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.

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

Vol.26 No.2 December 2021 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

Advisory Board

Editors-in-Chief

Associate Editors

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. 26, no. 2, December 2021

Contents

Papers

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 1-16

以遷移學習改善深度神經網路模型於中文歌詞情緒辨識

Using Transfer Learning to Improve Deep Neural Networks for Lyrics Emotion Recognition in Chinese

廖家誼*、林亞宣、林冠成、張家瑋+

Jia-Yi Liao, Ya-Hsuan Lin, Kuan-Cheng Lin, and Jia-Wei Chang

摘要

情緒是音樂資訊檢索中的重要屬性，目前深度學習方法已被廣泛用於自動音樂情緒辨識。音樂情緒辨識主要以歌曲情緒為主，部分研究關注英文歌詞，罕見對於中文歌詞情緒辨識的研究。本研究提出運用遷移式學習改善深度神經網路模型—BERT 預訓練模型在中文歌詞的情緒分類任務上。實驗結果顯示，直接使用 BERT 對中文維度情緒語料庫建立中文情緒分類模型，對中文歌詞情緒分類僅有 50%的準確度，若使用 BERT 對中文維度情緒字典與片語建立情緒分類模型再遷移至中文維度情緒語料庫，能達到 71%的歌詞情緒分類準確度。

Abstract

Emotion is an important attribute in music information retrieval. Deep learning methods have been widely used in the automatic recognition of music emotion. Most of the studies focus on the audio data, the role of lyrics in music emotion classification remains under-appreciated. Due to the richness of English language resources, most previous studies were based on English lyrics but rarely in Chinese. This study proposes an approach without specific training for the Chinese lyrics

* 國立中興大學資訊管理學系
Department of Management Information Systems, National Chung Hsing University.

+ 國立臺中科技大學資訊工程系
Department of Computer Science and Information Engineering National Taichung University of Science and Technology
E-mail: jiaweichang.gary@gmail.com
The author for correspondence is Jia-Wei Chang.

emotional classification task: using transfer learning to improve deep neural networks, BERT pre-training model, for the emotion classification in Chinese lyrics. The experimental results show that directly using BERT to build an emotion classification model of CVAT only reach 50% of the classification accuracy. However, using BERT with transfer learning from CVAW, CVAP, to CVAT can achieve 71% classification accuracy.

關鍵詞：自然語言處理，音樂情緒辨識，遷移學習，中文歌詞

Keywords: Natural Language Processing, Music Emotion Recognition, Transfer learning, Chinese Lyrics.

1. 緒論 (Introduction)

音樂和人類情緒相互影響，在生活中扮演不可或缺的角色。音樂的搜尋通常以歌曲標題、詞曲作者、演唱者和演奏流派進行檢索，然而，情緒可以作為音樂的一個新且重要的搜尋屬性。隨著音樂串流平台使用者和歌曲庫的爆炸式增長，傳統的由專家進行情緒標註已不能滿足實際需求，推薦系統需要更快速的標註方法，自動情緒辨識因此成為重要的議題。音樂情緒辨識(Music Emotion Recognition) 用於觀察音樂與人類情感之相關性、對音樂抽取特徵並加以分析找出音樂特徵與人類對於音樂情緒感知的關聯。目前機器學習和深度學習方法已被廣泛用於辨識音樂的情緒。

支持向量機(Support Vector Machine, SVM)和支持向量回歸(Support Vector Regression, SVR)等機器學習方法(Han *et al*., 2009)。基於歌詞和音訊的歌曲情緒檢測方法來計算效價(Valence)和喚醒(Arousal)進行音樂之情緒分類(Jamdar *et al*., 2015)。用卷積神經網路預訓練模型對每 30 秒剪輯的印度古典音樂進行音樂情緒分類(Sarkar *et al*., 2021)。上述研究大多都集中利用聲學特徵進行音樂情緒辨識並無討論歌詞對於情緒的影響。歌詞被賦予情緒，在引發人類的情緒以及預測音樂情緒扮演著重要的角色(Hu & Downie, 2010)。雖然旋律和歌詞會同時對聽眾產生影響，但聽眾對於歌詞內容的偏好能進一步反映聽眾的特徵和傾向(Qiu *et al*., 2019)。Agrawal *et al*. (2021)提出歌詞可視為一連串彼此相關的句子，需捕捉上下文和長期依賴的關係，並在研究使用 Transformer-based 的模型進行歌詞情緒辨識，在多個英文歌詞情緒資料集上取得良好的成果，上述的英文歌詞資料集皆基於 Russell (1980)的 Valence-Arousal 心理學環繞情感模型進行音樂情緒的標註。

本研究提出一中文歌詞情緒分類方法。首先，運用基於 Transformer 語言預訓練模型對中文維度情緒字典(CVAW)與中文維度情緒片語(CVAP)進行建模，其次將模型遷移至中文維度情緒語料(CVAT)，最後將模型直接用於無標註的歌詞文本進行情緒的自動標註。

本研究其餘章節的組織如下：第二節回顧了心理學維度情緒模型、基於 Transformer 之模型和遷移學習的相關文獻。第三節的方法論說明了本研究所使用的兩個資料集、文本預處理並解釋本研究提出的架構。第四節為本研究模型訓練和歌詞驗證的結果。第五節對實驗結果進行相關討論。最後，在第六節總結本研究的成果和未來改進方向。

2. 文獻回顧 (Literature Review)

2.1 維度情緒模型 (Dimensional Models of Emotion)

情緒被視為一個連續體而非離散的形容詞，維度模型比起分類模型有著較低的歧義 (Yang *et al*., 2008)。現有的研究大多採用 Russell (1980) 所提出的心理學環繞模型。Laurier *et al*. (2009) 的研究中表明，Russell 心理學情緒模型可以用於情緒分析或音樂情緒辨識任務。該維度模型的兩個維度的連續數值，分別為效價(Valence)和喚醒(Arousal)。效價(Valence)代表所有情緒體驗所固有的積極或消極，高效價(Valence)的歌曲聽起來更為積極、快樂，低效價(Valence)的歌曲聽起來較沮喪、憤怒。喚醒(Arousal)代表情緒的激動程度，歌曲的能量(energy)能對應於喚醒(Arousal)值，代表歌曲強度，能量高的歌曲通常越快速、響亮和強烈(Kim *et al*., 2011)。

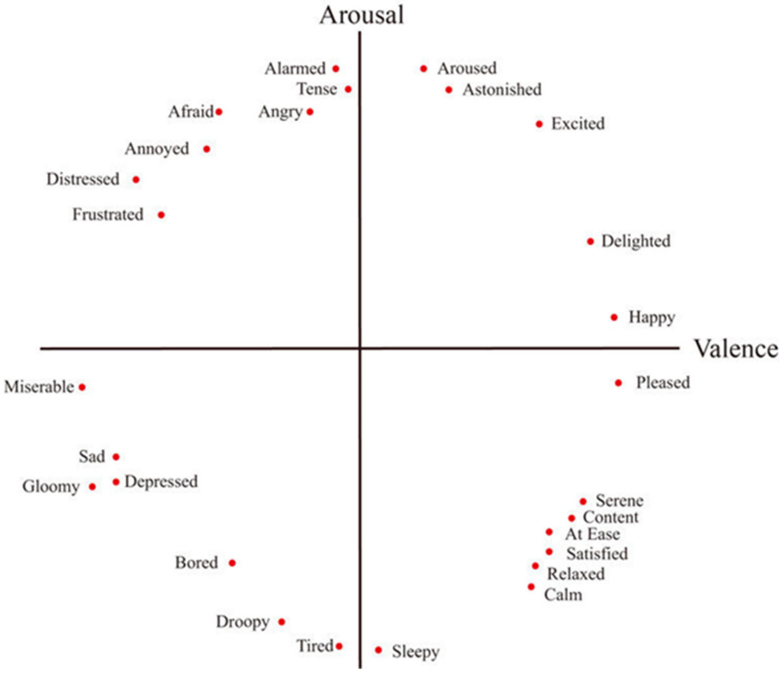


圖 1. Russell 提出之心理學維度情緒模型
[Figure 1. The circumplex model of affect (Russell, 1980)]

如圖 1 所示，情緒由效價(Valence)和喚醒(Arousal)兩個維度表示，情緒平面被分為四個象限，創建了四個情緒類別空間。在 Çano & Morisio (2017a)的研究中基於 Russell 維度情緒模型的四個象限將情緒分為四類別(Q1、Q2、Q3、Q4)，分別為快樂、憤怒、悲傷和輕鬆。因此，本研究在歌詞驗證的部分也依此方法將歌詞情緒分為四個象限類別。

2.2 基於Transformer之模型 (Transformer-based Model)

過去文本情緒分析是使用基於統計的詞袋模型和靜態特徵的詞向量模型將文本轉為向量特徵(Barry, 2017; Han *et al*., 2013)，但這些方法會遇到無法解讀多義詞的瓶頸。歌詞被視為是敘事而非彼此獨立的句子，需要捕捉上下文的依賴關係，在歌詞的音樂情緒分類任務上，若基於傳統詞典進行效果有限(Hu & Downie, 2010; Hu *et al*., 2009)，Abdillah *et al*. (2020)運用能捕捉時序關係的雙向長短期記憶(Long Short-Term Memory，LSTM)對MoodyLyrics 資料集(Çano & Morisio, 2017b)進行歌詞的情緒分類，但遞歸架構難以具備平行運算的能力。Transformer (Vaswani *et al*., 2017)則改變過去序列網路的做法，自注意力機制藉由 Scaled dot-product Attention 讓資料得以平行運算， 考慮詞在不同空間映射的重要性，允許 BERT (Devlin *et al*., 2018)預訓練模型在多項任務中取得突破，包含實體辯識、序列或句子對分類、問答等 11 種任務，使得 Transformer-based 的模型架構在自然語言領域中成為主流。在歌詞情緒辨識的應用上，Agrawal *et al*. (2021)的研究便是使用基於 Transformer 的語言模型作為情緒分類的基礎架構，在多個英文歌詞情緒資料集上達到良好的成果，展示 Transformer-based 方法的高效能。

2.3 遷移學習 (Transfer Learning)

在某些領域中標籤的標記昂貴，若原始資料中含有標籤的數量太少，容易造成模型過度擬合。遷移學習常用的兩個技巧：特徵萃取和微調。遷移學習的有效性催生了多種應用，例如：學習情緒辯識(Hung *et al*., 2019)、時間序列任務(Fawaz *et al*., 2018)、3D 醫學影像分析(Chen *et al*., 2019)。在自然語言處理領域，也常運用遷移學習的技巧對於預訓練模型進行模型微調或特徵萃取，Transformer-based 的預訓練模型，證明微調在無註釋語料上預訓練大規模語言模型的有效性。Hung & Chang (2021)則提到多層遷移學習的有效性，表明了不管在電腦視覺任務或自然語言處理任務，經遷移學習的結果會優於未經過遷移的結果，因此，本篇研究提出的模型架構基於語言預訓練模型對文本進行遷移學習。

3. 方法論 (Methodology)

3.1 資料集 (Datasets)

- 中文維度情緒資料集(Yu *et al*., 2016; Yu *et al*., 2017)：資料如表 1 所示，包含中文情緒字典(Chinese Valence-Arousal Words, CVAW)、中文維度情緒片語(Chinese Valence-Arousal Phrases, CVAP)以及中文情緒語料庫(Chinese Valence-Arousal Text, CVAT)三個。CVAW 包含 5,512 個中文情緒詞；CVAP 中每個片語結合修飾符和來自 CVAW 中的詞，共 2,998 個中文情緒片語；CVAT 則從 720 篇來自 6 種不同類別的網路文章蒐集而來，共 2,009 個句子。三個資料集的每個詞或句子皆包含效價(Valence)和喚醒(Arousal)。效價(Valence)的範圍從 1 到 9 其分別代表極端負面和極端正面的情緒，喚醒(Arousal)的範圍從 1 到 9 其分別代表平靜和激動，效價(Valence)和喚醒(Arousal)若為 5 則代表沒有特定傾向的中性情緒。

- 歌詞資料集：為本研究自行收集並標籤的資料集。標籤包含象限一(Q1)、象限二(Q2)、象限三(Q3)及象限四(Q4)。Q1 代表正向激昂共 43 首，Q2 代表負向激昂共 45 首，Q3 代表負向平靜共 43 首，Q4 代表正向平靜共 39 首。V 和 A 分別代表效價(Valence)和喚醒(Arousal)，V 標記 1 代表正向情緒、0 代表負向情緒，A 標記為 1 代表激昂情緒、0 代表平靜情緒。

表 1. 中文維度情緒資料集
[Table 1. The datasets of Chinese valence-arousal]

名稱	總數	範例文字	Valence	Arousal
中文維度情緒字典 CVAW	5,512	不爽	2.8	7.2
中文維度情緒片語 CVAP	2,998	非常可愛	8	7.313
中文維度情緒語料庫 CVAT	2,009	這種記錄難免空洞，虛構也顯得薄弱。	3	3.5

3.2 提出之架構 (Proposed Architecture)

本研究提出的模型架構如圖 2，透過 BERT 預訓練模型建立 CVAT 中文維度情緒模型，將此模型直接用於歌詞情緒的標記，驗證在未學習過歌詞文本的情況下模型的成效。本章總共有三個小節，第一小節說明資料預處理，第二小節介紹模型實作的細節以及實驗的參數設定，第三小結討論將模型應用於歌詞文本情緒驗證的方法。

3.2.1 資料預處理 (Data Preproceessing)

CVAW、CVAP 和 CVAT 皆採用資料集內的文字、效價(Valence)平均和喚醒(Arousal)平均。由於 CVAW、CVAP 的文字較短且類似，因此將兩個資料集合併成一個資料集，稱 CVAW+CVAP，以 8 比 2 拆分為訓練集跟測試集。BERT 模型有別於傳統文本的方法，會將標點符號視為一個特徵值進行訓練，因此 CVAT 文字不進行刪除標點符號的預處理。歌詞的資料集共 170 首，由三位標註者將每首歌曲的針對效價(Valence)和喚醒(Arousal)分別標註為正或負，以中性情緒為原點，依照效價(Valence)和喚醒(Arousal)的正跟負分標記到四個象限。BERT 能夠訓練的最大文本長度為 512，考慮到 CVAP 和 CVAW 的文字都在 10 字以內，而 CVAT 的文本分佈大多集中在 100 字以內，為了避免產生過於稀疏向量，最大文本長度設定為 256 而非 512。輸入 BERT 模型前必須在每個序列開頭加上特殊字元符號[CLS]，此特殊字元代表整個輸入序列的向量表示，在序列尾巴則加上特殊字元符號[SEP]作為文本的結束，每個中文字會對應到 BERT 中文字典的一個索引值稱為 Token id，為了讓每一則輸入序列的長度保持一致，若文字長度不足則會在文字序列後端填充特殊字元[PAD]，最後轉為向量的序列和目標值轉為張量(Tensor)至 BERT 模型進行訓練。

3.2.2 實施細節 (Implementation Details)

本研究以知名的深度神經網路模型—BERT(Devlin *et al.*, 2018)為基礎架構，並進一步提出了多輸出(Multi-output)與單輸出(Single-output)兩種模型訓練架構。如圖 2 所示，多輸出(Multi-output)架構為一個 BERT 模型共享權重，一次輸出效價(Valence)和喚醒(Arousal)兩個連續值，單輸出(Single-output)為一個 BERT 模型輸出單一個值(例如 Valence)。由於是效價(Valence)、喚醒(Arousal)的數值預測，因此模型訓練時的損失函數選擇使用均方誤差(Mean square error, MSE)。兩種模型架構都實驗兩種方法：(a)使用從 CVAW+CVAP 遷移至 CVAT 資料集的遷移學習方法。(b)從零訓練 CVAT 未遷移的方法。最後進行兩種方法的比較。本研究基於微調方法進行實驗，微調方法的優點在於模型的許多參數不需要重新學習，即使只有少量訓練樣本也能達到良好的效果。在模型微調方面，在 BERT 預訓練模型加上一層 Dropout 和一層線性分類層，優化器為 Adam，學習速率本研究嘗試多種學習速率進行實驗，微調模型適合較小的學習速率避免預訓練的權重被修改破壞，也在實驗中發現，若學習速率不夠小會導致損失(loss)無法降低，最後選擇了 1e-05、1e-06 和 5e-05 三個超參數進行近一步實驗及比較，每次訓練最大 Epoch 設定為 100，加入 Early Stopping 的機制，將耐心(Patience)設至為 10。

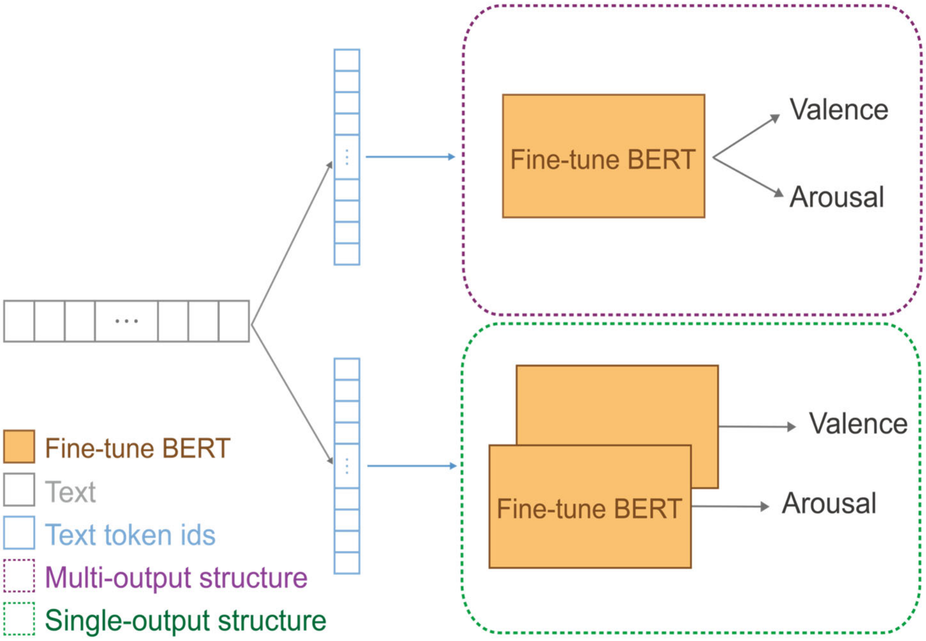


圖 2. 本研究提出之模型架構：包含兩種架構分別為單輸出與多輸出
[Figure 2. Training architecture of multi-output and single-output models]

3.2.3 歌詞情緒之分類 (Lyrics Emotion Classification)

此階段的目的在於驗證本研究提出的方法能在未學習過歌詞文本的情況下，能對於歌詞文本進行情緒的標註。首先，將歌詞文本進行與第一小節同樣的預處理後送入模型進行預測，輸出效價(Valence)和喚醒(Arousal)，其範圍為 0 到 9。本研究依照原資料集的敘述(Yu *et al*., 2016; Yu *et al*., 2017)，效價(Valence)和喚醒(Arousal)都以中性值 5 為閾值，因此，若預測出的效價(Valence)數值大於 5 表示模型預測該歌詞為正向情緒並標記為 1、效價(Valence)數值小於 5 則表示模型預測該歌詞為負向情緒並標記為 0，若預測喚醒(Arousal)數值大於 5 則表示模型預測該歌詞為激動情緒並標記為 1、喚醒(Arousal)值小於 5 表示模型預測該歌詞為平靜情緒並標記為 0。我們將效價(Valence)和喚醒(Arousal)標記之後的結果轉為四個象限 Q1、Q2、Q3 和 Q4 的情緒分類之結果，最後驗證其分類效果。

4. 實驗結果 (Experimental Result)

本章節將實驗結果分為兩個階段，第一階段是中文情緒的模型訓練結果，第二階段是驗證模型預測歌詞情緒的成效，每個段落包含在不同的模型架構和不同訓練方式的實驗結果。

4.1 中文情緒模型 (Chinese Emotion Model)

訓練模型的資料集切分皆以 8 比 2 進行，CVAP+CVAW 的訓練集和測試集分別為 6808 筆和 1702 筆。在多輸出(Multi-output)與單輸出(Single-output)架構的訓練結果，如表 2 所示，多輸出架構模型的均方誤差為 0.59126，單輸出模型的效價(Valence)和喚醒(Arousal)的均方誤差(MSE)分別為 0.3788 和 0.77339，兩個模型架構最佳的學習速率皆為 1e-05。

表 2. 在 CVAW + CVAP 資料集上之結果：包含多輸出(Multi-output) 與單輸出 (Single-output) 兩種架構之結果
[Table 2. Results on multi-output and single-output models]

架構名稱	輸出	學習速率 (Lr)	損失 (loss)	Epoch
多輸出架構 Multi-output	-	1e-5	**0.59126**	14
		1e-6	0.65283	32
		5e-5	0.69301	17
單輸出架構 Single-output	Valence	1e-5	**0.3788**	24
		1e-6	0.39498	35
		5e-5	0.51918	4
	Arousal	1e-5	**0.77339**	12
		1e-6	0.92874	19
		5e-5	1.8867	12

如表 3 所示在多輸出架構底下，從零訓練 CVAT 資料集(Training From Scratch)和從 CVAP+CVAW 資料集模型遷移至 CVAT 資料集(Transfer Learning)的結果來看，兩者同樣都在學習速率皆為 1e-05 的訓練效果最佳，經過遷移的均方誤差為 0.65696 優於未經遷移的 0.72025。經過遷移的模型結果比未經遷移效果好，在不同學習速率下，經遷移的 CVAT 在不同的學習速率下都優於未經遷移的結果，有經過遷移學習的 CVAT 收斂速度也比未經遷移學習的快。表 4 為在單輸出架構底下，從零訓練 CVAT 資料集(Training From Scratch)和從 CVAP+CVAW 資料集模型遷移至 CVAT 資料集(Transfer Learning)的結果，單輸出(Single-output)架構是將效價(Valence)和喚醒(Arousal)作為獨立的兩個輸出，首先比較效價(Valence)輸出的結果，未經遷移(Training From Scratch)的均方誤差(MSE)為 0.50338，而經遷移學習(Transfer Learning)的均方誤差(MSE)為 0.46624，顯示經遷移學習的 CVAT 其結果優於未經遷移的結果。經遷移學習的最佳學習速率為 1e-06，未經遷移的最佳學習速率為 1e-5，就算同樣都在 1e-5 的學習速率底下，經遷移學習的均方誤差 0.47898 依然是優於未經遷移的均方誤差 0.50338。比較輸出為喚醒(Arousal)的結果，經遷移學習的 CVAT 其均方誤差為 0.84259 優於未經遷移的 0.87107，兩者同樣都在學習速率為 1e-05 的時候得到最佳結果。

表 3. 多輸出架構(Multi-output)經遷移學習與未經遷移學習在 CVAP 資料集之結果
[Table 3. CVAP results on the multi-output model with/without transfer learning]

方法	學習速率 (Lr)	損失 (loss)	Epoch
從 0 訓練 CVAT From Scratch	1e-5	**0.72025**	10
	1e-6	0.73979	58
	5e-5	0.80925	10
經遷移學習 Transfer Learning	1e-5	**0.65696**	3
	1e-6	0.67836	22
	5e-5	0.70594	2

表 4. 單輸出架構(Single-output)經遷移學習與未經遷移學習在 CVAP 資料集之結果
[Table 4. CVAP results on the single-output model with/without transfer learning]

方法	輸出	學習速率 (Lr)	損失 (loss)	Epoch
從 0 訓練 CVAT From Scratch	Valence	1e-5	**0.50338**	12
		1e-6	0.51199	44
		5e-5	0.55236	6
	Arousal	1e-5	**0.87107**	5
		1e-6	0.93317	28
		5e-5	0.9303	10
經遷移學習 Transfer Learning	Valence	1e-5	0.47898	4
		1e-6	**0.46624**	15
		5e-5	0.53422	5
	Arousal	1e-5	0.84259	1
		1e-6	0.88142	7
		5e-5	0.93479	11

4.2 中文歌詞情緒模型之驗證 (Verification of Chinese Lyrics Emotion Model)

此段落討論前述的中文情緒模型應用於中文歌詞分類之結果，第一小節描述中文歌詞的分類結果，第二小節更進一步說明 Valence-Arousal 平面的分類結果。中文歌詞的情緒分類是將模型輸出的效價(Valence)和喚醒(Arousal)基於中性值 5 作為閾值，轉換為座標平面上的四個象限類別(Q1、Q2、Q3、Q4)。歌詞情緒分類結果如表 8 所示，包含歌名、歌詞、模型預測的 Valence 數值和 Arousal 數值、預測的標籤和真實標籤。

4.2.1 中文歌詞情緒分類之結果 (Chinese Lyrics Emotion Classification Results)

經遷移學習 CVAT 模型與未經遷移學習 CVAT 模型的歌詞情緒分類的混淆矩陣的結果，如表 5 所示，在經遷移學習的 CVAT 分類，Q1 有 26%容易被錯分成 Q2，19%會被錯分成 Q4，僅有 2.3%會被分成 Q3，也就是在 Q1 的情緒類別中，效價(Valence)和喚醒(Arousal)都有被分類錯的可能，效價(Valence)和喚醒(Arousal)同時被錯分的機率僅 2.3%。Q4 有 56% 被錯分為 Q3，被分成 Q2 的可能為 5%，僅有 2.5%會被分成 Q1，也就是在 Q4 的情緒類別中，僅效價(Valence)容易被分類錯誤。Q2 的情緒幾乎都能夠準確辨識，僅有 2%會被錯分成 Q3，只有 2%會因為喚醒(Arousal)被錯誤分類。Q3 有 16%會被錯分成 Q2，其餘的都能被正確分類，也就是在 Q3 的情緒只有 16%會因為喚醒(Arousal)被錯誤分類。

未經遷移的 CVAT 分類結果中，Q1 有 14%被錯分成 Q2，25.6%被錯分成 Q3，35%被錯分成 Q4，Q2 只有 62%分類正確，其餘 37%皆被錯分成 Q3，Q3 有 14%被錯分成 Q2，其餘分類正確，Q4 有 7.7%被錯分成 Q2，69%容易被錯分成 Q3 ，只有 29%分類正確。

經遷移學習 CVAT 模型與未經遷移學習 CVAT 模型的歌詞情緒分類結果，如表 6 所示，經遷移學習的 CVAT 模型在歌詞情緒分類的準確度為 0.71，標籤 Q1 和 Q4 的 F1-score 較低，分別為 0.69 和 0.51，而 Q2 和 Q3 的 F1-score 較高，分別為 0.83 和 0.72。未經遷移學習的 CVAT 模型在歌詞情緒分類的準確度為 0.50，同樣是標籤 Q1 和 Q4 的 F1-score 較低，分別為 0.41 和 0.29，而 Q2 和 Q3 的 F1-score 較高，分別為 0.64 和 0.55。比較經遷移學習的模型與未經遷移學習的模型，經遷移學習的模型中每一個情緒標籤的分類結果都優於未經遷移學習的模型，可得知到在訓練階段 CVAT 模型學習效果較佳的模型，應用在歌詞的情緒分類也能得到較佳的結果，表示經遷移學習的模型在 CVAW+CVAP 資料集中所學習到的中文情緒特徵，有助於提升模型在歌詞文本的情緒辨識能力。

表 5. 歌詞情緒分類之混肴矩陣：經遷移學習與未經遷移學習
[Table 5. Results on the model with/without transfer learning by confusion matrix]

Prediction by CVAT: Transfer Learning					
True		Q1	Q2	Q3	Q4
	Q1	**23**	8	1	11
	Q2	0	**44**	1	0
	Q3	0	7	**36**	0
	Q4	1	2	**19**	17
Prediction by CVAT: Training from Scratch					
True		Q1	Q2	Q3	Q4
	Q1	**11**	6	11	15
	Q2	0	**28**	17	0
	Q3	0	6	**37**	0
	Q4	0	3	**27**	9

表 6. 歌詞分類結果之分數：經遷移學習與未經遷移學習
[Table 6. Results on the model with/without transfer learning]

CVAT : Transfer Learning			
Label	Precision	Recall	F1-score
Q1	0.96	0.53	0.69
Q2	0.72	0.98	0.83
Q3	0.64	0.84	0.72
Q4	0.61	0.44	0.51
Accuracy	0.71		
CVAT: Training From Scratch			
Label	Precision	Recall	F1-score
Q1	1.00	0.26	0.41
Q2	0.65	0.62	0.64
Q3	0.40	0.86	0.55
Q4	0.38	0.23	0.29
Accuracy	0.50		

4.2.2 Valence-Arousal 分類之結果 (Valence-Arousal Plane Classification Result)

此小節將說明效價(Valence)和喚醒(Arousal)的預測結果。經遷移學習的 CVAT 模型在 Valence-Arousal 平面之分類結果，如表 7 所示效價(Valence)和喚醒(Arousal)的分類準確率都為 0.76。在效價(Valence)的負向情緒中，Recall 為 1，表示 80 首負向情緒的歌詞都被正確分類，Precision 為 0.67，表示被預測為負向的歌詞總共有 120 首，有 80 首被正確分為負向情緒，但有 40 首標籤應為正向情緒的歌詞被錯誤分類為負向情緒。效價(Valence)的正向情緒中，Recall 為 0.56，表示有 50 首正向情緒的歌詞被正確分類，但有 40 首正向情緒的樣本被分類為負向情緒，Precision 為 1，表示被預測為正向情緒的歌詞總共有 50 首，而 50 首皆被正確分類。在喚醒(Arousal)的激動情緒中，Recall 為 0.71，表示總共有 105 首歌詞應被分類為激動情緒，有 75 首被正確分類、30 首應該被分類為激動情緒的歌詞被錯誤分類為平靜情緒，Precision 為 0.88，表示被預測為激動情緒的歌詞總共有 85 首，其中，有 75 首歌詞被正確分類為激動情緒，但有 10 首被錯誤分類為激動情緒。在喚醒(Arousal)的平靜情緒中，Recall 為 0.85，表示共有 65 首歌詞應被分類為平靜情緒，有 55 首被正確分類，有 10 首應被分成平靜情緒的歌詞被錯誤分類為激動情緒，Precision 為 0.65，表示被預測為平靜情緒的歌詞總共有 85 首，55 首歌詞被正確預測為平靜情緒，但有 30 首被錯誤分類為平靜情緒。

表 7. 經遷移學習 CVAT 模型之 Valence-Arousal 平面分類之結果
[Table 7. Classification results on the four categories of valence-arousal]

Valence			
Label	Precision	Recall	F1-score
1 (+)	1.00	0.56	0.71
0 (-)	0.67	1.00	0.80
Accuracy	0.76		
Arousal			
Label	Precision	Recall	F1-score
1 (+)	0.88	0.71	0.79
0 (-)	0.65	0.85	0.73
Accuracy	0.76		

5. 討論 (Discussion)

從單輸出模型架構的結果發現喚醒(Arousal)的特徵較難學習，該結果在多個研究中都有提到(Malheiro *et al*., 2016; Yu *et al*., 2016; Çano & Morisio, 2017b)，發現在無論中文或者英文的資料上，文字的喚醒(Arousal)維度較難以區分，推測激動程度或強度在文字上較難以顯示出來。不論在多輸出還是單輸出的模型架構底下，經過遷移的結果都優於未經遷移的結果且能提高模型的收斂速度，證明在 CVAW 和 CVAP 兩個資料集所學習到的特徵，有助於模型對 CVAT 中文維度情緒語料庫的學習。在驗證模型能否應用於歌詞文本的實驗結果中觀察到，CVAT 訓練結果較佳的遷移模型，應用於歌詞文本分類使其結果也會較佳，優於未遷移的 CVAT 模型，顯示經遷移學習學到的特徵是有助歌詞文本的情緒辨識成果，並且在未學習過歌詞文本的狀況下達到 71%的準確率。

另外，我們推測分類錯誤的原因是歌詞中的情緒在不同段落中可能帶有不同情緒，因此，較難以分類為單一情緒類別。以表 8 中，其真實標籤(Label)為正向激動(Q1)但模型預測(Predict)為負向激動(Q2)的「美好」這首歌曲進行每個段落的歌詞情緒分析，結果如表 9 所示，若只看段落一的句子中「你哭著說再見」、「揮手和你道別」會解讀出字句裡包含負面情緒，模型如期預測為負向情緒，相反的在段落六的句子中，「你是全世界的美好」、「你有最美麗的微笑」、「比你更重要就是和你 一起變老」，句子間流露出滿足、幸福等正向的情緒，模型如期預測為正向情緒。在段落三與段落五當中，兩段歌詞的標籤都是正向激昂的(Q1)，兩段歌詞差別只在於最後一句的不同，分別是「我不知道也永遠不要知道」和「我不知道也不用知道」，比較兩句話對於模型預測喚醒(Arousal)程度的影響，可以看出段落三的激動情緒為 5.40，略高於段落五的 5.371，我們推測第三段的歌詞，其關鍵字眼「永遠」提高整句話的激昂程度。在段落二與段落四當也能看到

類似的結果，段落二比起段落四多一句「對不對」，「對不對」表達強調或質問的語氣，因此加強句子的喚醒(Arousal)程度，從模型預測結果來看，段落二的喚醒(Arousal)也從4.88 提升至 4.94。

總結以上的結果說明本研究提出的模型能辨別出不同句子的正、負向情緒和字句間喚醒(Arousal)程度的差別，同時，我們觀察到在某些歌詞中其實隱含多種情緒，若歌詞只進行單一類別標註可能會對於情緒的標記不夠全面，因為當不同標註者對於歌詞關注的地方不同，標註的情緒也會有所不同。

表 8. 中文歌詞情緒分類之結果
[Table 8. Classification results on Chinese lyrics emotion]

歌曲	歌詞（未顯示完整歌詞）	V	A	Predict	Label
美好	我看見風吹過你的臉 有那年的感覺 那一年你哭著說再見 而我揮手和你道別	4.936	5.038	Q1	Q1
厭世吉娃娃	吉娃娃 吉娃娃 我是吉娃娃 我討厭爸爸 我討厭媽媽 我更討厭這個世界啊	2.654	6.474	Q2	Q2
你	風輕輕 我聽見你聲音 你對著我叮嚀 要注意自己的心情 風輕輕 我聽見你聲音 你拿著傘靠近	5.800	4.836	Q4	Q4
哼情歌	在無關景要的場合都會想起這首歌 是因為你曾經哼唱著 在平淡無奇的眼神都會想起你呢	4.360	4.809	Q3	Q3

表 9. 歌曲「美好」依照句子分割之 Valence-Arousal 預測結果
[Table 9. Prediction of valence-arousal by the example sentences from a song]

段落	歌詞	V(+/-)	A(+/1)
1	我看見風吹過你的臉 有那年的感覺 那一年你哭著說再見 而我揮手和你道別	4.509(-)	5.07(+)
2	時間並沒有帶走一切 反而給我們更多洗鍊 如果分離製造了更多的想念 那我們是否該更加感謝	5.156(+)	4.88(-)
3	你是全世界的美好 你有最美麗的微笑 你問我有什麼比你重要 我不知道也永遠不要知道	5.113(+)	5.40(+)
4	時間並沒有帶走一切 反而給我們更多洗鍊 如果分離製造了更多的想念 那我們是否該更加感謝 對不對	4.583(-)	4.94(-)
5	你是全世界的美好 你有最美麗的微笑 你問我有什麼比你重要 我不知道 也不用知道	5.766(+)	5.371(+)
6	你是全世界的美好 你有最美麗的微笑 你問我有什麼比你重要 總算知道 比你更重要 就是和你 一起變老	6.094(+)	5.226(+)

6. 結論 (Conclusion)

本研究提出以基於 Transformer 的語言預訓練模型對中文情緒資料集進行學習，將中文情緒資料庫的模型直接用於歌詞的效價(Valence)和喚醒(Arousal)進行標註。在實驗中比較了有遷移學習與未經遷移學習的模型，實驗結果證明在中文情緒字典與中文情緒片語學習到的特徵，有助於中文情緒文本的學習。同時，本研究將經遷移學習及未遷移的模型用於歌詞的情緒分類，發現經遷移學習的模型結果優於未經遷移的模型，證明在中文情緒資料集學習結果較佳的模型，用於歌詞情緒分類其結果也會較佳。在不同模型架構底下，發現喚醒(Arousal)的情緒較難以學習。最後，我們注意到歌詞中的情緒在不同段落會帶有不同情緒，不僅模型在人工標註時，若對於關注到不同語句則會對於歌詞情緒有不同的判斷，在未來研究方向可以將一首歌的歌詞視為多種情緒以多標籤任務方式進行。

致謝 (Acknowledgements)

特別感謝科技部計畫(108-2218-E-025-002-MY3)，「人工智慧音樂家-運用深度嵌入方法與生成對抗網路與和諧性為導向的詞曲生成器之設計」對於本研究的支持。

參考文獻 (References)

Abdillah, J., Asror, I., & Wibowo, Y. F. A. (2020). Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting. *Jurnal RESTI* (Rekayasa Sistem Dan Teknologi Informasi), *4*(4), 723-729. https://doi.org/10.29207/resti.v4i4.2156

Agrawal, Y., Shanker, R. G. R., & Alluri, V. (2021). Transformer-based approach towards music emotion recognition from lyrics. arXiv preprint arXiv:2101.02051.

Barry, J. (2017). Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches. In *AICS*, 272-274.

Chen, S., Ma, K., & Zheng, Y. (2019). Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2018). Transfer learning for time series classification. arXiv e-prints, arXiv:1811 .01533.

Han, B. J., Rho, S., Dannenberg, R. B., & Hwang, E. (2009). SMERS: Music Emotion Recognition Using Support Vector Regression. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 651-656.

Han, Q., Guo, J., & Schuetze, H. (2013). Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 520-524.

Hu, X., & Downie, J. S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In *Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 619-624.

Hu, Y., Chen, X., & Yang, D. (2009). Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 123-128.

Hung, J. C., Lin, K. C., & Lai, N. X. (2019). Recognizing learning emotion based on convolutional neural networks and transfer learning. *Applied Soft Computing*, 84, 105724. https://doi.org/10.1016/j.asoc.2019.105724

Hung, J. C., & Chang, J. W. (2021). Multi-level transfer learning for improving the performance of deep neural networks: Theory and practice from the tasks of facial emotion recognition and named entity recognition. *Applied Soft Computing*, 109, 107491. https://doi.org/10.1016/j.asoc.2021.107491

Jamdar, A., Abraham, J., Khanna, K., & Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. arXiv preprint arXiv:1506.05012.

Kim, J., Lee, S., Kim, S., & Yoo, W. Y. (2011). Music mood classification model based on arousal-valence values. In *Proceedings of 13th International Conference on Advanced Communication Technology (ICACT 2011)*, 292-295.

Laurier, C., Sordo, M., Serra, J., & Herrera, P. (2009). Music Mood Representations from Social Tags. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 381-386.

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, *9*(2), 240-254. https://doi.org/10.1109/TAFFC.2016.2598569

Qiu, L., Chen, J., Ramsay, J., & Lu, J. (2019). Personality predicts words in favorite songs. *Journal of Research in Personality*, 78, 25-35. https://doi.org/10.1016/j.jrp.2018.11.004

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161- 1178. https://doi.org/10.1037/h0077714

Sarkar, U., Nag, S., Basu, M., Banerjee, A., Sanyal, S., Sengupta, R., & Ghosh, D. (2021). Neural Network architectures to classify emotions in Indian Classical Music. arXiv preprint arXiv:2102.00616.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on neural information processing systems (NIPS '17)*, 5998-6008.

Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, *16*(2), 448-457. https://doi.org/10.1109/TASL.2007.911513

Yu, L. C., Lee, L. H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. R., & Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, 540-545. https://doi.org/10.18653/v1/N16-1066

Yu, L. C., Lee, L. H., Wang, J., & Wong, K. F. (2017). IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases. In *Proceedings of the IJCNLP 2017, Shared Tasks*, 9-16.

Çano, E., & Morisio, M. (2017a). Music mood dataset creation based on last. fm tags. In *Proceedings of Fourth International Conference on Artificial Intelligence and Applications (AIAP 2017)*, 15-26. https://doi.org/10.5121/csit.2017.70603

Çano, E., & Morisio, M. (2017b). Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI '17)*, 118-124. http://dx.doi.org/10.1145/3059336.3059340

Computational Linguistics and Chinese Language Processing
Vol. 26, No.2, December 2021, pp. 17-34

A Pretrained YouTuber Embeddings for Improving Sentiment Classification of YouTube Comments

Ching-Wen Hsu[*], Hsuan Liu[*], and Jheng-Long Wu[*]

Abstract

Technology is changing the way we consume information and entertainment. YouTube streaming video services provide a discussion function that allows video publishers to know what matters most to the people they want to love their brand. Through comments, video publishers can better understand the audience's thoughts and even help video publishers improve their video quality. We propsoe a classifier based on machine learning and BERT to automatically detect YouTuber preferences, video preferences, and excitement levels. In order to make high performance of models, we use a pretrained YouTuber embeddings to enhance performance, which is trained in advance based on roughly 175,000 pieces of videos' comments that contain YouTubers' name. YouTuber embeddings can capture some of the semantics and character of the relation between YouTubers. Experimental results show that the performances of machine learning-based models with YouTuber embeddings have improved overall accuracy and F1-score on all sentiment classications. The result validates that YouTuber embedding training is significantly helpful when detecting audience sentiment towards YouTubers. On the contrary, BERT model cannot perfectly deal with the polarity classificational tasks when using YouTubers embeddings. However, the BERT model construction is more suitable for addressing multi-dimensional classification tasks, such as the five-labels classification task used in this task. Conclusion, the sentiment detection task on the YouTube can improve performance by the proposed multi-dimensional sentiment indicators and our solution to modify the structure on classifiers.

Keywords: YouTuber Embeddings, Sentiment Classification, Deep Learning, Pretrained Model

[*] School of Big data Management, Soochow University
E-mail: 06170139@gm.scu.edu.tw; 06170114@gm.scu.edu.tw; jlwu@gm.scu.edu.tw
The author for correspondence is Jheng-Long Wu.

1. Introduction

Due to the rapid rise of new media, streaming platforms and video providers have increased. According to one report, 68% of people prefer watching a video rather than reading a long product manual to acquire information. People change the way of their entertainment even daily habitual. No one wants to be tied to a TV schedule, so people nowadays favor subscribing to streaming video services, such as Netflix or YouTube, to enjoy watching videos anytime and anywhere. Also, mobile phone viewers or smartphone viewers have increased astonishingly. YouTube reports that mobile video consumption is rising with an impressive rate of 100 percent every year. The large amount of data captured by video platforms provides insights for video streaming apps, and video stream services make recommendations based on audience's viewing profiles. Because of High-speed internet connectivity, more and more people have been allowed to become YouTubers and create large volumes of high-quality videos. YouTube has 16 million active users in Taiwan monthly, and nearly 93% of users have visited YouTube. It seems that YouTube has played an increasingly important role in modern life and entertainment. Therefore, we aim to analyze audience's habitual preferences on consuming information and entertainment on YouTube.

According to the audience's watching records, YouTube can create customized recommended content, which means consumers' interests have been collected and analyzed by YouTube. On the contrary, YouTubers, who upload videos to the YouTube platform, also want to check their videos' performance. YouTube has provided several analysis functions such as average view duration, browsing history, variance in audience's demographics, Etc for YouTubers to check their channel's performance. However, it lacks sentiment analysis on the audience's comments. It is verified that public views, comments, and attitudes towards many events can be analyzed through social media (Heredia *et al.*, 2016). Public reviews on Amazon were used to evaluate users' opinions and determine the audience's preference by classifying opinions into negative, positive, and neutral (Bhatt *et al.*, 2015). Therefore, we deduced that YouTube could also serve as a sentiment analysis platform because it provides an increasing number of comments.

We utilized comments to monitor YouTube viewers' emotions in the previous task by designing three sentiment indicators, YouTuber preference, video preferences, and excitement level. In this task, we are not changing sentiment indicators but aim to optimize the result of sentiment detection, hoping to get higher overall accuracy to analyze audience's feelings. We not only use comments to monitor emotions, but we also consider characteristics in YouTube channels as an additional feature. Before restarting the experiment, we trained YouTubers' correlation and established YouTuber embeddings, a critical vector in determining what characteristic YouTubers shared between each other. Also, the similarity between different channels can be calculated by placing similar YouTubers close together in the embedding space.

The social sentiment is excellent in providing a better understanding of how their audience perceives the YouTuber channel or brand. In general, sentiment analysis focuses on determining positive, negative, or neutral emotions (Cunha *et al*., 2019). Therefore, before this task, we also conducted some experiments that used YouTube comments to identify users' positive, negative, or neutral emotions and how strong those emotions are. Unlike previous tasks, we change our method in the experiment stage. We combine comments and use our established YouTuber word embedding. Not only to capture emotions behind everything social viewers but also to measure YouTubers intimately by translating YouTubers' features into a relatively low-dimensional space. The analyzing result may help video loaders who want to identify their viewers' depth of feeling and provide a chance for YouTubers to engage with their viewers directly.

By modifying the structure of models that contain pre-trained YouTuber word embeddings as part of the sentence input, we expect a better model's performance than the previous tasks, not containing pre-trained YouTuber word embedding. Anticipate that YouTuber word embedding can provide additional information when analyzing sentiment tasks.

2. Related Work

Various models deal with text-based sentiment classification tasks. Machine learning-based models are used to address the text classification task (Zhang & Zheng, 2016). Other deep learning models have been used for sentiment analysis and obtained acceptable performances (Hassan & Mahmood, 2017). Recently, it has refreshed the best performance of using pre-trained language models, such as Bidirectional Encoder Representations from the Transformers (BERT) because its pre-trained method has captured linguistic structure from learning and detecting different tasks. Sun *et al.* (2019) have explored BERT pre-trained structure to deal with classification task and achieve excellent performance through the way of fine-tuning in the downstream tasks. In our previous task, we also fine-tuning BERT model to detect a multi-dimensional aspect of the audience's comments. Although the experiment results outperformed machine learning-based classifiers and even had similar outcomes in the deep learning-based classifier, it may lack task-specific knowledge and domain-related knowledge to further improve the BERT model's performance. Considering viewers may present different passion intensities through many kinds of channels they watch, so we take channels' information, which means the types and features of YouTubers, into consideration. For example, YouTubers who always share ironic videos, their viewers may reflect stronger emotions than educational videos. Peters *et al.* (2018) realized that word representations are key component in many neural language understanding models, so they introduced a new type of word representations which can deal with syntax and semantics. However, our way of dealing with complex characteristics of word use is adding YouTubers' information into each comment. We proposed pre-trained YouTuber embeddings to fully present domain-related knowledge in YouTube, so we can confirm whether characteristic of YouTubers can improve models' comprehension. Specifically, we concatenate the original sentence embedding and YouTuber embeddings which serve as additional features when analyzing comments' emotion tendency.

Compared with the limited dataset for training a relatedness between terms, more researchers have focused on using a pre-trained word embedding to understand semantic relatedness and similarity between terms in recent years. Zhu *et al*. (2017) show that increasing the size of datasets can identify more relations of biomedical terms even though it does not guarantee models' better precision. As a result, because of the small size of dataset, researcher often have to use pre-trained word embeddings to better capture meaningful vectors. Rezaeinia *et al*. (2017) have increased the accuracy on sentiment analysis research by using pre-trained word embeddings. Their method is experience different word representation methods, such as Part-of-Speech (POS) tagging techniques, lexicon-based approaches, and Word2Vec/Glove methods to compare their effectiveness.

Recently word embeddings methods have been widely applied in downstream models. Aydoğan & Karci (2020) used Word2Vec method on a large corpus of approximately 11 billion words to train word vectors, then applied to deep neural networks. The result did show that embedding method affected the rate of accuracy. Another research used pre-trained word embedding as a critical component for its downstream models. (Miyato *et al*., 2017). Cited from the above experiences, initially, we decided to utilize word vectors from the 2021 Wikipedia Chinese corpus to represent YouTuber similarity because of the large size of corpuses. However, we only focus on capturing the strong connection between each YouTuber and extracting characteristic behind YouTubers. According to our selected 25 YouTuber's channels, we select comments beneath each channel latest ten videos. Then, we filter these substantial comments by checking whether comments involve different YouTubers' names. Comments that up to standard are remained to train YouTubers embeddings. To compare whether the sentiment detection tasks can perform better by adding generating exact vectors, we propose a novel method, concatenating comments with YouTubers embeddings, to apply on classifiers.

Usherwood & Smit (2019) focus on comparing BERT and top classical machine learning approaches on a trinary sentiment classification task. Their task aims to verify whether BERT can perform state-of-the-art result when one only has 100-1000 labelled examples per class. As the result, BERT outperformed top classical machine learning algorithms even when training with 100 examples per class. Another research shows the superiority of BERT and support to use BERT as a default technique in NLP problems (González-Carvajal & Garrido-Merchán 2020). With similar task, we apply our own generated word vectors and go on the previous algorithms to determine if these approaches may represent the better result or even both BERT and machine learning-based methods are valid options.

3. Methodology

Figure 1 shows the proposed method for sentiment analysis and classification processes. Firstly, we collected the audience's comments from the YouTube platform and subsequently labeled these comments according to our designed three sentiment indicators. Data preprocessing works include transferring emojis to texts and establishing a YouTube-based dictionary for tokenization. Next, all comments are converted into vectors, and YouTuber embedding is prepared to concatenate in the proper layer according to models. Finally, by the experiment stage, we evaluate the performance of each classifier in three detection tasks and discuss a comparative study.

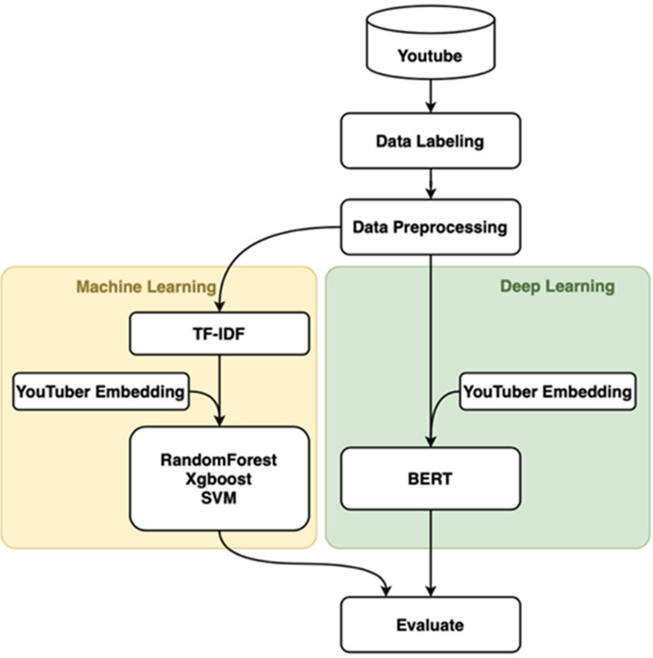


Figure 1. The process of the proposed sentiment analysis in this paper

3.1 Comment Collection

To cover the diversity of YouTube channels, we generated our dataset by selecting different types of YouTube channels. The composition of the selected videos' film creation types with game 1%, education 4%, DIY with 4%, science and technology with 5%, comedy 9%, entertainment with 28%, and blog with 49%. Through these 25 selected channels, we then filter five videos from each channel that have been highly popular or controversial since 2019 because people imminently show their interest in new tread and debatable topics. Therefore, the data source contains a total of 125 videos. In this way, we collected more controversial and polarizing comments, and it becomes easier for annotators to determine the sentimental tendency of comments. However, to avoid different accumulated numbers of comments in each video, we randomly remain 100 pieces of comments from each video. Thus, a total of 12500 pieces of comments is taken into consideration.

3.2 Definition of Sentiment Indicators

YouTube has provided a discussion function for audiences to express their opinion by clicking like or dislike bottom under the videos. However, positive or negative sentiment classifications cannot explain why the audience does not like the videos and what reason keeps the audience subscribing to a specific channel. There is no noticeable analysis of likes and dislikes opinion, so we design three indicators, YouTube preference, video preference, and excitement level, to investigate different aspects of the audience's comment.

- **YouTuber preference**: In the indicator of YouTuber preference, comments can roughly divide into non-relative and relative towards YouTubers. Excluding non-relative comments, the rest comments that talk about YouTubers' names or affairs can continue to dig into positive, negative, and neutral attitudes according to their comments' content.
- **Video preference**: The indicator categories are the same as YouTuber preference. Non-relative, unlike, neutral, and like are four categories used to judge Video preference. For example, comments that do not talk about video content will be labeled as non-relative comments.
- **Excitement level**: This indicator is designed into five categories, from barely excited to hyper excited. We classify the audience's speaking tone from no emotion to extreme emotion state step by step. In addition, we consider emojis a judgment in this indicator because people tend to use emojis as their comments. For example, the second level of Excited level means the audience can speak confidently and contain two types of emojis.

3.3 Sentiment Indicator Labeling

The main drawback of using own data sources is having to label our dataset. Therefore, the main objective is to address semantic comprehension gaps between annotators. We introduce some guidelines to properly annotate our comments. For example, watching videos before annotation is required because it might resonate powerfully with the audience's opinions. During the annotation process, we eliminate some non-relative comments, such as advertisements, comments that not using Mandarin, comments that post links to external web pages, and merely timestamps in the comments, to optimize the availability of the dataset. In the last part, we use the majority decision to filter out inconsistent labels unless each comment annotation is marked as the same point.

In Table 1, we use three methods to calculate agreement scores after labeling comments, which include Krippendorff's Alpha, Fleiss's Kappa, and Cronbach's Alpha. With

Krippendorff's Alpha method, due to the reason that values smaller than 0.667 represent as discard data, so our three indicators are shown not up to the standard. Fleiss's Kappa method stands for fair and moderate data because values between 0.21 to 0.6 are considered acceptable levels. Cronbach's Alpha method evaluates three indicators as outstanding labeling work because a value higher than 0.7 may show annotation agreement, let alone we get 0.9 on Excitement level. Therefore, two of the methods were qualified as acceptance results, and thus we provide an adequately labeled dataset to train and assess a given model.

Table 1. Annotation agreement scores for each indicator.

	YouTuber preference	Video preference	Excitement level
Krippendorff's Alpha	0.5829	0.4545	0.3898
Fleiss's Kappa	0.5840	0.4594	0.3928
Cronbach's Alpha	0.8520	0.7264	0.900

3.4 Text Preprocessing

We consider emojis as part of emotional expressions. The first step of text processing is to transfer emojis to text, so dealing with rich emojis is our priority. We transfer emojis to text by the package called “emojiswitch.” Then we establish a user-defined dictionary to recognize specific words, such as YouTubers’ names and the texts transferred from emojis. In this way, we go through word tokenization, and thus now we can accurately determine unique objects from a user-defined dictionary. After these two parts, we go through word tokenization using the current state-of-the-art word tokenization tool created by the Chinese Knowledge and Information Processing (CKIP) Group. This tool is available for dealing with tokenization in Mandarin. In previous task, after completing all these above steps, we can start model training and evaluating.

3.5 Training YouTuber Embedding

There are various sentiment analysis techniques, but recently, word embeddings have been widely used in sentiment classification tasks. Word2Vec and GloVe are among the most accurate and usable word embedding methods to convert words into meaningful vectors. Therefore, we trained YouTubers embedding, a dense vector representation of words that capture something about their meaning, to present meaningful vectors to understand the relationship between YouTubers.

To have the best results when using the generated embeddings, we selected ten newly released videos, due to October 2021, from 25 YouTube channels that are the primary data source in this task. The comments’ contents are selected based on having YouTubers’ names, whether lead actors/actresses or supporting actors/actresses. A total of 175,000 pieces of comments remains and applied to train YouTuber correlations. As a result, we present a YouTuber embedding dictionary that stores YouTubers' names and their corresponding 300-dimensional vector. This step aims to retrieve information about the audience's perceptions of different YouTubers because YouTubers' attitudes or behavior can stand for the character of the channel. In this way, the similarity between YouTubers has been predicted and presented in a low-dimensional vector. After training YouTuber embedding, we can use this vectorial representation to replace the YouTuber variable and obtain the corresponding vector from each comment. For example, we use each comment as a key to finding which YouTuber's channel is, and the YouTuber information can continue to map with its 300-dimensional vector. The next step is to apply this embedding; the input may be comment vectors after an additional YouTuber embedding to automatically train on classification models.

3.6 Training Classifiers

We propose a BERT-based model via constructing an additional embedding layer before calculating the probability distributions over categorical labels. In the beginning, we did not change the input; we sum the position embeddings, word embeddings, and segmentation embeddings for each token. Then we add YouTuber embeddings to each sequence after extracting the hidden state vector. Finally, using a SoftMax classifier to determine over categorical labels. Figure 2 shows the modified structure of BERT model. We only used comments to detect audience's emotions and did not change the structure of pretrained BERT model in the previous task. This time, we still remain comments and incorporate YouTube domain knowledge by adding YouTuber embedding to detect emotion variance more precisely.

Besides the BERT-based model, machine learning-based models: RandomForest, Xgboost, and SVM, are also used as a classifier to deal with dimensional sentiment analysis tasks. We transform comments into numerical vectors using TF-IDF, greatly improving the more basic methods like word counts in text analysis with machine learning. TF-IDF gives us a way to associate each word in a document with a number that represents how relevant each word is in that document. In the previous task, the TF-IDF score was fed to algorithms. However, we add a 300-dimensional vector, which stands for Youtubers' information, after retrieving the TF-IDF score of each comment at this time. Simply put, each comment may find their corresponded YouTubes' channel at first. Then, each channel can be mapped with our pre-trained YouTube word embeddings.

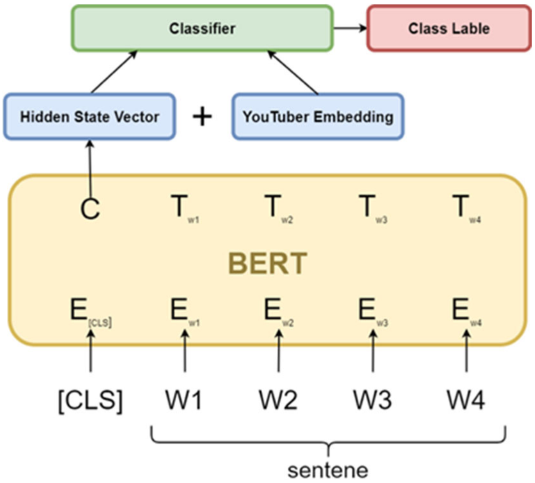


Figure 2. Structure of BERT model with YouTuber embedding.

3.7 Classification Tasks

Sentiment analysis is a fast-growing area and one of the well-known tasks of research in natural language processing (NLP) and text classifications. To better capture wide emotion variance on the audience's comment, we use three sentiment indicators and five modified models to train classifiers and analyze five targets, T1 to T5 in this task. The following elaborates the meaning of five tasks for our experiment.

- **T1**: Whether comments are related to YouTubers is a binary classification task. The data sources are generated from the result of the indicator, YouTubers preferences. By rearranging the category of the labeled datasets, we merge the annotation result of unlike, neutral, and like comments into related comments. In contrast, non-related comments remain to be. This classification task is aims to discover the motivation behind watching videos. If comments are talking about YouTubers' affairs, audience might pay attention to YouTubers.

- **T2**: Audience's sentiment towards YouTubers is an extended issue from an indicator of YouTuber preference. Exclude non-relative comments; we extract unlike, neutral, and like comments from the annotation result. Like to dislike can serve as an indicator for YouTubers to check the followers of his or her channel. Also, YouTubers can know what attractive they own or what causes them to make a nuisance.

- **T3**: Whether comments are related to videos, also be rearranged from the indicator, video preference. We duplicate the same techniques for whether comments are related to YouTubers but present a completely different meaning. This task may explain whether the contents of the video arouse discussion or become no interest to the audience. If the topic interested to the audience, it may show more relative comments towards video.

- **T4**: Audience's sentiment towards videos excludes non-relative comments from the indicator, video preference; The rest of the comments can deal with the audience's sentiment towards video. Even if watching the same channel, the different themes will captivate and engage different audiences. Therefore, this task may help YouTubers understand their audience's preferences within a specific channel.
- **T5**: Corresponding to the indicator of excitement level, T5 aims to analyze the audience's emotional ups and downs from barely excited to hyper excited, which can firmly confirm the degree of support from different audiences and affirm the audience's attitude towards specific issues.

4. Experiment

4.1 Dataset

Moving to the composition of annotated comments according to three indicators. We applied three indicators to five analysis tasks, so comments have also been rearranged into five datasets. When analyzing the target, whether comments are related to Youtubers, the proportion of the non-relative comments to the relative comments is three to one. It presented that audiences prefer talking about video content rather than YouTubers' affairs. At the same time, it comes out that the most significant piece of comments was labeled as like in audience's sentiment towards YouTubers, which is extracted from the above relative comments. This composition made sense because if people do not like someone, they may not notice their condition, even watching their channel. Next, relative comments in whether comments are related to video account for the majority in the task, and 60 percent of comments with a neutral attitude talked about the video's content. This proportion presented that the audience does not frequently present animosity on the YouTube platform within our selected channels. The fifth analyzing task, emotional ups and downs, revealed that the audience could express their health and happiness by commenting. The following table shows the proportion of data to our five tasks.

Table 2. Distribution of five tasks.

Task	Class	Number
T1	Non-Relative	8223 (75%)
	Non-Relative	2776 (25%)
T2	Unlike	287 (10%)
	Neutral	784 (28%)
	Like	1705 (61%)
T3	Non-Relative	1036 (10%)
	Relative	9775 (90%)
T4	Unlike	659 (7%)
	Neutral	5842 (60%)
	Like	3274 (33%)
T5	Barely excited	2788 (30%)
	Slightly excited	2478 (27%)
	Excited	2341 (25%)
	Fairly excited	1136 (12%)
	Hyper excited	471 (5%)

4.2 Experiment Design

This section presents multiple models in Table 4 that we experiment with. Except for BERT models that we followed it pre-trained parameters, other models have experimented with different parameters. We configure the best parameters on each model through experiments and then apply them to analyze different aspects of sentiment tasks. Also, we use 5-fold cross-validation to ensure the performance for all models. By fixedly setting k=5 to our dataset, 80% of data will be randomly selected for training and 20% for testing in each fold. In M1, M2, M3, M4, we set the number of epochs as 10 through the entire training dataset to make sure that the BERT model can have enough time to learn the pattern from social comments. After conducting experiments, we evaluate and interpret the performances of different models through the suitable metrics used for classification problems: overall accuracy. The results of social sentiment analysis are shown in the next section.

Table 3. There are four models use to solve three tasks.

Model	Description
M1	BERT model using *bert-base-multilingual-cased* pre-trained model
M2	BERT model using *distilbert-base-multilingual-cased* pre-trained model
M3	BERT model using *bert-base-multilingual-cased* pre-trained model + YouTuber embedding
M4	BERT model using *distilbert-base-multilingual-cased* pre-trained model + YouTuber embedding
M5	RandomForest
M6	Xgboost
M7	SVM
M8	RandomForest + YouTuber embedding
M9	Xgboost + YouTuber embedding
M10	SVM + YouTuber embedding

4.3 Experiment Result

Figure 3 and Figure 4 are the result of predicting the target, whether comments are related to YouTubers. The result shows that adding YouTuber embedding machine learning-based classifiers can better detect relative or non-relative comments towards YouTubers. On the contrary, after adding YouTuber embedding, the BERT model does not show better performances in the prediction result. We can also notice that M6 performed the worst in the previous task. However, it improved to become M9 and serve as the best classifier in the end.

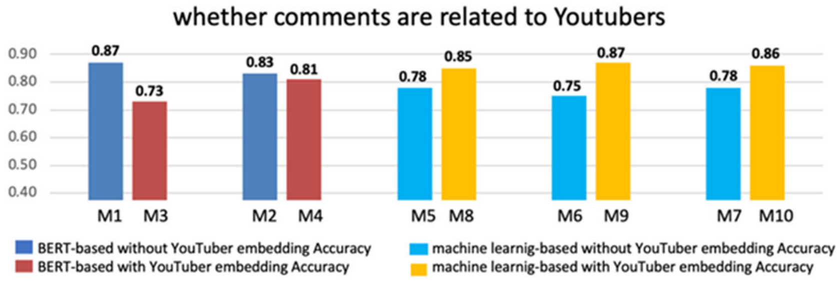


Figure 3. Models' accuracy on whether comments is related to YouTubers.

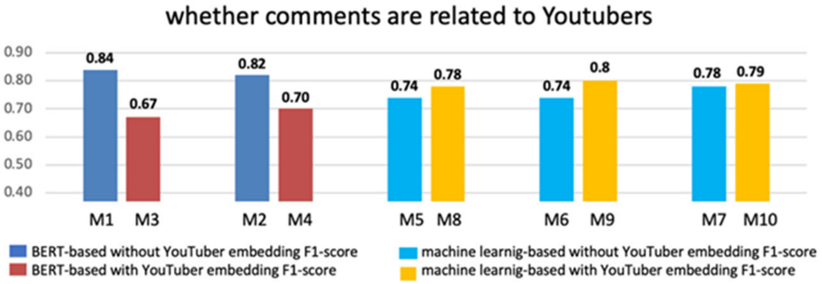


Figure 4. Models' F1-score on whether comments is related to YouTubers.

Figure 5. and Figure 6 show the result of the audience's sentiment towards YouTubers. The data in this detection task is comment about YouTubers' affairs, so we expected that adding YouTuber embedding after each comments can increase overall accuracy and F1-score. Machine learning-based classifiers proved the same result with our exception. The models 'performances have at least increased 7% in overall accuracy and 8% in F1-score. However, BERT, the variance seen from M2 to M4 surprisingly decrease.

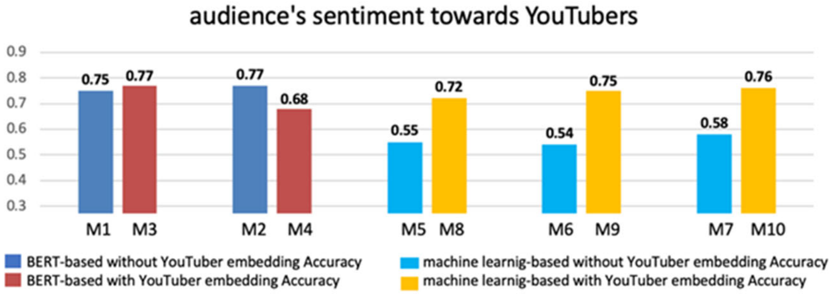


Figure 5. Models' accuracy on audience's sentiment towards YouTubers.

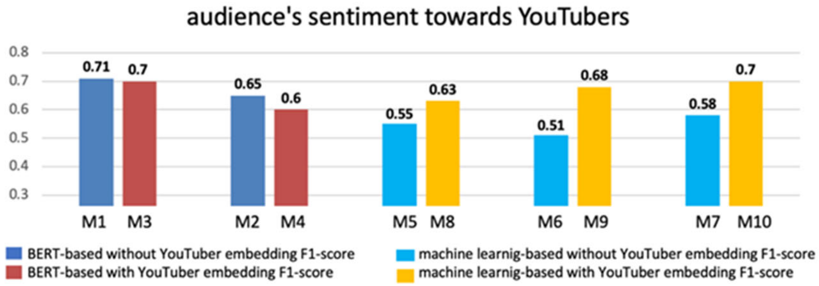


Figure 6. Models' F1-score on audience's sentiment towards YouTubers.

Figure 7 and Figure 8 are the result of predicting the target, whether comments are related to videos. We notice that overall accuracy in all models is upscale to nearly 90%. However, the improvement in F1-score is limited, only increasing smaller than 3% or even regressing in BERT method when adding YouTuber embedding. We deduce the small amount of increment or even getting worse because YouTubers' information has little relationship with determining relative or non-relative comments towards videos.

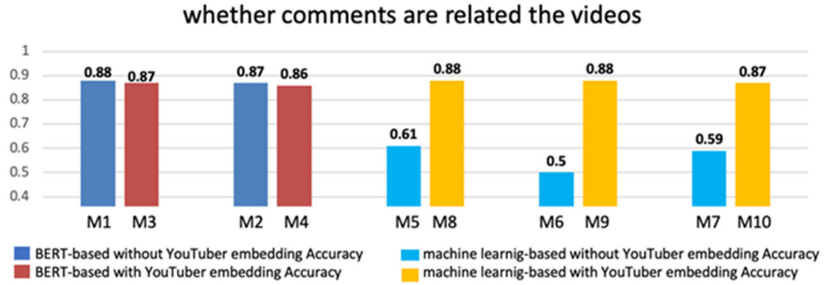


Figure 7. Models' accuracy on whether comments is related to videos

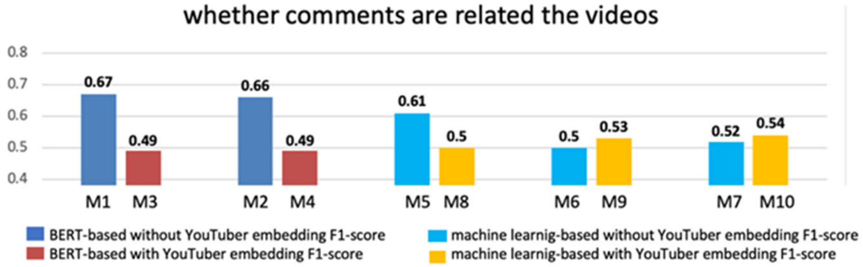


Figure 8. Models' F1-score on whether comments is related to videos

Figure 9 and Figure 10 are the result of predicting the audience's sentiment towards videos. Although data in this detection task is comments that discuss video content, the experiment result show that machine learning-based methods improved the predicted result after adding YouTuber embedding. In comparison, M4 and M4 do less well than before, decreasing from 5% to 10% and becoming the worst classifier.

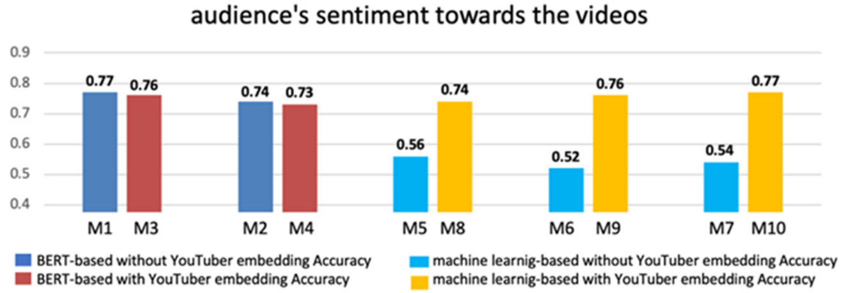


Figure 9. Models' accuracy on audience's sentiment towards videos.

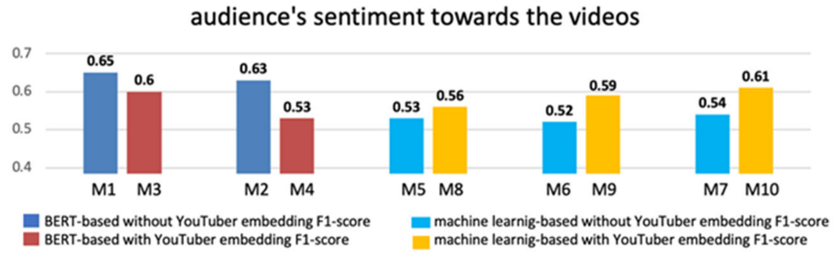


Figure 10. Models' F1-score on audience's sentiment towards videos.

Figure 11 and Figure 12 shows the result of predicting the audience's emotional ups and downs from their leaving comments. Compared with adding YouTuber embedding and without YouTuber embedding, the former method can improve model performance in machine learning-based methods. We deduce that the improvement may result from different types of YouTubers having different audiences. The more controversial YouTuber, the more excitement level may show in their audience's comments. For example, a YouTuber who prefers talking about political issues may vary their audience emotional variance than educational channels.

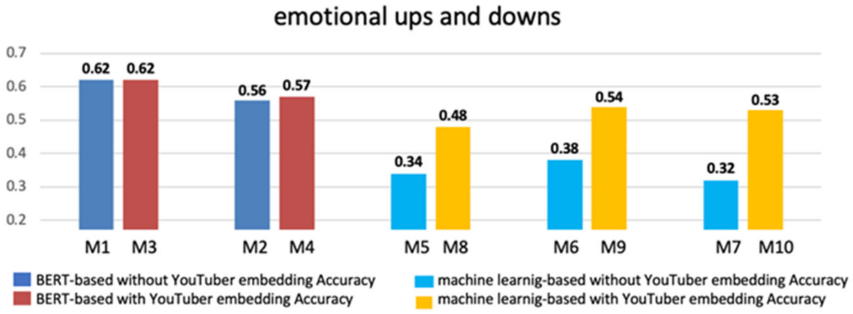


Figure 11. Models' accuracy on emotional ups and downs.

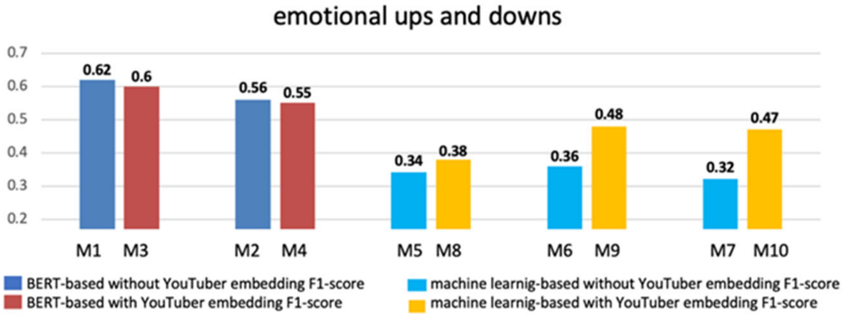


Figure 12. Models' F1-score on emotional ups and downs.

4.4 Discussion

In summary, there are three findings after we conducted experiments (1) Within machine learning-based models, the experiment results validate that adding YouTuber embedding is an effective way to identify audiences' emotions and depth of feeling. Also, we notice that YouTuber embedding is significantly helpful when detecting audience sentiment towards YouTubers. This result explains that we successfully trained YouTuber word embedding by using many comments with YouTubers' or guests' names who are invited on YouTuber's channel. (2) We notice that BERT neither improves the prediction score nor goes backward, a nearly ten percent decrease when predicting T1, T2, and T3. However, when predicting T5, two kinds of BERT (M3 and M4) do not regress their performance but remain top ranking. This result explains that BERT's model construction is more suitable for addressing multi-dimensional classification tasks. (3) Except for BERT models that performance well in determining audience's emotional ups and downs, BERT cannot perfectly deal with the polarity classificational tasks after adding YouTubers embedding. We also discover two characters that social media users own on the YouTube streaming platform. People prefer to discuss videos' content rather than YouTubes' affairs. In addition, people do not frequently present animosity in their comments; most people present their comments as neutral or barely excited attitudes.

5. Conclusion

This paper focuses on improving the over-all accuracy and F1-score on dimensional sentiment classification task. This time, we combine comments with YouTuber embeddings to train on the all classifiers. In machine learning-based classifiers, we use TF-IDF as sentence vectors and concatenate YouTuber Embedding in the last layer to fit in RandomForest, Xgboost, and SVM. On the contrary, we add YouTuber embeddings to the hidden state vector of BERT model. After that, we compare the above experiments' result with the previous tasks that only utilize comments as our data sources. Although BERT does not present a better prediction score on sentiment polarity problems, it perfectly deals with a muti-dimensional problem, the task of predicting the audience's excitement level. This result proves the superiority of BERT by achieving at least 10 % more in overall accuracy and F1-score than other classifiers. In comparison to the traditional machine learning classifiers, we identify that although machine learning models cannot perform as well as BERT before adding YouTuber embeddings, the performances of the machine learning-based classifiers can be dramatically improved after our proposed method which concatenating comments text with trained YouTubers embeddings to these classifier.

Analyzing the public's perception of YouTubers and the influence of their videos is a challenging task for researchers so far. Much work has been done in this paper, but it still has a long way to overcome some problems. In this research, we have emphasized the following problems in order to make our results improve. In the future, we could explore more information on YouTube, such as combining videos' cover photo as features, to optimize multiple-dimensional sentiment analysis tasks. In this way, even if imbalanced dataset, models may identify feature represented on the picture and capture different aspects of information that cannot present in context only. In addition, with the recent emergence of deep learning, an increasing number of researchers have started to use deep neural networks to deal with sentiment analysis, we may explore deep leering techniques to automated detect the audience's preference on social media. Last but not least, others indicators, such as whether the comments contain an ironic statement or whether the comments contain an erotic statement, can be added for analyzing other aspects of the audience's comments. The latter proposed indicator may serve as a guard for children's users, and the former indicator may prevent YouTubers from getting into conflict with their fans.

Reference

Aydoğan, M., & Karci, A. (2020). Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification. P*hysica A-statistical Mechanics and Its Applications*, 541, 123288.

Bhatt, A., Patel, A., Chheda, H., & Gawande, K. (2015). Amazon Review Classification and Sentiment Analysis. *International Journal of Computer Science and Information Technologies*, *6*(6), 5107-5110.

Cunha, A.A.L., Costa, M.C., & Pacheco, M.A.C. (2019). Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks. In *Proceedings of the 18th International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, 561-570. https://doi.org/10.1007/978-3-030-20912-4_51

González-Carvajal, S., & Garrido-Merchán, E.C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012

Hassan, A., & Mahmood, A. (2017). Deep learning approach for sentiment analysis of short texts. In *Proceedings of the 3rd international conference on control, automation and robotics (ICCAR 2017)*, 705-710. https://doi.org/10.1109/ICCAR.2017.7942788

Heredia, B., Khoshgoftaar, T. M., Prusa, J., & Crawford, M. (2016). Cross-domain sentiment analysis: An empirical investigation. In *Proceedings of the IEEE 17th International Conference on Information Reuse and Integration (IRI 2016)*, 160-165. https://doi.org/10.1109/IRI.2016.28

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer. L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Rezaeinia, S.M., Ghodsi, A., & Rahmani, R. (2017). Improving the Accuracy of Pre-trained Word Embeddings for Sentiment Analysis. arXiv preprint arXiv:1711.08609.

Sun, C., Qiu, X., Xu, Y. & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu & Y. Liu (eds.), *Chinese Computational Linguistics* (p./pp. 194-206), Cham: Springer International Publishing.

Usherwood, P., & Smit, S. (2019). Low-Shot Classification: A Comparison of Classical and Deep Transfer Machine Learning Approaches. arXiv preprint arXiv:1907.07543.

Zhu, Y., Yan, E. & Wang, F (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. *BMC Med Inform Decis Mak*, 17, 95. https://doi.org/10.1186/s12911-017-0498-1

Zhang, X., & Zheng, X. (2016). Comparison of text sentiment analysis based on machine learning. In *Proceedings of the 15th international symposium on parallel and distributed computing (ISPDC 2016)*, 230-233. https://doi.org/10.1109/ISPDC.2016.39

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 35-48

使用低通時序列語音特徵 訓練理想比率遮罩法之語音強化

Employing Low-Pass Filtered Temporal Speech Features for the Training of Ideal Ratio Mask in Speech Enhancement

陳彥同[*]、洪志偉[*]

Yan-Tong Chen and Jeih-weih Hung

摘要

在諸多基於深度學習之語音強化法中，遮罩式(masking-based)強化法求取一個遮罩與雜訊語音之時頻圖相乘、藉此使所得乘積之新時頻圖所含雜訊成分降低、以重建相對乾淨的語音訊號。在用以訓練遮罩之深度模型其輸入特徵的選取上，許多長期以來用以語音辨識的特徵、如梅爾倒倒頻譜、振幅調變時頻圖、感知線性估測係數等都是適合的選擇、可使訓練所得的遮罩達到有效的語音強化效果。另外，傳統上若將語音特徵之時序列作低通濾波處理，可以抑制雜訊所帶來的失真，因此，在本研究中，我們嘗試將各種語音特徵時序列，藉由離散小波轉換的方式加以低通濾波，再用它們來訓練語音遮罩的深度模型，探究其是否能使所學習之遮罩能對於原始雜訊語音之時頻圖有更佳的語音強化效果。在我們的初步實驗裡，在人聲雜訊環境中，我們發現上述之低通濾波所得之特徵序列、相較於原始特徵序列而言所學習而得的深度模型，能更有效地提升測試語音之品質與可讀性。

[*] 國立暨南國際大學電機工程學系
Department of Electrical Engineering, National Chi Nan University
E-mail: s109323508@mail1.ncnu.edu.tw; jwhung@ncnu.edu.tw

Abstract

The masking-based speech enhancement method pursues a multiplicative mask that applies to the spectrogram of input noise-corrupted utterance, and a deep neural network (DNN) is often used to learn the mask. In particular, the features commonly used for automatic speech recognition can serve as the input of the DNN to learn the well-behaved mask that significantly reduce the noise distortion of processed utterances. This study proposes to preprocess the input speech features for the ideal ratio mask (IRM)-based DNN by lowpass filtering in order to alleviate the noise components. In particular, we employ the discrete wavelet transform (DWT) to decompose the temporal speech feature sequence and scale down the detail coefficients, which correspond to the high-pass portion of the sequence. Preliminary experiments conducted on a subset of TIMIT corpus reveal that the proposed method can make the resulting IRM achieve higher speech quality and intelligibility for the babble noise-corrupted signals compared with the original IRM, indicating that the lowpass filtered temporal feature sequence can learn a superior IRM network for speech enhancement.

關鍵詞：語音強化、特徵時序列、低通濾波、理想比例遮罩法、小波轉換

Keywords: Speech Enhancement, Temporal Feature Sequence, Lowpass Filtering, Ideal Ratio Mask, Wavelet Transform

1. 緒論 (Introduction)

深度類神經模型與相關之學習演算法的高度發展，引發許多科技研究的空前突破與創新，過往的許多技術開發，常是基於解釋思維、在多次試錯之後找到一個可行方案，再對此可行方案賦予人們專業的解釋，然而深度學習則普遍基於統計思維、並不著重於方法在解釋上的合理性，而是嘗試將大量觀察（輸入）和對應結果（輸出）的關聯性藉由深度類神經網路加以詮釋，以期對於新的觀察能精準預測出對應的結果。

在語音處理的領域中，近年來基於深度學習所開發出的演算法也琳瑯滿目，且因訓練資料的可取得性越來越高，這些演算法在學習與預測結果的能力也隨之增強。以本研究著重的語音強化法為例，基於深度類神經模型之各式語音強化架構其表現常超越經典且富有高度理論根據的演算法，或是以後者的演算法的原型 (prototype) 出發，但配合深度類神經網路來有效學習該演算法的各項參數，使其語音強化效果更佳。

根據文獻(Wang *et al.*, 2014)，許多基於深度學習之語音強化法根據其訓練目標大致可以分為兩大範疇：對映式 (mapping) 與遮罩式 (masking)，前者直接求取一個對映函數，使此對映函數之理想輸出為乾淨語音的呈現式(特徵)，如時域訊號波形、時頻圖 (spectrogram) 或耳蝸時頻譜圖 (cochleagram)，後者是求取一個遮罩 (mask)，用以與原始輸入訊號或特徵呈現作點對點的相乘，使相乘後的訊號呈現式能趨近乾淨時的狀態。簡

單來說對映式所求取的函數，對於輸入訊號特徵的運算可以是任意由所使用之深度學習模型定義的非線性運算，而遮罩式所求取的函數運算，則簡化或限制為對輸入訊號特徵作乘法 (即加權運算)。二者各擅勝場，但近年來似乎是以遮罩式的語音強化更受重視與發展，相關的演算法包括了理想二元遮罩 (ideal binary mask, IBM) (Wang, 2005; Srinivasan *et al*., 2006)、理想比例遮罩 (ideal ratio mask, IRM) (Srinivasan *et al*., 2006)、頻譜強度遮罩 (spectral magnitude mask, SMM) (Wang *et al*., 2014)、複數理想比例遮罩 (complex ideal ratio mask, cIRM) (Williamson *et al*., 2016)、相位敏感型遮罩 (phase-sensitive mask, PSM) (Erdogan *et al*., 2015) 等。

在本研究中，主要是針對上述之遮罩式語音強化法加以改進，我們提出對於訓練遮罩模型的輸入雜訊語音的特徵時序列作簡單的預處理 (pre-processing)，使其包含的雜訊失真較低，以期在之後的訓練遮罩步驟能更加精確。而使用的預處理方法，是透過簡易的一階離散小波轉換 (discrete wavelet transform, DWT) (Mallat, 1999)]，將特徵時序列分為高低兩調變頻帶 (modulation frequency bands)，然後藉由一權重的相乘來降低高調變頻帶之序列的振幅，再將其與原始低調變頻帶序列搭配、透過一階反離散小波轉換 (inverse discrete wavelet transform, IDWT) 重建特徵序列，再使用此相當於透過低通濾波處理後的特徵序列來訓練遮罩模型。

上述低通濾波之處理，主要是基於先前諸多學者所提出的觀察(Kanedera *et al*., 1997; Chen & Bilmes, 2007)：乾淨語音特徵時序列主要分布頻率在 1 Hz 至 16 Hz 之間，以一般的音框取樣率 100 Hz 而言，特徵序列可包含的（調變）頻帶為[0,50 Hz]，因此後半頻帶鮮少包含語音成分，抑制此頻帶不會對語音造成明顯失真，但可有效抑制雜訊的干擾。

另外，基於文獻(Wang *et al*., 2018)所述，使用小波轉換分解語音特徵時序列、消除其細節係數 (detail coefficients，相當於調變高頻成分) 後重建之語音特徵，在雜訊環境下有明顯進步的語音辨識率，我們參照這樣的做法來實現前述之語音特徵序列的低通濾波處理，期許它對應的遮罩深度模型能得到更佳的語音強化效果。

2. 提出的新方法 (Proposed Method)

在本研究中，我們選擇加以研究改進的是理想比例遮罩(ideal ratio mask, IRM)法，此法通常是求取語音之一般時頻圖 (spectrogram) 或耳蝸時頻圖 (cochleagram) 對應的理想遮罩值：

$$M(m,f) = \frac{|s(m,f)|^2}{|s(m,f)|^2 + |d(m,f)|^2}, \tag{1}$$

其中，$|s(m,f)|^2$與$|d(m,f)|^2$分別代表了雜訊語音其時頻圖或耳蝸時頻圖在音框時間m與頻率f之時頻單位(time-frequency unit, T-F unit) 所對應的乾淨語音與純雜訊的能量，在人造訓練雜訊語句的準備上，由於事先可得知其乾淨語音及純雜訊的成分，因此可根據式(1)計算其理想比例遮罩的值，作為 IRM 深度模型的訓練目標。

在我們構思的新方法中，嘗試將用以訓練 IRM 深度模型所使用的語音特徵時序列，

加以低通濾波處理、藉此抑制其調變高頻的成分，再使用處理後的語音特徵來求取 IRM 深度模型，預期此 IRM 模型相對於原始特徵對應之 IRM 模型，能求取更佳的遮罩來抑制雜訊對語音時頻圖上的失真。

值得一提的是，我們使用離散小波轉換 (discrete wavelet transform, DWT) (Mallat, 1999; Wang *et al*., 2018)來執行上述的低通濾波處理，部分原因是在 DWT 其分解與重建的濾波器彼此互補，在分解與重建的過程中不會造成序列相位的失真，此相較於一般的低通濾波器而言存在優勢。

以下，我們敘述此新方法的步驟：

訓練階段：

步驟一：將訓練集 (training set) 中的任一雜訊干擾的語音$x[n]$，經音框化 (framing) 與窗化 (windowing) 切割成個別音框訊號$x_m[n]$後（m為音框索引），再將個別音框訊號轉換成語音特徵，如 amplitude modulation spectrogram (簡稱 AMS), relative spectral transformed perceptual linear prediction coefficients (簡稱 RASTA-PLP), mel-frequency cepstral coefficients (簡稱 MFCC) 及 Gammatone filterbank power spectra (簡稱 GF) 等。我們將對應的D維語音特徵向量以$\mathbf{x}_m$表示，$\mathbf{x}_m$為一$D \times 1$ 的行向量，假設該語句共切割成M個音框，則其對應的語音特徵矩陣可表示為：

$$\mathrm{X} = [\mathbf{x}_0 \quad \mathbf{x}_1 \quad \cdots \quad \mathbf{x}_{M-1}] \tag{2}$$

其尺寸為$D \times M$。

步驟二：上述之特徵矩陣X的任一第d個橫列向量

$$[\mathrm{X}_{d,0}\ \mathrm{X}_{d,1} \ldots \mathrm{X}_{d,M-1}] \tag{3}$$

以$\mathrm{X}_d[m]$代表之，其稱作X的第d維特徵時序列，尺寸為$1 \times M$，其中$1 \leq d \leq D$。

我們將任一維特徵時序列$\mathrm{X}_d[m]$以一階離散小波轉換加以分解如下：

$$[\mathrm{cA}_d[m], \mathrm{cD}_d[m]] = \mathbf{DWT}(\mathrm{X}_d[m]) \tag{4}$$

其中$\mathbf{DWT}$(.) 代表離散小波轉換 (discrete wavelet transform)、 $\mathrm{cA}_d[m]$與 $cD_d[m]$分別為轉換分解而得的近似係數(approximation coefficients)與細節係數(detail coefficients)，其可視為原始序列$\mathrm{X}_d[m]$之低通成分與高通成分，二者頻寬均約等於原始序列頻寬的一半，且點數減半。

步驟三：我們將上一步驟所得的細節係數c$D_d[m]$]乘上一個小於 1 的權重α，再與原近似係數相組合、經過反離散小波轉換重建第d維特徵時序列，表示如下：

$$\widetilde{\mathrm{X}}_d[m] = \mathbf{IDWT}([\mathrm{cA}_d[m], \alpha \times \mathrm{cD}_d[m]]) \tag{5}$$

其中$\widetilde{\mathrm{X}}_d[m]$為更新的特徵時序列，相較於原始特徵時序列$\mathrm{X}_d[m]$，$\widetilde{\mathrm{X}}_d[m]$包含較低的高通成分，因此應當包含較少雜訊造成的失真。

步驟四：參照一般 IRM 深度模型的訓練法，我們改以新的特徵序列$\{\tilde{X}_d[m], 1 \le d \le D\}$作為輸入，以理想 IRM 遮罩值為目標輸出，訓練 IRM 深度模型。值得注意的是，若式(4)中的權重$\alpha = 1$，則所訓練的 IRM 模型與原始（即使用原始特徵訓練）IRM 模型完全一致。

測試階段：

將測試之語句如同訓練語句之處理的前三個步驟、求取低通濾波之特徵時序列，將其通過訓練完成的 IRM 模型求取遮罩值，將遮罩值與原設定之對應的時頻圖作點乘積 (dot product)，即可得強化後的時頻圖，經由適當的反轉換重建成強化版的時域訊號。

3. 實驗設置 (Experimental Setup)

參照文獻(Wang *et al.*, 2014)所提供的程式碼[1]，我們使用了 TIMIT 資料庫的部分語句 (取樣頻率為 16 kHz) 來實驗評估我們所提出的方法，其中，訓練集包含了 5 位語者、每人 10 句共 50 個語句，而測試集則包含了與訓練集不同的 3 位語者、每人 10 句共 30 個語句。我們將訓練與測試語句摻入 babble 雜訊，訊雜比 (signal-to-noise ratio, SNR) 固定為 -2 dB。在訓練與測試 IRM 之深度模型上，輸入特徵的種類包含了 AMS、RASTA-PLP、MFCC 與 GF 四種，同時，我們將左右相鄰的 5 個音框 (frames) 串接成一個長向量，作為深度模型的輸入單位，深度模型之架構為全連結層(densely connected layers)網路，共包含 4 層隱藏層，每個隱藏層由 1024 個神經元 (neurons) 構成。目標是求取語音之耳蝸時頻圖 (cochleagram) 的遮罩，其每個音框設有 64 維，相當有 64 個通道(channel)。

在我們所提的新 IRM 訓練法上，對於輸入特徵之時序列之細節係數（高頻係數）所給予的權重α，分別設定為 0, 0.25, 0.50, 0.75，藉此觀察細節係數之壓抑程度對於 IRM 效果之影響（原始 IRM 所對應之權重$\alpha = 1$）。

在使用的離散小波轉換與反轉換中，我們使用 db2 小波函數。

在評估效能上，我們使用了 PESQ 分數(Rix *et al.*, 2001)作為語音品質 (quality) 的客觀指標、STOI 分數(Taal *et al.*, 2011)作為語音可讀性(intelligibility) 的客觀指標，PESQ 分數介於-0.5 與 4.5 之間， STOI 分數介於 0 與 1 之間，分數越高代表語音的品質/可讀性越佳。

4. 實驗結果與討論 (Experimental Results and Discussions)

在我們的評估實驗上，我們將分為三部分來呈現並討論，第一部分是對應於使用所有種類之輸入特徵組合所訓練及測試之 IRM 模型，第二部分是對應於使用單一種類之輸入特徵所訓練及測試之 IRM 模型，我們將在這兩部分中，探究所提新方法之低通濾波特徵時序列對於 IRM 效能的改變，第三部分則是藉由時頻圖的展示，觀察原始與更新之 IRM 所

[1] Matlab toolbox for DNN based speech separation .Retrieved from http://web.cse.ohio-state.edu/pnl/DNN_toolbox/

強化的語音訊號的差異。

4.1 使用所有種類之輸入特徵所得的IRM效能分析 (The IRM Results and Analyses for the Case using all kinds of Features)

首先，表 1 列出了測試雜訊語句在處理前、經由理想 IRM（遮罩直接由乾淨語音與摻雜之雜訊求得）及原始 IRM（使用原始輸入特徵訓練，並可能額外加入差量特徵）處理後所對應的 PESQ 與 STOI 的平均值。從此表中，我們可以看到：

1. 雜訊語句經過理想 IRM 處理後，在 PESQ 與 STOI 都得到了大幅的提升。
2. 原始 IRM 雖然也能帶來顯著的改進，但效果明顯與理想 IRM 有差距，這代表了藉由雜訊語音（特徵）中估測乾淨語音與雜訊成分之精準度仍有很大的進步空間。
3. 差量特徵的有無並未對於訓練而得 IRM 在 STOI 與 PESQ 的表現上有大幅影響，額外使用差量特徵甚至使 IRM 得到較低的 STOI 分數。

表 1.未處理語音與經過理想 IRM、原始 IRM$_1$（使用原特徵求取）、原始 IRM$_2$（使用原特徵與其差量特徵求取）處理後對應的 STOI 與 PESQ 平均分數，原特徵由四種特徵 (AMS, RASTA-PLP, MFCC, GF) 排列而得
[Table 1. The PESQ and STOI results for the baseline, oracle IRM, original IRM$_1$ (using the original combo static features) and IRM$_2$ (using the original combo static and delta features)]

	未處理語音	理想 IRM	原始 IRM$_1$	原始 IRM$_2$
STOI	0.6130	0.9004	0.6763	0.6658
PESQ	1.6081	2.6408	1.7755	1.7748

接下來，我們開始評估所提之新 IRM 訓練法，表 2 列出了在不使用差量特徵時，給定輸入特徵之時序列之高頻係數不同的權重α，經訓練之 IRM 所對應的 STOI 與 PESQ 分數，從此表中，我們有以下的發現：

1. 當使用我們提出之抑制調變高頻的特徵法時，多數α權重設定都得到了更佳的 STOI 與 PESQ 值（$\alpha = 0.25$在 STOI 分數除外，$\alpha = 0.25, 0.50$ 在 PESQ 分數除外），此初步驗證了此方法對於訓練更佳 IRM 模型、以抑制雜訊干擾有更好的效果。
2. 全然移除（設定 $\alpha = 0$）或少量移除（設定 $\alpha = 0.75$）調變高頻成分似乎是較佳選項，二者至少皆可使 PESQ 與 STOI 值提升，$\alpha = 0$得到最佳的 PESQ 值，而$\alpha = 0.75$則使 STOI 進步最大。

表 2. 未處理語音與經過理想 IRM、原始 IRM$_1$（使用原特徵求取）、不同權重α抑制調變高頻之 IRM（未搭配差量特徵）處理後對應的 STOI 與 PESQ 平均分數。原特徵由四種特徵 (AMS, RASTA-PLP, MFCC, GF) 排列而得

[Table 2. The PESQ and STOI results for the baseline, oracle IRM, original IRM$_1$ (using the original combo static features) and the lowpass-filtered IRM$_1$ (using the lowpass filtered combo static features with different assignments of parameter α)]

	原始 IRM$_1$	不同權重α抑制調變高頻之 IRM$_1$			
		0	0.25	0.50	0.75
STOI	0.6763	0.6767	0.6728	**0.6799**	0.6789
PESQ	1.7755	**1.7844**	1.7612	1.7717	1.7760

其次，表 3 列出了在額外使用差量特徵時，給定輸入特徵之時序列之高頻係數不同的權重α，經訓練之 IRM 所對應的 STOI 與 PESQ 分數，從此表中，我們有以下的發現：

1. 相較於原始 IRM 而言，使用較大權重α (0.75) 在 STOI 與 PESQ 上都有較明顯的改進，其他較小值的α設定值則並未一制性地得到明顯進步的效果，這可能原因是，當使用差量特徵時，差量特徵本身就已經抑制原始特徵的調變高頻成分，因此此時用較大的α值再對原始特徵的調變高頻成分小幅抑制，即可達到預期之進步效果。
2. 若我們將表 2 與表 3 的數據同時比較，發現達到最佳 STOI 值 (0.6799) 的是「不使用差量特徵、使用$\alpha = 0.50$ 之抑制調變高頻」的 IRM 法，而達到最佳 PESQ 值 (1.7996) 的則是「使用差量特徵、使用$\alpha = 0.75$ 之抑制調變高頻」的 IRM 法。

表 3. 未處理語音與經過理想 IRM、原始 IRM$_2$（使用原特徵與其差量特徵求取）、不同權重α抑制調變高頻之 IRM（有搭配差量特徵）處理後對應的 STOI 與 PESQ 平均分數。原特徵由四種特徵 (AMS, RASTA-PLP, MFCC, GF) 排列而得

[Table 3. The PESQ and STOI results for the baseline, oracle IRM, original IRM$_2$ (using the original combo static and delta features) and the lowpass filtered IRM$_2$ (using the lowpass filtered combo static and delta features with different assignments of parameter α)]

	原始 IRM$_2$	不同權重α抑制調變高頻之 IRM$_2$			
		0	0.25	0.50	0.75
STOI	0.6658	0.6639	0.6671	0.6615	0.6682
PESQ	1.7748	1.7819	1.7916	1.7589	1.7996

4.2 The IRM Results and Analyses for the Case using each Individual Kind of Features

在前一節中，我們已經呈現綜合四類特徵所得之 IRM 的效果，並初步驗證將特徵時序列低通濾波可以進一步強化 IRM。在本節裡，我們想進一步觀察各個類別的特徵（包含 AMS, RASTA-PLP, MFCC, GF）對於 IRM 效能之影響，同時我們也使用低通濾波來處理其序

列、進而比較濾波前與濾波後對於 IRM 效能的影響，表 4 與表 5 分別列出各種不同特徵搭配低通濾波對應之 IRM 所得之測試語句的 STOI 與 PESQ 分數，為了使整體效能優化起見，這裡我們把差量特徵一併加入，同時，我們將前一節四類特徵的組合（以"combo"表示）之結果列在表的最下一列，以供比較。從這兩個表之數據，我們有以下幾點的觀察與討論：

1. 對於語音可讀度指標 STOI 而言，不使用低通濾波之四類特徵中，以 MFCC 表現最佳（0.6740），甚至超越了組合特徵的結果（0.6658），然而，當配合低通濾波時，MFCC 可以達到更佳的 STOI 值，例如當使用$\alpha = 0.25$的權重時，MFCC 對應之 STOI 值可以進一步提升至 0.6772。此外，低通濾波處理並非對每一種特徵都能帶來改進，例如對於 AMS 特徵而言，不使用低通濾波所對應的原始 IRM 表現最好。
2. 對於語音品質指標 PESQ 而言，在不使用低通濾波之四類特徵中，MFCC 仍表現最佳（1.7966），超越了組合特徵（1.7748），而 AMS 特徵表現較不好，只有 1.6721 之 PESQ 值。然而，當配合低通濾波時，各種類特徵皆可以達到更佳的 PESQ 值，例如當使用$\alpha = 0.75$的權重時，MFCC 對應之 PESQ 值可以進一步提升至 1.7977。然而，獲得 PESQ 最佳之特徵是組合特徵配合$\alpha = 0.75$之低通濾波法，可達到 1.7996。

根據以上觀察，四類特徵的組合未必在 STOI 表現上優於單類特徵，而在 PESQ 表現上只能些許超越個別單類特徵，這可能原因在於某類特徵（如 AMS）在表現上與其他特徵差異較大，即使後端的深度模型在學習中理應能淡化這類特徵的負面影響，但是從測試結果上，多類特徵的組合並未發揮顯著的加成性。

表 4. 單一種類特徵的 STOI 分數比較，未處理語音與經過原始 IRM$_2$（使用原特徵與其差量特徵求取）、不同權重 α 抑制調變高頻之 IRM（有搭配差量特徵）處理後對應的 STOI 平均分數，其中"combo"表示四類特徵之組合

[Table 4. The averaged STOI results for the original IRM$_2$ (using the original static and delta features of single type) and the lowpass filtered IRM$_2$ (using the lowpass filtered static and delta features of single type with different assignments of parameter α)]

STOI 分數	原始 IRM$_2$	不同權重α抑制調變高頻之 IRM$_2$			
		0	0.25	0.50	0.75
AMS	**0.6472**	0.6430	0.6435	0.6458	0.6466
RASTAPLP	0.6559	0.6600	0.6607	**0.6611**	0.6556
MFCC	0.6740	0.6771	**0.6772**	0.6761	0.6770
GF	0.6695	**0.6698**	0.6667	0.6672	0.6692
combo	0.6658	0.6639	0.6671	0.6615	**0.6682**

表 5. 單一種類特徵的 PESQ 分數比較，未處理語音與經過原始 IRM_2（使用原特徵與其差量特徵求取）、不同權重 α 抑制調變高頻之 IRM（有搭配差量特徵）處理後對應的 <u>PESQ 平均分數</u>，其中"combo"表示四類特徵之組合
[Table 5. The averaged PESQ results for the original IRM_2 (using the original static and delta features of single type) and the lowpass filtered IRM_2 (using the lowpass filtered static and delta features of single type with different assignments of parameter α)]

PESQ 分數	原始 IRM_2	不同權重α抑制調變高頻之 IRM_2			
		0	0.25	0.50	0.75
AMS	1.6721	1.6705	1.6712	**1.6731**	1.6758
RASTA-PLP	1.7463	**1.7634**	**1.7634**	1.7630	1.7426
MFCC	1.7966	1.7870	1.7916	1.7946	**1.7977**
GF	1.7641	**1.7791**	1.7669	1.7635	1.7633
combo	1.7748	1.7819	1.7916	1.7589	**1.7996**

4.3 增加訓練及測試資料且使用單一種類之輸入特徵所得的IRM效能分析 (The IRM Results and Analyses for the Case using a Single Feature with More Training and Test Data)

在前一節中，我們可觀察出在各個類別的特徵中，單獨使用 MFCC 特徵的 IRM 效能明顯優於其他特徵，其同時使用低通濾波與差量特徵處理其序列可得到較佳的 STOI ($\alpha = 0.25$) 與 PESQ ($\alpha = 0.75$) 分數。在本節裡，我們想進一步觀察此表現良好的的 MFCC 特徵，若再增加 1 倍的資料數量 (其中，訓練集包含了 10 位語者、每人 10 句共 100 個語句，而測試集則包含了與訓練集不同的 6 位語者、每人 10 句共 60 個語句) 的情況下，其 IRM 的效能，同時觀察在使用我們所提出的低通濾波法對於 MFCC 特徵在此狀態下之 IRM 效能的影響，這一系列實驗結果分別列在表 6（無差量特徵）與表 7（有差量特徵）。

從表 6 與表 7 我們可以觀察出以下幾點：

1. 把表 6、7 與表 4、5 的數據相比較，我們可以看到增加訓練資料量可以同時使測試資料的 PESQ 與 STOI 的分數都明顯進步，進而驗證訓練資料的增加可以使 IRM 模型在語音強化的效果更好。
2. **當沒有使用差量特徵時**，若增加訓練語料，在 STOI 分數上，原始的 IRM 比使用低通濾波法對應的 IRM 效果較佳，代表此時低通率波處理並未帶來 STOI 分數的進步，然而在 PESQ 分數上，當配合低通濾波時，可以比原始 IRM 達到更佳的結果，例如當使用$\alpha = 0.75$的權重時，MFCC 對應之 PESQ 值可以進一步提升至 1.8192。然而，獲得 PESQ 最佳權重是$\alpha = 0$之低通濾波法，可達到 1.8214。
3. **當使用差量特徵時**，上一點的結果則剛好對調：即若增加訓練語料，在 PESQ 分數上，

原始的 IRM 比使用低通濾波法對應的 IRM 效果較佳，而在 STOI 分數上，當配合低通濾波時，可以比原始 IRM 達到更佳的結果，例如當使用$\alpha = 0.5$的權重時，MFCC 對應之 STOI 值可以進一步提升至 0.6880。然而，獲得 PESQ 最佳權重是$\alpha = 0$之低通濾波法，可達到 1.8214。

4. 當比較表 6 與表 7 的數據，我們可以清楚看到，額外使用差量特徵反而同時使 PESQ 與 STOI 的分數都降低，這結果似乎表明，在訓練資料增加時，差量特徵的參與並未對於 IRM 模型之訓練有正面的影響，這背後原因可能是此時 IRM 模型之複雜度應該進一步提高、以因應額外的差量特徵帶來的資料多樣性。如果在原始 IRM 模型架構的設定下，不使用差量特徵可能是較佳的選擇，同時配合低通濾波處理，可使 PESQ 分數進一步提升。

表 6. 未處理語音與經過原始 IRM$_1$（使用原 MFCC 特徵求取）、不同權重α抑制調變高頻之 IRM$_1$ 處理後對應的 STOI 與 PESQ 平均分數。原特徵由單一特徵 MFCC 而得
[Table 6. The averaged PESQ and STOI results for the original IRM$_1$ (using the original static MFCC features) and the lowpass filtered IRM$_1$ (using the lowpass filtered static MFCC features with different assignments of parameter α)]

MFCC 特徵	原始 IRM$_1$	不同權重α抑制調變高頻之 IRM$_1$			
		0	0.25	0.50	0.75
STOI	**0.6947**	0.6900	0.6926	0.6918	0.6928
PESQ	1.8182	**1.8214**	1.7996	1.8056	**1.8192**

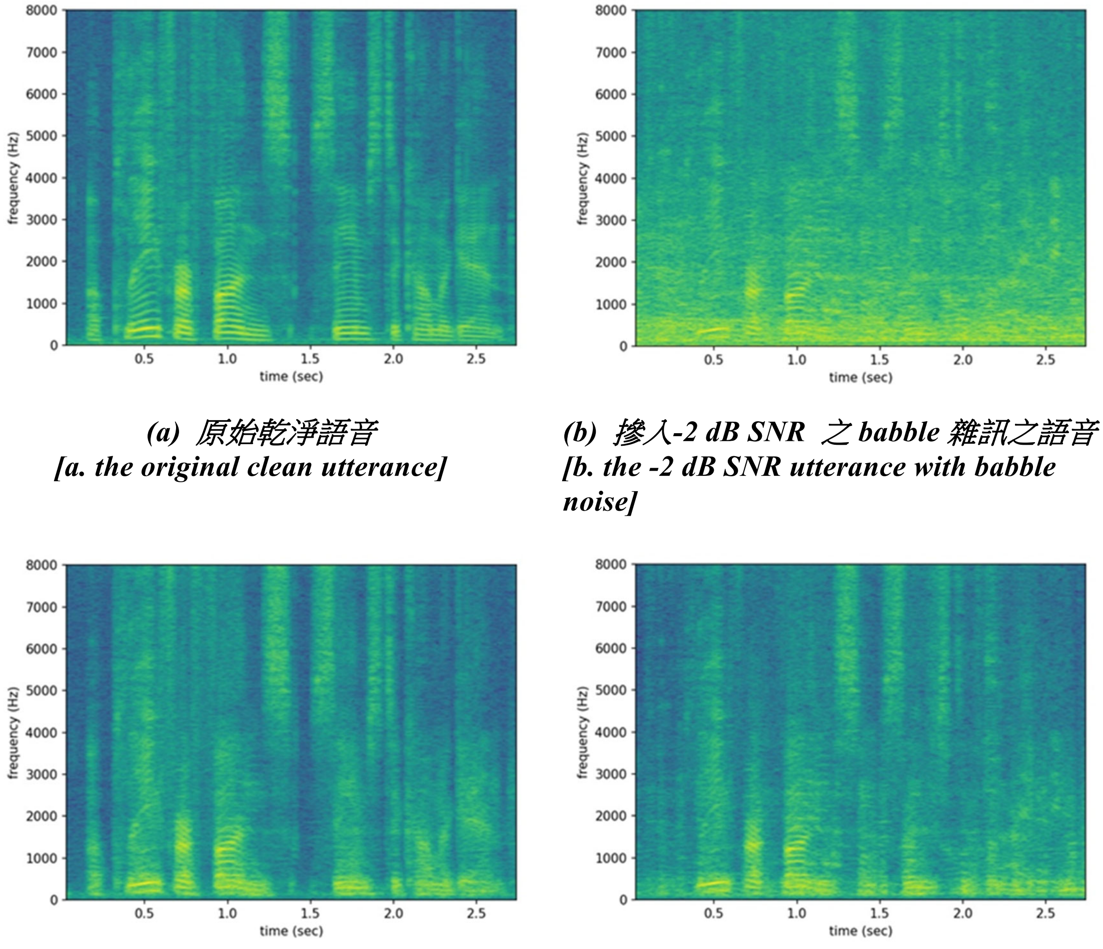
表 7. 未處理語音與經過原始 IRM$_2$（使用原 MFCC 特徵與其差量特徵求取）、不同權重α抑制調變高頻之 IRM$_2$（有搭配差量特徵）處理後對應的 STOI 與 PESQ 平均分數。原特徵由單一特徵 MFCC 而得
[Table 7. The averaged PESQ and STOI results for the original IRM$_2$ (using the original static and delta MFCC features) and the lowpass filtered IRM$_2$ (using the lowpass filtered static and delta MFCC features with different assignments of parameter α)]

MFCC 特徵	原始 IRM$_2$	不同權重α抑制調變高頻之 IRM$_2$			
		0	0.25	0.50	0.75
STOI	0.6863	0.6841	0.6840	**0.6880**	0.6837
PESQ	**1.8003**	1.7966	1.7966	1.7853	1.7972

4.4 使用時頻圖演示結果 (Spectrogram Demonstration for Each Method)

最後在這一小節，我們使用語音訊號的強度時頻圖(magnitude spectrogram)，來檢視原始 IRM 與我們提出之低通濾波特徵之 IRM 的強化效能，圖 1(a)-(f) 為一語句在各種狀態下所對應的強度時頻圖，首先，我們比較圖 1(a)與圖 1(b)，發現雜訊對於語音在時頻圖上

產生顯著的失真，接著，比較圖 1(b)與圖 1(c)可看出，理想的 IRM 可帶來顯著的語音強化效果，最後，觀察原始 IRM 與低通濾波特徵之 IRM 所對應的圖 1(d) 與 圖 2(e) ，相對於圖 1(b)，雜訊所造成的失真明顯降低，但效果並不如理想 IRM 所對應的圖 1(c)，例如在時間 0.1-0.3 秒之間的頻譜強度並未有效重建（在紅色框所標示區域），然而圖 1(e)的在此區域的頻譜重建程度稍優於圖 1(d)，根據此比較結果，我們似乎可看出，低通濾波特徵之 IRM 在此語句的處理上略優於原始 IRM。

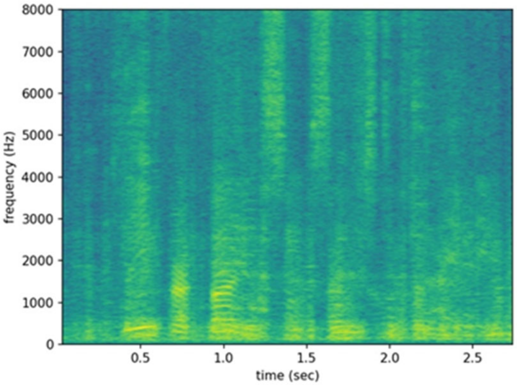


(a) 原始乾淨語音
[a. the original clean utterance]

(b) 摻入-2 dB SNR 之 babble 雜訊之語音
[b. the -2 dB SNR utterance with babble noise]

(c) 雜訊語音經由理想 IRM 處理之語音
[c. the oracle-IRM enhanced utterance]

(d) 雜訊語音經由原始 IRM 處理之語音
[d. the original-IRM enhanced utterance]



(e) 雜訊語音經由低通濾波 IRM 處理後之語音
[e. the lowpass-filtered IRM-enhanced utterance]

圖 1. 各種狀態下之語音強度時頻圖
[Figure 1. The magnitude spectrograms of an utterance at different conditions]

5. 結論與未來展望 (Conclusion and future works)

在本研究中，我們提出並初步驗證了當理想比例遮罩(IRM) 之深度模型使用低通濾波之語音特徵時序列來訓練時，相較於使用原特徵時序列訓練，可以得到更佳的語音強化效果。我們使用小波轉換來實現低通濾波的處理，其執行簡易但效果明顯，在未來工作上，我們初步規劃將此低通濾波的時序列處理用在訓練其他種類的語音強化深度模型之特徵上，檢視其是否也能更有效改進該模型的效能、提升語音之品質與可讀性。

致謝 (Acknowledgement)

本論文其部分初階實驗由本校畢業同學林子強 (Mr. Zi-Qiang Lin) 加以執行，特此致謝。

參考文獻 (References)

Chen, C., & Bilmes, J. (2007). MVA processing of speech features. *IEEE Trans. on Audio, Speech, and Language Processing*, *15*(1), 257-270. https://doi.org/10.1109/TASL.2006.876717

Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 780-712. https://doi.org/10.1109/ICASSP.2015.7178061

Kanedera, N., Arai, T., Hermansky, H., &. Pavel, M. (1997). On the importance of various modulation frequencies for speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, *3*, 1079-1082.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. (2nd ed.). Academic Press.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of of 26th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, 749-752. https://doi.org/10.1109/ICASSP.2001.941023

Srinivasan, S., Roman, N., & Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communications*, *48*(11), 1486-1501. https://doi.org/10.1016/j.specom.2006.09.003

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125-2136. https://doi.org/10.1109/TASL.2011.2114881

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi P. (eds) *Speech Separation by Humans and Machines*, (pp. 181-197). Springer. https://doi.org/10.1007/0-387-22794-6_12

Wang, S.-S., Lin, P., Tsao, Y., Hung, J.-W., & Su, B. (2018). Suppression by selecting wavelets for feature compression in distributed speech recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, *26*(3), 564-579. https://doi.org/10.1109/TASLP.2017.2779787

Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1849-1858. https://doi.org/10.1109/TASLP.2014.2352935

Williamson, D.S., Wang, Y., & Wang, D. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(3), 483-492. https://doi.org/10.1109/TASLP.2015.2512042

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 49-66

整合語者嵌入向量與後置濾波器於提升個人化合成語音之語者相似度

Incorporating Speaker Embedding and Post-Filter Network for Improving Speaker Similarity of Personalized Speech Synthesis System

王聖堯*、黃奕欽*

Sheng-Yao Wang, Yi-Chin Huang

摘要

近年來在語音合成的研究之中，單一語者的合成系統已經有著高品質的表現，但對於多語者系統來說，合成語音的品質與語者相似度仍是一大挑戰，本研究針對合成語音的品質與語者相似度兩個議題來建立出一套可合成多語者之文字轉語音系統，首先針對多語者的議題中，目標為透過少量樣本(Zero-Shot)來達成語者轉換，我們透過語者嵌入向量(Speaker Embedding)的引入來實作多語者語音合成系統，並比較針對不同任務所建立的語者嵌入向量的效果差異。在此我們比較了用於語者辨識(Speaker Verification)以及單純用於語音轉換(Voice Conversion)的語者嵌入向量。接著，為了提升合成的語者相似度以及語音品質，我們嘗試置換類神經網路架構中，作為提升頻譜的 Post-Net 的部分，在此處我們使用了一個後置濾波器(Post-Filter)的網路來取代，且比較和 Post-Net 所產生的頻譜差異以及探討其模型參數量之差異性。實驗結果表明，透過疊加性注意力機制來整合語者嵌入向量進入到類神經網路架構的語音合成系統的確能夠有效地產生具有目標語者的合成語音，並且在加入後置濾波器網路後能夠比傳統透過 Post-Net 的方式來強化合成語音的語者特性以及語音品質，且合成一般長度語音句的時間約為 2 秒鐘，已接近即時合成個人化語音之成

* 國立屏東大學資訊科學研究所
Department of Computer Science, National Pingtung University
E-mail: mike456852@gmail.com; ychuangnptu@mail.nptu.edu.tw

果。未來的研究方向會加入更多資訊來幫助語者嵌入向量在 TTS 的效能上改進。

Abstract

In recent years, speech synthesis system can generate speech with high speech quality. However, multi-speaker text-to-speech (TTS) system still require large amount of speech data for each target speaker. In this study, we would like to construct a multi-speaker TTS system by incorporating two sub modules into artificial neural network-based speech synthesis system to alleviate this problem. First module is to add the speaker embedding into encoding module of the end-to-end TTS framework while using small amount of the speech data of the training speakers. For speaker embedding method, in our study, two speaker embedding methods, namely speaker verification embedding and voice conversion embedding, are compared for deciding which one is suitable for the personalized TTS system. Besides, we substituted the conventional post-net module, which is conventionally adopted to enhance the output spectrum sequence, to a post-filter network, which is further improving the speech quality of the generated speech utterance. Finally, experiment results showed that the speaker embedding is useful by adding it into encoding module and the resultant speech utterance indeed perceived as the target speaker. Also, the post-filter network not only improving the speech quality and also enhancing the speaker similarity of the generated speech utterances. The constructed TTS system can generate a speech utterance of the target speaker in fewer than 2 seconds. In the future, other feature such as prosody information will be incorporated to help the TTS framework to improve the performance.

關鍵詞：多語者語音合成、語音轉換、語者識別、少量樣本、後置濾波器
Keywords: Multi-speaker Text-to-Speech, Voice Conversion, Speaker Verification, Zero-Shot, Post-Filter

1. 緒論 (Introduction)

就單一語者的語音合成技術來看，其合成技術已經能夠合成出逼真且自然的語音，並且不需要太多的語音數據及訓練時間，而為了擴展到其他語者，常見的方法有語音轉換和模型自適應兩種方法：

- 語音轉換： 透過更換不同語者訊息來達成目標，有基於 GAN 的 StarGAN-VC (Kameoka *et al.,* 2018) 和 CyCleGAN-VC (Kaneko *et al.,* 2018) 等方法，也有基於 AutoEncoder 的 AdaIN-VC (Chou *et al.,* 2019) 和 AutoVC (Qian *et al.,* 2019) 等方法，它們都有相當不錯的效果，唯一的侷限就是僅能更換語者不能更改內容。

- 模型自適應： 主要是在 TTS 系統中加入 Speaker ID Table 來使模型能夠依照 Speaker ID 生成對應語者的聲音，它既能更換內容也能更換語者，但是需要大量不同語者的語音數據以及較多的訓練時間來達成目標，且無法擴展到沒看過的語者。

基於語音轉換和模型自適應在多語者 TTS 上的不足，於是有著 (Jia *et al.*, 2018) 和 (Chien *et al.*, 2021) 等研究，將語音轉換或語者辨識這兩種方法取代模型自適應中的 Speaker ID Table 來使模型擴展到沒看過的語者。在本次研究中，我們將比較分別使用語音轉換和語者辨識這兩種任務所設計的語者嵌入向量作為我們 TTS 系統中語者的表示方式，並比較何者對於我們提出的架構更合適。

我們的 TTS 架構是基於 Google 所提出的自回歸模型 Tacotron 2 (Shen *et al.*, 2018)，它由三個神經網路區塊組成，每個區塊都有明確的目的以便我們進行改動：

- 編碼器: 將輸入的文字編碼成一種潛在表示，通常為了使模型擴展到多語者，會將文字潛在表示與語者嵌入向量串接。
- 解碼器: 於訓練期間，將文字潛在表示與目標頻譜的每個音框建立注意力對齊 (Chorowski *et al.*, 2015)，於推論期間，依據當前音框與文字潛在表示推測出下一個音框的值，直至注意力機制對齊到停止符號 (例如：文字中的句點)為止。
- Post-Net: 提升整體頻譜的品質。

Tacotron 2 的模型架構如圖 1 所示：

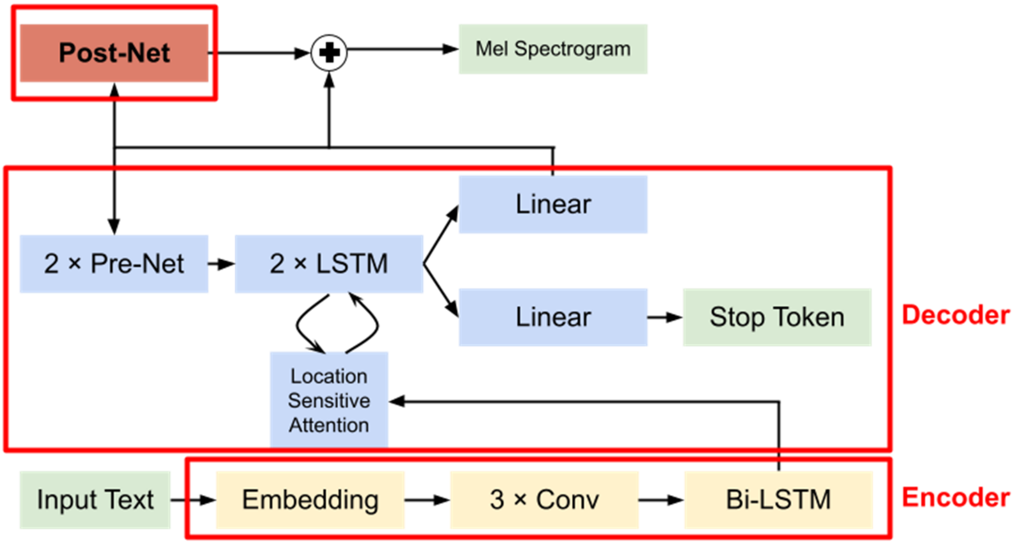


圖 1. Tacotron 2 模型架構
[Figure 1. Tacotron 2 model architecture]

Tacotron 2 整體架構對於目前神經網路的技術來說是相對舊的，隨著 Self-Attention (Vaswani *et al.*, 2017) 大量被運用於語音合成的任務上，改善了如 Tacotron 2 因使用 RNN 神經網路需要依照順序傳播的大量計算，如 Transformer TTS (Li *et al.*, 2019) 和 Fastspeech 2 (Ren *et al.*, 2020)；也有著各種注意力機制的方法被提出，以改善 Tacotron 2 舊有注意力機制訓練速度慢或是較長的句子會發生漏字或重複發音的問題，如 Forward Attention (Zhang *et al.*, 2018) 及 Dynamic Convolution Attention (Battenberg *et al.*, 2020)。因此，我們將運用近期的神經網路技術來更動 Tacotron 2 模型，期望模型訓練速度加快、合成語音品質的提升以及加強合成多語者語音的語者相似度。

我們將在第二章節闡述語者嵌入向量所用到的語音轉換及語者辨識模型，在第三章節闡述本次研究改動 Tacotron 2 的方法，第四章節闡述實驗結果，最後，在第五章節闡述本次研究的結論。

2. 語者嵌入向量 (Speaker Embedding)

2.1 語音轉換 (Voice Conversion)任務

在本次研究中，我們使用 AdaIN-VC 作為本次研究的語音轉換模型，雖然如之前所述，語音轉換的模型有很多種，但並不是都能提取出語者嵌入向量，如 StarGan 及 CycleGAN 等 GAN 模型雖然也是語音轉換，但它們是透過在訓練期間判別器 (Discriminator) 的約束，使得生成的語音接近內部語者，這無法提取出語者嵌入向量；屬於 AutoEncoder 模型的 AutoVC 也無法提取語者嵌入向量，因它是利用編碼層將語者訊息去除，並在解碼層加入 Speaker ID Table 來進行轉換的，而 AdaIN-VC (Adaptive Instance Normalization-Voice Conversion) 是一種將圖片風格轉換的技術套用到語音轉換上的 VAE 模型，它透過兩個編碼層將語音編碼成語者潛在表示及內容潛在表示，並透過解碼層組合兩者後生成轉換後的語音，我們可以藉由更換語者潛在表示來達到語音轉換的效果，其模型架構如圖 2 所示：

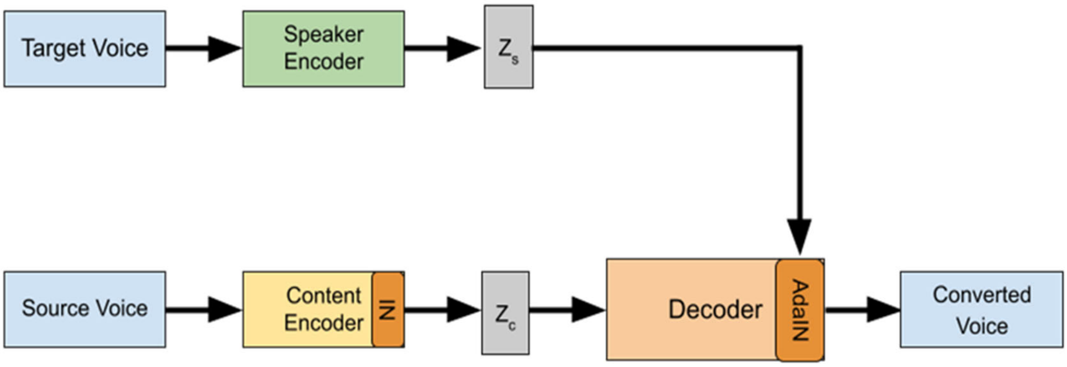


圖 2. AdaIN-VC 模型架構
[Figure 2. AdaIN-VC model architecture]

2.2 語者辨識 (Voice Verification)任務

我們使用 Learnable Dictionary Encoding (Cooper *et al.*, 2020) 簡稱 LDE，作為本次研究的語者辨識模型，它是基於 X-Vector (Snyder *et al.*, 2018) 所做的改進，並且在語者辨識的任務上以及多語者 TTS 系統上皆是優於 X-Vector 的。

X-Vector 的運作方式是將整個語音分成數個片段並透過數層卷積計算其輸出特徵，再將所有特徵取平均與標準差通過線性轉換來計算該語者的嵌入向量。LDE 與 X-Vector 不同的地方是 LDE 引入了數個 Dictionary Clusters，這些 Clusters 是需要透過神經網路去學習的，它們代表某些說話人的特徵，LDE 使 X-Vector 得到的輸出特徵與所有 Clusters 計算彼此差距的平均值與標準差來判斷該語音接近哪一個 Clusters，然後再進一步讓神經網路判斷該語者的嵌入向量，其模型架構如圖 3 所示：

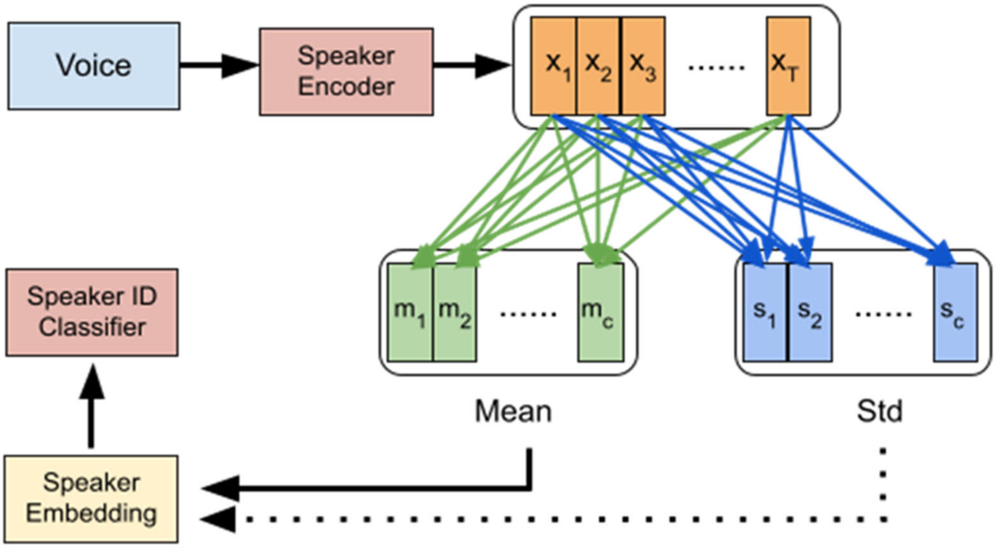


圖 3. LDE 模型架構
[Figure 3. LDE model architecture]

3. 研究方法 (Research Method)

3.1 編碼器 (Encoder)

基於原本的 Tacotron 2 架構，我們將 LSTM 的輸出降維降至 128 維並通過 Self-Attention 當作另一個輸出，Self-Attention 會將 LSTM 輸出的潛在表示進行全域相關性的連接，這些資訊將在解碼層幫助注意力機制更快的對齊，我們將原 LSTM 輸出稱為內容資訊 (Content Information)，另一個通過 Self-Attention 的輸出稱為長距離內容資訊 (Long-distance Content Information)，同時，為了使模型能夠合成多語者的語音，我們在這兩個潛在表示後方串接了語者嵌入向量，詳細架構如圖 4 所示：

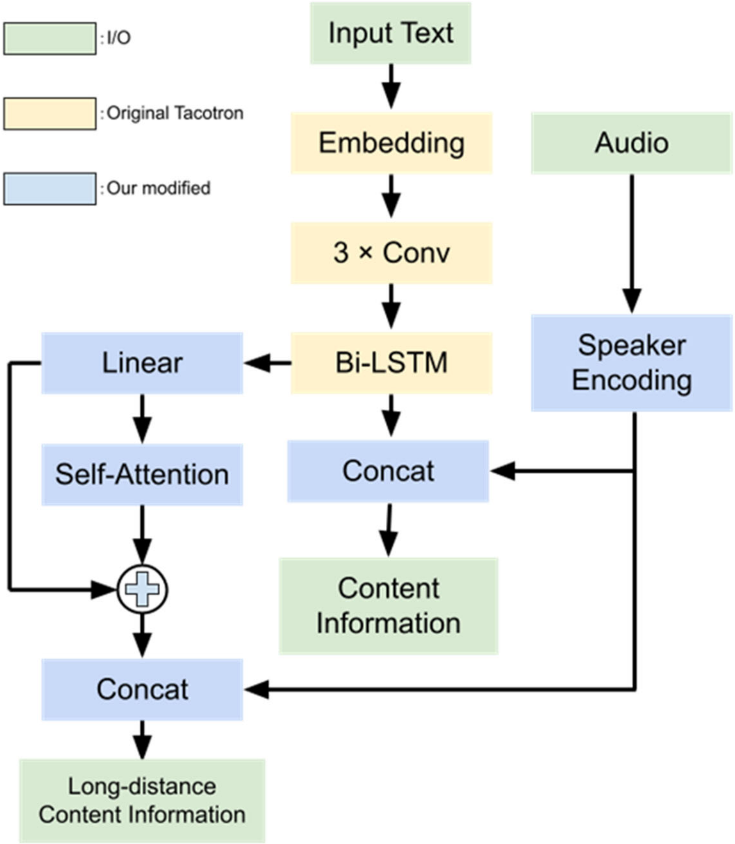


圖 4. 編碼器架構
[Figure 4. Encoder architecture]

3.2 解碼器 (Decoder)

解碼器部份我們做了較多的改動，首先，由於編碼層有兩個輸出，因此我們分別引入了兩個不同的注意力機制，我們為內容資訊引入了 Forward Attention 取代 Tacotron 2 舊有的注意力機制，它可以更快地引發對齊，並且能改善因長句所引發的重複發音或漏字的問題；長距離內容資訊則引入了 Bahdanau Attention (Bahdanau *et al.*, 2014)，它是一個傳統的 Additive Attention，因其架構較為簡單，可以快速地得到某些頻譜與文字的關係，這將能夠幫助 Forward Attention 更快地引發對齊，並且因為低維度的關係，它不會與 Forward Attention 競爭文字與頻譜間的對齊，在實驗結果會有更詳細地說明。

此外，為了加強語者嵌入向量對於模型的作用，我們在 Pre-Net 層加入了語者嵌入向量，透過神經網路的學習，能夠使模型更看重語者嵌入向量。

最後，在通過 LSTM 解碼後，我們又再一次引入 Self-Attention 將頻譜潛在表示的資訊進行全域相關性的連接，以幫助後續線性轉換更快地優化，其詳細架構如圖 5 所示：

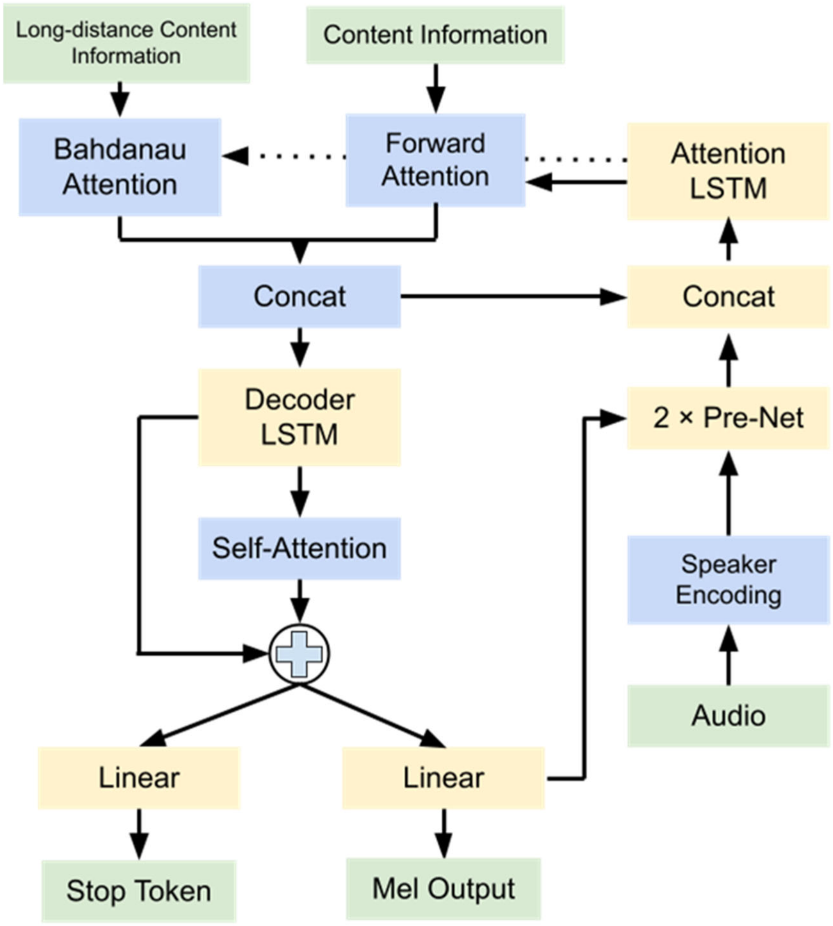


圖 5. 解碼器架構
[Figure 5. Decoder architecture]

3.3 Post-Net

原本 Post-Net 目的是為了改善頻譜重構的品質，在 Tacotron 2 的論文裡提到，有 Post-Net 的 MOS 評分是比較高的。

在本次研究中，我們額外引入了另一個架構 Diffwav (Kong, Z. *et al.*, 2020) 作為 Post-Filter 來與 Post-Net 比較。Diffwave 是 Nvidia 於 2020 年推出的 Vocoder，能夠將頻譜轉換成波形訊號，它的基礎理論是 Denoising Diffusion Probabilistic Model (Ho *et al.*, 2020)，簡稱 DDPM。DDPM 是一個馬可夫鍊 (Markov Chain) 模型，透過指定步數為目標添加高斯噪音直至目標變成高斯亂數，再透過朗之萬動力學 (Langevin Dynamics) 反向還原至目標。

我們利用上述的原理，將 Diffwave 修改成頻譜間的轉換的 Post-Filter，期望透過添加噪音能使生成的頻譜有著更多的細節，其運作流程如圖 6：

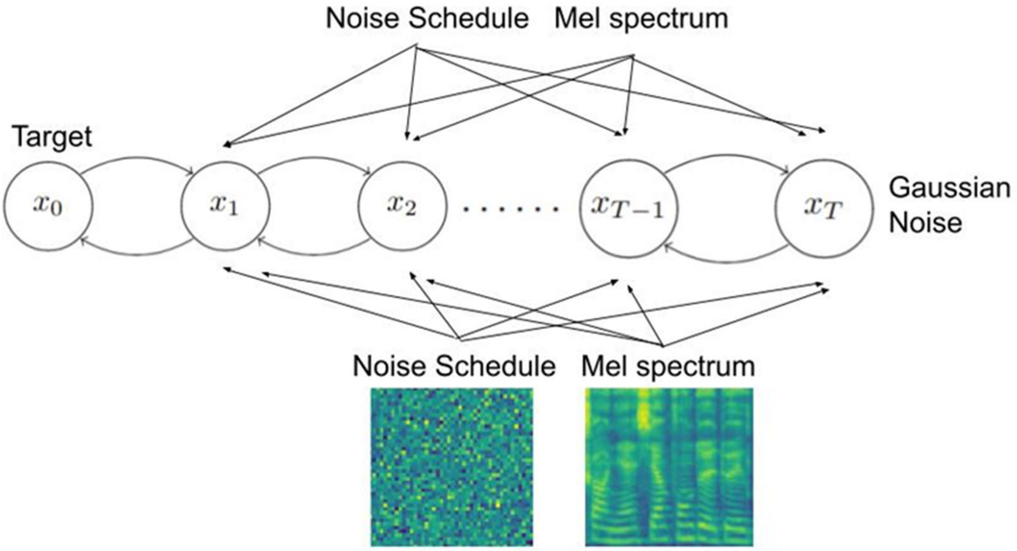


圖 6. Diffwave 流程
[Figure 6. Diffwave process]

如圖 6 所示，Diffwave 透過模型反覆運作並以噪音表 (Noise Schedule，強度由小到大的噪音) 與梅爾頻譜作為輸入條件使模型在訓練期間學習到如何透過輸入條件來添加噪音分佈破壞輸入目標；由於模型已經學得如何依照輸入條件添加噪音分佈，在推論期間，運用反函式的作法，將添加的噪音分佈除去，使輸入的高斯噪音逐漸還原成目標，其模型架構如圖 7 所示：

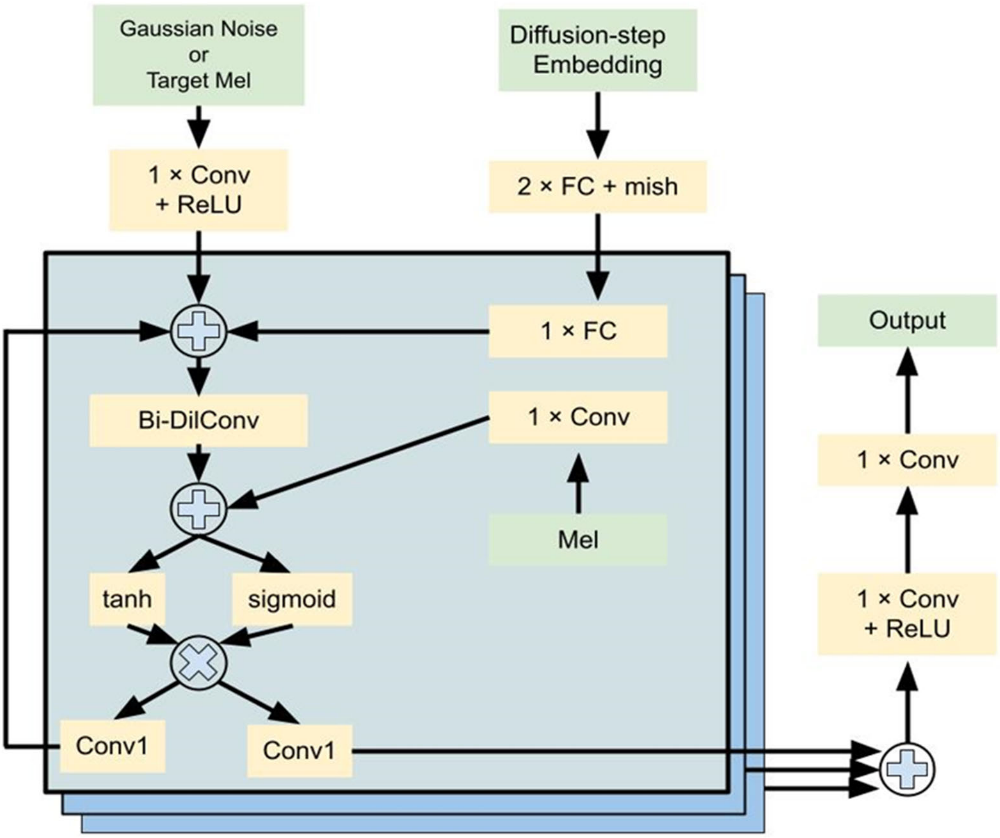


圖 7. Diffwave 模型架構
[Figure 7. Diffwave model architecture]

4. 實驗 (Experiments)

4.1 資料集 (Dataset)

我們使用 AISHELL-3 高保真中文語音數據庫作為本次實驗的資料集，共有 88035 個音檔，218 位語者，採樣率為 44.1kHz，16bit。我們將所有音檔下採樣至 22050Hz，並從中提取出 173 位語者(約佔整體語者 80%)，每位語者隨機取 100 句音檔作為訓練集，共 17300 個音檔，其餘 45 為語者當成未看過語者測驗模型合成外部語者的性能。

4.2 實驗設置 (Experimental Setups)

首先，我們使用 HiFiGAN (Kong, J. *et al.*, 2020) 作為本次實驗的 Vocoder，沒有重新訓練也沒有進行參數的微調，僅使用原作者實現的 Github 中所提供的預訓練模型。接著，我們利用資料集的音檔分別對於語音轉換的 AdaIN-VC 和語者辨識的 LDE 模型訓練，使其

生成 128 維度的語者嵌入向量。在我們提出改動的 Tacotron 2 模型架構之中，編碼層的輸出 Content Information 輸出維度仍維持 512 維，串接上語者嵌入向量後為 640 維；Long-distance Content Information 輸出維度為 128 維，串接上語者嵌入向量後為 256 維 。 在解碼層中，我們把語者嵌入向量升維至 256 維並以 Softsign 激活函數激活，於 Pre-Net 層中與頻譜相加，其餘設置皆按照原 Tacotron 2。

我們提出的 TTS 模型是在 Pytorch 神經網路框架上運行，並以 Nvidia GeForce RTX 2070 GPU 訓練，批量大小 (Batch Size) 設為 8，共訓練 208,000 個 Steps，約為 96 個 Epochs。

4.3 實驗結果 (Results)

4.3.1 語音品質 (Speech quality)

首先，我們使用客觀評測 (MOS) 來證實實驗結果，分別合成語音轉換和語者辨識所訓練的 TTS 系統各 10 個內部語者的音檔來比較品質，另外再合成各 10 個內部語者的音檔比較語者相似度，其結果如表 1：

表 1. 語音轉換和語者辨識的 MOS
[Table 1. MOS for Voice Conversion and Speaker Verification]

	Quality	***Similarity***
Tacotron 2 with VC	***2.67 ± 0.35***	***2.70 ± 0.41***
Tacotron 2 with SV	***2.54 ± 0.37***	***2.31 ± 0.18***

根據表 1 可以發現語音轉換提取出來的語者嵌入向量對於我們的 TTS 系統效果較好，因此進一步使用語音轉換的語者嵌入向量來比較 Post-Filter 與原始 Post-Net 的效果，結果如下表：

表 2. 比較 Post-Filter 與 Post-Net 的效果
[Table 2. Compare the effects of Post-Filter and Post-Net]

	Quality	***Similarity***
Post-Filter	***3.75 ± 0.35***	***3.75 ± 0.71***
Post-Net	***2.67 ± 0.71***	***2.50 ± 0.30***

接著我們使用 Mel Cepstral Distortions (MCD) 作為客觀評測的方法，隨機從內部語者與外部語者各挑選 5 個男性與女性語者，每個語者合成 10 個音檔來計算 MCD 值，結果如下表：

表 3. 計算 Post-Filter 與 Post-Net 的 MCD，值越小越好。
[Table 3. Calculate the MCD of Post-Filter and Post-Net, the smaller the value, the better.]

	Inside		***Outside***	
	Men	***Women***	***Men***	***Women***
Post-Filter	***6.99***	*7.30*	***8.15***	***8.65***
Post-Net	*7.31*	*7.98*	*9.20*	*9.11*

我們還使用 Resemblyzer 分析器計算不同性別在語音轉換上的語者空間，Resemblyzer 是一個透過神經網路來比較或分析語音的 Python 套件。研究中， 男性與女性每位語者皆合成 10 句 Post-Filter 和 Post-Net 個音檔與原語者比較，其結果如圖 8 和圖 9，我們可以從這兩張圖中發現，在內部語者中，合成的女性音檔都很接近原音檔，在男性音檔中則可以發現 Diffwave 較 Post-Net 接近原始音檔，不管是語音合成或語者辨識效果皆相似；在外部語者中，可以發現語音合成的語者空間較為集中，而語者辨識的語者空間較為發散，並且 Diffwave 比 Post-Net 稍微接近原始音檔。我們可以斷定語音合成任務以及針對 Post-Net 所提出的 Diffwave 架構對於我們的多語者 TTS 系統來說是更有幫助的。

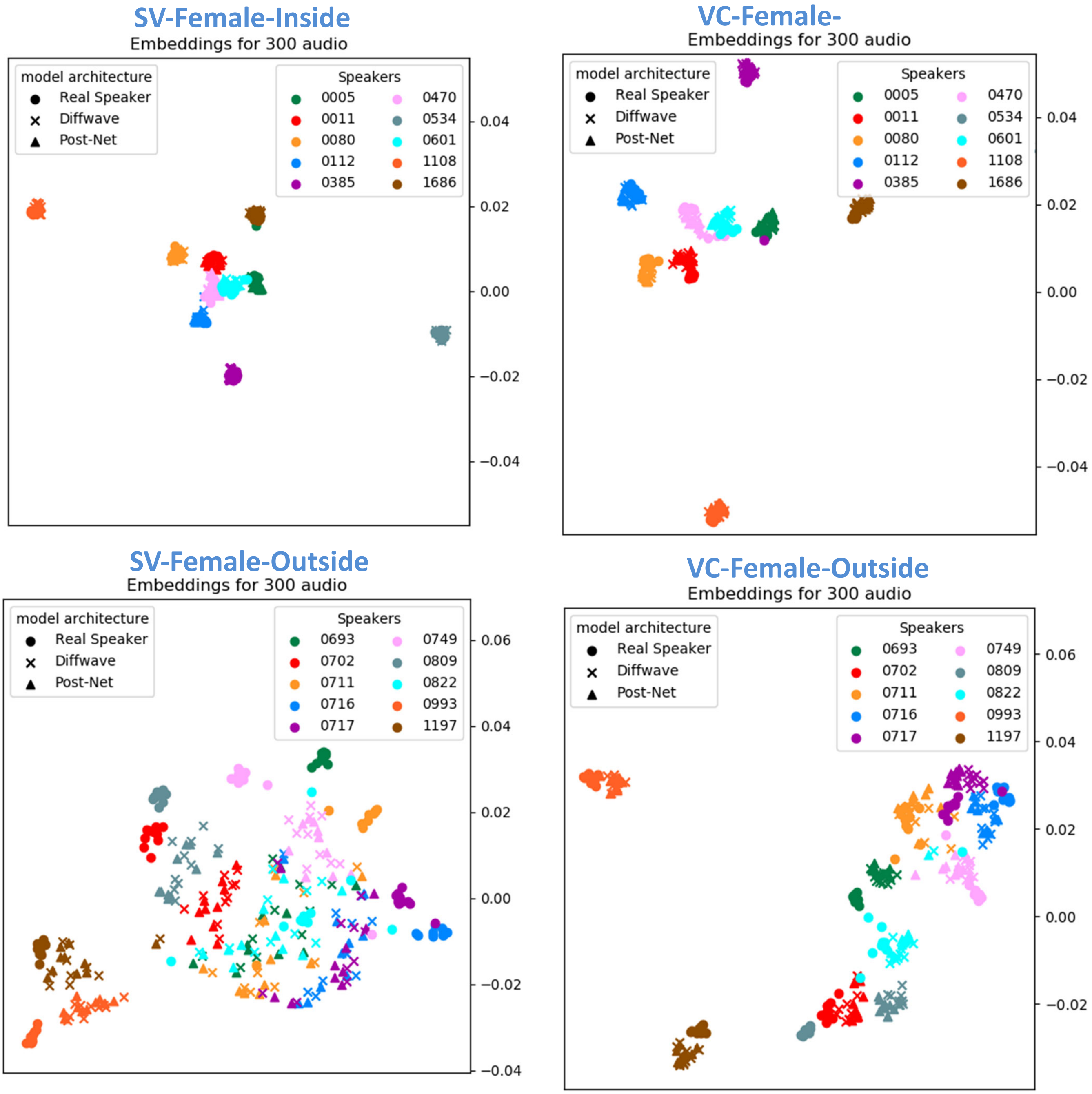
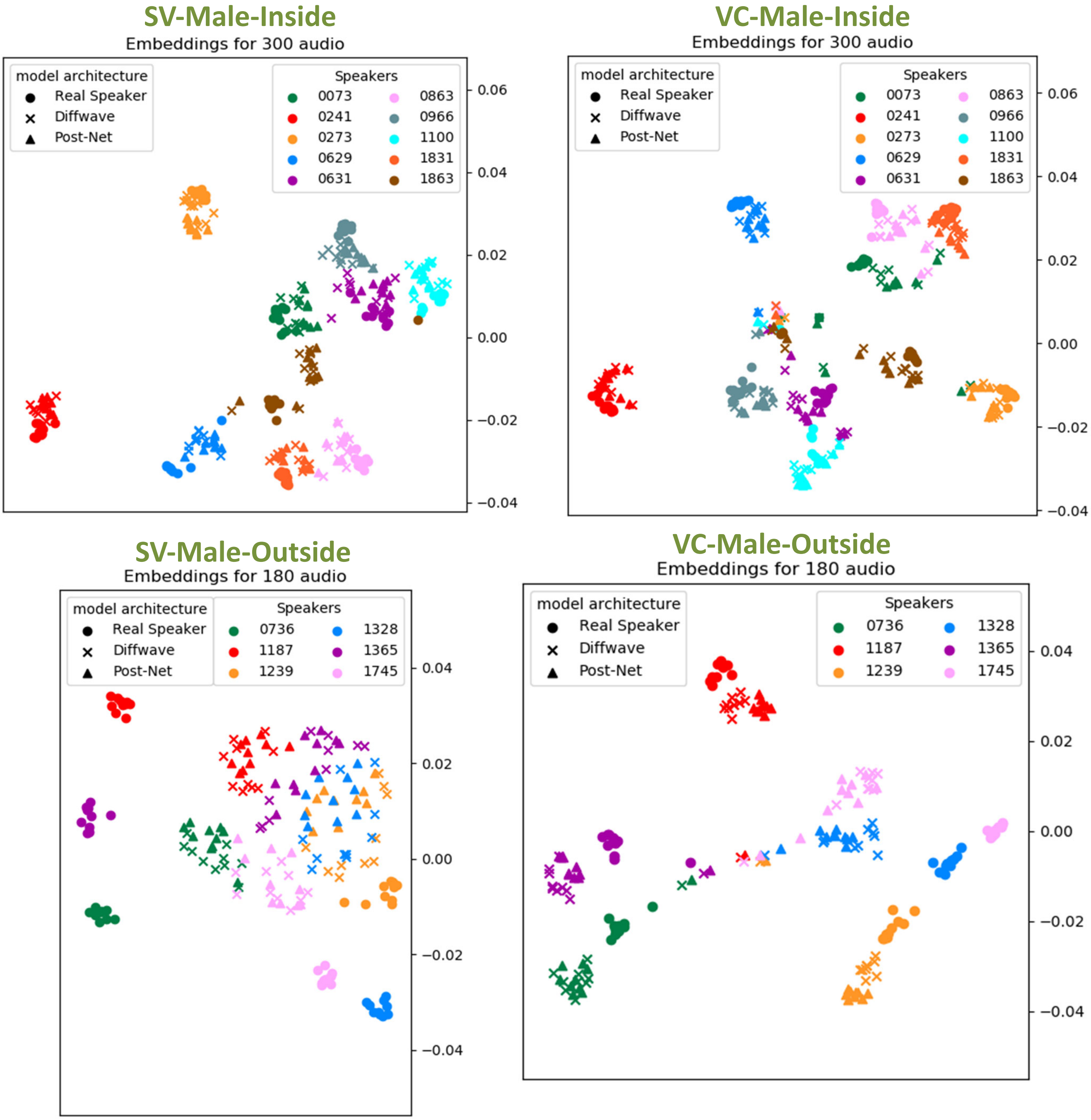


圖 8. 內外部女性語者的語者空間。

[Figure 8. Speaker space for inside and outside female speakers.]


SV-Male-Inside

VC-Male-Inside

SV-Male-Outside

VC-Male-Outside

圖 9. 內外部男性語者的語者空間。

[Figure 9. Speaker space for inside and outside male speakers.]

4.3.2 注意力機制的改動 (Change in Attention mechanism)

在我們所提出的架構中，Decoder 層引入了兩個注意力機制，分別為 Forward Attention 及 Bahdanau Attention，以 “今天天氣很好。(jin1 tian1 tian1 qi4 hen2 hao3.)” 為例子，對齊圖如下：

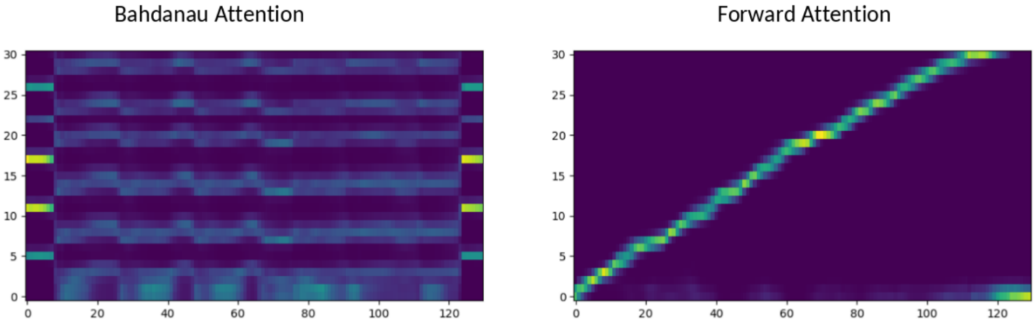


圖 10. 注意力對齊圖
[Figure 10. Attention alignment figure.]

可以看到圖 10 顯示出 Forward Attention 是斜對角的對齊圖，而 Bahdanau Attention 則包含了一些具有規律性的資訊，我們進一步針對 Bahdanau Attention 的對齊圖所顯示的資訊研究：

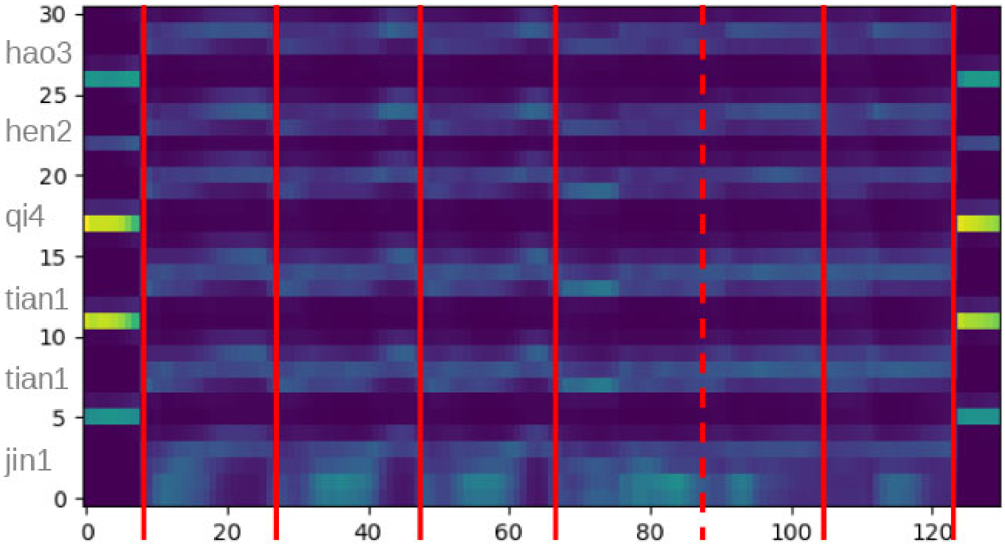


圖 11. 解析 Bahdanau Attention
[Figure 11. Parsing Bahdanau Attention.]

從圖 11 的紅線分段處，我們發現 Bahdanau Attention 提供了每段語音大概的音框範圍，圖中虛線左右處分別是”qi4”跟”hen2”的發音，由於它們主要都是氣音，導致分段沒有很明顯，而最左側及最右側對稱性的條紋可以判斷為空格資訊，即該片段是靜音的。

既然 Bahdanau Attention 夾帶每段語音大概的音框範圍資訊，那這些資訊是否能幫助模型快速建立對齊呢?下圖將顯示有無 Bahdanau Attention 的差異：

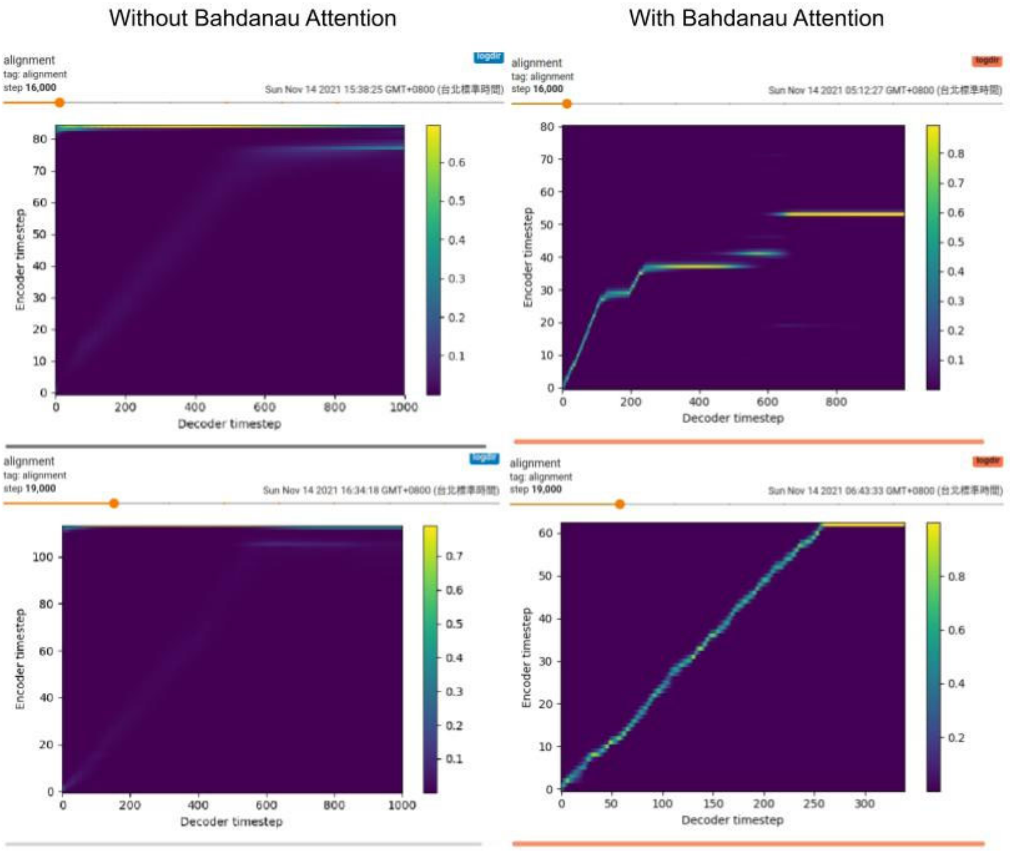


圖 12. Bahdanau Attention 能否幫助模型快速對齊？
[Figure 12. Can Bahdanau Attention help the model to align quickly?]

從圖 12 得知，模型訓練到 16000 個 Steps 時，儘管雙方都無法建立良好的對齊，但有 Bahdanau Attention 的對齊是優於沒有 Bahdanau Attention 的，在 19000 個 Steps 時，有 Bahdanau Attention 已經能建立對齊了，另一個則隱約有對齊線而已，因此可得知，Bahdanau Attention 加上 Forward Attention 的架構是能夠幫助模型快速地建立對齊。可以於我們的網站上聆聽樣本：https://babaili.github.io/rocling2021_demo/

5. 結論 (Conclusion)

在本次研究中，我們改進了多語者 Tacotron 2 的架構，透過加入語者嵌入向量便可合成未知語者的語音，並且比較語音轉換與語者辨識這兩個不同任務的語者嵌入向量用於 TTS 的成效，由實驗結果得知語音轉換的效果是優於語者辨識的，使用 Post-Filter 來提

升合成語音的語者相似度以及語音品質皆優於原始的 Post-Net，最後，於解碼層中添加第二個注意力機制有助於模型快速引發注意力對齊。未來的研究方向會加入更多資訊來幫助語者嵌入向量在 TTS 的效能上改進，例如：音韻(Prosody)資訊、發音(Articulation)資訊。

參考文獻(References)

Battenberg, E., Skerry-Ryan, R. J., Mariooryad, S., Stanton, D., Kao, D., Shannon, M., & Bagby, T. (2020). Location-relative attention mechanisms for robust long-form speech synthesis. In *Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6194-6198. https://doi.org/10.1109/ICASSP40776.2020.9054106

Chien, C. M., Lin, J. H., Huang, C. Y., Hsu, P. C., & Lee, H. Y. (2021). Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *Proceedings of ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8588-8592.

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., & Bengio, Y. (2015). Attention-based models for speech recognition. arXiv preprint arXiv:1506.07503.

Chou, J. C., Yeh, C. C., & Lee, H. Y. (2019). One-shot voice conversion by separating speaker and content representations with instance normalization. arXiv preprint arXiv:1904.05742.

Cooper, E., Lai, C. I., Yasuda, Y., Fang, F., Wang, X., Chen, N., & Yamagishi, J. (2020). Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *Proceedings of ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6184-6188. https://doi.org/10.1109/icassp40776.2020.9054535

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. arXiv preprint arXiv:1806.04558.

Kameoka, H., Kaneko, T., Tanaka, K., & Hojo, N. (2018). Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *Proceedings of 2018 IEEE Spoken Language Technology Workshop (SLT)*, 266-273. https://doi.org/10.1109/SLT.2018.8639535

Kaneko, T., & Kameoka, H. (2018). Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In *Proceedings of 2018 26th European Signal Processing Conference (EUSIPCO)*, 2100-2104. https://doi.org/10.23919/EUSIPCO.2018.8553236

Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. arXiv preprint arXiv:2010.05646.

Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761.

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019, July). Neural speech synthesis with transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6706-6713. https://doi.org/10.1609/aaai.v33i01.33016706

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,4779-4783. https://doi.org/10.1109/ICASSP.2018.8461368

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329-5333. https://doi.org/10.1109/ICASSP.2018.8461375

Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2019). Autovc: Zero-shot voice style transfer with only autoencoder loss. In *Proceedings of the 36th International Conference on Machine Learning(PMLR)*, 5210-5219.

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T. Y. (2020). Fastspeech 2: Fast and high-quality end-to-end text to speech. arXiv preprint arXiv:2006.04558.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems(NIPS'17)*, 5998-6008.

Zhang, J. X., Ling, Z. H., & Dai, L. R. (2018, April). Forward attention in sequence-to-sequence acoustic modeling for speech synthesis. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4789-4793. https://doi.org/10.1109/ICASSP.2018.8462020

Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 67-84

Answering Chinese Elementary School Social Studies Multiple Choice Questions

Chao-Chun Liang[*], Daniel Lee[†], Meng-Tse Wu[‡],

Hsin-Min Wang[*], and Keh-Yih Su[*]

Abstract

We present several novel approaches to answer Chinese elementary school social studies multiple choice questions. Although BERT shows excellent performance on various reading comprehension tasks, it handles some kinds of questions poorly, in particular *negation*, *all-of-the-above*, and *none-of-the-above* questions. We thus propose a novel framework to cascade BERT with preprocessor and answer-picker/selector modules to address these cases. Experimental results show the proposed approaches effectively improve the performance of BERT, and thus demonstrate the feasibility of supplementing BERT with additional modules.

Keywords: Natural Language Inference, Machine Reading Comprehension, Multiple Choice Question, Question and Answering.

1. Introduction

Machine reading comprehension (MRC) is a challenge for AI research, and is frequently adopted to seek desired information from knowledge sources such as company document collections, Wikipedia or the Web for a given question. To evaluate the capability of a MRC system, different test forms have been adopted in the literature (Qiu *et al*., 2019; Liu *et al*., 2019) such as binary choice, multiple choice (MC), multiple selection (MS), and cloze. Which test form to adopt usually depends on the format of the given benchmark/dataset. In this paper,

[*] Institute of Information science, Academia Sinica, Taipei, Taiwan
E-mail: {ccliang, whm, kysu}@iis.sinica.edu.tw

[†] Department of Computer Science Engineering, University of Michigan, Ann Arbor, Michigan, USA (His work was done during his summer internship in Institute of Information science, Academia Sinica).
E-mail: danclee@umich.edu

[‡] NYU Tandon School of Engineering, Brooklyn, NY, USA (His work was done during his research assistantship in Institute of Information science, Academia Sinica).
E-mail: michaelmoju@gmail.com

Table 1. Example social studies MC question.

Passage	三代同堂家庭是子女和父母、祖父母或外祖父母同住。
Question	「我和爸爸、媽媽、爺爺、奶奶住在一起。」是屬於哪一種類型的家庭？
Options	(1) 三代同堂家庭 (2) 單親家庭 (3) 隔代教養家庭 (4) 寄養家庭
Answer	(1) 三代同堂家庭

we solve MC questions about traditional Chinese primary school social studies. In this Chinese Social Studies MC (CSSMC) QA task, the system selects the correct answer from several candidate options based on a given question and its associated lesson manually constructed by Taiwan book publishers. Table 1 shows an example of CSSMC, where the passage is the corresponding supporting evidence (SE).

Previous work on answering MC questions can be divided into statistics-based approaches (Kouylekov & Magnini, 2005; Heilman & Smith, 2010) and neural-network-based approaches (Parikh *et al*., 2016; Chen *et al*., 2017). Recent pre-trained language models such as BERT (Devlin *et al*., 2019), XLNET (Yang *et al*., 2019), RoBERTa (Liu *et al*., 2019), and ALBERT (Lan *et al*., 2019) show excellent performance on different RC MC tasks. As BERT shows excellent performance on various English datasets (e.g., SQuAD 1.1 (Rajpurkar *et al*., 2016), GLUE (Wang *et al*., 2018), etc.), it is adopted as our baseline. Table 6 shows its performance given the gold SE.

After analyzing error cases, we observed that BERT handles the following question types poorly: (1) **Negation** questions, that is, questions with negation phrases such as 不可能 (unlikely). For this type of question, BERT selects the same answer for "小敏的媽媽目前在郵局服務，請問小敏的媽媽**可能**會為居民提供什麼服務？ (Xiaomin's mother serves at the post office. What kind of services could Xiaomin's mother provide to the residents?)" and "小敏的媽媽目前在郵局服務，請問小敏的媽媽**不可能**會為居民提供什麼服務？ (Xiaomin's mother serves at the post office. What kind of service could **not** Xiaomin's mother provide to the residents?)" (which differ only in the negation word 不 (**not**)). BERT evidently pays no special attention to negative words; however, any one of them would change the desired answer; (2) **All-of-the-above** (以上皆是) and **none-of-the-above** (以上皆非) questions, choices for which include either *All of the above* or *None of the above*. In both cases, the answer cannot be handled by simply by selecting the most likely choice without preprocessing

Table 2. Question types in CSSMC corpus.

Problem type	Questions
Negation	**Question**: 浩浩跟家人到臺東縣關山鎮遊玩，他不可能在當地看到什麼？ **Options**: (1)阿美族豐年祭 (2)環鎮自行車道 (3)油桐花婚禮 (4)親水公園
All of the above	**Question**: 在高齡化的社會裡，我們應該如何因應高齡化社會的到來？ **Options**: (1)制定老人福利政策 (2)提供良好的安養照顧 (3)建立健全的醫療體系 (4)以上皆是
None of the above	**Question**: 都市有公共設施完善、工作機會多等優點，常吸引鄉村地區哪一種年齡層的居民前往？ **Options**: (1)老人 (2)小孩 (3)青壯年 (4)以上皆非

the given choices. Table 2 shows an example of these question types.

The above phenomenon was also observed by Wu & Su (2020), who reported that BERT achieves superior results mainly by utilizing surface features, and that its performance degrades significantly when the dataset involves negation words. Moreover, it is difficult for BERT to learn the semantic meaning of all-of-the-above and none-of-the-above questions, which suggests that the listed candidate options are all correct or all incorrect, with a small amount of data.

However, it is difficult to pinpoint the sources of the problem and then find corresponding remedies within BERT, due to its complicated architecture (even its basic version includes 12 heads and 12 stacked layers). We thus prefer to keep its implementation untouched if the problem can be fixed by coupling BERT with external modules. Accordingly, we here propose a framework that cascades BERT with a preprocessor module and an answer-picker/selector module. The preprocessor module revises the choices for all-of-the-above and none-of-the-above questions, and the answer-picker/selector module (a postprocessor) determines the appropriate choices under the cases mentioned above. The above approach is inspired by Lin & Su (2021), who demonstrate that BERT learns natural language inference inefficiently, even for simple binary prediction; however, they also point out that task-related features and domain knowledge significantly help to improve BERT's learning efficiency.

For negation-type questions, instead of picking the highest-scoring choice as usual, the answer-picker/selector module selects the candidate with the lowest score. On the other hand, for all-of-the-above or none-of-the-above questions, we use a decision tree to select the answer, as illustrated in Figure 2. In these cases, the preprocessor module first replaces the original "all of the above" or "none of the above" choices with a new choice generated by concatenating all other choices together (before those candidates are sent to BERT). Take for example the second last row in Table 2: we replace "以上皆是 (all of the above)", the original last choice, with "制定老人福利政策^提供良好的安養照顧^建立健全的醫療體系 (Make welfare policies for elderly people^ Provide good nursing care^ Establish a sound medical system)".

We evaluate the proposed framework on a CSSMC dataset. The experimental results show the proposed approaches outperform the pure BERT model. This thus constitutes a new way to supplement BERT with additional modules. We believe the same strategy could be applied to other DNN models, which — despite good overall performance — are too complicated to customize for specific problems.

In summary, in this paper we make the following contributions: (1) We propose several novel approaches to supplement BERT to solve negation, all-of-the-above, and none-of-the-above questions. (2) Experimental results show that the proposed approach effectively improves performance, and thus demonstrate the feasibility of supplementing BERT with additional modules to fix given problems. (3) We construct and release a new Traditional Chinese Machine Reading Question and Answering dataset to assess the performance of RC MC models.

In comparison with our previous conference version (Lee *et al.*, 2020), this article describes additional "*Separately Judge then Select*" and "*Separately Judge Concatenation then Select*" experiments, which adopt a BERT entailment prediction model to handle each candidate option separately (details are provided in Sections 2.2.1 and 2.2.2) instead of jointly processing all candidate options together. We have also added Section 3 to describe the construction of the CSSMC dataset, which we adopt to compare different approaches.

2. Proposed Approaches

2.1 Problem Formulation

Given a social studies problem *Q* and its corresponding supporting evidence *SE*, our goal is to find the most likely answer from the given candidate set $A = \{A_1, A_2, \dots A_n\}$, where *n* is the total number of available choices or candidates, and A_i denotes the *i*-th answer candidate. This task is formulated as follows, where $\hat{A}$ is the answer to be chosen.

$$\hat{A} = \underset{i=1,\dots,n}{\arg\max}\, P(\mathrm{A}_i \mid Q, SE, A) \tag{1}$$

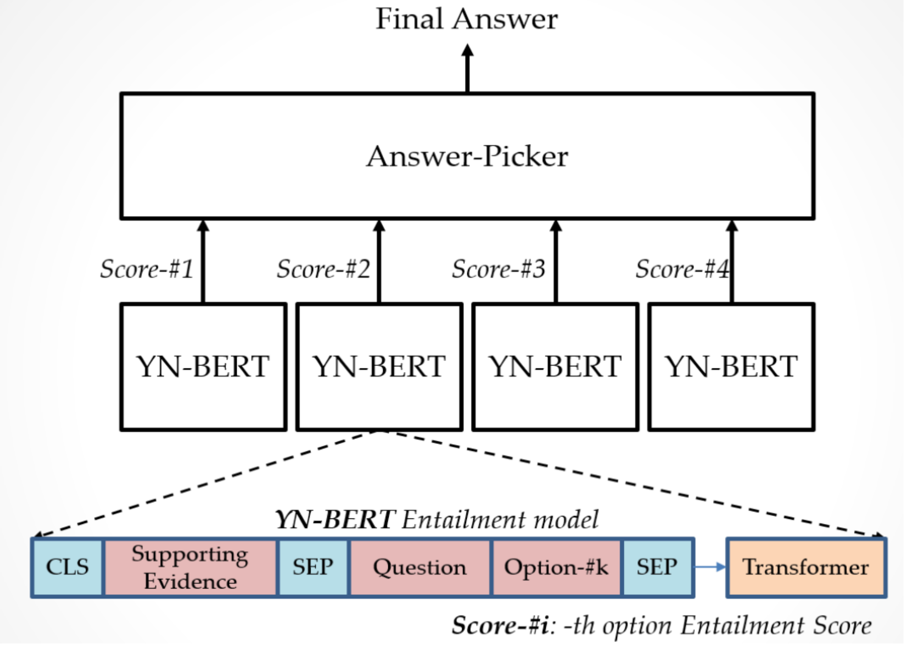


Figure 1. Architecture of proposed SJS approach.

2.2 Proposed Models

Three different approaches are proposed in which we use entailment prediction (Dagan *et al.*, 2005) to determine whether the candidate option is the correct answer to the question: (1) *Separately judge then select* (SJS), which considers each individual candidate option separately and then selects the final answer based on their output scores; (2) *Separately judge with concatenation then select* (SJCS), which adopts the framework of the first approach but first replaces the all-of-the-above (以上皆是) and none-of-the-above (以上皆非) answer choices with the concatenation of all the other remaining candidate options before entailment judgment; (3) *Jointly judge then select* (JJS), which jointly considers all candidate options to make the final decision. Details are provided below.

2.2.1 Separately Judge then Select (SJS)

Figure 1 shows the architecture of the proposed SJS approach, which consists of two main components: (1) the YN-BERT module, a fine-tuned BERT entailment prediction model (where YN denotes its output is a yes-no binary entailment judgment), and (2) the answer-picker module, which determines the final answer given the entailment judgment scores from four different YN-BERT modules. The input sequence is the concatenation of the associated supporting evidence, a given question, and a specific individual answer candidate/option. For each answer candidate, YN-BERT outputs an entailment judgment score

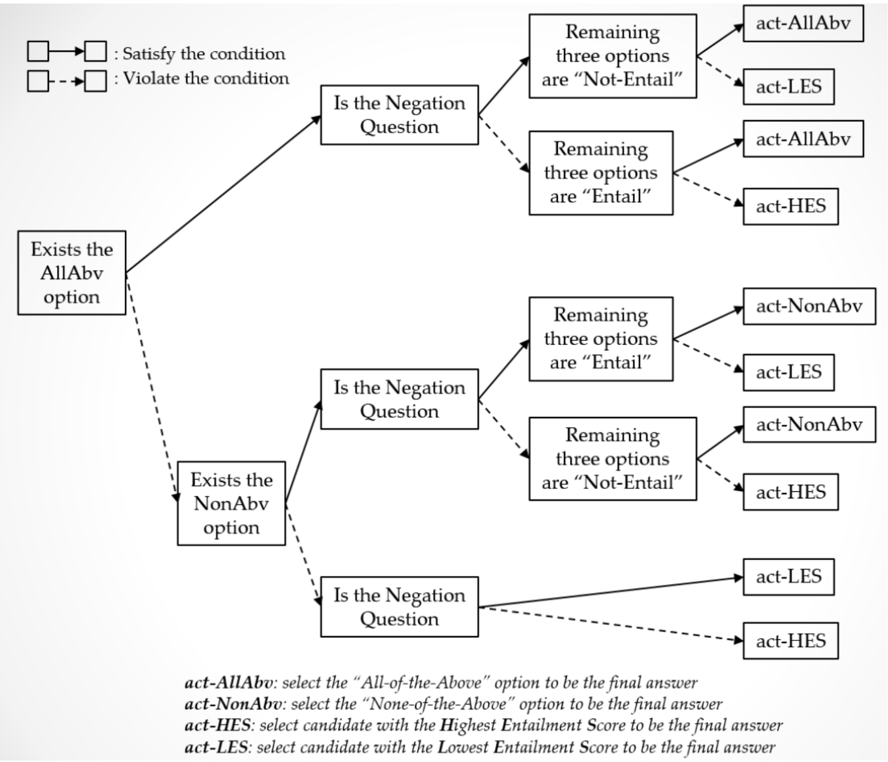
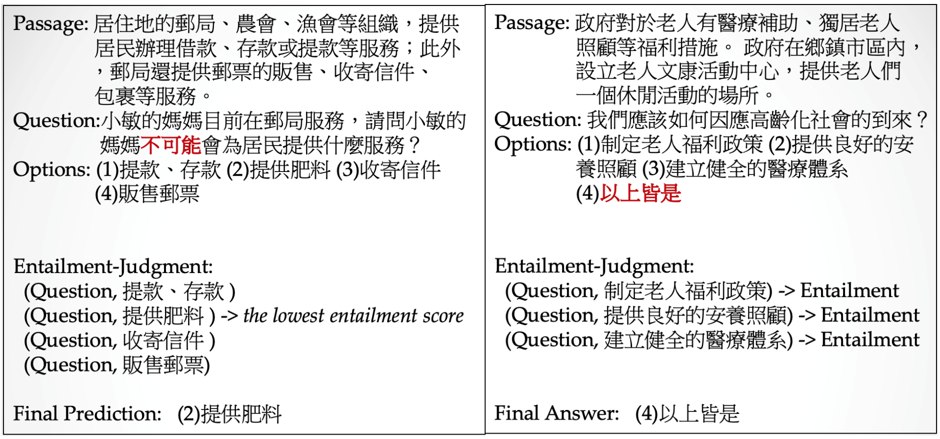

: Satisfy the condition
: Violate the condition
Exists the AllAbv option
Is the Negation Question
Remaining three options are "Not-Entail"
Remaining three options are "Entail"
act-AllAbv
act-LES
act-AllAbv
act-HES
Exists the NonAbv option
Is the Negation Question
Remaining three options are "Entail"
Remaining three options are "Not-Entail"
act-NonAbv
act-LES
act-NonAbv
act-HES
Is the Negation Question
act-LES
act-HES
act-AllAbv: select the "All-of-the-Above" option to be the final answer
act-NonAbv: select the "None-of-the-Above" option to be the final answer
act-HES: select candidate with the Highest Entailment Score to be the final answer
act-LES: select candidate with the Lowest Entailment Score to be the final answer

Figure 2. Decision tree for SJS approach. Each "act-xxx" is a specific action to be taken.

used to select either *Entail* or *Not-entail* (i.e., the judgment is *Entail* if the score exceeds 0.5, and *Not-entail* otherwise). *Entail* implies that the given answer candidate is entailed by the combination of the question and its associated supporting evidence. The answer-picker module considers the entailment judgment scores of the various choices and selects the most appropriate one based on the decision tree shown in Figure 2. Note that this decision tree is used only by the answer picker to make the final decision and is not involved in BERT's fine-tuning process.

A given question is classified as negative-type if it includes a negation word within a pre-specified negation word list, which is obtained from the CSSMC training data, and currently consists of {"不會 (will not)", "不能 (cannot)", "不得 (not allow)", "不是 (is not)", "不應該 (should not)", "不可能 (unlikely)", "不需 (do not need)", "不必 (do not need)", "不用 (do not need)", "沒有 (without)"}. Since the proposed approaches aim to supplement BERT, these negation words are manually picked from the error cases in the training data-set, on which BERT model make mistakes. Figure 3 shows the examples under two different inference mechanisms: (1) for a negation-type question (left figure), and (2) a question with all of the above option (right figure).



Passage: 居住地的郵局、農會、漁會等組織，提供居民辦理借款、存款或提款等服務；此外，郵局還提供郵票的販售、收寄信件、包裹等服務。
Question:小敏的媽媽目前在郵局服務，請問小敏的媽媽不可能會為居民提供什麼服務？
Options: (1)提款、存款 (2)提供肥料 (3)收寄信件 (4)販售郵票

Entailment-Judgment:
(Question, 提款、存款)
(Question, 提供肥料) -> *the lowest entailment score*
(Question, 收寄信件)
(Question, 販售郵票)

Final Prediction: (2)提供肥料

Passage: 政府對於老人有醫療補助、獨居老人照顧等福利措施。 政府在鄉鎮市區內，設立老人文康活動中心，提供老人們一個休閒活動的場所。
Question: 我們應該如何因應高齡化社會的到來？
Options: (1)制定老人福利政策 (2)提供良好的安養照顧 (3)建立健全的醫療體系 (4)以上皆是

Entailment-Judgment:
(Question, 制定老人福利政策) -> Entailment
(Question, 提供良好的安養照顧) -> Entailment
(Question, 建立健全的醫療體系) -> Entailment

Final Answer: (4)以上皆是

Figure 3. Two inference mechanisms under SJS framework.

2.2.2 Separately Judge with Concatenation then Select (SJCS)

Another approach adopts the framework of the first approach but first recasts “以上皆是 (all of the above)” and “以上皆非 (none of the above)’ answer candidates as the concatenation of all of the other options. Take for example the last row in Table 2: we replace “以上皆非”, the original last choice, with “老人^小孩^青壯年 (elderly people^children^young people)”. Afterwards, the answer-picker module selects the most appropriate choice based on the following rule: For negation questions, we select the answer candidate with the lowest entailment score; otherwise, we select that with the highest entailment score.

2.2.3 Jointly Judge then Select (JJS)

Shown in Figure 4, the system architecture of the JJS approach consists of three main components: (1) the preprocessor, which recasts “以上皆是 (all of the above)” and “以上皆非 (none of the above)” answer candidates as the concatenation of the other options (associated with the same question), as shown above, before inputting the question-choice-evidence combination into the BERT model; (2) the BERT-MC model, a typical fine-tuned BERT multiple-choice prediction model (Xu *et al*., 2019) described in Section 4.1; and (3) the answer selector, a candidate re-selector which for negation-type questions picks that answer candidate with the lowest score as opposed to that with the highest score (as for other question types).

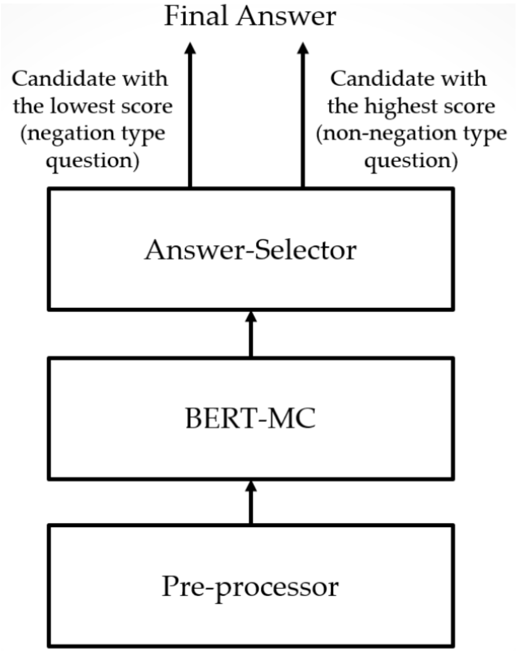

Final Answer
Candidate with the lowest score (negation type question)
Candidate with the highest score (non-negation type question)
Answer-Selector
BERT-MC
Pre-processor

Figure 4. System architecture of proposed "Jointly Judge then Select" framework.

3. Chinese Social Studies MRQA Dataset Construction

To evaluate the proposed approaches, we constructed a *Chinese Social Studies Machine Reading and Question Answering* (CSSMRQA) dataset, which is a superset of the CSSMC dataset mentioned above, to assess the capability of different Q&A systems (not just MC questions). This dataset consists of three question types: (1) yes/no questions, which ask whether the given question is a correct statement judged from the supporting evidence; (2) multiple-choice (MC) questions, which include four answer choices from which the correct one is to be chosen (here, this is the CSSMC dataset adopted in this paper); and (3) multiple-selection (MS) questions, which are similar to the multiple-choice questions but can contain more than one correct answer. Below we describe how they are constructed.

3.1 Corpus Collection

We first collected lessons for grades 3 to 6 from elementary-school social studies textbooks published in Taiwan. For each lesson, we collected relevant questions from leading publishing houses in Taiwan. We thus obtained 14,103 yes/no questions, 5347 MC questions, and 340 MS questions from a total of 255 lessons. We then annotated the supporting evidence to indicate what information is needed to answer each question. This is described in detail below.

3.2 Supporting Evidence (SE) Annotation

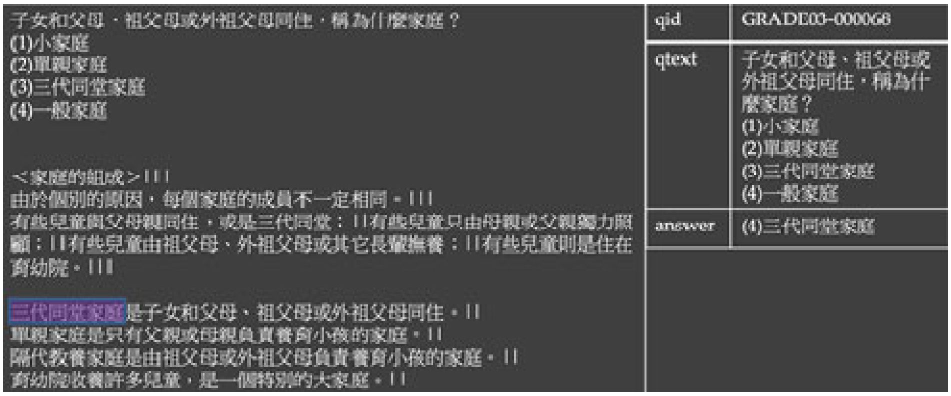
We hired two annotators to annotate the supporting evidence for each question. Supporting evidence is the content in the lesson (associated with the given question) which contains just the information necessary to answer the question. In the CSSMRQA dataset, each lesson comprises several paragraphs, and each paragraph comprises several sentences. Supporting evidence consists of one or more sentences.

We used Doccano (Nakayama *et al*., 2018), an open-source text annotation tool, as the platform for annotation. Doccano allows the user to highlight supporting words in the text (i.e., those words that provide hints to find the related passage). Given a question and its corresponding answer (also the lesson associated with the question), the annotators highlighted supporting words necessary to answer the question. Usually, these supporting words were words within the given question. Annotators were not allowed to annotate supporting words across sentence splitters or delimiters. Nonetheless, some questions lack suitable supporting evidence in the lesson. For example, students may rely on common sense (instead of textbook context) to answer the question, "班上同學有人亂丟垃圾，身為衛生股長的小玉可以怎麼做？ (1) 默默的跟在他們後面撿垃圾 (2) 勸告亂丟垃圾的同學，並請他們將垃圾撿起來 (3) 沒關係，等打掃時間再掃就好了 (4) 把垃圾藏在看不見的地方 (What can Xiaoyu (the Chief of Health) do when her classmate litters? (1) Pick up trash after them silently; (2) Advise the classmate who litters and ask him/her to pick up the litter; (3) It doesn't matter, just wait until the cleaning time; or (4) Hide litter out of sight)". In such cases, annotators found no suitable supporting words in the lesson and thus skipped SE annotation. Afterward, sentences that contain marked supporting words were annotated as supporting evidence. Table 3 shows the final results of SE annotation.

Table 3. Supporting evidence annotation in training/dev/test subsets.

Subset	Training	Dev	Test
Questions	3,879	780	778
Questions w/ SE	3,135	604	563
Questions w/o SE	744	176	215
Averaged SPs	1.09	1.16	1.14
Averaged SSs	3.17	2.94	2.73

*Questions w/o SE: the number of questions without supporting evidence
Averaged SPs: the average number of Supporting Paragraphs
Averaged SSs: the average number of Supporting Sentences



子女和父母、祖父母或外祖父母同住，稱為什麼家庭？
(1)小家庭
(2)單親家庭
(3)三代同堂家庭
(4)一般家庭

＜家庭的組成＞|||
由於個別的原因，每個家庭的成員不一定相同。|||
有些兒童與父母親同住，或是三代同堂；||有些兒童只由母親或父親獨力照顧；||有些兒童由祖父母、外祖父母或其它長輩撫養；||有些兒童則是住在育幼院。|||

三代同堂家庭是子女和父母、祖父母或外祖父母同住。||
單親家庭是只有父親或母親負責養育小孩的家庭。||
隔代教養家庭是由祖父母或外祖父母負責養育小孩的家庭。||
育幼院收養許多兒童，是一個特別的大家庭。||

qid	GRADE03-000068
qtext	子女和父母、祖父母或外祖父母同住，稱為什麼家庭？ (1)小家庭 (2)單親家庭 (3)三代同堂家庭 (4)一般家庭
answer	(4)三代同堂家庭

Figure 5. Multiple-choice question annotation.

Figure 5 shows an example of multiple-choice question annotation. Annotators first read both the question (qtext) and the correct answer (answer) from the right-hand side windows, and then highlight supporting words (marked with purple boxes) in the lesson. To prevent annotators from highlighting supporting word regions across sentences, we use special symbols as separators (||| for paragraphs and || for sentences).

4. Experiments

We conducted experiments on the above CSSMC dataset with the three proposed approaches. Table 4 shows the dataset statistics. For comparison, we used a typical BERT multiple-choice implementation (Xu *et al*., 2019) as our baseline.

4.1 Baseline: BERT-MC

For the baseline, we used the BERT-MC model, that is, BERT (Devlin *et al*., 2019) fine-tuned for the multiple-choice task as our baseline, as it is the most widely adopted state-of-the-art model (Xu *et al*., 2019). It was built by exporting BERT's final hidden layer into a linear layer and then taking a softmax operation. For details on the BERT-MC model, please see Xu *et al*. (2019). The BERT input sequence consists of "[CLS] SE [SEP] Question [SEP] Option-#*i* [SEP]", where Option-#*i* denotes the *i*-th option and [CLS] and [SEP] are special tokens representing the classification and the passage separators, respectively, as defined in Devlin *et al*. (2019). Figure 6 shows the architecture of the BERT baseline model.

Table 4. CSSMC dataset.

	Training	Dev	Test
Lessons	202	27	26
Questions	3,879	780	778
Averaged paragraphs/lesson	11.28	13.93	10.93
#Averaged entences/lesson	46.40	52.67	46.33

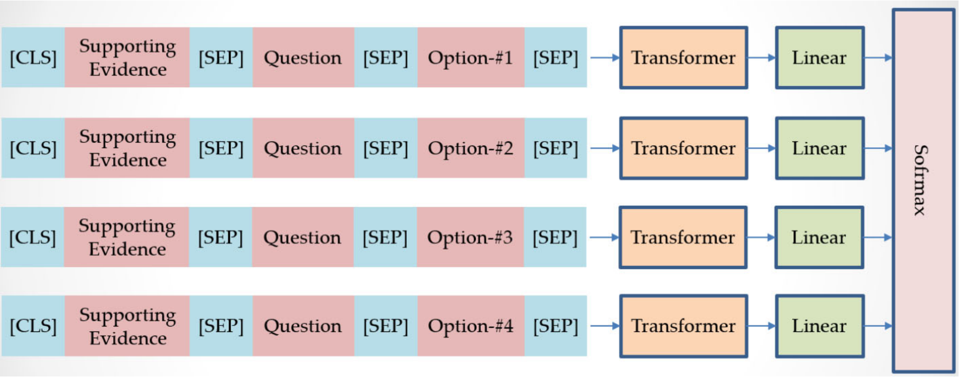


Figure 6. The architecture of the BERT-MC model (Xu et al., 2019).

4.2 Retrieved Supporting Evidence (SE) Dataset

SE is the corresponding shortest passage based on which the system can answer the given question. Given the annotation results described in Section 3.2, we find many questions that involve common-sense reasoning, for which no corresponding SEs can be found in the retrieved lesson. We denote as SE1 that set of questions for which SEs can be found in the retrieved lesson (this is termed GSE1 if it is also associated with gold SEs); the set of remaining questions is SE2. Table 5 shows the statistics for GSE1.

Table 5. CSSMS GSE1 (with gold SEs) subset statistics.

	Training	Dev	Test
Lessons	196	27	26
Questions (NEG[a]) (AllAbv&NonAbv[b])	3,135 (53) (332)	604 (14) (69)	563 (15) (56)
Averaged paragraphs/lesson	11.35	13.93	10.85
Averaged sentences/ Lesson	46.72	52.67	46.15

[a] NEG: number of negation-type questions.

[b.] AllAbv&NonAbv: number of AllAbv&NonAbv-type questions.

4.3 Results

We conducted two sets of experiments on the CSSMC dataset: (i) GSE1, based on SE1 with gold SEs, to compare the QA component performance of different models; and (ii) LSE, based on the whole dataset with all SEs directly retrieved from the Lucene search engine, to compare different approaches under a real-world situation. Each set covers six different models: (1) *BERT-MC Only*, (2) *SJS*, (3) *SJCS*, (4) *BERT-MC+Neg*, (5) *BERT-MC+AllAbv&NonAbv*, and (6) *BERT-MC+Neg+AllAbv&NonAbv*, where *BERT-MC Only* is the baseline model and *Neg* and *AllAbv&NonAbv* denote additional answer-selector and preprocessor modules for the negation and all-of-the-above/none-of-the-above question-types, respectively. We adopted the setting specified in Xu *et al.* (2019) for BERT training. All other models were trained using the following hyperparameters: (1) a maximum sequence length of 300; (2) a learning rate of 5e-5 with the AdamW optimizer (Loshchilov & Hutter, 2019); (3) 3 to 5 epochs. Table 6 compares the accuracy of various approaches; we report test set performance using the settings that corresponded to the best dev set performance.

4.3.1 Jointly Judge then Select (JJS)

In this scenario we sought to evaluate the QA component performance of six different models on the GSE1 subset (i.e., with gold SEs). The GSE1 column in Table 6 gives the test set accuracy rates of various approaches. As the *SJS* model has special handling for negation and “以上皆是 (all-of-the-above)” or “以上皆非 (none-of-the-above)” questions, it yields better performance than *BERT-MC Only* (0.862 vs. 0.849). The *SJCS* model further replaces the “以上皆是 (all-of-the-above)” and “以上皆非 (none-of-the-above)” options with the concatenation of the three other options. However, this degrades the baseline performance significantly, from 0.849 to 0.822. This is because the “以上皆是 (all-of-the-above)” and

Table 6. Test-set performance comparison.

	GSE1[a]	***GSE1-Neg***[b]	***GSE1-AllAbv&NonAbv***[c]	***LSE***[d]
BERT-MC only (baseline)	0.849	0.200	0.643	0.692
SJS	0.862	NA	NA	0.694
SJCS	0.822	NA	NA	0.661
BERT-MC **+ *Neg***	0.870	**0.400**	NA	0.695
BERT-MC **+ *AllAbv&NonAbv***	0.879	NA	**0.839**	0.719
BERT-MC **+ *Neg*** **+ *AllAbv&NonAbv*** ***(also JJS)***	**0.879**	NA	NA	**0.725**

[a] GSE1: SE1 subset with gold SEs.

[b] GSE1-Neg: Only negation-type questions within GSE1.

[c] GSE1-AllAbv&NonAbv: Only AllAbv&NonAbv-type questions within GSE1.

[d] LSE: <SE1+SE2> with all SEs retrieved from the Lucene search engine.

"以上皆非 (none-of-the-above)" options are closely related to the other three options. However, as it considers the concatenation option and the other three options independently, or separately, without using a complicated decision tree (specified in Figure 3), this approach is unable to take such correlation into account.

The *JJS* model (i.e., the last row in Table 6) addresses this problem by considering all of the options together simultaneously. Table 6 shows that it considerably outperforms the *SJCS* model by 5.7% (87.9% - 82.2%) on the test set, which shows that jointly processing all options together is essential after the concatenation step. The *BERT-MC+Neg* and *BERT-MC+AllAbv&NonAbv* models are also evaluated as an ablation analysis. Table 6 indicates they also outperform the *BERT-MC only* baseline by 2.1% (87.0% - 84.9%) and 3.0% (87.9% - 84.9%) on the test set, respectively, which shows the necessity of both the preprocessor and answer-selector modules.

Last, to explore the effects of the proposed approaches on specific question types, we conducted two additional experiments on two GSE1 subsets: (1) the *Neg-type only* subset, which contains only negation questions, to compare the performance between the BERT-MC only and *BERT-MC+Neg* approaches to evaluate the effectiveness of the answer-selector module; (2) the *AllAbv&NonAbv* only subset, which contains only *AllAbv* or *NonAbv* questions, to compare the *BERT-MC only* and *BERT-MC+AllAbv&NonAbv* approaches to evaluate the effectiveness of the proposed preprocessor. Table 6 clearly shows

Table 7. Error case of "BERT-MC+Neg" on "GSE1-Neg" subset.

SEs: 另外，隨著商業興盛，在府城、鹿港、艋舺等大城市，也出現由商人組成的「郊」。「郊」類似現代同業公會，成員除了經營貿易外，也積極參與地方的公共事務。

Question: 清朝統治臺灣時期，怎樣的人應該比較沒有共同的血緣？

Options: (1)參加同一個宗親會 (2)參加同一個祭祀公業 (3)參加同一個「郊」
(4) 在同一座宗祠祭祀祖先

Table 8. Error case of "BERT-MC+AllAbv&NonAbv" on "GSE1-AllAbv&NonAbv" subset.

SE:工業生產如果沒有適當處理，很容易破壞周遭環境，造成空氣汙染、噪音汙染、水質汙染、土地汙染等。例如：工業廢水或是家庭汙水直接排入河流，不僅危害河流生態，有毒物質如果流入大海，通過食物鏈進入人體，更會嚴重損害健康。

Question: "志忠家附近有一間工廠，時常將未經處理的汙水排入河川中，這樣可能會造成什麼後果？"

Options: (1)空氣汙染 (2)噪音 (3)水質汙染 (4)以上皆是

that the preprocessor (GSE1-Neg column) and answer-selector (GSE1-AllAbv&NonAbv column) modules effectively enhance *BERT-MC* on these two subsets (from 20% to 40%, and from 64.3% to 83.9%, respectively). The above experiments sufficiently demonstrate the effectiveness of our proposed approaches (unnecessary combinations are marked "NA" in Table 6).

The remaining errors in the *GSE1-Neg* and *GSE1-AllAbv&NonAbv* subsets are mainly due to that answering those questions requires further inference capability. Table 7 shows that we need to know that "商人 (businessmen)" are people without "共同的血緣 (blood relations)". Similarly, Table 8 shows that we need to know that "未經處理的汙水排入河川 (untreated sewage discharged into the river)" causes "水質汙染 (water pollution)".

4.3.2 LSE (SE1+SE2 with all SEs retrieved from Lucene)

Since the gold SE is not available for real-world applications, this scenario compares the system performance of different models in a real-world situation. That is, we evaluated various models with all the SEs retrieved from a search engine (i.e., *Apache Lucene* (https://lucene.apache.org/)). Furthermore, to support those questions for which no associated SEs from the lessons (i.e., the SE2 subset), we used Wikipedia as an external knowledge resource to provide SEs when possible. We first used Lucene to search the Taiwan elementary-school social studies textbook and Wikipedia separately to yield two different SEs, after which we constructed a fused SE by concatenating these two SEs with the format "**Textbook-SE [SEP] Wiki-SE**" where Textbook-SE and Wiki-SE denote the two SEs retrieved from the textbook and Wikipedia, respectively.

Table 9. Error types.

Error Type	Questions
Incorrect supporting evidence (52%)	**Wrong SE**: *清朝統治臺灣初期，漢人渡海來臺後，往往同鄉人聚居在一起，並且建築廟宇供奉共同信仰的神明。* **Question**: 臺灣有許多從中國移民來的漢人，來臺要渡過危險的臺灣海峽，所以什麼神明就被所有移民所共同信仰？ **Options**: (1)關公 (2)土地公 (3)媽祖 (4)三山國王
Requires advanced inference capability (48%)	**SE**: 刑法對傷害他人的行為加以處罰；民法則以損害賠償的方式，請問牛奶的保存期限過了沒？（相關法律：民法、消費者保護法、食品安全衛生管理法） **Question**: “小花在超市買到過期的餅乾，請問該超市的販售行為違反什麼法律？” **Options**: (1)刑法 (2)憲法 (3)教育基本法 (4)食品安全衛生管理法

Experimental results (the LSE column in Table 6) show that both the preprocessor and the answer selector effectively supplement BERT-MC; performance is improved further when they are jointly adopted (3.3% = 72.5% - 69.2%). Furthermore, the accuracy of the BERT-MC only model on LSE is significantly lower than that on GSE1 (69.2% vs. 84.9%), which clearly illustrates that extracting good SEs is essential in QA tasks. Last, to show the influence of incorporating Wikipedia, we conducted an experiment in which we used only Lucene to search the textbook. The BERT-MC+Neg+AllAbv&NonAbv model now drops to 70.4% (not shown in Table 6) from 72.5%, which shows that Wikipedia provides the required common sense for some cases.

5. Error Analysis and Discussion

We randomly selected 40 error cases from the test set of the *BERT-MC+Neg+AllAbv&NonAbv* model under the “*all SEs retrieved from Lucene*” scenario. We found that all errors come from two sources: (1) the correct support evidence was not retrieved (52%), and (2) the answer requires deep inference (48%). Table 9 shows an example for each category. For the first example, the retrieved SE is irrelevant to the question; our model thus fails to produce the correct answer. The second example illustrates that the model requires further inference capability to know that both “牛奶的保存期限過了沒 (Has the milk expired?)” and “在超市買到過期的餅乾 (I bought expired cookies in the supermarket)” are similar events related to “食品安全衛生管理法 (Act Governing Food Safety and Sanitation)”.

6. Related Work

Before 2015, most work on entailment judgment adopted statistical approaches (Kouylekov & Magnini, 2005; Heilman & Smith, 2010). In subsequent work, neural network models were widely adopted due to the availability of large datasets such as RACE (Lai *et al*., 2017) and SNLI (Bowman *et al*., 2015). Parikh *et al*. (2017) propose the first alignment-and-attention mechanism, achieving state-of-the-art (SOTA) results on the SNLI dataset. Chen *et al*. (2017) further propose a sequential inference model based on chain LSTMs which outperforms previous models. In recent work, pre-trained language models such as BERT (Devlin *et al*., 2019), XLNET (Yang *et al*., 2019), RoBERTa (Liu *et al*., 2019) and ALBERT (Lan *et al*., 2019) yield superior performance on MC RC tasks. However, these results are obtained mainly by utilizing surface features (Jiang & Marneffe, 2019). Besides, Zhang *et al*. (2020) propose a dual co-matching network to model relationships among passages, questions, and answer candidates to achieve SOTA results for MC questions. Also, Jin *et al*. (2020) propose two-stage transfer learning for coarse-tuning on out-of-domain datasets and fine-tuning on larger in-domain datasets to further improve performance. In comparison with those previous approaches, instead of adopting a new inference NN, our proposed approaches supplement the original BERT with additional modules to address two specific problems that BERT handles poorly.

7. Conclusion

We present several novel approaches to supplement BERT with additional modules to address problems with three specific types of questions that BERT-MC handles poorly (i.e., negation, all-of-the-above, and none-of-the-above). The proposed approach constitutes a new way to enhance a complicated DNN model with additional modules to pinpoint problems found in error analysis. Experimental results show the proposed approaches effectively improve performance, and thus demonstrate the feasibility of supplementing BERT with additional modules to fix specific problems.

Reference

Bowman, S. R., Angeli, G., Potts, C. & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632-642. https://doi.org/10.18653/v1/D15-1075

Chen, D. (2018). Neural Reading Comprehension and Beyond. (Doctoral Dissertation). Stanford Univ..

Chen, Q., Zhu, X., Ling, Z., Wei, S., Jiang, H. & Inkpen, D. (2017). Enhanced LSTM for Natural Language Inference. In *Proceedings of the 55th Annual Meeting of the*

Association for Computational Linguistics, 1657-1668. https://doi.org/10.18653/v1/P17-1152

Dagan, I., Glickman, O., & Magnini, B. (2005) The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, Springer,* 177-190. https://doi.org/10.1007/11736790_9

Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171-4186. https://doi.org/10.18653/v1/N19-1423

Heilman, M. & Smith, N. A. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics,* 1011-1019.

Jiang, N. & Marneffe, M.-C. D. (2019). Evaluating BERT for natural language inference: A case study on the CommitmentBank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* 6086-6091. https://doi.org/10.18653/v1/D19-1630

Jin, D., Gao, S., Kao, J. Y., Chung, T., & Hakkani-tur, D. (2020). MMM: Multi-stage multi-task learning for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 8010–8017. https://doi.org/10.1609/aaai.v34i05.6310

Kouylekov, M. & Magnini, B. (2005). Recognizing textual entailment with tree edit distance algorithms. In *Proceedings of the First Challenge Workshop Recognising Textual Entailment 2005,* 17-20.

Lai, G., Xie, Q., Liu, H., Yang, Y. & Hovy, E. (2017). RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* 785–794. https://doi.org/10.18653/v1/D17-1082

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P. & Soricut, R. (2019). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.

Lee, D., Liang, C. C. & Su, K. Y. (2020). Answering Chinese Elementary School Social Study Multiple Choice Questions. In *Proceedings of the 2020 International Conference on Technologies and Applications of Artificial Intelligence*.

Lin, Y.C. & Su, K.Y. (2021). How Fast can BERT Learn Simple Natural Language Inference? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 626-633. https://doi.org/10.18653/v1/2021.eacl-main.51

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

Liu, S., Zhang, X., Zhang, S., Wang, H., & Zhang, W. (2019). Neural Machine Reading Comprehension: Methods and Trends. *Applied Sciences, 9*(18)*,* 3698. https://doi.org/10.3390/app9183698

Loshchilov, I. & Hutter, F. (2019). Decoupled Weight Decay Regularization. In Proceedings of *International Conference on Learning Representations 2019*.

Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., & Liang, X. (2018). Doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano

Parikh, A. P., Tackstrom, O., Das, D. & Uszkoreit, J. (2016). A Decomposable Attention Model for Natural Language Inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 2249-2255. https://doi.org/10.18653/v1/D16-1244

Qiu, B., Chen, X., Xu, J., & Sun, Y. (2019). A Survey on Neural Machine Reading Comprehension. arXiv preprint arXiv:1906.03824.

Rajpurkar, P., Zhang, J., Lopyrev, K. & Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing,* 2383–2392. https://doi.org/10.18653/v1/D16-1264

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP,* 353–355. https://doi.org/10.18653/v1/W18-5446

Wu, T. M. & Su, K. Y. (2020). Making Negation-word Entailment Judgment via Supplementing BERT with Aggregative Pattern. In *International Conference on Technologies and Applications of Artificial Intelligence (TAAI 2020),* 17-22. https://doi.org/10.1109/TAAI51410.2020.00012

Xu, K., Tin, J., & Kim, J. (2019). A BERT based model for Multiple-Choice Reading Comprehension. Retrieved from http://cs229.stanford.edu/proj2019spr/report/72.pdf

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. & Le, Q. C. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of Advances in neural information processing systems 32 (NIPS 2019),* 5753–5763.

Zhang, S., Zhao, H., Wu, Y., Zhang, Z., Zhou, X., & Zhou, X. (2020). DCMN+: Dual co-matching network for multi-choice reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(05), 9563-9570. https://doi.org/10.1609/aaai.v34i05.6502

The individuals listed below are reviewers of this journal during the year of 2021. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Tao-Hsing Chang

Jia-Wei Chang

Ru-Yng Chang

Yung-Chun Chang

Miao-Hsia Chang

Kuan-Yu Chen

Alvin C.-H. Chen

Chung-Chi Chen

Yun-Nung (Vivian) Chen

Chen-Yu Chiang

Yu-Tai Chien

Hong-Jie Dai

Min-Yuh Day

Janice Fon

Zhao-Ming Gao

Mei-Ching Ho

Shu-Kai Hsieh

Yu-Lun Hsieh

Fu-Hui Hsieh

Wen-Lian Hsu

Hen-Hsen Huang

Yi-Chin Huang

Han-Chun Huang

Jeih-Weih Hung

Chih-Hao Ku

Win Ping Kuo

Wen-Hsing Lai

Ying-Hui Lai

Hong-Yi Lee

Cheng-Te Li

Chun-Yen Lin

Jen-Chun Lin

Szu-Yin Lin

Shih-Hung Liu

Jenn-Long Liu

Yi-Fen Liu

Shang-Pin Ma

Shu-chen Ou

Ming-Hsiang Su

Emily Chia-Yu Su

Wei-Ho Tsai

Chin-Chin Tseng

Chun-Wei Tung

Jenq-Haur Wang

Yu-Cheng Wang

Yu-Fang Wang

Jiun-Shiung Wu

Jheng-Long Wu

Jian-Cheng Wu

Jui-Feng Yeh

2021 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 26

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2021

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
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, 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
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan, R.O.C.

Tel.：886-2-2788-1638 Fax：886-2-2651-9386

E-mail: aclclp@aclclp.org.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：

社團法人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：請至本會網頁填妥入會申請表，填妥後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115022南港區舊莊街一段3巷34號1樓
電話：(02) 2788-1638　　傳真：(02) 2651-9386
E-mail：aclclp@aclclp.org.tw　網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

社團法人中華民國計算語言學學會
個人會員入會申請書

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

審查結果:

1. 年費：

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

2. 連絡處：

   地址：台北市115022南港區舊莊街一段3巷34號1樓
   電話：(02) 2788-1638　　　傳真：(02) 2651-9386
   E-mail：aclclp@aclclp.org.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 Banl:________________

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

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

CARD HOLDER SIGNATURE : ___________________________________

Address: ____

Tel.: ____________________________ E-mail : ________________________

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 886-2-2651-9386 or Mail this form to :**

Association for Computational Linguistics and Chinese Language Processing
1F., No. 34, Ln. 3, Sec. 1, Jiuzhuang St., Nankang Dist., Taipei City, 115022, Taiwan, R.O.C
E-mail: aclclp@aclclp.org.tw Website: http://www.aclclp.org.tw

社團法人中華民國計算語言學學

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$_________ □ 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 02-26519386 或郵寄至:
115022台北市南港區舊莊街一段3巷34號1樓 中華民國計算語言學學會 收
E-mail: aclclp@aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

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

10% member discount: ___________Total Due:__________

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

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

社團法人中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息為本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01　「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表 (甲)	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義 （中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊 (一年兩期)　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) 2788-1638

聯絡人：　黃琪 小姐、何婉如 小姐　　E-mail:aclclp@aclclp.org.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.

Online Submission: https://ijclclp.aclclp.org.tw/servlet/SignInHandler

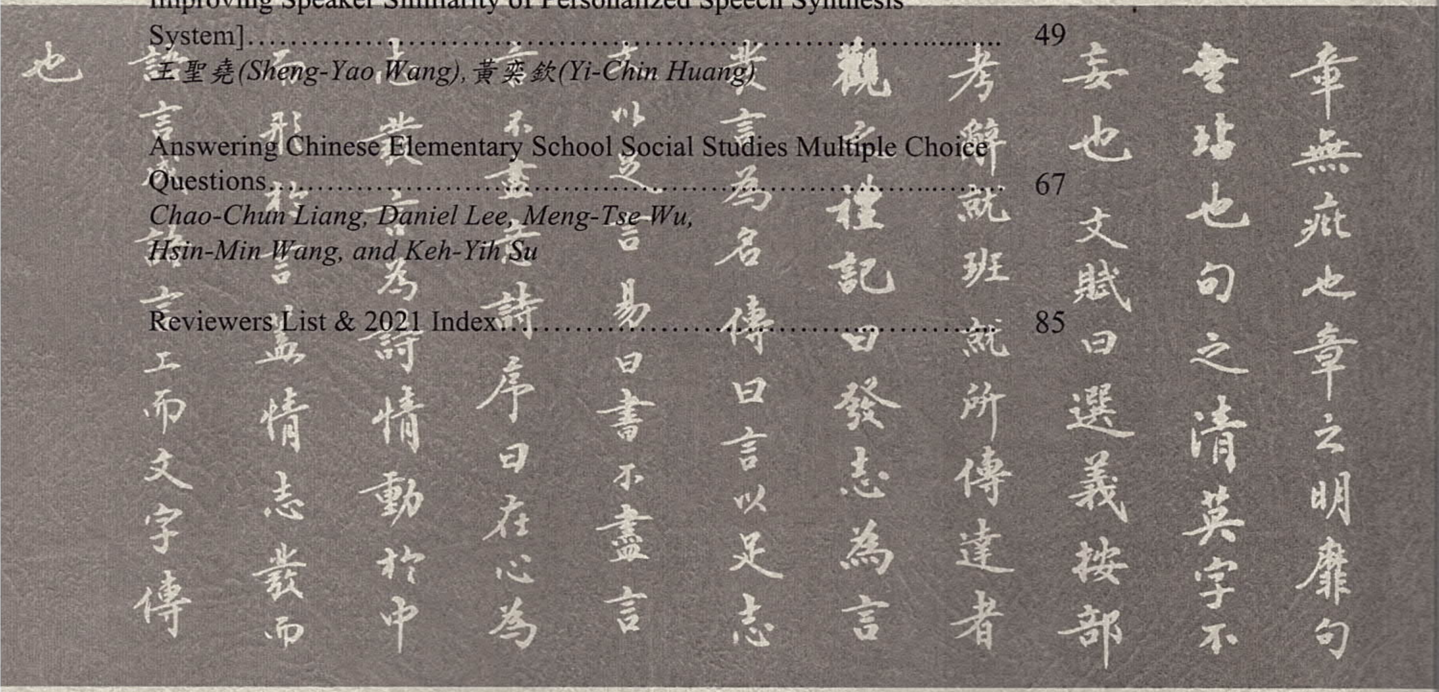
Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

For more information, please email to ijclclp@aclclp.org.tw

International Journal of
Computational Linguistics & Chinese Language Processing
vol.26 no.2 December 2021

Contents

Papers



ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing