International Journal of

Computational Linguistics & Chinese Language Processing

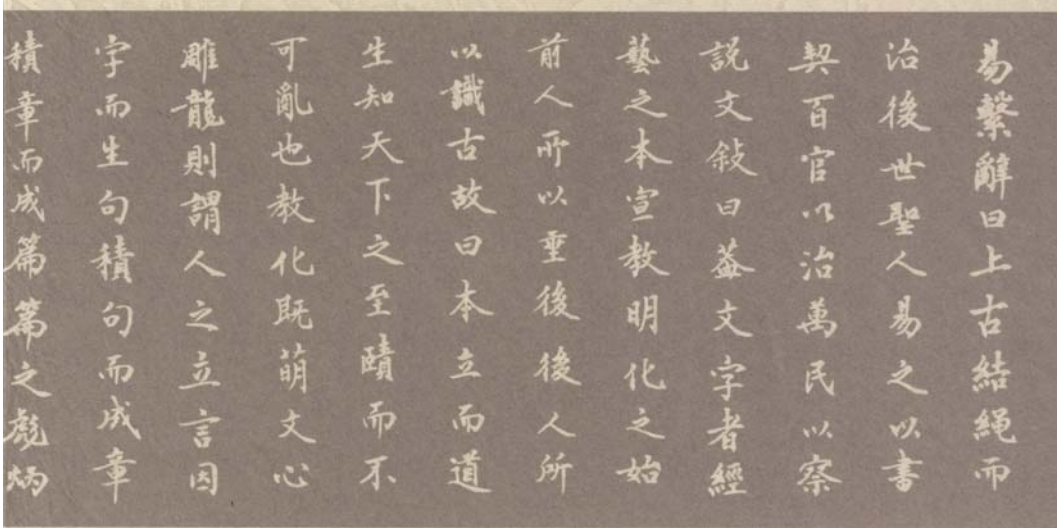
中文計算語言學期刊

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 XXV"

Guest Editor: Chia-Hui Chang, Chia-Ping Chen, and Jia-Ching Wang



Vol.18 No.4 December 2013 ISSN: 1027-376X

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. 18, no. 4, December 2013

Contents

Special Issue Articles:
Selected Papers from ROCLING XXV

Papers

Forewords

The 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013) was held at National Sun Yat-sen University, Kaohsiung on Oct. 4-5, 2013. ROCLING is the leading and most comprehensive conference on computational linguistics and speech processing in Taiwan, bringing together researchers, scientists and industry participants to present their work and discuss recent trends in the field. This special issue presents extended and reviewed versions of eight papers meticulously selected from ROCLING 2013, including 4 natural language processing papers and 4 speech processing papers.

The first paper done at Chaoyang University of Technology focuses entailment analysis for improving Chinese textual entailment recognition. By considering four special cases, the RTE system can be significantly improved. The second and third papers are applications of statistical machine translation in Chinese spelling check and English grammatical error correction. Both papers are research work from National Tsing Hua University. The fourth paper, from Academia Sinica, studies Chinese noun-noun compound and concludes that two nouns are either linked by semantic roles assigned by events or by static relations including) including meronymy, conjunction, and the host-attribute-value relation. The fifth paper is from National Cheng Kung University. This research work employs Hidden Markov Model-based synthesis approach to generate Mandarin songs with arbitrary lyrics and melody in a certain pitch range. The sixth paper is a joint work from National Taiwan University, National Tsing Hua University, and Chang Gung University. This research uses speech recognition and assessment to automatically find the potentially problematic utterances for preparing a Taiwanese speech corpus. The seventh paper is done by National Taiwan University of Science and Technology. For LMR-Mapping based voice conversion, this work places a histogram-equalization module and a target frame selection module immediately before and after the LMR based mapping. The eigth paper, from National Chi Nan University, presents a noise-robust speech feature representation method in speech recognition. This method applies linear predictive coding on the time series of cepstral coefficients and then removes the linear prediction error component.

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
Chia-Hui Chang,
Department of Computer Science and Information Engineering, National Central University, Taiwan
Chia-Ping Chen
Department of Computer Science and Engineering, National Sun Yat-Sen University, Taiwan
Jia-Ching Wang
Department of Computer Science and Information Engineering, National Central University, Taiwan

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 1-16

蘊涵句型分析於改進中文文字蘊涵識別系統

Entailment Analysis for Improving Chinese Recognizing Textual Entailment System

楊善順*、吳世弘*、陳良圃+、邱宏昇+、楊仁達+

Shan-Shun Yang, Shih-Hung Wu, Liang-Pu Chen,

Hung-Sheng Chiu, and Ren-Dar Yang

摘要

文字蘊涵是自然語言處理最近興起的研究課題。文字蘊涵識別(Recognizing Textual Entailment, RTE)可以應用到其他許多自然語言處理的研究中。在本文中將介紹我們在觀察 NTCIR-10-RITE-2 資料集後發現過去系統的缺陷，進而提出如何改進中文文字蘊涵系統的方法。過去的系統處理文字蘊涵多使用機器學習分類文題的方法，所有輸入句子都用同樣的分類器處理，對於某些特別的問題往往會產生誤判。我們認為應該針對於特定類型的問題做處理，增加系統可以處理的問題類型。實驗結果顯示配合之前提出的機器學習方法，增加四種特殊類型分類對特殊類型句子進行個別處理，可以有效改進系統，實驗結果系統在識別簡體中文蘊涵兩類的正確率從原本 67.86%提昇到 72.92%。

關鍵詞：中文文字蘊涵識別、蘊涵分析

*朝陽科技大學資訊工程系 Department of Computer Science and Information Engineering, Chaoyang University of Technology, Taichung, Taiwan (R.O.C)
E-mail: { s10027619; shwu}@cyut.edu.tw
The author for correspondence is Shih-Hung Wu.
+財團法人資訊工業策進會 Institute for Information Industry, Taipei, Taiwan (R.O.C)
E-mail: {eit; bbchiu; rdyang}@iii.org.tw

Abstract

Recognizing Textual Entailment (RTE) is a new research issue in natural language processing (NLP) research area. RTE can be a useful component in many NLP applications. In this paper, we introduce our finding on the entailment analysis of the NTCIR-10 RITE-2 dataset, and use the observation to improve our system. In the previous works, all the input pairs are treated equally in a standard classification architecture. We find that is not suitable for some special cases. We believe that by isolating the special cases and building separated classifiers, a RTE system can perform better. After implementing modules for four special cases into our system, the result is significantly improved from 67.86% to 72.92% on the binary class classification task.

Keywords: Chinese Recognizing Textual Entailment, Entailment Analysis

1. 緒論

文字蘊涵(Textual Entailment, TE)(Dagan *et al*., 2006)是自然語言處理(Natural Language Processing, NLP)最近興起研究議題，文字蘊涵識別目標為給定一個句子對(T1,T2)系統能夠準確的推斷這兩句子之間的蘊涵關係。因此文字蘊涵可以應用在自然語言處理其他領域研究中，例如問答系統、資訊抽取、資訊檢索、機器翻譯(Dagan & Glickman, 2004 ; Ou & Yao, 2010)等等。文字蘊涵最基本的方法就藉由句子字面上的資訊例如語意、句法(Hua & Dinga, 2011)等等字面上的相似性進而推斷句子是否有著蘊涵關係。因此利用這個特性，文字蘊涵有助於問答系統找到資料庫中與輸入問句最相近的問句進而回應最適當回答。以資訊檢索來說檢索詞的好壞對資訊檢索有著很大影響，藉由文字蘊涵找到與檢索詞相關的字詞(例如同義詞)加入檢索條件這樣可以讓使用者更容易找到使用者所需要的資訊。

目前文字蘊涵的研究分成兩種層面，首先兩類(Binary Class, BC)的任務的目標是單純判別 T1 與 T2 之間是否有蘊涵關係，但句子之間蘊涵關係並不能單純以有或沒有這麼簡單就區分開，因此為了表示不同情況下的句子之間蘊涵關係，NTCIR RITE 另外定義多類(Multi Class, MC)這項任務將句子之間的蘊涵作更為明確的分類。假設這個句子對具有蘊涵關係，我們可以很合理認為這兩個句子所表達是相同的意思，但有可能兩個句子如表 1 中正向蘊涵的例句一樣兩個句子所包涵的資訊數量不同，造成我們可以從 T1 句子可以推論出 T2 句子的完整的意思，但是不能從 T2 推論出 T1 句子完整的意思，這樣情況我們就稱正向蘊涵。反之如表 1 中雙向蘊涵的例句一樣 T1 句子可以推論出 T2 句子的含意，T2 也可以推論出 T1 句子完整的意思，兩個句子之間可以相互推論這樣的情況我們就稱為雙向蘊涵關係。假設句子對之間沒有蘊涵關係，我們可以很合理認為兩個句子所表達的意思不相同，但這並不完全正確的想法。如同表 1 矛盾例句一樣可能兩個句子所包涵的資訊大致相同只是少部份資訊例如:是與不是或是時間點不同造成句子的意

思互相衝突，這樣的情況我們就稱之為矛盾蘊涵。或是兩個句子本身包涵的資訊毫無關係這樣的情況我們就稱之為獨立蘊涵，藉由上述的四種蘊涵關係將句子之間的蘊涵關係細分，使得文字蘊涵系識別的研究更有其意義。

本篇重點在處理簡體中文與繁體中文方面的文字蘊涵，使用 NTCIR-10 RITE-2 所提供的訓練資料基於機器學習的方法 SVM 建立一個中文文字蘊涵系統。系統的一開始先將輸入的文句對資料進行預處理，由於處理是中文資料必須先進行斷詞以便接下來的工作，使用問題類型分類將可以個別處理的類型抽出特別處理，接著句子對經由特徵擷取的子系統取得各項特徵值，最後將取的特徵值使用 SVM 分類處理。

本篇接下來章節如下，在第二段介紹過去研究方法，第三段將介紹系統架構與預處理部分以及系統使用到的特徵值跟我們所觀察到特殊類型問題分析，第四段是實驗結果與討論最後是結論與未來工作。

表 1. 各種蘊涵關係例句

類型	例句
正向蘊涵 (F)	t1：異位性皮膚炎好發於具過敏體質的嬰幼兒、兒童及青少年，常見症狀是臉、頸、手肘窩、膝窩、或四肢背側等部位出現搔癢紅疹、皮膚變厚、變粗糙。
	t2：異位性皮膚炎產生的發紅皮疹好發於臉頰頸側、手肘窩或膝蓋等彎曲部位。
雙向蘊涵 (B)	t1：洋基球團保護選手的立意甚篤，要求「只投一場且不超過一百球」。
	t2：洋基球隊開出「只投一場且不超過一百球」的保護條件。
矛盾蘊涵 (C)	t1：印尼蘇門答臘西岸外海發生芮氏規模九的強震，主震與餘震所引發的海嘯席捲南亞諸國的海岸。
	t2：印尼蘇門答臘北部外海廿八日深夜發生芮氏規模八點七的強震。
獨立蘊涵 (I)	t1：中研院基因體中心正和何大一合作，研發新一代的禽流感基因疫苗。
	t2：中研院院士何大一最近正在研發禽流感基因疫苗。

2. 過去研究

之前的研究文獻中有許多不同的方法應用在英文文字蘊涵識別，例如定理證明或使用 WordNet 等等不同的詞意語料資源。在中文文字蘊涵方面(Wu *et al.*, 2011)等人參考其他語言的方法提出一個基礎機器學習利用機器翻譯效能評估的 BLEU 分數及句子長度做為特徵以及句子長度做為特徵來訓練分類器，建立基礎中文文字蘊涵識別系統，(Zhang *et al.*, 2011)等人提出加入語意相關資訊作為特徵處理，藉由上下位詞、同義詞與反義詞等資訊來進行的推論以及使用多個機器學習的方法，最後使用投票機制選出最合適蘊涵關係提高系統準確率。

隨後(Yang *et al.*, 2012)等人提出使用單語言機器翻譯的方法，藉由 GIZA++中字詞對齊的匹配值進行計算出句子之間相似度作為新特徵使用以有效提升正確率，但這個方法在處理兩類中效果較好再處理多類蘊涵關係時效果不如兩類，之後(Wang *et al.*, 2013)等人提出反向的矛盾字詞對齊，有效改善之前正向字詞對齊容易產生的誤判問題。

在文字蘊涵識別的推論時，各種語義資訊與上下文資訊是必要的處理。雖然 NTCIR-10 RITE-2 (Watanabe *et al.*, 2013) 任務目的在於評鑑各種語義/上下文處理系統，但這有一個問題注重具體語言現象的研究是不容易的達成。在 RITE-2 日文子任務的資料集中含包涵了識別 T1 和 T2 之前蘊涵關係必需的語言學現象，從兩類(BC)子任務資料集中擷取句子對作為樣本，建立具有語言學現象的句子，將這樣的句子加入資料集中作為單元測試使用。單元測試數據相當於多類(BC)任務資料集的一個子集。雖然這個資料集並不多，但您可以將它用於各種研究，包括分析 RITE 資料集中出現的語言學問題，評測每一個語言學現象識別正確度和為各種語言學現象的分類器訓練與測試資料。

3. 系統架構

我們的系統的系統流程圖如圖 1 所示，基本組成部分”預處理”、”斷詞”、”中文簡繁轉換”、”特殊類型分類”、個別” SVM 分類器”和最後”結果整合”。

3.1 預處理

3.1.1 表示格式正規化

在預處理的部份第一步就是句子的字詞格式正規化，我們系統預處理正規化模塊將括號中字詞視為括號前字詞的同義詞，例如”葉望輝(Stephen J. Yates)”，括號中的字詞”Stephen J. Yates”是另一個字詞”葉望輝”是相同意思。另外我們觀察訓練資料發現有部分句子之間對於時間表示格式不同，因此時間表示方法必須統一以便接下來的處理工作，如表 2 中所示的時間表示格式常見的例子，測試資料中數字的格式也相當不一致造成系統判讀上的困難，這個部份也是我們需要統一的部份。將句子正規化後的資料有著更高的相似度也更容易被識別，在機器翻譯方面也句子中的字詞更容易文字對齊進而提高兩個句子可翻譯機率。

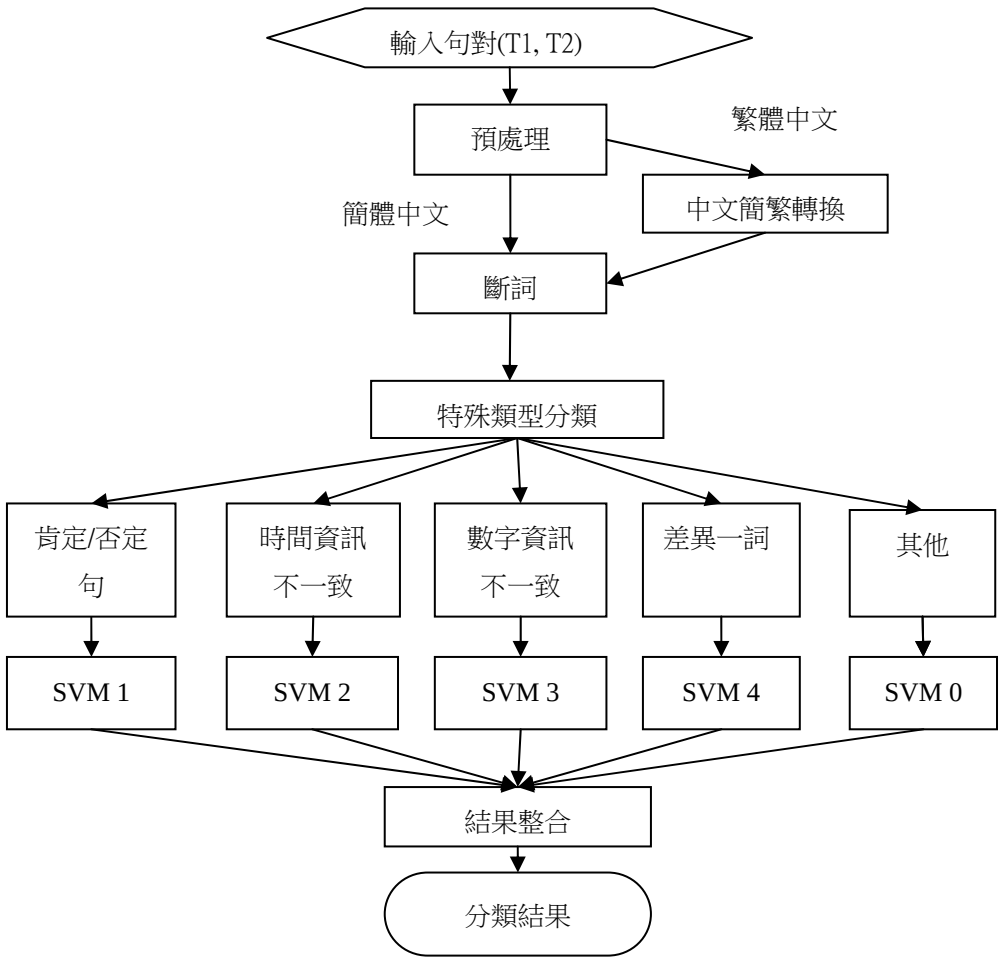


圖 1. 系統流程圖

表 2. 各種時間表示格式之例句

時間型態	時間表示方式
中文	一九九七年二月廿三日
數字全形	１９９７年２月２３日
數字半形	1997年2月23日
數字以「-」隔開	1999-05-07
數字以「/」隔開	1985/12/20
範圍	1999年延長至2001年

3.1.2 背景知識的替換

預處理的部份第二步是代替它們的同義詞，同義詞的資訊可以從”維基百科”、”HowNet (Liu & Li, 2002)中取得，其中從維基百科中可以取得句子必要的背景知識，包括明確的時間和地點的資訊作為同義詞替換補充所需的資訊。另外有關時間表示格式還有另一個難題就是歷史上不同朝代如何轉換成同一個時間表示，這是需要相當的背景知識才能正確的轉換。例如”乾隆”是西元 1735 年和”昭和”是西元 1925 年或是例如”北京奧運”擁有背景知識的人可以很輕易就知道”北京奧運”發生於西元 2008 年。所以時間表示式問題需要相當背景知識才能進行正規化，因此我們從維基百科取得各朝代資訊建立一個朝代資料庫，作預處理時偵測句子中是否含有朝代字詞，若有朝代字詞將其替換成對應年份後加上朝代字詞後數字減一，如”乾隆 3 年”將替換成”西元 1737 年”。另一個類似的問題是地名的表示問題有時候可能因作者的不同對於相同的地方有不同的表示方式，所以遇到這樣的情況我們必須進行地名正規化的步驟，例如縮寫”台、澎、金、馬”是指”台灣、澎湖、金門、馬祖”或是代稱，觀察過去的語料資料建立地名資料庫，偵測句子中是否含有地名字詞，若有地名字詞將其替換成統一字詞。

3.1.3 斷詞與中文簡繁轉換

由於我們使用的斯坦福剖析器只能處理簡體中文以及英文，因此我們的系統中使用的斷詞工具是 ICTCLAS 斷詞系統，這是由中國科學院計算技術研究所提供。該工具功能包括斷詞、詞性標註、NE 識別、新字詞識別，以及自訂字典。為了配合斷詞系統我們處理繁體中文時必須將繁體中文文句對轉換成簡體中文文句，一開始我們使用 Google 翻譯之後改使用我們自行開發的機器翻譯系統(Li *et al*., 2010)。

3.2 特徵提取

在我們的系統中使用到的特徵在表 3 列出以前的文字蘊涵識別工作(Yang *et al*., 2012)大多數可用的特徵。在前三個特徵測量 T1 和 T2 中根據一般中文類似的字元。Unigram recall、unigram precision、unigram F-measure 可以視為在 T1，T2 的字元比例和幾何平均，我們的系統使用 BLEU (Papineni *et al*., 2002)三個特點。Bleu (Zhou *et al*., 2006)當初是被設計來測量機器翻譯(machine translation)的品質。一個良好的機器翻譯需要包含適當、準確以及流暢的翻譯，我們的系統會將其翻譯為原來的文字 T1 和 T2 得到 log Bleu recall、log Bleu precision 和 log Bleu F measure values。

第七到第十這四個特徵是 T1 和 T2 的句子長度。我們的系統根據字元和字詞計算 T1 和 T2 的句子中長度的差異，並使用了這兩個特徵的絕對值在我們的系統中。

最後特徵是由 GIZA++(Och & Ney, 2003)字詞對齊分數，這是 T1 句子以單一語言機器翻譯到 T2 句子的機率。機器翻譯可該功能有助於 RTE (Quang *et al*., 2012)，我們的系統採用的單一語言機器翻譯作為一個特徵。在我們的系統中，我們使用 GIZA ++做為單一語言機器翻譯工具，GIZA++根據 IBM 模型製作而成，我們經由 GIZA++ 計算出兩個

輸入句子之間的匹配得分，在使用下列公式計算最後應用在 SVM 上的特徵值：

$$p_n = \frac{\log\left\{\prod_{i,j=0}^{i,j=\max}\left[p(t1_i|t2_j)\right]\right\}}{n} \quad (1)$$

公式中 $t1_i$ 與 $t2_j$ 分別代表 T1 與 T2 句子中各字詞， $p(t1_i \mid t2_j)$ 代表 $t1_i$ 與 $t2_j$ 字詞對齊的機率，使用 GIZA++計算句子中字詞對齊機率後連乘取 log 在除以連乘的次數 n 後就是使用在 SVM 的特徵值 p_n 。

表 3. 分類器使用的特徵

編號	特徵
1	Unigram recall
2	Unigram precision
3	Unigram F-measure
4	Log bleu recall
5	Log bleu precision
6	Log bleu F-measure
7	difference in sentence length (character)
8	absolute difference in sentence length (character)
9	difference in sentence length (term)
10	absolute difference in sentence length (term)
11	GIZA++

3.3 特殊類型處理

我們觀察 NTCIR10 RITE-2 所提供的訓練資料，發現有許多的句子是過去的系統所無法正確處理容易產生誤判。如表 4 所示我們檢視 NTCIR10 RITE-2 我的們系統正式評測結果整理過去系統容易誤判問題類型。

在表 4 中 Caes1 是一個獨立關係被系統誤判成雙向蘊涵的例子，我們發現兩個句子字面上非常相似，只是因為粗體部份不同使兩個句子變成獨立關係，假如對句子中單詞進行比對即可解決這個誤判可能。Caes2 是一個矛盾關係誤判成雙向蘊涵的例子，從字面上可以發現兩個句子非常相似，只是因為句子中數字資訊不相同產生互相矛盾關係的情況，在此可以知道我們系統對於數字資訊比對不足這是未來改進的方向，在觀察系統誤判題目時發現有許多如 Caes3 這種句子中包含字詞完全相同就誤判成正向的情形的存在，Caes4 的例子由於我們的系統不具有相關背景知識可以推論句子中的”倫敦”就指是”英國”的意思，因此將正向誤判成獨立關係，除了上面提到的同義詞問題，Caes5 中可以發現反義詞也是未來需要解決的問題之一。

表 4. 容易誤判問題類型例子

編號	例句
Case1	T1: **流感**病毒可在人体外存活**三到六**小时
	T2: **冠状**病毒通常可在人体外存活**二到三**小时
Case2	T1: 大陆已有**四百万**人感染爱滋病
	T2: 大陆有**八十五万**人感染艾滋病
Case3	T1: 美国奉行一中政策和遵守三公报的立场并未改变；切尼则进一步表示不支持台湾独立
	T2: 美国不支持台湾走向独立
Case4	T1: 申奥成功的伦敦当局，在爆炸案后立即宣布取消庆祝活动。
	T2: 英国已停止所有庆祝申奥成功的活动。
Case5	T1: 我国生物技术可以与美国等先进国家相提并论，**解决**“异种核转殖”的问题并不难。
	T2: 我国生物技术可能**造成**异种核转殖等问题。

4. 蘊涵句型分析

為了解決這些容易誤判的問題，我們認為應將句子依照問題類型進行分類再各自使用最適合的方法進行處理。在系統處理完預處理後，將可以特殊類型的句子挑選出來，使用我們開發的子系統做處理，處理後的結果在與過去使用的機器學習方法作整合，得到最後的結果，以下是我們以實做出來的特殊類型，表 5 是對應的例句。

4.1 肯定/否定句

我們觀察訓練集資料發現有些句子對，句子包涵的資訊幾乎完全相同，只因為否定詞就使得兩個句子具有是否關係形成矛盾或獨立蘊涵關係，針對這種類型的句子對我們可以使用自行開發的系統進行句子對是否具有是否關係字詞，假設句子對只有一句具有是否關係字詞，這樣可以認為句子對沒有蘊涵關係。

4.2 時間資訊不一致

訓練資料之中發現有些句子中含有時間的資訊，然而這些時間的資訊對句子的含意有很重要的意義，所以我們認為應該將這些句子挑選出來針對句子中的時間部份比較，針對這種類型的句子在之前系統前處理部分會對句子中的時間格式做正規化，再對句子的時間進行比對可以處理特定時間點的句子。

4.3 數字資訊不一致

訓練資料中有些句子中含有數字的資訊,然而這些數字資訊對句子的含意有很大的影響,因此我們認為應該將這些句子挑選出來針對句子中的數字部份比較,對這種類型的句子在之前系統前處理部分會對句子中的數字格式做正規化,針對數字部分進行比對就可以處理這類大部分的句子。

4.4 差異一詞

我們觀察訓練資料時發現有些句子字面上相當相似,例如主詞或受詞或是部份字詞被替換就會導致句子的意思都改變了,針對這類型的句子可以對句子字詞進行比對處理。

當然特殊類型的問題不止上述的幾種,我們也歸納出更多特殊類型有待未來完成。

4.5 同義詞

我們觀察訓練資料時發現有些句子對部份的字詞有著同義詞的關係,對於這類的句子對應該對句子先進行處理,使用 E-hownet 等語料庫將同義詞都替換成相同字詞就可以使兩個句子變得更相似,這樣判斷蘊涵關係可以更容易。

4.6 背景知識

我們觀察訓練資料時發現有些句子提供的資訊即使是人類,在沒有一定的背景知識情況下也無法正確的判斷是否有蘊涵關係,例如新德里是印度的首都沒有這樣的背景知識很容易將新德里與印度認為成不同的地方,對於這類型的句子可以使用外部資源例如維基百科等獲得對應的背景知識作處理。

4.7 句法調換

我們觀察訓練資料時發現有些句子,只是將句子的順序進行調換就導致句子的含意有所改變,當句子中的主詞與受詞有所改變會導致整個句子意思整個改變,針對這類型的句子可以使用如史丹佛剖析器將句子進行剖析得到語意角色關係進而偵測句子主詞與受詞關係以及連接詞也進行偵測是否為前後調換語意不變的詞如與、一起等等。

4.8 一詞多義與命名實體

我們觀察訓練資料時發現有些句子雖然句子的字詞很相似,但其實含意大大不相同例如”沒完沒了”本身是一個成語但是這個詞在其他地方有著其他意思如歌名、電影名,正因如此這樣字詞造成句子整個含意大大不同其蘊涵關係也變成獨立關係了。

4.9 句子縮減

我們觀察訓練資料時發現有些句子將部份資訊去除掉就形成另一個句子,造成某個句子包涵著另一個句子所有的資訊,這樣兩個句子就有著方向性的蘊涵關係,對句子進行剖

析取得句子的剖析樹應用編輯距離的方法來取得句子重要資訊是否有所改變或相同來認定句子是否為縮減關係。

4.10 邏輯推理

我們觀察訓練資料時發現有些句子可以從句子的資訊可以合理推理出與另一個句子的含意，這樣的句子的我們初步將它分類在句子推理這一類。

表 5. 特殊類型例子

類型	例句
肯定/否定句	T1 319 槍擊案中，陳總統受槍擊時沒忘了穿防彈衣。
	T2 319 槍擊案中，陳總統受槍擊時沒穿防彈衣。
	兩句內容大都相同但因為 T1 句子含有反義詞造成文句對有著矛盾關係。
時間資訊不一致	T1 茱莉安德魯絲 1930 年出生
	T2 1935 年出生於英國的茱莉・安德魯絲。
	兩句由於時間有所不同造成句子表達的意思互相矛盾。
數字資訊不一致	T1 娜拉提諾娃一共獲得 18 座大滿貫金杯。
	T2 娜拉提洛娃一共取得了 58 個大滿貫的金杯。
	兩句由於數字有所不同造成句子訊息互相矛盾。
差異一詞	T1 一九七一年，印度協助東巴基斯坦脫離巴基斯坦成為孟加拉，結果印巴戰事再起。
	T2 一九七一年，印度協助西巴基斯坦脫離巴基斯坦成為孟加拉，結果印巴戰事再起。
	兩個句子可能只差異一個字詞就導致兩個句子的蘊涵關係改變，如例子兩個句子的主/受詞不同使得兩個句子的蘊涵變成互相獨立。
同義詞	T1 金泳三在一九九二年當選韓國第十四屆大統領。
	T2 金泳三在一九九二年當選韓國第十四屆總統。
	兩句中字詞雖然字面上有所不同但在語意上是相同的，因此這是雙向蘊涵關係。
背景知識	T1 2005 年全球恐怖攻擊活動不斷是第三世界向歐美霸權宣戰。
	T2 2005 年全球恐怖攻擊活動不斷是回教世界向歐美霸權宣戰。
	兩句中字詞雖然字面上有所不同但在有相當背景知識的人看來兩句表達的意思是相同，因此這是雙向蘊涵關係。
句法調換	T1 松花江汙染事件導致俄羅斯對大陸民眾反感日增
	T2 松花江汙染事件導致大陸對民眾俄羅斯反感日增

	兩句由於句子調換造成主詞與受詞關係改變表達的意思也有所改變。
一詞多義與命名實體	T1 馮小剛與和陳凱歌，吵的沒完了沒了
	T2《沒完沒了》是一部由馮小剛導演
	T1 與 T2 中都出現”沒完沒了”這個詞，然而所具有的涵義大大不同。
句子縮減	T1 在 1964 年 10 月中共成功試爆第一顆原子彈後，加快了中國本身的核子工業發展
	T2 1964 年中共第一次試爆原子彈
	T1 的句子部份資訊被刪掉但是不影響 T1 主要想表達的含意依然與 T2 含意相同，因此這是正向蘊涵關係。
邏輯推理	T1 張學良 1900 年 6 月 3 日出生，1920 年官拜少將
	T2 張學良廿歲官拜少將
	從 T1 句子可以得到關鍵字進而可以推理出 T2 句子的內容，因此這是一個正向蘊涵關係。

這些新的特殊類型目前需要使用人工的方式挑選出來，在未來將開發自動分類系統，然而有部分特殊類型並不像之前已完成的類型可以輕易程式化處理，例如:邏輯推理、背景知識、同義詞等特殊類型必須需要依靠語料資料才能夠正確的區分出來。

如表6所示RITE-2訓練與測試集資料中我們所分類所有的特殊類型句子中的數量，然而統計結果並非完全準確，這是因為一個句子對可能同時擁有多個特殊類型特性。

表 6. 特殊類型在訓練集與測試集涵蓋數量

特殊類型	**訓練集**	**測試集**
肯定/否定句	15(1.82%)	42(5.37%)
時間資訊不一致	43(5.28%)	60(7.68%)
數字資訊不一致	42(5.15%)	83(10.62%)
差異一詞	73(8.9%)	82(10.4%)
同義詞	53(6.5%)	43(5.5%)
背景知識	5(0.6%)	11(1.4%)
句法調換	67(8.2%)	52(6.6%)
一詞多義與命名實體	115(14.1%)	91(11.6%)
句子縮減	290(35.6%)	159(20.3%)
邏輯推理	111(13.6%)	86(11%)
粗略涵蓋	99.75%	90.47%

5. 實驗與討論

在這個章節中，我們將介紹 NTCIR10 RITE-2 測試集進行的幾個實驗設定的實驗結果。

5.1 實驗資料

本篇我們使用到的測試資料取自日本第十屆 NTCIR 研討會中 RITE-2(Recognizing Inference in Text-2)比賽子項目的開發資料(Development Data)以及公開測試集(open text)，在開發資料中 814 個文句對，另一個公開測試資料部份有 781 個文句對。

RITE-2 這個效能評鑑任務(task)，其目的在評鑑系統自動偵測句子之間「推論關係」的效能。句子之間的關係有很多種，比如說蘊涵(entailment)、意譯(paraphrase)以及矛盾(contradiction) 等。本次參與 RITE-2 評鑑的語言包含日文、簡體中文以及繁體中文，每個語言都包含數個「子任務」(subtask)。參與評鑑的系統需辨識給定的兩個文本(text)，輸出二選一(蘊涵與非蘊涵)或四選一(正向、雙向、獨立、矛盾)的蘊涵關係標記(label)。

5.2 實驗結果

我們在 NTCIR10 RITE-2 的正式評測簡體中文(CS)與繁體中文(CT)結果如表 7 表 8 所示，本次 NTCIR10 RITE-2 我們一共嘗試三種不同實驗設定，首先 CYUT-01 我們使用之前系統實驗結果效果最好實驗設定使用了上一個章節提到的前 10 個做測試，實驗設定 CYUT-02 延續 CYUT-01 的實驗設定加入之前提出的單一語言機器翻譯的方法作為的十一個特徵使用，實驗設定 CYUT-03 使用上述提到的 11 個特徵在加入一個是非句規則判斷作為的 12 個特徵使用，我們簡體中文以及繁體中文的實驗都是使用相同的實驗設定，我們在 NTCIR10 RITE-2 簡體中文這個項目取得第三名不錯成績，但在繁體中文這個項目表現比較不亮眼，相較於簡體中文繁體中文只得到第 8 名的成果。

表 7. NTCIR10 RITE-2 簡體中文正式評測結果

	BC (%)	**MC (%)**
CYUT-01	61.17	40.37
CYUT-02	63.11	42.52
CYUT-03	**67.86**	**40.37**

表 8. NTCIR10 RITE-2 繁體中文正式評測結果

	BC (%)	**MC (%)**
CYUT-01	**55.16**	**25.60**
CYUT-02	52.64	26.26
CYUT-03	51.58	23.51

5.2.1 改進實驗一

我們系統簡體中文與繁體中文使用相同的實驗流程但是如表 7 和表 8 所示，實驗結果有著顯著的不同，雖然簡體中文與繁體中文使用的測試資料不同，可能會造成兩個實驗結果有所不同，但我們認為不只這個因素可以造成如此大的差異。比較兩種中文的實驗流程，我們的系統處理繁體中文相對於簡體中文多了將繁體中文翻譯成簡體中文的步驟，深入研究翻譯過後繁體中文測試資料可以發現翻譯效果不佳造成之後抽取特徵時得到錯誤的數值造成之後 SVM 的誤判，因此改使用我們自行開發的機器翻譯系統，解決之前翻譯錯誤產生的空格與術語錯誤的問題提高系統效能，替換後的實驗結果如表 9 所示，在兩類(BC)任務提高 6.02 正確率以及多類(MC)任務則是提高 9.49 正確率。

表 9. 替換簡繁轉換系統實驗結果

簡繁轉換系統	BC (%)	MC (%)
Google 翻譯	53.64	26.26
CYUT 開發系統	**59.54(+6.02)**	**35.75(+9.49)**

5.2.2 改進實驗二

在前面章節提到特殊類型問題是過去系統所無法適當處理的問題，因此針對特殊類型問題進行開發個別專屬系統處理是有所必要，本次將兩類(BC)任務中特殊類型問題句子對抽出來進行個別實驗，實驗結果如表 10 所示，我們可以發現特殊類型使用特別開發的子系統作處理可以提高其正確率。

表 10. 個別特殊類型子系統做兩類(BC)實驗結果

	Case1 肯定/否定句	Case2 時間資訊不一致	Case3 數字資訊不一致	Case4 差異一詞
正式評測	52.38%	58.33%	52.25%	47.59%
個別特殊處理	71.42%	70%	71.25%	60.97%

在表 10 的實驗結果顯示，特殊類型子系統有助於提升兩類(BC)系統效果，因此我們認為處理多類(MC)時也可以提高系統效能，接著我們將特殊類型處理的方法加入系統之中做實驗，我們選擇在正式評測中效果最好的 CYUT-03 做測試，實驗結果如表 11 所示。

表 11. 加入特殊類型方法簡體中文實驗結果

項目	**BC (%)**	**MC (%)**
Cyut	67.86	40.37
Cyut+case1	68.88	42.08
Cyut+case1+case2	69.78	43.45
Cyut+case1+case2+case3	71.63	45.13
Cyut+case1+case2+case3+case4	**72.92**	**45.92**

5.3 實驗討論

在其他語言的文字蘊涵處理研究中也有與我們相似的類型處理例如在日文中(Shibata *et al.*, 2013)等人使用到的同義詞、上下位詞與我們所沒有的反義詞處理，(Makino *et al.*, 2013)等人使用了命名實體名稱作處理，在英文方面有許多文字蘊涵相關的研究也使用到 WordNet 提取同義詞與上下位詞作為擴增資訊，我們參考其他語言使用到的方法來個別開發子系統作處理，在 NTCIR10 RITE-2 日文測試資料中參雜了部份特殊語言現象，如同義詞、倒裝句、代詞等等更貼近現實人類使用的語言，因此特殊類型的處理是未來研究方向。

從實驗結果可以發現我們所提出的特殊類型問題需個別處理的方法有助於提升系統效果，然而我們可以 case4 的差異一詞改進的幅度不如其他子系統，仔細比較前後兩個系統結果後，發現差異一詞子系統雖然可以將原本錯誤的問題處理正確但也會將原本正確問題誤判，這是因為句子本身不只擁有一種特殊類型以及兩個句子之間差異詞對句子含義重要性不一致，子系統過度針對不重要的詞造成誤判。

6. 結論與未來工作

本文報告了我們的系統針對 NTCIR10 RITE-2 正式評測的結果所做的改進，在最後的實驗結果可以看到我們與之前系統改進幅度，改進簡體中文兩類系統，其成果從原本的 67.86%提升到 72.92%。改進簡體中文多類系統，其成果從原本的 40.37%提升到 45.92%。在繁體中文任務的結果發現一個好的簡繁轉換系統有助於我們系統改進，然而簡體中文與繁體中文之間還是有著些微差異有待克服。

在分析過去系統實驗結果後發現許多是以前的方法無法處理的問題，在其他語言的文字蘊涵研究中提出許多方法來增加系統可處理的句子數量，因此我們參考其他語言增加處理的方法提出特殊類型問題分類處理的構想，在個別特殊類型實驗結果證明我們提出的特殊類型問題因個別處理的想法是正確的，藉由個別問題使用個別子系統進行處理以提升系統整體效能，然而我們目前提出的特殊類型分類並非完美，還需要規畫更貼近問題本身的分類類型與建立更多特殊類型子系統來減少因為多重特殊類型造成系統的誤判，針對一些類型問題還需要收集更多語料資料建立資料集。

致謝

研究依經濟部補助財團法人資訊工業策進會「102 年度 數位匯流服務開放平台技術研發計畫 (4/4) 」辦理。本研究感謝國科會贊助部分經費，計畫編號為 NSC 100-2221-E-324-25-MY2，謹此致謝。

參考文獻

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL recognizing textual entailment challenge.

Dagan, I., & Glickman, O. (2004). Probabilistic textual entailment: Generic applied modeling of language variability. In *Proceedings of the Workshop on Learning Methods for Text Understanding and Mining*, Grenoble, France, 2004.

Hua, D. B., & Dinga, J. (2011). Study on Similar Engineering Decision Problem Identification Based on Combination of Improved Edit-Distance and Skeletal Dependency Tree with POS. *Systems Engineering Procedia*, *1*, 406-413.

Liu, Q., & Li, S. J. (2002). Word Similarity Computing Based on How-net. *International Journal of Computational Linguistics and Chinese Language Processing*, *7*(2), 59-76.

Li, M. H., Wu, S. H., Zeng, Y. C., Yang, P. C., & Ku, T. (2010). Chinese Characters Conversion System based on Lookup Table and Language Model. *International Journal of Computational Linguistics and Chinese Language Processing*, *15*(1), 19-36.

Makino, T., Okajima, S., & Iwakura, T. (2013). FLL: Local Alignments based Approach for NTCIR-10 RITE-2. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013.

Ou, Y. C. (2010). Recognize Textual Entailment by the Lexical and Semantic Matching. *Computer Application and System Modeling, 2010 International Conference* on V2-500 -504.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19-51.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on ACL*, 311-318, Philadelphia, PA.

Quang, M., Pham, N., Nguyen, L. M., & Shimazu, A. (2012). Using Machine Translation for Recognizing Textual Entailment in Vietnamese Language. *2012 IEEE RIVF International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future*.

Shibata, T., Kurohashi, S., Kohama, S., & Yamamoto, A. (2013). Predicate-argument Structure based Textual Entailment Recognition System of KYOTO Team for NTCIR-10 RITE-2. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013

Wu, S. H., Huang, W. C., Chen, L. P., & Ku, T. (2011). Binary-class and Multi-class Chinese Textural Entailment System Description in NTCIR-9 RITE. In *Proceedings of the NTCIR-9 workshop*, Tokyo, Japan, 6-10 Dec., 2011

Wang, X. L., Zhao, H., & Lu, B. L. (2013). BCMI-NLP Labeled-Alignment-Based Entailment System for NTCIR-10 RITE-2 Task. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013.

Watanabe, Y., *et al*. (2013). Overview of the Recognizing Inference in Text (RITE-2) at the NTCIR-10 Workshop. In *Proceedings of the NTCIR-10 conference*, Tokyo, Japan, 18-21 June, 2013.

Yang, S. S., Wu, S. H., Chen, L. P., Hsieh, W. T., & Chou, S. T. (2012). Improving Binary-class Chinese Textural Entailment by Monolingual Machine Translation Technology. In *Proceedings of the IEEE IRI 2012*, Las Vegas, USA, 8 Aug, 2012.

Zhang, Y., *et al*. (2011). ICRC_HITSZ at RITE: Leveraging Multiple Classifiers Voting for Textual Entailment Recognition. In *NTCIR-9 RITE, in Proceedings of the NTCIR-9 workshop*, Tokyo, Japan, 6-10 Dec., 2011.

Zhou, L., Lin, C. Y., & Hovy, E. (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the Conference on EMNLP*, 77-84, Sydney, Australia, 2006.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 17-30

Integrating Dictionary and Web N-grams for Chinese Spell Checking

Jian-cheng Wu*, Hsun-wen Chiu+, and Jason S. Chang*,+

Abstract

Chinese spell checking is an important component of many NLP applications, including word processors, search engines, and automatic essay rating. Nevertheless, compared to spell checkers for alphabetical languages (*e.g.*, English or French), Chinese spell checkers are more difficult to develop because there are no word boundaries in the Chinese writing system and errors may be caused by various Chinese input methods. In this paper, we propose a novel method for detecting and correcting Chinese typographical errors. Our approach involves word segmentation, detection rules, and phrase-based machine translation. The error detection module detects errors by segmenting words and checking word and phrase frequency based on compiled and Web corpora. The phonological or morphological typographical errors found then are corrected by running a decoder based on the statistical machine translation model (SMT). The results show that the proposed system achieves significantly better accuracy in error detection and more satisfactory performance in error correction than the state-of-the-art systems.

Keywords: Chinese Spelling Detection, Chinese Spelling Correction, Chinese Similar Characters, Ngram, Language Model, Machine Translation.

1. Introduction

Chinese spell checking is a task involving automatically detecting and correcting typographical errors (typos), roughly corresponding to misspelled words in English. In this paper, we define typos as Chinese characters that are misused due to shape or phonological similarity. Liu *et al*. (2011) shows that people tend to unintentionally generate typos that sound similar (*e.g.*, *措折 [cuo zhe] and 挫折 [cuo zhe]), or look similar (*e.g.*, *固難 [gu nan] and 困難 [kun nan]). On the other hand, some typos found on the Web (such as forums

* Department of Computer Science, National Tsing Hua University
E-mail: { wujc86; jason.jschang}@gmail.com
+ Department of Institute of Information Systems and Applications, National Tsing Hua University
E-mail: chiuhsunwen@gmail.com

or blogs) are used deliberately for the purpose of speed typing or just for fun. Therefore, spell checking is an important component for many applications, such as computer-aided writing and corpus cleanup.

The methods of spell checking can be classified broadly into two types: rule-based methods (Ren *et al*., 2001; Jiang *et al*., 2012) and statistical methods (Hung & Wu, 2009; Chen & Wu, 2010). Rule-based methods use knowledge resources, such as a dictionary, to identify a word as a typo if the word is not in the dictionary and to provide similar words in the dictionary as suggestions. Simple rule-based methods, however, have their limitations. Consider the sentence "心是很重要的。" [xin shi hen zhong yao de] which is correct. Nevertheless, the two single-character words "心" [xin] and "是" [shi] are likely to be regarded as an error by a rule-based model for the longer word "心事" [xin shi] with identical pronunciation.

Data-driven, statistical spell checking approaches appear to be more robust and perform better. Statistical methods tend to use a large monolingual corpus to create a language model to validate the correction hypotheses. Considering "心是" [xin shi], the two characters "心" [xin] and "是" [shi] are a bigram with high frequency in a monolingual corpus, so we may determine that "心是" [xin shi] is not a typo after all.

In this paper, we propose a model that combines rule-based and statistical approaches to detect errors and generate the most appropriate corrections in Chinese text. Once an error is identified by the rule-based detection model, we use the statistic machine translation (SMT) model (Koehn, 2010) to provide the most appropriate correction. Rule-based models tend to ignore context, so we use SMT to deal with this problem. Our model treats spelling correction as a kind of translation, where typos are translated into correctly spelled words according to the translation probability and language model probability. Consider the same case "心是很重要的。" [xin shi hen zhong yao de]. The string "心是" [xin shi] would not be incorrectly replaced with "心事" [xin shi] because we would consider "心是" [xin shi] to be highly probable, according to the language model.

The rest of the paper is organized as follows. We present the related work in the next section. Then, we describe the proposed model for automatically detecting the spelling errors and correcting the found errors in Section 3. Section 4 and Section 5 present the experimental data, results, and performance analysis. We conclude in Section 6.

2. Related work

Chinese spell checking is a task involving automatically detecting and correcting typos in a given Chinese sentence. Previous work typically takes the approach of combining a confusion set and a language model. A rule-based approach depends on dictionary knowledge and a

confusion set, a collection set of certain characters consisting of visually and phonologically similar characters. On the other hand, statistical-based methods usually use a language model, which is generated from a reference corpus. A statistical language model assigns a probability to a sentence of words by means of n-gram probability to compute the likelihood of a corrected sentence.

Chang (1995) proposed a system that replaces each character in the sentence based on the confusion set and estimates the probability of all modified sentences according to a bigram language model built from a newspaper corpus before comparing the probability before and after substitution. They used a confusion set consisting of pairs of characters with similar shape that were collected by comparing the original text and its OCR results. Similarly, Zhuang *et al*. (2004) proposed an effective approach using OCR to recognize a possible confusion set. In addition, Zhuang *et al.* (2004) also used a multi-knowledge based statistical language model, the n-gram language model, and Latent Semantic Analysis. Nevertheless, the experiments by Zhuang *et al*. (2004) seem to show that the simple n-gram model performs the best.

In recent years, Chinese spell checkers have incorporated word segmentation. The method proposed by Huang *et al*. (2007) incorporates the Sinica Word Segmentation System (Ma & Chen, 2003) to detect typos. With a character-based bigram language model and the rule-based methods of dictionary knowledge and confusion sets, the method determines whether the word is a typo or not. There are many more systems that use word segmentation to detect errors. For example, in Hung and Wu (2009), the given sentence is segmented using a bigram language model. In addition, the method also uses a confusion set and common error templates manually edited and provided by the Ministry of Education in Taiwan (MOE, 1996). Chen and Wu (2010) modified the system proposed by Hung and Wu (2009) by combining statistic-based methods and a template matching module generated automatically to detect and correct typos based on the language model.

Closer to our method, Wu *et al*. (2010) adopted the noise channel model, a framework used both in spell checkers and in machine translation systems. The system combined a statistic-based method and template matching with the help of a dictionary and a confusion set. They also used word segmentation to detect errors, but they did not use existing word segmentation, as Huang *et al*. (2007) did, because that might regard a typo as a new word. They used a backward longest first approach to segment sentences with an online dictionary sponsored by MOE (MOE, 2007), and a templates with a confusion set provided by Liu *et al*. (2009). The system also treated Chinese spell checking as a kind of translation by combining the template module and translation module to get a higher precision or recall.

In our system, we also treat the Chinese spell checking problem as machine translation, but we use a different method of handling word segmentation to detect typos and translation

model, where typos are translated into correctly spelled words.

3. Method

In this section, we describe our solution to the problem of Chinese spell checking. In the error detection phase, the given Chinese sentence is segmented into words. (Section 3.1) The detection module then identifies and marks the words that may be typos. (Section 3.2) In the error correction phase, we use the statistical machine translation (SMT) model to translate the sentences containing typos into correct ones (Section 3.3). In the rest of this section, we describe our solution to this problem in more detail.

3.1 Modified Chinese Word Segmentation System

Unlike English text, in which sentences are sequences of words delimited by spaces, Chinese texts are represented as strings of Chinese characters (called Hanzi) with word delimiters. Therefore, word segmentation is a pre-processing step required for many Chinese NLP applications. In this study, we also perform word segmentation to reduce the search space and probability of false alarms. After segmentation, sequences of two or more singleton words are considered likely to contain an error. Nevertheless, over-segmentation might lead to falsely identified errors, which we will describe in Section 3.2. Considering the sentence "除了要有超世之才，也要有堅定的意志" [chu le yao you chao shi zhi cai, ye yao you jian ding de yi zhi], the sentence is segmented into "除了/要/有/超世/之/才/，/也/要/有/堅定/的/意志." The part "超世之才" [chao shi zhi cai] of the sentence is over-segmented and runs the risk of being identified as containing a typo. To solve the problem of over-segmentation, we used additional lexical items to reduce the chance of generating false alarms.

3.2 Error Detection

Motivated by the observation that a typo often causes over-segmentation in the form of a sequence of single-character words, we target the sequences of single-character words as candidates for typos. To identify the points of typos, we take all n-grams consisting of single-character words in the segmented sentence into consideration. In addition to a Chinese dictionary, we also include a list of web-based n-grams to reduce false alarms due to the limited coverage of the dictionary.

When a sequence of singleton words is not found in the dictionary or in the web-based character n-grams, we regard the n-gram as containing a typo. For example, "森林 的 芳 多 精" [sen lin de fang duo jing] is segmented into consecutive singleton words: bigrams such as "的 芳" [de fang], and "芳 多" [fang duo] and trigrams such as "的 芳 多" [de fang duo] and "芳 多 精" [fang duo jing] are all considered as candidates for typos since those n-grams are not found in the reference list.

3.3 Error Correction

Once we generate a list of candidates of typos, we attempt to correct typos using a statistical machine translation model to translate typos into correct words. When given a candidate, we first generate all correction hypotheses by replacing each character of the candidate typo with similar characters, one character at a time.

Take the candidate "氣份" [qi fen] as example, the model generates all translation hypotheses according to a visually and phonologically confusion set. Table 1 shows some translation hypotheses. The translation hypotheses then are validated (or pruned from the viewpoint of SMT) using the dictionary.

Table 1. Sample "translations" for the candidate "氣份" [qi fen].

Replaced character	氣		份	
Translations	汽份	泣份	氣分	氣忿
	器份	契份	氣憤	氣糞
	企份	憩份	氣奮	氣吩
	訖份	氮份	氣扮	氣汾
	迄份	粥份	氣芬	氣氛

The translation probability *tp* is a probability indicating how likely a typo is to be translated into a correct word. *tp* of each correction translation is calculated using the following formula:

$$tp(candi,\ trans) = \log_{10}\left(\frac{freq(trans)}{freq(trans) - freq(candi)} * \gamma\right)^{\substack{if\ trans\ in\ ngrams \\ otherwise = 0}} \quad (1)$$

where *freq(trans)* is the frequency of translation, *freq(candi)* is the frequency of the candidate, and γ is the weight of different error types: visual or phonological.

Take "氣份" [qi fen] from "不/一樣/的/氣/份" [bu/yi yang/de/qi/fen] for instance, the translations with non-zero *tp* after filtering are shown in Table 2. Only two translations are possible for this candidate: "氣憤" [qi fen] and "氣氛" [qi fen].

Table 2. Translations for "氣份" [qi fen] with corresponding translation probability and language model probability (log).

Translations	Frequency	LM probability	tp
氣憤	48	-4.96	-1.20
氣氛	473	-3.22	-1.11

We use a simple, publicly available decoder written in Python to correct potential spelling errors found in the detection module. The decoder reads one Chinese sentence at a

time and attempts to "translate" the sentence into a correctly spelled one. The decoder translates monotonically without reordering the Chinese words and phrases using two models — the translation probability model and the language model. These two models read from a data directory containing two text files containing a translation model in GIZA++ (Och & Ney, 2003) format and a language model in SRILM (Stolcke *et al*., 2011) format. These two models are stored in memory for quick access.

The decoder invokes the two modules to load the translation and language models and decodes the input sentences, storing the result in output. The decoder computes the probability of the output sentences according to the models. It works by summing over all possible ways that the model could have generated the corrected sentence from the input sentence. Although, in general, covering all possible corrections in the translation and language models is intractable, a majority of error instances can be "translated" effectively via the translation model and the language model.

4. Experimental Setting

Our systems were designed to provide wide coverage spell checking for Chinese. As such, we trained our systems using a dictionary, a compiled corpus, and Web scale n-grams. We evaluated our systems on the sentence level. Finally, we used an annotated dataset to provide human judges the ability to evaluate the quality of error detection and correction. In this section, we first present the details of data sources used in training (Section 4.1). Then, Section 4.2 describes the test data. Section 4.3 describes the systems evaluated and compared. The evaluation metrics for the performance of the systems are reported in Section 4.4.

4.1 Data Sources

To train our model, we used several corpora, including Sinica Chinese Balanced Corpus, TWWaC (Taiwan Web as Corpus), a Chinese dictionary, and a confusion set. We describe the data sets in more detail below.

Sinica Corpus

"Academia Sinica Balanced Corpus of Modern Chinese," or "Sinica Corpus," is the first balanced Chinese corpus with part-of-speech tags (Huang *et al*., 1996). The current size of the corpus is about 5 million words. Texts are segmented according to the word segmentation standard proposed by the ROC Computational Linguistic Society. Each segmented word is tagged with its part of speech. We used the corpus to generate the frequency of bigrams, trigrams, and 4-grams for training the translation model and to train the n-gram language model.

TWWaC (Taiwan Web as Corpus)

We used TWWaC for obtaining more language information. TWWaC is a corpus gathered from the Web under the .tw domain, containing 1,817,260 Web pages that consist of 30 billion Chinese characters. We use the corpus to generate the frequency of all character n-grams for n = 2, 3, 4 (with frequency higher than 10). Table 3 shows the information of n-grams in Sinica Corpus and TWWaC.

Table 3. The information of n-grams in Sinica corpus and TWWaC.

N-gram	Sinica Corpus Types	TWWaC Types
2-gram	66,778	2,848,193
3-gram	45,382	13,745,743
4-gram	12,294	17,191,359

Words and Idioms in a Chinese Dictionary

From the dictionaries and related books published by Ministry of Education (MOE) of Taiwan, we obtained two lists, one is the list of 64,326 distinct Chinese words (MOE, 1997)[1], and the other one is the list of 48,030 distinct Chinese idioms[2]. We combined the lists into a Chinese dictionary for validating words with lengths of 2 to 17 characters.

Confusion Set

After analyzing erroneous Chinese words, Liu *et al*. (2011) found that more than 70% of typos were related to the phonologically similar character, about 50% are morphologically similar, and almost 30% are both phonologically and morphologically similar. We used the ratio as the weight for the translation probabilities. In this study, we used two confusion sets generated by Liu *et al*. (2011) and provided by SIGHAN 7 Bake-off 2013: Chinese Spelling Check Shared Task as a full confusion set, based on loosely similar relation.

In order to improve the performance, we expanded the sets slightly and also removed some loosely similarly relations. For example, we removed all relations based on non-identical phonologically similarity. After that, we added the similar characters based on similar phonemes in Chinese phonetics, such as “ㄣ，ㄥ” [en, eng], “ㄤ，ㄢ” [ang, an], “ㄕ，ㄙ” [shi, si], and so on. We also modified the similar shape set, so we checked the characters by comparing the characters in Cangjie codes (倉頡碼) and required strong shape similarly. Two characters differing from each other by at most one symbol in Cangjie code were considered as strongly similar and were retained. For example, the code of “徵” [zheng] and “微” [wei]

[1] Chinese Dictionary http://www.edu.tw/files/site_content/m0001/pin/biau2.htm?open

[2] Chinese Idioms http://dict.idioms.moe.edu.tw/cydic/index.htm

are strongly similar in shape, since in their corresponding codes“竹人山土大”and“竹人山山大”, differ only in one place.

4.2 Test Data

We used the official dataset from SIGHAN 7 Bake-off 2013: Chinese Spelling Check to evaluate our systems. This dataset contains two parts: 350 sentences with errors and 350 sentences without errors, extracted from student essays that covered various common errors. The dataset was released in XML format with the information of sentences, wrong position, typos, and correction. A sample is shown below:

<DOC Nid="00001">
<P>我看過許多勇敢的人，不怕措折地奮鬥，這種精神值得我們學習。</P>
<TEXT>
<MISTAKE wrong_position=13>
<WRONG>措折</WRONG>
<CORRECT>挫折</CORRECT>
</MISTAKE>
</TEXT>
</DOC>

We found that all of the sentences with errors contain exactly one typo and that most errors were either similar in pronunciation or shape. Therefore, the confusion set was suitable for error correction. We generated the sentence with/without error and the correct answer from XML format. In this data, more than 80% of errors were characters with identical pronunciation, almost 20% of errors were characters with similar shape, and 40% of errors involved both phonological and visual similarity. Hence, we focused on detecting and correcting these two common types of errors in our study.

4.3 Systems Compared

Recall that we propose a system to detect and correct typos in Chinese based broadly on statistical machine translation. We experimented with different resources as kinds of language models to detect typos: dictionary entries, a compiled corpus, and Web corpus. The four detection systems evaluated are:

—Dictionary (**DICT**): A dictionary is used to detect unregistered words as errors.

—Corpus (**CORP**): A word list from a reference corpus is used to detect unseen words as errors.

—Web corpus (**WEB**): A character n-gram of Web corpus is used to detect unseen n-grams as errors.

—Dictionary and Web corpus (**DICT+WEB**): A dictionary combining a character n-gram of Web corpus is used to detect unregistered words as errors.

To correct typos, we used a character confusion set to transform the detected typos and generate the "*translation*" hypotheses with translation probability. These hypotheses were pruned using a Chinese dictionary before running the MT decoder in order to reduce the load on the decoder. The scope of this confusion set and the weights associated with translation probability clearly influenced the performance of our system. We evaluated and compared four different confusion set and weight settings. The four correction systems evaluated are:

—Full confusion set (**FULL+WT**): A broad confusion set with loosely similar relations in character sound and shape was used to generate mapping from a detected typo to its correction. Different weights were used in modeling probability for sound and shape based mapping.

—Confusion set with identical sound (**SND+WT**): A broad confusion set with identical sounds and loosely similar shape relations was used to generate mapping. Different weights were used in modeling probability for sound and shape based mapping.

—Restricted confusion set with identical sound and strong similarly shape (**SND+SHP**): A broad confusion set with identical sounds and strongly similar shape relations was used to generate mapping. Sound and shape were given the same weight.

—Restricted confusion set with different weights (**SND+SHP+WT**): A broad confusion set with identical sounds and strongly similar shape relations was used to generate mapping. Different weights were used in modeling probability for sound and shape based mapping.

4.4 Evaluation Metrics

To assess the effectiveness of the proposed system, we used test data to experiment with our system. We also exploited several language resources, including TWWaC, Sinica Corpus, a Chinese dictionary, and the confusion set, in the proposed system to detect errors and correct errors. The Chinese Word Segmentation System produces the word segmentation result with the help of a Chinese dictionary to improve the proposed system. To evaluate our system, we used the precision rate and recall rate, which are defined as follows:

$$Precision = C / S \tag{2}$$

$$Recall = C / N \tag{3}$$

where N is the number of error characters, S is the number of characters translated by the proposed system, and C is the number of characters translated correctly by the proposed system. We also compute the corresponding F-score as:

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (4)$$

5. Evaluation Results

In this section, we report the results of the experimental evaluation using the methodology described in the previous section. We evaluated detection, as well as correction, for many systems with different language resources and settings. During this evaluation, we tested our systems on 350 sentences containing at least one typo, provided in SIGHAN Bake-off 2013: Chinese Spelling Check. Table 4 shows the precision, recall, and F-score for four detection systems, while Table 5 shows the same metrics for four correction systems.

Table 4. The comparison of the detection with different references.

System	Precision	Recall	F-score
DICT	.91	.52	.66
CORPUS	.90	.46	.61
WEB	.93	.47	.63
WEB+DICT	**.95**	**.56**	**.71**

Table 5. The comparison of the correction experiment.

System	Precision	Recall	F-score
FULL+WT	.53	.51	.52
SND+WT	.74	.57	.65
SND+SHP	.90	.55	.68
SND+SHP+WT	**.95**	**.56**	**.70**

As can be seen in Table 4, using the Web corpus (**WEB**) achieves higher precision than the dictionary (**DICT**) or compiled corpus (**CORPUS**) with slightly lower recall. Using the dictionary (**DICT**) leads to the highest recall but slightly lower precision. By combining the dictionary and Web corpus (**WEB+DICT**), we achieve the best precision, recall, and F-score.

Table 5 shows that using the full confusion set with loosely similar sound and shape relation leads to the lowest recall and precision in error correction (**FULL**). By restricting the sound confusion to identical sound and the shape confusion to strongly similar shape, we can improve precision dramatically, with a small increase in recall (**SND** and **SND+SHP**).

We can further improve the precision and recall by applying different weights in modeling the probability of sound and shape based hypotheses (**SND+SHP+WT**). Since typos are more often related to sound confusion than shape, giving higher weight to sound confusion

indeed leads to further improvement in both precision and recall. Previous works typically have used only a language model to correct errors, but we compute language model probability and translation probability, resulting in more effective error correction. For this reason, we were placed among the top scoring systems in the SIGHAN Bake-off 2013.

In order to test whether the system can produce false alarms as rarely as possible, when handling the sentences with typos, we tested our systems on a dataset with an additional 350 sentences without typos. The best performing system (**SND+SHP+WT**) obtained a precision rate of .91, recall rate of .56, and F-score of .69 in correction. The results show that this system is very robust, maintaining a high precision rate in different situations.

The recall of our system is limited by the dictionary that we used to correct a typo. For example, the typo "七彈場" [qi tan chang], which is detected by the model, is not corrected to "漆彈場" [qi tan chang] because it is a new term and not found in the Chinese dictionary we used. To correct such errors, we could use Web-based character n-grams, which are more likely to contain such new terms or productive compounds not found in a dictionary.

6. Conclusions and Future Work

Many avenues exist for future research and improvement of our system. For example, new terms can be automatically discovered and added to the Chinese dictionary to improve both detection and correction performance. Part of speech tagging can be performed to provide more information for error detection. Named entities can be recognized in order to avoid false alarms. A supervised statistical classifier can be used to model translation probability more accurately. Additionally, an interesting direction to explore is using Web n-grams in addition to a Chinese dictionary for correcting typos. Yet another direction of research would be to consider errors related to a missing or redundant character.

In summary, we have proposed a novel method for Chinese spell checking. Our approach involves error detection and correction based on the phrasal statistical machine translation framework. The error detection module detects errors by segmenting words and checking word and phrase frequency based on a compiled dictionary and Web corpora. The phonological or morphological spelling errors found then are corrected by running a decoder based on the statistical machine translation model (SMT). The results show that the proposed system achieves significantly better accuracy in error detection and more satisfactory performance in error correction than the state-of-the-art systems. The experimental results show that the method outperforms previous works.

References

Chang, C.-H. (1995). A new approach for automatic Chinese spelling correction. In *Proceedings of Natural Language Processing Pacific Rim Symposium*, 278 - 283.

Chen, Y.-Z. (2010). *Improve the detection of improperly used Chinese characters with noisy channel model and detection template*. Master thesis, Chaoyang University of Technology.

Huang, C.-R., Chen, K.-j., & Chang, L.-L. (1996). Segmentation standard for Chinese natural language processing. In *Proceedings of the 1996 International Conference on Computational Linguistics (COLING 96)*, *2*, 1045 - 1048.

Huang, C.-M., Wu, M.-C., & Chang C.-C. (2007). Error detection and correction based on Chinese phonemic alphabet in Chinese text. In *Proceedings of the 4th International Conference on Modeling Decisions for Artificial Intelligence (MDAI IV)*, 463 - 476.

Hung, T.-H. (2009). *Automatic Chinese character error detecting system based on n-gram language model and pragmatics knowledge base*. Master thesis, Chaoyang University of Technology.

Jiang, Y., *et al*. (2012). A rule based Chinese spelling and grammar detection system utility. *2012 International Conference on System Science and Engineering (ICSSE)*, 437 - 440, 30 June - 2 July 2012.

Koehn, P. (2010). *Statistical Machine Translation*. United Kingdom: Cambridge University Press.

Liu, C.-L., Tien, K.-W., Lai, M.-H., Chuang, Y.-H., & Wu, S.-H. (2009). Phonological and logographic influences on errors in written Chinese words. In *Proceedings of the Seventh Workshop on Asian Language Resources (ALR7), the Forty Seventh Annual Meeting of the Association for Computational Linguistics (ACL'09)*, 84 - 91.

Liu, C.-L., Lai, M.-H., Tien, K.-W., Chuang, Y.-H., Wu, S.-H., &Lee, C.-Y. (2011). Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications. *ACM Trans. Asian Lang, Inform. Process*, *10*(2), Article 10, pages 39, .

Ma, W.-Y., & Chen, K.-J. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. In *Proceedings of ACL, Second SIGHAN Workshop on Chinese Language Processing*, *17*, 168 - 171.

MOE. (1997). *MOE word frequency table*, Taiwan: Ministry of Education.

MOE. (2007). *MOE Dictionary new edition*. Taiwan: Ministry of Education.

MOE. (1996). *Common errors in Chinese writings*. Taiwan: Ministry of Education.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19 - 51.

Ren, F., Shi, H., & Zhou, Q. (2001). A hybrid approach to automatic Chinese text checking and error correction. *2001 IEEE International Conference on Systems, Man, and Cybernetics*, *3*, 1693 - 1698, 07 - 10 Oct. 2001.

Stolcke, A., Zheng, J., Wang, W., & Abrash, V. (2011). SRILM at Sixteen: Update and Outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, Dec. 2011.

Wu, S.-H., Chen, Y.-X., Yang, P.-c., Ku, T., & Liu, C.-L. (2010). Reducing the false alarm rate of Chinese character error detection and correction. In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP 2010)*, 54 - 61, 28 - 29 Aug. 2010.

Zhuang, L., Bao, T., Zhu, X., Wang, C., & Naoi, S. (2004). A Chinese OCR spelling check approach based on statistical language models. *2004 IEEE International Conference on Systems, Man and Cybernetics*, 5, 4727 - 4732, 10 - 13 Oct. 2004.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 31-44

Correcting Serial Grammatical Errors based on N-grams and Syntax

Jian-cheng Wu*, Jim Chang*, and Jason S. Chang*

Abstract

In this paper, we present a new method based on machine translation for correcting serial grammatical errors in a given sentence in learners' writing. In our approach, translation models are generated to translate the input into a grammatical sentence. The method involves automatically learning two translation models that are based on Web-scale n-grams. The first model translates trigrams containing serial preposition-verb errors into correct ones. The second model is a back-off model, used in the case where the trigram is not found in the training data. At run-time, the phrases in the input are matched and translated, and ranking is performed on all possible translations to produce a corrected sentence as output. Evaluation on a set of sentences in a learner corpus shows that the method corrects serial errors reasonably well. Our methodology exploits the state-of-the art in machine translation, resulting in an effective system that can deal with many error types at the same time.

Keywords: Grammatical Error Correction, Serial Errors, Machine Translation, N-grams, Language Model

1. Introduction

Many people are learning English as a second or foreign language: it is estimated there are 375 million English as a Second Language (ESL) and 750 million English as a Foreign Language (EFL) learners around the world, according to Graddol (2006). Three times as many people speak English as a second language as there are native speakers of English. Nevertheless, non-native speakers tend to make many kinds of errors in their writing, due to the influence of their native languages (*e.g.*, Chinese or Japanese). Therefore, automatic grammar checkers are needed to help learners improve their writing. In the long run, automatic grammar checkers also can help non-native writers learn from the corrections and

* Department of Computer Science, National Tsing Hua University
E-mail: {wujc86; jim.chang.nthu; jason.jschang}@gmail.com

gradually gain better command of grammar and word choices.

The grammar checkers available in popular word processors have been developed with a focus on native speaker errors, such as subject-verb agreement and pronoun reference. Therefore, these word processors (*e.g.*, Microsoft Word) often offer little or no help with common errors causing problems for English learners (*e.g.*, missing, unnecessary, or wrong article, preposition, and verb form) as described in The Longman Dictionary of Common Errors, second edition (LDOCE) by Heaton and Turton (1996). The LDOCE is the result of analyzing errors encoded in the Longman Learners' Corpus.

The LDOCE shows that grammatical errors in learners' writing can either appear in isolation (*e.g.*, the wrong proposition in "*I want to improve my ability of* [*in*] *English*.") or consecutively (*e.g.*, the unnecessary preposition immediately followed by a wrong verb form in "*These machines are destroying our ability of thinking* [*to think*]."). We refer to two or more errors appearing consecutively as *serial errors*.

Previous works on grammar checkers either have focused on handling one common type of error exclusively or handling it independently in a sequence of errors. Nevertheless, when an error is not isolated, it is difficult to correct the error when another related error is in the immediate context. In other words, when serial errors occur in a sentence, a grammar checker needs to correct the first error in the presence of the second error (or *vice-versa*), making correction difficult to achieve. These errors could be corrected more effectively if the corrector recognized them as serial errors and attempted to correct the serial errors at once.

Consider an erroneous sentence, "*I have difficulty to understand English*." The correct sentence should be "*I have difficulty in understanding English*." It is hard to correct these two errors one by one, since the errors are dependent on each other. Intuitively, by identifying "*difficulty to understand*" as containing serial errors and correcting it to "*difficulty in understanding*," we can handle this kind of problem more effectively.

Input: *I have difficulty to understand English.* .

Phrase table of translation model:		Back-off translation model:	
difficulty of understanding	||| difficulty in understanding ||| 0.86	difficulty of VERB+ing	||| difficulty in VERB+ing ||| 0.34
difficulty to understand	||| difficulty in understanding ||| 0.86	difficulty to VERB	||| difficulty in VERB+ing ||| 0.34
difficulty with understanding	||| difficulty in understanding ||| 0.86	difficulty with VERB+ing	||| difficulty in VERB+ing ||| 0.34
difficulty in understand	||| difficulty in understanding ||| 0.86	difficulty in VERB	||| difficulty in VERB+ing ||| 0.34
difficulty for understanding	||| difficulty in understanding ||| 0.86	difficulty for VERB+ing	||| difficulty in VERB+ing ||| 0.34
difficulty about understand	||| difficulty in understanding ||| 0.86	difficulty about VERB+ing	||| difficulty in VERB+ing ||| 0.34

Output: *I have difficulty in understanding English.*

Figure 1. Example session of correcting the sentence, "I have difficulty to understand English."

We present a new system that automatically generates a statistical machine translation model based on a trigram containing a word followed by preposition and verb or by an infinitive in web-scale n-gram data. At run-time, the system generates multiple possible trigrams by changing a word's lexical form and preposition in the original trigram. Example trigrams generated for "*difficulty to understand*" are shown in Figure 1. The system then ranks all of these generated sentences and use the highest ranking sentence as suggestion.

The rest of the paper is organized as follows. We review the related work in the next section. Then, we describe our method for automatically learning to translate a sentence that may contain preposition-verb serial errors into a grammatical sentence (Section 3). In our evaluation, we describe how to measure the precision and recall of producing grammatical sentences (Section 4) in an automatic evaluation (Section 5) over a set of marked sentences in a learner corpus.

2. Related Work

Grammatical Error Detection (GED) for language learners has been an area of active research. GED involves pinpointing some words in a given sentence as ungrammatical and offering correction if necessary. Common errors in learners' writing include misuse of articles, prepositions, noun number, and verb form. Recently, the state-of-the-art research on GED has been surveyed by Leacock *et al.* (2010). In our work, we address serial errors in English learners' writing which are simultaneously related to the preposition and verb form, an aspect that has not been dealt with in most GED research. We also consider the issues of broadening the training data for better coverage and coping with data sparseness when unseen events happen.

Although there are over a billion people estimated to be using or learning English as a second or foreign language, common English proofreading tools do not target specifically the most common errors made by second language learners. Many widely-used grammar checking tools are based on pattern matching and at least some linguistic analysis, based on hand-coded grammar rules (Leacock *et al.*, 2010). In the 1990s, data-driven, statistical methods began to emerge. Statistical systems have the advantage of being more intolerant of ill-form, interlanguage, and unknown words produced by the learners than the rule-based systems.

Knight and Chander (1994) proposed a method based on a decision tree classifier to correct article errors in the output of machine translation systems. Articles were selected based on contextual similarity to the same noun phrase in the training data. Atwell (1987) used a language model of a language to represent correct usage for that language. He used the language model to detect errors that tend to have a low language model score.

More recently, researchers have looked at grammatical errors related to the most common prepositions (9 to 34 prepositions, depending on the percentage of coverage). Eeg-Olofsson and Knuttson (2003) described a rule-based system to detect preposition errors for learners of Swedish. Based on part-of-speech tags assigned by a statistical trigram tagger, 31 rules were written for very specific preposition errors. Tetreault and Chodorow (2008), Gamon *et al.* (2008), and Gamon (2010) developed statistical classifiers for preposition error detection. De Felice and Pulman (2007) trained a voted perceptron classifier on features of grammatical relations and WordNet categories in an automatic parse of a sentence. Han *et al.* (2010) found that a preposition error detection model trained on correct and incorrect usage in a learner corpus works better than using well-formed text in a reference corpus.

In the research area of detecting verb form errors, Heidorn (2000) and Bender *et al.* (2004) proposed methods based on parse tree and error templates. Lee and Seneff (2008) focused on three cases of verb form errors: subject-verb agreement, auxiliary agreement, and verb complement. The first two types are isolated verb form errors, while the third type may involve serial errors related to preposition and verb. Izumi *et al.* (2003) proposed a maximum entropy model, using lexical and POS features, to recognize a variety of errors, including verb form errors. Lee and Seneff (2008) used a database of irregular parsing caused by verb form misuse to detect and correct verb form errors. In addition, they also used the Google n-gram corpus to filter out improbable detections. Both Izumi *et al.* (2003) and Lee and Seneff (2008) obtained a high error correction rate, but they did not report serial errors separately, making comparison with our approach is impossible.

In a study more closely related to our work, Alla Rozovskaya and Dan Roth (2013) introduced a joint learning scheme to jointly resolve pairs of interacting errors related to subject-verb and article-noun agreements. They showed that the overall error correction rate is improved by learning a model that jointly learns each of these interacting errors.

3. Method

Correcting serial errors (*e.g.*, "*I have difficulty to understand English.*") one error at a time in the traditional way may not work very well, but previous works typically have dealt with one type of error at a time. Unfortunately, it may be difficult to correct an error in the context of another error, because an error could only be corrected successfully within the correct context. Besides, such systems need to correct a sentence multiple times, which is time-consuming and more error-prone. To handle serial errors, a promising approach is to treat serial errors together as one single error.

3.1 Problem Statement

We focus on correcting serial errors in learners' writing using the context of trigrams in a sentence. We train a statistical machine translation model to correct learners' errors of the types of a content word followed by a preposition and a verb using web-scale n-grams.

Problem Statement: We are given a sentence $S = w_1, w_2, \ldots, w_n$, and web-scale n-gram, *webgram*. Our goal is to train two statistical machine translation model *TM* and back-off model TM_{bo} to correct learners' writing. At run-time, trigrams (w_i , w_{i+1}, w_{i+2}) in S (i =1, n-2) are matched and replaced using *TM* and the back-off model TM_{bo} to translate S into a correct sentence T.

In the rest of this section, we describe our solution to this problem. First, we describe the strategy to train *TM* (Section 3.2) and TM_{bo} (Section 3.3) using *webgrams*. Finally, we show how our system corrects a sentence at run-time using *TM*, TM_{bo}, and a language model *LM* (Section 3.4).

3.2 Generating *TM*

We attempt to identify trigrams that fit the pattern of serial errors and correction we are dealing with in *webngram*, and we group the selected trigrams by their content words and verb lemmas. Our learning process is shown in Figure 2. We assume that, within each group, the low frequency trigrams are probably errors that should be replaced by the most frequent trigram: a *one construction per collocation* constraint. For example, when expressing "*difficulty*" and "*to understand*," any NPV constructs with low frequency (*e.g.*, "*difficulty for understanding*" and "*difficulty about understanding*") are erroneous forms of the most frequent trigram "*difficulty in understanding*". Therefore, we generate *TM* with such phrase to phrase translations accordingly.

(1) Select *trigrams* related to serial errors and corrections from *webngram* (Section 3.2.1)

(2) Group the selected trigrams by the first and last word in the *trigrams* (Section 3.2.2)

(3) Generate a phrase table for the statistical machine translation models for each group (Section 3.2.3)

Figure 2. Outline of the process used to generate TM.

3.2.1 Select and Annotate Trigrams

We select four types of trigrams (t_1, t_2, t_3) from *webngram*, including noun-prep-verb (NPV), verb-prep-verb (VPV), adj-prep-verb (APV), and adverb-prep-verb (RPV). We then annotate the trigrams with types and lemmas of content words t_1 and t_3 (*e.g.*, "*accused of being* 230633" becomes "*VPV, accuse be, accused of being* 230633). Figure 3 shows some sample annotated trigrams.

VPV, *accuse be*, accused of being	230,600
VPV, *accuse kill*, accused of killing	83,100
VPV, *accuse have*, accused of having	78,500
VPV, *accuse use*, accuse of using	45,200
VPV, *accuse* murder, accused of murdering	40,032
VPV, *accuse be*, accused to be	10,200
VPV, *accuse prove*, accused to prove	3,600

Figure 3. Sample annotated trigrams

VPV, *accuse be*, accused of being	230,600
VPV, *accuse be*, accused to be	10,200
VPV, *accuse be*, accused of is	2,841
VPV, *accuse be*, accuse of being	2,837
VPV, *accuse be*, accused as being	929
VPV, *accuse be*, accused of was	676
VPV, *accuse be*, accused from being	535

Figure 4. Sample trigram group

accused to be	||| accused of being	||| 0.93
accused of is	||| accused of being	||| 0.93
accuse of being	||| accused of being	||| 0.93
accused as being	||| accused of being	||| 0.93
accuse of was	||| accused of being	||| 0.93
accused from being	||| accused of being	||| 0.93

Figure 5. Sample phrase translations for a trigram group

3.2.2 Group Trigrams

We then group the trigrams by types, the first words, and the verb lemmas. See Figure 4 for a sample VPV group of trigrams. This step should bring together the trigrams containing serial errors and their correction. Note that we assume certain serial errors will have a correction of the same length here, which is true in most cases.

3.2.3 Generate Rules

For each group of annotated trigrams, we then generate phrase and translation pairs with

probability as follows. Recall that we assume that the higher the count of the trigram, the more likely the trigram is to be correct. So, we generate "l_1, l_2, l_3 ||| h_1, h_2, h_3 ||| p ," where h_1, h_2, h_3 is the trigram with the highest frequency count; l_1, l_2, l_3 is one of the trigrams with lower frequency count; and p denotes the probability of l_1, l_2, l_3 translating into h_1, h_2, h_3. We define p=(highest frequency count)/(group frequency count).

3.3 Generating TM_{bo}

In addition to the surface-level translation model *TM*, we also build a back-off model as a way of coping with cases where the trigram (t_1, t_2, t_3) is unseen in *TM*. The idea is to assume the complement (t_2, t_3) of t_1 tends to be in a certain syntactic form regardless of the verb t_3, as dictionaries typically would describe the usage of "*accuse*" in terms of "*accuse somebody of doing something*." Our learning process for TM_{bo} is shown in Figure 9.

VPV, *accuse VERB*, accused of *VERB*-ing	230,600
VPV, *accuse VERB*, accused of *VERB*-ing	83,100
VPV, *accuse VERB*, accused of *VERB*-ing	78,500
VPV, *accuse VERB*, accuse of *VERB*-ing	45,200
VPV, *accuse VERB*, accused of *VERB*-ing	40,032
VPV, *accuse VERB*, accused to *VERB*	10,200
VPV, *accuse VERB*, accused to *VERB*	3,600

Figure 6. Sample annotated trigrams

VPV, *accuse VERB*, accused of *VERB-ing*	870,600
VPV, *accuse VERB*, accused to *VERB*	50,200
VPV, *accuse VERB*, accuse to *VERB*	20,200

Figure 7. Sample trigram group

accused to VERB ||| accused of VERB-ing ||| 0.47
accused of VERB ||| accused of VERB-ing ||| 0.47

Figure 8. Sample back-off translations

(1) Select trigrams with specific forms from Web 1T n-gram
(2) Reform trigrams W3 to W3's lexical
(3) Group the selected trigrams using the first word
(4) Group the selected trigrams using the first word

Figure 9. Outline of the process used to generate TM_{bo}

3.3.1 Generalize Trigrams

First, we generalize the annotated trigrams (see Section 3.2.1) by replacing the verb form with its part of speech designator (*i.e.*, replace “accuse” with VERB, and replace “accusing” with VERB-ing).

3.3.2 Sum Counts

In this step, we group the identically transformed trigrams and sum up the frequency counts. See Figure 6 for sample results.

3.3.3 Group Trigrams of the Same Context

We then group the trigrams by type and by the first word (context). See Figure 7 for a sample “accuse P V” group of trigrams.

3.3.4 Generate Rules

For each group of generalized trigrams, we then generate the phrase and translation pair with the probability as described in Section 3.2.3. See Figure 8 for a sample of back-off translations.

3.4 Run-time Correction

If one loads *TM* and TM_{bo} into memory before the decoding process (generating, ranking, and selecting translations), that would take up a lot of memory and slow the process of matching phrases to find translations. Therefore, we generate phrase translations on the fly for the given sentence before decoding. Our process of decoding to correct grammatical errors is shown in Figure 10.

(1) Tag the input sentence with part of speech information in order to find trigrams that fit the type of serial errors

(2) Search *TM* and generate translations for the input phrases

(3) Search TM_{bo} and generate translations for the input phrases

(4) Run statistical machine translation

Figure 10. Outline of the process used to correct the sentence at run-time

3.4.1 Tag the Input Ssentence

We use a POS tagger to tag the input sentence, and we identify trigrams (t_1, t_2, t_3) consisting of a content word followed by a preposition and verb (belonging to the NPV, VPV, APV, or RPV types we described in Section 3.2.1).

3.4.2 Search TM and Generate Translation Rules

We then search for the group of trigrams (indexed by POS type and t_1, t_3) in *TM* containing the trigrams (t_1, t_2, t_3), found in Step 3.4.1. We find the trigram (h_1, h_2, h_3) with the highest count in that group. With that, we can dynamically add the translation, "t_1, t_2, t_3 ||| h_1, h_2, h_3 ||| 1.0" to the cache of *TM* in memory (*e.g.*, "difficulty to understand ||| difficulty in understanding ||| 1.0") to speed up the subsequent decoding process.

3.4.3 Search TM_{bo} and Generate Translation Rules

Just like in 3.4.2, we use t_1 and its part of speech p_1 to search TM_{bo} for the generalized trigram group that matches (t_1, t_2, t_3). We then find the most frequent generalized trigram (h_1, h_2, h_3) in that group. After that, we need to specialize (h_1, h_2, h_3) for t_3 by replacing h_3 with the verb form of t_3 for the designator h_3, resulting in (h_1, h_2, h'_3). Consider the generalized trigram "accused of VERB-ing" and t_3 = "murder," the specialized trigram would be "accused of murdering." Finally, we add "t_1, t_2, t_3 ||| h_1, h_2, h'_3 ||| 1.0" (*e.g.*, "accused to murder ||| accused of murdering ||| 1.0") to the cache of *TM* in memory for the same purpose of speeding up decoding.

3.4.4 Decode the Input Sentence without Reordering

Finally, we run a monotone decoder with the cache *TM* and a language model *LM*. By default, any word not in *TM* will be translated into itself.

4. Experimental Setting

Our system *DeeD* (Don'ts-to-Do's English-English Decoder) was designed to correct preposition-verb serial errors in a given sentence written by language learners. Nevertheless, since large-scale learner corpora annotated with errors are not widely available, we have resorted to Web scale n-grams to train our system, while using a small annotated learner corpus to evaluate its performance. In this section, we first present the details of training *DeeD* for the evaluation (Section 4.1). Then, Section 4.2 lists the grammar checking systems that we used in our evaluation and comparison. Section 4.3 introduces the evaluation metrics for the performance of the systems, and details of the sentences evaluated and performance judgments are reported in Section 4.4.

4.1 Training DeeD

We used the Web 1T 5-grams (Brants & Franz, 2006) to train our systems. Web 1T 5-grams is a collection that contains 1 to 5 grams calculated from a 1 trillion words of public Web pages provided by Google through the Linguistic Data Consortium (LDC). There are some ten

million unigrams, 3 hundred million bigrams, and around 1 billion trigrams to fivegrams. We obtained 104,537,560 trigrams, containing only words in the General Service List (West, 1954) and Academic Word List (Coxhead, 1999). These trigrams were further reduced to 4,486,615 entries that fit the patterns of four types of serial errors and corrections: an adjective, noun, verb, or adverb followed by a preposition (or infinitive *to*) and a verb.

To determine the part of speech of words in the n-gram, we used the most frequent tag of a given word in BNC to tag words in the trigram.

4.2 Grammar Checking Systems Compared

Once we have trained *DeeD* as described in Section 3, we evaluated its performance using two datasets. The first dataset contained sentences written by an ESL or EFL learner with the serial errors with corrections. The second dataset contained mostly correct sentences in British National Corpus (BNC) with mostly published works written by native, expert speakers.

The first testset is a subset of the Cambridge Learner Corpus, the CLC First Certificate Exam Dataset (CLC/FCE). This dataset contains 1,244 exam essays written by students who took the Cambridge ESOL First Certificate in English (FCE) examination in 2000 and 2001. For each exam script, the CLC/FCE Dataset includes the original text annotated with error, type, and correction. From the 34,893 sentences in the 1,244 exam essays, we extracted 118 sentences that contained the serial errors in question. Other types of errors were replaced with corrections in these sentences.

The second testset is a random sample of 1000 sentences containing trigrams that fit the error patterns also used to evaluate our system. The four system and testset combinations evaluated are:

—Learner corpus without back-off model (**LRN**): The proposed system using only the surface-level translation model was tested on the first testset obtained from a learner corpus.

—Learner corpus with back-off model (**LRN-BO**): The proposed system with the additional back-off model was tested on the first testset obtained from a learner corpus.

—BNC without back-off model (**BNC**): The proposed system using only the surface-level translation model was tested on the first testset obtained from the British National Corpus.

—BNC with back-off model (**BNC-BO**): The proposed system without the back-off model was tested on the first testset obtained from the British National Corpus.

4.3 Evaluation Metrics

English correction systems usually are compared based on the quality and completeness of correction suggestions. We measured the quality using the metrics of precision, recall, and error rate. For the first testset, we measured precision and recall rates while, for the second

testset, we measured the error rate (false alarms). We define precision and recall as:

$$\text{Precision} = C/S \quad (1)$$

$$\text{Recall} = C/N \quad (2)$$

where N is the number of serial errors, S is the number of corrections our system found, and C is the number of corrections where our system was correct. We also computed the corresponding F-score. Error rate was used in the second dataset described above, and we define the error rate as follows:

$$\text{Error Rate} = E/T \quad (3)$$

where E is the number of corrections our system found (which are all wrong, since we were testing sentences with no errors) and T is the number of sentences tested.

5. Evaluation Results

In this section, we report the results of the evaluation using the dataset and environment mentioned in the previous section. During this evaluation, 118 sentences with serial errors were used to evaluate the two systems: **LRN** and **LRN-BO**. Table 1 shows the average precision, recall, and F-score of **LRN** and **LRN-BO**. As we can see, **LRN** performs better in precision, which is reasonable since the back-off model corrects errors without the information of the verb involved. **LRN-BO** performs better in recall because the back-off model applies when the original model does not cover the case. Overall, **LRN-BO** performs better in F-score.

Table 1. Average precision, recall, and F-score of LRN and LRN-BO

	F-Score	Precision	Recall
LRN	0.43	0.71	0.31
LRN-BO	0.45	0.68	0.33

Table 2. Average error rate of BNC and BNC-BO

	Error Rate
BNC	0.10
BNC-BO	0.13

During this evaluation, 1000 sentences in BNC that fit the pattern of serial errors but in fact do not contain errors, were used to evaluate the same two systems: **BNC** and **BNC-BO**. Table 2 shows the average error rate of BNC and **BNC-BO**. It is not surprising that **BNC** performs better than **BNC-BO**, since **BNC** always makes fewer corrections than **BNC-BO**. Nevertheless, **BNC-BO** is only slightly worse than **BNC**.

6. Conclusions

Many avenues exist for future research and improvement of our system. For example, spell checking can be done before correcting grammatical errors. Context used to "translate" the serial errors can be enlarged from one word to two or more words (immediately or closely) preceding the errors. We can also add one more level of backing off for the context word preceding the serial errors: from surface word to lemma or from a proper name to named entity type (PERSON, PLACE, ORGANIZATION). We also can improve the accuracy of part of speech tagging used in applying the back-off model.

Additionally, an interesting direction to explore is extending this approach to handle other types of isolated and serial errors commonly found in learners' writing. Yet another direction of research would be to consider corrections resulting in more or fewer words (*e.g.*, one less word as in **spend time for work* vs. *spend time working*). Or, we could also combine n-gram statistics from different types of corpora: a Web-scale corpus, a reference corpus, and a learner corpus. For example, the translation probability can be determined via statistical classifier training on the learner corpus with features extracted from n-grams of multiple corpora.

In summary, we have introduced a new method for correcting serial errors in a given sentence in learners' writing. In our approach, a statistical machine translation model is generated to attempt to translate the given sentence into a grammatical sentence. The method involves automatically learning two translation models based on Web-scale n-grams. The first model translates trigrams containing serial preposition-verb errors into correct ones. The second model is a back-off model for the first model, used in the case where the trigram is not found in the training data. At run-time, the phrases in the input are matched using the translation model and are translated before ranking is performed on all possible translation sentences generated. Evaluation on a set of sentences in a learner corpus shows that the method corrects serial errors reasonably well. Our methodology exploits the state of the art in machine translation, resulting in an effective system that can deal with serial errors at the same time.

References

Atwell, E. S. (1987). How to detect grammatical errors in a text without parsing it. In *Proceedings of the Third Conference of the European Association for Computational Linguistics (EACL)*, 38-45, Copenhagen.

Bender, E. M., Flickinger, D., Oepen, S., & Baldwin, T. (2004). Arboretum: Using a precision grammar for grammar checking in CALL. In *Proceedings of the Integrating Speech Tech- nology in Learning/Intelligent Computer Assisted Language Learning*

(inSTIL/ICALL) Symposium: NLP and Speech Technologies in Advanced Language Learning Systems, Venice.

Brants, T., & Franz, A. (2006). *The Google Web 1T 5-gram corpus version 1.1*. LDC2006T13.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, *34*(2), 213-238.

De Felice, R., & Pulman, S. G. (2009). Automatic detection of preposition errors in learner writing. *CALICO Journal*, *26*(3), 512-528.

Eeg-Olofsson, E., & Knuttson, O. (2003). Automatic grammar checking for second language learners - the use of prepositions. In *Proceedings of the 14th Nordic Conference in Computational Linguistics* (NoDaLiDa).

Gamon, M. (2010). Using mostly native data to correct errors in learners' writing. In *Proceedings of the Eleventh Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Los Angeles.

Gamon, M., Gao, J., Brockett, C., Klementiev, A., Dolan, W. B., Be-lenko, D., & Vanderwende, L. (2008). Using contextual speller techniques and language modeling for ESL error correction. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 449-456, Hyderabad, India.

Graddol, D. (2006). *English next: Why global English may mean the end of 'English as a Foreign Language*.' UK: British Council.

Han, N.-R., Tetreault, J., Lee, S.-H., & Ha, J.-Y. (2010). Using error-annotated ESL data to develop an ESL error correction system. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, Malta.

Heidorn, G. E. (2000). Intelligent writing assistance. In R. Dale, H. Moisl, and H. Somers, editors, *Handbook of Natural Language Processing*, 181-207. Marcel Dekker, New York.

Izumi, E., Uchimoto, K., Saiga, T., Supnithi, T., & Isahara, H. (2003). Automatic error detection in the Japanese learners' English spoken data. In *Companion Volume to the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, 145-148.

Knight, K., & Chander, I. (1994). Automated postediting of documents. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI)*, 779-784, Seattle.

Leacock, C. *et al*. 2010. Automated grammatical error detection for language learners. *Synthesis Lectures on Human Language Technologies*, *3*(1), 1-134.

Lee, J., & Seneff, S. (2006). Automatic grammar correction for second-language learners. In *Proceedings of the Ninth International Conference on Spoken Language Processing (Interspeech)*, 1978-1981.

Lee, J., Tetreault, J., & Chodorow, M. (2009b). Human evaluation of article and noun number usage: Influences of context and construction variability. In P*roceedings of the Third Linguistic Annotation Workshop (LAW)*, 60-63, Suntec, Singapore.

Rozovskaya, A., & Roth, D. (2013). Joint Learning and Inference for Grammatical Error Correction, In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 791-802.

West, M. (1953). *A General Service List of English Words*. London: Longman, 1953.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts, In *Proceedings of the Annual Meeting of the Association for Computational Linguistic*.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 45-62

A Semantic-Based Approach to Noun-Noun Compound Interpretation

You-shan Chung* and Keh-Jiann Chen*

Abstract

In this project, we have studied Chinese noun-noun compounds (NNCs) and have found that N1 and N2 are linked either by semantic roles assigned by events (complex relations) or by static relations (simple relations), including meronymy, conjunction, and the host-attribute-value relation. Using data from the FrameNet and E-HowNet, we have found that, for NNCs of either type, the major semantic relations between the two components are limited enough to allow computational implementation. Regarding simple relations, most conjunction pairs have been listed in E-HowNet,and so are host-attribute-value sets. The E-HowNet Taxonomy also makes identification of meronymy possible. As for NNCs involving complex relations, each component's semantic role, along with the events that assign these roles, can be restored through mappings to corresponding frame elements (FEs) in entity and to event frames and lexical units (LUs) in FrameNet's frames, respectively, that represent the concept the NNC conveys.

Keywords: Noun-noun Compounds, Automatic Interpretation, Extended HowNet (E-HowNet), FrameNet

1. Introduction

Noun-noun compounds (henceforth NNC) are compounds composed of two nouns. For example:

麵包刀 *mianbao-dao* 'bread knife'
衛星城市 *weixin-chengshi* 'satellite city'
金融股 *jinrong-gu* 'stocks in the financial sector'
秋蟹 *qiu-xie* 'autumn crab'

* Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-mail: {yschung, kchen}@iis.sinica.edu.tw

腳踏車輪胎 *jiao-ta-che luntai* ‘bicycle tire’
卵石地板 *luan-shi diban* ‘pebble floor’
鐘錶 *zhong-biao* ‘clock and watch’
鐵桌 *tie-zhuo* ‘iron table/desk’
車速 *che-su* ‘car speed’

While the part-of-speech (POS) of NNCs usually is nominal, their interpretations seem so diverse that some researchers even contend that they are completely determined by context (*e.g.* Dowty, 1979; reviewed in Copestake & Lascarides, 1997).

Nevertheless, the majority of researchers believe that there is at least some degree of regularity in NNC interpretation. This regularity is often reported to be at least partially universal as well (Levi, 1978; Søgaard, 2005). There are three popular theories along these lines, which are not mutually exclusive. First, there is a limited set of semantic relations between the two component nouns, N1 and N2 (Levi, 1978; as well as computational works that implemented her theory, *e.g.* Copestake & Lascarides, 1997; Søgaard, 2005; Copestake & Briscoe, 2005; Huang, 2008). Second, N1 and N2 are the arguments of an event that bridges them and by which they are assigned semantic roles (Levi, 1978; Leonard, 1984; Ryder, 1994). Third, the two component nouns sometimes are linked through similarity in some aspect, resulting in metaphorical readings.

Nevertheless, these accounts generally have the following four problems. First, the semantic relations they proposed or adopted tend to be not specific enough. Levi (1978), for instance, proposed nine semantic relations between N1 and N2, which she called Recoverably Deleted Predicates (RDP), including CAUSE, HAVE, MAKE, USE, BE, IN, FOR, FROM, and ABOUT. These RDPs, however, appear to be too general to be informative, especially with prepositional ones like IN and FOR, as NNCs linked by the same preposition belong to the same semantic categories only in a very broad sense.

Second, some of the studies resolve only limited or sporadic semantic categories, while others are questionable in terms of their correct prediction rate. For example, the fourteen semantic relations Li and Thompson (1981) proposed do not seem to make up a meaningful and discrete inventory of semantic relations, while Huang’s (2008) combinational patterns for three major categories of physical objects (*i.e.* animals, plants, and artifacts) are each based on the analysis of only six morphemes, raising concerns about generality.

The third problem is that the classifying criteria mostly are left unaccounted for; thus, they appear arbitrary. For example, Levi (1978) sees the two components of *lemon peel* and *apple seed* as linked by the predicates HAVE and FROM, respectively, but such a distinction

between the two NNCs may not be without controversy.

The last problem is that bridging does not seem to be eventive or by prepositions in the following three situations: first, the host-attribute-value relation (*e.g.* 鐵桌 *tie-zhuo* 'iron table/desk,' 車速 *che-su* 'car speed') with two special subclasses, where N1 denotes time (*e.g.* 秋蟹 *qiu-xie* 'autumn crab') or N1 denotes space (*e.g.* 倫敦地鐵 *Lundun-ditie* 'London Underground'); second, meronymy, or part-whole relation (part-whole: *e.g.* 雙底船 *shuang-di chuan* 'double-bottom,'; whole-part: *e.g.* 腳踏車輪胎 *jiao-ta-che luntai* 'bicycle tire'); and third, conjunction (*e.g.* 鐘錶 *zhong-biao* 'clock and watch,' 禮樂 *li-yue* 'manners and music') .

Before we go on, we need to explain the definition of Chinese NNCs adopted in this study. Unlike in English, formal similarity in Chinese does not entail a shared POS. For example, the first component in 希臘國歌 *xila guo-ge* 'the national anthem of Greece,' 希臘菜 *xila-cai* 'Greek dish,' and 月費 *yue-fei* 'monthly fee' corresponds to adjective forms in their English equivalents. Nevertheless, we include these various forms in our analysis since such formal differences do not reflect conceptual differences, as Levi (1978) has argued for this at length and also included adjectives in her analysis of what she called "complex nominal," or "NNCs" in our terms.

Addressing the aforementioned four problems, we used a knowledge base that we believe could help decide the precise semantic relations for both event-linked and non-event-linked NNCs, which is FrameNet (https://framenet.icsi.berkeley.edu/fndrupal/). In essence, the theory behind FrameNet is that lexical units (LU) evoke concepts represented by "frames," which are each composed of a set of frame elements (FE), *i.e.* the overtly-realized semantic roles assigned by the frame's LUs. Some LUs evoke entity concepts, while others evoke eventive ones. Since many entities in FrameNet have frames, we think it might be possible to map more NNC-productive N2s in our database, along with the NNCs they derive, to corresponding entity frames in FrameNet.

We have two research questions. First, with a corpus and FrameNet, we investigate whether there are only limited bridging verbs and semantic relations between the two component nouns of a NNC. Second, are there semantic relations between N1 and N2 that do not involve bridging events?

2. Complex Relations

As mentioned in the Introduction, many researchers hold that an NNC's component nouns are the arguments of an event that bridges them and by which they are assigned semantic roles. Levi (1978) regards all N1s and N2s as subjects and objects of nine linking predicates, with

one component entity doing something to the other. Below are her examples[1] and their Chinese equivalents:

CAUSE: *e.g.* malarial mosquitoes (瘧蚊 *nue-wen*)

HAVE: *e.g.* picture book (圖畫書 *tuhua-shu*), apple cake (蘋果蛋糕 *pingguo dangao*), gunboat (砲艇 *pao-ting*), industrial area (工業區 *gongye-qu*), imperial bearing (貴族氣質 *guizu-qizhi*)

MAKE: *e.g.* honeybee (蜜蜂 *mi-fong*), daisy chains (雛菊鍊 *chuju-lian*)

USE: *e.g.* steam iron (蒸氣電熨斗 *zhengqi dian-yundou*), solar generator (太陽能發電機 *taiyang-neng fadian-ji*)

BE: *e.g.* target structure (目標結構 *mubiao-jiegou*), ceiling price (天價 *tian-jia*), queen bee (女王蜂 *nu-wang fong*), satellite nation (衛星國家 *weixing-guojia*), phantom limb (幽靈肢 *youling-zhi*)

IN: *e.g.* field mouse (田鼠 *tian-shu*), autumnal rains (秋雨 *qiu-yu*)

FOR: *e.g.* horse doctor (馬醫 *ma-yi*), arms budget (武器預算 *wuqi-yusuan*), nasal mist (鼻腔噴霧 *bi-qiang pen-wu*)

FROM: *e.g.* olive oil (橄欖油 *ganlan-you*), test-tube baby (試管嬰兒 *shi-guan yinger*), apple seed (蘋果籽 *pingguo-zi*), rural visitors (鄉間訪客 *xiangjian fang-ke*)

ABOUT: *e.g.* tax law (稅法 *shui-fa*), criminal policy (刑事政策 *xingshi-zhengce*)

Levi says NNCs are all linked by one of the nine predicates, with the two components being their arguments; however, we believe that some NNCs simply involve more static relations and some relations are not covered by the above nine predicates. One instance that involves a missing static relation is, for example, the highly-productive shape relation, *e.g.* dragon boat (龍舟 *long-zhou*). In the following sections, we will use evidence of both language instinct and FrameNet data to support the distinction between simple and complex relations.

3. Motivating Simple Relations

Besides event-bridging relations, we propose simple relations, where N1 and N2 are not interacting participants of an event. Despite their shared syntactic and semantic properties, instances of simple relations have not been recognized as a distinct category, as observed by Liu (2008) and by Chung and Chen (2010).

[1] Only NNCs within the scope of this paper are listed.

We identified three types of simple relations, as opposed to complex ones:

(1) N1 and N2 denote two of the three elements of a host-attribute-value set

(a) Temporal N1

N1 denotes time:

e.g. 晨霧 *chen-wu* 'morning mist' (value+host), 秋蟹 *qiu-xie* 'autumn crab' (value+host), 午夜列車 *wuyie-lieche* 'midnight train' (value+host)

N1 denotes frequency:

e.g. 月費 *yue-fei* 'monthly fee' (value+host)

(b) Locational N1

e.g. 希臘菜 *xila-cai* 'Greek dish' (value+host), 倫敦地鐵 *Lundun-ditie* 'London Underground' (value+host), 台北人 *Taibei-ren* 'Taipei people' (value+host)

(c) Others

e.g. 鐵桌 *tie-zhuo* 'iron table/desk' (value+host), 法式 *fa-shi* 'French-style' (value+attribute), 電價 *dian-jia* 'electricity price' (host+attribute), 金塊 *jin-kuai* 'gold bricks' (host+value), 衣服堆 *yifu-dui* 'heap of clothes' (host+value), 車速 *che-su* 'car speed' (host+attribute)

(2) Meronymy (*i.e.* part-whole relation)

N1 denotes part; N2 denotes whole:

e.g. 雙底船 *shuang-di chuan* 'double-bottom,'

N1 denotes whole; N2 denotes part:

e.g. 腳踏車輪胎 *jiao-ta-che luntai* 'bicycle tire,' 腸道 *chang-dao* 'intestine canal'

(3) Conjunction

e.g. 手腳 *shou-jiao* 'hands and feet,' 鐘錶 *zhong-biao* 'clock and watch,' 警民 *jing-min* 'the police and the people'

In (1a), the N1 usually denotes the value of the semantic role "time" of an event related to the N2. In 午夜列車 *wuyie-lieche* 'midnight train' and 秋蟹 *qiu-xie* 'autumn crab,' the temporal values are 午夜 *wuyie* 'midnight' and 秋 *qiu* 'autumn,' respectively. The two NNCs either can be elaborated to mean 'trains that travel at midnight' and 'crabs that reach maturity in autumn,' or can be simply put as 'trains at midnight' and 'crabs in autumn,' omitting the events. In (1b), locational N1s usually denote place names. (1a) and (1b) are similar in that understanding of the NNCs does not depend on figuring out the bridging events

that decide the semantic roles of the component nouns.

It should be noted, however, that the nature of the event that takes place in the time or space denoted by N1 can be less than straightforward. Sometimes, this indeterminacy is caused by the meaning shift of individual components. Take 秋葵 *qiu-kui* 'okra' for example. Even native speakers may have no idea what happens to the N2 '葵' *kui* in autumn (*i.e.* 秋 *qiu* 'autumn'). This is because 葵 *kui* may not be as familiar a vegetable to modern people as it was when the compound was coined. Sometimes, meaning extension allows multiple readings of a word. For example, in antiquity, when international travel was essentially impossible, 希臘人 *xila-ren* 'Greeks' usually lived and stayed in Greece, but nowadays 希臘人 *xila-ren* 'Greeks' and 希臘菜 *xila-cai* 'Greek dishes' can reach far beyond the national borders.

Nevertheless, while the bridging event can be obscure or diverse, NNCs with temporal or locational N1s share one common characteristic: Some bridging event(s) exists, but it does not have to be clearly identified to enable sufficient understanding.

Finally, (1c) consists of host-attribute-value relations other than time and space. As argued by Chung and Chen (2010) in line with Liu (2008), objects and events are characterized by the attributes they have, and attributes are characterized in turn by values. For the examples in (1c), the morphemes 式 *shi* 'style,' 價 *jia* 'price,' and 速 *su* 'speed' are attributes and 鐵 *tie* 'iron' and 法 *fa* 'French' are attribute-values of material and style, respectively. In other words, both objects and events (collectively called "hosts") generally are associated with some attributes and attributes are associated with values. For example, artifacts, which are a subclass of objects, have the attribute "material," and "iron" is a kind (value) of material.

Given that N1 usually specifies N2, it is natural for value and host, value and attribute, and host and attribute to form NNCs in order to modify the host and attribute or to name the relevant host of an attribute.

As for (2), in 雙底船 *shuang-di chuan* 'double-bottom' and 腳踏車輪胎 *jiao-ta-che luntai* 'bicycle tire,' N1 and N2 are not interacting participants of an event. Likewise, in (3), N1 and N2 assume parallel roles in situations like 手腳看起來很乾淨 *Shou-jiao kan-qilai hen ganjing* 'Hands and feet look tidy,' 修理鐘錶 *xiuli zhong-biao* 'repair a clock (watch),' 警民合作打擊犯罪 *Jing-min hezuo daji fanzui* 'The police and the people join hands to fight crime.'

4. Mapping NNCs to FrameNet's Frames

We chose NNC-productive N2s (*i.e.* those that form NNCs with various types of N1s) from our Prefix-Suffix Database (http://140.109.19.103/affix/), sorted them according to their semantic categories and the situations their derived NNCs described, and matched these situations with FrameNet's frames.

To the extent that frames represent concepts, to map NNCs to frames is to identify the concepts NNCs convey. The corporal data to date have indicated that N2s of nine semantic categories are NNC-productive. They are: "people," "people of different vocations," "food," "clothing," "container," "vehicle," "wealth," "text," and "road." We have listed the most common relations between N1 and N2 for each category at the appendix. These categories each can be mapped to one or more entity frames, where the N2 is represented by an FE that usually has the same name as the frame itself and the N1 by another FE of the frame. Below are some examples of such mappings. (Frame names have all capital letters, while FEs have only the initial letters as capital letters.)

Simple relation (subclass: host-attribute-value)

FOOD

N1-N2=Material-Food

e.g. 玉米餅 *yumi-bing* 'corn cake,' 綠豆糕 *lu-dou gao* 'green beans cake,' 牛肉湯 *niu-rou tang* 'beef soup,' 奶茶 *nai-cha* 'milk tea,' 蘋果汁 *pingguo-zhi* 'apple juice,' 花生醬 *huasheng-jiang* 'peanut butter'

CLOTHING

N1-N2=Material-Clothing

e.g. 草鞋 *cao-xie* 'straw shoes,' 木鞋 *mu-xie* 'wooden shoes,' 皮鞋 *pi-xie* 'leather shoes,' 膠鞋 *jiao-xie* 'plastic shoes,' 豹皮帽 *bao-pi mao* 'leopard-skin hat,' 毛衣 *mao-yi* 'sweater,' 布衫 *bu-shan* 'cotton shirt'

Simple relation (subclass: meronymy)

VEHICLE_SUBPARTS

N1-N2=Part-Whole

e.g. 雙底船 *shuang-di chuan* 'double-bottom,' 鐵殼船 *tie-ke chuan* 'iron ship'

BUILDING_SUBPARTS

N1-N2=Whole-Building_part

e.g. 院牆 *yuan-qiang* 'yard wall,' 屋簷 *wu-yian* 'roof'

Complex relation

PEOPLE_BY_VOCATION

N1-N2=Person-Type

e.g. 弓箭手 *gong-jian-shou* 'archer,' 樂師 *yue-shi* 'musician,' 水電工 *shui-dian-gong* 'utilities technician'

MONEY

N1-N2=Buyer-Money

e.g. 家長費 *jiazhang-fei* 'parental fee'

N1-N2=Goods-Money

書款 *shu-kuan* 'money for buying books,' 田租 *tian-zu* 'land rent'

The above mappings show that NNCs that involve simple (as well as complex) relations correspond to FE pairs in FrameNet's entity frames. Take 玉米餅 *yumi-bing* 'corn cake' for example. The NNC can be mapped to FOOD, with the N2 餅 *bing* 'cake' denoting the FE "Food" and the N1 玉米 *yumi* 'corn' denoting "Material," which is another FE of the frame.

For NNCs of complex relations, besides an entity frame, the N1 usually can be mapped to another event frame, a point we will return to in Section 6.

5. Results

We have two findings attested to by the behavioral patterns of the nine semantic categories of N2s and their derived NNCs. First, NNCs generated by N2s of the same semantic category mostly correspond to one or a few conceptually-related frames. Second, some of the relations mapped are simple and some are complex, with N2 categories varying in their tendencies to denote simple and complex semantic relations.

5.1 Mapped to Entity Frames, Bridged by a Few Events, and Involving Limited Semantic Relations

We noticed that, when N1 and N2 are bridged by events, they usually can be mapped to both an entity frame and one or more event frames. We also found that common bridging events that link N1s to a N2 for each semantic category of N2 are limited.

For example, some of the NNCs the N2 category "money" derives include 中資 *zhong-zi* 'China capital,' 車款 *che-kuan* 'money for buying a car,' and 所費 *suo-fei* 'institute fund,' which we identified to belong to the entity frame "MONEY," where the N1s in the above three examples can be mapped to the FE "Use" and the N2 to "Money." Meanwhile, we found these N1s labeled as FEs in at least two event frames, which are "COMMERCE_BUY" and

"COMMERCE_SELL, where the three N1s 中 *zhong* 'China,' 車 *che* 'car,' and 所 *suo* 'institute,' correspond to the FEs Buyer, Goods, and Seller, respectively, are all core FEs of the event frames. Since the range of LUs and FEs for each frame usually is limited, the range of possible interpretations is more or less restricted for each NNC.

Below are all of the LUs and some of the FEs of these two event frames. (Not all non-core FEs are listed.)

COMMERCE_BUY

LUs: *buy.v, purchase_(act).n, purchase.v*

Core FEs: Buyer, Goods, Seller

Non-core FEs (not exhaustively listed): Manner, Means, Money, Purpose, Purpose_of_Goods, etc.

COMMERCE_SELL

LUs: *auction.n, auction.v, retail.v, retailer.n, sale.n, sell.v, vend.v, vendor.n*

Core FEs: Buyer, Goods, Seller

Non-core FEs (not exhaustively listed): Manner, Means, Money, Rate, Unit, etc.

5.2 N2 Categories Vary in Tendencies to Involve Simple and Complex Relations

Most of the N2 categories we have analyzed so far have produced both simple and complex-type NNCs. Below are two entity frames, CLOTHING and VEHICLE, which correspond to the N2 categories "clothing" and "vehicle". Each frame has at least one simple and one complex relation, which differ in frequency. The simple ones are labeled with their subclasses (and FEs[2]); the complex ones are labeled with the relevant FEs, which refer to the FEs that occur most or second-most often. (The frame names have all capital letters, while FEs only have initial capital letters.)

[2] FrameNet sometimes has FEs that we consider the simple type as well.

CLOTHING (衣 *yi* 'clothes,' 服 *fu* 'clothes,' 裝 *zhuang* 'clothes,' 帽 *mao* 'hat,' 鞋 *xie* 'shoes,' etc.)

Most frequent semantic relation: simple_host-attribute-value

The relevant FE(s) of N1

- As realized in CLOTHING (entity frame): Material
- (Not realized in event frames)

e.g. 草鞋 *cao-xie* 'straw shoes,' 木鞋 *mu-xie* 'wooden shoes,' 皮鞋 *pi-xie* 'leather shoes,' 膠鞋 *jiao-xie* 'plastic shoes,' 豹皮帽 *bao-pi-mao* 'leopard-skin hat,' 毛衣 *mao-yi* 'sweater'

Second-most frequent semantic relation: complex (*i.e.* eventive)

The relevant FE(s) of N1

- As realized in CLOTHING (entity frame): Wearer
- As realized in WEARING (event frame): Wearer

e.g. 女鞋 *nu-xie* 'women's shoes,' 僧鞋 *seng-xie* 'monk's shoes,' 法衣 *fa-yi* 'judge's robe,' 官服 *guan-fu* 'official robe,' 童裝 *tong-zhuang* 'children's clothes,' 學士服 *xue-shi-fu* 'Bachelor's gown'

VEHICLE (車 *che* 'vehicle,' 船 *chuan* 'ship,' etc.)

Most frequent semantic relation: complex (*i.e.* eventive)

The relevant FE(s) of N1

- As realized in VEHICLE (entity frame): Use
- As realized in BRINGING (event frame): Theme

e.g. 娃娃車 *wa-wa-che* 'kindergarten school bus,' 砂石車 *sha-shi che* 'gravel truck,' 客船 *ke-chuan* 'passenger ship,' 貨船 *huo-chuan* 'cargo ship'

Second-most frequent semantic relation: (complex, simple[3])

Complex

The relevant FE(s) of N1

- As realized in VEHICLE (entity frame): Means-of-propulsion
- (Not realized in event frames)

e.g. 電車 *dian-che* 'trolley bus,' 人力車 *ren-li-che* 'rickshaw'

Simple_meronymy

[3] For VEHICLE, the complex and simple relations have about the same second-highest frequencies.

The relevant FE(s) of N1

- As realized by VEHICLE (entity frame): Part
- (Not realized in event frames)

e.g. 雙底船 *shuang-di-chuan* 'double-bottom,' 鐵殼船 *tie-ke-chuan* 'iron ship'

5.3 The Coverage of the Identified Semantic Relations

As shown in Table 1, the average coverage of the semantic relations that FrameNet and E-HowNet have is 94.2% for the 1,153 compositional NNCs in the Prefix-Suffix Database. Below is the individual coverage of each N2 category.

Table 1. The average coverage of the semantic relations for the nine semantic categories of N2

Category	**Coverage**
Road	40/40 (100%)
Text	121/121 (100%)
People	241/243 (99.2%)
People of Different Vocations	46/48 (95.8%)
Wealth	72/72 (100%)
Container	411/427 (96.3%)
Food	60/86 (69.8%)
Clothing	42/47(89.3%)
Vehicle	53/69 (76.8%)
Mean	**1086/1153 (94.2%)**

Table 2 shows the average coverage of the three and five most frequent semantic relations. For the mapped percentage of each fine-grained relation for the nine categories, please refer to Appendix B.

We found that the top three most frequent semantic relations account for about eighty percent of the NNC instances. Meanwhile, the five most frequent relations on average have about 8% better coverage than the top three.

Table 2. The average coverage of the three and five most frequent semantic relations

Category \ Coverage	Top3	Top5
Road	67.5%	92.5%
Text	100%	100%
People	86.8%	94.2%
People of Different Vocations	83.3%	89.5%
Wealth	100%	100%
Container	94.7%	96.3%
Food	69.8%	69.8%
Clothing	72.2%	89.2%
Vehicle	49.2%	65.1%
Mean	**80.4%**	**88.5%**

Nevertheless, we noticed individual differences among N2's categories, with "food" and "vehicle" having a much lower coverage than others. Also, although we considered compositional NNCs only, there are still some relations that we lack labels for in FrameNet and E-HowNet. Some of these instances include metaphors, *e.g.* 野雞車 *yie-ji che* 'unlicensed car,' 霸王車 *bawang-che* 'unpaid ride'; apposition, *e.g.* 酒吧車 *jiuba-che* 'bar van,' 袍服 *pao-fu*, 'robe,' 靶船 *ba-chuan* 'target ship'; and those whose N1 indicates a general "use" relation unlike the other fine-grained mappings, *e.g.* 商輪 *shang-lun* 'merchant vessel,' 交通船 *jiaotung-chuan* 'commuter ship.'

6. Discussion

In this section, we will relate the two findings to our two research questions.

First, are there only limited bridging verbs and semantic relations between the two component nouns?

In the nine categories we investigated, the NNCs' bridging verbs, as well as the possible semantic roles that N1s and N2s take, are very limited, with an average coverage of over ninety percent. Even the least covered category "food" has 69.8% of its instances accounted for.

These findings support previous studies proposing that N1 and N2 often are bridged by events (Levi, 1978; Leonard, 1984; Ryder, 1994), that bridging events are limited (Levi, 1978;

Copestake & Lascarides, 1997; Copestake & Briscoe, 2005), and that the semantic relations are limited as well (Søgaard, 2005; Huang, 2008).

Second, are there semantic relations between N1 and N2 that do not involve bridging events?

To understand NNCs of simple relations does not require the identification of what one component entity does to the other. FrameNet data also suggest that bridging events sometimes are absent. We say this because we found that, among the NNCs that a N2 derives, N1s that involve complex relations usually can be mapped to FEs in eventive frames that the bridging event represents, while those that involve simple ones do not. For example, although "Material" is a productive static FE in the entity frame CLOTHING, it is not among the FEs of the eventive frame DRESSING, which describes the process and state of putting and having clothes on. In contrast, "Wearer" and "Body_location," which are also FEs of CLOTHING but involve complex relations, also assume FEs in DRESSING as arguments of LUs like "dress-up" and "put-on." Such distributional differences of FEs mean that the N1s represented by them are also distributed differently, resulting in NNCs contrasting in simple and complex terms.

While the simple-complex distinction also is attested to in a corpus-based framework like FrameNet, it seems that it is not recognized as a distinct class in Levi's widely-adopted system. While it appears that Levi (1978) considers some simple NNCs under the predicate HAVE, the status of other simple NNCs is unclear. For example, *imperial bearing* is classified as an instance of HAVE and paraphrased as 'have the bearing of an emperor.' Nevertheless, it seems that HAVE does not cover all the simple relations, as she defines the predicate as roughly corresponding to the semantic roles of "productive," "constitutive," and "compositional," which do not exhaust all simple relations. Moreover, some simple instances fall under her other predicates. For example, *apple seed* is considered an instance of FROM. We think FrameNet as a mapping means helps sort simple NNCs under semantic relations like Levi's predicates.

With regards to implementation, the findings indicate that simple and complex NNCs should be processed differently. For simple NNCs, host-attribute-value sets, place names, temporal expressions, and conjunction pairs to some degree can be exhaustively listed, as we have done in our knowledge base, Extended-HowNet, reducing identification of simple relations to table-checking. The E-HowNet taxonomy can also detect meronymy relations. For complex NNCs, the inventory of LUs and their argument FEs in FrameNet's frames narrows down the possible interpretations of NNCs.

We believe such mappings can complement the inadequacies of frameworks like Levi's (1979). First, the designation of FrameNet makes NNCs' readings more specific, as frames use fine-grained FEs and LUs are real words. Similarly, classification can be FE-based. For example, *lemon peels* and *apple seed* both belonging to the FE pair Whole-Part can be a reason for them to be grouped under the same predicate; for example, HAVE. Another classifying criterion is the simple-complex distinction. For example, to analyze the example in a different way, NNCs of the HAVE type can be defined as being made up of FE pairs like Whole-Part or Part-Whole and belonging to the simple type. Along the same vein, her IN category may involve NNCs with N1s of the FEs Time and Location, which in turn define the simple subclass of time and space. Finally, since frames are motivated by semantic and syntactic differences between words, they are expected to grow in coverage with more words' behaviors analyzed and new frames annotated.

7. Conclusion

The current study shares the insights with previous researchers that NNCs usually describe a limited range of situations and that the meaning of an NNC is compositional, while putting forth the idea that the range of semantic relations for event-bridging NNCs usually is clustered around the head, *i.e.* N2. We attained such findings by mapping the situations sorted by N2's semantic categories to frames from FrameNet, which is based on corpus-attested thematic patterns. We also noted that N1 and N2 sometimes are bridged in non-eventive ways. Both eventive and non-eventive cases can be interpreted through mapping to resources like FrameNet and E-HowNet.

Acknowledgement

This research was supported in part by the National Science Council under a Center Excellence Grant NSC 99-2221-E-001-014-MY3.

References

Chung, Y. S., & Chen, K. J. (2010). Analysis of Chinese morphemes and its application to sense and part-of-speech prediction for Chinese compounds. *ICCPOL 2010*, California, USA, 2010.

Copestake, A., & Lascardies, A. (1997). Integrating symbolic and statistical representations: The lexicon pragmatics interface. In *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL (ACL-EACL'97)*, Madrid, 1997, 136-43.

Huang, H. J. (2008). *Strategies in comprehending Mandarin Chinese noun-noun compounds with animals, plants, and artifacts as constituents*. MA thesis. National Cheng-Kung University, 2008.

Leonard, R. (1984). *The Interpretation of English Noun Sequences on the Computer*, Amsterdam: North-Holland.

Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals*. New York: Academic Press.

Liu, C. H. (2008). *Xiandai Hanyu Shuxing Fanchou Yianjiu* (*現代漢語屬性範疇研究*). Chengdu: Bashu Books.

Ryder, M. E. (1994). *Ordered Chaos: The Interpretation of English Noun-Noun Compounds*, University of California Press, Berkeley, CA.

Søgaard, A. (2005). Where does the meaning of compounds and possessives come from? A contrastive view. *The 3rd International Conference in Contrastive Semantics and Pragmatics*, Shanghai, China.

Appendix A: Examples[4] of mappings of N2-based NNC categories to FrameNet's entity and event frames (To avoid visual cluster, subclasses of simple relations are indicated as numbered in Section 3)

Simple_(1c/1a[5])

Telic/Use[6] + Clothing

[4] In part because of limited space and in part for demonstrative purpose only, we did not list examples of two of the nine productive semantic categories, "vehicle" and "container," neither did we exhaust all the instances of the other seven categories.

[5] The N1s here can be seen as either spatial (1a) or an important attribute of PEOPLE (1c).

[6] Sometimes called "Type" in FrameNet.

Appendix B: The mapped percentage of each N1 semantic role for the nine categories:

("Others" refers to instances we could not map with existing semantic role labels from FrameNet and E-HowNet.)

N2 category	**Road**		**Text**	
N1's role /N1-N2 elation, type number, and percentage	Theme	10 (25%)	Text	114 (94.2%)
	Conjunction	10 (25%)	Medium	7 (5.8%)
	Meronymy	7 (17.5%)		
	Material	6 (15%)		
	Path	4 (10%)		
	Name	3 (7.5%)		
Total mapped instances	**40/40 (100%)**		**121/121 (100%)**	

N2 category	**People**		**People of Different Vocations**	
N1's role /N1-N2 relation, type number, and percentage	Origin	167 (68.7%)	Telic/Use	35 (72.9%)
	Ethnicity	32 (13.2%)	Place_of_ Employment	3 (6.2%)
	Affiliation	12 (4.9%)	Contract_Basis	2 (4.2%)
	Hobby	10 (4.1%)	Compensation	2 (4.2%)
	Vocation	8 (3.3%)	Ethnicity	1 (2.1%)
	Material	5 (2.1%)	Rank	1 (2.1%)
	Appearance	4 (1.6%)	Compensation	2 (4.2%)
	Time	3 (1.2%)		
	Others	2 (0.8%)	Others	2 (4.2%)
Total mapped instances	**241/243 (99.2%)**		**46/48 (95.8%)**	

N2 category	Wealth		Container	
N1's role /N1-N2 relation, type number, and percentage	Goods	59 (81.9%)	Material	224 (52.5%)
	Seller	7 (9.7%)	Content	166 (38.9%)
	Buyer	6 (8.3%)	Meronymy	14 (3.3%)
			Shape	4 (0.9%)
			Location	3 (0.7%)
			Others	16 (3.7%)
Total mapped instances	**72/72 (100%)**		**411/ 427 (96.3%)**	

N2 category	Clothing		Vehicle	
N1's role /N1-N2 relation, type number, and percentage	Material	17 (36.2%)	Theme	17 (24.6%)
	Wearer	10 (21.3%)	Means-of-propulsion	10 (14.5%)
	Sub_region	7 (14.9%)	Location	7 (10.1%)
	Conjunction	5 (10.6%)	Possessor	6 (8.7%)
	Location	3 (6.4%)	Meronymy	5 (7.2%)
	Others	5 (10.6%)	Shape	3 (4.3%)
			Itinerary	2 (2.9%)
			Conjunction	2 (2.9%)
			Material	1 (1.4%)
Total mapped instances	**42/47 (89.3%)**		**53/69 (76.8%)**	

N2 category	Food	
N1's role /N1-N2 relation, type number, and percentage	Constituent_Part	53 (61.6%)
	Conjunction	4 (4.7%)
	Shape	3 (3.5%)
	Others	26 (30.2%)
Total mapped instances	**60/ 86 (69.8%)**	

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 63-80

HMM-based Mandarin Singing Voice Synthesis Using Tailored Synthesis Units and Question Sets

Ju-Yun Cheng*, Yi-Chin Huang*, and Chung-Hsien Wu*

Abstract

Fluency and continuity properties are essential in synthesizing a high quality singing voice. In order to synthesize a smooth and continuous singing voice, the Hidden Markov Model-based synthesis approach is employed in this study to construct a Mandarin singing voice synthesis system. The system is designed to generate Mandarin songs with arbitrary lyrics and melody in a certain pitch range. In this study, a singing voice database is designed and collected, considering the phonetic converge of Mandarin singing voices. Synthesis units and a question set are defined carefully and tailored the meet the minimum requirement for Mandarin singing voice synthesis. In addition, pitch-shift pseudo data extension and vibrato creation are applied to obtain more natural synthesized singing voices.

The evaluation results show that the system, based on tailored synthesis units and the question set, can improve the quality and intelligibility of the synthesized singing voice. Using pitch-shift pseudo data and vibrato creation can further improve the quality and naturalness of the synthesized singing voices.

Keywords: Mandarin Singing Voice Synthesis, Hidden Markov Models, Vibrato

1. Introduction

In recent years, Mandarin text-to-speech synthesis systems have been proposed and have achieved satisfactory performance (Ling, 2012; Wu, 2007). These systems are able to synthesize fluent and natural speech, even with personal characteristics (Huang, 2013). Recently, singing voice synthesis has been one of the emerging and popular research topics. Such systems enable computers to sing any song.

There are two main methods in the research on corpus-based singing voice synthesis. The first one is the sample-based approach. The principle of this method is to use a large database

* Department of Computer Science and Information Engineering, National Cheng Kung University, Taiwan
E-mail: { carrie771221; ychin.huang; chunghsienwu}@gmail.com

of recordings of singing voices that are further segmented into units. In the synthesis phase, based on a given score with the lyrics, the system then searches and selects appropriate sub-word units for concatenation. VOCALOID (Kenmochi, 2007) is such a singing voice synthesizer that enables the user to input lyrics and the corresponding melody. Given the score information, the system selects the necessary samples from the Singer Library and concatenates them to produce the synthesized singing voice. Finally, the system performs pitch conversion and timbre manipulation to generate smoothed concatenated samples. The software was originally only available in English and Japanese, but VOCALOID 3 has added support for Spanish, Chinese, and Korean. A Mandarin singing voice system using a unit selection method was proposed in (Zhou, 2008). Singing units in this method are chosen from a singing voice corpus with the lyrics of the song and the musical score information embedded in a MIDI file. To improve the synthesis quality, synthesis unit selection and the prosody and amplitude modification are applied. This system uses a Hanning window to smooth instances where speech segments were concatenated. Although the unit selection method is able to synthesize high quality speech at the waveform level, the concatenation-based methods suffer from the discontinuity problem at the boundaries between concatenated units. As different samples that make up the singing voice are recorded in different pitches and phonemes, discontinuity might exist in the resulting singing voice.

The other method is the statistical approaches, where hidden Markov models (HMMs) (Oura, 2010; Saino, 2006) are the most widely used. Acoustic parameters are extracted from a singing voice database and modeled by the context-dependent HMMs. The acoustic parameters are generated by the concatenated HMM sequence. Finally, vocoded waveforms of the singing voice are generated from the inverse filter of the acoustic parameters. Sinsy (Oura, 2010) is a free-online HMM-based singing voice synthesis system that provides Japanese and English singing voices. Users can obtain synthesized singing voices by uploading musical scores. Synthesizing singing voices based on HMMs sound blurred due to the limitation of the current vocoding technique. Nevertheless, it can generate a smooth and stable singing voice, and its voice characteristics can be modified easily by transforming the parameters appropriately.

In addition to the concatenation-based method and statistical method, there are also some other methods proposed to generate a Mandarin singing voice, *e.g.*, Harmonic plus Noise Model (HNM) (Gu, 2008), which adopted HNM parameters of a source syllable to synthesize singing syllables of diverse pitches and durations. This method can generate singing voices with good quality. Nevertheless, the discontinuity problem occurring at the concatenation points is still a major problem. Speech-to-singing method (Saitou, 2007) is another approach. Instead of synthesizing from a singing database, the speech-to-singing method converts speech into a singing voice by a parameter control model. Similarly, text-to-singing (lyrics-to-singing)

synthesis (Li, 2011) is used to generate synthesized speech of input lyrics by a TTS system followed by a melody control model that converts speech signals into singing voices by modifying the acoustic parameters. These two methods are based mainly on conversion rules that could be patchy.

Research on speech and singing synthesis has been closely linked, but there are important differences between the two methods with respect to the generated voices. The major parts of singing voices are voiced segments, whereas speech consists of a relatively large percentage of unvoiced sounds (Kim, 2003). Besides, fluency and continuity in singing voices are very important properties. In order to synthesize a smooth and continuous singing voice, an HMM-based synthesis approach is adopted in this study to build our singing voice synthesis system. To the best of our knowledge, the currently available HMM-based singing voice synthesis systems have not been applied to the Mandarin singing voice. By carefully defining and tailoring the synthesis units and the question set, a Mandarin singing voice synthesis system based on HMM-based framework has been constructed successfully in this study.

The rest of the paper is organized as follows. The proposed HMM-based singing voice synthesis system is introduced in Section 2. Section 3 consists of subjective and objective evaluations of the proposed system, compared to the original HMM-based singing voice synthesis system. Concluding remarks and future work are given in Section 4.

2. Proposed Mandarin Singing Voice Synthesis System

In recent years, the number of studies on HMM-based speech synthesis has grown. Some research has made progress on prosody improvement (Hsia, 2010; Huang, 2012) to obtain more natural speech. Recently, an HMM-based method has been applied to singing voice synthesis (Saino, 2006). There are more combinations of contextual factors in singing voice synthesis than that in speech synthesis. Applying a unit selection method to singing voice synthesis is quite difficult, because it needs a huge number of singing voices. On the contrary, an HMM-based system can be constructed using a relatively small amount of training data. As a result, the HMM-based approach is easier for constructing a singing voice synthesizer.

The system proposed in this study is based on the HMM-based approach that was developed by the HTS working group (Zen, 2007). The proposed structure of the singing synthesis system based on HMM is shown in Figure 1.

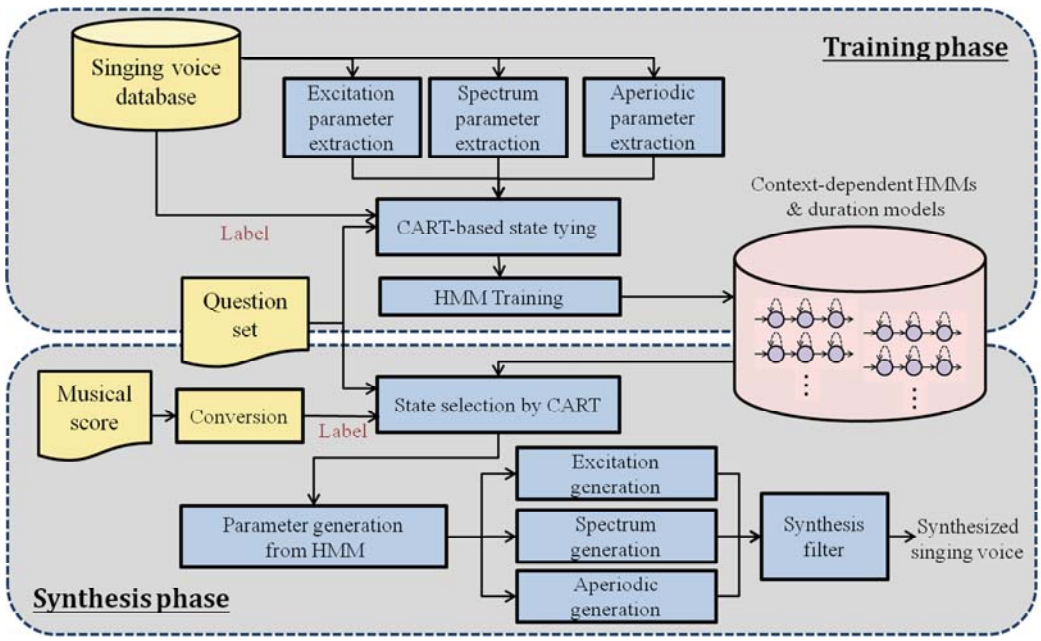


Figure 1. Structure of the HMM-based Singing Voice Synthesis System

In the training phase of the proposed system, excitation, spectral, and aperiodic parameters are extracted from a singing voice database. Lyrics and notes of the songs in the singing corpus are considered as contextual information for generating context-dependent label sequences. Then, the sequences are split and clustered with context-dependent question sets and the context-dependent HMM models are trained based on the clustered phone segments. In the synthesis phase, a musical score and the lyrics to be synthesized also are converted into a context-dependent label sequence. Based on the label sequence, a sequence of parameters, consisting of excitation, spectral, and aperiodic parameters, corresponding to the given song is obtained from the concatenated context-dependent HMMs. Finally, the obtained parameter sequences are synthesized to generate the singing voice.

2.1 Model Definition

Singing is the act of producing musical sounds with one's voice, and one main difference between a singing voice and speech is the use of the tonality and rhythm of a song. Therefore, the contextual factors should consist of not only linguistic information but also note information. In addition, the cue information obtains the actual timing of each phone in the singing data. The details of the model definition are described in the following section.

2.1.1 Linguistic Information

In the HMM-based Mandarin speech synthesis, "segmental Tonal Phone Model, STPM" (Huang, 2004) is often adopted to define the HMM-based phone models. Only a relatively small number of phone models are defined to characterize all Mandarin tonal syllables. Furthermore, in order to represent the five lexical tones for Mandarin syllables, each Mandarin syllable is defined to consist of three parts, based on phonology (Lin, 1992), as: **C+V1+V2**. In the phonological structure, **C** denotes the first extended initial phone and the following units (**V1** and **V2**) are tonal final phones. Tonal final phone conveys tonal information using the extended tone notations, such as H (high), M (middle), and L (low), *i.e.*, Tone 1: H+H, Tone 2: L+H, Tone 3: L+L, Tone 4: H+L, and Tone 5: M+M).

Although STPM can describe all pitch patterns in Mandarin speech, pitch patterns in singing voices are quite different from read speech. Figure 2 shows the pitch contours (blue lines) of the read speech and singing voice of the same sentence produced by the same person. As the figure shows, the pitch contour of the read sentence is controlled by the tone of each syllable. In contrast, the pitch contour of a singing sentence is relatively flat and corresponds to the musical notes of the corresponding syllables. The musical note is more of a requirement than the tones of the syllables for the pitch contour in singing voice. Therefore, the definition of each syllable for a singing voice is redefined as **C+V**, where **C** is still the extended initial sub-syllable and **V** is the final sub-syllable without tonal information.

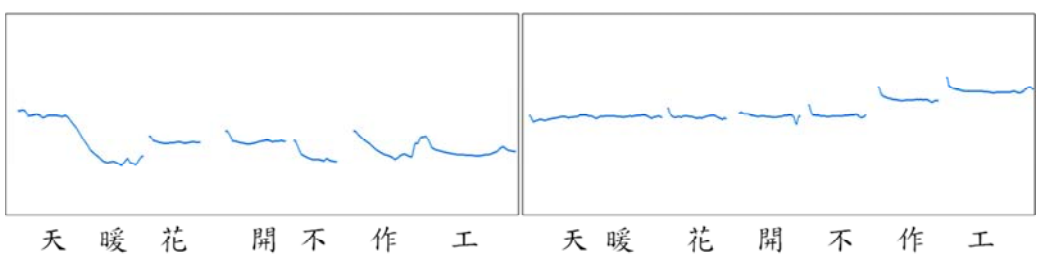


Figure 2. An example of the read speech and singing voice for the sentence "天暖花開不作工," which is uttered and sung by the same person.

Rhythm is one major difference between read speech and a singing voice. Vowels usually convey the rhythm of a singing voice since the vocal tract remains open while uttering a vowel, allowing the resonance frequencies of the vocal tract to remain stable. Because of these characteristics, vowels are probably one of the most important factors to represent a good singing voice. A Mandarin syllable consists of two parts: *initial* and *final*. The *initial* part is optional and is composed of consonants. The *final* part, namely vowels, includes *medial* and *rime*. The *medial* is located between the *initial* and the *rime*. The *medial* phonologically is connected with the *rime* rather than the *initial*. So, in the definition of singing sub-syllable

models, each *medial* is combined with a *final*. Furthermore, the combination of *medial* and *rime* is collectively known as a *final*, and some examples are listed in Table 1. For the singing model definition, the phonetic annotation is based on the Hanyu Pinyin. Note that the tone information (Arabic numbers) of the original tonal syllable is ignored for the *initial* or *final* models in the singing sub-syllable definition. Besides, we define the *final* models with *medial* as separate models to ensure that each vowel can have a specific model representing this property.

Table 1. Examples of finals with medial

	Tonal Syllable	C	V
ㄉㄧㄠˋ	diau4	d	iau
ㄌㄨㄛˇ	luo3	l	uo
ㄒㄩㄝˊ	shiueh2	sh	iueh

The syllable with only an *initial* is generally followed by an empty *rime* “ㄭ” . The empty *rime* does not have word phonetic annotation. In order to represent this property, we define a phoneme “zr” as the empty *rime* of the retroflex, which is connected only to the retroflex class of *initial* phonemes. Correspondingly, the phoneme “sr” is the empty *rime* of the alveolar, which is connected only to the alveolar class of *initial* phonemes.

In general, a long duration note is sung differently from short duration note. For shorter notes, temporal variation is relatively small and stable. Nevertheless, temporal variation of a longer note is much larger and unstable. Lengthening a syllable with a short duration note cannot precisely represent the expression of syllable with long duration note. So, when the word corresponds to a half note or above, the *finals* followed by an “L” are defined to denote the long duration model.

According to the above rules, 95 Mandarin signing sub-syllables are obtained according to the definition for singing voice. There are 21 *initial* sub-syllables, 18 *final* sub-syllables (2 of *finals* are empty *final* phonemes), 20 *medials* combined with *final* sub-syllables, and 36 long duration sub-syllables. In addition to the 95 signing sub-syllables, the silence and pause models are further included. Silence is an unvoiced segment in the beginning and the end of a song. Pause is an unvoiced segment in the middle of a song.

2.1.2 Note Information

In addition to lyrical information, note information is one of the vital factors for singing voice synthesis. Contextual factors of note information consist of three categories to fully describe singing characteristics, including pitch and duration of the note and the song structure. Note pitch refers to the melody of a song and determines if the song sounds great or not. In this

category, absolute pitch, relative pitch, pitch difference between previous and current notes, and pitch difference between current and next notes are included. Duration is the length of a note and is one of the bases of rhythm. In this category, the length of the note can be expressed by three kinds of standards. Song structure means which overall musical form or structure the song adopted and the order of the musical score. Different note positions in the measure or phrase may have different expressions due to breathing. In this category, beat, tempo, key of the song, and position of each note are included.

2.1.3 Cue Information

Cue information considered in the contextual factors consists of the timing and the length of a sub-syllable. We manually segment all of the songs at the sub-syllable level. The timing information of a sub-syllable measured based on a time interval of 0.1 seconds will be converted into the absolute length of the note. The position of note identity in the measure or phrase is also converted according to the cue information.

2.2 Question Set for Decision Trees

Based on unit definition and contextual factors, we define five categories for the questions in the question set. The five categories of the question set are sub-syllable, syllable, phrase, song, and note. The details of the question set are described as follows.

(1) Sub-syllable: (current sub-syllable, preceding one and two sub-syllables, and succeeding one and two sub-syllables) Initial/final, final with medial, long model, articulation category of the initial, and pronunciation category of the final

(2) Syllable: The number of sub-syllables in a syllable and the position of the syllable in the note

(3) Phrase: The number of sub-syllables/syllables in a phrase

(4) Song: Average number of sub-syllables/syllables in each measure of the song and the number of phrases in this song

(5) Note: The absolute/relative pitch of the note; the key, beat, and tempo of the note; the length of the note by syllable/0.1 second/thirty-second note; the position of the current note in the current measure by syllable/0.1 second/ thirty-second note; and the position of the current note in the current phrase syllable/0.1 second/thirty-second note

2.3 Baseline Model

There are 5364 different questions defined in the question set. The HMMs for the baseline Mandarin singing voice synthesis system were trained based on the entire question set, and the resulting clustered HMMs are shown in Table 2 and Table 3. As shown in these tables, the

number of leaf nodes in the tree clustered using fundamental frequency (F0) is 3951. The number of each state for the clustered F0 models is shown in Table 2. The most frequently used questions for every clustered tree of each state were sub-syllable types, position of note in measure or phrase, and phrase level.

The number of leaf nodes in the trees for mel-cepstral coefficients (mgc) is 2844. The number of the leaf nodes in each state is shown in Table 3. The most frequently used questions are the same as the results for F0.

Table 2. Number of leaf nodes in each state in F0 tree

State	State 2	State 3	State 4	State 5	State 6
Number of nodes	1146	509	366	626	1304

Table 3. Number of leaf nodes in each state in mel-cepstral coefficient tree

State	State 2	State 3	State 4	State 5	State 6
Number of nodes	244	849	938	604	209

2.4 System Refinement

The baseline of singing voice system can synthesize arbitrary songs, but it still has a lot of room to improve. The approaches we implemented to refine our system include question set modification, singing voice database extension using pitch-shift pseudo data, and vibrato creation.

2.4.1 Pitch-Shift Pseudo Data

Pitch is highly related to the notes and the sounds we hear when someone is singing. The quality of the song strongly depends on the accurate pitch of all notes produced by the singer. The quality of the HMM-based synthesized singing voices depends strongly on the training data, owing to its statistical nature. Therefore, the singing database should cover the pitch range of the notes in the song. Using the pitch-shift pseudo data, it is helpful to cover the missing pitch of sub-syllables and increase the size of the training data. We examine whether all Mandarin sub-syllables we defined cover the whole pitch range (C4~B4) or not. Since shifting too much frequency of a note will change the timbre, the missing pitches of sub-syllables could be obtained using the nearby notes from other songs.

2.4.2 Question Set Modification

The parameters generated from the clustered HMMs are highly correlated to the speech quality of the synthesized singing voice. A large number of contextual factors are not suitable when the size of the training data is not large enough to be clustered by various contextual

factors, and this may cause a data sparseness problem. The selection of the question set is crucial for generating proper models. In the baseline system, the most frequently used questions in the trees for F0 and mel-cepstral coefficients are sub-syllables types, position of note, and phrase level. Nevertheless, our singing database is not large enough to obtain every contextual factor. Thus, the question set should be tailored to remove some unsuitable questions. The removed questions consist of three types, including duplicate questions, indirect questions, and relative questions.

Duplicate questions refer to times when the note length can be represented by two types of units, 0.1 second and thirty-second note. Although 0.1 second is an absolute length and thirty-second note is the relative pitch of the recorded waveform, both units describe the same information. So, we delete the note length question with 0.1 second. Indirect question means the questions at the level of phrase and song, which are called paralinguistic information. These questions do not directly represent the information of one note, because they are mainly about how many sub-syllables and syllables there are in phrases and the average numbers of sub-syllables and syllables in each measure of the songs. The essential information of a note is its pitch and length, so the questions about position of note are also indirect questions. The paralinguistic information, however, could be useful when the size of corpus is large. Every song has different keys, so the standard of the relative pitch also is different. Two notes with the same relative pitch may have different absolute pitch values. Therefore, we delete the question sets related to relative pitch.

Furthermore, we modify the absolute pitch questions by keeping the questions with absolute answers and remove the questions with comparative answers. Thus, we can ensure that the leaf node that is divided by the absolute pitch questions can be clustered with the same absolute pitch.

2.4.3 Vibrato Creation

Vocal vibrato is a natural oscillation of musical pitch, and singers employ vibrato as an expressive and musically useful aspect of the performance. Adding vibrato can make the synthesized singing voice more natural and expressive. The frequency and the amplitude can be considered since they are the two fundamental parameters affecting the characteristic sound of a vibrato effect. The method to create vibrato is to vary the time delay periodically (Zölzer, 2002), and it uses the principle of Doppler Effect. Our system implemented the vibrato effect by a delay line and a low frequency oscillator (LFO) to vary the delay.

3. Evaluations

3.1 Singing Voice Database

For the construction of the signing voice database, the musical scores from nursery rhymes and children's songs are considered as the candidates. The major selection criterion for choosing the songs is the phonetic coverage for synthesizing universal Mandarin singing voices. The lyrics of the selected songs should cover all of the sub-syllables in Mandarin. A total of 74 songs were selected. Some of the selected songs have two or more versions with the same melody but different lyrics. Considering the variation of pitch and timbre, a female singer who has been participating in a singing contest and is a member of the a cappella team was invited as the signer to provide a stable and natural-sounding signing voice. The singer used the built-in microphone of a MAC notebook for recording. The songs were recorded using Audacity. The environment where the signer recorded was quiet. Noises, including the metronome, were not allowed. Besides, each song has two versions in order to increase the quantity of the database. The singing data with low signal-to-noise ratio or energy exceeding a limit were not included. The amplitude of all singing data was normalized. The overview of this database is summarized in Table 4. To improve the quality of the database, sub-syllable boundaries and musical scores were manually corrected.

Table 4. Details of NCKU singing voice database

Songs	Nursery rhymes (children's songs) Total 148 songs
Singer	One female
Pitch range	C4~B4
Version	2
Total time	About 102 minutes
Sample rate	48 kHz
Resolution	16 bits
Channels	Mono

3.2 Experimental Conditions

Singing voice signals were sampled at a rate of 48 kHz and windowed by a 25ms Blackman window with a 5ms shift. Then, mel-cepstral coefficients were obtained from the STRAIGHT algorithm (Kawahara, 2006). The feature vectors consisted of spectrum, excitation, and aperiodic factors. The spectrum parameter vectors consisted of 49-order STRAIGHT

mel-cepstral coefficients, including the zero-th coefficient, their delta, and delta-delta coefficients. The excitation parameter vectors consisted of log F0, its delta, and delta-delta.

A seven-state (including the beginning and ending null states), left-to-right Hidden Semi-Markov Models (HSMM) (Zen, 2007) was employed, in which the spectral part of the state was modeled by a single diagonal Gaussian output distribution. The excitation stream was modeled with multi-space probability distributions HSMM (MSD-HSMM), each of which consisted of a Gaussian distribution for "voiced" frames and a discrete distribution for "unvoiced" frames.

The term *Riffs and runs* implies a syllable with multiple notes. In other words, it is a quick articulation of a series of pitches sustained on a single vowel sound. In the proposed method, the generation of riffs and runs repeats the last final in previous words to mimic the singing skill.

Furthermore, in the middle of a song, vibrato is combined with the amplitude in 4E-4 millisecond, frequency in 6 Hz, and start timing in 25% of the sub-syllable. At the end of a song, vibrato is combined with the amplitude in 8E-4 millisecond, frequency in 5 Hz, and the start timing is at the position of 50% of the sub-syllable.

3.3 Evaluation Results

To evaluate the constructed Mandarin singing voice synthesis system, we conducted a subjective listening test. Ten songs not included in the training data were divided into two parts. Therefore, we obtained 20 parts for testing. The testing waveforms generated by different systems were presented to the subjects in a random order. 12 native Mandarin speaking subjects were asked to participate in the evaluation test. Mean Opinion Score and Preference test were used as evaluation measures for the subjective test.

In order to evaluate the effectiveness of the refinements we proposed, four different settings of the synthesis models were used. These models were evaluated on the effect of the refinements, *i.e*. question set modification and inclusion of pitch-shift pseudo data. The settings and the descriptions are described in Table 5.

Table 5. Four different settings of models and their descriptions

Model	Description
Baseline	All question set
QM	Question set modification
PS	Pitch shift pseudo data
QM+PS	Question set modification and pitch shift pseudo data

3.3.1 Pitch Contour Comparison

Figure 3 shows the Mandarin singing voice synthesis system can generate the F0 patterns similar to the actual F0 patterns of the musical score. Figure 4 shows the Mandarin singing voice synthesis system can generate the pitch contour of the synthesized singing voice with almost the same as the pitch contour of the original singing voice. Nevertheless, some of the singing phenomena, such as overshoot and preparation were smoothed after HMM training.

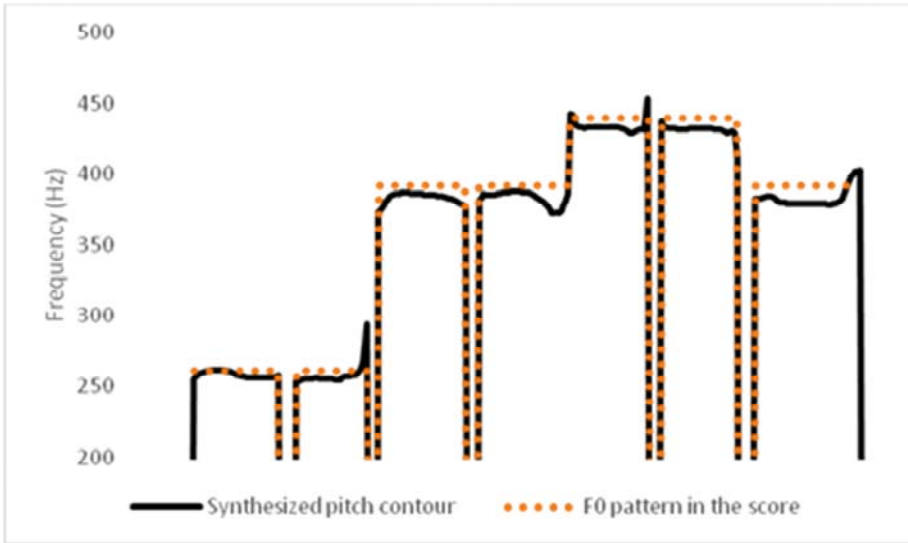


Figure 3. Comparison with generated F0 patterns and F0 patterns in the score

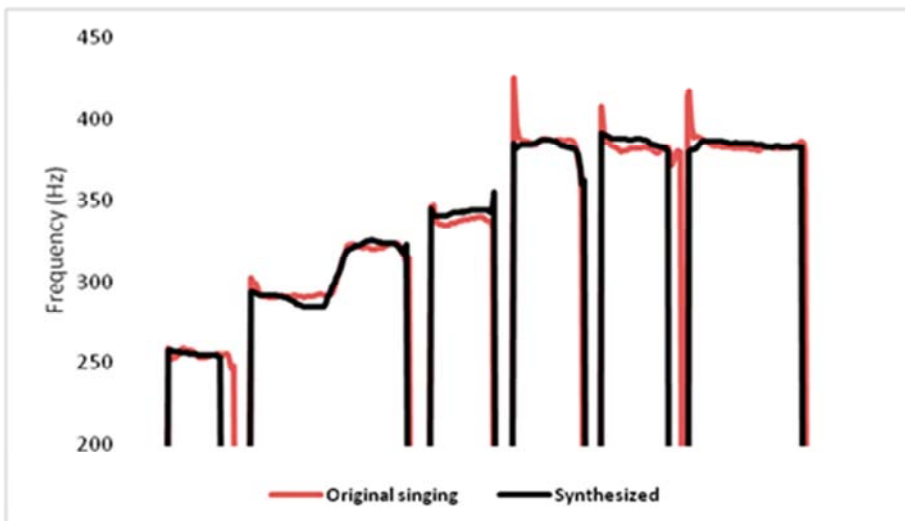


Figure 4. Comparison between the original singing and the synthesized singing pitch contours

3.3.2 Preference Test

We evaluated the nature of the synthesized singing voice with a long duration model. Figure 5 shows the system with the long duration model has 62% preference, which is higher than 38% for the system without the long duration model. This shows that the long duration model can improve the nature of phones with long duration. Therefore, all of the evaluated systems use the long duration model in the following tests.

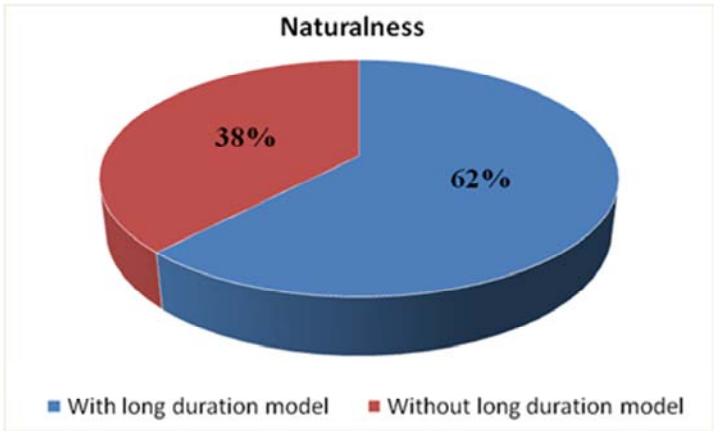


Figure 5. Result of preference test with long duration model

In addition, we evaluated the nature of the synthesized singing voice with vibrato. The preference result is shown in Figure 6. The subjects only slightly preferred the synthesized singing voice with vibrato over that without vibrato. The main reasons are that two combinations of parameter settings are insufficient and that different pitches and situations must correspond to different combinations of vibrato parameters. Moreover, vibrato is not essential in children' songs. Subjects preferred simple over skillful singing styles in these kinds of songs.

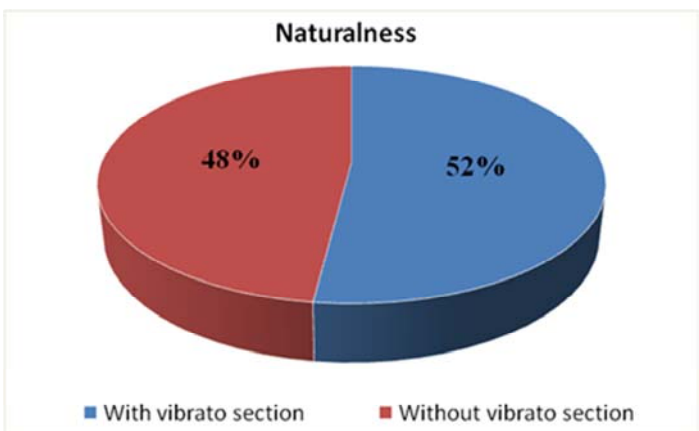


Figure 6. Result of preference test with/without vibrato

3.3.3 Mean Opinion Scores (MOS)

The MOS of quality in four evaluation settings is shown in Figure 7, and the MOS of intelligibility is shown in Figure 8.

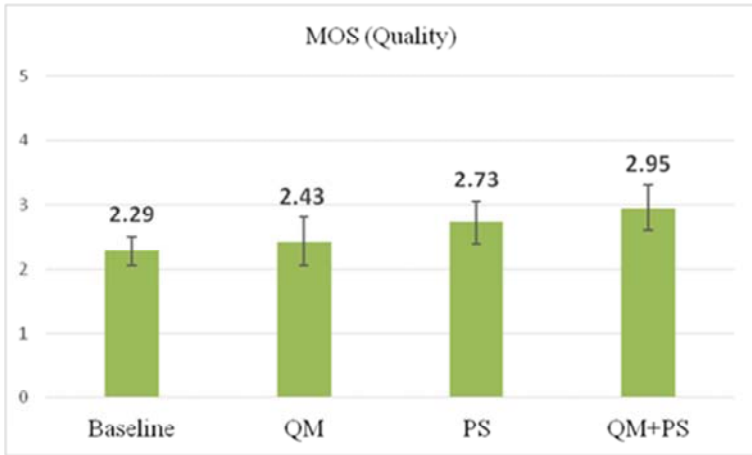


Figure 7. MOS of the synthesized singing voice in quality

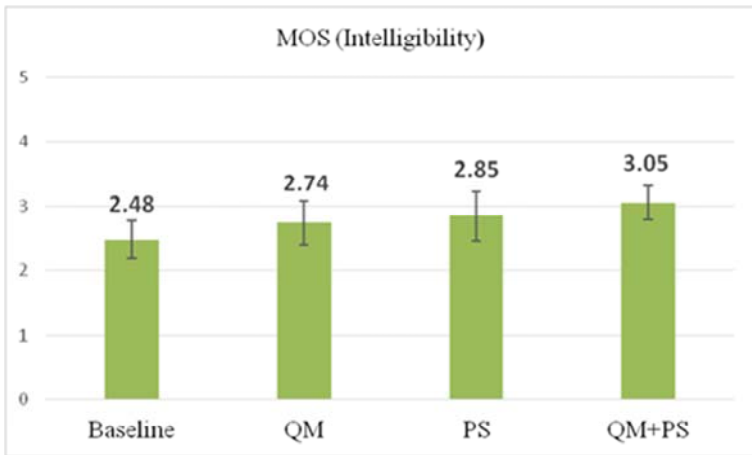


Figure 8. MOS of the synthesized singing voice in intelligibility

The results show that the baseline singing voice system has the lowest MOS, because the training data is insufficient for clustering using a large number of questions and because some sub-syllables are not covered in some of the pitch frequencies. After question modification, MOS is 2.43 in quality and 2.74 in intelligibility, which are higher than those for the baseline system. The PS model has MOS of 2.73 in quality and 2.85 in intelligibility, which are higher than those for the baseline system. This shows that adding pitch-shift pseudo data is one of the useful refinements. Finally, the MOS of QM+PS model is 2.95 in quality and 3.05 in intelligibility. These scores are higher than those for the PS model with the modified question

set and the QM model with pitch-shift pseudo data. According to the results, we can conclude that, although all question sets take all of the contextual factors into account, some contextual information might not be included in the corpus, which may cause bad clustering results. By tailoring the question set appropriately, the system can improve the quality and intelligibility of the synthesized singing voice. In addition, adding pitch-shift pseudo data also can improve the quality of the synthesized singing voice.

4. Conclusions and Future Work

In this paper, a corpus-based Mandarin singing voice synthesis system based on hidden Markov models (HMMs) was implemented. We defined the Mandarin phone models and the question set for model clustering. Linguistic information and musical information both are modeled in the context-dependent HMM. Furthermore, three methods were employed to refine the constructed system, *i.e*. question set modification, pitch-shift pseudo data, and vibrato creation. Experimental results show that the proposed system could synthesize a satisfactory singing voice. The performance of the corpus-based synthesis system is highly dependent on the training corpus, and the quality of the corpus can directly affect the synthesized voice quality. The environment for data recording should be professional and silent, such as in an anechoic chamber or using sound-absorbing equipment. Furthermore, the training corpus should be as large as possible to cover all contextual factors. Although our singing database was designed with high phonetic coverage and enhanced by adding pseudo data for better pitch coverage, there are some factors that were not covered, such as the coverage of duration and higher level information.

Besides, a more accurate model is essential for synthesizing a better singing voice. Model clustering should be categorized and labeled with priority, since some factors are more important than others for singing characteristics. The process of clustering decision trees should be guided based on the priority of clustering questions to obtain a more accurate model.

The singer's timbre and pronunciation are also important factors that affect synthesized singing voice quality. The nasal tone of a singer's voice might cause acoustic information disappearance when uttering syllables with higher pitches. Unclear utterances also cause the synthesized singing voice to become unintelligible. For further improvement, these problems should be carefully considered in order to generate better synthesized singing voices.

References

Gu, H.-Y., & Liau, H.-L. (2008). Mandarin Singing Voice Synthesis Using an HNM Based Scheme. *International Congress on Image and Signal Processing (CISP)*, 347-351.

Hsia, C.-C., Wu, C.-H., & Wu, J.-Y. (2010). Exploiting Prosody Hierarchy and Dynamic Features for Pitch Modeling and Generation in HMM-Based Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1994-2003.

Huang, Y.-C., Wu, C.-H., & Chao, Y.-T. (2013). Personalized Spectral and Prosody Conversion using Frame-Based Codeword Distribution and Adaptive CRF. *IEEE Trans. Audio, Speech, and Language Processing*, *21*(1), 51-62.

Huang, Y.-C., Wu, C.-H., & Weng, S.-T. (2012). Hierarchical prosodic pattern selection based on Fujisaki model for natural mandarin speech synthesis. *2012 8th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 79-83.

Huang, C., Shi, Y., Zhou, J., Chu, M., Wang, T., & Chang, E. (2004). Segmental tonal modeling for phone set design in Mandarin LVCSR. *Proceedings of ICASSP 04*, 901-904.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology*, *27*(6), 349-353.

Kenmochi, H., & Ohshita, H. (2007). VOCALOID-Commercial singing synthesizer based on sample concatenation. *INTERSPEECH 2007*, 4009-4010.

Kim, Y. E. (2003). *Singing Voice Analysis/Synthesis*. Ph.D. dissertation, Massachusetts Institute of Technology.

Li, J., Yang, H., Zhang, W., & Cai, L. (2011). A Lyrics to Singing Voice Synthesis System with Variable Timbre. *Applied Informatics and Communication Communications in Computer and Information Science*, 225, 186-193.

Lin, T., & Wang, L.-J. (1992). *Phonetic Tutorials*. Beijing University Press, 103-121.

Ling, Z.-H., Xia, X.-J., Song, Y., Yang, C.-Y., Chen, L.-H., & Dai, L.-R. (2012). *The USTC System for Blizzard Challenge 2012*. Blizzard Challenge Workshop.

Oura, K., Mase, A., Yamada, T., Muto, S., Nankaku, Y., & Tokuda, K. (2010). Recent Development of the HMM-bases Singing Voice Synthesis System-Sinsy. *The 7th ISCA Speech Synthesis Workshop*, 211-216.

Saino, K., Zen, H., Nankaku, Y., Lee, A., & Tokuda, K. (2006). An HMM-based Singing Voice Synthesis System. *International Conference on Spoken Language Processing (ICSLP)*, 1141-1144.

Saitou, T., Goto, M., Unoki, M., & Akagi, M. (2007). Speech-to-Singing Synthesis: Converting Speaking Voices to Singing Voices by Controlling Acoustic Features Unique to Singing Voices. *Applications of Signal Processing to Audio and Acoustics Workshop*, 215-218.

Wu, C.-H., Hsia, C.-C., Chen, J.-F., & Wang, J.-F. (2007). Variable-length unit selection in TTS using structural syntactic cost. *IEEE Trans. Audio, Speech, Lang. Process.*, *15*(4), 1227-1235.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., & Tokuda, K. (2007). The HMM-based Speech Synthesis System (HTS) Version 2.0. *The 6th ISCA Workshop on Speech Synthesis*, 294-299.

Zen, H., Tokuda, K. T., Masuko, T., Kobayasih, T., & Kitamura, T. (2007). A Hidden Semi-Markov Model-Based Speech Synthesis System. *IEICE Trans. Inf. & Sys*., *90*(5), 825-834.

Zhou, S.-S., Chen, Q.-C., Wang, D.-D., & Yang, X.-H. (2008). A Corpus-Based Concatenative Mandarin Singing voice Synthesis System. *2008 International Conference on Machine Learning and Cybernetics*, 2695-2699.

Zölzer, U. (2002). *DAFX- Digital Audio Effects*. John Wiley & Sons, Chapter 3, 68-69.

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 81-96

使用語音評分技術輔助台語語料的驗證

Using Speech Assessment Technique for the Validation of Taiwanese Speech Corpus

李毓哲*、王崇喆*、陳亮宇+、張智星#、呂仁園‡

Yu-Jhe Li, Chung-Che Wang, Liang-Yu Chen,

Jyh-Shing Roger Jang, and Ren-Yuan Lyu

摘要

本論文的主要研究為使用語音辨識及結合語音評分，對未整理的台語語料進行初步的篩選。藉由機器先過濾掉有問題的音檔，如錄音音量過小、太多雜訊、錄音音檔內容有誤等情形，取代傳統人工聽測費時的做法。本論文可分為三個階段，分別是：「基礎聲學模型訓練」、「語音評分與錯誤原因標記」及「效能評估」。

於基礎聲學模型訓練階段，以長庚大學提供的台語語料 ForSD (Formosa Speech Database)為材料，使用隱藏式馬可夫模型（Hidden Markov Model, HMM）進行聲學模型的訓練。聲學模型單位分別為：單音素聲學模型（Monophone acoustic model）、音節內右相關雙連音素聲學模型（Biphone acoustic model）及音節內左右相關三連音素聲學模型（Triphone acoustic model），其針對測試語料進行自由音節解碼辨識網路（Free syllable decoding）的音節辨識率（Syllable accuracy）最佳結果分別為：27.20%、43.28%、45.93%。

*國立清華大學資訊工程學系 Dept. of CS, NTHU, Taiwan
E-mail: {yujhe.li; geniusturtle}@mirlab.org
+國立清華大學資訊與應用研究所 ISA, NTHU, Taiwan
E-mail: davidson.chen@mirlab.org
#國立臺灣大學資訊工程學系 Dept. of CSIE, NTU, Taiwan
E-mail: jang@mirlab.org
‡長庚大學資訊工程學系 Dept. of CSIE, CGU, Taiwan
E-mail: renyuan.lyu@gmail.com

於語音評分與錯誤原因標記階段，將於基礎聲學模型訓練階段已訓練好的左右相關三連音素聲學模型，對待整理的語料進行語音評分，而將其評分結果依照門檻值分為三部分，分別為低分區、中間值區及高分區。且針對低分區部分語料進行人工標記，標記其錯誤原因，再對其擷取特徵，使用支持向量機（Support Vector Machine, SVM）訓練出分類器，最後以該分類器對低分區語料進行二次檢驗，將低分區語料分為可用語料及不良語料。

於效能評估階段，將原先訓練語料分別加入「未整理語料」、「中間值區及高分區語料」、「高分區語料」進行聲學模型的訓練，比較篩選語料前、後效能，其音節辨識率結果分別為：40.22%、41.21%、44.35%。

由結果看來，經過篩選後語料所訓練出的聲學模型與未經篩選語料所產生的聲學模型，其辨識率的差別最高可達 4.13%，證實本論文所提的方法，藉由語音評分確實能有效的自動篩選掉有問題的語句。

關鍵詞：台語語料整理、隱藏式馬可夫模型、語音評分、語音辨識、支持向量機

Abstract

This research focuses on validating a Taiwanese speech corpus by using speech recognition and assessment to automatically find the potentially problematic utterances. There are three main stages in this work: acoustic model training, speech assessment and error labeling, and performance evaluation.

In the acoustic model training stage, we use the ForSD (Formosa Speech Database) ,provided by Chang Gung University (CGU), to train hidden Markov models (HMMs) as the acoustic models. Monophone, biphone (right context dependent), and triphone HMMs are tested. The recognition net is based on free syllable decoding. The best syllable accuracies of these three types of HMMs are 27.20%, 43.28%, and 45.93% respectively.

In the speech assessment and error labeling stage, we use the trained triphone HMMs to assess the unvalidated parts of the dataset. And then we split the dataset as low-scored dataset, mid-scored dataset, and high-score dataset by different thresholds. For the low-scored dataset, we identify and label the possible cause of having such a lower score. We then extract features from these lower-scored utterances and train an SVM classifier to further examine if each of these low-scored utterances is to be removed.

In the performance evaluation stage, we evaluate the effectiveness of finding problematic utterances by using 2 subsets of ForSD, TW01, and TW02 as the

training dataset and one of the following: the entire unprocessed dataset, both mid-scored and high-scored dataset, and high-scored dataset only. We use these three types of joint dataset to train and to evaluate the performance. The syllable accuracies of these three types of HMMs are 40.22%, 41.21%, 44.35% respectively.

From the previous result, the disparity of syllable accuracy between the HMMs trained by unprocessed dataset and processed dataset can be 4.13%. Obviously, it proves that the processed dataset is less problematic than unprocessed dataset. We can use speech assessment automatically to find the potential problematic utterances.

Keywords: Taiwanese Corpus Validation, Hidden Markov Model, Speech Assessment, Support Vector Machine.

1. 緒論

傳統語料的整理往往需要耗費相當的時間以及需要具有專業知識背景的人員進行人工聽測。本論文的研究即是針對台語語料使用語音評分輔助機器篩選掉不良的語料，如空白音檔、文本有誤...等錯誤類型，取代傳統費時的人工聽測方法，藉此減少人工檢查的時間，加快語料庫的建立。

本論文首先會進行基礎聲學模型的訓練，藉由基礎聲學模型對未整理的語料進行語音評分，並依照分數門檻值將未整理的語料劃分為低分區語料、中間值區語料及高分區語料，最後分別將未經篩選語料與經篩選語料加入原始訓練語料重新進行聲學模型的訓練，以測試語料的辨識結果來評估語料整理的程度。

而對於低分區語料，我們特別對它進行人工標記，標記該音檔的不良類型，並依觀測到的不良類型對其擷取特徵，使用支持向量機（Support Vector Machine, SVM）訓練分類器。最後我們則使用該分類器對低分區語料再次檢驗為可用語料或是不良的語料，減少需要人工檢驗的語料數目。底下圖 1 為語料整理系統流程圖：

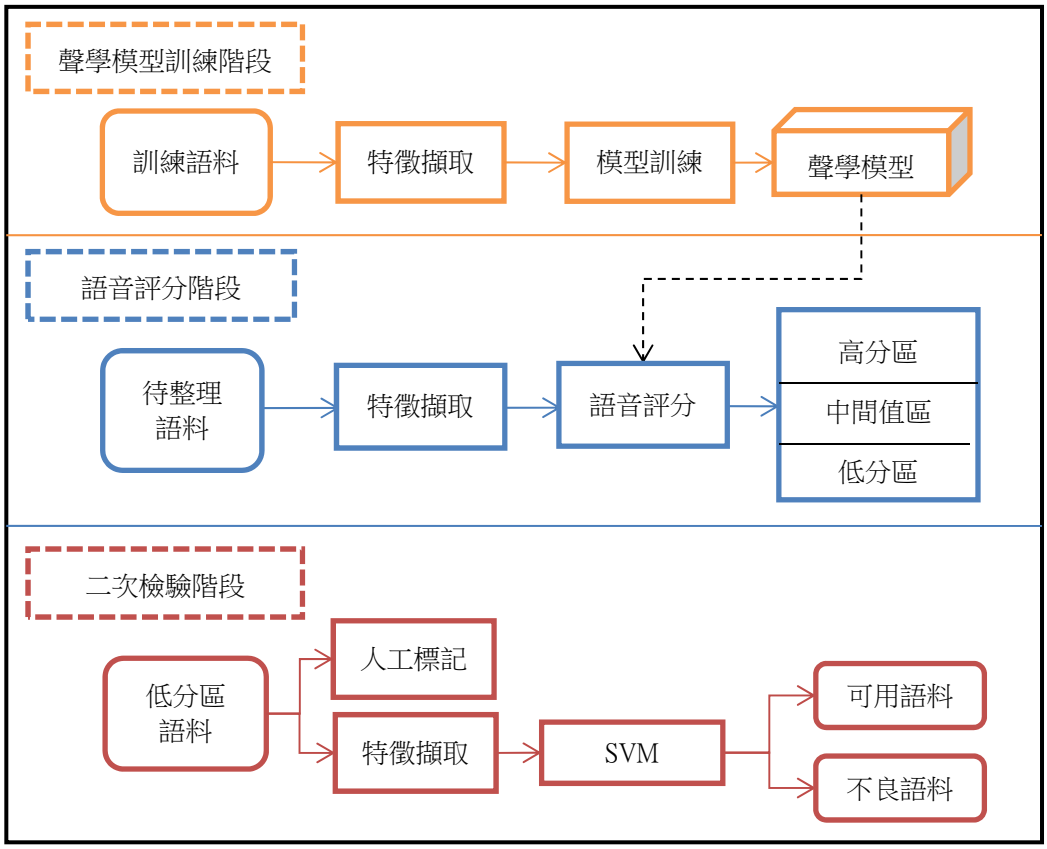


圖 1. 語料整理系統流程圖

因此，本論文的研究方向於如何使用電腦為輔助工具，對未整理語料藉由語音評分進行初步的篩選，並標記出不良類型，且以觀測到的不良類型為特徵產生分類器，對語料進行第二階段的檢驗。以下第二節將介紹論文之相關研究。第三節將詳述論文方法，提出使用語音評分輔助語料的驗證方法。第四節為本論文所提出的方法進行實驗，並對實驗的結果進行探討。第五節為本論文的結論以及未來研究方向。

2. 相關研究

2.1 標音系統與模型訓練

本論文內採用的標音系統為福爾摩沙音標（Formosa Phonetic Alphabet, ForPA）。ForPA 拼音是台灣地區華語、台語、客語三語共用的標音系統，在華語的音素有 37 個，台語有 56 個，兩種語言音素聯集共有 63 個，交集共有 32 個（朱晴蕾等，2010）。進行聲學模型之訓練時，梅爾倒頻譜係數（Mel-frequency Cepstral Coefficients, MFCCs）（Davis & Mermelstein, 1980）和對數能量（Log energy）做為語音特徵；並以隱藏式馬可夫模型（Hidden Markov Model, HMM）來建立不含聲調的聲學模型。

2.2 語音評分

語音評分（Speech assessment）（李俊毅，2002；陳宏瑞，2011）能藉由聲學模型對錄音進行評分，在本論文中以評分後的分數來評量音檔的與文本間的相似程度。但依據前人研究（黃武顯，2007），在某些狀況下語音評分的分數並不合理，因此在本論文中加入了三種分數調整的扣分機制，以降低評分時不合理情形出現的機率。

音節之音框個數差距過大:正常情況下,每個音節所占有的音框個數差距不會太大。當語句中所占的音框個數最長的音節與所占的音框個數最短的音節，音框個數差距達 3 倍以上時，則將分數減少為原先分數的 80%。

音節中連續音素之音框數目過小：在正常情況下，很少出現一個音節中連續數個音素其音框各數皆為 3 個音框，相當於音素的最小音框個數。當一個音節中出現連續兩個音素的音框數皆為 3 個音框，則將分數減少為原先分數的 80%。

音節數目不一：在本論文中我們允許評分時出現漏字的情形，即可跳過某個音節，而不強制對文本內的所有音節都做強迫對位的動作，因此系統評分的切音結果可能會出現缺字的情況。而在（李俊毅，2002; 陳宏瑞，2011）當中所提到的分數計算方法，僅針對切音的結果來進行分數的計算，並不會考慮到輸入語句的所有音節。當經評分後之切音結果的音節數目小於正確文本的音節數目時，則將分數依照切音結果的音節數目與輸入語句之音節數目之比例來調整。例如，原先輸入語句為「如果咱 e 先賢無 e 話 /ru-ger-lan-e-sen-hen-bher-e-ue 」，共 9 個音節，但切音結果後的語句卻為「ger-lan-e-sen-hen-bher-e-ue」,共 8 個音節,因此若調整前為 80 分,則調整後為 80×8/9=71 分。

3. 研究方法

3.1 基礎聲學模型訓練

本論文使用 HTK（Hidden Markov Model Toolkit）（Young, 2009）訓練聲學模型和調整特徵參數。於基礎聲學模型訓練階段，使用 HTK 對訓練語料擷取特徵後，分別訓練出三種不同音素單位的聲學模型，包含了單音素（Monophone）、音節內右相關雙連音素（Biphone）及音節內左右相關三連音素（Triphone）。

3.2 語音辨識結合語音評分

本論文中使用 HMM 聲學模型與語音評分及扣分機制作結合，藉由評分後的分數，當作音檔內容與文本內拚音內容的相似性參考。如果說錄音音檔的品質越好的話，藉由穩定的聲學模型便可以辨識出音檔的語句內容，評分分數也會較高；反之，如果錄音音檔的品質不好，例如有雜訊、錄音音量太小、音檔內容與實際文本不同、音檔片段切音沒切好...等，則會增加辨識的困難度，相對的評分分數會偏低。因此我們將評分分數的高低做為該音檔是否為優良語料的參考依據。

在對待整理語料進行語音評分的方面，首先我們以（李俊毅，2002; 陳宏瑞，2011）

當中的方法，算出該語句的評分分數，再使用二-（二）所提到的扣分機制，產生該語句最終的評分分數。依據不同的分數門檻值，我們將評分後的語料分為低分區、中間值區及高分區。其中低分區內的語料普遍為不良的語料，但仍然可能存在著因為分數誤被低評的可用語料，因此我們藉由人工觀察，更深入地去分析低分區語料的錯誤類型；而依據評分的結果，高分區內的語料具有相當的公信力是屬於較優良的語料，或許當中也有可能出現該錄音音檔的評分分數被高評的的情況，但是因為有扣分機制的加入，該狀況會是極少數。而中間值區的語料屬於比較模糊的區域。

3.3 低分區語料二次檢驗

由於低分區內的語料可能因為語音評分及扣分機制的誤判，造成可用語料的分數被低評。因此本論文提出使用二次檢驗的方法，藉由 SVM 產生的分類器再次檢驗低分區語料，將語料分類為可用的語料和不良的語料。

由於我們事先不知道該語料是否為好的語料但分數被低評或者是不良語料，因此我們需要先對部分的低分區語料進行人工標記，標記其音檔問題，並將已標記的語料分為訓練語料及測試語料。接著再以標記過程中語料常見的錯誤類型當作特徵，藉由參數的調整產生出分類效果不錯的分類器。最後針對低分區內的所有語料使用該分類器進行二次檢驗，並將其被歸類為可用的語料進行人工檢驗。

藉由使用分類器進行二次檢驗的動作可降低誤刪除可用語料的機率，同時不需要針對低分區內的所有語料進行人工檢查，只需對過濾後的可用語料作人工檢驗即可，藉此減少人工檢查時間，加快語料庫的建立。

4. 研究結果與分析

4.1 語料簡介

本論文之語料來源為長庚大學於 2001 至 2003 年間，執行國科會委託計畫「台灣地區多語語音辨認之研究暨多語語音資料庫之建立」所蒐集的台語語料庫 ForSD (Formosa Speech Database)，本論文使用其中的 TW01,TW02,以及 TW03 等三個子集合（Lyu *et al*., 2004）。該語料分為訓練語料與測試語料兩個集合，本論文主要也以語料中分配好的集合做為訓練與測試。表 1 為該語料的相關數據。

表 1. 語料資訊

	訓練語料	測試語料
語料名稱	ForSD-TW01、ForSD-TW02	ForSD-TW01、ForSD-TW02
錄音格式	單聲道，16kHz，16bits	單聲道，16kHz，16bits
錄音者	600 人，男 317 人、女 283 人	26 人，男 13 人、女 13 人
錄音句數	117047 句，男 61908 句、女 55139 句	3072 句，男 1549 句、女 1523 句
錄音時間	32.58 小時	0.98 小時

本論文待整理語料來源為 ForSD 語料中的 TW03 子集合（廖子宇等，2012）。該語料依照錄音者身份可分為老師及學生語料，共由 4 位老師和 671 位學生錄製而成。其中男生錄音人數少於女生錄音人數，比例約為 1:1.5。錄音內容為文章段落內的句子。表 2 為待整理語料的相關數據。

表 2. 待整理語料資訊

	待整理語料
語料名稱	ForSD-TW03
錄音格式	單聲道，16kHz，16bits
錄音者	675 人，男 263 人（老師 2 人、學生 261 人）、女 412 人（老師 2 人、學生 410 人）
錄音句數	205311 句，男 82099 句、女 123212 句
錄音時間	136.14 小時

4.2 辨識網路與效能評估

本論文採用自由音節解碼（Free syllable decoding）做為辨識網路，其中欲辨識音節的個數為 853 個，而音節至下個音節的對數機率值為-20。該辨識網路前後音節設定為 silence，中間為欲辨識的音節，如圖 2 所示。

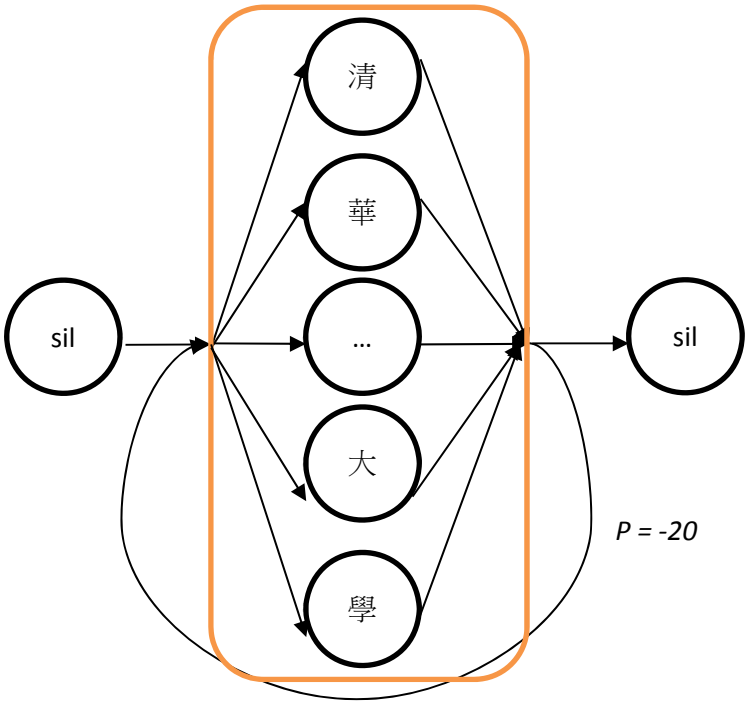


圖2. 辨識網路結構

正常情況下，當訓練時使用的語料量越多，則能訓練出較穩定的聲學模型，辨識效果也會較佳。但如果訓練語料內夾雜著不良的語料，比如有雜音、音檔與文本內容不符合、有咳嗽聲或是有異常停頓等情形出現時，這些不良的語料則會在訓練聲學模型時對其他的語料產生影響，造成訓練出來的聲學模型較不穩定，對辨識結果會產生影響。因此，在這裡我們使用辨識率來做為語料整理乾淨程度的效能評估方法，對篩選後語料與未經篩選語料所產生的聲學模型之辨識率進行比較。

在本論文中，我們採用音節辨識率（Syllable accuracy）做為語料整理乾淨程度的評估方法。底下為其計算公式：

$$Syllable\ accuracy = \frac{N - D - S - I}{N} \times 100\%$$

其中 N 為正確文本內的總音節數目；D 為遺失的音節數目，指的是出現在正確文本內，但在辨識的結果裡卻沒有辨識出來的音節數目；S 為替換的音節數目，指的是將某一個音節辨識成另一個音節的數目；I 為插入的音節數目，指的是辨識結果中除了正確的音節外，多出了不該出現的音節數目。

4.3 基礎聲學模型訓練

本實驗的目的為使用 ForSD-TW01、ForSD-TW02 訓練語料，藉由參數的調整訓練出穩定、辨識率不錯的基礎聲學模型，以供後續語音評分使用。

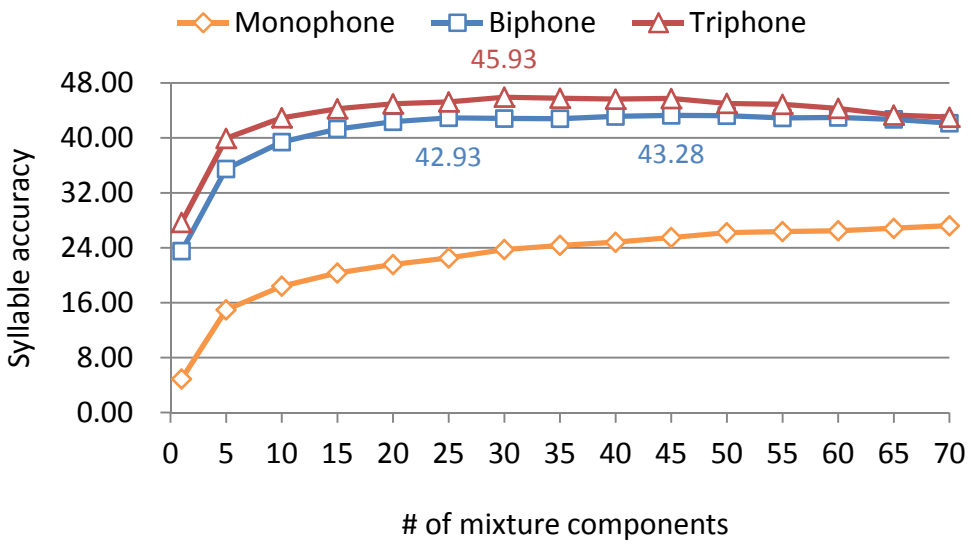


圖 3. Monophone、Biphone、Triphone 聲學模型於不同高斯混和數下之辨識結果

圖 3 列出了 Monophone、Biphone、Triphone 聲學模型於每個狀態下高斯混和數以 5 的倍數，從[1 1 1]至[70 70 70]提升的辨識結果，其中，每個狀態包含三個 Streams，silence 模型的高斯混和數為其他模型的兩倍。

在本實驗中，Triphone 聲學模型於高斯混和數為[30 30 30]時的音節辨識率達到最高，為 45.93%，符合於目前台語語音辨識研究的水平，之後便開始下降；而 Biphone 聲學模型於高斯混和數為[25 25 25]及[45 45 45]時達到高點，之後便開始下降；而 Monophone 聲學模型的音節辨識率則呈現持續上升的趨勢，但明顯與 Biphone、Triphone 聲學模型差距過大。說明了當模型的高斯混和數過高時，則會對該模型過度符合（Over fit），造成辨識率下降。由以上結果，我們選擇高斯混和數為[30 30 30]的 Triphone 聲學模型當作基礎聲學模型，於語音評分階段時使用。

4.4 對待整理語料進行語音評分

本實驗的目的為使用實驗一所產生的基礎聲學模型對待整理語料進行語音評分，並依照分數門檻值將待整理語料分為低分區、中間值區和高分區語料三個區塊。我們首先由前一節所述的基礎聲學模型，對待整裡語料進行強迫對位（Force alignment）及語音評分。在語音評分的過程中，如 2.2 節所述，我們允許忽略文本中欲辨識的音節。得到評分分數後，我們依照分數門檻值將待整理語料分為低分區、中間值區及高分區語料。分數門檻值設定為 60、80 分，將評分分數低於 60 分的語料歸為低分區、評分分數高於 80 分以上的語料歸為高分區、而評分分數為 60 分以上但未達 80 分的語料歸為中間值區。圖 4 為待整理語料進行分區後的句數與時間分布圖，由於一般來說，語者較不會念得特別好或特別壞，所以大部分語料的評分分數落點於 60 至 80 分間。

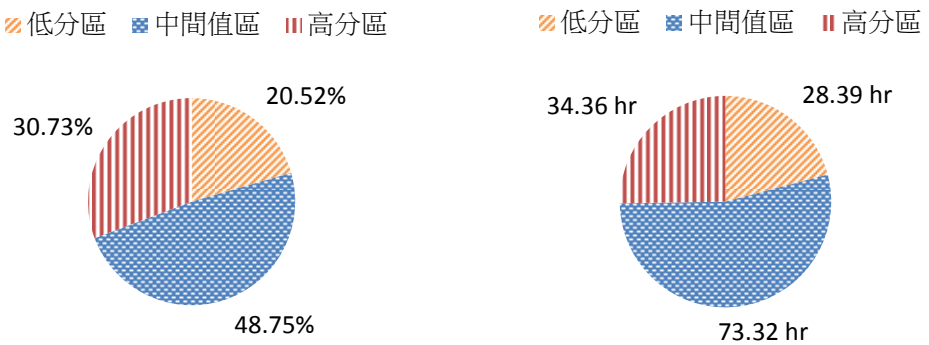


圖 4. 低分區、中間值區、高分區之語料句數（左）與時間（右）分布圖

4.5 未經篩選語料與經篩選語料之聲學模型訓練

本實驗的目的為比較使用未經篩選語料與經語音評分篩選後語料所訓練出的聲學模型，其中使用音節辨識率作為語料整理乾淨程度的評估方法。

本實驗使用上一小節的結果，將原先 ForSD-TW01、ForSD-TW02 訓練語料分別加入未整理語料、中間值區和高分區語料、高分區語料，來進行初始化模型訓練，最終產生三個不同訓練語料組合的聲學模型。於特徵擷取與模型訓練時，使用的基本參數設定皆與基礎聲學模型訓練時相同。

表 3、4、5 分別為加入未經篩選語料、加入中、高分區語料、加入高分區語料所訓練出來的聲學模型其辨識結果。其中，由於模型單位為 Triphone，故語料的多寡對模型數量會稍有影響。從實驗結果可得知，將原始訓練語料加入未整理語料訓練出的聲學模型，其音節辨識率為 40.22%；而加入中間值區及高分區語料訓練出的聲學模型，其音節辨識率為 41.21%；而使用高分區語料訓練出的聲學模型其音節辨識率為 44.35%。此結果比加入待整理語料前的辨識率稍低的原因為，在錄音者及句型等方面，原始訓練語料與測試語料較為接近，但待整理語料與這兩者相差較遠，故將待整理語料加入訓練時，辨識率會稍微降低。

表 3. 未經篩選語料其聲學模型之辨識結果

模型單位	模型數量	總 Mixture 數	Deletion	Substitution	Insertion	Correction	Accuracy
Triphone	1,544	417,150	0.86%	53.04%	5.89%	46.10%	40.22%

表 4. 中間值區及高分區語料其聲學模型之辨識結果

模型單位	模型數量	總 Mixture 數	Deletion	Substitution	Insertion	Correction	Accuracy
Triphone	1,539	415,800	0.82%	51.15%	5.82%	47.03%	41.21%

表 5. 高分區語料其聲學模型之辨識結果

模型單位	模型數量	總 Mixture 數	Deletion	Substitution	Insertion	Correction	Accuracy
Triphone	1,527	412,560	0.95%	49.45%	5.25%	49.60%	44.35%

從實驗結果看來，加入語音評分篩選後的語料所訓練出的聲學模型比加入未經篩選語料所訓練出的聲學模型穩定，其音節辨識率差距可達 4.13%。說明了訓練時使用的語料並非數量越多越好，仍然需要考慮語料的品質。而單純加入高分區語料的辨識效果又比加入中、高分區語料來的好。其原因為中間值區本身屬於一個較模糊的地帶，雖然該區域語料量約為高分區的 1.6 倍，但語料內夾雜不良語料的機率也比高分區高。由此，再次證實了訓練聲學模型時語料品質的重要性。所以經由本實驗，證實了語音評分確實能有效的自動篩選掉有問題的語句。

4.6 低分區語料二次檢驗

由於進行語音評分時，如果使用的聲學模型不夠穩定，則容易產生誤評，造成優良的語料但評分分數卻偏低。本實驗的目的即是使用 SVM 分類器，對低分區的語料進行第二次的檢驗，將低分區的語料分類為可用語料和不良語料。同時藉由前處理人工標記的過程中，歸類出音檔問題的原因。

由於我們事先並不曉得語料為可用語料或是不良語料，因此需要先進行人工標記的工作。標記完成後，再將標記好的檔案切分為訓練檔案和測試檔案，以進行分類器的訓練與效能的評估。

於人工標記階段，我們將低分區內的語料以 10 分為間隔，分別對評分分數為 0 至 10 分、10 至 20 分、...、50 至 60 分的前 250 個音檔進行標記的動作，共標記 1500 個音檔。標記過程中，除了標記為可用語料、不良語料外，同時也標記出歸類為不良語料的原因。標記完成後，以 2 比 1 的比例將已標記好的 1500 個檔案分為訓練檔案和測試檔案。表 6 為訓練檔案、測試檔案的相關數據。其中，標記的 1500 個檔案中有 2 個音檔因為標頭檔損壞，無法使用。

表 6. 低分區二次檢驗分類器之訓練、測試檔案相關數據

	訓練檔案		測試檔案	
標記結果	可用語料	不良語料	可用語料	不良語料
音檔數	246	754	100	398
總音檔數	1,000		498	

於 SVM 分類器訓練階段，我們使用 LIBSVM（Lin, 2013）為工具進行分類器的訓練。於特徵擷取時，我們以人工標記過程中觀察到的不良原因為參考，分別取音檔前、後五個音框的音量（1~10 維）、音檔前、後五個音框音量的一階微分（11~20 維）、語音評分分數（21 維）、依語音評分結果，語句內最長音節和最短音節的比例（22 維）、依語音評分結果，切音結果音節數與文本音節數的比例（23 維）、依語音評分結果，音節的平均音量（24 維）、依語音評分結果，每單位時間內的音節數目（25 維），共 25 維特徵。表 7 為 SVM 分類器的參數設定。

表 7. SVM 分類器參數設定

參數項目	**設定內容**
Cross validation	5-fold
Kernel function	RBF kernel
C	8.0
γ	0.125
Feature dimension	25 dimension

於人工標記的過程中，我們將不良語料其不良的原因歸為 14 種類型，其中一個音檔內可能出現多種錯誤類型。表 8 列出了 1500 個標記檔案的不良原因。由該統計結果看來，標記的音檔內最常見的標記為可用語料、音檔片段切音有誤、音檔音節數目小於文本音節數目、有雜音...等。

表 8. 標記音檔的不良原因及出現次數

不良原因	出現次數(百分比)	不良原因	出現次數(百分比)
可用語料	346 (23.07%)	文本和聽打結果不同	80 (5.33%)
音檔片段切音有誤	333 (22.20%)	音節發音錯誤	62 (4.13%)
音檔音節數小於文本音節數	231 (15.40%)	有爆破音	41 (2.73%)
有雜音	177 (11.80%)	說話速度過快	18 (1.20%)
空白音檔	166 (11.07%)	異常停頓	12 (0.80%)
音檔內容與文本內容不符合	125 (8.33%)	音檔有問題	7 (0.47%)
音量過小	90 (6.00%)	有笑聲	3 (0.20%)

底下列出了上述不良原因的詳細說明：

A. 可用語料：標記為可用語料的語料其音檔內容與文本沒有異狀，為可用的語料。圖 5 為文本內容為「我答應了/ ghua-da-ing-liau」的波形圖，該錄音音檔的內容與文本內容符合，但其評分分數卻低於 60；而造成其低分的原因為音素強制對位時有誤，產生誤評。

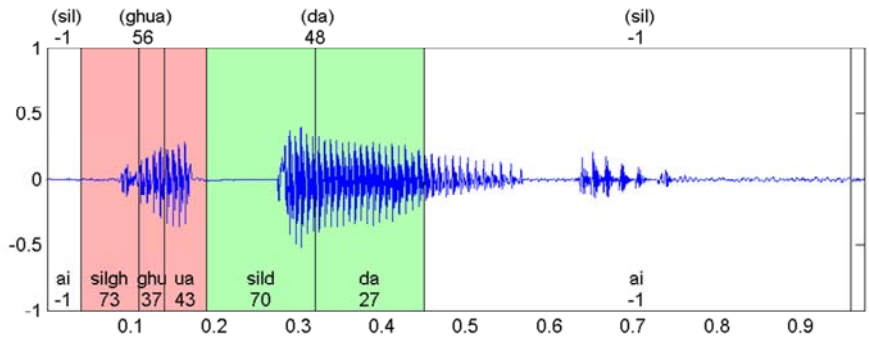


圖 5. 標記為可用語料之波形圖

B. 音檔片段切音有誤：於標記錄音音檔片段的開頭與結尾時，音檔的開頭或結尾恰好切割在音節的中間，造成音檔內前、後音節發音的不完整。以圖 6 為例，該錄音音檔內容為「著會凍真真正正 e 組織/ dier- e- dang- zin- zin- ziann- ziann- e- zo- zit」，但在標記該錄音音檔片段的開頭時，卻切割在音節 dier 中間，造成部分音素的發音遺失。

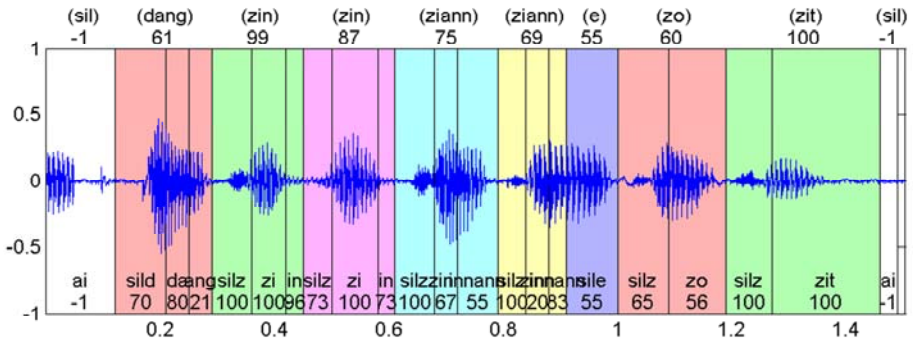


圖 6. 音檔片段切音有誤之波形圖

C. 音檔音節數小於文本音節數：文本內部分音節在錄音音檔內沒有發音，例如文本內容為「有一寡人出門遊覽/u-zit-gua-lang-cut-mng-iu-lam」，但音檔內容卻只發出「寡人出門遊覽/ gua-lang-cut-mng-iu-lam」，少了「有一/u-zit」兩個音節。

D. 有雜音：錄音音檔內，人聲裡頭夾雜著雜音。

E. 空白音檔：錄音音檔內沒有人聲，只有爆破音或純雜音。

F. 音檔內容與文本內容不符合：錄音音檔的內容和真正文本內容不符合，例如文本內容為「東吳大學嘛 m 是東湖大學 dang-gho-dai-hak-ma-m-si-dang-o-dai-hak」，但音檔內容為「東吳大學嘛 m 是嘛 m 是東湖大學/dang-gho-dai-hak-ma-m-si-ma-m-si-dang-o-dai-hak」，多了「嘛 m 是/ma-m-si」三個音節，與文本內容部分不符合。

G. 音量過小：錄音音檔的音量過小。

H. 文本和聽打結果不同：文本的內容和聽打的內容不同，例如文本內容為「殘殺 in-e 族人」，但聽打內容為「zan-sat-zok-rin」，沒拚出「in-e」兩個音節；而錄音音檔又為依據文本內容錄製而成，造成音檔內容和聽打結果對應上有問題。

I. 音節發音錯誤：錄音音檔內部分音節的發音有誤。

J. 有爆破音：錄音過程中，麥克風出現爆音狀況。

K. 說話速度過快：錄音音檔內的說話速度過快，人耳無法辨識所錄製的內容。

L. 異常停頓：錄音音檔內出現異常的停頓，或是某個音節的尾音拖的特別長。

M. 音檔有問題：錄音音檔的標頭檔有問題，或是格式有誤，造成無法讀取音檔。

N. 有笑聲：錄音音檔內夾雜笑聲。

使用由訓練檔案進行 5-fold CV（Cross validation）所得到的最佳參數：C=8.0、γ =0.125，其分類器對測試檔案分類結果的辨識率（Accuracy）為 82.33%。底下為辨識率的計算公式：

$$Accuracy = \frac{\text{分類正確檔案數目}}{\text{測試檔案總數目}} \times 100\%$$

表 9 為本實驗中分類器對測試檔案的混淆矩陣（Confusion matrix）。從表中我們可以求出該分類器針對找出不良語料為目的的錯誤接受率（False Accept Rate, FAR）為 70%；錯誤拒絕率（False Reject Rate, FRR）為 4.52%。以剔除不良語料為目的的狀況下，使用該分類器確實能有效的篩選掉不良語料。

表 9. 混淆矩陣

		預測值	
		可用語料	不良語料
實際值	可用語料	30	70
	不良語料	18	380

藉由 SVM 分類器，我們可以對低分區內所有的語料進行分類，將語料分類為可用語料和不良語料，其分類結果可用語料為 8570 個、不良語料為 33398 個。由實驗的結果，我們只需對被分類為可用語料的音檔進行人工檢驗，從中挑選出實際可用的語料。

同時，我們使用該分類器對高分區語料進行分類，發現被歸類為不良語料的音檔內容，確實也有出現上述人工檢驗觀察到的不良原因。由此結果，說明了我們能藉由該分類器觀察到語音評分無法找出的錯誤。因此，我們可更進一步的對被歸類為不良語料的音檔進行人工檢驗，將有問題的語料剔除。

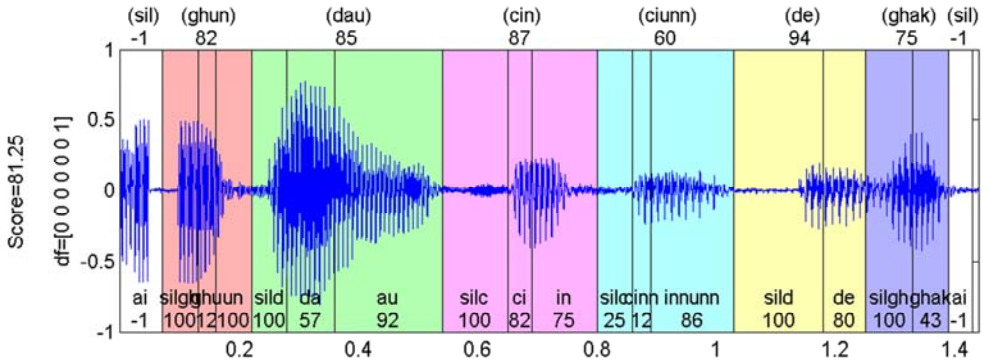


圖 7. 高分區語料音檔片段切音有誤

圖 7 為高分區語料音檔片段切音有誤範例，該錄音音檔的文本內容為「阮兜親像地獄/ghun-dau-cin-ciunn-de-ghak」，語音評分分數為 81.25，但在切割音檔片段時，第一個音節「阮/ghun」的部分音素卻被切掉，造成音節不完整。

5. 結論與未來研究方向

5.1 結論

本論文藉由參數的調整，使用效果最佳的基礎聲學模型對未整理的語料進行語音評分，並依照評分後的分數將語料分為低分區、中間值及高分區，其中我們對低分區的語料進行更深入的觀察。

於未經篩選語料與經篩選語料之聲學模型訓練的實驗結果，加入未篩選語料所產生的聲學模型，其音節辨識率為 40.22%；而加入篩選後語料所產生的聲學模型，其音節辨識率最佳為 44.35%；其辨識率的差別可達 4.13%。證實了藉由語音評分確實能有效的自動過濾掉有問題的語句。

而在觀測低分區語料過程中，我們將語料的不良類型歸類為 14 種，並以語料不良原因為特徵，進行 SVM 分類器的訓練，其辨識效果為 82.33%；我們能利用該分類器對低分區語料再挑選出可用語料，對中、高分區語料再剔除掉不良語料，降低語音評分誤判的機會。同時，由於該分類器僅考慮語料不良的原因，並沒有針對特定語言，因此可適用於各種語言做語料檢測使用。

5.2 未來研究方向

在語音評分部分，聲學模型穩定的程度會影響評分的結果。因此，如何在基礎聲學模型訓練階段，藉由參數的調整產生辨識效果較好的聲學模型會是要點。或許可以試著將待整理語料切成數分，依序對切割的語料進行語音評分、剔除不良語料、加入訓練語料進行聲學模型訓練，並反覆此一過程。

在效能評估部分，目前的測試語料為 ForSD-TW01、ForSD-TW02 測試語料，其中內容多為短詞，而 ForSD-TW03 語料多為長句。可以試著從高分區語料抽取部分語料加入測試語料，增加效能評估的客觀性。

在語料整理部分，雖然 ForSD-TW01 與 ForSD-TW02 是經過人工整理的語料，但也可以利用此研究解果，對語料進行重新分析，以得到更穩定的訓練結果。

致謝

本論文經費來源由國科會計畫 NSC 99-2221-E-007 -049 -MY3，以及 NSC 99-2221-E-182-029-MY3 所提供。

參考文獻

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences, *IEEE International Conference on Acoustics*, 1980.

Lin, C.-J. (2013). http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html, 2013

Lyu, R.-y., Liang, M.-s., & Chiang, Y.-c. (2004). Toward Construction A Multilingual Speech Corpus for Taiwanese (Min-nan), Hakka, and Mandarin, *International Journal of Computational Linguistics and Chinese Language Processing*, *9*(2), 1-12.

Young, S. (2009). The HTK Book version 3.4, Microsoft Corporation, 2009.

朱晴蕾、呂道誠、呂仁園(2010)。混合語言之語音的語言辨認。*ISCSLP*。

李俊毅(2002)。*語音評分*。清華大學研究所碩士論文，新竹。

陳宏瑞(2011)。*使用多重聲學模型以改良台語語音評分*。清華大學研究所碩士論文，新竹。

黃武顯(2007)。*基於 32 位元整數運算處理器之華語語音評分的改良與研究*。清華大學研究所碩士論文，新竹。

廖子宇、呂仁園、高明達、江永進、張智星(2012)。台語文字與語音語料庫之建置，*ROCLING 2012*，102-111。

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 97-114

基於音段式 LMR 對映之語音轉換方法的改進

Improving of Segmental LMR-Mapping Based Voice Conversion Method

古鴻炎*、張家維*

Hung-Yan Gu and Jia-Wei Chang

摘要

基於線性多變量迴歸(linear multivariate regression, LMR)頻譜對映之語音轉換方法，轉換出的頻譜包絡仍然存在過度平滑(over smoothing)的現象，因此本論文研究在音段式 LMR 頻譜對映之前加入直方圖等化(HEQ)的處理，並且在 LMR 頻譜對映之後加入目標音框挑選的處理，希望藉以提升轉換出語音的品質。在此，直方圖等化處理包含兩個步驟，首先是把離散倒頻譜係數(DCC)轉換成主成分分析(PCA)係數，接者把 PCA 係數轉換成累積密度函數(CDF)係數；目標音框挑選則是依據一個音框的音段類別編號、及 LMR 對映出的 DCC 向量，到目標語者相同音段類別所收集的音框群中，去搜尋出距離較小的目標語者 DCC 向量、並且取代原先對映出的 DCC 向量，如此以避免發生頻譜包絡之過度平滑現象。對於直方圖等化與目標音框挑選，我們以外部平行語料(未參加模型參數訓練)來量測語音轉換之平均 DCC 誤差，當加入直方圖等化後會使誤差值變大一些，而當加入目標音框挑選後則會使誤差值變大得更多。不過，VR (variance ratio)值量測及主觀聽測的結果卻是相反的方向，亦即直方圖等化可使語音品質提升一些，而目標音框挑選則可使語音品質獲得更為明顯的提升。這種誤差距離值和語音品質聽測之間的不一致性，我們設法去尋找了它的原因，所找到的一個理由在內文裡說明。

關鍵詞：語音轉換、線性多變量迴歸、直方圖等化、目標音框挑選、離散倒頻譜係數

*國立臺灣科技大學資訊工程系 Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology
E-mail: {guhy, m9815064}@mail.ntust.edu.tw

Abstract

Spectral over-smoothing is still observable in the converted spectral envelope when linear multivariate regression (LMR) based spectrum mapping is adopted to convert voice. Therefore, in this paper, we study to place a histogram-equalization (HEQ) module immediately before LMR based mapping and to place a target frame selection (TFS) module immediately after LMR based mapping. These two modules are intended to promote the quality of the converted voice. Here, HEQ processing includes the two steps: (a) transform discrete cepstral coefficients (DCC) into principal component analysis (PCA) coefficients; (b) transform PCA coefficients into cumulated density function (CDF) coefficients. As to TFS, an input frame is first processed to obtain its converted DCC and its segment-class number. Then, the group of target-speaker frames corresponding to the same segment-class number is searched to find a target frame whose DCC are sufficiently close to the converted DCC. Next, the converted DCC are replaced by the DCC of the target frame found. In experimental evaluation, the outside parallel sentences (not used in model-parameter training) are used to measure average cepstral distances (ACD) between the converted DCC and the target DCC. When the HEQ module is added, the value of ACD would be increased a little. Furthermore, the value of ACD would be apparently increased when the TFS module is added. Nevertheless, according to the measured VR (variance ratio) values and the scores of subjective listening tests, the quality of the converted voice will become better when HEQ is added, and become much better when TFS is added. As to the reasons for why the measured ACD values and the perceived converted-voice qualities are inconsistent, we have found one possible cause which can explain why this inconsistency may occur.

Keywords: Voice Conversion, Linear Multivariate Regression, Histogram Equalization, Target Frame Selection, Discrete Cepstral Coefficients.

1. 緒論

把一個來源語者(source speaker)的語音轉換成另一個目標語者(target speaker)的語音，這種處理稱為語音轉換(voice conversion) (Abe *et al*., 1988; Valbret *et al*., 1992; Stylianou *et al*., 1998)，語音轉換可應用於銜接語音合成處理，以獲得多樣性的合成語音音色。去年我們曾嘗試以線性多變量迴歸(linear multivariate regression, LMR)來建構一種頻譜對映(mapping)的機制(古鴻炎等，2012)，然後用於作語音轉換，希望藉以改進傳統上基於高斯混合模型(Gaussian mixture model, GMM)之頻譜對映機制(Stylianou *et al*., 1998)常遇到的一個問題，就是轉換出的頻譜包絡(spectral envelope)會發生過度平滑(over smoothing)的現象。我們經由實驗發現，音段式(segmental) LMR 頻譜對映機制不僅在平均轉換誤差

上可以比傳統 GMM 頻譜對映機制獲得一些改進，並且轉換出語音的音質也比傳統 GMM 對映的稍好一些。不過，整體而言音段式 LMR 對映機制所轉換出的頻譜包絡，仍然存在有過度平滑的現象，而使得轉換出的語音仍然令人覺得有一些悶悶的，而不像真人發音那樣清晰。前面提到的”音段式” LMR，是指我們對於訓練語料中不同的韻母、有聲聲母(如/m, n, l, r/)的語音要分別去建立各自的 LMR 矩陣，這是為了避免發生一對多(one to many)對映的問題(Godoy *et al*., 2009)，而造成某些相鄰的音框之間，相鄰音框所轉換出的頻譜卻出現劇烈的頻譜形狀差異(即頻譜不連續)，而不連續的頻譜很可能導致怪音(artifact sound)被合成出來。

去年我們研究的基於 LMR 頻譜對映之語音轉換系統，其主要的處理流程如圖 1 所示，來源語者發出的語音先分割成一序列的音框，然後對各個音框去估計它的 40 階 DCC (discrete cepstral coefficients) 倒頻譜係數(Cappé & Moulines, 1996; Gu & Tsai, 2009)及偵測出基頻值；接著，依據各音框的 DCC 係數，可作有聲聲母與韻母的音段(segment)偵測，先前我們曾提出一種基於音段式 GMM 與最大似然率(maximum likelihood)的音段自動偵測方法(Gu & Tsai, 2011)，實驗顯示即使挑選到錯誤但近似的音段，也仍可轉換出正確的語音，由於在此我們把焦點放在 LMR 對映方塊，所以音段偵測方塊暫時以讀取標記(label)檔案的方式來進行；LMR 對映就是把 LMR 矩陣乘以輸入的 DCC 向量而求得輸出的 DCC 向量，至於 LMR 矩陣的訓練方法，則可參考我們去年的論文(古鴻炎等，2012)；之後，LMR 對映出的 DCC 向量、及以平均值與標準差轉換出的基頻值，兩者就可送給 HNM (harmonic plus noise model)語音再合成方塊，以合成出轉換後的語音信號，關於使用諧波加雜音模型(HNM)作語音信號合成的細節，可參考前人的論文(Stylianou, 1996; Gu & Tsai, 2009)。

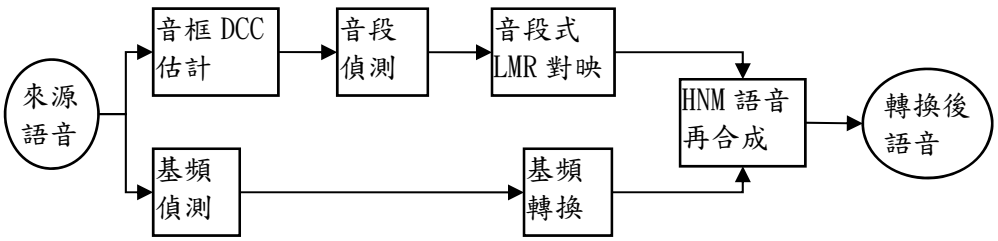


圖 1. 基於 LMR 頻譜對映之語音轉換的主要處理流程

為了提升轉換出語音的音質，我們開始思考在 GMM 對映與 LMR 對映之外，是否還有其它種類的對映方法？後來我們想到一種似乎可行的頻譜對映方法，就是以直方圖等化(histogram equalization, HEQ)來取代 LMR 對映。直方圖等化雖然起源於影像處理領域，但是近年來被應用於語音辨識領域(Torre *et al*., 2005; Lin *et al*., 2007)，用以降低環境噪音造成的訓練語音和測試語音之間的頻譜不匹配(mismatch)問題，而使得辨識率獲得了明顯的改進。有鑑於此，我們覺得在語音轉換的問題上，來源與目標語者之間有著差異的頻譜形狀而呈現出差異的音色，這可想像是因為來源語音通過了某一種特殊的通訊通道而使得其頻譜形狀被轉換成目標語音的形狀，以致於造成來源與目標語音之間的頻譜

不匹配。因此在觀念上應可應用直方圖等化的處理，來模仿前述的通訊通道之特性，以把來源語音的頻譜轉變成目標語音的頻譜，所以我們構想了如圖 2 所示的基於直方圖等化之語音轉換的處理流程。

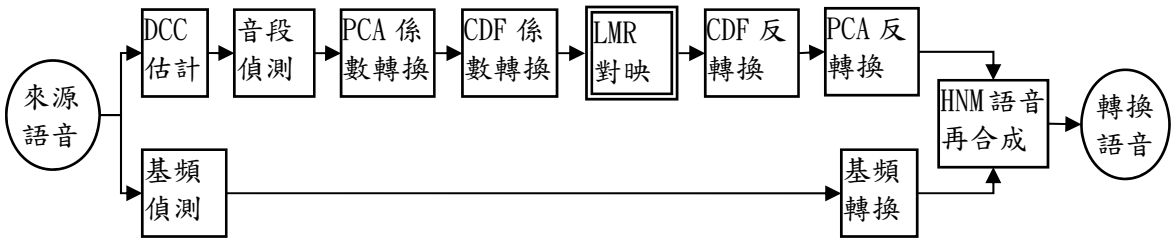


圖 2. 基於直方圖等化及 LMR 對映之語音轉換流程

在圖 2 的處理流程中，我們不直接拿 DCC 係數去作直方圖等化，即計算 CDF (cumulative density function)係數，我們的觀點是，一個音框各維度的 DCC 係數之間有明顯的相關性存在，而直方圖等化卻是對特徵的各維度獨立去進行，這恐將降低直方圖等化的功用，因此我們決定對各個音段類別所屬的音框 DCC 向量先進行主成分分析 (principle component analysis, PCA) (Jolliffe, 2002)，再依據主成分向量把 DCC 係數轉換成 PCA 係數，如此將可讓一個音框各維度的 PCA 係數之間變成是獨立的。此外，圖 2 中的 LMR 對映方塊，一開始時是未被加入的，不過經由初步的測試實驗發現，當沒有作 LMR 對映的處理時，轉換出語音的音色雖可達到部分近似目標語者的音色，但是仍存在明顯的音色落差，因此我們遂決定把 LMR 對映方塊加上去，以提升音色相似度。

對於圖 1 處理流程會遇到的頻譜包絡過於平滑的情況，雖然前人曾經提出至少兩種的改進方法，即全域變異數(global variance, GV)之變異數調整方法(Toda *et al*., 2007)、和頻率軸校正(frequency warping)的方法(Erro *et al*., 2010; Godoy *et al*., 2012)，但是 Toda 等人的方法(Toda *et al*., 2007)和 Erro 等人的方法(Erro *et al*., 2010)都是針對 GMM 對映所設計的，而 Godoy 等人的方法(Godoy *et al*., 2012)則不是針對 GMM 對映或 LMR 對映所設計的。因此我們就從另外一個方向去思考圖 1 流程的改進作法，在參考 Dutoit 等人的論文(Dutoit *et al*., 2007)之後，我們想到的一個作法是，在圖 1”LMR 對映”方塊之後插入”目標音框挑選”的方塊。既然經過 GMM 或 LMR 對映得到的頻譜包絡會發生過度平滑的現象，那麼就不要直接拿 LMR 對映得到的頻譜係數去作語音再合成處理，而要改變成依據**來源音框**(來源語者音框)的音段類別、及對映出的頻譜特徵係數(如 DCC)，去對同一音段類別的**目標音框**(目標語者音框)群作搜尋，以找出頻譜特徵很相似(或距離很小)的目標音框，然後把找出的目標音框的頻譜係數拿去取代對映出的頻譜係數，如此就可免除發生頻譜包絡過度平滑的問題。由於被找出的目標音框不是經由頻譜對映而得到，所以在此也稱它為**真實音框**(真實語音的音框)，此外，目標音框的音段分類與收集是在訓練階段進行，所以轉換階段就可直接去作搜尋與挑選。當圖 1 插入”目標音框挑選”的方塊之後，一種基於 LMR 對映及目標音框挑選之改進的語音轉換處理流程就如圖 3 所示。

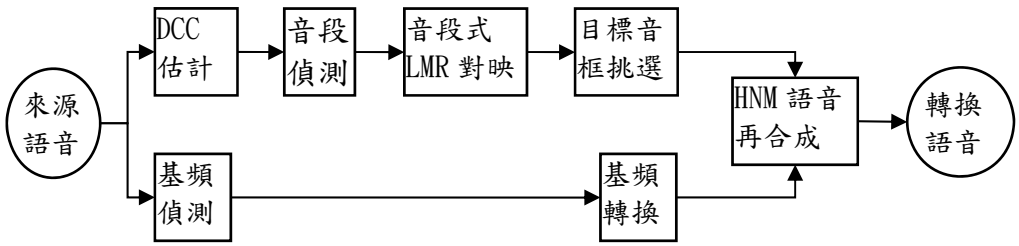


圖 3. 基於 LMR 對映及目標音框挑選之語音轉換流程

除了分別去加入直方圖等化和目標音框挑選的處理動作，我們也考慮了另外一種處理流程，就是同時把這兩種處理動作加入圖 1 的處理流程中，如此轉換出的語音是否可以獲得最好的音色相似度及語音品質？這將會第四節中作實驗探討。此外，在圖 1、2、3 裡都出現的 DCC 估計之方塊，表示我們採用離散倒頻譜係數(DCC)(Cappé & Moulines, 1996; Gu & Tsai, 2009)作為頻譜特徵參數，並且階數設為 40 階，即一個音框要計算出 c_0, c_1, c_2, …, c_{40} 等 41 個係數，但是只拿 c_1, c_2, …, c_{40} 去作頻譜轉換的處理。當轉換出各個音框的 DCC 係數之後，我們就可依據各音框的 DCC 係數去計算出頻譜包絡(Cappé & Moulines, 1996; Gu & Tsai, 2009)，然後再依據頻譜包絡、轉換出的基頻值，去設定該音框的 HNM 模型之諧波參數和雜音參數(Gu & Tsai, 2009; Stylianou, 1996)，之後就可拿這些參數去合成出語音信號(Gu & Tsai, 2009; Stylianou, 1996)。

2. PCA係數轉換與直方圖等化

若要依據圖 2 的處理流程來進行語音轉換的處理，則各音框在求取 DCC 係數之後，接著就要作 PCA 係數轉換和 CDF 係數轉換的動作，然後在 LMR 對映之後，還要作 PCA 反轉換和 CDF 反轉換的動作，以將頻譜特徵還原成 DCC 係數。因此，在這一節就說明 PCA 係數轉換和 CDF 係數轉換的細節。

2.1 PCA係數轉換

要能夠把一個來源音框的 DCC 係數轉換成 PCA 係數，則在訓練階段就要先對來源語者各個音段類別所收集到的 DCC 向量作 PCA 分析，以求取來源語者各個音段類別的主成分向量。相對地，要能夠把一個 LMR 對映後音框的 PCA 係數反轉換成 DCC 係數，則在訓練階段也要先對目標語者各個音段類別所收集到的 DCC 向量作 PCA 分析，以求取目標語者各個音段類別的主成分向量。然而關於 PCA 分析的作法，我們曾經思索的一個疑問是，雖然直覺上我們會認為來源音框和目標音框應該要分開去收集，並且分開去作 PCA 分析以求取各自的主成分向量，但是，為什麼不能夠把同一音段類別的來源音框和目標音框放在一起去作 PCA 分析？又為什麼不讓來源音框和目標音框共用一組主成分向量呢？因此，我們將以實驗評估的方式來探討此一疑問。

PCA 分析是由 K. Pearson 於 1901 年提出，在 1933 年時再由 H. Hotelling 加以發展

(Hotelling, 1933)。PCA 轉換是一種正交變換，它可以將原本維度間相關的原始數據轉換成各維度獨立的新數據，再者作 PCA 轉換後的新數據，它們的總變異數(variance)與原始數據集的總變異數相等，也就是說 PCA 轉換能保留原始數據的訊息。

2.1.1 主成分分析

對於某一音段類別的所有訓練語音作音框切割及求取 DCC 係數，以建立一個 40 維 DCC 係數的數據集，接著再對這個數據集作 PCA 分析以得到該種音段的主成分向量，詳細的分析流程如下:

(a) 假設某一音段類別的訓練語音總共可切成 M 個音框，而每個音框經由計算可得到一個 DCC 係數的向量，然後把全部音框的 DCC 向量並列成各欄(column)的方式，表示成大小為 $L\times M$ 的矩陣 $\Gamma=[\Gamma_1,\Gamma_2,\ldots,\Gamma_M]$ ，其中 L 表示 DCC 係數的階數，M 的值大於 L。

(b) 接著求出這 M 個音框之 DCC 向量的平均向量 Ψ ， Ψ 代表著這 M 個音框共有的 DCC 向量成分。

(c) 將第 i 個音框的 DCC 向量作標準化，即減去平均向量 Ψ ，而得到一個差值向量 Φ_i 。

(d) 使用所有的差值向量 Φ_i ，來計算出一個共變異矩陣 Λ 。

$$\Lambda=\sum_{i=1}^{M}\Phi_i\Phi_i^{\mathrm{T}} \tag{1}$$

(e) 對矩陣 Λ 求其特徵值(eigen value) λ_i 與特徵向量(eigen vector) γ_i 。

$$\Lambda\cdot\gamma_i=\lambda_i\cdot\gamma_i,\ \ i=1,2,\ldots,L \tag{2}$$

(f) 求得特徵向量 γ_i 後，進一步對 γ_i 作正規化，以取得 L 個主成分基底向量 μ_i 。

$$\begin{aligned}\upsilon_i&=\sqrt{(\gamma_{i1})^2+(\gamma_{i2})^2+\cdots+(\gamma_{iL})^2},\ \ i=1,2,\ldots L\\ \mu_i&=\left[\frac{\gamma_{i1}}{\upsilon_i},\frac{\gamma_{i2}}{\upsilon_i},\ldots\frac{\gamma_{iL}}{\upsilon_i}\right]^{\mathrm{T}},\ \ i=1,2,\ldots,L\end{aligned} \tag{3}$$

2.1.2 主成分係數轉換

當我們對某一個音段類別做完主成分分析後，就可得到該類別的 DCC 平均向量 Ψ 、L 個主成分基底向量 μ_i 。接著，要把各個音框的 DCC 係數轉換成 PCA 係數，首先把一個音框的 DCC 向量 Γ_i 減去 DCC 平均向量 Ψ 而得到差值向量 Φ_i ，再將 Φ_i 分別投影到各個主成分基底向量 μ_j ，投影公式為:

$$\omega_{ij}=\mu_j^{\mathrm{T}}\cdot\Phi_i,\ \ j=1,2,\ldots,L \tag{4}$$

如此就可得到 DCC 向量 Γ_i 的 L 個主成分係數(亦稱為 PCA 係數)，再用以形成 L 維度的主成分係數(PCA 係數)之向量：

$$\Omega_i=\left[\omega_{i1},\omega_{i2},\ldots\omega_{iL}\right]^{\mathrm{T}} \tag{5}$$

2.1.3 主成分係數反轉換

在圖 2 的處理流程中，“PCA 反轉換”方塊就是要將轉換後的 PCA 係數還原到 DCC 係數的向量空間，以得到轉換後的 DCC 係數。假設我們取得一序列音框的 PCA 係數之向量，則首先要知道各個音框分別所屬的音段類別，如此才能對各個音框分別去作還原，令第 i 個音框所屬的音段類別之編號為 k，則我們就要取出訓練階段目標語者在第 k 類音段所計算出的 DCC 平均向量 Ψ、及 L 個主成分基底向量 μ_j，來把轉換後的 PCA 向量 Ω_i 還原成轉換後的 DCC 向量 Γ_i ，如公式(6)所示:

$$\Gamma_i = \Psi + \sum_{j=1}^{L} \mu_j \cdot \omega_{ij} \tag{6}$$

2.2 直方圖等化

直方圖等化所指的是圖 2 流程裡“CDF 係數轉換” 與 “CDF 反轉換”兩方塊的處理。要能夠把一個來源音框的 PCA 係數轉換成 CDF 係數，則在訓練階段就要先對來源語者各個音段類別所收集到的 PCA 向量作 HEQ 分析，以建造來源語者各個音段類別的 HEQ 表格。相對地，要能夠把一個 LMR 對映後音框的 CDF 係數反轉換成 PCA 係數，則在訓練階段也要先對目標語者各個音段類別所收集到的 PCA 向量作 HEQ 分析，以建造目標語者各個音段類別的 HEQ 表格。這裡提到 HEQ 表格，意謂我們採取基本的表格法來建立 PCA 係數和 CDF 係數之間的直方圖等化關係。

2.2.1 HEQ表格建造

選定一個來源(或目標)語者的音段類別，令該類別裡收集到的音框總數為 M，則將 M 個維度為 L 的 PCA 係數向量作為輸入資料，依照下列步驟來建造 HEQ 表格:

(a) 令區間數為 N，並且對各個維度 i，i=1, 2, …, L，分別作下列步驟的處理。

(b) 將 M 個音框中所有位於第 i 維度的 PCA 係數挑出，然後依係數值作由小到大之排序，排序後則把 M 個 PCA 係數依順序且平均地分配到 N 個區間。

(c) 區間編號 j 從 1 變到 N，對於第 j 個區間內的 PCA 係數，挑選排序位於中間(median)的 PCA 係數數值，然後記錄該 PCA 係數值為 Fp_i^j，並且記錄其對應的 CDF 值為 Fc_i^j ，CDF 值就是該 PCA 係數在全體(M 個)係數排序中的順序值除以 M。

(d) 記錄第 i 維度 PCA 係數的最大值為 Fp_i^{N+1} ，且記錄其對應的 CDF 值為 $Fc_i^{N+1}=1$ ；此外，記錄第 i 維度 PCA 係數的最小值為 Fp_i^0 ，且記錄其對應的 CDF 值為 $Fc_i^0=\frac{1}{M}$ 。

當所有維度都完成上述步驟，則該音段類別的 HEQ 表格就建立完成了。對於區間數 N 的選擇，我們在評估實驗裡嘗試了 32, 64, 128 等三種。HEQ 表格建造後的外觀為何？在此舉一個簡化的例子，設有 20 個音框，PCA 係數向量維度為 1 維，且 PCA 係數序列

排序後為 1, 2, …, 20，若設定的區間數為 N=4，則建造出的 HEQ 表格如下所列。

表 1.一個簡化的 HEQ 表格例子

區間 j	0	1	2	3	4	5
Fp_1^j	1(min)	3	8	13	18	20(max)
Fc_1^j	0.05	0.15	0.4	0.65	0.9	1

2.2.2 CDF係數轉換

假設有一個音框的 PCA 係數向量 $P=[P_1,P_2,\cdots,P_L]$要被轉換，而該音框所屬的音段類別資訊，已經在圖 2 的“音段偵測”方塊決定出來，所以我們可以取出該音段類別的來源音框所訓練出的 HEQ 表格，然後以線性內插的方式來計算出該音框的 CDF 係數向量 $Q=[Q_1,Q_2,\cdots,Q_L]$，線性內插之公式如下：

$$Q_i = Fc_i^j + \left(Fc_i^{j+1} - Fc_i^j\right)\cdot\left[\frac{\left(P_i - Fp_i^j\right)}{\left(Fp_i^{j+1} - Fp_i^j\right)}\right],\ i=1,2,\ldots L. \tag{7}$$

公式(7)中 i 表示維度編號，Fp_i^j、Fc_i^j 分別為 HEQ 表格裡所記錄的第 j 區間的 PCA 係數值、CDF 值，並且假設我們已作過搜尋而得知 P_i 的值落於 Fp_i^j 與 Fp_i^{j+1} 之間。

2.2.3 CDF反轉換

假設有一個音框的 CDF 向量 $Q=[Q_1,Q_2,\cdots,Q_L]$要被反轉換成 PCA 係數向量，而該音框所屬的音段類別資訊，已經在圖 2 的“音段偵測”方塊決定出來，所以我們可以取出該音段類別的目標音框所訓練出的 HEQ 表格，然後以線性內插的方式來計算出該音框的 PCA 係數向量 $P=[P_1,P_2,\cdots,P_L]$，線性內插之公式如下：

$$P_i = Fp_i^j + \left(Fp_i^{j+1} - Fp_i^j\right)\cdot\left[\frac{\left(Q_i - Fc_i^j\right)}{\left(Fc_i^{j+1} - Fc_i^j\right)}\right],\ i=1,2,\ldots L. \tag{8}$$

公式(8)中 i 表示維度編號，Fp_i^j、Fc_i^j 分別為 HEQ 表格裡所記錄的第 j 區間的 PCA 係數值、CDF 值，並且假設我們已作過搜尋而得知 Q_i 的值落於 Fc_i^j 與 Fc_i^{j+1} 之間。

3. 目標音框挑選

在訓練階段，我們可預先把目標語者的訓練語音依據標示檔的資訊拿去作音段分類，並且對各種音段分別作音框的收集，之後在轉換階段，就可依據所偵測出的音段代號去取出對應的音框集，再依據所轉換出的 DCC 向量去作真實音框的搜尋與挑選。

令 Y_1, Y_2, …, Y_T 是一序列 T 個被轉換出的 DCC 向量，轉換可以是直接經由圖 3 “LMR 對映”方塊得到，或是 LMR 對映後再作 CDF 反轉換與 PCA 反轉換而得到(圖 2

的流程)。為了改進轉換出的語音的品質，所以在此要依據 Y_t 及其對應的音段類別代號 $I(t)$，從目標語者的 $I(t)$音段的音框集去挑選出一個非常靠近 Y_t 的真實音框的 DCC 向量 Z_t。然而挑選 Z_t 的準則，不僅只是考慮 Y_t 與 Z_t 的匹配距離 dist(Y_t, Z_t)，也要考慮相鄰音框之間的連接距離 dist(Z_{t-1}, Z_t)，以避免發生頻譜之不連續，而導致怪音被合成出來。在本論文裡，距離函數 dist(·, ·)是量測幾何距離。除了依循 Dutoit 等人的論文(Dutoit *et al*., 2007)去考慮音框連接的距離，我們還更加考慮了另外一種距離量測，即動態頻譜(dynamic spectral)距離，以把轉換出的相鄰兩 DCC 向量之間的頻譜改變 $\Delta Y_t = Y_t - Y_{t-1}$ 納入考慮。在此，動態頻譜距離是量測 dist(ΔY_t, ΔZ_t)，而 $\Delta Z_t = Z_t - Z_{t-1}$ 表示相鄰兩個挑選出的 DCC 向量之間的頻譜改變量。

依據前述的三種距離，即匹配距離、連接距離與動態頻譜距離，我們發展了一種基於動態規劃的演算法來作目標音框的挑選。首先，對於各個轉換出的 DCC 向量 Y_t，我們依其音段編號 $I(t)$，從第 $I(t)$個音框集去尋找出 K 個最靠近 Y_t (即離 Y_t 的距離最小)的真實音框的 DCC 向量，在此 K 的值設為 16。接著，令 $U(t, i)$表示從時刻 1 到時刻 t 的最小的累積距離，而條件是在時刻 t 時所挑選到的目標音框必須是 K 個中的第 i 個。如此，我們就可得到如下的遞迴公式:

$$U(t,i) = \min_{0 \le j < K} \left[U(t-1, j) + \alpha \cdot dist(Z_{t-1}^{j}, Z_t^{i}) + \alpha \cdot dist(Y_t - Y_{t-1}, Z_t^{i} - Z_{t-1}^{j}) \right] + dist(Y_t, Z_t^{i}), \quad (9)$$

其中 α 是加權常數，我們經過試驗後將它的值設為 0.5， Z_t^i 表示時刻 t 時所尋找出的 K 個音框中的第 i 個音框 DCC 向量。另外，前人論文(Dutoit *et al*., 2007)中曾提到一個技巧，當 Z_t^i 和 Z_{t-1}^j 被檢查出是來自同一次發音的相鄰音框時，就機動地把公式(9)中 α 的值改設為 0，以便優先選取相鄰的目標音框來提升頻譜連接的自然性。在此我們也應用了這個技巧，並且把條件放寬，就是當 Z_t^i 和 Z_{t-1}^j 不是直接相鄰而是存在另一個音框在它們之間，我們也接受此一情況而會把 α 的值機動地改設為 0。

當到達最後時刻 T 時，全部路徑中的最小累積距離 A(T)可以下列公式來計算，

$$A(T) = \min_{0 \le j < K} \left[U(T, j) \right], \quad (10)$$

此外，我們可再作回溯(backtrack)處理，以找出在最佳路徑上各個時刻 t 所選到的目標音框編號 $k(t)$，然後把 t 時刻所選到的第 $k(t)$個目標音框的 DCC 向量，拿去取代被轉換出的 DCC 向量 Y_t。

4. 測試實驗

我們邀請了二位男性和二位女性錄音者，其中二位男性以 MA 和 MB 為代號，而另二位女性則以 FA 和 FB 為代號。請四位錄音者分別到隔音錄音室去錄製 375 句(共 2,926 個音節)之國語平行語料，取樣率設成 22,050Hz，這 375 句的語料中，前 350 句被拿來作模型參數的訓練之用，而剩下的 25 句則保留作為外部測試之用。在此我們實驗了四種語者配對方式，分別是(a)MA 至 MB、(b)MA 至 FA、(c)FA 至 MA、(d)FA 至 FB，這四種配對方式中，前者就當來源語者，而後者則當目標語者。

4.1 語音轉換系統之訓練

首先，我們操作 HTK (HMM tool kit)軟體，經由強制對齊(forced alignment)來作自動標音，把一個語句的各個聲母、韻母的邊界標示出來，然後操作 WaveSurfer 軟體，以檢查自動標記的邊界是否有錯，有錯則作人工更正。接著，依據各個聲、韻母的拼音符號標記和邊界位置，就可作音段切割和分類的動作，我們一共分成 57 類，即 21 類聲母和 36 類韻母。

對於各個語音音框，我們先計算零交越率(ZCR)，以把 ZCR 很高的無聲(unvoiced)音框偵測出來；再使用一種基於自相關函數及 AMDF 的基週偵測方法(Kim *et al*., 1998)，來偵測剩餘音框的音高頻率。之後，把一個語者發音中有聲(voiced)音框偵測出的音高頻率值收集起來，據以算出該語者音高的平均值及標準差，而平均值及標準差就是本論文所使用的音高參數。在此一個音框的長度設為 512 個樣本點(23.2ms)，而音框位移則設為 128 個樣本點(5.8ms)。此外，對於一個音框的頻譜係數，我們使用先前發展的 DCC 估計程式(Gu & Tsai, 2009)來計算出 41 維的 DCC 係數。

在訓練 LMR 對映矩陣之前，我們逐一對各個聲、韻母類別所收集的平行發音音段作 DTW 匹配，以便為來源語者音段所切出的各個音框，去目標語者之平行音段內找出正確的音框來對應。然後，把各個平行音段的音框序列串接起來，就可為一個聲、韻母類別準備好一序列的來源音框和目標音框的 DCC 向量對應組合，(S_i, R_i)，i=1, 2, …, Nr，其中 S_i 表示第 i 個來源音框的 DCC 向量，R_i 表示第 i 個經 DTW 配對到的目標音框的 DCC 向量，Nr 表示此一序列的音框總數。再來，依照所建構系統的結構，若是如圖 3 的流程，則各個聲、韻母類別的一序列的來源與目標音框對應的 DCC 向量組合，就可直接拿去訓練計算 LMR 對映所需的對映矩陣(古鴻炎等，2012)；然而當系統的結構是如圖 2 所示的流程時，則各個聲、韻母類別的 DCC 向量組合序列，(S_i, R_i)，i=1, 2, …, Nr，其中各個組合的 S_i 與 R_i 就必須先作 PCA 係數轉換和 CDF 係數轉換，以形成 CDF 係數的向量組合，然後才拿去訓練 LMR 對映之矩陣。

設 $\tilde{S}$ 、 $\tilde{R}$ 矩陣的定義如下所列，

$$\tilde{S} = \begin{bmatrix} S_1 & S_2 & \cdots & S_{Nr} \\ 1, & 1, & \cdots & 1 \end{bmatrix}, \quad \tilde{R} = \begin{bmatrix} R_1 & R_2 & \cdots & R_{Nr} \\ 1, & 1, & \cdots & 1 \end{bmatrix}, \tag{11}$$

其中各行的 S_i 與 R_i 都被附加一列的常數 1，以增加一個常數項至多變量線性迴歸的各個維度裡，如此，LMR 對映所需的最佳(least squared error)對映矩陣 $\tilde{M}$，就可以下列公式(古鴻炎等，2012)來求得，

$$\tilde{M} = \tilde{R} \cdot \tilde{S}^{\mathrm{t}} \cdot (\tilde{S} \cdot \tilde{S}^{\mathrm{t}})^{-1} . \tag{12}$$

然後，我們就可用矩陣 $\tilde{M}$ 來作 LMR 對映，即令 $[Y^{\mathrm{t}}, 1]^{\mathrm{t}} = \tilde{M} \cdot [X^{\mathrm{t}}, 1]^{\mathrm{t}}$，其中 X 表示一個來源語者音框的 DCC 或 CDF 係數向量，而 Y 表示經由 LMR 對映出的係數向量。

4.2 共用主成分向量之測試

圖 2 的處理流程裡，PCA 係數轉換與 PCA 反轉換兩個處理方塊，若讓兩者共用一組主成分向量是否會比較好？原先不共用主成分向量的情況，表示“PCA 係數轉換”方塊使用的主成分是由**來源**音框作完音段分類後再作 PCA 分析得到，而“PCA 反轉換”方塊使用的主成分則是由**目標**音框作完音段分類後再作 PCA 分析得到；若是共用主成分向量，就表示同一音段類別的來源音框和目標音框要放在一起去作 PCA 分析，以求得共用的一組主成分向量。

我們以量測語音轉換的平均轉換誤差的方式，來比較共用與不共用主成分向量之優劣。在此，我們只拿平行語料最後的 25 句來作語音轉換之外部測試，當一個來源音框經過轉換而得到 DCC 向量之後，我們就可量測此 DCC 向量與對應的目標音框 DCC 向量之間的幾何距離，這樣的距離也稱為轉換誤差，當把全部音框的轉換誤差加總及取平均，就可算出平均的轉換誤差。此外，我們也把圖 2 流程裡的直方圖等化(即 CDF 係數轉換與反轉換)分成三種情況來作實驗，就是分別設定區間的數量 *N* 為 32、64、與 128，經過實驗量測後，我們得到如表 2 所示的平均轉換誤差值。

從表 2 的轉換誤差平均值可以看出，圖 2 中的 PCA 係數轉換與反轉換方塊若是使用共用的 PCA 主成分向量，則平均轉換誤差可從 0.5447 降到 0.5414，這說明了使用共用的 PCA 主成分向量，可以略微提升來源與目標音框之間 PCA 係數的相關性，而稍微減小 LMR 對映的誤差。此外，關於直方圖等化的區間數的設定，依據表 2 的轉換誤差平均值可知，設為 64 區間或 128 區間是沒有差異的。

表 2. 共用與不共用主成分向量之平均轉換誤差

誤差 配對	不共用 PCA 向量			共用 PCA 向量		
	32 區間	64 區間	128 區間	32 區間	64 區間	128 區間
MA=> MB	0.5442	0.5438	0.5442	0.5389	0.5389	0.5389
MA=> FA	0.5159	0.5158	0.5156	0.5155	0.5154	0.5154
FA => MA	0.5387	0.5386	0.5384	0.5369	0.5344	0.5344
FA => FB	0.5807	0.5806	0.5805	0.5773	0.5768	0.5768
平均	**0.5449**	**0.5447**	**0.5447**	**0.5422**	**0.5414**	**0.5414**

4.3 PCA轉換之必要性測試

對於圖 2 的流程裡，加入“PCA 係數轉換”與“PCA 反轉換”方塊是否為必要的？在此我們以量測語音轉換的平均轉換誤差的方式，來比較 PCA 係數轉換加入與不加入的優劣，所用的測試語料和誤差的量測方式，和 4.2 節裡敘述的一樣，亦即使用平行語料最後 25 句來作外部測試，並且量測轉換得到的 DCC 向量與對應的目標音框 DCC 向量之間的幾何距離，再計算全部音框的平均誤差。此外，直方圖等化也分成三種區間數來作實驗，即 32、64、與 128 個區間。經過實驗量測後，我們得到如表 3 所示的平均轉換誤差值，

其中右邊三欄的數值是取自表 2 的右邊三欄。

表 3. 作與不作 PCA 係數轉換之平均轉換誤差

誤差 配對	不作 PCA 係數轉換			作 PCA 係數轉換		
	32 區間	64 區間	128 區間	32 區間	64 區間	128 區間
MA=> MB	0.5454	0.5450	0.5446	0.5389	0.5389	0.5389
MA=> FA	0.5177	0.5172	0.5171	0.5155	0.5154	0.5154
FA => MA	0.5410	0.5402	0.5399	0.5369	0.5344	0.5344
FA => FB	0.5826	0.5825	0.5823	0.5773	0.5768	0.5768
平均	**0.5467**	**0.5462**	**0.5460**	**0.5422**	**0.5414**	**0.5414**

從表 3 的數值可以看出,作 PCA 係數轉換的確可使得語音轉換的誤差平均值下降，在 64 區間直方圖等化的情況下，平均轉換誤差可從 0.5462 降到 0.5414，這說明了直方圖等化之前先作 PCA 係數轉換是有用的、需要的。

4.4 目標音框挑選之轉換誤差

目標音框挑選可用以避免發生頻譜過度平滑的問題，其詳細的作法已在第三節說明。在此我們依據圖 3 之處理流程，測試目標音框挑選是否可以讓語音轉換的平均誤差減少？是否可以比圖 2 處理流程的好？圖 3 流程的語音轉換方法，我們稱為基本型目標音框挑選法，此外，我們也測試了另外一種語音轉換方法，稱為複合型目標音框挑選法，就是在圖 2 流程中“PCA 反轉換”與“HNM 語音再合成”兩方塊之間插入“目標音框挑選”之方塊，至於直方圖等化(CDF 轉換與反轉換)所用的區間數，這裡就設為 64。

對於前述的基本型與複合型目標音框挑選法，我們使用的測試語料和誤差的量測方式，和 4.2 節裡敘述的一樣，亦即使用平行語料最後 25 句來作外部測試，並且量測轉換得到的 DCC 向量與對應的目標音框 DCC 向量之間的幾何距離，再計算全部音框的平均誤差。經過實驗量測後，我們得到如表 4 所示的平均轉換誤差值，由表 4 可知基本型目標音框挑選的轉換誤差平均值會變大成為 0.6029，這明顯比表 3 的 0.5414 增加了許多；再者，複合型目標音框挑選的轉換誤差平均值也變得更大，0.6121。根據這二個變大很多的誤差平均值，直覺上會讓人認為基本型與複合型目標音框挑選法，所轉換出的語音應會在音色相似度和語音品質上衰減很多，然而實際上當我們去聽轉換出的語音時，發現經由基本型或複合型目標音框挑選所轉換出的語音，語音品質卻是會變得更為清晰(應是使用真實音框 DCC 的緣故)，並且音色相似度也沒有衰減。所以，基於量測兩 DCC 向量之間幾何距離的轉換誤差平均值，其數值大小和語音品質之間似乎不是正比例的關係。

表 4. 目標音框挑選之平均轉換誤差

誤差 配對	基本型	複合型
MA=> MB	0.5990	0.6087
MA=> FA	0.5706	0.5791
FA => MA	0.5925	0.6032
FA => FB	0.6493	0.6574
平均	**0.6029**	**0.6121**

前述的**不一致性**情況，即誤差距離變大反而得到更好的語音品質，是什麼原因造成的？為了瞭解其原因，我們就找一些目標音框來觀察它們的頻譜包絡曲線。對於各個目標音框，我們把 LMR 對映出的 DCC 向量、經目標音框挑選得到的 DCC 向量、及該目標音框的 DCC 向量，計算出三者的頻譜包絡曲線並且畫出來作比較，結果我們發現了一個現象可用以解釋前述的不一致性。一個例子如圖 4 所示，圖 4 中的虛線代表/song/音節的一個目標音框的頻譜包絡線，淺灰色實線代表 **LMR 對映**得到的 DCC 向量所算出的頻譜包絡線，深黑色實線則代表**目標音框挑選**得到的 DCC 向量所算出的頻譜包絡線，比較這三條包絡線，我們可發現在橫軸頻率範圍 2,500 Hz 至 4,500 Hz 之間，深黑色實線的形狀比起淺灰色實線的形狀較為接近虛線曲線的共振峰起伏，所以這可以解釋為什麼目標音框挑選能夠改進轉換出語音的品質；此外，在橫軸頻率範圍 5,500 Hz 至 11,000 Hz 之間，淺灰色實線會比深黑色實線更為靠近虛線曲線，所以這可以解釋為什麼 LMR 對映所導入的轉換誤差，會比**目標音框挑選**所導入的轉換誤差來得小。

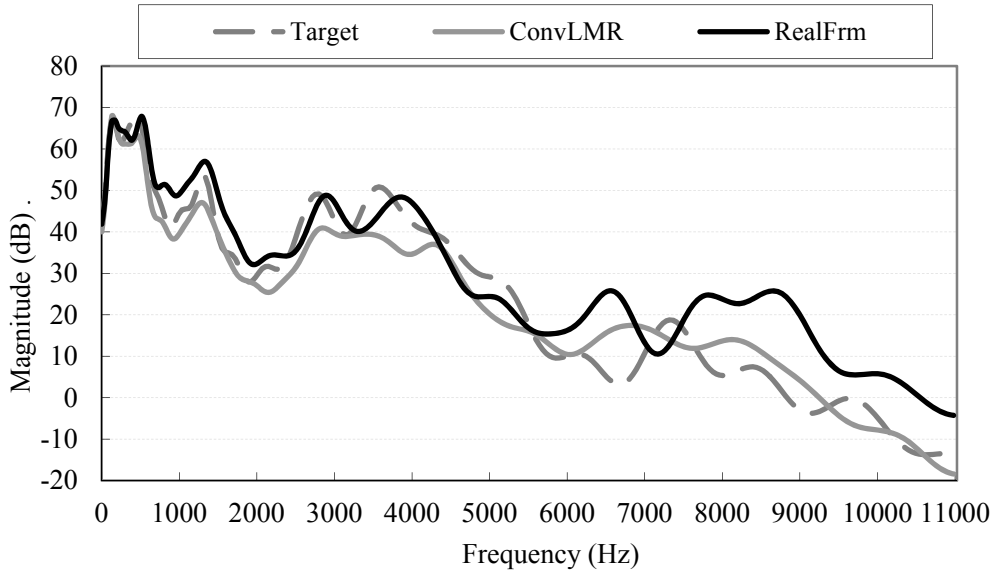


圖 4. 音節/song/的一個音框的三條頻譜包絡曲線

轉換後音框與目標音框的頻譜係數之間，誤差距離平均值的大小並不能夠代表語音品質的好壞，這樣的情形在前人的研究中已經注意到了，所以 Godoy 等人(Godoy *et al*., 2012)採用以變異數比值(variance ratio, VR)來量測轉換後語音的品質，變異數比值的量測公式為:

$$VR = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{L}\cdot\sum_{k=1}^{L}\frac{\hat{\sigma}_i^k}{\sigma_i^k}\ , \tag{13}$$

其中 C 表示音段的類別數，L 表示頻譜特徵向量的維度，$\hat{\sigma}_i^k$ 表示轉換後音框中第 i 類音段第 k 維頻譜係數的變異數，σ_i^k 則表示目標音框第 i 類音段第 k 維頻譜係數的變異數。

對於前面提到的四種處理流程，即作與不作直方圖等化、作與不作目標音框挑選之四種組合，我們依據公式(13)去量測轉換後音框與目標音框之間的變異數比值，結果得到如表 5 所示 VR 值。由表 5 的 VR 值可發現，若不作目標音框挑選，則平均 VR 值只有 0.2 左右，但是當加入目標音框挑選之後，就可讓平均 VR 值提升到 0.5 以上，所以客觀上來看，目標音框挑選之動作應可讓語音品質獲得明顯的提升。至於直方圖等化，做了此種處理反而讓 VR 值下降一些，而 VR 值下降一些是否在主觀聽測上就會感覺到語音品質的衰退? 這尚需進行聽測實驗來驗證。

表 5. 變異數比值之比較

	無 目標音框挑選		有 目標音框挑選	
配對	DCC+LMR	HEQ+LMR	DCC+LMR	HEQ+LMR
MA=> MB	0.2463	0.1671	0.5893	0.5245
MA=> FA	0.1994	0.1290	0.5182	0.4485
FA => MA	0.2367	0.1775	0.5814	0.5383
FA => FB	0.2063	0.1375	0.5648	0.5303
平均	**0.2222**	**0.1528**	**0.5634**	**0.5104**

4.5 語音品質主觀聽測

我們使用未參加模型訓練的來源語句，來準備 4 組作語音品質聽測的音檔，這 4 組音檔的代號是 VD、VH、WD、WH，並且每一組中含有兩個音檔，分別是使用 MA=>MB 與 MA=>FA 之語者配對來作語音轉換而產生出的音檔，在此以_1 與_2 之代號來作區分。代號 VD 與 VH 中的 V 表示未作目標音框挑選，而 WD 與 WH 中的 W 則表示有作目標音框挑選；此外，VD 與 WD 中的 D 表示直接拿 DCC 向量去作 LMR 對映，就如圖 1 之處理流程，而 VH 與 WH 中的 H 表示 DCC 向量要先作 PCA 係數轉換及 CDF 係數轉換，然後才作 LMR 對映，就如圖 2 之處理流程。這 4 組音檔可從如下網頁去下載試聽: http://guhy.csie.ntust.edu.tw/vcHeqLmr/。

使用這 4 組音檔，我們先編排成二項的聽測實驗，第一項聽測實驗裡，受測者先、後點播(VD_1, VH_1)與(VD_2, VH_2)兩對音檔來試聽，然後受測者分別給每一對音檔一

個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞；第二項聽測實驗裡，受測者先後點播(WD_1, WH_1)與(WD_2, WH_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞。在二項聽測實驗裡，受測者都是同樣的 12 位學生，他們大部分都不熟悉語音轉換之研究領域，至於評分的標準是，2 (-2)分表示右(左)邊音檔的語音品質比左(右)邊音檔的明顯地好，1 (-1)分表示右(左)邊音檔的語音品質比左(右)邊音檔的稍為好一點，0 分表示分辨不出左、右兩音檔的語音品質。在二項聽測實驗之後，我們將受測者所給的評分作整理，結果得到如表 6 所示的平均評分。從表 6 的二項平均評分(即 0.583 與 0.375)可得知，評分分數都是正值，表示先作直方圖等化再作 LMR 對映，比起 DCC 向量直接作 LMR 對映會得到更好一些的語音品質；此外，第二項聽測的平均評分(0.375)，比起第一項聽測的平均評分(0.583)要稍微低一點，表示在作過目標音框挑選的處理之後，直方圖等化所帶來的語音品質改進，就會變得較不明顯。

表 6. 語音品質聽測--比較 DCC 與 HEQ

	DCC vs. HEQ (無 目標音框挑選)	DCC vs. HEQ (有 目標音框挑選)
平均評分 AVG (STD)	0.583 (0.776)	0.375 (0.824)

接著，我們再將前述的 4 組音檔作編排以進行另二項聽測實驗，在第三項聽測實驗裡，受測者先、後點播(VD_1, WD_1)與(VD_2, WD_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞；在第四項聽測實驗裡，受測者先後點播(VH_1, WH_1)與(VH_2, WH_2)兩對音檔來試聽，然後受測者分別給每一對音檔一個評分，以顯示左邊音檔的語音品質比起右邊音檔的品質是好或壞。在第三、第四項聽測實驗裡，受測者也共有 12 位學生，他們大部分不熟悉語音轉換之研究領域，至於評分的標準與分數範圍則和前一段所說的一樣。在這二項聽測實驗之後，我們將受測者所給的評分作整理，結果得到如表 7 所示的平均評分。從表 7 二項平均評分 0.917 與 1.125 可得知，只要加入目標音框挑選的處理，就可讓轉換出語音的品質獲得明顯的提升，並且這樣的提升要比表 6 裡的更明顯很多，所以這二項聽測實驗的結果，和表 5 裡量測出的 VR 值是相互呼應的。

表 7. 語音品質聽測--比較有、無目標音框挑選之差異

TFS (Target Frame Selection)	TFS_no vs. TFS_yes (DCC+LMR)	TFS_no vs. TFS_yes (HEQ + LMR)
平均評分 AVG (STD)	0.917 (0.584)	1.125 (0.680)

5. 結論

我們研究改進了線性多變量迴歸(LMR)頻譜對映為基礎的語音轉換方法，在處理流程中加入**直方圖等化**及**目標音框挑選**之處理步驟，用以提升轉換出語音的品質。當我們在圖一流程的 DCC 估計與 LMR 對映之間插入“直方圖等化”處理(包含 PCA 係數轉換與 CDF 係數轉換)之後，雖然語音轉換的平均誤差距離會由 0.5382 (古鴻炎等，2012)變大成為 0.5414，但是主觀聽測實驗的結果顯示，轉換出語音的品質卻是比未加直方圖等化時的好，所以直方圖等化處理可用以紓解 LMR 對映所造成的頻譜過度平滑之問題。此外，關於來源語者和目標語者是否應共用主成分向量的疑問，實驗的結果顯示，讓兩語者共用主成分向量是比較好的作法，可讓語音轉換的平均誤差從 0.5447 減小成 0.5414。

另一種改進語音品質的方法是，在圖 1 流程的 LMR 對映與 HNM 語音再合成之間插入“目標音框挑選”之處理，雖然語音轉換的平均誤差距離會由 0.5382 變大成為 0.6029，但是客觀 VR 值的量測及主觀聽測實驗的結果都顯示，轉換出語音的品質確實是明顯地提升了，不論 LMR 頻譜對映方塊之前有否作過直方圖等化的處理，所以“目標音框挑選”比起“直方圖等化”，對於轉換出語音之品質提升更為有功效，並且 VR 值大體上可反應出語音的品質。另外，對於平均誤差距離愈大反而得到愈好的語音品質，這種不一致性的情況，我們觀察一些音框的頻譜包絡曲線後發現，轉換出之語音聽起來比較模糊者，通常其頻譜包絡在 2,500 Hz 至 4,500 Hz 之頻率範圍，會顯現過度平滑的情形，並且比起清晰者較為遠離目標頻譜包絡曲線；然而在 5,000 Hz 之後的頻率範圍，雖然模糊者的頻譜包絡也是顯現過度平滑的情形，但是比起清晰者卻較為接近目標頻譜包絡曲線，所以會計算出比較小的誤差距離。

致謝

感謝國科會計畫之經費支援，國科會計畫編號 NSC 101-2221-E-011-144。

參考文獻

Abe, M., Nakamura, S., Shikano, K., & Kuwabara, H. (1988). Voice Conversion through Vector Quantization. *Int. Conf. Acoustics, Speech, and Signal Processing*, 1, 655-658.

Cappé, O., & Moulines, E. (1996). Regularization Techniques for Discrete Cepstrum Estimation. *IEEE Signal Processing Letters*, *3*(4), 100-102.

Dutoit, T., Holzapfel, A., Jottrand, M., Moinet, A., Perez, J., & Stylianou, Y. (2007). Towards a Voice Conversion System Based on Frame Selection. *Int. Conf. Acoustics, Speech, and signal Processing*, Honolulu, Hawaii, 513-516.

Erro, D., Moreno, A., & Bonafonte, A. (2010). Voice Conversion Based on Weighted Frequency Warping. *IEEE trans. Audio, Speech, and Language Processing*, 18, 922-931.

Godoy, E., Rosec, O., & Chonavel, T. (2009). Alleviating the One-to-many Mapping Problem in Voice Conversion with Context-dependent Modeling. *Proc. INTERSPEECH 2009*, 1627-1630.

Godoy, E., Rosec, O., & Chonavel, T. (2012). Voice Conversion Using Dynamic Frequency Warping with Amplitude Scaling, for Parallel or Nonparallel Corpora. *IEEE trans. Audio, Speech, and Language Processing*, 20, 1313-1323.

Gu, H. Y., & Tsai, S. F. (2009). A Discrete-cepstrum Based Spectrum-envelope Estimation Scheme and Its Example Application of Voice Transformation. *International Journal of Computational Linguistics and Chinese Language Processing*, *14*(4), 363-382.

Gu, H. Y. , & Tsai, S. F. (2011). An Improved Voice Conversion Method Using Segmental GMMs and Automatic GMM Selection. *Int. Congress on Image and Signal Processing,* Shanghai, China, 2395-2399.

Hotelling, H. (1933). Analysis of a Complex of Statistical Variables into Principal Components. *Journal of Educational Psychology*, *24*(6), 417-441.

Jolliffe, I. T. (2002). *Principal Component Analysis*, second edition, New York: Springer-Verlag, 2002.

Kim, H. Y. *et al*. (1998). Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter. *20-th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Hong Kong, China, 1998.

Lin, S. H., Yeh, Y. M., & Chen, B. (2007). A Comparative Study of Histogram Equalization (HEQ) for Robust Speech Recognition. *International Journal od Computational Linguistics and Chinese Language Processing*, *12*(2), 217-238.

Stylianou, Y. (1996). *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*, Ph.D. thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, 1996.

Stylianou, Y., Cappé, O., & Moulines, E. (1998). Continuous Probabilistic Transform for Voice Conversion. *IEEE trans. Speech and Audio Processing*, *6*(2), 131-142.

Toda, T., Black, A. W., & Tokuda, K. (2007). Voice Conversion Based on Maximum-likelihood Estimation of Spectral Parameter Trajectory. *IEEE trans. Audio, Speech, and Language Processing*, 15, 2222-2235.

Torre, A., Peinado, A. M., Segura, J. C., Perez-Cordoba, J. L., Bentez, M. C., & Rubio, A. J. (2005). Histogram Equalization of Speech Representation for Robust Speech Recognition. *IEEE trans. Speech and Audio Processing*, *13*(3), 355-366.

Valbret, H., Moulines, E., & Tubach, J. P. (1992).Voice Transformation Using PSOLA Technique. *Speech Communication*, *11*(2-3), 175-187.

古鴻炎、張家維、王讚緯(2012)。以線性多變量迴歸來對映分段後音框之語音轉換方法。，*第 24 屆自然語言與語音處理研討會*，中壢，Session 1 (speech processing)。

Computational Linguistics and Chinese Language Processing
Vol. 18, No. 4, December 2013, pp. 115-132

雜訊環境下應用線性估測編碼於特徵時序列之強健性語音辨識

Employing Linear Prediction Coding in Feature Time Sequences for Robust Speech Recognition in Noisy Environments

范顯騰*、曾文俞*、洪志偉*

Hao-teng Fan, Wen-yu Tseng, and Jeih-weih Hung

摘要

在本論文裡，我們提出了一種藉由線性估測編碼來強化語音辨識中特徵之抗噪性的新方法，在此方法中，根據線性估測編碼技術，將語音倒頻譜特徵時間序列分解出估測誤差成分後，將此估測誤差成分從原特徵序列扣除，所得的新特徵序列，相對於原始特徵序列而言，發現具有更佳的雜訊強健性，在 Aurora-2 此包含各類雜訊之數字語料庫的實驗環境下，經過各種預強健化處理之倒頻譜語音特徵，再進一步藉由我們所提之新方法處理後，都能得到更佳的辨識效能，且在線性估測階數很低的情況下，就可有效提升辨識率，顯示了我們可以高效率地執行實現所提之新技術。

關鍵詞：線性估測編碼、特徵時間序列、雜訊強健性。

Abstract

In this paper, we present a novel method to extract noise-robust speech feature representation in speech recognition. This method employs the algorithm of linear predictive coding (LPC) on the feature time series of mel-frequency cepstral coefficients (MFCC). The resulting linear predictive version of the feature time

*國立暨南國際大學電機工程學系 Department of Electrical Engineering, National Chi Nan University
E-mail: { s99323904; s100323553}@mail1.ncnu.edu.tw; jwhung@ncnu.edu.tw

series, in which the linear prediction error component is removed, reveals more noise-robust than the original one, probably because the prediction error portion corresponding to the noise effect is alleviated accordingly. Experiments conducted on the Aurora-2 connected digit database shows that the presented approach can enhance the noise robustness of various types of features in terms of significant improvement in recognition performance under a wide range of noise environments. Furthermore, a low order of linear prediction for the presented method suffices to give promising performance, which implies this method can be implemented in a quite efficient manner.

Keywords: Noise Robustness, Speech Recognition, Linear Predictive Coding, Temporal Filtering.

1. 緒論

本論文是探討與發展降低各種外在環境存在之雜訊干擾所對應的強健性演算法。在近幾十年來，無數的學者先進對於此雜訊干擾問題提出了豐富眾多的演算法，也都可對雜訊環境下的語音辨識效能有所改進，我們把這些方法略分成兩大範疇：

（1）強健性語音特徵參數（robust speech feature）求取

這類方法主要目的在抽取不易受到外在環境干擾下而失真的語音特徵參數，或從原始語音特徵中儘量削減雜訊造成的效應，其中常應用的是探究語音訊號與雜訊干擾不同的特性、藉由凸顯其差異將二者儘量分離，這類的方法可使用語音訊號的不同領域（domain）上，分別發揮不同的效果，例如常見的時域、頻域、對數頻域與倒頻譜之時間序列域等。比較知名的方法有：頻域上的頻譜消去法（spectral subtraction, SS）（Boll, 1979）、韋納濾波器法（Wiener filtering, WF）（Plapous *et al*., 2006）、對數頻域上的對數頻譜平均消去法（logarithmic spectral mean subtraction, LSMS）（Gelbart & Morgan, 2001）與基於雙聲道語音之片段線性補償法（Stereo-based Piecewise Linear Compensation for Environments, SPLICE）（Deng *et al*., 2003），倒頻譜之時間序列域上的倒頻譜平均消去法（cepstral mean subtraction, CMS）（Furui, 1981）、倒頻譜增益正規化法（cepstral gain normalization, CGN）（Yoshizawa *et al*., 2004）、倒頻譜平均值與變異數正規化法（cepstral mean and variance normalization, CMVN）（Tiberewala & Hermansky, 1997）、倒頻譜統計圖正規化法（cepstral histogram normalization, CHN）（Hilger & Ney, 2006）、倒頻譜形狀正規化法（cepstral shape normalization, CSN）（Du & Wang, 2008）、倒頻譜平均值與變異數正規化結合自動回歸動態平均濾波器法（cepstral mean and variance normalization plus auto-regressive-moving average filtering, MVA）（Chen & Bilmes, 2007）等，附帶一提的是，近幾年來本語音實驗室針對倒頻譜之時間序列域開發了許多此類的強健型包括了：廣義對數域調變頻譜平均值正規化法（generalized-log magnitude spectrum mean normalization, GLMSMN）（Hsu *et al*., 2012）、調變頻譜指數權重法（modulation spectrum exponential weighting, MSEW）（Hung *et al*., 2012a）、調變頻譜替代法（modulation

spectrum replacement, MSR）（Hung *et al*., 2012b）、調變頻譜濾波法（modulation spectrum filtering, MSF）（Hung *et al*., 2012b）、分頻帶調變頻譜補償（Sub-band modulation spectrum compensation）（Tu *et al*., 2009）等，這些方法跟前人所提的許多技術幾乎都可以有良好的加成性、對於語音特徵更好的強健性加強效果。

（2）語音模型調適法（speech model adaptation）

此類的方法則是藉由少量的應用環境語料或雜訊，來對原始的語音模型中的統計參數作調整，降低模型之訓練環境與應用環境之不匹配的情況，而它特點之一是在於無需對於待辨識的語音或其特徵作消噪等強健的處理。較有名的語音模型調適技術包含了：最大後機率法則調適法（maximum a posteriori adaptation, MAP）（Gauiain & Lee, 1994）、平行模型合併法（parallel model combination, PMC）（Hung *et al*., 2001）、向量泰勒級數轉換（vector Taylor series transform, VTS）（Moreno *et al*., 1996）與最大相似度線性回歸法調適（maximum likelihood linear regression, MLLR）（Leggetter & Woodland, 1995）等。

本論文較集中討論與發展的是上述的第一類方法，簡單來說，我們將提出一套作用於倒頻譜時間序列域的強健性技術，稱作線性估測編碼濾波法（linear prediction coding-based filtering, LPCF），此方法主要是應用線性估測（linear prediction）的原理，來擷取語音特徵隨著時間變化的特性、進而凸顯語音的成分、抑制雜訊的成分。實驗結果將顯示，LPCF 此方法作用於原始倒頻譜特徵或經過許多上述之強健性技術預處理後的倒頻譜特徵，都可帶來明顯進步的辨識率。

本論文其他之各章節之內容結構安排如下：第二節中，將簡介線性預估編碼（linear prediction coding, LPC）之處理的原理與應用發展，並藉其推演出所提出之基於 LPC 的特徵強健技術。第三節包含了實驗環境設定，第四節則為辨識實驗結果與討論，主要內容為包含所提新技術的各種強健性法在一系列雜訊環境下的語音辨識結果，以及相關討論。最終之第五節則為結論與未來展望，其對本論文內容做一結論，並敘說未來可研究之方向。

2. 基於線性預估編碼之特徵時間序列濾波技術

在本章節中我們將分成二個小節、分別介紹 LPC 之基本原理、及本篇論文所提出之基於 LPC 的語音特徵序列之線性濾波技術。

2.1 LPC的原理介紹

線性預估編碼（王小川，2004）技術普遍運用於訊號處理與資料分析的領域中，此方法的基本背景假設，在於時間（或位置）相近的訊號點彼此存在著相關性，每一個訊號點可以由相近的訊號點藉由線性組合（linear combination）加以逼近或估測，而線性組合中各相鄰訊號點所使用的係數（parameter）即稱為線性估測係數（linear prediction coefficients）。在語音訊號處理的應用上，由於語音訊號在短時距的範圍內有變化緩慢的特性，相鄰訊號點之間的相關性很高，所以許多擷取語音特性的理論或技術，皆是根

據線性估測法則來加以發展推論，其應用範圍廣及語音訊號傳輸上的編碼、語音辨識與語者辨識之特徵等。

在諸多的語音相關主題上，藉由 LPC 可將語音訊號中大致地分離成聲帶與聲道兩大成分，可能是最為廣泛且為人知的應用，以下，我們就簡單地藉由數學推導方式，介紹 LPC 的分析及求取線性估測係數的步驟。

LPC 法將一段時域（time domain）上的訊號 $x[n]$ 用以下數學式表示：

$$x[n]=\sum_{k=1}^{P} a_k x[n-k]+e[n] \quad , \tag{1}$$

$x[n]$ 為在特定時間 n 時的訊號值，其由位於時間軸上 n-1, n-2,..n-P 這串 P 點的訊號值以線性組合（加權組合）來近似，每一點所使用的權重係數（稱做線性預估係數）分別為 $a_1, a_2, \cdots, a_P$, $e[n]$ 則為在近似過程中的誤差訊號，換言之，下式的訊號 $\hat{x}[n]$ 為線性估測的訊號：

$$\hat{x}[n]=\sum_{k=1}^{P} a_k x[n-k] \tag{2}$$

P 稱作線性估測的階數（order）。而式（1）與（2）中 $x[n]$ 與 $\hat{x}[n]$ 兩者之間的差量就是線性估測的誤差，

$$e[n]=x[n]-\hat{x}[n] \tag{3}$$

由之前的陳述，可明顯得知，線性預估的過程是希望原始訊號與預估訊號之間的誤差越小越好，而預估訊號與原始訊號逼近的程度，恰是由線性預估係數$\{a_k\}$所決定，在標準之線性估測理論中，$\{a_k\}$是藉由最小化將誤差訊號 $e[n]$ 的最小均方值（mean squared value）所決定：

$$E=\frac{1}{N}\sum_{n=0}^{N-1} e^2[n]=\frac{1}{N}\sum_{n=0}^{N-1}(x[n]-\sum_{k=1}^{P} a_k x[n-k])^2 \tag{4}$$

其中，N 為計算誤差之訊號點數，將上式對每一個線性預估係數 a_k 作偏微分，並將偏微分後的結果設為 0，就可解出每個線性預估係數 a_k 的最佳值，步驟如下：

$$\frac{\partial E}{\partial a_l}=\frac{1}{N}\sum_{n=0}^{N-1}\{2(x[n]-\sum_{k=1}^{P} a_k x[n-k])x[n-l]\}=0 \tag{5}$$

整理（5），可得出：

$$r_x[l]-\sum_{k=1}^{P} a_k r_x[|l-k|]=0, \qquad l=1,2,...,P \tag{6}$$

其中，$r_x[l]$ 為 $x[n]$ 的自相關係數，定義為：

$$r_x[l]=\sum_{n} x[n]x[n-l] \tag{7}$$

最後將（6）展開排列成矩陣方程式可以得到以下公式：

$$\begin{bmatrix} r_x[0] & r_x[1] & \cdots & r_x[P-1] \\ r_x[1] & r_x[0] & \cdots & r_x[P-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x[P-1] & r_x[P-2] & \cdots & r_x[0] \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} r_x[1] \\ r_x[2] \\ \vdots \\ r_x[P] \end{bmatrix} \tag{8}$$

也可以表示成：

$$\mathbf{R}_x\mathbf{a} = \mathbf{r}_x \ , \tag{9}$$

其中，$\mathbf{R}_x$ 為自相關函數矩陣（autocorrelation function matrix）：

$$\mathbf{R}_x = \begin{bmatrix} r_x[0] & r_x[1] & \cdots & r_x[P-1] \\ r_x[1] & r_x[0] & \cdots & r_x[P-2] \\ \vdots & \vdots & \ddots & \vdots \\ r_x[P-1] & r_x[P-2] & \cdots & r_x[0] \end{bmatrix}$$

$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_P \end{bmatrix}^T$ ，為代表線性預估參數 a_k 的向量。

根據上式，線性預估係數向量可直接由下式求得：

$$\mathbf{a} = \mathbf{R}_x^{-1}\mathbf{r}_x \, , \tag{10}$$

而由於上式需使用到反矩陣運算，複雜度較高，實際作法上，藉由矩陣 $\mathbf{R}_x$ 的特殊性質，有一高效率的演算法，稱作 L-D 遞迴法（Levinson-Durbin recursion）（王小川，2004）來求取線性預估係數向量 **a** 。

一般而言，若要得到 R_x^{-1} ，傳統的作法是利用高斯消去法去求解，但是高斯密度的複雜度是 $O(n^3)$，n 所代表的是未知數的總數，而 L-D 遞迴法（Levinson-Durbin recursion）的作法是利用 Toeplitz 矩陣的特性，使得對角線都是對稱的情況下，去求取每一個線性估測係數，而且複雜度是 $O(n^2)$ ，比起高斯消去法還要來的快速有效。

2.2 LPC技術運用於特徵時間序列域之處理

從前一小節，我們可以清楚了解 LPC 的原理在時域上推理過程，在本節中，我們將介紹本論文主要提出的新方法：即把 LPC 分析技術應用於語音特徵時間序列上，嘗試求取新的語音特徵序列，使其相對於原始特徵序列更具有雜訊強健性。

在我們提出的新方法中，當原始語音特徵時間序列以 $x[n]$ 表示時，**新特徵時間序列為 $x[n]$ 經過 LPC 分析所得到的線性預估序列 $\hat{x}[n]$** ，具體的步驟如下：

步驟一：將原始語音特徵時間序列以 $x[n]$ 作 P 階之線性估測，如式(2)，求取最佳之線性估測係數 $\{a_k, 1 \le k \le P\}$ 。

步驟二：經由下式求得新的特徵時間序列：

$$\hat{x}[n] = \sum_{k=1}^{P} a_k x[n-k] \tag{11}$$

從前面的討論明顯知道，新序列 $\hat{x}[n]$ **是在 P 階的線性估測的模型架構下，最逼近原序列** $x[n]$ 的序列，兩者之間的誤差序列 $e[n]$ 能量可達到最小。

上述的新方法雖然看似簡易，卻有許多合理的原因可顯示新序列相對於原始序列而言，包含了較少的失真、或對於雜訊更具強健性，分述如下：

1. 如前面章節對於 LPC 理論的描述，可知在語音分析中，原始訊號 $x[n]$ 與預估訊號 $\hat{x}[n]$ 之間的誤差訊號可能是週期性訊號或是白色雜訊，一般的線性迴歸模型（auto-regression model, AR model）也是建立在誤差訊號（激發訊號）本身是白色雜訊的假設下。將其套用於我們這裡分析的語音特徵時間序列 $x[n]$ 中，可合理推測線性預估序列 $\hat{x}[n]$ 相當於扣除了 $x[n]$ 其中部份無法線性估測的近似雜訊成份或週期性訊號成份，然而一般從語音特徵時間序列的軌跡，很少出現週期性的現象，因此我們可較確定的是，藉由 LPC 對於原始特徵序列 $x[n]$ 的分解，我們可將其分佈於全頻帶的白色雜訊成份加以消除或減低，而使新特徵序列 $\hat{x}[n]$ 包含較少的失真成份。
2. 繼續前一點的敘述，誤差序列 $e[n]$ 可能是週期性訊號（頻譜亦成週期性）或白色雜訊（頻譜呈現平坦之形狀），但根據許多前人的研究，語音特徵時間序列其頻譜（即調變頻譜）的主要成份是集中於中低頻率上（以音框取樣率 50 Hz 為例，重要的語音調變頻譜約在 1 Hz 至 20 Hz 之間），因此誤差序列不太可能包含語音特徵序列的重要資訊，亦即將其扣除，至少無損於語音辨識的精確度。

針對以上我們所提出的新方法以及推論，我們將以一句乾淨語音為例，將其原始 MFCC 之 c_1 特徵序列做 LPC 分析求取線性預估序列，其中線性估測的階數 P 為 3，對應時序列訊號如圖 1(a)所示，而二者對應的功率頻譜密度如圖 1(b)所示。從這兩張圖顯示，在乾淨環境下，LPCF 所得的特徵時間序列跟原始時間序列無論在時域與頻域都十分接近，表示 LPCF 法不會明顯破壞原始乾淨語音特徵。另外，我們也可約略看出，LPCF 對特徵序列產生的頻譜效應較偏於強調高通成分，只是並不十分明顯。

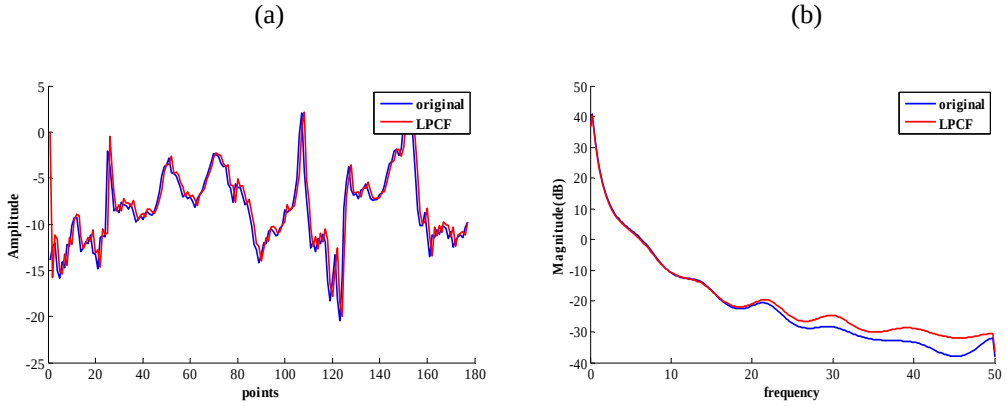


圖 1. 乾淨環境下，MFCC 和經 LPCF 處理後的 c_1 特徵之(a)序列波形圖 (b)功率頻譜密度（PSD）圖

之後，我們將上述乾淨語音加入雜訊，使其訊雜比變為 0 dB，在此狀況下，同樣求取原始 MFCC 之 c_1 特徵序列及其做 LPC 分析所得的線性預估序列，在時域與功率頻譜域的曲線分別顯示於圖 2(a)與圖 2(b)。從圖 2(a)可看出，二者曲線並沒有太大差異，顯示 LPCF 可能無法有效抑制 MFCC 中雜訊造成的失真，但從圖 2(b)可看到，LPCF 約略呈現強調低頻、抑制高頻的效果，此現象應該對於語音特徵的強健性有所幫助。

(a) (b)

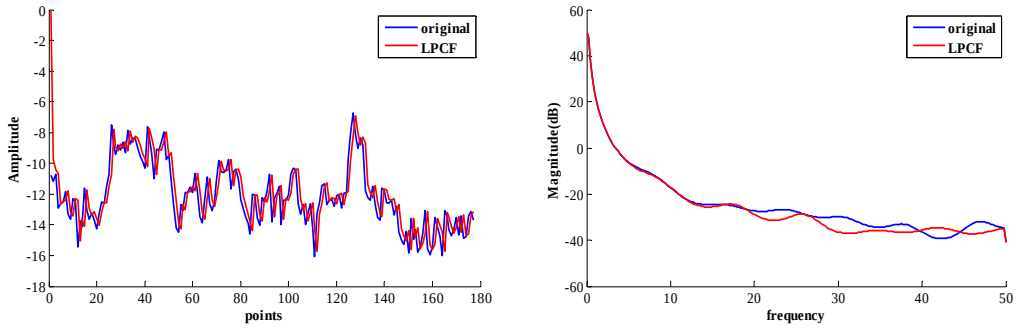


圖 2. 訊雜比 0 dB 之雜訊環境下，MFCC 和經 LPCF 處理後的 c_1 特徵之(a)序列波形圖 (b)功率頻譜密度（PSD）圖

類似之前的繪圖，我們另外求取了以下幾種語音特徵在時域與功率頻譜域中的波形圖，分述如下：

(1) 乾淨環境下，經過 CMVN 預處理後的特徵、及其再經過 LPCF 處理後的特徵，其時間序列與 PSD 圖，繪於圖 3(a)與圖 3(b)。

(a) (b)

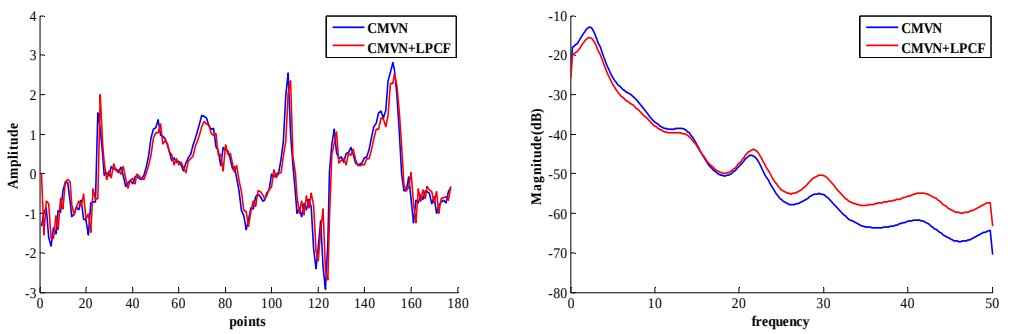


圖 3. 乾淨環境下，CMVN 和經 LPCF 處理後的 c_1 特徵之(a)序列波形圖 (b)功率頻譜密度（PSD）圖

(2) 訊雜比為 0 dB 之雜訊環境下經過 CMVN 預處理後的特徵、及其再經過 LPCF 處理後的特徵，其時間序列與 PSD 圖，繪於圖 4(a)與圖 4(b)。

(a) (b)

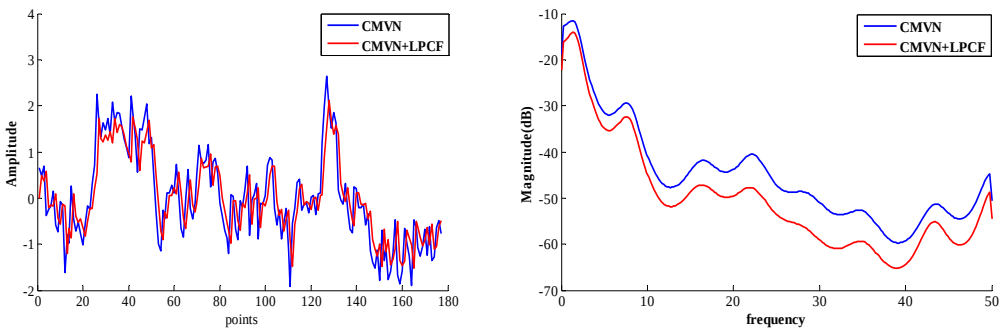


圖 4. 訊雜比 0 dB 之雜訊環境下，CMVN 和經 LPCF 處理後的 c_1 特徵之(a)序列波形圖 (b)功率頻譜密度（PSD）圖

(3) 乾淨環境下，經過 CHN 預處理後的特徵、及其再經過 LPCF 處理後的特徵，其時間序列與 PSD 圖，繪於圖 5(a)與圖 5(b)。

(a) (b)

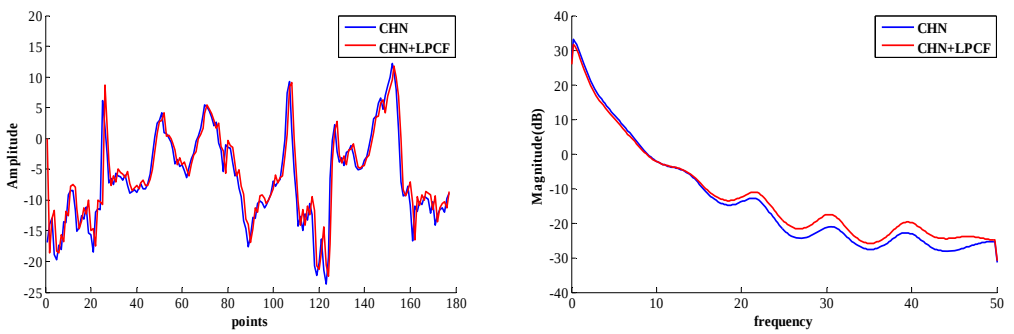


圖 5. 乾淨環境下，CHN 和經 LPCF 處理後的的 c_1 特徵之(a)序列波形圖 (b)功率頻譜密度（PSD）圖

(4) 訊雜比為 0 dB 之雜訊環境下經過 CHN 預處理後的特徵、及其再經過 LPCF 處理後的特徵，其時間序列與 PSD 圖，繪於圖 6(a)與圖 6(b)。

從這些圖，我們可以看出以下的幾個現象：

(1) 無論是乾淨或雜訊干擾的環境下，原特徵序列與 LPCF 處理後的序列在時域上下起伏的波形十分類似，這顯示了 LPCF 不會明顯改變原始特徵序列的相位（phase）。

(2) LPCF 所對應的濾波器頻率響應的形式是隨著原特徵序列而變的，並非固定為高通或低

通濾波器的效應，這代表 LPCF 對應的是資料趨向（data-driven）的濾波處理，能因應所處理的特徵時間序列做調適（adaptation）。

(3) 無論是 CMVN 與 CHN 欲處理的特徵，LPCF 在乾淨環境下所對應的是稍傾於高通形式的全通濾波器（allpass filter），但在雜訊環境下則明顯是低通濾波器，這再次呼應了前面的敘述，在乾淨狀態下，LPCF 儘量保持原有語音特徵的特性，但在雜訊干擾環境下，LPCF 則可抑制特徵序列的高頻成分，強調低頻成分，對於凸顯語音成分的功能顯著，此在之後章節的實驗分析將可以看到。

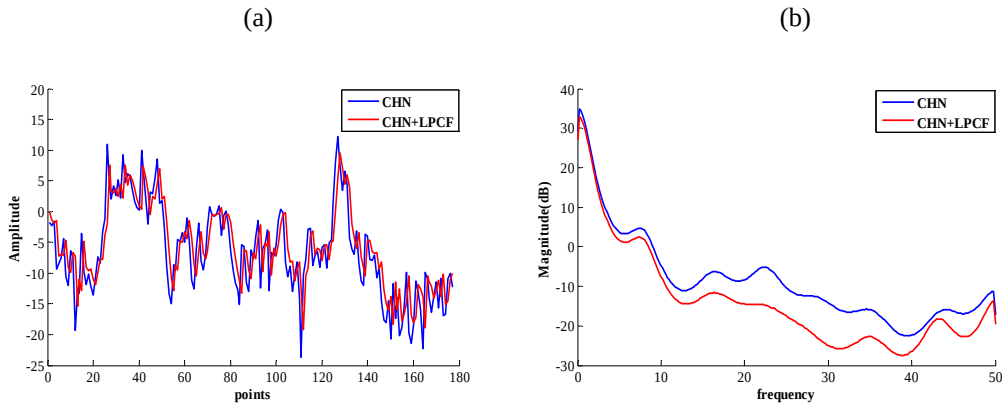


圖 6. 訊雜比 0 dB 之雜訊環境下，CHN 和經 LPCF 處理後的 c_1 特徵之(a)序列波形圖 (b)功率頻譜密度（PSD）圖

最後，我們再利用 Aurora-2 的 Set A 其不同雜訊程度影響下語音特徵 c_1 序列的 PSD 之平均的比較，來觀察 LPCF 法所能達到的效果。圖 7(a)與圖 7(b)分別為「MFCC 處理」與「MFCC 加上 LPCF 處理」後在三種訊雜比環境的語音特徵 c_1 的 PSD 平均之曲線，圖 8(a)與圖 8(b)分別為「CMVN 處理」與「CMVN 加上 LPCF 處理」後在三種訊雜比環境的語音特徵 c_1 的 PSD 平均之曲線，圖 9(a)與圖 9(b)分別為「CHN 處理」與「CHN 加上 LPCF 處理」後在三種訊雜比環境的語音特徵 c_1 的 PSD 平均之曲線，從這些圖形可以看出，當加入 LPCF 處理後，特別對於 CMVN 與 CHN 法預處理的特徵而言，各訊雜比環境下的 PSD 平均曲線都能較為接近，代表了 LPCF 與 CMVN 及 CHN 有良好的加成性，可以進一步降低 CMVN 或 CHN 預處理後剩餘的失真，進而使不同訊雜比下的特徵特性更為匹配，唯有 MFCC 加入 LPCF 處理後，並未十分明顯的使特徵更為匹配。顯示 LPCF 法不太適合作在原始 MFCC 特徵上。

(a) (b)

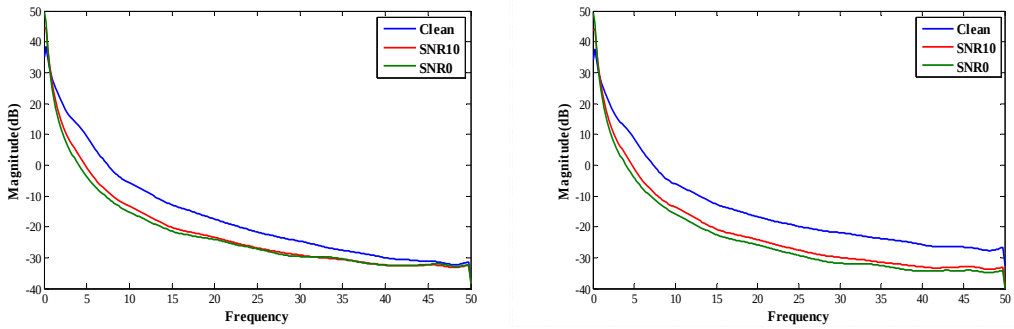


圖7. 不同訊雜比下，Set A 之1001 句的MFCC 其(a)原始c_1特徵序列與(b)經LPCF 處理後之 c_1 特徵序列 的平均功率頻譜密度

(a) (b)

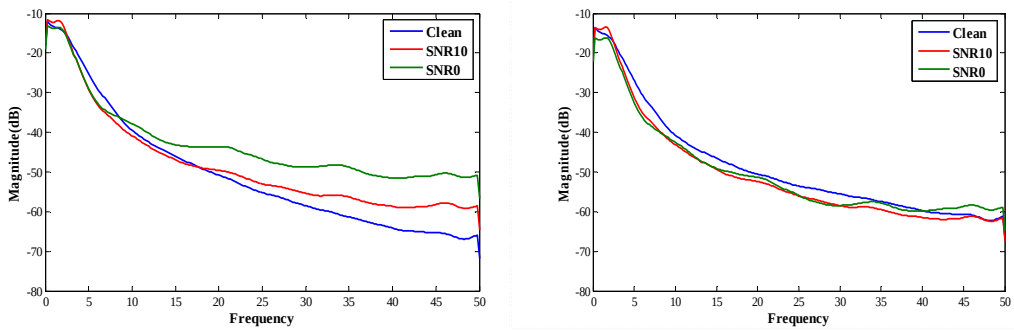


圖8. 不同訊雜比下，Set A 之1001 句之(a)經CMVN 處理後之 c_1 特徵序列 (b) 經 CMVN+LPCF 處理後之 c_1 特徵序列 的平均功率頻譜密度

(a) (b)

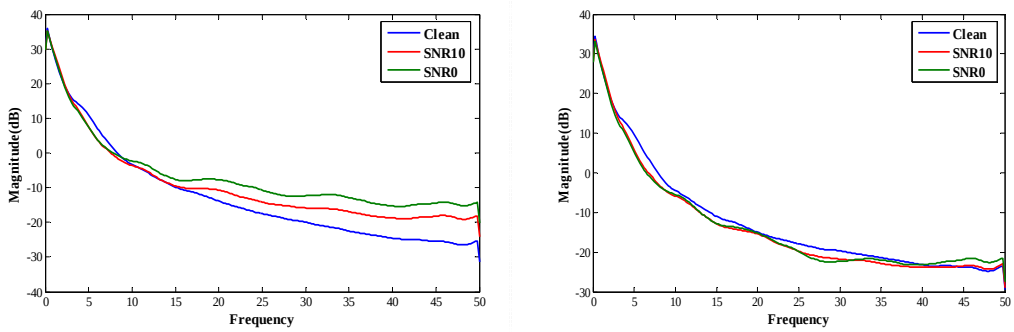


圖9. 不同訊雜比下，Set A 之1001 句之(a)經CHN 處理後之 c_1 特徵序列 (b) 經 CHN+LPCF 處理後之 c_1 特徵序列 的平均功率頻譜密度

3. 實驗環境設定

本論文之實驗中所採用的語音資料庫為歐洲電信標準協會（European Telecommunication Standard Institute, ETSI）所發行的 AURORA 2.0（Hirsch & Pearce, 2000）語音資料庫，內容包含美國成年男女以人工方式錄製的一系列連續英文數字字串，在我們所採取的乾淨模式訓練、多元雜訊模式測試（clean-condition training, multi-condition testing）之實驗架構中，用以訓練聲學模型之語句為 8440 句乾淨語句，唯其包含了 G.712 通道之通道效應。測試語料則包含了三個子集合：Sets A 與 B 的語句摻雜了加成性雜訊，Set C 則同時包含加成性雜訊與摺積性雜訊，Sets A 與 B 各包含 28028 句語音，Sets C 包含 14014 句語音。加成性的雜訊種類分別為：地下鐵（subway）、人類嘈雜聲（babble）、汽車（car）、展覽館（exhibition）、餐廳（restaurant）、街道（street）、機場（airport）、火車站（train station）等雜訊，並以不同程度的訊雜比（signal-to-noise ratio, SNR）摻雜，分別為：clean、20 dB、15 dB、10 dB、5 dB、0 dB 與-5 dB；而通道效應分為 G.712 與 MIRS 兩種通道標準，由國家電信聯盟（international telecommunication Union, ITU）（Hirsch & Pearce, 2000）所訂定而成。

上述之用以訓練與測試的語句，我們先將其轉換成梅爾倒頻譜特徵參數（mel-frequency cepstral coefficients, MFCC），作為之後各種強健性方法的基礎特徵（baseline feature），建構 MFCC 特徵的過程主要是根據 AURORA 2.0（Hirsch & Pearce, 2000）資料庫中的設定，最終 MFCC 特徵包含了 13 維的靜態特徵（static features）附加上其一階差分與二階差分的動態特徵（dynamic features），共 39 維特徵。值得一提的是，本論文之後所提的強健性技術，皆是作用於 13 維的靜態特徵上，再由更新後的靜態特徵求取 26 維的動態特徵。新提出的 LPCF 的演算法，藉此得到更佳的辨識精確度。

在聲學模型上，我們採取連續語音辨識中常見的隱藏式馬可夫模型（hidden Markov model, HMM），並採用由左到右（left-to-right）形式的 HMM，意即下一個時間點所在的狀態只能停留在當下的狀態或下一個鄰近的狀態，狀態的變遷隨著時間由左至右依序前進。此外，模型中的狀態觀測機率函數為連續式高斯混合機率函數（Gaussian mixtures），所以此模型又稱為連續密度隱藏式馬可夫模型（continuous-density hidden Markov model, CDHMM）。我們採用了 HTK（HTK, n.d.）軟體來訓練上述的 HMM，在模型單位的選取上，採用前後文獨立（context independent）的模型樣式，所得之聲學模型包含了 11 個數字（oh, zero, one, …, nine）與靜音的隱藏式馬可夫模型，每個數字的 HMM 皆包含了 16 個狀態，而每個狀態由 3 個高斯混合函數組成。

4. 實驗數據與討論

本節將由三部分所組成，在第一與第二部分，我們固定新提出的 LPCF 法所用的線性估測階數為 2，分別作用於 MFCC 基礎特徵、與經過各種強健性演算法預處理後的特徵上，探討其對辨識率的改進程度。第三部分則是變化 LPCF 法中的線性估測係數，觀察其對於辨識率的影響。

4.1 階數為2之LPCF法運用於MFCC基礎特徵

表 1 列出了線性估測階數為 2 之 LPCF 法作用於 MFCC 基礎特徵所得之辨識率。將此表的數據與 MFCC 基礎特徵所得的辨識率相比較，我們看到在 Set A 與 Set B 這兩組雜訊環境下，LPCF 法可以使 MFCC 達到更佳的辨識結果，平均進步率約在 4%，可見 LPCF 法可以提升 MFCC 特徵在加成性雜訊干擾下的強健性，然而，在 Set C 此同時包含通道失真與加成性雜訊的環境下，LPCF 法並未帶來實質的進步，凸顯了此方法較不適用於通道干擾下的語音辨識。

表 1. 原始 MFCC 基礎特徵與其經 LPCF 法（階數為 2）處理後的特徵，在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率(%)平均比較

	Set A	Set B	Set C	Avg
MFCC	59.24	56.37	67.53	59.75
LPCF	63.90	61.96	66.44	63.63

4.2 階數為2之LPCF法運用於經CMVN、CHN或MVA預處理之MFCC特徵

三種著名的時間序列處理技術：CMVN、CHN 與 MVA 皆能明顯提升 MFCC 之雜訊強健性、得到較高的辨識率，因此，這裡我們將 LPCF 法作用於經 CMVN、CHN 或 MVA 法預處理後的 MFCC 特徵上，觀察 LPCF 法是否能夠使它們的辨識率進一步提升，值得注意的是，此時 LPCF 法所使用的線性估測係數必須由預處理後的新特徵求得，而非直接採取原始 MFCC 之 LPCF 法所運用的線性估測係數。跟第一部分相同的是，此時 LPCF 之線性估測階數仍然固定為 2。其辨識率分別列於表 2、3 與 4。從這三個表的數據，我們得到以下的觀察結果：

1. 無論是作用哪一種方法預處理後的特徵，LPCF 法在三組雜訊環境（Sets A, B, C）下皆能明顯提升其平均辨識率，例如就整體平均辨識率而言，LPCF 法能使 CMVN、 CHN 與 MVA 預處理之特徵分別提升了 3.38%、2.2%與 0.87%，此代表了 LPCF 能與這些著名的時序域強健性技術有良好的加成性，換言之，LPCF 可進一步降低這些技術處理後殘餘的雜訊不匹配成分，進而得到更佳的辨識率。
2. 不同於作用於原始 MFCC 之結果，LPCF 作用於上述三種預處理技術後的特徵時，也能使 Set C 此組包含了通道失真與加成性雜訊的語音，辨識率有所提升，其中可能原因在於，三種預處理技術有效降低通道效應後，LPCF 法接著把加成性雜訊造成的失真作有效的抑制，進而在辨識上有較好的效果。
3. 在三種預處理技術的比較上，LPCF 對於 MVA 特徵的辨識率提升，明顯較其對 CMVN 與 CHN 特徵來的小，其中可能原因是，MVA 法中已經結合了一個形式為 ARMA 的低通濾波器，之後再結合 LPCF 之類似的濾波處理，改進的效應較不明顯。

表 2. CMVN 法與 MVN 串聯 LPCF 法（階數為 2），在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率(%)平均比較

	Set A	Set B	Set C	Avg
CMVN	73.83	75.01	75.09	74.55
MVN+LPCF	77.14	78.84	77.67	77.93

表 3. CHN 法與 CHN 串聯 LPCF 法（階數為 2），在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率(%)平均比較

	Set A	Set B	Set C	Avg
CHN	81.42	83.34	81.51	82.21
CHN+LPCF	77.14	78.84	77.67	77.93

表 4. MVA 法與 MVA 串聯 LPCF 法（階數為 2），在不同組別之下、取 5 種訊雜比(20 dB, 15 dB, 10 dB, 5 dB 與 0 dB)之辨識率(%)平均比較

	Set A	Set B	Set C	Avg
MVA	78.15	79.17	79.12	78.75
MVA+LPCF	78.15	79.17	79.12	78.75

4.3 變化LPCF法中線性估測之階數（order）產生的效應

在第一部分與第二部分中，我們使用的 LPCF 法，其中線性估測的階數固定為 2，從其實驗結果，我們觀察到使用很低的階數就能使 LPCF 法發揮不錯的效能、有效改善 CMVN、CHN 與 MVA 預處理後的 MFCC 語音特徵之強健性。在這一部分，我們進一步將 LPCF 法的線性估測的階數加以變化，探討此變化對於辨識率產生的影響。類似之前的兩部分，變化線性估測階數的 LPCF 法會分別作用於原始 MFCC 特徵、以及經 CMVN、CHN 或 MVA 預處理後的特徵。

首先，我們觀察不同階數之 LPCF 法作用於 MFCC 原始特徵的效應，我們將作用於 MFCC 原始特徵的 LPCF 法其階數分別設為 2, 3, 4, …, 10，進而執行對應的辨識實驗。表 5 列出了不同階數之 LPCF 法所得的辨識率結果，從此表的數據來看，當階數為 3 時，可得到最佳的辨識率，但與我們先前所設定的階數為 2 相較，辨識率的提升上並未十分明顯，我們也看到，增加階數反而使辨識率逐漸下降，此現象的可能原因在於，增加階數使 LPCF 法對應之濾波器的長度增加，進而使濾波器輸出訊號的暫態響應（transient response）變長，相對於輸入訊號有更長的延遲（delay），在要求輸入訊號（即原始特徵時間序列）與輸出訊號（即藉由 LPCF 更新後的特徵序列）二者長度一致的前提下，達到穩態(steady state)的輸出訊號的區域明顯變短，而無法發揮 LPCF 法所預期的效果。因此總結而論，當作用於原始 MFCC 時，變化 LPCF 之階數並未能對於預期的強健功能有明顯的改善。

表 5. 不同階數之 LPCF 法作用於 MFCC 基礎特徵，在 10 類雜訊環境與 5 種訊雜比之總平均辨識率(%)比較

LPCF 階數	2	3	4	5	6	7	8	9	10
辨識率(%)	63.63	**63.74**	63.47	63.08	62.81	62.68	62.48	62.02	61.94

接著，我們把不同階數之 LPCF 法作用於經 MVN、CHN 或 MVA 預處理後的特徵，進而執行對應的辨識實驗。表 6、7 與 8 列出了所得的辨識率結果，從這三個表的數據，並與表 2 至 4 的數據比較，我們有以下的發現及討論：

1. 就 CMVN 預處理之特徵而言，最佳的 LPCF 之階數為 3，相較於不作 LPCF 的結果，總辨識率可提升 4.19%，而跟原始階數為 2 的數據相較，階數設為 3 可使平均辨識率提升約 1%。而 LPCF 法使用超過 3 的階數時，對應之辨識率並無顯著的下降，在所選定之階數範圍 2 至 10 之中，CMVN 法與 LPCF 法的組合皆比單一 CMVN 法得到更佳的辨識效能。
2. 就 CHN 預處理之特徵而言，最佳的 LPCF 之階數為 4，相較於不作 LPCF 的結果，總辨識率可提升 2.65%，而跟原始階數為 2 的數據相較，階數設為 4 可使平均辨識率提升約 0.5%。其他結果與上一點關於 CMVN 與 LPCF 之組合的結果很類似。足見簡易的 LPCF 法（階數較少）在與 CHN 法結合時，就有近乎最佳的效能表現。
3. 就 MVA 預處理之特徵而言，最佳的 LPCF 之階數為 10，此現象與前兩點所述之 CMVN 及 CHN 法較不同，但仔細觀察，可看出當與 MVA 法結合時，不同階數之 LPCF 法所得到的辨識率十分接近，最佳平均值與跟原始階數為 2 的數據相較，也只提升了 0.03% 至 0.13%，足見此時 LPCF 的階數對其效能影響甚微。如同前面的討論，LPCF 法與 MVA 法的加成性較低，但是二者結合仍比單一 MVA 法在提升 MFCC 之辨識率的表現上來的好。

表 6. 不同階數之 LPCF 法作用於 CMVN 預處理特徵，在 10 類雜訊環境與 5 種訊雜比之總平均辨識率(%)比較

LPCF 階數	2	3	4	5	6	7	8	9	10
辨識率(%)	77.93	**78.74**	78.28	78.11	77.86	77.80	77.80	77.75	77.82

表 7. 不同階數之 LPCF 法作用於 CHN 預處理特徵，在 10 類雜訊環境與 5 種訊雜比之總平均辨識率(%)比較

LPCF 階數	2	3	4	5	6	7	8	9	10
辨識率(%)	84.41	84.82	**84.86**	84.81	84.71	84.66	84.58	84.59	84.63

表 8. 不同階數之 LPCF 法作用於 MVA 預處理特徵，在 10 類雜訊環境與 5 種訊雜比之總平均辨識率(%)比較

LPCF 階數	2	3	4	5	6	7	8	9	10
辨識率(%)	79.62	79.25	79.51	79.59	79.69	79.58	79.72	79.71	**79.73**

5. 結論

在本論文中，我們所提出一套基於線性估測編碼的新方法（LPCF），應用於倒頻譜序列上，此新方法看似簡易，卻有許多合理的原因可顯示新序列包含了較少的失真或對於雜訊更具強健性。許多改良倒頻譜序列的新特徵如 CMVN、CHN 和 MVA，這些方法都能明顯地改善雜訊帶來不匹配，但是這些方法在處理雜訊成分較低的語音特徵序列時，會對語音成分造成失真。所以我們除了探討 LPCF 應用於原始的 MFCC 特徵辨識效能變化以外，我們也將上述的這幾種方法所產生之抗雜訊能力強的特徵再經由我們所提出的 LPCF，並且從實驗數據觀察，我們可以得知不論是哪一種特徵，都能藉由 LPCF 來提升辨識率，並且也可從功率頻譜密度圖得知線性估測編碼濾波器法確實可抑制特徵序列的高頻成分，強調低頻成分。

過去的線性估測的演算法多用於頻譜的估測上，這種使用方法通常必須要將線性估測階數調整為 8~12 階才能有較好的表現，但是本論文中線性估測用於倒頻譜序列上，則採用很小的階數，就可以得到抗雜訊能力較高的特徵。

在實際應用上，過去很多文獻裡，線性估測編碼已是現今普及的數位音訊處理技術，其主要優點是低位元率與高壓縮率，但常只是侷限於傳送語音訊號，而我們所提出的 LPCF 方法可以運用於語音特徵的傳輸上，藉由適當之 LPC 階數的選擇，使其語音特徵傳遞時，不會造成辨識效果的降低，甚至可以提升語音特徵的抗噪能力、使傳輸的語音特徵，同時具備傳輸效率與強健效能的優點。

在未來展望中，由於我們所提出的 LPCF 法的缺點之一，在於需要整句語音的特徵皆已接收到後，才能精確地估測 LPCF 的參數，未來我們希望針對這缺點加以改善，另外我們希望更進一步的研究我們所提之 LPCF 法相關的理論基礎，並且可以利用動態調適的方法來求取 LPCF 法中的階數，然而提升此法的效能，此外，我們也將廣泛地測試 LPCF 法，使其能更進一步運用於其它干擾與失真環境的特徵強健性之改善上。

參考文獻

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

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

Deng, L., Droppo, J., & Acero, A. (2003). Recursive estimation of non-sta-tionary noise using iterative stochastic approximation for robust speech recognition. *IEEE Transactions on Speech Audio Process*, *11*(6), 568-580.

Du, J., & Wang, R. (2008). Cepstral shape normalization for robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 4389-4392.

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

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

Gelbart, D., & Morgan, N. (2001). Evaluating long-term spectral subtraction for reverberant ASR. In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, 103-106.

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

Hirsch, H. G., & Pearce, D. (2000). The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In *Proceedings of the 2000 Automatic Speech Recognition Challenges for the new Millenium*, 181-188.

Hsu, C. H., Fang, H. T., & Hung, J. W. (2012). The study of q-logarithmic modulation spectral normalization for robust speech recognition. In *Proceedings of International Conference on System Science and Engineering*, 183-186.

Hung, J. W., Fan, H. T., & Lian, Y. C. (2012). Modulation spectrum exponential weighting for robust speech recognition. In *Proceedings of International Conference on ITS Telecommunications*, 812-816.

Hung, J. W., Tu, W. H., & Lai, C. C. (2012). Improved modulation spectrum enhancement methods for robust speech recognition. In *Proceedings of Signal Processing*, *92*(11), 2791-2814.

Hung, J. W., Shen, J. L., & Lee, L. S. (2001). New approaches for domain transformation and parameter combination for improved accuracy in parallel model combination techniques. *IEEE Transactions on Speech and Audio Processing*, *9*(8), 842-855.

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

Moreno, P. J., Raj, B., & Stern, R. M. (1996). A vector Taylor series approach for environment-independent speech recognition. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 2, 733-736.

Plapous, C., Marro, C., & Scalart, P. (2006). Improved signal-to-noise ratio estimation for speech enhancement. *IEEE Transactions on Acoustics Speech and Signal Processing*, *14*(6), 2098-2108.

Tiberewala, S., & Hermansky, H. (1997). Multiband and adaptation approaches to robust speech recognition. In *Proceedings of European Conference on Speech Communication and Technology*, 25(1-3), 2619-2622.

Tu, W. H., Huang, S. Y., & Hung, J. W. (2009). Sub-band Modulation Spectrum Compensation for Robust Speech Recognition. In *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding*, 261-265.

Yoshizawa, S., Hayasaka, N., Wada, N., & Miyanaga, Y. (2004). Cepstral gain normalization for noise robust speech recognition. In *Proceedings of IEEE International Conference on Acoustics Speech and Signal Processing*, 1, I-209-212.

王小川(2004)。《語音訊號處理》，*全華科技圖書*。

Retrieved from http://htk.eng.cam.ac.uk/

The individuals listed below are reviewers of this journal during the year of 2013. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Cheng-Hsien Chen
Chia-Ping Chen
Man-ni Chu
Hong-Jie Dai
Tirthankar Dasgupta
Feng-fan Hsien
Jeih-weih Hung
Wen-Hsing Lai
Ying-Shing Li
Hui-shan Lin
Kikuo Maekawa
Shu-chen Ou
Ho-hsien Pan
Samir Rahman
Thoudam Doren Singh
Yu Tsao
Hsu Wang
Jia-Ching Wang
Jui-Feng Yeh

2013 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 18

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2013.

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

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

中　華　民　國　　年　　月　　日</td></tr>
</table>

審查結果:

1. 年費：

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

2. 連絡處：

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

3. 本表可自行影印

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

Name: _____(Please print) Date: _____

Please debit my credit card as follows: US$ _____

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

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

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

CARD HOLDER SIGNATURE: _____

Phone No.: _____E-mail: _____

Address: _____

PAYMENT FOR

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

Quantity Wanted: _____

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

Quantity Wanted: _____

US$ ____ ❑ Publications:_____

US$ ____ ❑ Text Corpora: _____

US$ ____ ❑ Speech Corpora:_____

US$ ____ ❑ Others: _____

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

US$ _____ = Total

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

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

中 華 民 國 計 算 語 言 學 學 會
信用卡付款單

姓名: _____(請以正楷書寫)　　日期:：_____

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

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

卡片後三碼：_____（卡片背面簽名欄上數字後三碼）

持卡人簽名：_____(簽名方式請與信用卡背面相同)

通訊地址：_____

聯絡電話：_____E-mail：_____

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$_____ ❑ 中文計算語言學期刊(IJCLCLP) _____

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

NT$_____ ❑ 中研院詞庫小組技術報告_____

NT$_____ ❑ 文字語料庫 _____

NT$_____ ❑ 語音資料庫 _____

NT$_____ ❑ 光華雜誌語料庫1976~2010

NT$_____ ❑ 中文資訊檢索標竿測試集/文件集

NT$_____ ❑ 會員年費：❑續會　❑新會員　❑終身會員

NT$_____ ❑ 其他: _____

NT$_____ = 合計

填妥後請傳真至 02-27881638 或郵寄至:

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

Publications of the Association for Computational Linguistics and Chinese Language Processing

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

10% member discount: _____Total Due:_____

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

Name (please print): _____ Signature: _____

Fax: _____ E-mail: _____

Address：_____

中華民國計算語言學學會

相關出版品價格表及訂購單

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

※ 此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

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

訂購者：_____ 收據抬頭：_____

地　　址：_____

電　　話：_____ E-mail: _____

Information for Authors

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

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

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

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

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

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

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

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

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

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

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

Here shows an example.

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

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

Please visit the following websites for details.

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

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

No page charges are levied on authors or their institutions.

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

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

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

International Journal of
Computational Linguistics & Chinese Language Processing
vol.18 no.4 December 2013

Contents

Special Issue Articles:
Selected Papers from ROCLING XXV

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing