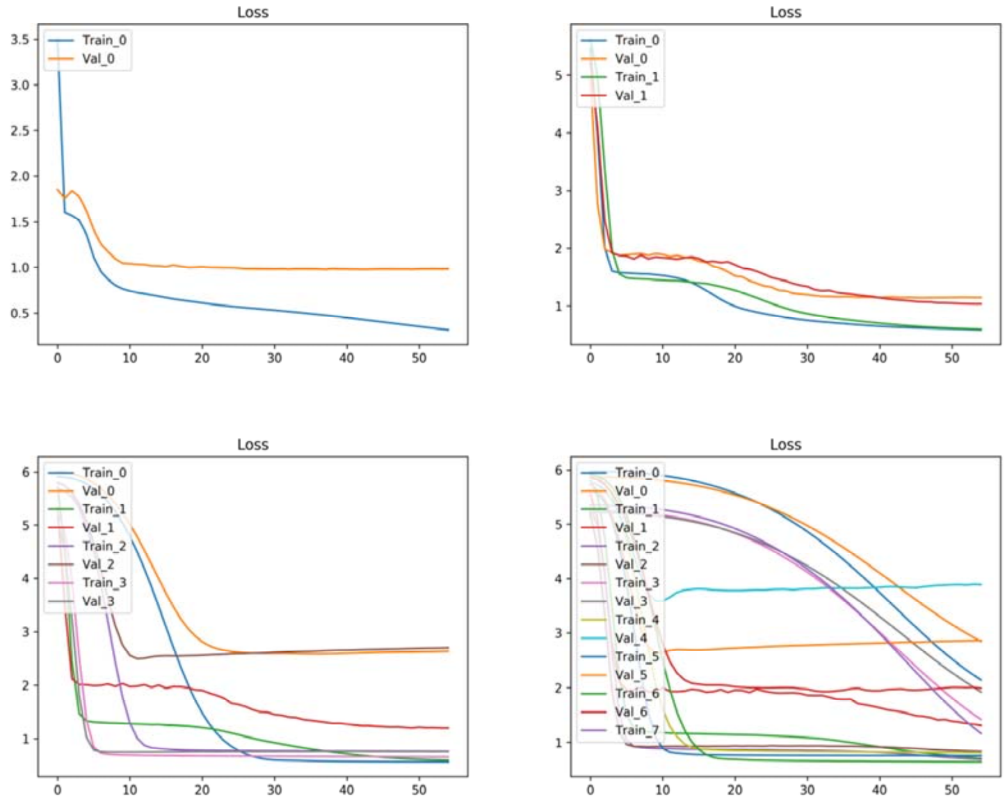


圖 5. 不同集群下，訓練集和驗證集的損失值
[*Figure 5. Loss in training set and developing set with differnent clusters*]

為了進一步研究這次分群對結果的影響，我們將網路訓練的損失值(loss)視覺化，記錄在圖 5。縱軸為損失值，橫軸為迭代次數。不同的線代表不同網路訓練時的兩個值，訓練集(Training Set)和驗證集(Validation Set)的損失值，由左圖到右圖，分別是不分群，以及二、四和八個集群的訓練結果，有些集群只分到一個查詢，集群大小為一就不切成訓練集和驗證集。從上圖中可以看出，分成二個集群的訓練過程較為穩定，訓練集和驗證集的損失值皆為穩定地下降。分成四個集群後，訓練情況有些不同，雖然訓練集的損失值仍是穩定下降，但驗證集的損失值中表現較差，尤其是第零群和第二群。分成八個集群後，這樣的現象就更為明顯，訓練集的損失值依舊穩定下降，但驗證集的損失值則沒有呈穩定的下降曲線，有些集群的損失值甚至有些上升，可以看出明顯地過度訓練(Overfitting)。除了過多的集群可能會造成過度訓練以外，我們也觀察到分越多集群，部分集群的表現(訓練集和驗證集)下降的比原先較少集群的多(如分成四個集群的第三個集群)，但這樣的表現並不是穩定的結果，可能是原先將查詢分群的依據是詞袋模型，過短的查詢造成語意不清，導致分群效果不彰，部分集群沒訓練好，連帶影響整體的成果。儘管這次實驗的結果不如預期，但我們依舊發現，透過簡單的查詢分類，可以更有效地

在 NRM 的架構下訓練模型，損失值可以降得比原先集群較少的狀況下更低(如分成四個集群的第三個集群)。因此用於捕捉使用者意向(Intent)的訓練方式，讓網路得到更多資訊的情況下，能更有效地訓練 NRM。總結先前所做的實驗，以上種種實驗揭示了一些訊息，不論是在語音文件檢索或資訊檢索的領域上，我們提出的 NRM 都可得到更好的結果，與最新穎的語言模型比較起來毫不遜色。

5. 結論 (CONCLUSIONS)

在這篇論文中，我們提出一個建立在 NRM 之上的查詢意向探索方法。在語音文件檢索的任務中，NRM 的方法能夠在不需要繁雜的準相關回饋的處理下，得到一個更有鑑別力的查詢語言模型，大幅提升檢索的效能。實驗的結果也證實，這樣的重構查詢語言模型的技術，比起過往的相關技術，檢索效能皆能穩定的勝出。儘管初步加入查詢意向的結果不盡理想，但實驗結果揭示仍有訓練 NRM 查詢意向的可能。未來的工作，我們希望能夠嘗試一些更複雜的類神經網路模型(如摺積神經網路(CNN)、遞歸神經網路(RNN)等)來作為 NRM 的骨幹。此外，我們也將嘗試加入一些新的特徵(如句法或韻律)，觀察網路獲得更多資訊的情況下能否增益學習效果。最後則是提供更複雜的查詢意向方法，以求更為細緻的查詢結果。

參考文獻 References

Baeza-Yates, R. & Ribeiro-Neto, B. (2011). *Modern information retrieval: the concepts and technology behind search*. Boston, MA: Addison-Wesley Professional.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, *3*(4-5), 993-1022.

Chelba, C., Hazen, T. J. & Saraclar, M. (2008). Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, *25*(3), 39-49. doi: 10.1109/MSP.2008.917992

Chen, B., Chen, K.-Y., Chen, P.-N. & Chen, Y.-W. (2012). Spoken document retrieval with unsupervised query modeling techniques. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(9), 2602-2612. doi: 10.1109/TASL.2012.2208628

Chen, K. Y., Lee, H. S., Wang, H. M., Chen, B. & Chen, H. H. (2014). I-vector Based Language Modeling for Spoken Document Retrieval. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7083-7088. doi: 10.1109/ICASSP.2014.6854974

Chen, K. Y., Liu, S. H., Chen, B. & Wang, H. M. (2016). Learning to Distill: The Essence Vector Modeling Framework. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016)*, 358-368.

Chen, K. Y., Liu, S. H., Chen, B. & Wang, H. M. (2017). A locality-preserving essence vector modeling framework for spoken document retrieval. In *Proceedings of ICASSP 2017*, 5665-5669. doi: 10.1109/ICASSP.2017.7953241

Chen, K. Y., Liu, S. H., Chen, B., Wang, H. M. & Chen, H. H. (2016). Exploring the use of unsupervised query modeling techniques for speech recognition and summarization. *Speech Communication, 80*, 49-59. doi: 10.1016/j.specom.2016.03.006

Chen, Y. W., Chen, K. Y., Wang, H. M. & Chen, B. (2017). Exploring the use of significant words language modeling for spoken document retrieval. In *Proceedings of INTERSPEECH 2017*. doi: 10.21437/Interspeech.2017-612

Croft, W. B. & Lafferty, J. (2003). *Language modeling for information retrieval*. Dordrecht, the Netherlands : Kluwer Academic Publishers. doi: 10.1007/978-94-017-0171-6

Dehghani, M., Azarbonyad, H., Kamps, J., Hiemstra, D. & Marx, M. (2016). Luhn revisited: significant words language models. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM '16)*, 1301-1310. doi: 10.1145/2983323.2983814

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, 39*(1), 1-38.

Guo, J.-F., Fan, Y., Ai, Q. & Croft, W. B. (2016). A deep relevance matching model for ad-hoc retrieval. In *Proceedings of CIKM '16*, 55-64. doi: 10.1145/2983323.2983769

Huang, P. S., He, X., Gao, J., Deng, L., Acero, A. & Heck, L. (2013). Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of CIKM '13*, 2333-2338. doi: 10.1145/2505515.2505665

Huang, C. L., Ma, B., Li, H. & Wu, C.-H. (2011). Speech indexing using semantic context inference. In *Proceedings of INTERSPEECH 2011*, 717-720.

Jones, K. S., Walker, S. & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments (Parts 1 and 2). *Information Processing and Management, 36*(6), 779-840. doi: 10.1016/S0306-4573(00)00015-7

Kingma, D. & Ba, J. (2015). ADAM: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Lavrenko, V. & Croft, W. B. (2001). Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '01)*, 120-127. doi: 10.1145/383952.383972

Linguistic Data Consortium. (2000). Project of Topic Detection and Tracking.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of ICML '14*, 1188-1196.

Lee, L. S. & Chen, B. (2005). Spoken document understanding and organization. *IEEE Signal Processing Magazine*, *22*(5), 42-60. doi: 10.1109/MSP.2005.1511823

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development, 2*(2), 159-165. doi: 10.1147/rd.22.0159

Manning, C. D., Raghavan, P. & Schutze, H. (2008). *Introduction to Information Retrieval*. New York, NY: Cambridge University Press.

Meng, H., Chen, B., Khudanpur, S., Levow, G.-A., Lo, W.-K., Oard, D., …Wang, J. (2004). Mandarin-English information (MEI): investigating translingual speech retrieval. *Computer Speech and Language*, *18*(2), 163-179. doi: 10.1016/j.csl.2003.09.003

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS '13*, 3111-3119.

Mitra, B., Diaz, F. & Craswell, N. (2017). Learning to match using local and distributed representations of text for web search. In *Proceedings of WWW '17*, 1291-1299. doi: 10.1145/3038912.3052579

Ponte, J. M. & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of SIGIR '98*, 275-281. doi: 10.1145/290941.291008

Salton, G., Wong, A. & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620. doi: 10.1145/361219.361220

Shen, D., Pan, R., Sun, J. T., Pan, J. J., Wu, K., Yin, J., & Yang, Q. (2006). Query enrichment for web-query classification. *ACM Transactions on Information Systems (TOIS)*, *24*(3), 320-352. doi: 10.1145/1165774.1165776

Song, F. & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of CIKM '99*, 316-321. doi: 10.1145/319950.320022

Zhai, C. & Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM '01*, 403-410. doi: 10.1145/502585.502654

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 17-30

Question Retrieval with Distributed Representations and Participant Reputation in Community Question Answering

Sam Weng*,+, Chun-Kai Wu#, Yu-Chun Wang‡ and

Richard Tzong-Han Tsai+

Abstract

In recent years, community-based question and answer (CQA) sites have grown rapidly in number and size. These sites represent a valuable source of online knowledge; however, they often suffer from the problem of duplicate questions. The task of question retrieval (QR) aims to find previously answered semantically similar questions in CQA archives. Nevertheless, synony- mous lexical variations pose a big challenge for question retrieval. Some QR approaches address this issue by calculating the probability of correlation between new questions and archived questions. Much recent research has also focused on surface string similarity among questions. In this paper, we propose a method that first builds a continuous bag-of-words (CBoW) model with data from Asus's Republic of Gamers (ROG) forum and then determines the similarity between a given new question and the Q&As in our database. Unlike most other methods, we calculate the similarity between the given question and the archived questions and descriptions separately with two different features. In addition, we factor user reputation into our ranking model. Our experimental results on the ROG forum dataset show that our CBoW model with reputation features outperforms other top methods.

* AsusTek Computer Inc., Taiwan

+ Department of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: thtsai@csie.ncu.edu.tw

Department of Computer Science, National Tsing Hua University, Taiwan

‡ Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts, Taiwan
E-mail: ycwang@dila.edu.tw

The authors for correspondence are Yu-Chun Wang and Richard Tzong-Han Tsai.

Keywords: Question Retrieval, QR, Community-based Question and Answer, CQA

1. Introduction

Over the past decade, there has been a proliferation of online user forums and community question and answer (CQA) sites such as Yahoo! Answers, Quora and Baidu Zhidao. These sites provide a platform for people to discuss questions and solutions to common problems in a wide variety of fields, and they have generated massive amounts of data. Question retrieval (QR) is the task of sorting through this data to find previously answered questions in CQA archives that are similar to a user's current query.

A major challenge for QR is matching the user's query to its lexical variations in the dataset. For example, the system needs to be able to estimate the similarity between synonymous keywords and phrases like "blue screen", "BSOD" and "system crash" that may all refer to the same event. Four main approaches have been proposed to deal with synonyms, such as language model information retrieval (LMIR) (Ponte & Croft, 1998; Song & Croft, 1999; Zhai & Lafferty, 2004), language model with category smoothing (LMC) (Cao, Cong, Cui, Jensen & Zhang, 2009), translation-based language modeling (TBLM) (Xue, Jeon & Croft, 2008), and distributed-representation- based language modeling (DRLM) (Mikolov, Chen, Corrado & Dean, 2013).

Language model information retrieval (LMIR) estimates probabilities of word sequences between query and candidate question. Another approach, language model with category smoothing (LMC), represents each question category as a dimension in a vector space. In both LMIR and LMC, words are represented as indexed in a vocabulary, and similarity of words is ignored. Still another approach is translation-based language modeling (TBLM), which uses QA pairs to learn semantically related words to improve traditional IR models. The basic assumption is that QA pairs are parallel texts and the relationship of words can be established through word-to-word translation probability. In practical use, however, TBLM may take too long to learn a translation table. Finally, distributed-representation-based language modeling (DRLM) uses distributed representations of data to replace the word-to-word translation probability in TBLM with the probability calculated using word2vector. DRLM further combines the similarity of a word vector and a category vector as the final retrieval model.

In this paper, we propose a method that first builds a continuous bag-of-word (CBoW) model with data from the Asus Republic of Gamers (ROG) forum and then determines the similarity between a given new question and the Q&As in our database. Unlike most other studies, we calculate the similarity between the given question and the archived questions and descriptions separately with two different features. In addition, we factor user reputation into our ranking model. Our experimental results on ROG forum dataset show that our CBoW model with reputation features outperforms other top methods.

2. Related Work

In this section, we offer an overview of existing community-based question retrieval models.

2.1 Language Model for Information Retrieval (LMIR)

In recent years, LMIRs and their extensions have been widely used for question retrieval on community Q&A data (Ponte & Croft, 1998; Song & Croft, 1999; Zhai & Lafferty, 2004). It is a statistical way to estimate the probabilities of word sequences between query and candidate questions. Measurement can be expensive, since sentences can be arbitrarily long and the size of a corpus needs to be very large. In practice, the statistical language models are often approximated by N-gram models. The unigram model makes a strong assumption that every single word occurs independently, and consequently, the probability of a word sequence becomes the product of probabilities of the individual words. The bigram and trigram models take the local context into consideration. As for the bigram, the probability of a new word depends on the probability of the previous word. While for a trigram, the probability of a new word depends on the probabilities of the previous two words. The basic language modeling approach (unigram language model) has performed quite well empirically in several information retrieval tasks (Ponte & Croft, 1998; Song & Croft, 1999; Zhai & Lafferty, 2004) and has also performed quite well in question search (Zhai & Lafferty, 2004). The basic idea is to estimate a language model for each query and then rank candidates by likelihood of the query according to the estimated model. Given a query q and a candidate Q, the ranking function is as follows:

$$P(q|Q) = \sum_{w\in q}(1-\lambda)P_{ml}(w|Q) + \lambda\, P_{ml}(w|C) \tag{1}$$

Where q is the queried question and w is a word in it. Q is an archived question and C is whole data collection. $P_{ml}(w|Q)$ presents the maximum likelihood estimated of w in Q. $P_{ml}(w|C)$ is a smoothing item which is calculated as the maximum likelihood in a large corpus C. The smoothing item avoids zero probability when the words appear in q but not in Q. λ is a parameter ranging from 0.0 to 1.0.

2.2 Language Model with Category Smoothing (LMC)

On most community question and answer sites, each question belongs to one or several categories by askers' tagging actions. Category information of archived questions is utilized such that category-specific frequent words will play an important role in comparing the relevancy of archived questions across categories to a query (Cao *et al.*, 2009). Instead of finding patterns among individual words, a language model may be designed to discover relationships among word groupings or categories. This idea can be realized as follows: the category language model is first smoothed with the whole question collection, and then the question language model is smoothed with the category model. To utilize category information, LMC expands LMIR with a new smoothing value estimated from questions under the same category. Given a user search question q and a candidate Q, LMC is described as follows:

$$P(q|Q) = \sum_{w\in q}(1-\lambda)P_{ml}(w|Q) + \lambda P_s(w|C) \tag{2}$$

$$P_s(w|Q) = (1-\beta)P_{ml}(w|C) + \beta P_{ml}(w|Cat(Q)) \tag{3}$$

In this, w means a word in the query question q, and λ and β are two different smoothing parameters. $Cat(Q)$ denotes the category of the candidate question Q, which is usually a root or a leaf category in the category hierarchy of a community Q&A site. $P_{ml}(w|Q)$ is the maximum likelihood estimate of w in Q. $P_{ml}(w|Cat(Q))$ means the maximum likelihood estimate for w in $Cat(Q)$. $P_{ml}(w|C)$ represents the maximum likelihood estimate of w in a large corpus C. λ and β both range from 0.0 to 1.0.

2.3 Translation Model (TM)

The idea of statistical machine translation was first introduced by Warren Weaver in 1949. The basic idea is based on a string-to-string noisy channel model. The channel converts a sequence of words from one language (such as Spanish) into another (such as Chinese) according to the probability distribution. The channel operations are movements, duplications and translations, applied to each word independently. The movement is conditioned only on word classes and positions in the string, and the duplication and translation are conditioned only on the word identity. Statistical translation models were initially word-based (Models 1-5 from IBM hid- den Markov model from Stephan Voge and Model 6 from Franz-Joseph Och), but significant advances were made with the introduction of phrase-based models. In word-based translation, the fundamental unit of translation is a word in natural language. Previous work (Berger, Caruana, Cohn, Freitag & Mittal, 2000; Xue *et al*., 2008) consistently reported that word-based translation models yielded better performance than traditional methods (such as language model) for question retrieval. These models exploited a modeling framework. The ranking function can be written as follows:

$$P(q|Q) = \prod_{w\in Q}(1-\lambda)P_{tr}(w|Q) + \lambda P_s(w|Q) \tag{4}$$

$$P_{tr}(w|Q) = \sum_{t\in Q} P(w|t)P_{ml}(t|Q) \tag{5}$$

Where $P_{ml}(w|C)$ and $P_{ml}(t|Q)$ can be estimated similarly as in the language model above, $P(w|t)$ denotes the translation probability that w is a translation of t, and it is assumed that the probability of self-translation is 1, meaning that $P(w|t) = 1$.

2.4 Translation-Based Language Model (TBLM)

A recent approach to question retrieval is the translation-based language model (TBLM) (Xue *et al.*, 2008), which combines LMIR and TM. It has been shown that this model achieves better performance than both LMIR and TM. TBLM uses word-to-word translation probabilities estimated from questions to find semantically similar questions. The TBLM ranking score is computed as follows:

$$P(q|Q) = \prod_{w\in q}(1-\lambda)P_{mx}(w|Q) + \lambda P_{ml}(w|C) \tag{6}$$

$$P_{mx}(w|Q) = (1-\beta)P_{ml}(w|Q) + \beta P_{tr}(w|Q) \tag{7}$$

$$P_{tr}(w|Q) = \sum_{v\in Q} P_{tp}(w|v)P_{ml}(v|Q) \tag{8}$$

Where λ and β are two different smoothing parameters controlling the translation component's impact, and $P_{tp}(w|v)$ is the translation probability from word w in query question to word v in historical candidate question Q. The difference between TM and TBLM is that TBLM calculates with one extra element $(1-\beta)P_{ml}(w|Q)$.

2.5 Word Embedding Learning

A distributed representation (word embedding) (Mikolov *et al.*, 2013) stores the same contextual information in a low-dimensional vector. Every word is now represented by a *D* dimensional vector, where *D* is a relatively small number (usually between 50 and 1000). Each dimension of the embedding corresponds to a semantic or grammatical attribute of the words. The hope is that similar words get to closer to each other in that space. In place of counting word co-occurrences, the vectors can be learned.

The basic algorithm starts from a random vector for each word in the vocabulary. It then crosses a large corpus, and at each step, observes a target word and its context. The vector of the target word and the context word will be updated to bring them close together in the vector space, thus increasing the similarity between them. Other vectors will be updated to become more distant from the target word. After the processing, the vectors become meaningful, representing similar words with similar vectors. The advantage of word embedding is that it allows the model to generalize sequences that do not appear in the set of training data but are similar in terms of their features.

3. Approach

In this section, we will describe the proposed approach consisting of three parts: (1) word embedding learning: given a forum data collection, questions are treated as basic units. Each word in a question is transformed into a word vector. (2) score generation: once the word vectors are learned, question retrieval can be performed by calculating the similarity between a query question and a candidate question. (3) utilizing reputation information: we enhance the ranking function by introducing reputation points of each archived question's participants.

3.1 Word2vec

Word2vec is an open-source software program that was created by a team of researchers led by Tomas Mikolov at Google[1]. It is a group of related models that are used to produce word embeddings. This tool provides an efficient implementation of the continuous bag-of-words (CBoW) and skip-gram architectures for computing vector representation of words. Using the CBoW architecture, the model predicts the current word by using the context words. The order of context words does not influence prediction. The input could be w_{i-2}, w_{i-2}, w_{i+1}, w_{i+2}, the previous words and the following words of the current word w_i. With the skip-gram architecture, the model uses the current word to predict the context words. The input is w_i and the output would be w_{i-2}, w_{i-2}, w_{i+1}, w_{i+2}. Furthermore, the context words are not limited to the immediate context. Training instances could be created by skipping a constant number of words in w_i's context–for instance, w_{i-4}, w_{i-3}, w_{i+3}, w_{i+4}.

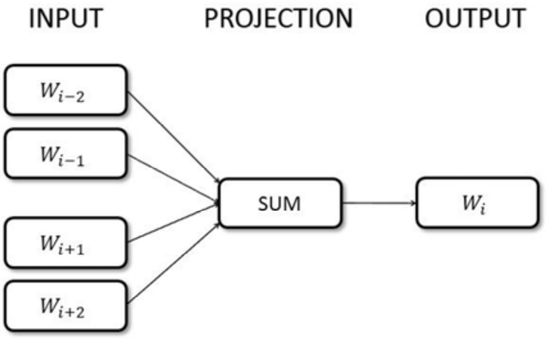


Figure 1. CBoW model & skip-gram model.

According to Mikolov *et al.* (Mikolov *et al.*, 2013) and some previous studies (Bansal, Gimpel & Livescu, 2014), the CBoW model performs better in text classification, especially suitable for documents containing very few infrequent words. In addition, training the CBoW model is much faster than the skip-gram model. Based on our preliminary analysis, our ROG forum dataset is composed of 421 thousand posts, most of the posts contain very few infrequent words. We decide to adopt the CBoW model.

3.2 Ranking Function for Question Title and Description

Once the word embedding is learned, questions can be represented by word vectors. Semantic similarities between query questions and archived questions represented by CBoW are

[1] https://code.google.com/archive/p/word2vec/

believed to be more accurate. We calculate *q*'s vector as follows:

$$V_{sen}(q) = \frac{1}{L_{qv}} \sum_{w \in q} v(w) \tag{9}$$

$$v(w) = \begin{cases} V_{cbow}(w), & w \in |V_{cbow}| \\ \text{NULL}, & \text{otherwise} \end{cases} \tag{10}$$

$$L_{qv} = \sqrt{Len_v} \tag{11}$$

$$Len_v = \sum_{e \in V_{sen}(q)} e^2 \tag{12}$$

Where *w* is each word in question *q*, we retrieve the vector of *w* in the training vocabulary. *e* is the value of each dimension in the vector. After getting the V_{sen}, we can calculate the similarity score by this method:

$$S(q, Q) = V_{sen}(q) \cdot V_{sen}(Q) = \sum_{i=1}^{D} e(q_i) \cdot e(Q_i) \tag{13}$$

Here *D* is the dimension size of the sentence vector. $e(q_i)$ and $e(Q_i)$ are the values of each dimension in the query question *q* and archived question *Q*. In our study, we treat the title and description fields of a forum question as two different parts. Several previous approaches such as (Zhang, Wu, Wang, Zhou & Li, 2016; Zhou, He, Zhao & Hu, 2015) combine title and description into one. Ideally, users should describe their main question in the title field and write a more detailed situation in the description field. Often, people write something with no clear connection to their problem in the title field, such as "Need Help!", "Not Happy" or "Error Code". From these kinds of titles, it is hard to interpret what the user's true question is. Even if the problem is clearly depicted in the description yet the title is unclear, combining a meaningless title with a particular description might actually lower the ranking score. Based on the facts mentioned above, we propose a prototype ranking function to measure title and description scores separately as follows:

$$R(q, Q) = \alpha \times S_{title}(q, Q) + \beta \times S_{desc}(q, Q) \tag{14}$$

Where $S_{title}(q,Q)$ is the score of the title and $S_{desc}(q,Q)$ is the score of the description between input question *q* and archived question *Q*. α and β are both free tuning parameters for finding the balance between title and description. Here we let $\alpha + \beta = 1$.

3.3 Utilizing User Reputation in The Forum

User reputation in its simplest form is a ranking of how the community scores a user's contributions to the forum. A user's reputation is given by other forum participants who read the user's posts. Positive reputation should be given to people whose posts are meaningful, helpful and thoughtful. Negative reputation should be given to users posting something that detracts from the conversation. So we improve the ranking function with an extra element:

reputation of participants. The new measurement is described as follows:

$$R(q,Q) = \alpha \times S_{title}(q,Q) + \beta \times S_{desc}(q,Q) + \gamma \times RPU(Q) \quad (15)$$

$$RPU(Q) = \frac{1}{\#u}\sum_{u\in Q} RP(u) \quad (16)$$

In this formula, we extend the function from above and add the reputation point. γ is a tuning parameter for reputation. $RPU(Q)$ is the summation of the reputation points of the users participating in the discussion of Q. Any one of the participants may post several answers in the same thread. To avoid too many reputation points from the same forum user, we only add each participant's reputation point once. To ensure fairness for newer post, we average the reputation point by the number of participants. Here we let $\alpha + \beta + \gamma = 1$. The reputation system of the ASUS ROG forum offers all participants a fairer and equal platform. Each registered user can anonymously offer points to anyone who posts an appropriate and useful answer under a discussion thread. There is only one way to gain points, when someone approves of the post. This is much more objective so we think it is a suitable factor to evaluate candidate questions.

4. Experiment and Evaluation

In this chapter, we present experiments to evaluate the performance of the proposed approach for question retrieval.

4.1 Data Sets

We collect data sets from the official ASUS Republic of Gamers discussion forum[2]. Unlike the general questions on other community sites, people discuss PC-related technical topics on the ROG forum such as overclocking, tweaking and cooling. For our experimental dataset, we extracted 42,899 threads and 420,983 posts archived in the ROG forum. Each thread consists of a title, a description and the discussion of the participants. For question retrieval, we look at not only titles and descriptions fields but also the reputation of participants.

4.2 Validation Set and Test Set

We assume that the title and description of threads already provide enough information for users to understand. We created a test set from the ROG forum by using the Lucene search engine with the default and BM25 similarity scoring functions to index all data from the ROG forum. All questions are stemmed and lowercased. Stopwords, HTML and forum tags are also removed. We randomly choose questions from the database with title length greater than 25 characters so that the title would be more likely to be meaningful. Also, the same criterion is

[2] http://rog.asus.com/forum

applied to the query questions. Then we retrieved 10 candidate questions from the corresponding indexed data using default and BM25 similarity ranking algorithms in Lucene. After retrieval, we labeled the relevance for the candidate questions regarding to the input queries. If a candidate question is considered semantically similar to the query, it will be labeled as relevant; otherwise it will be labeled as irrelevant. We use the labeled dataset of default similarity as the validation set and the dataset of BM25 as the test set. The validation set is used for tuning parameters of different models, whereas the test set is used for evaluating how well the models rank relevant candidates and irrelevant candidates.

4.3 Word2vec Training

In our experiments, we trained word embedding with a whole discussion dataset from the ROG forum site. Before training word embedding, some pre-processing was executed. Each character was converted to lowercase. Forum tag language, redundant spaces and duplicate symbols were removed. Finally, every word was stemmed and stopwords were removed. Here, we trained the word embedding by using the CBoW method. The parameters we set for training are as follows: 200 dimensions for the size of word vectors, and a max skip length between words of 8.

4.4 Baselines

In this paper, we implement several methods to be the baseline for comparison.

4.4.1 Language Model for Information Retrieval (LMIR)

$$P(q|Q) = \prod_{w \in q} (1-\lambda) P_{ml}(w|Q) + \lambda P_{ml}(w|C) \tag{17}$$

LMIR (Ponte & Croft, 1998; Zhai & Lafferty, 2004) is based on the probability of each word in query question *q* that appears in candidate question *Q* and the large collection *C*.

4.4.2 Language Model with Category Smoothing (LMC)

$$P(q|Q) = \prod_{w \in q} (1-\lambda) P_{ml}(w|Q) + \lambda P_s(w|Q) \tag{18}$$

$$P_s(w|Q) = (1-\beta) P_{ml}(w|C) + \beta P_{ml}(w|Cat(q)) \tag{19}$$

LMC (Cao *et al.*, 2009) extends LMIR by introducing the probability of each word in *q* that appears in the category of candidate *Q*.

4.4.3 Distributed Representation Based Language Model (DRLM)

The last compared configuration employs DRLM (Zhang *et al.*, 2016), which considers creating a retrieval model with learned representations of words. It borrows the idea from

TBLMs and incorporates word-to-word similarity calculated with the learned vectors into LMIR. The model finds the top N similar words for each word with Cosine similarity and defines a word-to-word similarity function as:

$$P_{sim}(w_i|w_j) = \begin{cases} \frac{e^{v(w_i)\cdot v(w_j)}}{\sum_{w' \in Sim(w_j)} e^{v(w')\cdot v(w_j)}}, & \text{if } w_i \in Sim(w_j) \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

Here, $Sim(w_j)$ represents the top N similar words of w_j, and $P_{sim}(w_i|w_j)$ is the translation probability from w_i to w_j. The idea is to replace the translation probability $P_{tp}(w|v)$ in TBLM with $P_{sim}(w_i|w_j)$. DRLM is also combined with a word-category similarity function, which is defined as:

$$Scat(w|c) = \frac{e^{v(w)\cdot v(c)}}{\sum_{w' \in V} e^{v(w')\cdot v(c)}} \tag{21}$$

Therefore, given a query question q and a candidate question Q, the DRLM retrieval model can be represented as follows:

$$P(q|Q) = \prod_{w \in q} (1-\lambda) P_{mx}(w|Q) + \lambda P_s(w|Q) \tag{22}$$

$$P_{mx}(w|Q) = (1-\alpha) P_{ml}(w|Q) + \alpha P_{sim}(w|Q) \tag{23}$$

$$P_s(w|Q) = (1-\beta) P_{ml}(w|C) + \beta Scat(w|Cat(q)) \tag{24}$$

$$P_{sim}(w|Q) = \sum_{v \in Q} P_{sim}(w|v) P_{ml}(v|Q) \tag{25}$$

4.5 Evaluation Metrics

In order to evaluate the performance of different models, we used mean average precision (MAP), and precision at K (P@3, P@1) as evaluation measures. These measures are widely used in the literature for question retrieval in community-based Q&A.

4.6 Main Results

In this section, we present the experimental results on our test sets of the ROG forum data. We compare Lucene, LMIR, LMC and DRLM_nocat against our approach. The number of dimensions of word2vec training is set to 200. We have implemented LMIR, LMC and DRLM models based on the original papers and set all the tuning parameters on our dataset. Table 2 shows the best tuning parameters of title, description and reputation for each approach.

Table 1 shows question retrieval performance in terms of different evaluation metrics. DRLM_nocat is better than Lucene except on P@1. By using only title similarity (Forum- T) or content similarity (Forum-C), our system obtains a comparative score to those of other state-of-the-art methods. After using both title and content scores (Forum-TC), our system performs better than DRLM_nocat. This indicates that considering titles and descriptions separately improves accuracy of similarity scores between questions. We also test our methods with Wiki trained data. In Wiki-T, we use only title score as in Forum-T. In Wiki-TC, we use both scores of title and description as in Forum-TC. Table 3 shows that Wiki performs the worst, indicating that in-domain training data is more effective than out-of-domain training data for word2vec training. Finally, we can see that Forum-TCR outperforms all other methods. It takes advantage of Forum-TC and participants' reputation.

4.7 Positive and Error Cases

One of the positive cases is the input query is "Fan Xpert 2 issues: access violation at address 0040b590...". After our ranking function, the 10th question: "Fan Xpert II Problem" is raised to be the first one. Because both their descriptions describe they can't start the application normally. The error case is like that the input is "Maximus V Extreme fans running after shutdown.", and the ranking dropped the first result: "PC doesn't power off when shutdown" to 9th. We found both questions said the fans are still spinning after PC shutting down. But the archived question does not mention this in its title. So our system gives high description score but low title score for the correct archived question.

Table 1. Performance of the state-of-the-art methods and our proposed methods.

	MAP	P@3	P@1
Lucene	0.423	0.272	0.408
LMIR	0.446	0.333	0.408
LMC	0.446	0.333	0.408
DRLM_nocat	0.468	0.340	0.398
LMIR TC	0.449	0.344	0.439
LMC TC	0.447	0.337	0.439
DRLM_nocat TC	0.456	0.34	0.459
Forum-T	0.441	0.323	0.418
Forum-C	0.473	0.354	0.408
Forum-TC	0.487	0.354	0.439
Forum-TCR	0.507	0.367	0.510

Table 2. Best parameters.

	Title	Description	Reputation
LMIR TC, LMC TC, DRLM_nocat TC	0.3	0.7	N/A
Forum-TC	0.2	0.8	N/A
Forum-TCR	0.4	0.5	0.1

Table 3. Comparison of using the in-domain word2vec and the out-domain word2vec.

	MAP	P@3	P@1
Lucene	0.423	0.272	0.408
Wiki-T	0.363	0.262	0.286
Wiki-TC	0.369	0.276	0.265

5. Conclusion

This paper proposes to learn vector representation for question retrieval in community forums and to exploit participant reputation to improve retrieval. We believe that title and description fields should be analyzed separately. Unlike baseline methods, our approach calculates the similarity between the query question and each archived question’s title as well as the similarity between the query question and each archived question’s description. As mentioned above, we create a retrieval model which combines user reputation and the learned word embedding representation of title and description. Evaluation results on our ROG forum dataset indicate that our proposed approach can dramatically enhance cQA question retrieval.

Acknowledgement

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST 106-2633-E-002-001), National Taiwan University (NTU-106R104045), Intel Corporation, and Delta Electronics, and Advantech.

Reference

Bansal, M., Gimpel, K. & Livescu, K. (2014). Tailoring Continuous Word Representations for Dependency Parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics(ACL)*, *2*, 809-815. doi: 10.3115/v1/P14-2131

Berger, A., Caruana, R., Cohn, D., Freitag, D. & Mittal, V. (2000). Bridging the lexical chasm: statistical approaches to answer-finding. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR '00)*, 192-199. doi: 10.1145/345508.345576

Cao, X., Cong, G., Cui, B., Jensen, C. S. & Zhang, C. (2009). The use of categorization information in language models for question retrieval. In *Proceedings of the 18th ACM conference on Information and knowledge management(CIKM '09)*, 265-274. doi: 10.1145/1645953.1645989

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrived from arXiv preprint arXiv:1301.3781

Ponte, J. M. & Croft, W. B. (1998). A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR '98),* 275-281. doi: 10.1145/290941.291008

Song, F. & Croft, W. B. (1999). A general language model for information retrieval. In *Proceedings of the eighth international conference on Information and knowledge management(CIKM '99)*, 316-32. doi: 10.1145/319950.320022

Xue, X., Jeon, J. & Croft, W. B. (2008). Retrieval models for question and answer archives. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval(SIGIR '08)*, 475-482. doi: 10.1145/1390334.1390416

Zhai, C. & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, *22*(2), 179-214. doi: 10.1145/984321.984322

Zhang, K., Wu, W., Wang, F., Zhou, M. & Li, Z. (2016). Learning Distributed Representations of Data in Community Question Answering for Question Retrieval. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining(WSDM '16)*, 533-542. doi: 10.1145/2835776.2835786

Zhou, G., He, T., Zhao, J. & Hu, P. (2015). Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 250-259.

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 31-46

探究使用基於類神經網路之特徵於文本可讀性分類

Exploring the Use of Neural Network based Features for Text Readability Classification

曾厚強*、陳柏琳*、宋曜廷+

Hou-Chiang Tseng, Berlin Chen and Yao-Ting Sung

摘要

可讀性通常指的是閱讀題材可以被讀者理解的程度：當閱讀材料愈能夠被讀者所理解時，就愈能夠產生好的學習效果。為了能夠幫助讀者去適配符合自己閱讀能力的文件，研究人員長久以來持續發展各種能夠自動且精準地估測文本可讀性的模型來達到此目標。可讀性分類通常是透過分析文件上的資訊來轉化成一組可讀性特徵，再利用這些可讀性特徵來訓練出可讀性模型，以便能預測未知文件的可讀性。然而，傳統的可讀性模型所使用的特徵都需要根據專家的經驗來進行選取，這卻也限制其實用性。近年來隨著表示學習法技術的蓬勃發展，訓練可讀性模型所需要的特徵可以不再需要仰賴專家，這也使得可讀性模型的發展有了一個嶄新的研究方向。因此，本論文嘗試以卷積神經網路以及快速文本兩種技術分別來自動地擷取文本特徵，以訓練出一個能夠分析跨領域文件的可讀性模型，並可以因應文件內容多元主題的特性。經與現有方法的一系列實驗比較後，其結果確認了本論文所提可讀性模型的效能優勢。

關鍵字: 可讀性、詞向量、卷積神經網路、表示學習法、快速文本。

Abstract

Text readability refers to the degree to which a text can be understood by its readers: the higher the readability of a text for readers, the better the the

* 國立臺灣師範大學資訊工程系
Department of Computer Science & Information Engineering, National Taiwan Normal University
E-mail: ouartz99@gmail.com, berlin@cise.ntnu.edu.tw

+ 國立臺灣師範大學教育心理與輔導學系
Department of Educational Psychology and Counseling, National Taiwan Normal University
E-mail: sungtc@ntnu.edu.tw

comprehension and learning retention can be achieved. In order to facilitate readers to digest and comprehend documents, researchers have long been developing readability models that can automatically and accurately estimate text readability. Conventional approaches to readability classification is to infer a readability model using a set of handcrafted features defined a priori and computed from the training documents, along with the readability levels of these documents. However, the use of handcrafted features requires special expertise and its applicability also is limited. With the recent advance of representation learning techniques, we can efficiently extract salient features from dcouments without recourse to specialized expertise, which offers a promising avenue of research on readability classification. In view of this, we in this paper propose two novel readability models built on top of a convolutional neural network based representation and the so-called fastText representation, respectively, which have the capability of effectively analyzing documents belonging to different domains and covering a wide variety of topics. A series of emperical experiments seem to demonstrate the utility of the proposed models in relation to several existing methods.

Keywords: Readability, Word Vector, Convolutional Neural Network, Representation Learning, fastText.

1. 緒論 (Introduction)

一般而言，可讀性(Readability)是指閱讀材料能夠被讀者所理解的程度(Dale & Chall, 1949; Klare, 1963, 2000; Mc Laughlin, 1969)，當讀者閱讀高可讀性的文件時，會產生較好的理解及學後保留效果(Klare, 1963, 2000)。由於文件的可讀性在知識傳遞扮演極為重要的角色，因此西方的可讀性公式發展的非常早，如：1923 年 Bertha 等人就提出方法來探討教科書中字彙難度的問題(Bertha & Pressey, 1923)。另外，Vogel 和 Washburne 在 1928 年則是提出一個 Winnetka Formula 來評量小孩讀物的可讀性(Vogel & Washburne, 1928)。由於可讀性相關的研究非常重要。因此，據 Chall 與 Dale 在 1995 年的統計，到 1980 年為止相關的可讀性公式就已經超過 200 多個可讀性公式(Chall & Dale, 1995)。這些傳統的可讀性研究大多使用較淺層的語言特徵來發展線性的可讀性公式，例如 Flesch Reading Ease 採用詞彙的平均音節數與平均的句子長度(Flesch, 1948)或 Chall 和 Dale 計算難詞在文章中的比率(Chall & Dale, 1995)等，都是傳統可讀性公式代表之一。然而，傳統可讀性公式所採用的淺層語言特徵，並不足以反映文件難度。Graesser、Singer 和 Trabasso 便指出，傳統語言特徵公式無法反映閱讀的真實歷程，文件的語意語法只是文件的淺層語言特徵，沒有考量文件的凝聚特性(Graesser, Singer & Trabasso, 1994)。Collins-Thompson 亦指出傳統可讀性公式僅著重在文件的表淺資訊，而忽略文件重要的深層特徵。這也讓傳統可讀性公式在預測文本可讀性的結果常遭受到質疑(Collins-Thompson, 2014)。直到今日，可讀性的研究仍持續不斷。研究人員為了克服傳統可讀性公式的缺點，嘗試利用更細緻的

機器學習演算法來發展出非線性的可讀性模型，並納入更多元的可讀性指標來共同評量文本的可讀性，以提升可讀性模型的效能(Petersen & Ostendorf, 2009; Feng, Jansche, Huenerfauth & Elhadad, 2010; Sung *et al*., 2015)。

然而可惜的是，研究人員發現採用一般語言特徵的可讀性模型在應用於特定領域文時，一般語言特徵並無法判斷詞彙在不同領域文本時背後所代表的意義。其原因在於特定領域文本的內容著重在闡述領域的「知識概念」，而這樣子的描述方式有別於一般語文的敘述文或故事體的結構。Yan 等人就明確指出在計算美國大型醫學資料庫(Medical Subject Headings, MeSH)中的專業術語去探討，發現語言特徵公式的音節數、字長與醫學類專業詞彙的困難度無相關。換句話說，採用淺層語言特徵的可讀性模型並無法反映特定領域文件中專業術語的難度(Yan, Song & Li, 2006)。針對一般語言特徵無法表徵特定領域知識結構的問題，開始有學者針對這個議題進行研究。例如，Yan 等人利用本體論的技術將美國國家醫學資料庫(Medical Subject Headings, MeSH)的醫學符號階層資料庫作為概念資料庫，從中找出每一個醫學類文件中的概念，並計算概念到此樹狀結構最底部的距離，得出每篇文件概念深度指標(Document Scope) (Yan *et al*., 2006)。Borst 等人則是利用詞表的方式將每個詞彙的「類別複雜度」與「詞頻」兩個分數加總來計算詞彙複雜度，作為評估醫學類線上文件詞彙、句子及文件難度的依據(Borst, Gaudinat, Grabar & Boyer, 2008)。

由上述的研究可知，不論是過去一般語言特徵或是針對特定領域文本的知識結構所設計的文件表示(Document Representation)技術，長久以來都需要仰賴專家來研發，有著耗時費力等問題。近年來，有所謂表徵學習(Representation Learning)方法可以自動從原始資料中去抽取有用的資訊，能有助於建立分類模型和預測測試資料(Goodfellow, Bengio & Courville, 2016)。使得模型所需要的特徵可以逐漸不需仰賴專家，成功開啟了另一個研究的方向。因此本研究將基於近年來熱門的表徵學習法來自動從文本中抽取出可讀性模型所需要的特徵，訓練一個能夠分析跨領域文件的可讀性模型。本論文的內容安排如下：第二節將描述目前採用表示學習法來研發可讀性特徵或是訓練可讀性模型的相關研究。第三節將基於表示學習技術來訓練出一個能夠同時分析不同領域文件的可讀性模型。第四節將呈現本論文所提出可讀性模型的效能。最後第五節是總結及未來研究的方向。

2. 相關研究 (Related Work)

在文本可讀性的研究中，潛在語意分析(Latent Semantic Analysis, LSA)是早期非常受歡迎的語意分析技術之一(Landauer & Dumais, 1997; Landauer, Foltz & Laham, 1998)。其技術如圖 1 所示，潛在語意分析僅需要將詞彙-文章矩陣利用奇異值分解(singular value decomposition, SVD) 將維度縮減，便可以擷取出語料庫的語意空間來表達文件潛藏語意屬性，在取得潛在語意空間(U)後便可以去測量任意二個詞彙、句子、段落及文章之間的語意相似度。在過去，已經許多學者利用潛在語意分析這種表徵學習法應用在可讀性的相關研究。如 Graesser 等人在 Coh-Metrix 3.0 中提供了八個跟 LSA 相關的指標來測量句子或篇章的相似程度(Graesser, McNamara, Louwerse & Cai, 2004)。Truran 等人則是利用

潛在語意分析技術來研究醫學臨床文章的可讀性(Truran, Georg, Cavazza & Zhou, 2010)。François 和 Miltsakaki 利用潛在語意分析去計算詞彙之間的凝聚性來當成可讀性模型的語意指標，以分類法文為第二學習語言的書本可讀性(François & Miltsakaki, 2012)。Kireyev 和 Landauer 使用潛在語意分析來觀察字的成熟度(Word Maturity)，以估測出詞彙的年級難度(Kireyev & Landauer, 2011)。Chang 等人利用潛在語意分析於社會科和自然科的教科書，利用相似度的做法來歸納出每個年級的特定知識概念(Chang, Sung & Lee, 2013)。

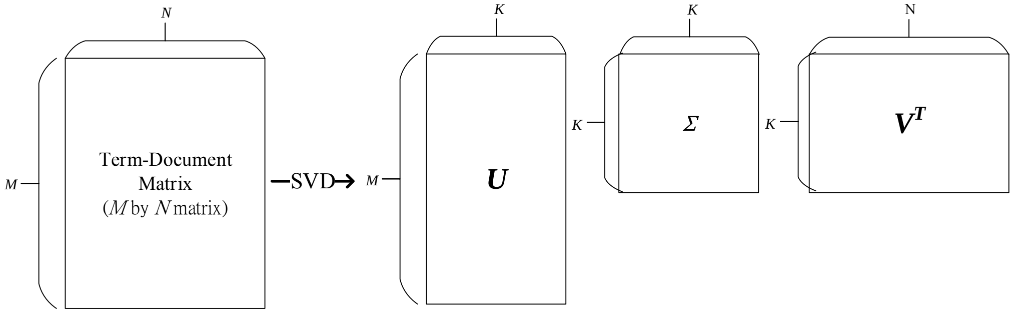


圖 1. 潛在語意分析運用奇異值分解抽取潛在語意空間
[Figure 1. Latent semantic analysis using singular value decomposition to extract the latent semantic space.]

另一個常被應用於可讀性研究的表示學習法是：詞向量，詞向量表示的觀念最早由 Hinton 在 1986 年所提出，又被稱為詞表示(Word Representation or Word Embedding) (Hinton, 1986)。Bengio 在 2003 年提出回饋式類神經網路語言模型(Feed-forward Neural Network Language Model (FFNNLM)的訓練架構，從文件中詞彙前後相鄰的關係來求取詞向量表示(Bengio, Ducharme, Vincent & Jauvin, 2003)。而近期 Google 所發表的 Word2vec 則可視為 FFNNLM 的後繼方法(Mikolov, Chen, Corrado & Dean, 2013)。然而跟 FFNNLM 架構不一樣的是，Word2vec 去除了 FFNNLM 在訓練時最耗時的非線性隱藏層，僅保留輸入層、投影層和輸出層，使其架構更加簡單。Word2vec 提供了二種訓練方式，分別是連續詞袋模型(Continuous Bag-of-words Model, CBOW)及略詞模型(Skip-gram Model, Skip-gram)。連續詞袋模型主要的精神是由目標詞之外的前後文來預測目標詞的機率；而略詞模型的訓練方式正好相反，它是由目標詞本身來去預測前後文的機率，二種訓練模型示意圖如圖 2(a)及圖 2(b)所示。在 Word2Vec 中不論是連續詞袋模型還是略詞模型，在輸出層都可以採用 Hierarchical Softmax 或是 Negative Sampling 兩種模式來增進訓練的效能。其中 Hierarchical Softmax 指的是將訓練資料中不同詞彙都建置霍夫曼樹(Huffman tree)上，使得根節點(root)到每個詞彙都是唯一的路徑，接著在訓練的過程中，不斷得更新霍夫曼樹上每個節點所對應的權重外，也逐步更新詞彙所對應的向量。而 Negative Sampling 則是捨棄了霍夫曼樹的作法；在訓練前除了原本的正例的樣本外，還額外選了數個負例的樣本，在訓練的過程中不斷更新權重，使得正例樣本的機率最大化外，也同時降低了負例樣本的機率，讓詞彙所對應的向量可以逐步獲得修正。在目前也有學者已

經將 Word2vec 應用於可讀性模型，例如 Liu 等人便將 Word2vec 當成可讀性模型裡其中的一個特徵，以分析中、小學國文科教科書及優良課外讀物的可讀性(Liu, Chen, Tseng & Chen, 2015)。Tseng 等人則是將 Word2vec 結合支向量機(Tseng, Sung, Chen & Lee, 2016a)或深層類神經網路(Tseng, Hung, Sung & Chen, 2016b)發展出一個能夠同時分析國文科、社會科及自然科等不同領域文本的可讀性模型。

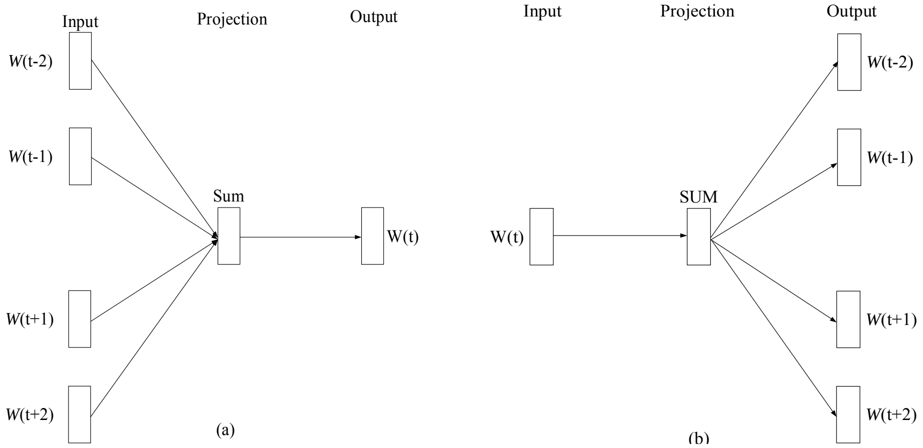


圖 2. (a)連續詞袋模型訓練演算法。(b)略詞模型訓練演算法。
[Figure 2. (a) The continuous bag-of-words model; (b) The skip-gram model.]

由上述可知研究可知不論是潛在語意分析或是 Word2vec，都可以在不需領域專家的介入之下，依據其演算法自動從文本中抽取出可讀性模型所需要的特徵。近年來表示學習法仍蓬勃發展，因此，本論文將嘗試以卷積神經網路(Convolutional Neural Network, CNN) (LeCun, 1989)或快速文本(festText) (Joulin, Grave, Bojanowski & Mikolov, 2016)等不同的表示學習法來自動抽取文本特徵的技術，訓練出一個能夠分析跨領域文件的可讀性模型。

3. 基於表示學習技術之可讀性模型 (Readability Model Based on Representation Learning Techniques)

3.1 卷積神經網路 (Convolutional Neural Network)

卷積神經網路是一種分層式的分類類型結構，它的每個模組都是由卷積層(Convolutional Layer)和池化層(Pooling Layer)來組成(LeCun, 1989)，通過模組不斷的疊加或是加上多層的深層類神經網路後，以形成深度的模型。整個卷積神經網路透過三個重要的思想來幫助改進機械學習的系統：稀疏交互(Sparse Interactions)、參數共享(Parameter Sharing)及等變表示(Equi-variant Representations) (Goodfellow *et al.*, 2016)。稀疏交互又稱為稀疏連接(Sparse Connectivity)，主要是利用數個核(Kernel)基於自定的大小(Kernel Size)來局部連結兩層網路，使得整個模型所要儲存的參數變少，可以有效的減少計算量而提高計算效率。參數共享指的是在同一個核中，每一個元素在不同位置所作用的權重都是相同的。這意

味著在卷積運算的過程中，模型只需要學習一組固定的參數即可，而不是對於每個元素在不同的位置所用的權重都是獨立的。因此這也將大幅度提高模型的訓練效能。而參數共享的機制再加上適當的池化策略，也促成了卷積類神經網路對於局部平移有一些不變的特性可以應用於圖像的處理或語音辨識，尤其是在關心某個特徵是不是有出現，而不是關心它出現的具體位置時(Goodfellow *et al*., 2016)。目前卷積類神經網路已經成功被應用於圖像分析(Cireşan, Meier, Gambardella & Schmidhuber, 2010; Cireşan, Meier, Masci & Schmidhuber, 2011; Ciresan, Giusti, Gambardella & Schmidhuber, 2012)、語音辨識(Abdel-Hamid, Deng & Yu, 2013; Deng, Abdel-Hamid & Yu, 2013; Deng *et al*., 2013)和自然語言處理(Kim, 2014; Zhang & Wallace, 2015; Johnson & Zhang, 2014)。本研究也嘗試將卷積類神經網路用來自動抽取可讀性模型所需要的特徵，並利用深層類神經網路訓練出可讀性模型，其架構如圖 3 所示，在訓練的過程中將使用到 Dropout (Srivastava, Hinton, Krizhevsky, Sutskever & Salakhutdinov, 2014)的技巧來避免模型過度適配(overfitting)外，並利用 rectified linear units (ReLU) (Nair & Hinton, 2010)作為的激發函數(active function)，以避免典型的梯度消失(gradient vanish)問題。

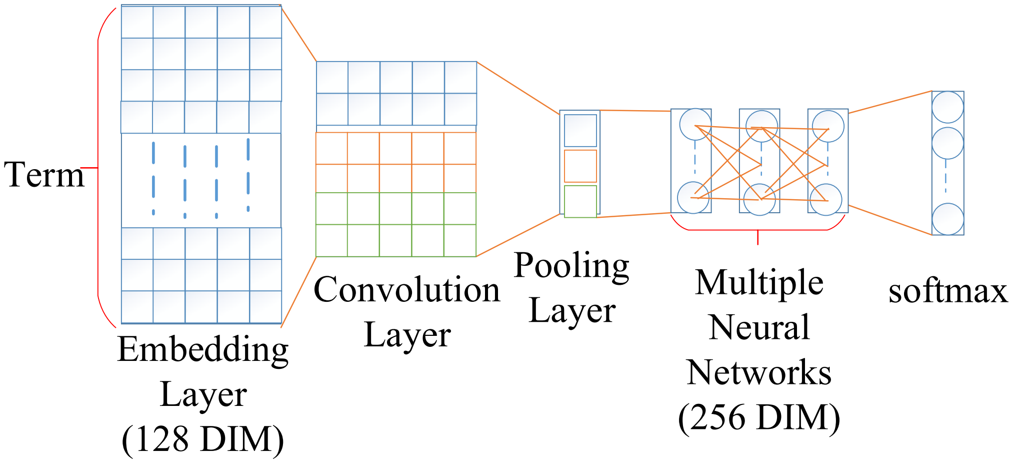


圖 3.基於卷積類神經網路之可讀性模型架構
[Figure 3. A Framework of Readability Model Based on Convolutional Neural Networks.]

3.2 快速文本(fastText)方法 (fastText Library)

繼 Word2vec 之後，Joulin 等人持續改變 Word2vec 的架構發展出快速文本(Joulin *et al*., 2016)。快速文本與 Word2vec 一樣有連續詞袋模型和略詞模型兩種架構，是屬於基於一個長度來看詞彙之間的關係，而透過滑動視窗(Sliding Window)來進行訓練的技巧可以彈性適用於不同長度的文本。但不一樣的地方在於將目標詞彙改換成訓練資料的類別，其示意圖如圖 4(a)及圖 4(b)所示。除此之外，在輸入層方面，也由 Word2vec 的 unigram 改

成 n-gram 的架構以供模型訓練時有更多的彈性。至於快速文本的輸出層是要採用 Hierarchical Softmax 或 Negative Sampling 則跟 Word2vec 一樣，可以自由選用。。而快速文本這樣子的訓練技巧也有別於潛在語意分析與 Word2vec，直接將分類器的需求巧妙的整合至訓練的演算法之中。因此，本研究也嘗試將快速文本採用連續詞袋演算法來訓練可讀性模型。

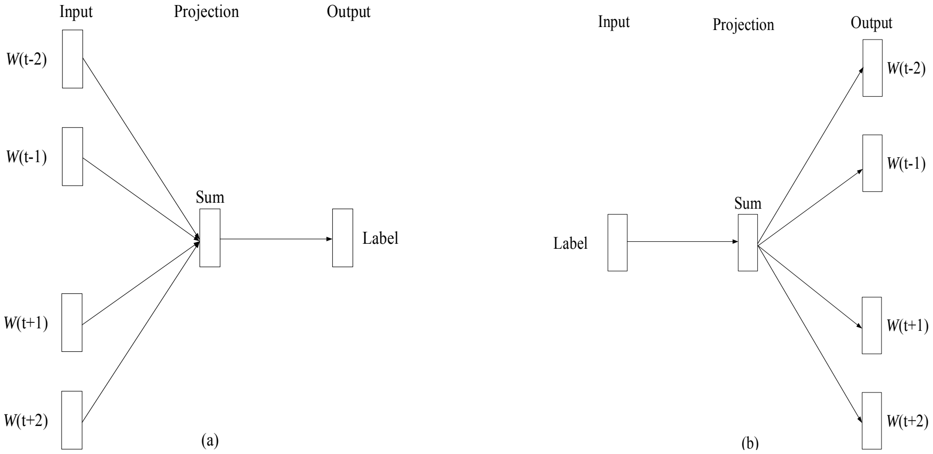


圖 4. (a)快速文本之連續詞袋模型。(b)快速文本之略詞模型。
[Figure 4. (a) Continuous Bag-of-words Model of fastText. (b) Skip-gram Model of fastText.]

4. 實驗及結果 (Experimental Results)

4.1 實驗材料 (Materials)

本研究材料選自 98 年度臺灣 H、K、N 三大出版社所出版的 1-12 年級審定版的國語科、社會科、自然科及體育和健康教育等四個領域的教科書全部共計 6,230 篇，各版本教科書均經由專家根據課程綱要編制而成，其實驗材料的年級分佈如表 1 所示。

表1. 實驗材料在各年級的數量分佈
[Table 1. The statistics of the dataset.]

年級	1	2	3	4	5	6	7	8	9	10	11	12
社會科	0	0	80	74	85	81	389	407	325	340	331	270
自然科	0	0	72	67	67	62	172	175	157	211	355	295
國文科	24	67	61	71	69	70	37	34	28	84	41	47
體育及健康教育科	125	125	121	144	149	150	79	91	85	197	139	177

4.2 訓練可讀性模型 (Train Readability Models)

首先將實驗材料利用 WECAn(Chang, Sung & Lee, 2012)來進行中文斷詞的前處理程序，接著再將訓練資料利用卷積類神經網路或快速文本來抽取出可讀性模型所需要的特徵，而它的類別就是課文所屬的年級，並未再細分文本是屬於何種領域。本研究利用 Keras (Chollet, 2015)予以實作，整個實驗流程皆採用 5-fold 交互驗證的方式如圖 5 所示。在挑選訓練資料的時候，本研究分別從四大領域的文本中各年級文本數量依比例亂數挑選，接著將這些訓練資料輸入至斷詞的程序，各文本經過斷詞後的結果並不含詞性的標記(e.g.「今天天氣很好」將會被斷詞成「今天 天氣 很 好」)。接著在訓練模型時，本研究將文本最大的長度設為 1,000，若文本的長度超過 1,000 時，程式將會自動截斷該文本後續的資訊。而在經過多次模型的參數調校，本研究將卷積神經網路和快速文本的嵌入層分別設為 128 與 100，而由於本研究並還未使用預訓練(pre-training)的技巧，因此嵌入層的權重將隨著模型訓練的過程中一起訓練。

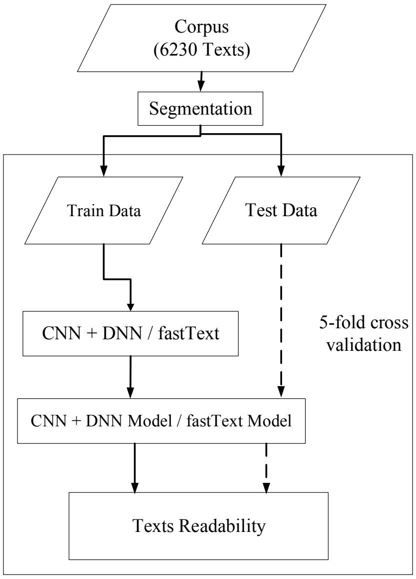


圖 5. 可讀性模型訓練及測試流程圖
[Figure 5. FlowChart of Training and Testing Readability models.]

4.3實驗結果 (Results)

本論文分別採用卷積神經網路或快速文本來訓練可讀性模型，並與過去的研究進行比較。可讀性模型準確率如表 2，而錯誤矩陣分別如表 3 及表 4 所示。除了呈現準確率之外，本研究也放寬上、下一個年級的標準來統計出鄰近準確率，以觀察可讀性模型錯誤預測的程度是否嚴重。我們可以發現不論是以類神經網路或是快速文本來訓練可讀性模型，其準確率和鄰近準確率皆比支向量機(Support Vector Machine, SVM) (Vapnik & Chervonenkis, 1974)還好。而我們也可以發現不論是採用卷積神經網路、快速文本及

Word2vec 皆可以有效的表徵不同領域的文本來當成可讀性特徵，使得訓練出來的可讀性模型可以具有領域一般化的能力。

表 2. 基於卷積神經網路及快速文本之可讀性模型效能比較
[Table 2. Comparison Performance of Readability Models Based on Convolutional Neural Network and fastText.]

適用年級	適用領域	可讀性特徵	分類器	準確率	鄰近 準確率
1-12 年級	國語科、社會科、自然科及體育和健康教育共計 6,230 篇	卷積神經網路	類神經網路 (一層)	67.62%	86.76%
		快速文本		69.63%	86.01%
		Word2vec	支向量機 (Tseng *et al*., 2016b)	61.33%	82.2%
		Word2vec	類神經網路 (一層) (Tseng *et al*., 2016b)	66.95%	85.26%

表 3. 卷積神經網路可讀性模型之錯誤矩陣
[Table 3. Confusion Matrices of the Convolutional Neural Network based Readability Model.]

		模型預估年級												準確率(%)
		1	2	3	4	5	6	7	8	9	10	11	12	
實際年級	1	74	52	21	0	1	0	0	0	0	0	0	1	49.66
	2	42	92	42	13	2	0	0	0	1	0	0	0	47.92
	3	6	36	197	63	17	11	1	2	1	0	0	0	58.98
	4	4	13	51	223	40	20	5	0	0	0	0	0	62.64
	5	0	1	22	54	165	88	29	4	2	3	1	1	44.59
	6	0	2	7	34	95	176	18	6	10	11	1	3	48.48
	7	2	0	1	2	16	26	526	36	26	39	1	2	77.70
	8	0	0	0	2	3	29	42	540	31	43	12	5	76.38
	9	1	0	0	0	1	23	29	29	454	41	15	2	76.30
	10	5	0	0	0	5	15	26	23	19	586	92	61	70.43
	11	0	0	0	0	2	7	7	9	9	142	630	60	72.75
	12	4	0	0	0	0	5	8	4	3	122	93	550	69.71

表 4. 快速文本可讀性模型之錯誤矩陣
[Table 4. Confusion Matrices of the fastText Readability Model.]

		模型預估年級												準確率(%)
		1	2	3	4	5	6	7	8	9	10	11	12	
實際年級	1	87	47	12	3	0	0	0	0	0	0	0	0	58.39
	2	46	97	43	6	0	0	0	0	0	0	0	0	50.52
	3	28	15	219	52	12	6	1	0	0	0	1	0	65.57
	4	10	5	62	209	33	25	10	1	1	0	0	0	58.71
	5	2	5	28	43	184	75	27	3	0	2	0	1	49.73
	6	0	4	22	23	79	188	18	5	13	7	2	2	51.79
	7	1	1	3	2	19	22	560	26	5	28	2	8	82.72
	8	10	0	0	1	12	21	28	571	12	32	7	13	80.76
	9	10	0	0	0	8	18	24	11	461	36	16	11	77.48
	10	3	4	0	1	7	11	47	39	13	555	95	57	66.71
	11	0	1	0	0	4	4	16	20	24	113	619	65	71.48
	12	1	3	0	0	1	4	11	6	4	84	87	588	74.52

此外，本論文也針對卷積類神經網路的可讀性模型去加深類神經網路的層數，以跟過去學者的結果進行比較。其結果如表 5 所示，本研究發現，以卷積類神經網路為特徵的可讀性模型，其效能並未隨著類神經分類器的層數增加而上升，且最佳的準確率低於 Word2vec 為特徵的可讀性模型 0.97%。然而，其鄰近準確率卻反而高過 0.65%。以整體而言，以 Word2vec 為特徵的可讀性模型其準確率是比較高的，但以卷積類神經網路為特徵的可讀性模型其鄰近準確率是較高的。最後綜合表 2 和表 5 而言，我們可以發現快速文本的準確率仍是所有可讀性模型中最高的，但鄰近準確率卻也是最低的。

表 5. 深層類神經網路隱藏層的數量對於可讀性模型的影響
[Table 5. The Influence of the Number of Hidden Layers of Neural Networks on the Readability Model.]

適用年級	適用領域	分類器-類神經網路層數量	可讀性特徵			
			卷積類神經網路		Word2vec (Tseng *et al.*, 2016b)	
			準確率(%)	鄰近準確率(%)	準確率(%)	鄰近準確率 (%)
1-12 年級	國語、社會、自然、體育和健康教育共計 6,230 篇	1	**67.62**	**86.76**	66.95	85.26
		2	67.09	85.99	**68.59**	**86.11**
		3	66.5	86.31	68.33	85.54

5. 結論 (Conclusions and Future work)

過去可讀性模型所採用的特徵大多需要專家去設計，有著耗時費力等問題。有鑑於此，本論文基於表示學習演算法，提出以卷積類神經網路或快速文本來自動抽取文本的特徵去訓練可讀性模型，並以實證證明其效能與具領域一般化的能力。除此之外，以本研究的實驗材料而言，與支向量機同屬於淺層結構的機械學習演算法：快速文本，其效能並不輸給深層結構的機械學習演算法。針對此點發現，本研究未來將會納入更多的訓練資料及不同深、淺層結構的機械學習演算法來加以探討對於可讀性模型的影響。

除此之外，本研究也發現不同架構的可讀性模型所呈現出來的結果有很大的差異，如快速文本準確率雖然是最高的，但從表 3 和表 4 的比較可以發現，快速文本對於某些年級的文本在預測錯誤時，其錯誤誤差的程度非常嚴重；反觀卷積類神經網路預測錯誤誤差的程度就相對集中。而針對模型預測部分文本產生嚴重的誤差，本研究認為可能的原因是：對於體育和健康教育這個領域的教科書而言，為了讓國小低年級的幼童可以盡早認識與自己切身相關的知識，如：身體構造、身體自主權及生活環境、疾病…等等議題。雖然遣詞用字早就超過該年級的識字難度(以國文科相應年級課文所教授的生字而言)，但經由老師的介紹及圖片和注音的輔助，使得學生是可以理解文本的內容。相較之下，表徵學習法單純從文字所獲得的資訊就相當有限，因此當上述這些低年級的文本當成訓練資料時，一些高年級的測試資料如果用字是簡單時(如：白話文、介紹體育器材、介紹運動規則…等等)，這些文本很容易被誤判成國小低年級就可以閱讀。因此在未來的研究中，除了整合不同類型的類神經網路模型的優點來促使可讀性模型在預測錯誤時，其誤差也能夠盡可能的往適讀年級集中外；也將納入更多的特徵以輔助目前可讀性模型不足的地方。

Reference

Abdel-Hamid, O., Deng, L. & Yu, D. (2013). Exploring convolutional neural network structures and optimization techniques for speech recognition. In *Interspeech 2013,* 3366-3370.

Bertha, A. L. & Pressey, S. L. (1923). A method for measuring the" vocabulary burden" of textbooks. *Educational Administration and Supervision*, *9*, 389-398

Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

Borst, A., Gaudinat, A., Grabar, N. & Boyer, C. (2008). Lexically-based distinction of readability levels of health documents. *Acta Informatica Medica*, *16*(2), 72-75.

Chang, T. H., Sung, Y. T. & Lee, Y. T. (2012). A Chinese word segmentation and POS tagging system for readability research. In *Proceedings of the 42nd Annual Meeting of the Society for Computers in Psychology*.

Chang, T. H., Sung, Y. T. & Lee, Y. T. (2013). Evaluating the difficulty of concepts on domain knowledge using latent semantic analysis. In *Proceedings of 2013 International Conference on Asian Language Processing (IALP),* 193-196. doi: 10.1109/IALP.2013.58

Chall, J. S. & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, Mass: Brookline Books.

Chollet, F. (2015). Keras: Deep learning library for theano and tensorflow. *URL: https://keras.io*.

Cireşan, D. C., Meier, U., Masci, J. & Schmidhuber, J. (2011). A committee of neural networks for traffic sign classification. In *Proceedings of The 2011 International Joint Conference on Neural Networks (IJCNN),* 1918-1921. doi: 10.1109/IJCNN.2011.6033458

Ciresan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. (2012). Deep neural networks segment neuronal membranes in electron microscopy images. In *Proceedings of the 25th International Conference on Advances in neural information processing systems(NIPS'12),* 2843-2851.

Cireşan, D. C., Meier, U., Gambardella, L. M. & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, *22*(12), 3207-3220. doi: 10.1162/NECO_a_00052

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, *165*(2), 97-135.

Dale, E. & Chall, J. S. (1949). The concept of readability. *Elementary English*, *26*(1), 19-26.

Deng, L., Abdel-Hamid, O. & Yu, D. (2013. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 6669-6673.

Deng, L., Li, J., Huang, J. T., Yao, K., Yu, D., Seide, F., ... & Acero, A. (2013). Recent advances in deep learning for speech research at Microsoft. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 8604-8608.

Feng, L., Jansche, M., Huenerfauth, M. & Elhadad, N. (2010). A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters (COLING '10)*, 276-284.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221-233. doi: 10.1037/h0057532.

François, T. & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas?. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations (PITR '12)*, 49-57.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning (adaptive computation and machine learning series)*. Cambridge, MA: The MIT Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, *Instruments, & Computers*, *36*(2), 193-202.

Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, *101*(3), 371-395. doi: 10.1037/0033-295X.101.3.371

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, *1*, 1-12.

Johnson, R. & Zhang, T. (2014). Effective use of word order for text categorization with convolutional neural networks. Retrieved from *arXiv preprint arXiv:1412.1058*.

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2016). Bag of tricks for efficient text classification. Retrived from *arXiv preprint arXiv:1607.01759*.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751. Retrieved from arXiv preprint arXiv:1408.5882.

Kireyev, K. & Landauer, T. K. (2011). Word maturity: Computational modeling of word knowledge. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 1, 299-308.

Klare, G. R. (1963). *Measurement of readability*. Ames, IA: Iowa State University Press.

Klare, G. R. (2000). The measurement of readability: useful information for communicators. *ACM Journal of Computer Documentation (JCD)*, *24*(3), 107-121. doi: 10.1145/344599.344630

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211-240. doi: 10.1037/0033-295X.104.2.211

Landauer, T. K., Foltz, P. W. & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2-3), 259-284. doi: 10.1080/01638539809545028

LeCun, Y. (1989). Generalization and network design strategies. In R. Pfeifer, Z. Schreter, F. Fogelman, & L. Steels (Eds.), *Connectionism in perspective*. Zurich, Switzerland: Elsevier.

Liu, Y. N., Chen, K. Y., Tseng, H. C. & Chen, B. (2015). A Study of Readability Prediction on Elementary and Secondary Chinese Textbooks and Excellent Extracurricular Reading Materials. *In Proceedings of the 27th Conference on Computational Linguistics and Speech Processing (ROCLING 2015),* 71-86. [In Chinese]

Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrived from *arXiv preprint arXiv:1301.3781*.

Nair, V. & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, 807-814.

Petersen, S. E. & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer speech & language*, *23*(1), 89-106. doi: 10.1016/j.csl.2008.04.003

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research*, *15*(1), 1929-1958.

Sung, Y. T., Chen, J. L., Cha, J. H., Tseng, H. C., Chang, T. H. & Chang, K. E. (2015). Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior research methods*, *47*(2), 340-354. doi: 10.3758/s13428-014-0459-x.

Truran, M., Georg, G., Cavazza, M. & Zhou, D. (2010). Assessing the readability of clinical documents in a document engineering environment. In *Proceedings of the 10th ACM symposium on Document engineering (*DocEng '10), 125-134. doi: 10.1145/1860559.1860585

Tseng, H. C., Hung, H. T., Sung, Y. T. & Chen, B. (2016). Classification of Text Readability Based on Deep Neural Network and Representation Learning Techniques. *In Proceedings of 28th Conference on Computational Linguistics and Speech Processing (ROCLING 2016),* 255-270. [In Chinese]

Tseng, H. C., Sung, Y. T., Chen, B. & Lee, W. E. (2016). Classification of text readability based on representation learning techniques. *In Proceedings of the 26th Annual Meeting of the Society for Text & Discourse*.

Vapnik, V. N. & Chervonenkis, A. Y. (1974). *Teoriya raspoznavaniya obrazov. Statisticheskie problemy obucheniya* (Theory of pattern recognition. Statistical problems of learning). Moscow, Russia: Nauka.

Vogel, M. & Washburne, C. (1928). An objective method of determining grade placement of children's reading material. *The Elementary School Journal*, *28*(5), 373-381.

Yan, X., Song, D. & Li, X. (2006). Concept-based document readability in domain specific information retrieval. In *Proceedings of the 15th ACM international conference on Information and knowledge management (CIKM '06)*, 540-549. doi: 10.1145/1183614.1183692

Zhang, Y. & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Retrieved from *arXiv preprint arXiv:1510.03820*.

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 47-62

改進的向量空間可適性濾波器用於聲學回聲消除

Acoustic Echo Cancellation Using an Improved Vector-Space-Based Adaptive Filtering Algorithm

李尤進[*]、曹昱[+]、錢膺仁[#]

Jin Li-You, Yu Tsao and Ying-Ren Chien

摘要

在回聲消除系統的應用中，濾波器係數是否能有效的快速更新是相當重要的關鍵，而收斂的效果也是影響回聲是否能消除乾淨的重要因素。因此向量空間可適性濾波器被提出，其結合機械學習向量空間的想法，引進可適性演算法中，達到有效的快速收斂的目標，但其運算複雜度也相對提高。然而為對應到現實生活中應用，運算複雜度將會是較為重要的考量。因此本篇提出改進的向量空間可適性濾波器(Improved Vector-space Adaptive Filter)與改進的向量空間仿射投影符號演算法(Improved Vector-space Affine Projection Sign Algorithm)，藉由重新設計向量空間以及濾波器係數合成的架構，將運算的矩陣維度降低，並運用組合演算法的想法，對仿射投影符號演算法與改進的向量空間仿射投影符號演算法進行組合，達到在任何環境下皆能快速且穩定收斂的目標且比起向量空間可適性濾波器有著更低的運算複雜度和更好的收斂速度與收斂效果，提升在現實生活中應用地可行性。

關鍵詞：回聲消除系統、可適性濾波器、向量空間可適性濾波器、機器學習、組合演算法、仿射投影符號演算法

[*] 國立臺灣大學電信工程研究所
Graduate Institute of Communication Engineering, National Taiwan University
E-mail: d05942004@ntu.edu.tw

[+] 中央研究院資訊科技創新研究中心
Research Center for Information Technology Innovation, Academia Sinica
E-mail: yu.tsao@citi.sinica.edu.tw

[#] 國立宜蘭大學電機工程學系
Department of Electrical Engineering, National Ilan University
E-mail: yrchien@niu.edu.tw

Abstract

To eliminate acoustic echo, the convergence rate and low residual echo are very important to adaptive echo cancelers. Meanwhile, an affordable computational complexity has to be considered as well. In this paper, we proposed the improved vector space adaptive filter (IVAF)and Improved Vector-space Affine Projection Sign Algorithm (IVAPSA). The proposed can be divided into two phases: offline and online. In the offline phase, IVAF constructs a vector space to incorporate the prior knowledge of adaptive filter coefficients from a wide range of different channel characteristics. Then, in the online phase, the IVAF combines the conventional APSA and IVAPSA algorithms, where IVAPSA computes the filter coefficients based on the vector space obtained in the offline phase. By leveraging the constructed vector space, the proposed IVAF is able to fast converge and achieve a better echo return loss enhancement performance. Moreover, the computational complexity is less than a comparable work.

Keywords: Acoustic echo cancellation, Adaptive Filter, Vector-space Adaptive Filter, Machine Learning, Combined Algorithm, Affine Projection Sign Algorithm.

1. 緒論 (Introduction)

在免持通訊系統或是進行遠端會議時，通常會運用麥克風作為語音的接收，再利用揚聲器進行撥放。然而在這樣的環境下，揚聲器的聲音容易被麥克風所接收造成回聲的產生，進而嚴重的影響到通訊系統的使用品質與語音訊號的可理解性。針對上述的問題，運用可適性濾波器(Adaptive Filter, AF)的聲學回聲消除系統(Acoustic Echo Cancellation, AEC)被提了出來，並廣泛的應用在實際的通信系統中(Ahgren, 2005) (Faller & Tournery, 2006) (Hansler & Schmidt, 2006) (Wada & Juang, 2012) (Haykin, 2003) (Wada & Juang, 2009) (van Waterschoot & Moonen, 2011)。

通常，AF 的方法是利用濾波器係數來近似回聲路徑的脈衝響應(又稱未知房間系統響應)，並且一步步的從麥克風的訊號預估並去除回聲。圖 1 便是以 AF 為基礎的 AEC 系統架構，其中 $x(n)$、$y(n)$、$d(n)$和 $e(n)$分別是 Far-end 輸入訊號(本地端)、回聲訊號、Near-end 輸入訊號(對方端)以及估計誤差。而 $\mathbf{h}$ 與 $\hat{\mathbf{h}}$ 分別是未知房間系統響應和濾波器係數。而 AEC 系統通過可適性濾波器產生出的估計回聲訊號 $\hat{y}(n)$，將 Near-end 輸入訊號與之相減去除訊號中的回聲部分 $y(n)$，達到消除回聲的目的。AF 的演算法有很多，最小均方演算法(Least Mean Square, LMS)與正規化最小均方演算法(Normalized Least Mean Square, NLMS)便是其中最為基本的，有著較低的運算複雜度，因此被廣泛的應用著(Widrow & Stearns, 1985) (Chien & Chu, 2014) (Soria *et al.*, 2004) (Tandon, Ahmad & Swamy, 2004) (Feuer & Weinstein, 1985) (Chien & Zeng, 2013) (Huang & Lee, 2012)。然而，上述兩個方法卻有著致命的缺點，對於彩色輸入訊號的處理能力不好，無法有好的濾波器係數估計。因此仿射投影演算法(Affine Projection Algorithm, APA) (Liao & Khong, 2010)

(Shin, Sayed & Song, 2004) (Hwang & Song, 2007) (Gil-Cacho, van Waterschoot, Moonen & Jensen, 2012)發展了出來，改善了上述問題，強化了可適性演算法對彩色輸入訊號的處理能力。但是在改善的同時，其運算複雜度上升，且對於雜訊的影響過於敏感，以至於現今較少在現實中被運用。仿射投影符號演算法(Affine Projection Sign Algorithm, APSA) (Shao, Zheng & Benesty, 2010) (Yoo, Shin & Park, 2014) (Shin, Yoo & Park, 2012)在其後被發表，比起 APA 有著更低的運算複雜度、較好的抗雜訊影響以及保持好的對彩色輸入訊號處理效果，現今也在學術討論中被廣泛的討論著。

在 AF 的應用中，演算法的收斂速度與收斂性能是相當重要的關鍵。在傳統的方法中，大多靠著對步階值的調整來影響收斂狀態，但使用這種方法無法同時顧及收斂性能與收斂速度。因此於 2015 年，向量空間可適性濾波器(Vector-space Adaptive Filter, VAF) (Tsao, Fang & Shiao, 2015)被提出，其以創新的想法將機械學習的概念加入 AF 當中，使得 AF 能擁有學習應用的能力。在離線階段可藉由 AF 的事前訓練得到先驗知識群 **H**，而這些先驗知識群便是一組組在不同環境設定下收斂好的濾波器係數；接著在在線階段中，更新權重向量的係數對先驗知識群的向量空間做線性組合，得到濾波器輸出。由於在先驗知識群中已有著收斂完成的各種環境係數，因此 VAF 可以比起傳統的 AF 有更快的收斂速度與收斂性能，由上述可知，先驗知識群的設計對於 VAF 來說相當重要。

然而，VAF 在線階段中的向量空間設計會造成運算複雜度大幅度的提高，無法有效的拿來作為為之，所以本篇提出了改進的向量空間可適性濾波器(Improved Vector-space Adaptive Filter, IVAF)，重新設計在線向量空間與濾波器輸出合成的結構，並搭配 APSA 提出改進的向量空間仿射投影符號演算法(Improved Vector-space Affine Projection Sign Algorithm, IVAPSA)，改善 VAF 運算複雜度過高之問題，並保持其優異之性能，詳細內容由章節四說明。

2. 演算法與系統架構 (System model and adaptive algorithm)

在這個章節將會介紹使用 AF 處理 AEC 問題的基本系統架構,以及目前較為新穎的仿射投影符號演算法,也是本篇提出方法的基石。

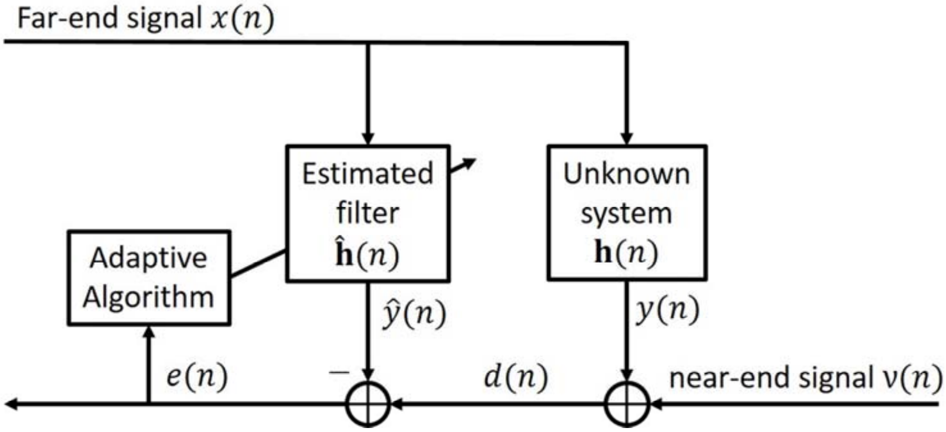


圖 1. AEC 系統架構示意圖
[Figure 1. Structure of acoustic echo cancellation using adaptive filter]

2.1 系統架構 (System model)

一個標準的 AEC 系統架構如圖 1 所示，我們定義未知房間系統響應為$\boldsymbol{h} = [h_0 h_1 \cdots h_{L-1}]^T$，而可適性濾波器的濾波器係數則是$\widehat{\boldsymbol{h}}(n) = [\hat{h}_0(n)\hat{h}_1(n) \cdots \hat{h}_{L-1}(n)]^T$，也稱之為估計系統響應，其中 L 是濾波器的長度，n 為時間指標。

我們令 Near-end 麥克風收到的信號 $d(n)$為:

$$d(n) = \boldsymbol{x}^T(n)\boldsymbol{h} + v(n) \tag{1}$$

其中 $v(n)$是由雜訊訊號與 Near-end 語音所組成。$\boldsymbol{x}(n) = [x_1 x_2 \cdots x_{n-L+1}]^T$Far-end 輸入的語音訊號，經過房間系統響應 $\boldsymbol{h}$ 變成回聲訊號 $y(n)$

$$y(n) = \boldsymbol{x}^T(n)\boldsymbol{h} \tag{2}$$

而可適性濾波器的輸出信號亦是估計的回聲訊號$\hat{y}(n)$為：

$$\hat{y}(n) = \boldsymbol{x}^T(n)\widehat{\boldsymbol{h}}(n-1) \tag{3}$$

則 $e(n)$是估計的回聲訊號與 Near-end 麥克風收到的訊號 $d(n)$相減完所剩下的估計誤差

$$e(n) = d(n) - \hat{y}(n) \tag{4}$$

2.2 仿射投影符號演算法 (APSA)

在仿設投影符號演算法(APSA)中，輸入訊號 $\mathbf{X}(n)$是一個 $L \times p$ 的矩陣，而輸入訊號 $\boldsymbol{d}(n)$為一個 p 維的向量

$$\mathbf{X}(n) = [\boldsymbol{x}(n)\boldsymbol{x}(n-1) \cdots \boldsymbol{x}(n-p+1)]^T$$
$$\boldsymbol{d}(n) = [d(n)d(n-1) \cdots d(n-p+1)]^T \tag{5}$$

其中 p 是仿射投影階數，而輸出訊號亦是一個 p 維的向量

$$\boldsymbol{y}(n) = [y(n)y(n-1)\cdots y(n-p+1)]^T \tag{6}$$

得到的估計誤差向量為

$$\boldsymbol{e}(n) = \boldsymbol{d}(n) - \mathbf{X}^T(n)\widehat{\boldsymbol{h}}(n-1) \tag{7}$$

上式中$\boldsymbol{e}(n) = [e(n)e(n-1)\cdots e(n-p+1)]^T$。而後驗估計誤差向量為

$$\boldsymbol{e}_o(n) = \boldsymbol{d}(n) - \mathbf{X}^T(n)\widehat{\boldsymbol{h}}(n) \tag{8}$$

由(Shao *et al.*, 2010)可以得到 APSA 的估計響應$\widehat{\boldsymbol{h}}(n)$疊代更新公式為

$$\widehat{\boldsymbol{h}}(n) = \widehat{\boldsymbol{h}}(n-1) + \mu\frac{\mathbf{X}(n)\mathrm{sgn}(\boldsymbol{e}(n))}{\|\mathbf{X}(n)\mathrm{sgn}(\boldsymbol{e}(n))\|_2+\delta} \tag{9}$$

而$0 < \delta \leq 1$為一個很小的正歸化常數，以避免分母為零，μ 是更新步階值。

APSA 演算法藉由 L1-*norm* 的架構推導出了更新公式，解決了以往仿射投影系列的演算法運算複雜度過大之問題，在步階值 μ 較小時，可以不怕脈衝式雜訊的影響，有著相當優秀的抗雜訊能力。但是也因為 L1-*norm* 的架構，演算法之收斂速度大幅下降的缺點。

3. 向量空間可適性濾波器(VAF) (Tsao *et al.*, 2015)

VAF 其系統架構如圖 2 所示，比較與圖 1 的差異，可以看到 VAF 的系統架構多出了向量空間 **H** 的結構，幫助更新可適性濾波器的係數。VAF 可以分為兩個階段，離線與在線。

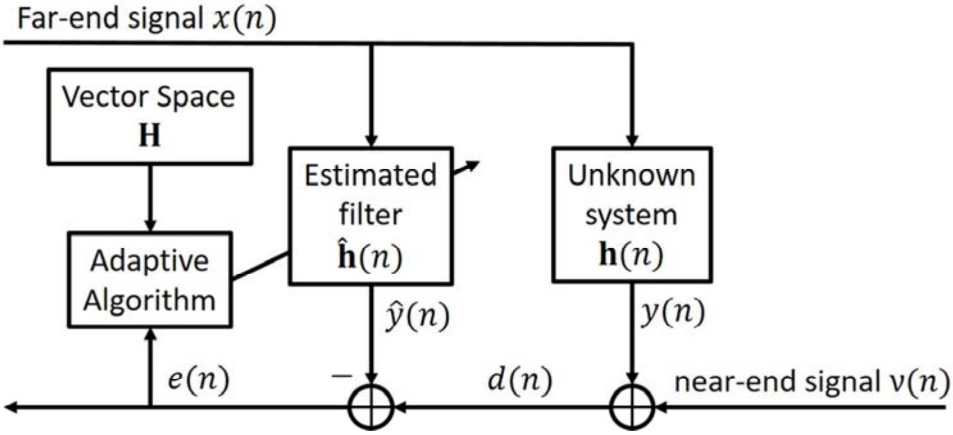


圖 2. VAF-AEC 系統架構示意圖

[Figure 2. Structure of acoustic echo cancellation using vector space based adaptive filter]

3.1 離線階段 (Offline phase)

對於離線階段的目標，是建構一個有效的擁有廣泛通道特性的濾波器輸出係數先驗知識群。該先驗知識群是由 K 組先驗知識所組成，這 K 組先驗知識便是以求得的房間響應向量，其長度為 L

$$\mathbf{H} = [\boldsymbol{h}^1 \boldsymbol{h}^2 \cdots \boldsymbol{h}^K] = \begin{bmatrix} h_0^1 & \cdots & h_0^K \\ \vdots & \ddots & \vdots \\ h_{L-1}^1 & \cdots & h_{L-1}^K \end{bmatrix} \tag{10}$$

其中$\boldsymbol{h}^K$是先驗知識，而 **H** 便是先驗知識群。

3.2 在線階段 (Online phase)

先驗知識群矩陣 **H** 可以看做是由不同房間響應特性所組成的向量空間，在其後方加上單位對角線矩陣$\mathbf{I}_{L\times L}$重新組合成新的向量空間 **S**

$$\mathbf{S} = [\mathbf{H}|\mathbf{I}_{L\times L}] \tag{11}$$

單位對角線矩陣的用意在於，假使 $\boldsymbol{h}$ 無法靠先驗知識群合成，也能倚靠後面的單位對角線矩陣使得演算法保持持續收斂。VAF 中令$\hat{\boldsymbol{h}}(n)$的合成為

$$\hat{\boldsymbol{h}}(n) = \boldsymbol{S}\boldsymbol{w}(n) \tag{12}$$

其中$\boldsymbol{w}(n)$是 K+L 維的權重向量。而 VAF 便是運用權重向量的調整，由先驗知識群合成出最恰當的濾波器輸出，而這種結合先驗知識以促進在線估計的相同技術已被證實在許多的問題中視為有效的(Kuhn, Junqua, Nguyen & Niedzielski, 2000) (Tsao & Lee, 2009) (Belhumeur, Hespanha & Kriegman, 1997)。

但在這樣設計的向量空間下會產生矩陣過大的問題，進而造成運算複雜度大幅提高；也因為單位對角線矩陣的關係，會影響到收斂。

4. 改進的向量空間可適性濾波器 (IVAF)

由於 VAF 向量空間設計的問題，因此 IVAF 在在線階段的向量空間設計上以及濾波器係數的合成上進行了改進。

IVAF 以先驗知識群直接當作合成的向量空間，並將濾波器係數的合成改為由兩個演算法做組合，一個是傳統的 APSA，另一個則是 IVAPSA，並藉由組合係數$\lambda(n)$進行組合。

$$\hat{\boldsymbol{h}}(n) = \lambda(n)\mathbf{H}\hat{\boldsymbol{a}}(n) + \big(1-\lambda(n)\big)\hat{\boldsymbol{b}}(n) \tag{13}$$

其中$\hat{\boldsymbol{a}}(n)$是只有 K 維的權重向量，$\hat{\boldsymbol{b}}(n)$則是 APSA 的濾波器係數。

這樣的設計會大幅降低矩陣的大小，以解決運算複雜度過大之問題，並藉由組合 APSA 保持單位對角線矩陣的用意。而組合係數的設計則是參考(Li-You, 2016)

$$\lambda(n) = \mathbb{E}\left[\frac{e(n)\big(e(n)-\rho(n)\big)}{\big(e(n)-\rho(n)\big)^2}\right] \approx \frac{\sigma_e^2(n)-R_e(n)}{\sigma_e^2(n)-2R_e(n)+\sigma_\rho^2(n)} \tag{14}$$

其中$\sigma_e^2(n)$、$\sigma_\rho^2(n)$為平均的瞬時誤差功率，$R_e(n)$則是誤差訊號的交相關值

$$R_e(n) = \mathbb{E}[e(n)\rho(n)] \tag{15}$$

藉由此組合係數的組合，便能得到最適當的組合比例。當先驗知識群能提供到收斂的幫助時$\lambda(n)$會趨近於 1，使得 IVAPSA 的輸出佔較大部分；反之則是使 APSA 的含量較多。然而由於 APSA 與 IVAPSA 之收斂是獨立的，因此加上一個係數繼承的條件(16)式，使得 IVAPSA 的收斂效果可以延伸到 APSA 上，保持持續收斂的效果。

$$\widehat{\boldsymbol{b}}(n) = \alpha\widehat{\boldsymbol{b}}(n) + (1-\alpha)\widehat{\boldsymbol{a}}(n),\ \text{if}\ \lambda(n) > \beta \tag{16}$$

其中α為趨近於 1 的一個值，β則是繼承門檻值。

4.1 改進的向量空間仿射投影符號演算法 (IVAPSA)

為得到 IVAPSA 權重向量$\widehat{\boldsymbol{a}}(n)$的疊代更新公式，套用 APSA 的推導算法。新的後驗估計誤差為

$$\boldsymbol{\rho}_o(n) = \boldsymbol{d}(n) - \mathbf{X}^T(n)\mathbf{H}\widehat{\boldsymbol{a}}(n) \tag{17}$$

藉由此後驗估計誤差便能得到新的成本函數

$$\mathcal{L}\big(\,\widehat{\boldsymbol{a}}(n),\Lambda\big) = \|\boldsymbol{\rho}_o(n)\|_1 + \Lambda[\|\mathbf{H}\widehat{\boldsymbol{a}}(n) - \mathbf{H}\widehat{\boldsymbol{a}}(n-1)\|_2^2 + \epsilon^2] \tag{18}$$

其中Λ為拉格朗乘數，而$0 < \epsilon \leq 1$。接著將$\mathcal{L}\big(\,\widehat{\boldsymbol{a}}(n),\Lambda\big)$對$\widehat{\boldsymbol{a}}(n)$偏微分後可以得到

$$2\Lambda\mathbf{H}^T[\mathbf{H}\widehat{\boldsymbol{a}}(n) - \mathbf{H}\widehat{\boldsymbol{a}}(n-1)] = \mathbf{H}^T\mathbf{X}(n)\text{sgn}(\boldsymbol{\rho}(n)) \tag{19}$$

在上式中，$\boldsymbol{\rho}(\text{n})$為新的誤差錯誤

$$\boldsymbol{\rho}(n) = \boldsymbol{d}(n) - \mathbf{X}^T(n)\mathbf{H}\widehat{\boldsymbol{a}}(n-1) \tag{20}$$

將(19)式重新整理過後得到

$$2\Lambda = \epsilon\|\mathbf{X}(n)\text{sgn}(\boldsymbol{\rho}(n))\|_2 \tag{21}$$

最後，便能獲得權重向量的更新公式

$$\widehat{\boldsymbol{a}}(n) = \widehat{\boldsymbol{a}}(n-1) + \mu'\mathbf{J}\mathbf{H}^T\frac{\mathbf{X}(n)\text{sgn}(\boldsymbol{\rho}(n))}{\|\mathbf{X}(n)\text{sgn}(\boldsymbol{\rho}(n))\|_2+\delta} \tag{22}$$

而$\mathbf{J} = [\mathbf{H}^T\mathbf{H}]^{-1}$，其值能在離線階段先求得，$\mu'$是新的更新步階值。有了權重向量$\widehat{\boldsymbol{a}}(n)$的更新公式後，就能藉由(13)式得到$\widehat{\boldsymbol{h}}(n)$的更新。

在運用先驗知識群作為向量空間合成的設計下，可以使可適性濾波器的估計系統響應達到更快收斂的功效。IVAF 藉由化簡向量空間的設計來克服 VAF 運算複雜度過高之問題，並運用組合係數將兩種演算法結合，解決化簡向量空間帶來的缺陷，達到改進 VAF 方法不適用於實際應用得缺點。

5. 實驗 (Experiment)

在本章中將會比較傳統 APSA 演算法與 VAPSA 演算法的回聲消除效能差異，採用回聲往返耗損增強(Echo Return Loss Enhancement, ERLE) (Rages & Ho, 2002) (Sukhumalwong & Benjangkaprasert, 2006)作為評比標準。ERLE 顧名思義是用來比較回聲殘響的大小。因此，在聲學回聲消除的論文較常被運用。ERLE 越大表示著剩餘之回聲訊號越少，反之，越接近 0 代表回聲殘留的情形越嚴重，其數學式如下：

$$\text{ERLE} = 10\log_{10}\left(\frac{\sum_{i=1}^{L} d^2(i)}{\sum_{i=1}^{L} e^2(i)}\right) \tag{23}$$

而評判運算複雜度之方法是利用 Matlab 之內建運算時間指令作為基準，該評判方法是相對比較，因為會受硬體之影響，導致不同電腦花費之時間不同，但以相對比例來說是客觀的。

5.1 實驗設置 (Simulation setup)

對於未知房間系統響應的設計，本篇運用了 RIR 工具(Habets, 2006)進行模擬。在 RIR 中有五項可以設定的參數，分別是麥克風位置、揚聲器位置、反射次數、反射係數(RC)、空間大小(RS(長,寬,高))。由於在大部分的通訊條件下，麥克風位置與揚聲器位置在本章實驗中只對 RC 與 RS 進行調整，其他參數皆是定值，麥克風位置設定(1, 0.4, 0.6)，揚聲器位置(1, 1, 1)、反射次數為 1 次。

Far-end 輸入訊號的部分有兩種，第一種採用 6000 點的彩色高斯訊號作為輸入，其數學模型如下

$$G(z) = \frac{1}{1-0.9z^{-1}} \tag{24}$$

第二種採用一段語音訊號作為輸入，該訊號為一女性英文語音，被輸入在 Aurora-4 資料庫中 (Hirsch & Pearce, 2000) (Macho *et al.*, 2002) (Parihar, Picone, Pearce & Hirsch, 2004)，檔案名"01zc020d.wv1"，其訊號如圖 3 所示。

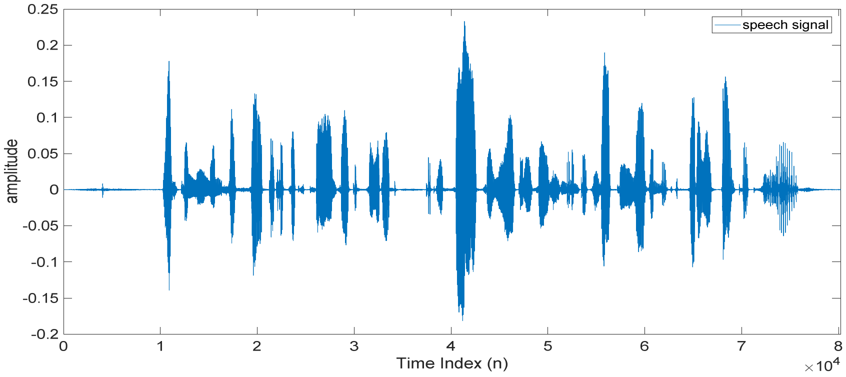


圖 3. 輸入語音訊號圖
[Figure 3. The input speech signal]

為準備先驗知識群的訓練，採用了(24)式的彩色高斯訊號作為輸入，並設定背景雜訊為 SNR= 30dB的白高斯雜訊。

本章中訓練了一組先驗知識群，並且其係數長度 L 為 100，準備了 50 種不同的 RIR 模擬結果訓練，包含了 5 種 RC 係數大小與 10 種 RS 做設定，如表 1；

表 1. 模擬訓練 RIR 50 組參數設定表
[Table 1. Configuration of training data sets]

Data Set	Room Size (RS)	Reflection Coefficient(RC)
Training Set	(1.1, 1.1, 1.1)	-0.91
	(1.2, 1.2, 1.2)	-0.82
	(1.3, 1.3, 1.3)	-0.73
	(1.4, 1.4, 1.4)	-0.64
	(1.5, 1.5, 1.5)	-0.55
	(1.6, 1.6, 1.6)	
	(1.7, 1.7, 1.7)	
	(1.8, 1.8, 1.8)	
	(1.9, 1.9, 1.9)	
	(2.0, 2.0, 2.0)	

表 2.模擬測試 RIR 3 組參數設定表
[Table 2. Configuration of testing data sets]

Data Set	Room Size (RS)	Reflection Coefficient(RC)
Test Set A	(1.1, 1.1, 1.1)	-0.91
Test Set B	(1.1, 1.1, 1.1)	-0.75
Test Set C	(1.1, 1.2, 1.3)	-0.75

並在表 2 中設定實驗時的參數，TEST SETA 是只 RC 與 RS 都在訓練的設定中，意指未知房間系統響應包含在先驗知識中；TEST SETB 則令 RC 不再訓練的設定中；TEST SETC 最為困難，其設定完全不再訓練之中，代表先驗知識完全沒有見過該未知房間系統響應。而可適性濾波器的其他係數設定如下：

步階值$\mu = 0.01$，仿射投影階數$p = 16$，$\delta = 10^{-6}$，$\alpha = 0.99999$，$\beta = 0.9$。

5.2 實驗結果 (Simulation result)

本章中主要比較 4 種可適性濾波器之效能，分別為傳統 APSA、VAF、以及本篇提出之 IVAF 與 IVAPSA，其中 IVAPSA 為 IVAF 組合係數$\lambda(n) = 1$之結果，亦代表不受 APSA 影響之 IVAF 之效能。以彩色高斯訊號作為輸入，加入 SNR= 30dB的背景白雜訊。

由圖 4 可以看到 TEST SET A 的實驗結果，在先驗知識群含有的環境設定中，IVAF 之收斂效果比起 VAF 來的更加的快速，IVAF 在接近 2000 點時已收斂，而 VAF 只比 APSA 法好了一點，且 IVAPSA 與 IVAF 的差異也可看出先驗知識群對於收斂效果的幫助非常大，IVAPSA 在 1000 點時就已穩定收斂，但 IVAF 有 APSA 的幫助下收斂的比 IVAPSA 還要再低；在運算時間的花費上，VAF 總共運算了 1.44 秒，而 IVAF 只花費了 0.5 秒，明顯比 VAF 快了接近 3 倍的時間。

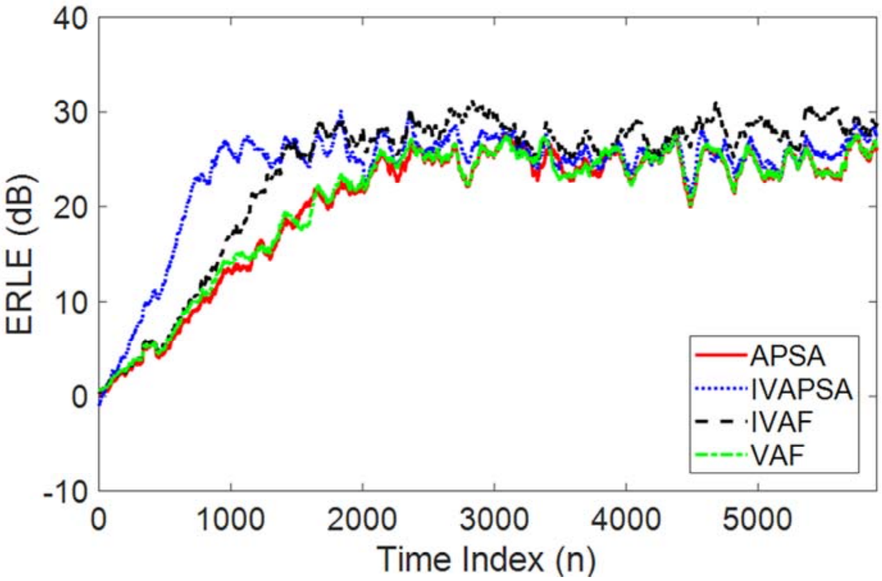


圖 4. TEST SET A 之 ERLE 比較圖
[Figure 4. ERLE of Test setA using Color Gaussian signal be the input.]

而在 TEST SETB 的環境設定下，VAF 花費了 1.48 秒，而 IVAF 則是使用了 0.52 秒的時間，收斂狀況如圖 5，IVAF 亦是能保持著相當快速的收斂狀態，IVAPSA 依舊有著最快收斂的效果，且比起 VAF 的收斂情形來的好很多。由此可以發現 RC 對於先驗知識群的可靠性影響不大，還是能使先驗知識群的向量空間有好的發揮，幫助可適性演算法進行收斂。

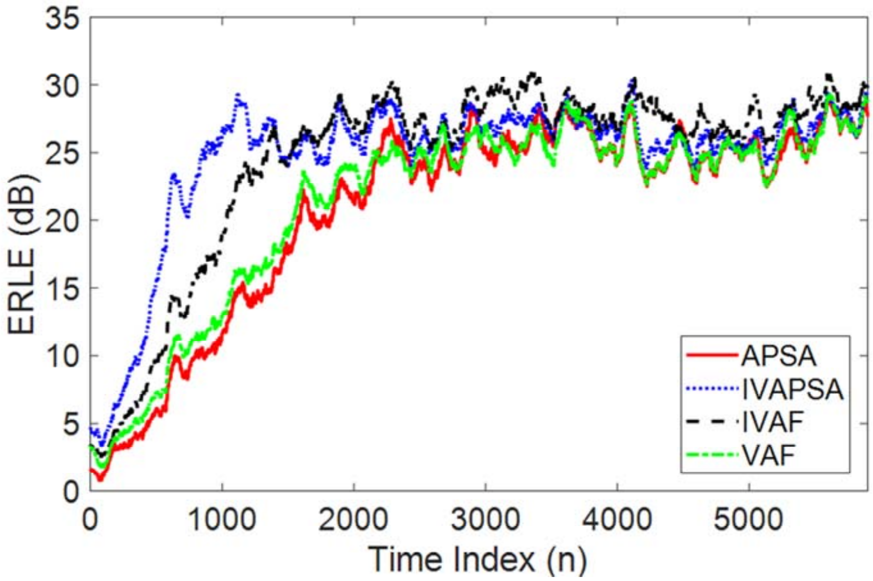


圖 5. TEST SET B 之 ERLE 比較圖
[Table 5. ERLE of Test setB using Color Gaussian signal be the input.]

則在 TEST SAETC 的實驗中，設定的 RC 與 RS 皆不再先驗知識群中，因此先驗知識群就會無法提供收斂幫助。如圖 6 所示，IVAPSA 由於完全是靠先驗知識群進行收斂，所以在這個設定中完全無法進行有效的收斂，甚至濾波器係數發散。反之，VAF 能保持著穩定的收斂，不會收到先驗知識群的影響，而這也是 VAF 加入單位矩陣的原因。再看到 IVAF 亦能持續穩定的收斂，並且收斂的效果大多優於 VAF 與 APSA。而且 VAF 在此花費了 1.51 秒的運算，IVAF 則是使用了 0.53 秒。

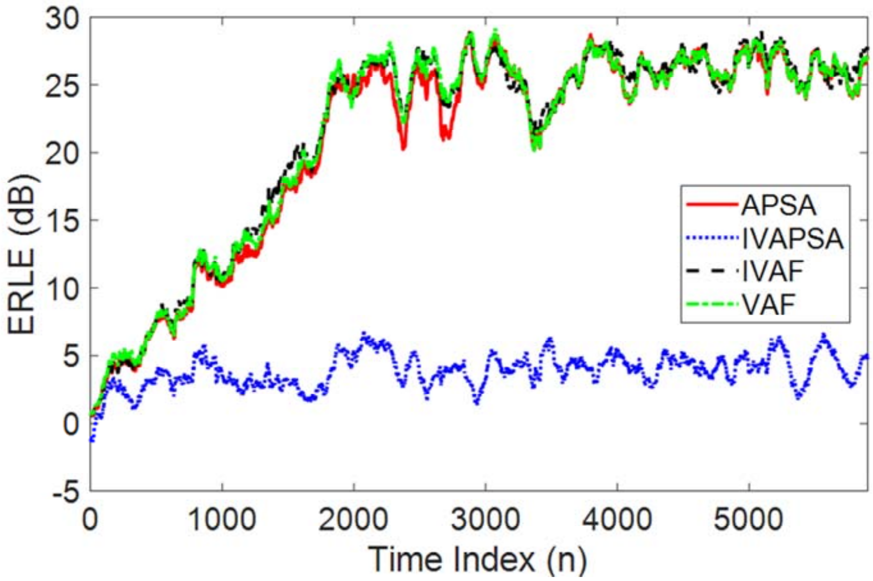


圖 6. TEST SET C 之 ERLE 比較圖
[Figure 6. ERLE of Test setC using Color Gaussian signal be the input.]

在上述三個實驗結果，可以看到 IVAF 明顯改善了 VAF 的收斂情形；在先驗知識群含有的環境設定中，有著比 VAF 更加快速並且收斂更好的效能；而在先驗知識群不含的情境下，也能持續穩定的收斂，不會受到先驗知識群的影響而發散；並且 IVAF 皆能比起 VAF 減少了需多的運算時間。

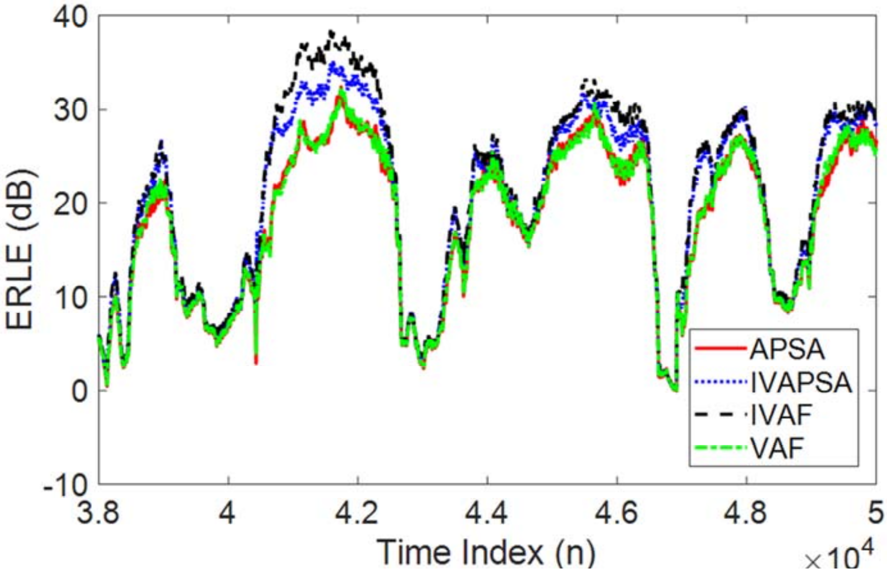


圖 7. 語音輸入採用 TEST SETA 環境設定之 ERLE 比較圖
[Figure 7. ERLE of Test setA using speech signal be the input have 80202 points]

由於本篇主要是將其應用在 AEC 的系統中，因此實際測試 IVAF 模擬 AEC 的狀況，在接下來的實驗採用上述之語音訊號作為輸入，並以 TEST SETA 作為環境設定，且為能更好的觀察收斂狀況，只擷取 38000~50000 點之間做觀察。如圖 7 所示，可以觀察到就算是在真實語音資料的輸入下，IVAF 依舊比 VAF 有更好的收斂速度以及收斂效能，尤其在 40000~43000 點間可以明顯看到 IVAF 的優秀表現，與 VAF 相比 ERLE 最大差了將近 8dB；並且在這個實驗中 IVAF 總共花費了 6.8 秒的時間運算，而 VAF 花費了 19.5 秒的運算時間。

由這些實驗結果可以得知本篇提出之 IVAF 能有效的改善 VAF 的收斂問題，提升收斂速度並且保持收斂的穩定，並且比起 VAF 能花費更少量的運算時間，因此能更實用的應用在 AEC 上。

6. 結論 (Conclusion)

隨著科技的進步，免持系統與遠端通訊已成為人類相當熟悉的應用，而其中衍生出的 AEC 問題更是現在急需解決的。而評判 AEC 是否做得好的標準就是觀察 AF 的收斂速度與收斂效能，且為能應用在現實生活中，運算複雜度也不可以太大，因此本篇提出之方法針對上述三個重點進行設計。在上章實驗中可以看到本篇提出之改進的向量空間可適性濾波器，不管是在彩色高斯的輸入下，又或是真實語音輸入的設定下，皆能有效的改善了以往 VAF 收斂受到限制且運算過於複雜之問題，並且在降低運算複雜度的前提下能有效的提高收斂速度與收斂效能，達到上述三個重點的要求，更容易實現並解決 AEC 之問題。

在未來的規劃中，將會進行更多的實驗模擬，例如空間響應變換，更大量的練習資料以及使用真實聲學回聲進行實驗，並且會詳細計算演算法的運算複雜度來更好的證明 IVAF 方法的優點。

參考文獻 References

Ahgren, P. (2005). Acoustic echo cancellation and doubletalk detection using estimated loudspeaker impulse responses. *IEEE Transactions on Speech and Audio Processing*, *13*(6), 1231-1237. doi:10.1109/TSA.2005.851995

Belhumeur, P. N., Hespanha, J. P. & Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, *19*(7), 711-720. doi: 10.1109/34.598228

Chien, Y.-R. & Chu, S.-I. (2014). A fast converging partial update LMS algorithm with random combining strategy. *Circuits, Systems, and Signal Processing*, *33*(6), 1883-1898.

Chien, Y.-R. & Zeng, W.-J. (2013). Switching-based variable step-size approach for partial update lms algorithms. *Electronics Letters*, *49*(17), 1801-1803. doi: 10.1049/el.2013.1762

Faller, C. & Tournery, C. (2006). Robust acoustic echo control using a simple echo path model. In *Proceedings of 2006 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, 281-284. doi:10.1109/ICASSP.2006.1661267

Feuer, A. & Weinstein, E. (1985). Convergence analysis of lms filters with uncorrelated gaussian data. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*(1), 222-230. doi: 10.1109/TASSP.1985.1164493

Gil-Cacho, J. M., van Waterschoot, T., Moonen, M. & Jensen, S. H. (2012). Nonlinear acoustic echo cancellation based on a parallel-cascade kernel affine projection algorithm. In *Proceedings of 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* 33-36. doi: 10.1109/ICASSP.2012.6287810

Habets, E. A. P. (2006). *Room impulse response generator*. Retrieved from Technische Universiteit Eindhoven, Tech. Rep, 2006.

Hansler, E. & Schmidt, G. (2006). *Topics in Acoustic Echo and Noise Control*. Berlin, Germany: Springer-Verlag.

Haykin, S. (2003). *Adaptive Filter Theory*. Upper Saddle River, NJ: Prentice-Hall.

Hirsch, H.-G. & Pearce, D. (2000). The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of the ISCA workshop ASR2000*, 181-188.

Huang, H. C. & Lee, J. (2012). A new variable step-size nlms algorithm and its performance analysis. *IEEE Transactions on Signal Processing*, *60*(4), 2055-2060. doi: 10.1109/TSP.2011.2181505

Hwang, K. Y. & Song, W. J. (2007). An affine projection adaptive filtering algorithmwith selective regressors. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *54*(1), 43-46. doi: 10.1109/TCSII.2006.883215

Kuhn, R., Junqua, J.-C., Nguyen, P. & Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing, 8*(6), 695-707. doi: 10.1109/89.876308

Liao, L. & Khong, A. W. (2010). Sparseness-controlled affine projection algorithm for echo cancelation. In *Proceedings of the Second APSIPA Annual Summit and Conference*, 355-361.

Li-You, J. (2016). *A Study of Convex Combined Adaptive Filtering Algorithms* (Master’s thesis). Available from http://hdl.handle.net/11296/8tqk4w. [In Chinese]

Macho, D., Mauuary, L., Noé, B., Cheng, Y. M., Ealey, D., Jouvet, D., ...Saadoun, F. (2002). Evaluation of a noise-robust dsr front-end on aurora databases. In *Proceedings of INTERSPEECH 2002*, 17-20.

Parihar, N., Picone, J., Pearce, D. & Hirsch, H. G. (2004). Performance analysis of the aurora large vocabulary baseline system. In *Proceedings of 2004 12th European Signal Processing Conference*, 553-556.

Rages, M. & Ho, K. C. (2002). Limits on echo return loss enhancement on a voice coded speech signal. In *Proceedings of the 2002 45th Midwest Symposium on Circuits and Systems*, *2*, 152-155. doi: 10.1109/MWSCAS.2002.1186820

Shao, T., Zheng, Y. R. & Benesty, J. (2010). An affine projection sign algorithm robust against impulsive interferences. *IEEE Signal Processing Letters, 17*(4), 327-330. doi: 10.1109/LSP.2010.2040203

Shin, H.-C., Sayed, A. H. & Song, W.-J. (2004). Variable step-size nlms and affine projection algorithms. *IEEE signal processing letters*, *11*(2), 132-135. doi: 10.1109/LSP.2003.821722

Shin, J., Yoo, J. & Park, P. (2012). Variable step-size affine projection sign algorithm. *Electronics Letters*, *48*(9), 483-485. doi: 10.1049/el.2012.0751

Soria, E., Calpe, J., Chambers, J., Martinez, M., Camps, G. & Guerrero, J. D. M. (2004). A novel approach to introducing adaptive filters based on the lms algorithm and its variants. *IEEE transactions on education*, *47*(1), 127-133. doi:10.1109/TE.2003.822632

Sukhumalwong, S. & Benjangkaprasert, C. (2006). Adaptive echo cancellation using variable step-size algorithm lattice filters. In *Proceedings of TENCON 2006 - 2006 IEEE Region 10 Conference*, 1-4. doi: 10.1109/TENCON.2006.343852

Tandon, A., Ahmad, M. O. & Swamy, M. N. S. (2004). An efficient, low-complexity, normalized lms algorithm for echo cancellation. In *Proceedings of the 2nd Annual IEEE Northeast Workshop on Circuits and Systems, 2004. NEWCAS 2004*, 161-164. doi: 10.1109/NEWCAS.2004.1359047

Tsao, Y. & Lee, C.-H. (2009). An ensemble speaker and speaking environment modeling approach to robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(5), 1025-1037. doi: 10.1109/TASL.2009.2016231

Tsao, Y., Fang, S. H. & Shiao, Y. (2015). Acoustic echo cancellation using a vector-space-based adaptive filtering algorithm. *IEEE Signal Processing Letters*, *22*(3), 351-355. doi: 10.1109/LSP.2014.2360099

van Waterschoot, T. & Moonen, M. (2011). Fifty years of acoustic feedback control: State of the art and future challenges. In *Proceedings of IEEE*, *99*(2), 288-327. doi:10.1109/JPROC.2010.2090998

Wada, T. S. & Juang, B.-H. (2009). Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement. In *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009 (WASPAA '09)*, 205-208. doi:10.1109/ASPAA.2009.5346494

Wada, T. S. & Juang, B.-H. (2012). Enhancement of residual echo for robust acoustic echo cancellation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 175-189. doi:10.1109/TASL.2011.2159592

Widrow, B. & Stearns, S. D. (1985). *Adaptive Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Yoo, J., Shin, J. & Park, P. (2014). Variable step-size affine projection sign algorithm. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *61*(4), 274-278. doi: 10.1109/TCSII.2014.2305013

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 63-72

基於鑑別式自編碼解碼器之錄音回放攻擊偵測系統

A Replay Spoofing Detection System Based on Discriminative Autoencoders

吳家隆*、許祥平*、呂滸鼎+、曹昱+、李鴻欣#、王新民#

Chia-Lung Wu, Hsiang-Ping Hsu, Yu-Ding Lu, Yu Tsao,

Hung-Shin Lee and Hsin-Min Wang

摘要

在此論文中，我們提出了一個基於鑑別式自編碼解碼器的神經網路模型，對語者辨識系統的錄音回放攻擊進行自動偵測，也就是判斷語者辨識系統所收到的音訊內容是屬於真實的人聲或是由錄音機所回放出來的人聲。在語者辨識領域中，以人為的聲音造假對語者辨識系統進行的攻擊稱之為欺騙攻擊(Spoofing Attack)。有鑑於深度類神經網路模型已被廣泛應用在語音處理相關問題，我們期望能夠應用相關模型在此類問題上。在所提出的鑑別式自編碼解碼器模型中，我們利用模型的中間層來達到特徵抽取的目的，並且提出新的損失函數，使得中間層的特徵將依照資料的標記結果做分群，因此新的特徵將具有能鑑別真偽人聲的資訊，最後再利用餘弦相似度來計算所抽取的特徵與真實的人聲相近與否，得到偵測的結果。我們採用 2017 Automatic Speaker Verification Spoofing and Countermeasures Challenge(ASVspoof-2017)所提供的資料庫進行

* 法務部調查局
Investigation Bureau, Ministry of Justice
E-mail: m38025@mjib.gov.tw

+ 中央研究院資訊創新科技研究中心
Research Center for Information Technology Innovation, Academia Sinica
E-mail: jolu.citi@gmail.com; yu.tsao @iis.sincia.edu.tw
The author for correspondence is Yu Tsao.

中央研究院資訊科學研究所
Institute of Information Science, Academia Sinica
E-mail: hungshinlee@gmail.com

測試，所提出的系統在開發數據集上得到了很好的成效，與官方所提供的測試方法相比，其準確度約有 42 %的相對進步幅度。

關鍵字：語者辨識，語者辨識攻擊，回放攻擊偵測，鑑別式自編碼解碼器，深度類神經網路

Abstract

In this paper, we propose a discriminative autoencoder (DcAE) neural network model to the replay spoofing detection task, where the system has to tell whether the given utterance comes directly from the mouth of a speaker or indirectly through a playback. The proposed DcAE model focuses on the midmost (code) layer, where a speech utterance is factorized into distinct components with respect to its true label (genuine or spoofed) and meta data (speaker, playback, and recording devices, etc.). Moreover, the concept of modified hinge loss is introduced to formulate the cost function of the DcAE model, which ensures that the utterances with the same speech type or meta information will share similar identity codes (i-codes) and higher similarity score computed by their i-codes. Tested on the development set provided by ASVspoof 2017, our system achieved a much better result, up to 42% relative improvement in the equal error rate (EER) over the official baseline based on the standard GMM classifier.

Keywords: Speaker Verification, Speakser Verification Attack, Spoofing Attack, Discriminative Autoencoder, Deep Neural Network.

1. 緒論 (Introduction)

在近幾年內，自動語者辨識的準確度已經有了顯著的提升，自動語者辨識系統也已經廣泛地應用在日常生活中，像是行動裝置或是個人語音裝置的登入系統。然而，由於語音合成(Speech Synthesis)、語音轉換(Voice Conversion)、文字轉語音(Text-to-speech)及錄音回放等技術的進步(Abe, Nakamura, Shikano & Kuwabara, 1990) (Chen, Ling, Liu & Dai, 2014) (Van Santen, Sproat, Olive & Hirschberg, 2013) (Ze, Senior & Schuster, 2013)，電腦越來越能夠模仿人類所發出來的聲音，因此，這些技術確實造成了自動語者辨識系統的潛在危機。

在 ASVspoof-2015 比賽中，其主要的任務目標為訓練一個系統來分辨真實人聲(Genuine)及由語音合成器產生的合成人聲(Spoofing) (Wu *et al*., 2015)；故這次比賽所提供的數據庫中，主要組成為文字轉語音和語音合成的語料，並為與文本無關(Text independent)的內容，但並不包括由錄音機所播放出的回放數據。為了能夠訓練出更好的系統來面對更多的語者辨識攻擊(Alegre, Amehraye & Evansdoi, 2013) (Alam, Kenny, Bhattacharya & Stafylakis, 2015) (Xiao *et al*., 2015) (Villalba, Miguel, Ortega & Lleida, 2015)，ASVspoof-2017 所提供的資料集就包含了由多個不同的播放設備所播出來的人聲錄音回放，參賽單位需要面對並解決新的問題。ASVspoof-2017 所提供的基本系統架構(Kinnunen *et al*., 2017) 以常數 Q 倒頻譜係數(Constant Q Cepstral Coefficients, CQCC)為特徵抽取的參考方法，並且利用了二類的高斯混合模型(Gaussian Mixture Models, GMM)作為分類器。此方法對語音合成及文字轉語音這類的攻擊有不錯的成效，故將此方法沿用至回放攻擊的偵側。

近年來，我們可以看到深度學習方法被廣泛應用在語者辨識系統中，並且取得了相當不錯的成績(Yamada, Wang & Kai, 2013) (Sarkar, Do, Le & Barras, 2014) (Lei, Scheffer, Ferrer & McLaren, 2014) (Kenny, Gupta, Stafylakis, Ouellet & Alam, 2014) (Variani, Lei, McDermott, Lopez Moreno & Gonzalez-Dominguez, 2014) (Lin, Mak & Chien, 2017)，但回放攻擊偵測系統則是一個尚未解決的問題。在這篇論文中，我們提出了一個基於深度自編碼解碼器(Chen, Sun, Rudnicky & Gershmandoi, 2016) (Bone, Lee & Narayanan, 2014) (Huang, Wu, Su & Fu, 2017) (Chung, Wu, Shen, Lee & Lee, 2016) (Richardson, Reynolds & Dehakdoi, 2015)的架構來作為偵測系統的基礎，稱之為鑑別式自編碼解碼器(Discriminative Autoencoders, DcAE) (Lee *et al*., 2017) (Yang *et al*., 2017)，我們嘗試使用自編碼解碼器來替換傳統分類方法，並且設計目標函數來偵測語者辨識攻擊。自編碼解碼器為一種對稱的神經網路架構，在此架構中分成編碼層以及解碼層，目標是希望輸出端能夠經由訓練重建輸入端的資料，為了最小化重建誤差，並且達到偵測語者辨識攻擊。我們也提出了一個新的損失函數，稱之為合頁損失(Hinge Loss)，來訓練這個鑑別式自編碼解碼器。根據這個損失函數，每一個經由音訊所輸入的音框將會編碼成相同的 identity codes(i-codes)，而這個 i-codes 就會具有辨別種類的能力，因此所有相同類別的 i-codes 則會表示成近似的特徵，使得 i-codes 可以利用簡易的分類法便可以輕易地分別數據。此外，在合頁損失的部分我們加入了可調動的邊界函數，此邊界函數可以限制自編碼解碼器所更新的權重，因此每次的更新將可以注重在最令模型困惑的數據上，提高模型泛化(Generalization)的能力。

本論文的主要貢獻包括：第一，i-codes 是一個全新的特徵表示法，並且此特徵能夠有效地偵測語者辨識攻擊，並達到有效分類的效果。第二，在鑑別式自編碼解碼器中，中間層具有聚集相同類型語句的能力，此方法為一個新穎且有效的類神經網路架構。第三，在多任務的鑑別式自編碼解碼器中，我們能夠結合額外的資料來改善我們所提出的系統架構，並且提高泛化的能力。

本論文的後續安排如下：第二節簡單介紹了 ASV-spoof 2017 Challenge；第三節敘述了我們所提出的兩個基於自編碼解碼器的架構；第四節介紹實驗語料與設定以及評估的方法跟結果，最後一節則是結論與未來研究方向。

2. 任務描述 (Task Description)

ASVspoof 是一個設定為偵測語者辨識攻擊的比賽，而在 2017 年，則注重在解決錄音回放攻擊，其中所提供的語料庫是來自於 RedDots 的語料庫跟此語料的錄音回放版本。在

這次比賽所提供的訓練數據中，包含了十位男性的語者，其中分別製作成 1508 句的真實語料和 1508 句的回放錄音語料，並且在錄音回放的設備中，利用了六種錄音設備以及三種播放設備。在開發數據中，包含了八個不同的語者和 760 筆真實語料以及 950 筆回放錄音語料，而在錄音回放設備中，則是有十種錄音及播放設備的組合。

在 ASVspoof-2017 中，主辦方所提供的評估標準為無門檻值的相等錯誤率(Equal Error Rate, EER)，意即在計算偵測系統時，在錯誤接收率(False Acceptance Rate, FAR)以及錯誤拒絕率(False Rejection Rate, FRR)相等時所得到的錯誤率。愈好的偵測系統，就會有愈低的相等錯誤率。

3. 系統描述 (System Description)

3.1 特徵抽取 (Feature Extraction)

根據 ASVspoof-2017 所提供的官方文件，常數 Q 倒頻譜係數對語者辨識攻擊偵測有良好的效果(Todisco, Delgado & Evans, 2016)，因此我們使用了常數 Q 倒頻譜係數作為特徵抽取的方法。常數 Q 倒頻譜係數為一個基於常數 Q 轉換(constant Q transform, CQT)的特徵表示法，這個方法最初使用在音樂訊號處理，並且在抽取的過程確保了頻譜上所有頻域的資訊都能夠被有效的保留下來，其計算過程可用下式表達：

$$\mathrm{CQCC(p)} = \sum_{l=1}^{L} \log|X^{CQ}(l)|^2 \cos\left[\frac{p\left(l\text{-}\frac{1}{2}\right)\pi}{L}\right] \tag{1}$$

其中 X^{CQ} 為經過 CQT 轉換輸入訊號後所得到的值，而 $p = 0, 1, \cdots, L\text{-}1$ ，l 則是取樣的頻帶(Frequency Bin)。

3.2 鑑別式自編碼解碼器 (Descriminative Autoencoder)

在自編碼解碼器中，主要的架構分為兩部分，編碼器 f 以及解碼器 g，在編碼器部分，自編碼解碼器會在隱藏層中建立出特別的特徵表示法 (Goodfellow, Bengio & Courville, 2016)，並且使得解碼器能夠利用這樣的特徵達到重建輸入資訊的效果，也就是 $\mathrm{X} \xrightarrow{f} H \xrightarrow{g} X$，其中 H 為資料經過隱藏層所產生出來特徵表示法，為了訓練自編碼解碼器中的參數，來達到重建的效果，則需要利用重建誤差來更新參數，重建誤差的表示如下：

$$F_r(X) = \frac{1}{|X|} \sum_{x \in X} ||y\text{-}x||_2^2 \tag{2}$$

其中 $y = g(f(x))$ 是重建出來的結果，$||\cdot||_2$ 則是 2-範數(2-norm)，$|X|$ 則是樣本的數量，由於此模型會限制並且強迫輸入值可以被複製到輸出，因此可以從輸入資料樣本中學到有用的資料特性。

根據 (Krizhevsky, Sutskever & Hinton, 2012)，我們可以假設 H 中包含了許多有價值的特徵，並且也包含了代表回放聲音的特徵，H_g 為真實錄音語料的集合，H_s 則為錄音回

放語料的集合，因此我們提出了兩個特別的合頁損失來使得隱藏層中的特徵更加具有代表性；這兩個合頁損失適用於更新隱藏層，其表示如下：

$$F_P(H) = \frac{1}{H}\sum_{h_i,h_j\in H_g|h_i,h_j\in H_s} f（M_p\text{-} < h_i, h_j > ）^2 \quad (3)$$

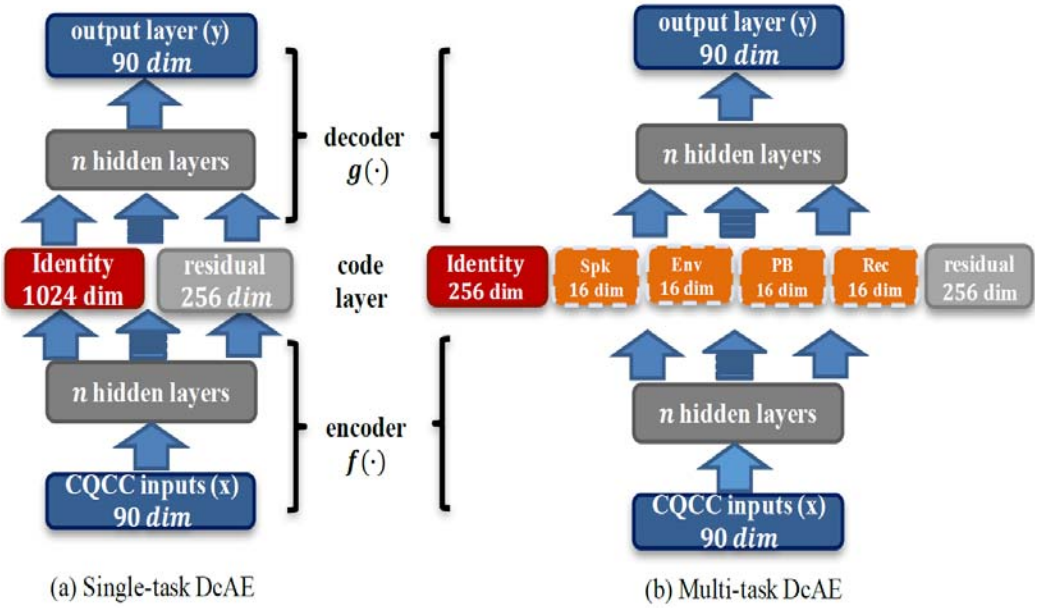
$$F_n(H) = \frac{1}{H}\sum_{h_i,h_j\in H_g|h_i,h_j\in H_s} f（ < h_i, h_j > \text{-}M_n）^2 \quad (4)$$

其中 F_p 為正合頁損失，F_n 為負合頁損失。在經過合頁損失訓練的隱藏層中，將會提供一個獨立的子空間來分離真實說話人聲或是錄音回放人聲，使我們能夠達到分離兩種類別的目標；因為我們專注在解決資料的複雜性，合頁損失能夠有效地面對這類型的問題。舉例來說，式(3)中，相同類別之間的內積若是小於邊界 M，代表此資料配對結果相似度高但不具有代表性，將會更新這個模型中神經元的權重，使同一類別之間內積能夠提高，反之亦然，使得鑑別式自編碼解碼器能夠更新權重達到想達到的目的，進而分辨出輸入的資料為真實人聲與否，在式(3)及式(4)中 f 則為 ReLU (Luong, Le, Sutskever, Vinyals & Kaiser, 2015)，這個非線性轉換方程式能夠有效限制更新權重與否。最後，結合式(2)、式(3)及式(4)，在訓練過程中，模型的目標函數則如下所表示：

$$\alpha\big(F_r(X)\big) + \beta（F_p(H) + F_n(H)） \quad (5)$$

其中 α 控制重建誤差所佔的權重，β 則控制了隱藏層在此模型中佔的重要性。

在論文中，我們將鑑別式自編碼解碼器分成兩種不同的架構，如圖 1 所示。



(a) Single-task DcAE (b) Multi-task DcAE

圖 1. 鑑別式自編碼解碼器示意圖。
[Figure 1. The two kinds of architecture in DcAE]

這兩個架構中，都是將輸入 *x* 經由模型中的 i-codes 轉換成 *y*，然而隱藏層的設計卻截然不同；圖左所表示的為單任務鑑別式自編碼解碼器，在此模型中，中間層被分成兩個部分，一個部分我們稱之為鑑別層，另一個則為剩餘層，在鑑別層中，我們加入了設計好的合頁損失，使得經由鑑別層所產生出來的特徵具有分開類別且避免同類別資料發散的能力，使得更能被辨別，而剩餘層則是希望能夠使解碼的部分較容易達到重建，另一方面，在多任務的鑑別式自編碼解碼器中，則是模仿多任務的類神經網路模型 (Liu *et al*., 2015) (Glorot & Bengio, 2010)，多任務模型具有利用額外的訓練資訊，使得模型泛化能力上升以及有效提升預測準確度，而在此，我們將額外資訊分成語者，錄音環境，錄音設備，以及回放設備，這些資訊皆能從 ASVspoof-2017 的官方資料中獲得，在多任務模型中，除了剩餘層外，我們都加入了合頁損失來讓每層都能夠跟著目標函數更新，提供更多的額外資訊，幫助主要目標來分類真實人聲或是回放人聲。此外，在編碼層及解碼層中則加入了隱藏層，使得此模型具有更多的參數，達到所謂深度學習的效果。

4. 實驗設定以及實驗結果 (Experiment Setting and Result)

在這一節，我們將會一一介紹實驗的設定以及所達到的結果，並且能夠加以討論，首先，在此篇論文中，我們會利用高斯混合模型以及類神經網路模型當作比較的標準，對於高斯混合模型，ASVspoof-2017 官方所提供的系統架構如下：利用常數 Q 倒頻譜係數抽取語料的特徵，並且利用此特徵訓練一個高斯混合模型，使得此模型具有分類的能力；另外，在類神經網路架構中，我們使用了三層隱藏層，並且每層有著 1024 個神經元，相同地，也利用了常數 Q 倒頻譜係數作為抽取語料的特徵表示法，此類神經網路的輸出則是用二元分類來判斷是否為真實人聲，另外，由於我們有提出多任務的鑑別式自編碼解碼器，於是我們也訓練了一個多任務類神經網路來當作比較的標準。在常數 Q 倒頻譜係數的設定中，我們抽取了 90 維的特徵向量作為所有系統的輸入向量，並且抽取完特徵後，我們使其標準正規化，並且使用訓練數據當作標準特徵來轉換開發數據。

在鑑別式自編碼解碼器中，我們將架構中，最中間的層分成了鑑別層以及剩餘層，在鑑別層中，常數 Q 倒頻譜係數將會被編碼為 1024 維的 i-codes，這裡所產生的 i-codes 將會具有分類是否為回放攻擊的能力，而特徵在剩餘層中則被編碼為 256 維；在合頁損失中，我們將正向邊界設定為 10，負向邊界設定為-10。在自編碼解碼器中，所有權重的初始值則是利用 Glorot Uniform 作為初始化設定，在所有的神經元中，我們則是選擇 tanh 作為激發函數(Activation Function)，除了最後一層用來還原數據則是利用選用了 linear 作為激發函數，另外，梯度下降的最佳化演算法則選用 Adaptive Moment Estimation(Adam) (Kingma & Ba, 2014)，使其能快速且有效的達到最佳化的目標。在多任務的鑑別式自編碼解碼器中，我們增加了中間層的數量，使得輸入特徵被額外編碼為其他的資訊，在這些額外的編碼層中，則設定為 16 個神經元，最後，我們則是利用向量內積來計算 i-codes 之間的相似度，因此可以利用此結果作為分類的依據，進而計算相等錯誤率。

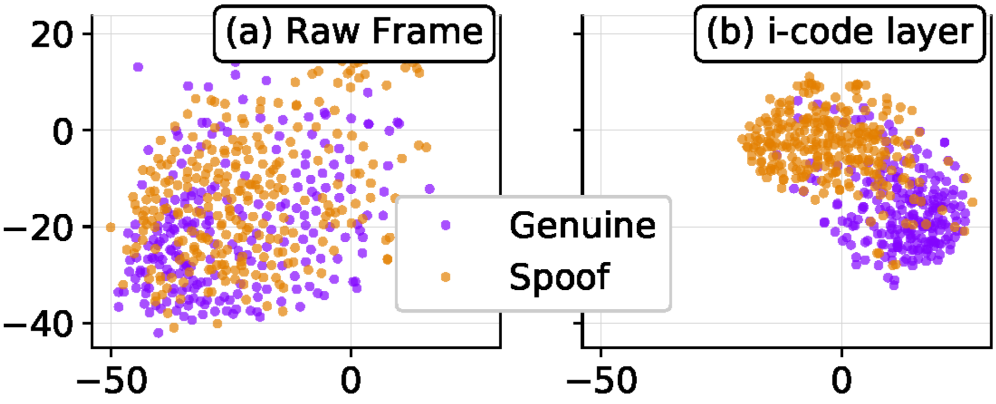


圖 2. 利用 t-SNE 表示在單任務鑑別式自編碼解碼器中，i-codes 的分佈情形。
[Figure 2. The result of single-task DcAE that t-SNE maps high dimension i-codes into 2 dimension.]

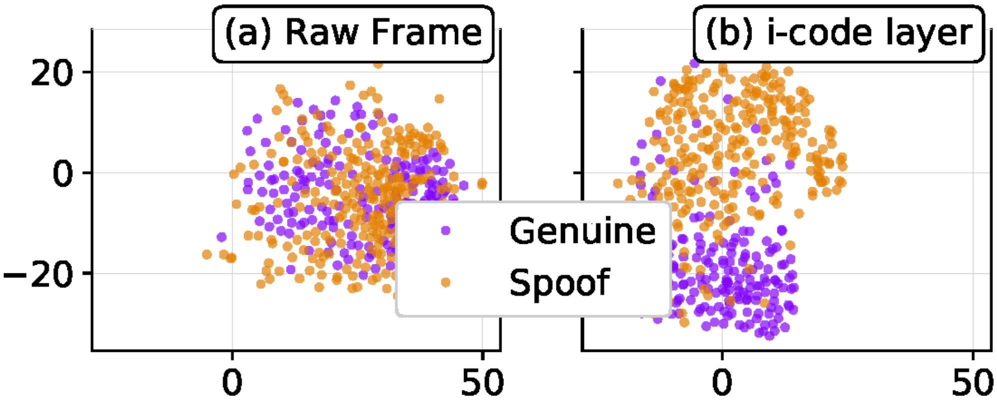
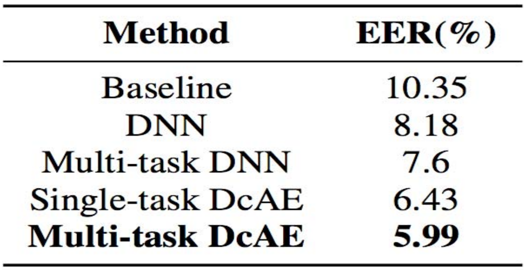


圖 3. 利用 t-SNE 表示在多任務鑑別式自編碼解碼器中，i-codes 的分佈情形。
[Figure 3. The result of multi-task DcAE that t-SNE maps high dimension i-codes into 2 dimension.]

表 1. 相等錯誤率在開發數據集上的結果。
[Table 1. Summary of development result in ASVspoof Task]



Method	**EER(%)**
Baseline	10.35
DNN	8.18
Multi-task DNN	7.6
Single-task DcAE	6.43
Multi-task DcAE	**5.99**

由表 1 可得知，由高斯混合模型作為 Baseline 方法的效果遠差於鑑別式自編碼解碼器所達到的效果，另外對於一般的深度類神經網路來說，我們的模型也能有較好得效果，在視覺化呈現的部分，如圖 2、圖 3，我們利用了 t-Distributed Stochastic Neighbor Embedding(t-SNE) (van der Maaten & Hinton, 2008) 來表示，由此圖可發現，經由鑑別層所產生出來的 i-codes 確實達到了使輸入數據依照目標分離的效果，由此可見，經由簡單的向量內積來計算即可快速的分辨輸入語料是否為真實人聲。

5. 結論 (Conclusion)

在這篇論文中，我們提出了一個全新的鑑別式自編碼解碼器來參與這次的ASVspoof-2017，在這個新的架構中，我們加入了新設計的目標函數，使得我們可以利用新的特徵表示法 i-codes 來達到辨別是否為真實人聲的目標，另外，我們同時利用了多任務模型來增強預測的結果，最後，相比於高斯混合模型以及深度類神經網路，鑑別式自編碼解碼器更能夠使資料具有辨別的價值。在未來，我們希望能夠加以延伸發展此模型，以建立一個更泛用的系統架構。

致謝 (Acknowledgement)

特別感謝由法務部調查局所提供的贊助來完成此篇論文研究(Grant number: 106-1301-05-04-01-3)。

參考文獻 References

Abe, M., Nakamura, S., Shikano, K. & Kuwabara, H. (1990). Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, *11*(2), 71-76.

Alam, M. J., Kenny, P., Bhattacharya, G. & Stafylakis, T. (2015). Development of CRIM system for the automatic speaker verification spoofing and countermeasures challenge 2015. In *Proceedings of Interspeech 2015*, 2072-2076.

Alegre, F., Amehraye, A. & Evansdoi, N. (2013). Spoofing countermeasures to protect automatic speaker verification from voice conversion. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP.2013.6638222

Bone, D., Lee, C.-C. & Narayanan, S. (2014). Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features. *IEEE Transactions on Affective Computing*, *5*(2), 201-213. doi: 10.1109/TAFFC.2014.2326393

Chen, L.-H., Ling, Z.-H., Liu, L.-J. & Dai, L.-R. (2014). Voice conversion using deep neural networks with layer-wise generative training. *IEEE/ACM Transactions on Audio, Speech and Language Processing(TASLP)*, *22*(12), 1859-1872. doi: 10.1109/TASLP.2014.2353991

Chen, Y.-N., Sun, M., Rudnicky, A. I. & Gershmandoi, A. (2016). Unsupervised user intent modeling by feature-enriched matrix factorization. In *Proceedings of 2016 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6150-6154. doi:10.1109/ICASSP.2016.7472859

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y. & Lee, L.-S. (2016). Audio Word2Vec: Unsupervised Learning of Audio Segment Representations Using Sequence-to-Sequence Autoencoder. In *Proceedings of Interspeech 2016*. doi:10.21437/Interspeech.2016-82

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of The Thirteenth International Conference on Artificial Intelligence and Statistics*, *9*, 249-256.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT press.

Huang, K.-Y., Wu, C.-H., Su, M.-H. & Fu, H.-C. (2017). Mood detection from daily conversational speech using denoising autoencoder and LSTM. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5125-5129. doi:10.1109/ICASSP.2017.7953133

Kenny, P., Gupta, V., Stafylakis, T., Ouellet, P. & Alam, J. (2014). Deep neural networks for extracting Baum-Welch statistics for speaker recognition. In *Proceedings of Odyssey 2014*, 293-298.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representation*. Retrieved from https://sarXiv preprint arXiv:1412.6980

Kinnunen, T., Evans, N., Yamagishi, J., Lee, K. A., Sahidullah, Md., Todisco, M. & Delgado, H. (2017). ASVspoof 2017: automatic speaker verification spoofing and countermeasures challenge evaluation plan. *Training, 10*, 1508.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in neural information processing systems*, 1106-1114.

Lee, H.-S., Lu, Y.-D., Hsu, C.-C., Tsao, Y., Wang, H.-M. & Jeng, S.-K. (2017). Discriminative autoencoders for speaker verification. In *Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5375-5379. doi:10.1109/ICASSP.2017.7953183

Lei, Y., Scheffer, N., Ferrer, L. & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP.2014.6853887

Lin, W.-w., Mak, M.-W. & Chien. J.-Z. (2017). Fast scoring for PLDA with uncertainty propagation via i-vector grouping. *Computer Speech & Language*, *45*, 503-515. doi:10.1016/j.csl.2017.02.009

Liu, X., Gao, J., He, X., Deng, L., Duh, K. & Wang, Y.-Y. (2015). Representation learning using multi-task deep neural networks for semantic classification and information retrieval. In *proceedings of HLT-NAACL 2015*.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O. & Kaiser, L. (2015). Multi-task sequence to sequence learning. In *Proceedings of ICLR 2016*. Retrived from https://arXiv preprint arXiv:1511.06114

Richardson, F., Reynolds, D. & Dehakdoi, N. (2015). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), 1671-1675. doi:10.1109/LSP.2015.2420092

Sarkar, A. K., Do, C.-T., Le, V.-B. & Barras, C. (2014). Combination of cepstral and phonetically discriminative features for speaker verification. *IEEE Signal Processing Letters*, *21*(9), 1040-1044. doi: 10.1109/LSP.2014.2323432

Todisco, M., Delgado, H. & Evans, N. (2016). A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients. In *Proceedings of Odyssey 2016*. doi: 10.21437/Odyssey.2016-41

van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579-2605.

Van Santen, J. P. H., Sproat, R., Olive, J. & Hirschberg, J. (2013). *Progress in speech synthesis*. New York, NY: Springer Science & Business Media.

Variani, E., Lei, X., McDermott, E., Lopez Moreno, I. & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4080-4084. doi: 10.1109/ICASSP.2014.6854363

Villalba, J., Miguel, A., Ortega, A. & Lleida, E. (2015). Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In *Proceedings of Interspeech 2015*, 2067-2071.

Wu, Z., Kinnunen, T., Evans, N. W. D., Yamagishi, J., Hanilci, C., Sahidullah, M. & Sizov, A. (2015). ASVspoof 2015 - the first automatic speaker verification spoofing and countermeasures challenge. In *Proceedings of Interspeech 2015*.

Xiao, X., Tian, X., Du, S., Xu, H., Siong, C. E. & Li, H. (2015). Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge. In *Proceedings of Interspeech 2015*.

Yamada, T., Wang, L. & Kai, A. (2013). Improvement of distant-talking speaker identification using bottleneck features of DNN. In *Proceedings of Interspeech 2013*.

Yang, M.-H., Lee, H.-S., Lu, Y.-D., Chen, K.-Y., Tsao, Y., Chen, B. & Wang, H.-m. (2017). Discriminative autoencoders for acoustic modeling. In *Proceedings of Interspeech 2017*.

Ze, H., Senior, A. & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. doi: 10.1109/ICASSP.2013.6639215

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 73-86

以知識表徵方法建構台語聲調群剖析器[1]

A Knowledge Representation Method to Implement A Taiwanese Tone Group Parser

張佑竹*

Yu-Chu Chang

摘要

聲調群剖析器是台閩語語音輸出系統的主要元件之一。本文提出聲調管轄假說，主張先將句內語詞定調，亦能決定台閩語聲調群分界的觀點，並以聲調群剖析器實作加以驗證。除了敘述如何應用預設調型、預設詞類和模式三種標記符號，將語言知識和經驗轉換為知識庫，並說明經由推論引擎與知識庫的連結，完成語詞定調的運作過程。目前內部測試平均變調正確率為 98.5%。外部測試平均變調正確率為 94%。本研究的實驗數據也顯示一個重要的線索：符號系統標記比規則推論對變調正確性有相對較高的貢獻率。

關鍵詞：台灣話，變調，聲調群剖析器，知識表徵，模擬

Abstract

A tone group parser could be one of the most important components of the Taiwanese text-to-speech system. In this paper, we offered the hypothesis of tonal government to emphasis the idea that if the allotone selection can be made for each word in a sentence then the tone groups will be separated within the sentence and supported our viewpoint with the implementation of a Taiwanese tone group parser. In addition to the description of using the symbol system to convert language expertise and heuristic knowledge into a knowledge base to cope with a frame-based corpus and a tone sandhi processor, the procedure of connecting the

[1] 本論文承蒙中正大學語言學研究所麥傑教授提供諸多建議，謹此致謝。

* 國立中正大學語言學研究所
Institute of Linguistics, National Chung Cheng University
E-mail: poirotdavid@yahoo.com.tw

inference engine and the knowledge base to make allotone selection was also discussed. In the current version of the tone group parser, the average accuracy of inside test is 98.5%. The average accuracy of outside test is 94%. The experiment data of the study also reveals an important clue: the marking of the symbol system makes a higher contribution rate to the tone sandhi accuracy than the rule inference.

Keywords: Taiwanese, Tone Sandhi, Tone Group Parser, Knowledge Representation, Simulation

1. 緒論 (Introduction)

聲調群(Tone groups)是台閩語[2]的基本韻律結構。聲調群剖析器也是台閩語語音輸出系統的主要元件之一(Liim, 2004; 田村志津枝, 2010)。本文首先從音韻和語言結構的觀點探索台閩語的特質，將台閩語如何藉著變調(Tone sandhi)形成獨特聲調群結構的衍生過程和當代音韻-句法界面的相關研究相互印證，提出聲調管轄假說(Tonal-government hypothesis)來主張先將句內語詞定調，亦能決定聲調群分界的觀點。隨後說明以台閩語變調習得(Tone sandhi acquisition)理論為基礎，將知識表徵(Knowledge representation)技術和語詞屬性分析加以整合，製作聲調群剖析器的方法。

2. 文獻回顧 (Literature Review)

2.1 從音韻和語言結構觀點看台閩語特質 (The Nature of Taiwanese Language from Phonological and Structural View)

台閩語為聲調語言。變調是指語詞聲調受到相鄰語詞影響而改變聲調調值的現象，常見於中國各地的語言。比較特別的是台閩語語詞具有普遍的變調現象。每個台閩語語詞皆有本調[3](Lexical tone)和變調(Sandhi tone)兩種調型(Tone form)。語詞或聲調群的最後音節讀本調，其餘音節讀變調(Chiu, 1931; 王育德, 1955)。因此，若且唯若一個或一組語詞中僅有最後一個音節讀本調，則此一語詞或詞組即為聲調群。換句話說，台閩語語句就是聲調群的集合。聲調群不僅是組成台閩語語句的句法單元(Syntax unit)，同時也是完整的語義單位(Semantic unit)和韻律結構(Prosodic structure)。台閩語很可能是唯一在語句內以變調方式建立聲調群結構的自然語言(Chang, 2009)。

[2] 台灣話源自中國閩南方音。公元 2006 年教育部公告台灣閩南語羅馬字拼音方案，台灣話得以文字公開傳承。本文例句，採用教育部台灣閩南語（以下簡稱台閩語）拼音加註聲調值。引述論文中的 Taiwanese 或台灣語也併譯為台閩語。

[3] 本調 Lexical tone 亦稱 juncture tone，變調 sandhi tone 亦稱 context tone。

2.1.1 台閩語變調與聲調群的形成 (Taiwanese Tone Sandhi and the Formation of Tone Group)

台閩語的聲調、語義和句法之間有密切的關係。聽話者可以從相同的語句中以聲調來辨別不同的詞類(POS)和語義。例如，本調相同的「ke^{55}(雞/加)」在例句(1)和(2)裡，有不同的調型、詞類和語義。

(1) Tsit^{31} tsiah^{31}（變調）ke^{55}（本調，名詞）tsit^{54} kong^{55}-kin^{55}.（這隻雞重一公斤。）

(2) Tsit^{31} tsiah^{31}（本調）ke^{55}（變調，動詞）tsit^{54} kong^{55}-kin^{55}.（這隻多重一公斤。）

這個例句讓我們注意到人腦雖然可以就台閩語同音異形漢字「雞/加」和語境來分析語義和句法結構，但是對電腦而言，(1)和(2)的羅馬拼音完全相同，只有自主語義(Autonomous semantic mapping)確定以後，才能決定「ke^{55}」的調型或進行句法分析。這部分屬於高階人工智能(Strong AI)的範疇，也是對話系統必須面對的困境。

例句(3)和(4)說明數量詞「tsit^{32} king^{55}（這間）」的調型視前後文語境而定。

(3) Tsit^{32} king^{55}（本調）u^{33} tsai^{55} hue^{55}.（這間房屋有種花。）

(4) Tsit^{32} king^{55}（變調）u^{33} tsai^{55} hue^{55} e^{23} tshu^{31} si^{33} guan^{53} tau^{55}.（這間有種花的房屋是我家。）

從例句(5)(6)(7)和(8)可以觀察經由不同形式的插入方式所造成的聲調群變化(Chang, 2009)。不論插入的形式為何，每個台閩語語句最終都會形成以聲調群組合而成的結構。

(5) [A^{55}-bi^{53}] [beh^{32} khi^{31} Tai^{23}-pak^{32}.]（阿美要去台北。）有兩個聲調群[4]。

插入變調語詞「siunn^{33}（想）」後，聲調群數目不變。

(6) [A^{55}-bi^{53}] [siunn^{33} beh^{32} khi^{31} Tai^{23}-pak^{32}.]（阿美想要去台北。）

插入本調語詞「pai^{31}-it^{32}（星期一）」後，聲調群數目增加為三個。

(7) [A^{55}-bi^{53}] [pai^{31}-it^{32}] [beh^{32} khi^{31} Tai^{23}-pak^{32}.]（阿美星期一要去台北。）

[4] 符號［ ］表示聲調群的分界。

插入聲調群「tse^{33} gu^{23}-tshia^{55}（坐牛車）」後，聲調群數目增加為三個。

(8) [A^{55}-bi^{53}] [beh^{32} tse^{33} gu^{23}-tshia^{55}] [khi^{31} Tai^{23}-pak^{32}.]（阿美要坐牛車去台北。）

就句法分析的觀點而言，聲調群必定是 XP(X Phrase)，然而並非所有的 XP 都是聲調群。聲調群可能是可以轉換為 XP 的先驅結構。

2.1.2 音韻-句法界面 (The Phonology-syntax Interface)

間接指涉假說(Indirect reference hypothesis)指出音韻規則並非由句法直接影響，而是經由韻律結構做為連結音韻和句法的媒介(Selkirk, 1986)。這種現象在台閩語尤其明顯。語言學習者為了習得句法結構的資訊而應用韻律訊息(Prosodic cue)建立韻律結構。如果兒童可以在台閩語句子中標記出聲調群的位置，就可習得有用的句法相關知識(Tsay, 1999)。聲調群必定是台閩語語言習得的重要線索。

2.2 語言習得的模擬 (The Simulation of Language Acquisition)

Norman Geschwind 對於語言功能如何在大腦皮層的特定區域運作，指出腦皮質裡至少有兩個區塊對語言能力有重大影響；這些區塊被精確規劃用來處理言語資訊(Geschwind, 1979)。即使他關於語言能力主要依賴左半區的理論或有爭議，人類在大腦記憶裡儲存語彙的功能則毋庸置疑。嬰幼兒從語音感知(Perception)中學習語言。及至成長，這種感知機制仍然存在(Eimas, 1985)。音節、語詞、片語或韻律單位很可能儲存在腦皮質的記憶區塊。因此我們假設台閩語語者在學習母語的過程中，將語詞的詞類和調型標記在記憶裡。

台閩語語詞當中有的語詞讀變調，有的語詞讀本調，也有為數不少的語詞需視詞類和語境才能定調。若是依照單音節語詞讀變調，複音節語詞讀本調的單一規則將一篇文章粗略定調，可以得到大約 70%的變調正確率。實務上，語詞調型的選擇往往和語詞的詞類、相鄰語詞和變調規則有關。這類需要規則處理的語詞，通常也是語者用來決定是否延伸語意的工具；語詞讀變調，表示該語詞指涉的語意尚待完成。讀本調的語詞則是聲調群的分界點，也是一個完整語意單位的結束。某些語詞要讀本調或變調通常取決於說話者，他必須在說話前的瞬間作出反應。

韻律導引假說(The prosodic bootstrapping hypothesis)說明兒童如何學會使用韻律訊息幫助自己界定聲調群，尋找句法結構並習得變調。這種技巧讓他得以交互使用語詞的本調和變調兩種形式(Tsay, 1999)。值得注意的事實是，以台閩語做為母語的使用者，即使無法察覺變調規則的存在或未曾認真學習句法，依然可以精確地處理變調並應用人腦剖析器辨識聲調群。同樣地，嬰兒在初學台閩語時，不只不認得語詞也不懂得句法結構。

2.3 知識表徵方法的應用 (The Application of Knowledge Representation Method)

Marvin Minsky 認為用來解決問題的系統(Problem-solving system)可能是認知過程的模擬。他提出框架理論的應用(Minsky, 1975)以後，知識表徵便成了人工智能研究的焦點技術。知識系統涵蓋知識庫(Knowledge base)、推論引擎(Inference engine)和開發介面三部分。知識庫包含目標(Goals)，規則(Rules)以及領域專業知識。推論引擎負責規則推論程序及策略管制(Chang, 1992)。開發介面則用來和使用者溝通或與其他系統做連結。

2.4 與本文相關的聲調群研究(Current Research of the Taiwanese Tone Groups)

現代學者在二十世紀中葉就注意到台閩語聲調群研究的重要性（王育德, 1955）。語言學家用句法分析來辨識台閩語聲調群(Cheng, 1968; Chen, 1987; Lin, 1994)。從語音實驗尋找台閩語變調作為韻律分界的證據(Tsay, Myers & Chen, 2000)或是經由研究台閩語鼻音化探討韻律階層的聲調群分界，提出聲調群分界是台閩語的韻律單位的主張(Pan, 2003)。資訊工程學者也試圖建立台閩語語音輸出系統(Liang, Yang, Chiang, Lyu&Lyu, 2004)或運用詞類標記和變調規則處理台閩語變調(Iunn, Lau, Tan-Tenn, Lee & Kao, 2007)。Pan 指出本調語詞是聲調群的分界。若要將句內語詞定調必須先界定聲調群(Pan, 2003)。我們則認為若能將句內所有的語詞定調，亦能界定聲調群，因此提出聲調管轄假說來說明台閩語聲調群、詞類與聲調調型間的關係。

3. 聲調管轄假說(Tonal-government Hypothesis)

Selkirk 的間接指涉假說指出句法結構並非直接限制音韻規則，而是透過韻律結構，作為媒介影響音韻的變化。其間的關係為句法->韻律結構->音韻(Selkirk, 1986)。然而台閩語語句由聲調群組成，詞類和聲調調型極有可能藉由變調規則改變韻律結構。例句(9)和(10)顯示，khuann^{31} 的不同調型，在詞類不變的情況下，形成不同的聲調群結構和語義。從例句(11)和(12)，也顯示單音節方位詞的前詞讀變調，複音節方位詞的前詞讀本調的規律性。lai^{33} 和 lai^{33}-te^{53}，因為音節數不同而影響前詞的調型和韻律結構。

(9) [Li^{53} khuann^{31}（動詞，讀本調）] [kam^{53} u^{33}]?（你覺得有沒有？）

(10) [Li^{53} khuann^{31}（動詞，讀變調）kam^{53} u^{33}]?（你看得懂嗎？）

(11) [Kong^{55}-hng^{23}（讀變調）-lai^{33}] [u^{33} tsit^{54} tsiah^{32} kau^{23}].（公園裡有一隻猴子。）

(12) [Kong^{55}-hng^{23}(讀本調)] [lai^{33}-te^{53}] [u^{33} tsit^{54} tsiah^{32} kau^{23}].(公園裡面有一隻猴子。)

雖然音韻直接影響句法結構的情況在台閩語裡並非常態，聲調變化造成韻律結構改變的現象卻屢見不鮮。從圖 1 可以看出除了韻律結構影響音韻的變化之外，詞類、變調規則、聲調調型也會影響或主導韻律結構的形成。值得注意的是在聲調管轄假說涵蓋的

範域內，韻律結構與構成音韻變化的因素間存在著明顯的遞迴(Recursive)現象。以下的章節將摘要敘述聲調群剖析器的實作方法，藉以驗證聲調管轄假說存在的可能性。

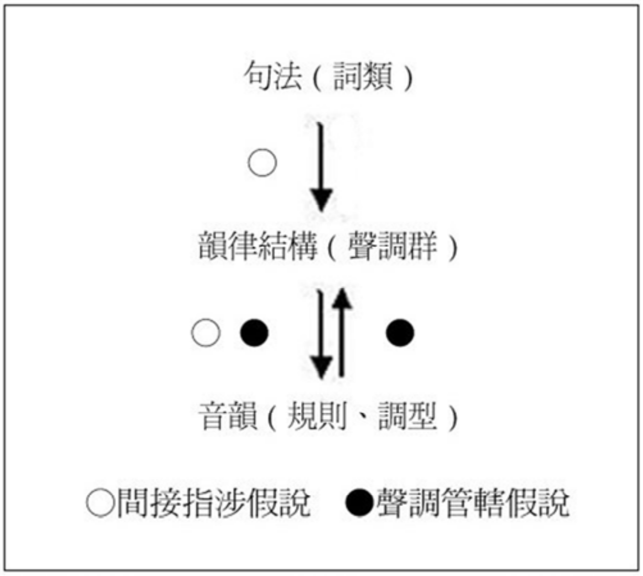


圖 1.台閩語句法、韻律結構和音韻間的關係及相關假說的適用範域
[Figure 1. The relationship among syntax, prosodic structure and phonology in Taiwanesewith the related hypotheses]

4. 從台閩語語句擷取聲調群的方法(The Method to Capture the Tone Groups from the Taiwanese Sentences)

Tsay 對於台閩語變調習得的論述凸顯回饋機制對模擬系統的重要性。Pan 的實驗明確指出聲調群是台閩語習得的重要關鍵。這些研究激發我們以個人電腦製作人工聲調群剖析器的構想。至於應用符號系統尋找韻律訊息，以聲調群解決多重 POS 語詞變調問題的靈感則來自 Selkirk 的間接指涉假說。

我們的構想是，一旦句內語詞被賦予正確調型，讀本調的語詞就是聲調群分界。實作方法就是以台閩語變調習得、間接指涉假說和聲調管轄假說等論述做為基礎，採取預設聲調調型為主,預設詞類及前後詞調型模式為輔的規則推論策略,將句內語詞定調後，從台閩語語句擷取聲調群。圖 2 是台閩語聲調群剖析器內，以知識庫為基礎的專家系統基本架構示意圖。系統由變調規則庫、框架語料庫和推論引擎組成。這個專家系統將被用來推論句內每個語詞的調型值。基於聲調群的生成有明顯的遞迴現象，系統從語句擷取的聲調群或聲調群前詞也能經由遞迴機制回饋至語料庫。

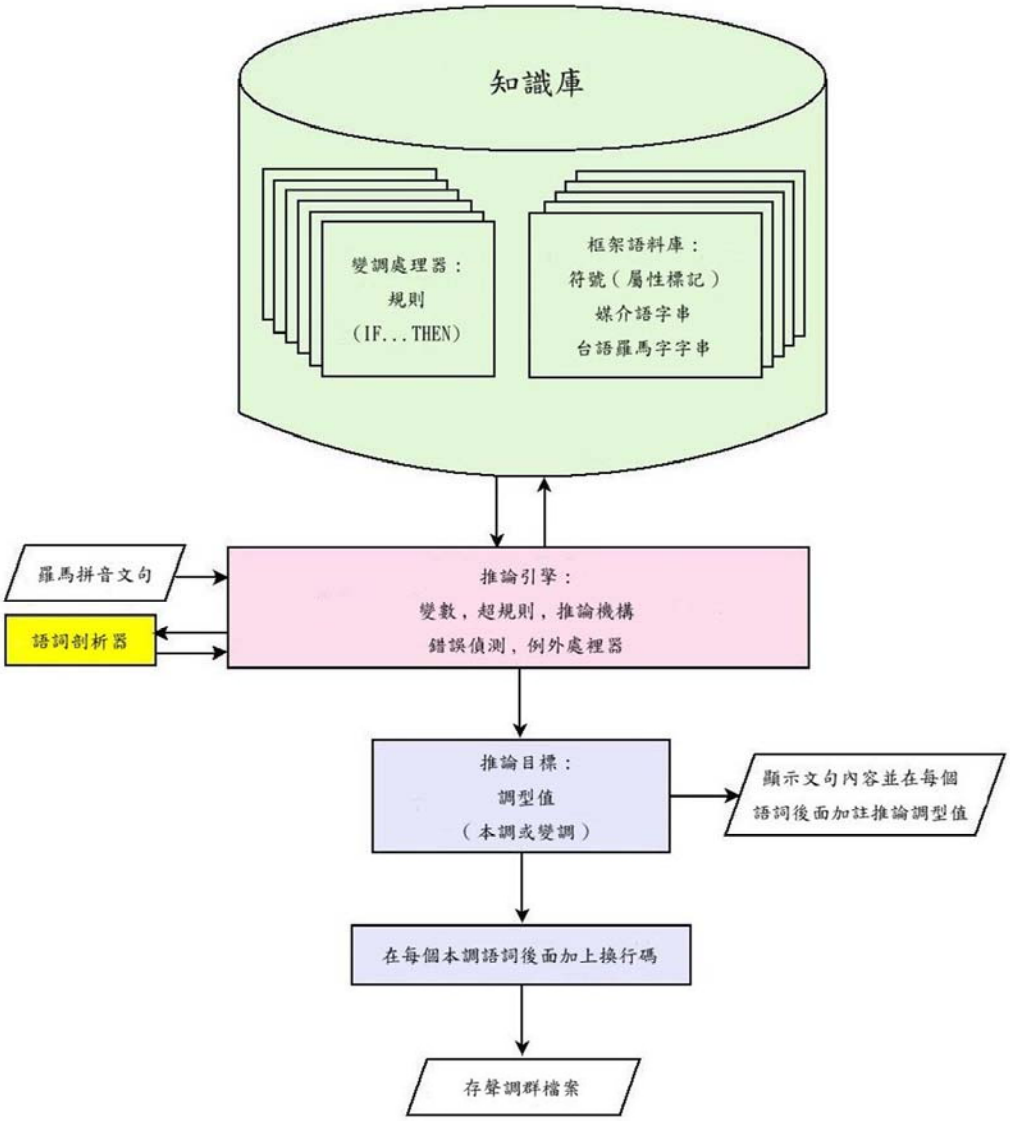


圖 2. 台閩語聲調群剖析器內以知識庫為基礎的專家系統基本架構示意圖
[Figure 2. Basic structure of the Taiwanese Tone group parser]

5. 製作過程摘要(The Schema of Implementation)

實作程式適用於個人電腦 Windows XP/Windows 7 作業系統[5]。製作過程摘要分述如下：

[5] Windows XP / Win7 相容的台閩語聲調群剖析器版本，可從 https://vikon.myweb.hinet.net/ttgpe.htm 下載。

5.1 語言專業知識的轉換 (The Transformation of Linguistic Expertise and Heuristic Knowledge)

語言專業知識和經驗主要用於規則庫和語料庫。所需語料從詞典、專業文獻和田野調查工作等資訊中萃取後，由台閩語專家和知識工程師應用符號系統將語料進行標記並建立變調規則庫。符號系統的設計概念，是在早期制定或修改變調規則的過程中所衍生的創意。目前建立的符號系統是以預設調型(Default mark of tone form)、預設詞類(Default POS)和模式(Mode)三種標記組成。語料庫內的每個資料錄都賦予一組包含這三種語詞屬性的符號。這組符號被用來連結語料庫和變調處理器(Tone sandhi processor)。處理台閩語變調時，藉著符號系統和規則推論，即使同音異義詞或兼有多種詞類的語詞也能經由變調處理程序，予以定調。語料庫和變調處理器建構完成時，語言專業知識也就轉換成為知識庫。下列章節將說明組成符號的三個語詞屬性以及應用推論引擎將符號連結語料庫和變調處理器的推論過程。

5.1.1 第一個屬性：預設調型記號 (The First Attribute: Default Mark of Tone Form)

台閩語有許多兼具兩種詞類的語詞，或由兩種詞類組合而成的組合詞等。這些語詞要讀本調或變調，需視前後詞及語詞在句內的關係位置而定。若不經由變調處理程序進行規則推論，無法定調。標記預設調型記號可以篩檢具有固定調型的語詞並排除不必要的規則推論。台閩語語詞預設調型記號及處理方式如表 1。

表 1. 台閩語語詞預設調號及處理方式
[Table 1. The list of default mark of tone form]

預設調型記號	預設調型	適用語詞	處理方式
0	固定讀本調	只讀本調的語詞	不需推論
1	預設讀變調	單音節語詞	以規則推論
2	預設讀本調	預設讀本調的語詞如詞組，輕聲詞，外國語	以規則推論
3	固定讀變調	只讀變調的語詞	不需推論
#	本調或變調	可能讀本調或變調的語詞	以規則推論
&	固定讀本調	聲調群或聲調群集合	不需推論

5.1.2 第二個屬性：預設詞類記號 (The Second Attribute: Default POS Mark)

台閩語有大量的組合詞，其組成元素與句法和構詞有密切關連。組合詞經由變調導致聲調的轉變，提供語者和聽者重要的訊息來區別不同的語義或句法結構。然而組合詞也讓台閩語詞類標記更加困難。因此我們使用定義較為寬鬆的預設詞類(DPOS)做為第二個屬性。預設詞類有 n（名詞/數詞）、v（動詞）、a（形容詞）、c（連接詞）、m（介詞）、d（副詞）、x（助動詞）、p（代名詞）、u（量詞）、s（語尾詞）、e（方位詞）、g

（動名詞）、k（連綴動詞）、&（聲調群）等標記。

5.1.3 第三個屬性：模式記號 (The Third Attribute: Mode Mark)

預設調型被標記為 1、2 或 # 的語詞或詞組需經規則推論。其中受到前後語詞影響的語詞無法經由預設調型和預設詞類定調。必需藉著一組應用二階及三階布林驗證(Boolean verification)的模式記號進行規則推論。二階模式記號有 a(-01)、b(-11)、c(-00)、d(-10)、e(10-)、f(11-)、g(00-)、h(01-)。三階模式記號有 j(000)、k(010)、m(101)、n(111)、p(001)、q(011)、r(100)、s(110)等。x 則用於不需模式記號的語詞。其中「0」代表本調，「1」代表變調，「-」用於二階模式以標識非相關前後詞位置。例如(-01)為本詞讀本調，後詞讀變調的二階模式。(101)為前詞及後詞讀變調，本詞讀本調的三階模式。

5.2 框架結構語料庫製作 (The Construction of a Frame-based Corpus)

現存的自然語言或因通行已久，或因約定成俗，無法以電腦做邏輯常規處理。因此必須建立電腦化的人工語言做為媒介。媒介語的符號和構詞法不受自然語言約束，也能建立與自然語言的映射(Mapping)機制。使用媒介語的好處是語詞的音節變調可以預行轉換，不需再經系統處理。

以物件-屬性-值(Object-attribute-value)來表徵知識是建構語料庫常見的方法。在框架語料庫裡，語詞或詞組可視為一個物件。描述聲調性質的預設調型則是物件的屬性，而賦予預設調型記號的「2」就是屬性的值。物件和其屬性間自然形成一種階層結構(Hierarchy structure)。因此台閩語語料庫可以採取一般資料庫的資料結構並將預設調型、預設詞類和模式合成一組符號。每個資料錄有符號、媒介語字串和以數字標調的台閩語羅馬字字串三個欄位。欄位間以逗號區隔，例如「2nx,kangte,kang1-te7」。符號標記由三個字母構成，「2nx」用於預設讀本調且不需布林驗證的名詞語詞或名詞組合詞。台閩語羅馬字串可以是單音節語詞、複音節語詞、片語、聲調群或詞組。媒介語則是和台閩語羅馬字對應的字串。語料庫內的所有語料經過詞頻統計、屬性和功能分析後，被依序存放到與大腦長程記憶類似的個別變數陣列，讓系統和規則可以隨時取用。

資料錄通常以台閩語羅馬字的音節數排序，音節數較多者優先。資料錄如何排序，在實務上頗為困難。統計常用語料的詞頻，或可作為排序的參考，但是兩者間並非唯一相關。由於台閩語構詞尚未標準化，音節連寫或分寫規則相當複雜，設計語料庫搜尋演算法時必須考慮構詞容錯機制，以確保進行推論程序時，系統得以順暢運作。

5.3 變調處理器的設計 (The Design of the Tone Sandhi Processor)

在實務上，變調處理器以語詞的多元屬性進行規則推論。所有規則必須預行分類並予優先定序。分類時依據相關詞類區分為數個主要區段(section)。系統先從第一區段開始推論，如有必要再推論的目標語詞，則轉往下一區段繼續進行。相關規則執行完畢後可直接跳到最後區段進行除錯、布林驗證或終結推論等動作。這種推論程序稱為正向連結(Forward chaining)。所有區段內的規則可用迴圈執行若干次，直到字串陣列裡的目標語詞逐一經過推論後，取得適當的調型值。

由於部分規則會處理語詞前後詞間的關係，因此在推論過程中無可避免地，會受到規則間的交互影響而改變已推論的前後詞調型值。規則多寡也會影響剖析器的執行效率與變調正確率。規則越多，推論所需時間越長，交互影響也更顯著。所以語料庫或規則庫更新時都必須執行內部測試(Inside test)，以免造成顧此失彼的窘境。

5.4 推論引擎與知識庫間的運作過程 (The Operating between Inference Engine and Knowledge Base)

推論引擎主要用來存取語料庫資訊、啟動及控制變調處理程序。內建的搜尋演算法，可將指定的羅馬拼音文句和語料庫的語詞做比對，轉換成媒介語文句。這些文句再經語詞剖析器切割為媒介語字串存入陣列。每個媒介語字串可從記憶體（變數陣列）取得相關屬性值並被賦予一個目標參數，也就是系統要進行推論的預設調型。參數取值的方法不外採用預設值、由系統推論取值或從語料庫取值。變調處理器被推論引擎呼叫時，字串陣列內的媒介語就會依序進行推論程序。推論機制開始就語料庫提供的資訊和變調處理器各個規則的條件部分進行比對。當規則的 IF 部分與相關資訊相符，則 THEN 部分的指令就被執行。如果變調處理器已經沒有其他規則被啟動，推論程序就中止並完成目標推論。隨後開始下一個字串的推論程序，直到字串陣列所有媒介語都完成推論作業，並將目標參數推論值回傳推論引擎。系統進行推論時可能遭遇資訊不足的情況，這時具有較高優先的超規則(Metarules)可從變調處理器直接偵錯、設定或變更相關屬性值。推論引擎預設的超規則也可研判是否需要修改變調處理器內的推論結果。這種功能通常用來修正錯誤的推論程序或處理例外狀況。

5.5 語意識別、構詞容錯與機器學習 (Semantic Identification, Fault Tolerance and Machine Learning)

語意識別的方法不外歧義消除或多義選一。台閩語有部分語詞可用一般規則來定調。例如「gah^{32}」的前詞都讀變調。同音異義詞如「ti^{33}」，雖然兼有名詞（箸）和介詞（在）的多重詞類屬性，仍然能用相關規則來區別詞類，予以定調。然而對於同形異音異義的詞組或片語，例如「e^{23}e^{23}」有「的鞋」或「鞋的」兩種不同的語音和語義，必須具備修改規則變數的能力，才能加以分辨。在測試變調處理器時，我們針對上述片語或詞組，制定規則讓機器研判上下文來選擇正確的語音，完成定調。下面的輸出例句是聲調群剖析器用來呈現初階人工智能(Weak AI)的部分範例。(1)代表變調，(2)代表本調。

(13) Tsit^{32} siang^{55} (1) e^{23} (2) e^{23} (1) e^{23}-bin^{33} (2) si^{33} (1) nng^{23}-a^{53}-phue^{23} (2) tso^{31}--e^{23} (2).
（這雙鞋的鞋面是二榔皮製成的。）

(14) Tsit^{32} siang^{55} (1) e^{23}-bin^{33} (2) si^{33} (1) nng^{23}-a^{53}-phue^{23} (2) e^{23} (1) e^{23} (2) si^{33} (1) gua^{53}-e^{23} (2).（這雙鞋面為二榔皮的鞋子是我的。）

由於台閩語語詞連寫或分寫常會影響變調的推論結果。建立構詞容錯機制也是剖析器設計的重點之一。至於如何讓機器學習研判語境，不論是應用統計機率分析或是直接以規則引導，都必須依賴大量的知識、人力與計算資源，目前只能進行局部實驗。我們無法確知兒童的台閩語變調習得是否全然依循經驗法則，然而(13)(14)的定調結果，或可實證台閩語聲調群剖析器經由變調規則來分辨語音，繼而選擇語義的可行性。

5.6 功能測試與實驗結果 (Function Testing and the Result of Experiment)

一旦句內語詞被指定為本調，聲調群即可被切割出來，所以讀入新的文章就能產生新的聲調群或聲調群前詞做為語料庫的回饋單元。由於聲調群不需推論即可定調，因此回饋機制能夠提昇聲調群剖析器的變調正確率和執行效率。聲調群剖析器得到的回饋愈多，變調正確率也愈高。就像兒童初學語言一樣，藉著遞迴回饋機制，聲調群剖析器可以不斷地進化。這種設計在實務上也能用來驗證(Tsay, 1999)和(Pan, 2003)關於聲調群的論述。

本研究的測試程序有兩種，一種是針對特定語詞或規則設計的除錯測試，另一種程序是針對剖析器進行的正確率與整體效率測試。內部測試語料包括十篇一般文章和五篇用來測試特定語詞和規則的文句。外部測試(Outside test)語料來自隨機擷取的國小台閩語課本口語語句。除了語音判定以外，受測文章的語詞也加註推論調型值，可以計算變調正確率。目前內部測試平均變調正確率為 98.5%。外部測試平均變調正確率為 94%。

程式開發期間，我們以內部測試語料做為初次回饋試材來更新知識庫。當內部測試平均變調正確率接近 98.5%或變調正確率開始收斂時，同步進行兩種知識庫功能試驗。第一種測試只用語料庫的符號系統標記，不做規則推論。第二種測試完全不用知識庫，直接將單音節語詞標變調，複音節語詞標本調。下列兩個公式用來計算規則推論貢獻率和符號系統標記貢獻率。兩種測試都用相同的內部測試語料。

規則推論貢獻率 = 內部測試正確率 - 第一種測試正確率

符號系統標記貢獻率 = 第一種測試正確率 - 第二種測試正確率

一般文章第一種測試的平均變調正確率為 91.41%。第二種測試的平均變調正確率為 75.87%。實驗結果顯示語料庫的規則推論對變調正確率有 7.09%的貢獻率。符號系統標記則提供 15.54%的貢獻率。相關數據列於表 2。特定文句第一種測試的平均變調正確率為 86.33%。第二種測試的平均變調正確率為 60.35%。實驗結果顯示語料庫的規則推論對變調正確率有 12.17%的貢獻率。符號系統標記則提供 25.98%的貢獻率。相關數據列於表 3。

表 2. 一般文章變調測試實驗數據表
[Table 2. Tone sandhi experiment data for the general articles]

檔案 編號	第一種測試			第二種測試		
	正確語詞數 (A)	語詞總 (B)	正確率 (A/B)	正確語詞數 (C)	語詞總數 (D)	變調正確率 (C/D)
1	396	430	92.09 %	343	430	79.77 %
2	200	224	89.29 %	157	224	70.09 %
3	307	347	88.47 %	254	347	73.20 %
4	546	603	90.55 %	454	603	75.29 %
5	201	219	91.78 %	153	219	69.86 %
6	98	105	93.33 %	68	105	64.76 %
7	1006	1088	92.46 %	869	1088	79.87 %
8	607	669	90.73 %	508	669	75.93 %
9	178	203	87.68 %	153	203	75.37 %
10	613	654	93.73 %	487	654	74.46 %
加總	**4152**	**4542**	**91.41 %**	**3446**	**4542**	**75.87 %**

表 3. 特定文句變調測試實驗數據表
[Table 3. Tone sandhi experiment data for the special files]

檔案 編號	第一種測試			第二種測試		
	正確語詞數 (A)	語詞總數 (B)	正確率 (A/B)	正確語詞數 (C)	語詞總數 (D)	正確率 (C/D)
11	411	489	84.05 %	263	489	53.78 %
12	663	773	85.77 %	463	773	59.90 %
13	606	667	90.85 %	413	667	61.92 %
14	456	556	82.01 %	336	556	60.43 %
15	675	771	87.55 %	490	771	63.55 %
加總	**2811**	**3256**	**86.33 %**	**1965**	**3256**	**60.35 %**

由於特定文句內需要用規則推論的語詞所佔比率較一般文章為高，所以符號系統標記和規則推論對變調正確率的貢獻率也相對較高，與預期相符。此外，實驗數據也提供一個重要的線索：不論一般文章或特定文句都顯示符號系統標記比規則推論對變調正確率有相對較高的貢獻率。就台閩語聲調群的感知而言，長程記憶內的語詞訊息，可能比短期記憶的規則更重要，也更有效率。幼兒在台閩語習得的過程中，從語詞學會使用韻律訊息幫助自己界定聲調群的假說，也與我們的實驗結果相符。

6. 結論(Conclusion)

台灣人的小孩在學習母語的過程中，可以經由聲調群來習得句法結構方面的知識。合理的假設是當越來越多的語言知識累積在兒童的腦海時，一個高效率的語詞變調處理機制也逐步建構完成。台閩語聲調群剖析器可說是一種人工智能的實驗平臺。我們設計一個符號系統並加以改良，做為將語言專業知識和經驗轉換為知識庫的重要工具，用來建構台閩語語料庫和變調處理器，將變調習得的模擬功能與語音輸出系統連結並完成測試。這種應用語言學理論建構台閩語聲調群剖析器的方法，在實務上是以知識工程技術來建立變調習得的模擬環境。先前認為若能將句內所有的語詞定調，就能將聲調群從台閩語語句切割出來的構想得以實作完成，不僅見證人工智能發展工具可以協助人類探索語言的認知功能來瞭解語言習得的過程，同時也呈現聲調管轄假說的可能性。我們嘗試以有限的語料和規則處理無限的文句，然而受限於知識庫的規模與計算資源，目前這個聲調群剖析器還不能處理自主語義和部份語句的定調問題。本研究若能進行監督式學習(Supervised learning)的模擬，或是完成聲調群及聲調群前詞的自動回饋機制，做為變調錯誤的最終解決方案，將來或可提升智慧型機器人的語音輸出功能。

參考文獻(References)

Chang, T. Y. (1992). *A multimedia-based bilingual instructional system using an expert system shell*. (Unpublished master's thesis). University of Central Missouri, Warrensburg, MO.

Chang, Y. C. (2009). An introduction to Taiwanese Speech Notepad. Retrieved from https://archive.org/details/TaiwaneseSpeechNotepadenglishVersion.

Chen, M. Y. (1987). The syntax of Xiamen tone sandhi. *Phonology*, *4*(1), 109-149.

Cheng, R. (1968). Tone sandhi in Taiwanese. *Linguistics*, *41*, 19-42.

Chiu, B. M. (1931). The phonetic structure and tone behaviour in Hagu (commonly known as the Amoy dialect) and their relation to certain questions in Chinese linguistics. *T'oung Pao*, *28*(1), 245-342. doi: 10.1163/156853231X00105

Eimas, P. D. (1985). The perception of speech in early infancy. *Scientific American*, *252*(1), 46-52.

Geschwind, N. (1979). Specialization of the human brain. *Scientific American*, *241*(3), 180-199.

Iunn, U. G., Lau, K. G., Tan-Tenn, H. G., Lee, S. A., & Kao, C. Y. (2007). Modeling Taiwanese Southern-Min Tone Sandhi Using Rule-Based Methods. *International Journal of Computational Linguistics and Chinese Language Processing, 12*(4), 349-370.

Liang, M. S., Yang, R. C., Chiang, Y. C., Lyu, D. C., & Lyu, R. Y. (2004). A Taiwanese text-to-speech system with applications to language learning. In *proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04),* 91-95. doi: 10.1109/ICALT.2004.1357381

Lin, J. W. (1994). Lexical government and tone group formation in Xiamen Chinese. *Phonology, 11*(2), 237-276. doi: 10.1017/S0952675700001962

Liim, K. (2004). Medical Education and Research in Taiwanese Language Since 1990. Paper presented at the *Symposium on Medical Taiwanese*, Kaohsiung Medical University.

Minsky, M. (1975). A framework for representing knowledge. In P. Winston (Ed.), *The Psychology of Computer Vision*. 211-277. New York, NY: McGraw-Hill.

Pan, H. H. (2003). Prosodic hierarchy and nasalization in Taiwanese. In *Proceedings of the 15th ICPhS*, 575-578.

Selkirk, E. O. (1986). *Phonology and syntax: the relationship between sound and structure*. Cambridge, MA: MIT press.

Tsay, J. (1999). Bootstrapping into Taiwanese tone sandhi. In *Chinese Languages and Linguistics V, Symposium Series of the Institute of History and Philology*, *2*(5), 311-333. Taipei, Taiwan: Academia Sinica.

Tsay, J., Myers, J., & Chen, X. J. (2000). Tone sandhi as evidence for segmentation in Taiwanese. In *Proceedings of the 30th Child Language Research Forum*. 211-218.

王育德（1955）。台灣語の聲調。*中國語學*，*41*，3-11。[Ong, I.T. (1955). Taiwanese Tones. *Journal of Chuugoku Gogaku*, *41*, 609-617.]

田村志津枝（2010）。*初めて台湾語をパソコンに喋らせた男—母語を蘇らせる物語*。東京：現代書館。[Tamura, S.(2010).*Hajimete Taiwango o pasokon ni shaberaseta otoko: bogo o yomigaeraseru monogatari*.Tokyo, Japan: Gendai Shokan.]

Computational Linguistics and Chinese Language Processing
Vol. 22, No. 2, December 2017, pp. 87-98

A Novel Trajectory-based Spatial-Temporal Spectral Features for Speech Emotion Recognition

Chun-Min Chang[*], **Wei-Cheng Lin**[*] **and Chi-Chun Lee**[*]

Abstract

Speech is one of the most natural form of human communication. Recognizing emotion from speech continues to be an important research venue to advance human-machine interface design and human behavior understanding. In this work, we propose a novel set of features, termed trajectory-based spatial-temporal spectral features, to recognize emotions from speech. The core idea centers on deriving descriptors both spatially and temporally on speech spectrograms over a sub-utterance frame (e.g., 250ms) - an inspiration from dense trajectory-based video descriptors. We conduct categorical and dimensional emotion recognition experiments and compare our proposed features to both the well-established set of prosodic and spectral features and the state-of-the-art exhaustive feature extraction. Our experiment demonstrate that our features by itself achieves comparable accuracies in the 4-class emotion recognition and valence detection task, and it obtains a significant improvement in the activation detection. We additionally show that there exists complementary information in our proposed features to the existing acoustic features set, which can be used to obtain an improved emotion recognition accuracy.

Keywords: Emotion Recognition, Speech Processing, Spatial-Temporal Descriptors, Mel-Filter Bank Energy

1. Introduction

The blooming of research effort in affective computing (Picard, 1997) in the past decade has started to enable machines to become capable toward sensing and synthesizing emotional expressive behaviors. Numerous technological applications, e.g., advanced human-machine interface (Bach-y Rita & Kercel, 2003; Swartout *et al.*, 2006) and interactive robotic design(Hollinger *et al.*, 2006; Hogan, Krebs, Sharon & Charnnarong, 1995), and even

[*] Department of Electrical Engineering, National Tsing Hua University
E-mail: cmchang@gapp.nthu.edu.tw; winston810719@gmail.com; cclee@ee.nthu.edu.tw

emerging cross-cutting research fields, e.g. social signal processing (Vinciarelli, Pantic & Bourlard, 2009) and behavioral signal processing (Narayanan & Georgiou, 2013), have all benefited from the vast amount of research advancements in affective computing. Speech is the most natural form of human communication that encodes both linguistic content and paralinguistic information (Schuller *et al.*, 2013), e.g., emotion (Nwe, Foo & De Silva, 2003; Scherer, 2003), gender (Childers & Wu, 1991), age (Dobry, Hecht, Avigal & Zigel, 2011), personality (Mairesse & Walker, 2006), etc. Development of suitable algorithms to robustly model emotional content in speech continues to be a prevalent topic in emotion recognition research.

There exists a vast amount of research in modeling speech acoustics for emotion recognition, topics ranging from lowlevel feature engineering, machine learning algorithms, to even joint feature-label representations (Calvo, D'Mello, Gratch & Kappas, 2014; Lee & Narayanan, 2005; Mower, Matarić & Narayanan, 2011). In this work, we aim at proposing a new set of long-term low-level features, named trajectory-based spatial-temporal spectral features, derived directly from the speech spectrograms to perform emotion recognition. Most of the current speech-based emotion recognition rely on extracting a set of commonly-used short-durational features (acoustic low-level descriptors - LLDs), e.g., those could be related spectral features (e.g., MFCCs), prosodic characteristics (e.g., pitch intonation), voicing quality (e.g., jitter), Teager-energy operater etc (Schuller *et al.*, 2007). Then, depending on the choice of emotion recognition framework, researcher would either apply global statistical functionals to be used in statics discriminative framework (e.g., support vector machine (Campbell, Sturim & Reynolds, 2006) or deep neural network (Kim, Lee & Provost, 2013) or using time-series model on these short durational low-level descriptors (e.g., hidden Markov model (Nwe *et al.*, 2003; Li *et al.*, 2013) in order to incorporate the feature's temporal characteristics to perform utterance-level emotion recognition.

Our proposed features are inherently different with the underlying inspiration coming from the dense trajectory-based video descriptors extraction approach (Wang, Kläser, Schmid & Liu, 2013). Dense trajectory video descriptors are extracted by first densely tracking important points on images over a frame (usually 0.5 - 1s) to forms a set of trajectories. The spatial-temporal descriptors for each trajectory can then be computed to obtain the final set of features. By modeling both the trajectory's temporal course and spatial changes over time, these descriptors have been shown to obtain superior improvement in tasks such as event (Oneata, Verbeek & Schmid, 2013) and motion (Wang, Kläser, Schmid & Liu, 2011) recognition than other key-points based image feature extraction. Our core concept, hence, centers around treating an audio file essentially as a sequence of spectrograms. Then, we compute a suite of spatial-temporal descriptors for each trajectory, i.e., a trajectory refers to a spectral energy profile across a time-frame (i.e., 250ms) of a Mel-filter bank output (MFB). In this work, we utilize these descriptors, i.e., trajectory-based spatial-temporal spectral features, to perform speech-based emotion recognition.

To the best of our knowledge, vast majority of the works in the speech emotion recognition literature have utilized short durational (25ms) LLDs which do not share the same concept with our proposed features. There are a few works that utilized auditory perception-inspired modulation spectral features (Chaspari, Dimitriadis & Maragos, 2014; Chi, Yeh & Hsu, 2012), i.e., temporal characteristic of spectral energy, for emotion recognition; these modulation spectrum features have been demonstrated to be robust under noisy conditions compared to features such as MFCCs and fundamental frequencies. In this work, we perform utterance-level categorical (4-emotion classes) and dimensional (valence and activation) emotion recognition on the USC IEMOCAP database (Lee, Mower, Busso, Lee & Narayanan, 2011). We additionally construct two set of features to compare our trajectory-based spatial-temporal spectral features (Traj-ST) to:

- *Conv-PS*: applying statistical functionals over a frame of conventional acoustic feature set
- *OpEmo-Utt*: state-of-the-art exhaustive utterance-level feature extraction using the OpenSmile toolbox (Eyben, Wöllmer & Schuller, 2010)

Our proposed features obtain comparable unweighted average recall on the task of 4-class emotion recognition and significantly outperform on the task of activation recognition compared to *Conv-PS* and *OpEmo-Utt*. Furthermore, by fusing *Traj-ST* with either *Conv-PS* and/or *OpEmo-Utt*, we achieve an improved recognition rate for the 4-class emotion recognition. It demonstrates the complementary information that our proposed features possess when combining with the well established acoustic features for emotion recognition. The rest of the paper is organized as follows: section 2 describes the database and the trajectory-based spatial-temporal spectral features, section 3 describes experimental setups and results, and section 4 is the conclusion and future work.

2. Research Methodology

2.1 The USC IEMOCAP Database

We utilize a well-known emotion database, the USC IEMOCAP database (Busso *et al.*, 2008), for this work. The database consists of 10 actors grouping in pairs to engage in dyadic face-to-face interactions. The design of the dyadic interactions is meant to elicit natural multimodal emotional displays from the actors. The utterances are annotated with both categorical emotion labels (e.g., angry, happy, sad, neural, etc) and dimensional representations (e.g., valence, activation, and dominance) on the scale of 1 to 5. The categorical labels per utterance are annotated by at least 3 raters, and the dimensional attributes are annotated by at least 2 raters. Given the spontaneous nature of this database and the inter-evaluator agreement is about 0.4, this database remains to be a challenging emotion database for algorithmic advancement. In this work, we conduct two different emotion recognition tasks on this database: 1) four-class emotion recognition 2) three-levels of valence and activation dimension recognition. For the categorical emotion recognitions, the four emotion classes are happy, sad, neutral and angry, and we consider samples with the label of ‘excited’ to be the same as ‘happy’. The labels are determined based on the majority vote. The three levels of valence and activation are defined as: low (0 - 1:67), mid (1:67-3:33), and high (3:33-5), where the value of each sample is computed based on the average of the raters. The following lists the number of samples for each type of labels,

- **Four-Emotion Classes:**

  happy: 531, sad: 576, neutral: 411, angry: 378

- **Arousal Dimension:**

  low: 331, mid: 1228, high: 337

- **Valence Dimension:**

  low 653, mid: 820, high:423

2.2 Trajectory-based Spatial-Temporal Spectral Features

Figure 1 depicts the complete flow of our trajectory-based spatial-temporal spectral features extraction approach. Given an audio file, the following is the steps of the feature extraction:

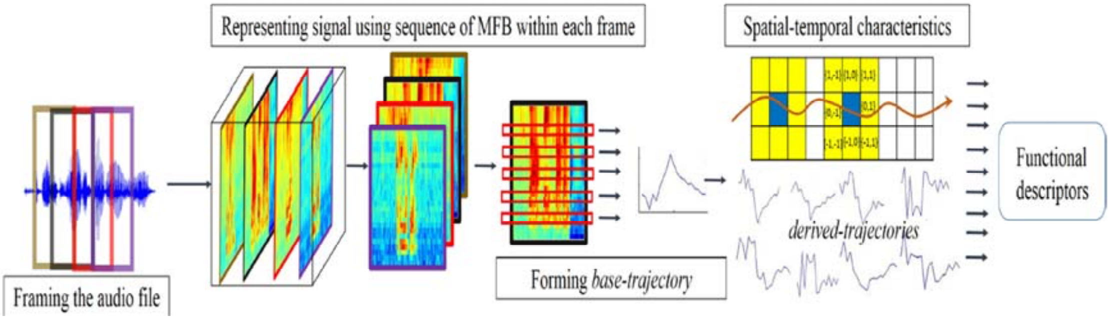

Representing signal using sequence of MFB within each frame
Spatial-temporal characteristics
Functional descriptors
derived-trajectories
Forming base-trajectory
Framing the audio file

Figure 1. It demonstrates the complete flow of trajectory-based spatial-temporal spectral feature extraction: framing the utterances, representing the signal within each frame using a sequence of MFB, forming base-trajectory of each MFB coefficient, computing grid-based spatial-temporal characteristics and derive 8 additional derived-trajectory, finally frame-level features are extracted by computing 4 statistical functionals on these 9 X 26 trajectories.

(1) Framing the signal:

Segment the entire utterance into regions of frames, where each frame is of length L (L = 250ms , 150ms).There is a 50% overlap between frames.

(2) Representing the segment:

Represent the signal within each frame using a sequence of 26 Mel-filter bank energy (MFB) output - can also be imaged as spectrogram. The window size for MFB is set to be 25ms with 50% overlap. The upper bound of frequency for MFB computation is capped at 3000 Hz.

(3) Forming base-trajectory:

The energy profile for each of the 26 filter output form a base-trajectory over the duration of each frame.

(4) Computing spatial-temporal characteristics:

For each base-$trajectory_i$, at t = 1, we compute the first-order difference with respect to its neighboring grid (8 total: marked as yellow in Figure 1); then we move along the time axis and compute these grid differences until we reach the end of frame. Hence, we obtain 8 extra trajectories (so called, derived-trajectories) to form a total of 9 trajectories (1 base-trajectory+8 derived-trajectories) per frame for each of the 26 filter outputs (a real example of trajectories can be seen in Figure 1).

(5) Frame-level spatial-temporal descriptors:

We derive the final frame-level trajectory-based spatial-temporal descriptors by applying 4 statistical functionals, i.e., maximum, minimum, mean, and standard deviation, on a total of 26 X 9 trajectories - forming the final set of 936 features per frame.

The basic idea of our newly-proposed features is to essentially track spectral energy changes within a long-durational frame in both the directions of frequency-axis (spatial) and time-axis. Since the framework is inspired from video descriptor’s extraction approach, the physical meaning related to speech production/perception can be difficult to establish. However, this framework provides a straightforward approach to quantify various inter-relationship between spectral-temporal characteristics in the speech signal directly from the time-frequency representations without much higher-level processing.

3. Experimental Setup and Results

In this work, we conduct the following two experiments on the emotion recognition tasks mentioned in section 2.1:

- **Exp I:** Comparison and analysis of our proposed *Traj-ST* with *Conv-PS* and *OpEmo-Utt* features in the three emotion recognition tasks
- **Exp II:** Analysis of recognition accuracy after fusion of *Traj-ST* with *Conv-PS* and/or *OpEmo-Utt* features in the three emotion recognition tasks

The *Conv-PS* feature extraction approach is similar to the *Traj-ST*, but instead of computing spatial-temporal characteristics on trajectories of Mel-filter bank output, we compute 45 low-level descriptors including fundamental frequency (f0), intensity (INT), MFCCs, their delta, and delta-delta every 10ms. We then derive the frame-level features by applying the 7 statistical functionals (max, min, mean, standard deviation, kurtosis, skewness, inter-quantile range) on these LLD features. This results in a total of 315 features per frame for Conv-PS. *OpEmo-Utt* is an exhaustive utterance-level feature set (i.e., emoLarge.config in the Opensmile toolbox) that has been used across many paralinguistic recognition tasks (Schuller *et al.*, 2013; Schuller *et al.*, 2014). It includes 6668 features in total per utterance. All features are znormalized with respect to an individual speaker. All evaluation is done via leave-one-speaker-out cross validation, and the accuracy is measured in unweighted average recall. Univariate feature selection based on ANOVA test is carried out for both *Traj-ST* and *Conv-PS* feature sets.

3.1 Recognition Framework

In Exp I, for *Traj-ST* and *Conv-PS* feature sets, we use Gaussian Mixture Model (M = 32) to generate a probabilistic score, $p_{i;t}$, for each class label at the frame-level, and then we perform utterance-level recognition using the following simple rule:

$$\underset{i \in \text{classes}}{\arg\max} \sum_{t=1}^{N} p_{i,t}$$

where *i* refers to the class label, *t* refers to the frame index, and N refers to the total number of frames in an utterance. For *OpEmo-Utt*, since it is a large-dimensional utterance-level feature vector, we utilize the GMM-based method after performing principal component analysis (90% of variance) and also linear-kernel support vector machine multi-class classifier.

In Exp II, the fusion methodology of *Traj-ST* with *Conv-PS* and *OpEmo-Utt* is depicted in Figure 2. The fusion framework is based on logistic regression. For *Traj-ST* and *Conv-PS*, the fusion is operated on the statistical functionals, i.e., mean, standard deviation, max, and min, applied on the $p_{i;t}$; and for *OpEmo-Utt*, the fusion is operated on the decision scores outputted from the one-vs-all multiclass support vector machine.

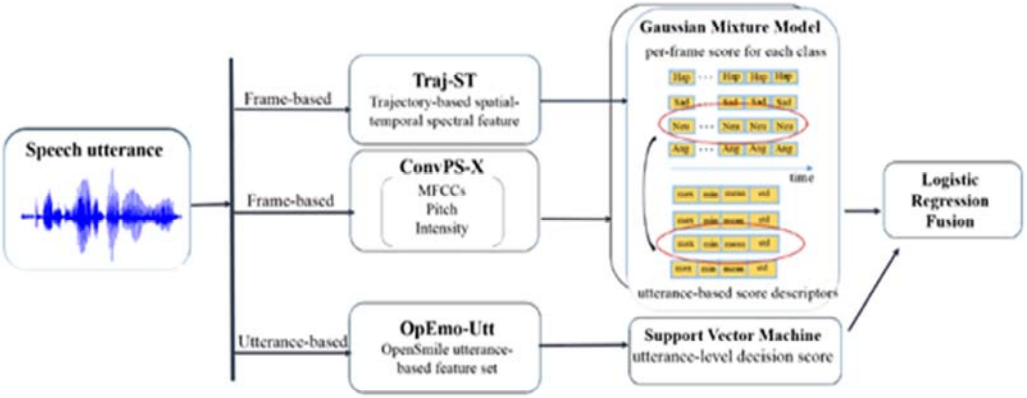

Speech utterance
Frame-based
Frame-based
Utterance-based
Traj-ST
Trajectory-based spatial-temporal spectral feature
ConvPS-X
MFCCs
Pitch
Intensity
OpEmo-Utt
OpenSmile utterance-based feature set
Gaussian Mixture Model
per-frame score for each class
Hap
Sad
Neu
Ang
time
utterance-based score descriptors
Support Vector Machine
utterance-level decision score
Logistic
Regression
Fusion

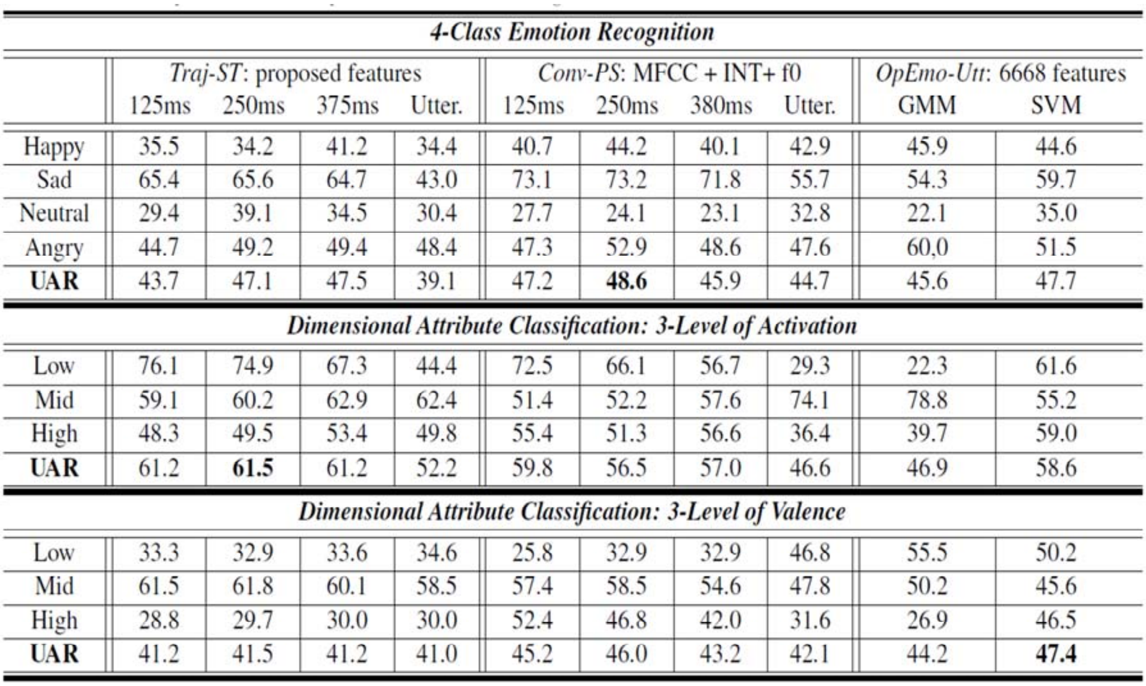
Figure 2. It depicts the fusion method for the three feature sets. Frame-based features are fused using statistical functionals of probabilistic scoring outputted from GMM model, utterance-level features are fused using decision score directly from the SVM classifier.The final fusion model utilized is logistic regression.

3.2 Exp I: Results and Discussions

Table 1 summarizes the detailed results of Exp I. For *Traj-ST* and *Conv-PS*, we report UARs of GMM model with different frame-length utilized for feature extraction, i.e., 125ms, 250ms, 375ms, and full-utterance length. For *OpEmo-Utt*, we report both UARs on using GMM and SVM models.

There are a couple points to note in the results. In the four-class emotion recognition task, *Traj-ST* compares comparably to *OpEmo-Utt* (47.5% vs. 47.7%), while the best accuracy is achieved by *Conv-PS* (48.6%). In the three-level valence recognition tasks, the best accuracy achieved is by using *OpEmo-Utt* (47.4%), where *Traj-ST* and *Conv-PS* do not perform well. Lastly, our proposed *Traj-ST* feature set performs significantly better than both *Conv-PS* and *OpEmo-Utt* on the task of three-level activation recognition. It achieves a recognition rate of 61.5%, which is an 1.7% improvement absolute over *Conv-PS* and 2.9% over *OpEmo-Utt*. By running the three types of emotion recognition tasks, it seems to be evident that each set of these features indeed possess a distinct amount and quality of emotional contents. *OpEmo-Utt* seems to perform the best for valence, possibly due to the complex nature on the perception of the degree of valence (i.e., requiring exhaust features to be extracted at the utterance-level). Although it has been demonstrated in the past that acoustic-related features often encodes more information in the activation dimension (Yildirim *et al*., 2004), it is quite still promising that see our proposed features, *Traj-ST*, are even more effective in predicting the overall perception of activation than these two other feature sets.

Table 1. It summarizes the detailed results of Exp I for three different emotion recognition tasks: 4-class emotion recognition, 3-level activation/valence recognition. For Traj-ST and Conv-PS, we report UARs of GMM model with different frame-length utilized for feature extraction. For OpEmo-Utt, we report both UARs on using GMM and SVM models.



	Traj-ST: proposed features				*Conv-PS*: MFCC + INT+ f0				*OpEmo-Utt*: 6668 features	
	125ms	250ms	375ms	Utter.	125ms	250ms	380ms	Utter.	GMM	SVM
4-Class Emotion Recognition										
Happy	35.5	34.2	41.2	34.4	40.7	44.2	40.1	42.9	45.9	44.6
Sad	65.4	65.6	64.7	43.0	73.1	73.2	71.8	55.7	54.3	59.7
Neutral	29.4	39.1	34.5	30.4	27.7	24.1	23.1	32.8	22.1	35.0
Angry	44.7	49.2	49.4	48.4	47.3	52.9	48.6	47.6	60,0	51.5
UAR	43.7	47.1	47.5	39.1	47.2	**48.6**	45.9	44.7	45.6	47.7
Dimensional Attribute Classification: 3-Level of Activation										
Low	76.1	74.9	67.3	44.4	72.5	66.1	56.7	29.3	22.3	61.6
Mid	59.1	60.2	62.9	62.4	51.4	52.2	57.6	74.1	78.8	55.2
High	48.3	49.5	53.4	49.8	55.4	51.3	56.6	36.4	39.7	59.0
UAR	61.2	**61.5**	61.2	52.2	59.8	56.5	57.0	46.6	46.9	58.6
Dimensional Attribute Classification: 3-Level of Valence										
Low	33.3	32.9	33.6	34.6	25.8	32.9	32.9	46.8	55.5	50.2
Mid	61.5	61.8	60.1	58.5	57.4	58.5	54.6	47.8	50.2	45.6
High	28.8	29.7	30.0	30.0	52.4	46.8	42.0	31.6	26.9	46.5
UAR	41.2	41.5	41.2	41.0	45.2	46.0	43.2	42.1	44.2	**47.4**

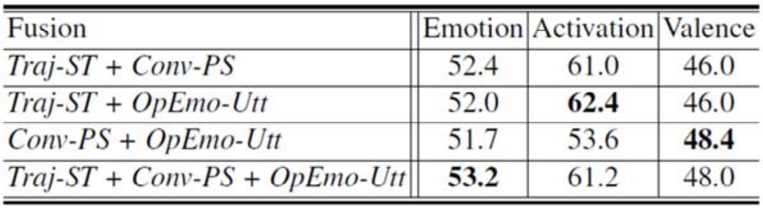
The frame duration also plays an important role in achieving the optimal accuracy for *Traj-ST* (also for *Conv-PS*). Our empirical finding seems to implicate that a duration of roughly 250ms is the optimal frame-duration - a result that corroborates findings in the previous use of long-term spectral features for emotion recognition (Chaspari *et al*., 2014; Chi *et al*., 2012). Furthermore, the feature selection output from *Traj-ST* shows that the top three directions of spatial-temporal characteristics are the {0,0} - base-trajectory, {1,0} - higher-spatial-equivalent-temporal directional trajectory, and {1,-1} - higher-spatial-earlier-temporal directional trajectory. These three constitutes 50% of the selected features. It is interesting to see that modeling not just the temporal changes but also the spatial (i.e., in the direction of frequency) can be beneficial; in specific, additional investigation will also need to be carried out to understand the reason to the finding that there seems to be a higher emotional discriminability in these specified trajectories, which quantify the spectral energy changes in direction toward higher-frequency bands.

In summary, we show that our novel feature set compares comparably to the state-of-art usage of exhaustive feature extractions in discrete 4-class categorical emotion recognition and outperforms significantly in the 3-level activation recognition.

3.3 Exp II: Results and Discussions

Given that in Exp I, each set of features seem to be capable of recognizing different representation of emotions. A natural experiment is to fuse the three different set of features. Table 2 lists the various fusion results. *OpEmo-Utt* refers to fusing the outputted decision scores from the SVM model.

Table 2. Exp II summary results on fusion of three different feature sets: Traj-ST, Conv-PS, OpEmo-Utt. The number presented is computed using UAR



Fusion	Emotion	Activation	Valence
Traj-ST + Conv-PS	52.4	61.0	46.0
Traj-ST + OpEmo-Utt	52.0	**62.4**	46.0
Conv-PS + OpEmo-Utt	51.7	53.6	**48.4**
Traj-ST + Conv-PS + OpEmo-Utt	**53.2**	61.2	48.0

There are a couple observations to be made with the results. The first is that fusion of different feature sets all improves the best single-feature set's result. In specifics, the best fusion accuracy of 4-class emotion recognition is 53.2% (4.6% absolute improvement over the best single-feature set) obtained by fusing all three sets of features; the best fusion result for 3-level valence is 48.4% (1.0% absolute improvement over the best single-feature set, *OpEmo-Utt*); lastly, the best fusion result for 3-level activation is 62.4% (0.9% absolute improvement over the best single-feature set, *Traj-ST*). We see that our newly propose features, *Traj-ST*, are indeed capable of additionally improve the recognition rate for categorical emotion recognition and activation level detection under this fusion framework - signifying the complementary information of our features possess with regard to emotional content that is originally lacking in these two well-established state-of-arts feature sets.

In summary, we have demonstrated that our novel trajectory-based spatial-temporal spectral features can be utilized in combination with the two popular and well-established acoustic feature sets in order to obtain improved emotion recognition rate.

4. Conclusions

In this work, we propose a novel set of low-level acoustic features derived directly from the spectrograms in order to characterize the long-term spatial-temporal information of the speech signal. We carry out emotion recognition experiments on both categorical emotion attributes and dimensional representations using the proposed features. Our experiments show that the newly-proposed feature set compares comparably to the well-established low-level acoustic descriptors and state-of-the-art exhaustive feature extraction approach on the categorical emotion recognition, and it outperforms on the task of activation level recognition. Furthermore, by fusing these trajectory-based spatial-temporal features, it improves the overall emotion recognition accuracy. Overall, it is quite promising to see these low-level features do possess emotional discriminatory power beyond the exhaustive set of established acoustic parameters.

There are several future directions. One of the immediate future direction is that we observe these features do not possess enough modeling power of the valence dimension; one of the possible causes may due to the fact that the grid-based differential operator may only capture the spatial-temporal interrelationship locally, and the statistical functionals may not be enough to quantify the suprasegmental information. We will immediately extend this grid-based differential operator to incorporate a wider range both in time and in space with different scales to capture the valence-related acoustic properties. Secondly, one of our main goals is to minimize the effort of raw signal processing required as we derive these features. We will replace the MFB portion of spectral representation to even lower-level (or employ sparse representation to ensure robustness) time-frequency representation while maintaining low computational complexity. Lastly, on the longer term, once these (raw) low-level features are developed, they can be suitable inputs to deep learning algorithms to learn various hierarchical representations of speech acoustic that are relevant for emotion perception. The ability to robustly recognize emotion will continue to be at the fore-front of developing human-centric applications

Acknowledgment

Thanks to Ministry of Science and Technology (103-2218-E-007-012-MY3) for funding.

References

Bach-y Rita, P. & Kercel, S. W. (2003). Sensory substitution and the human-machine interface. *Trends in cognitive sciences*, *7*(12), 541-546. doi:10.1016/j.tics.2003.10.013

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S.,...Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335-359. doi: 10.1007/s10579-008-9076-6

Calvo, R. A., D'Mello, S., Gratch, J. & Kappas, A. (2014). *The Oxford handbook of affective computing*. Oxford, England: Oxford University Press.

Campbell, W. M., Sturim, D. E. & Reynolds, D. A. (2006). Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308-311. doi: 10.1109/LSP.2006.870086

Chaspari, T., Dimitriadis, D. & Maragos, P. (2014). Emotion classification of speech using modulation features. In *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, 1552-1556.

Chi, T.-S., Yeh, L.-Y. & Hsu, C.-C. (2012). Robust emotion recognition by spectro-tempora modulation statistic features. *Journal of Ambient Intelligence and Humanized Computing*, *3*(1), 47-60. doi: 10.1007/s12652-011-0088-5

Childers, D. G. & Wu, K. (1991). Gender recognition from speech. Part ii: Fine analysis. *The Journal of the Acoustical society of America*, *90*(4), 1841-1856. doi: 10.1121/1.401664

Dobry, G., Hecht, R. M., Avigal, M. & Zigel, Y. (2011). Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 1975-1985. doi: 10.1109/TASL.2011.2104955

Eyben, F., Wöllmer, M. & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, 1459-1462. doi: 10.1145/1873951.1874246

Hogan, N., Krebs, H. I., Sharon, A. & Charnnarong, J. (1995). *U.S. Patent No. 5,466,213A*. Cambridge, MA: Massachusetts Institute Of Technology.

Hollinger, G. A., Georgiev, Y., Manfredi, A., Maxwell, B. A., Pezzementi, Z. A. & Mitchell, B. (2006). Design of a social mobile robot using emotion-based decision mechanisms. In *Proceedings of f the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3093-3098. doi: 10.1109/IROS.2006.282327

Kim, Y., Lee, H. & Provost, E. M. (2013). Deep learning for robust feature generation in audiovisual emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3687-3691. doi: 10.1109/ICASSP.2013.6638346

Lee, C. M. & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293-303. doi: 10.1109/TSA.2004.838534

Lee, C.-C., Mower, E., Busso, C., Lee, S. & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*(9), 1162-1171. doi: 10.1016/j.specom.2011.06.004

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I.,...Sahli, H. (2013). Hybrid deep neural network-hidden markov model (dnn-hmm) based speech emotion recognition. In *Proceedings of 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 312-317. doi: 10.1109/ACII.2013.58

Mairesse, F. & Walker, M. (2006). Words mark the nerds: Computational models of personality recognition through language. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 543-548.

Mower, E., Matarić, M. J. & Narayanan, S. (2011). A framework for automatic human emotion classification using emotion profiles. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(5), 1057-1070. doi: 10.1109/TASL.2010.2076804

Narayanan, S. & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. In *Proceedings of IEEE Inst Electr Electron Eng.*, *101*(5), 1203-1233. doi: 10.1109/JPROC.2012.2236291

Nwe, T. L., Foo, S. W. & De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, *41*(4), 603-623. doi: 10.1016/S0167-6393(03)00099-2

Oneata, D., Verbeek, J. & Schmid, C. (2013). Action and event recognition with fisher vectors on a compact feature set. In *Proceedings of ICCV '13 Proceedings of the 2013 IEEE International Conference on Computer Vision*, 1817-1824. doi: 10.1109/ICCV.2013.228

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT press.

Scherer, K. R. (2003). Vocal communication of emotion: A review ofresearch paradigms. *Speech communication*, *40*(1-2), 227-256. doi: 10.1016/S0167-6393(02)00084-5

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J.,...Aharonson, V. (2007). The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *Proceedings of INTERSPEECH 2007*, 2253-2256.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., MüLler, C. & Narayanan, S. (2013). Paralinguistics in speech and languagestate-of-the-art and the challenge. *Computer Speech & Language*, *27*(1), 4-39. doi: 10.1016/j.csl.2012.02.005

Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F.,...Zhang, Y. (2014). The interspeech 2014 computational paralinguistics challenge: cognitive & physical load. In *Proceedings of INTERSPEECH 2014*, 427-431.

Swartout, R. W., Gratch, J., Hill Jr, R. W., Hovy, E., Marsella, S., Rickel, J. & Traum, D. (2006). Toward virtual humans. *AI Magazine*, *27*(2), 96-108. doi: 10.1609/aimag.v27i2.1883

Vinciarelli, A., Pantic, M. & Bourlard, H. (2009). Social signal processing :Survey of an emerging domain. *Image and Vision Computing*, *27*(12), 1743-1759. doi: 10.1016/j.imavis.2008.11.007

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. (2011). Action recognition by dense trajectories. In *Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3169-3176. doi: 10.1109/CVPR.2011.5995407

Wang, H., Kläser, A., Schmid, C. & Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, *103*(1), 60-79. doi: 10.1007/s11263-012-0594-8

Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Busso, C., Deng, Z.,...Narayanan, S. S. (2004). An acoustic study of emotions expressed in speech. In *Proceedings of INTERSPEECH 2004*, 2193-2196.

The individuals listed below are reviewers of this journal during the year of 2017. 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
Yung-Chun Chang
Chien Chin Chen
Alvin C.-H. Chen
Tai-Shih Chi
Chih-Yi Chiu
Hsiu-Min Chuang
Hong-Jie Dai
Zhao-Ming Gao
Hung-Yan Gu
Jia-Fei Hong
Shu-Kai Hsieh
Jen-Wei Huang
Jeih-Weih Hung
Chih-Chung Kuo
Chi-Chun Lee

Hong-Yi Lee
Bor-Shen Lin
Shu-Yen 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

2017 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 22

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2017

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

SUBJECT INDEX

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

會員類別	□終身 □個人 □學生	會員編號	（由本會填寫）		
姓　　名		性別		出生日期	年　月　日
				身分證號碼	
現　　職		學　　歷			
通訊地址	□□□				
戶籍地址	□□□				
電　　話		E-Mail			
			申請人:　　　　　　（簽章） 中　華　民　國　　年　　月　　日		

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

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

nternational Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites ubmission of original research papers in the area of computational linguistics and speech/text processing of atural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously npublished and cannot be under consideration elsewhere. Submissions should report significant new research esults in computational linguistics, speech and language processing or new system implementation involving ignificant theoretical and/or technological innovation. The submitted papers are divided into the categories of egular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full etails. Short papers can focus on a smaller research issue. And survey papers should cover emerging research ends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no rict length limitation on the regular and survey papers. But it is suggested that the manuscript should not xceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced 4 pages. All contributions will be anonymously reviewed by at least two reviewers.

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

yle for Manuscripts: The paper should conform to the following instructions.

Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size 11 points or larger.

Title and Author: The first page of the manuscript should consist of the title, the authors' names and stitutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such er words as the orthography of the language requires beginning with a capital letter. The author's name ould appear below the title.

Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords equired. The abstract should not only indicate the scope of the paper but should also summarize the author's nclusions.

Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-nd margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).

Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with perscript numbers. Footnotes may appear at the end of manuscript

Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or itten clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in entheses.

References: All the citations and references should follow the APA format. The basic form for a reference ks like

Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

e shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

e basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. ruton, 1996).

ase visit the following websites for details.

APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)

APA Stytle (http://www.apastyle.org/)

page charges are levied on authors or their institutions.

al Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply l manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

line Submission: http://www.aclclp.org.tw/journal/submit.php

ase visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.22 no.2 December 2017

Contents

Special Issue Articles: Selected Papers from ROCLING XXIX

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing