The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 137-151

基於深度類神經網路之多模式情感偵測初步探討

A Preliminary Study on Deep Learning Neural Networks-based Multi-Model Sentiment Detection

陳泰融 Tai-Rong Chen, 廖元甫 Yuan-Fu Liao
國立臺北科技大學電子工程系
Department of Electronic Engineering, National Taipei University of Technology
t106368030@ntut.edu.tw, yfliao@mail.ntut.edu.tw

潘振銘 Chen-Ming Pan, 郭姿秀 Tzu-Hsiu Kuo
Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan
chenming@cht.com.tw, gaga820402@cht.com.tw

Matúš Pleva, Daniel Hládek
Department of Electronics and Multimedia Communications, Technical University of Košice, Slovakia
matus.pleva@tuke.sk, daniel.hladek@tuke.sk

摘要

為慶祝登月計劃五十週年，德州大學達拉斯分校（UTDallas）將登月任務中，所有太空人與任務中心間的通訊對話錄音進行數位化，發行 Fearless Steps Corpus 語料，並舉辦 Fearless Steps Challenge 競賽，希望能增進各種語音處理相關技術發展。本論文即針對其中的語音情緒偵測任務，進行初步探討。主要想法是同時考慮語音訊號中包含的聲學與語意資訊，提出基於深度類神經網路之多模式語音情緒偵測模型，用以偵測語音訊號中傳達的情緒狀態。實際做法包括(1)利用捲積神經網絡(Convolutional Neural Network, CNN)，從聲學頻譜自動求取情緒特徵參數，與（2）以雙向編碼變換器（Bidirectional Encoder Representation from Transformers, BERT），求取語音逐字稿的文字語意特徵參數。再將此兩類特徵參數向量融合，以強化系統的情緒狀態偵測效能。最後由正式比賽結果發現，我們的系統的情緒狀態偵測正確率達到 73.11%，在所有隊伍提交中的 20 個結果中，排第三名，不但超越主辦單位提供的基準參考系統(49.75%)，並只差第一名(74.07)不到 1%。

關鍵詞：情感檢測，CNN，BERT, 多模式情感檢測

一、簡介

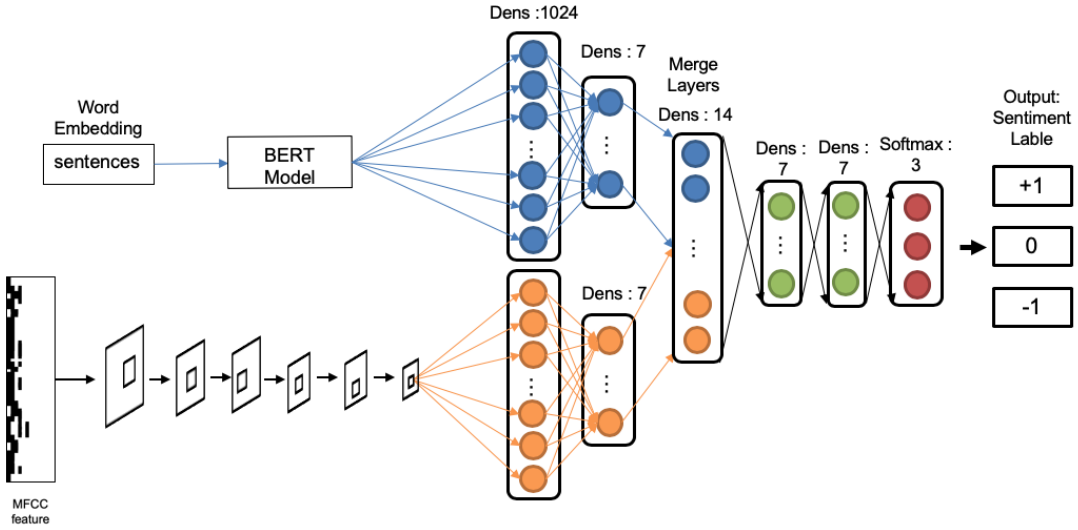
本論文針對 Fearless Steps Challenge 競賽中的 sentiment detection 任務，進行語音情感偵測初步探討。Fearless Steps Challenge 比賽，是為了慶祝登月計劃 50 週年所舉辦的大規模競賽。由德州賽拉達分校將登月任務中所有通訊對話數位化，並發行 Fearless Steps Corpus 語料，支援各個競賽項目，提供大量的訓練資料及測試資料，因為此項比賽主要是希望可以使用自然環境當中所錄製的資料庫進行比賽，所以 Fearless Steps Corpus 的語音資料，是真實太空任務中，太空人與任務中心的通訊對話錄音。

我們會選擇參加此項競賽，主要是因為目前大部分可取得的情緒相關語料庫，大都是由演員表演的，且語料都過於完美或是過於乾淨，導致在這些語料庫上所獲得的結果，不見得可以反應實際運用時的情境。而 Fearless Steps Challenge 的語音資料，是真實太空任務中，太空人與任務中心的通訊對話錄音，因此會有許多自然的雜訊和對話。最重要的是，此 Fearless Steps Corpus 語料庫總共包含 100 小時的語料，而且情緒標籤都是經由人工標註驗證，因此研究獲得的結果會更加有公信力。

傳統上針對語音情緒偵測，通常專注於先提取情緒的低階聲學特徵[3-14]。一些廣泛使用的頻譜特徵是 Mel-Frequency Cepstral Coefficients（MFCC）[1]、線性預測倒譜係數或是音高軌跡。然後再用高斯混合模型、支持向量機或是馬爾可夫模型進行情緒辨認。而若想從語意來求取情緒特徵參數，則需要先有語音辨認器，將語音轉成逐字稿，再以自然語言處理方式，例如以 word-to-vector 求取特徵向量，再以類神經網路進行情緒辨認。然而，人類情感與聲學低階特徵的表現，實際上不見得一致。而若用逐字稿，則通常會有語音辨認錯誤，影響最終判斷的情形。

針對 Fearless Steps Challenge 比賽，我們在進行初步實驗測試時，發現若單獨只用聲音製作模型，或是單獨使用文字訓練模型，所得到效果都有所不足。主要是語音中的情緒特徵，可能同時表現在音色、語氣或是文字用語上。因此在比賽當中，我們除了分別嚐試對於聲音和逐字稿抽取其隱含的情緒相關特徵，並希望以多模式神經網路，將兩者的特徵參數進行結合，同時以聲音中與逐字稿中的情緒特徵來建立模型，以提升情緒偵測的正確率。

因此，本論文提出如圖一的多模式情緒偵測模型。主要想法是同時考慮語音訊號中包含的聲學與語意資訊，提出基於深度類神經網路之多模式語音情緒偵測模型，用以偵測語音訊號中傳達的情緒狀態。實際做法包括（1）利用捲積神經網絡（Convolutional Neural Network, CNN），從聲學頻譜自動求取情緒特徵參數，與（2）以雙向編碼變換器（Bidirectional Encoder Representation from Transformers, BERT），求取語音逐字稿的語意特徵參數。再將此兩類特徵參數向量融合，以強化系統的情緒狀態偵測效能。圖一為我們進行情緒偵測的框架結構：



圖一、多模式神經網路模型架構圖

此系統的運作包含三個模組，包括：（1）我們將原始語音信號轉換為類似圖像的頻譜圖方式，作為 CNN 的輸入[2]。因此，可以使用以大量語音語料預訓練的深度 CNN 模型進行學習，擷取高級聲學情緒特徵。（2）對於逐字稿的多個連續段落，可以用以大規模文字數據集預訓練的 BERT 模型進行訓練，萃練高階的語意情緒特徵。（3）由 2D-CNN 和 BERT 學習的聲學和語意情緒特徵參數，被集成在多模式的融合網絡中。最後，我們採用多模式的最後一個隱藏層的輸出作為分段的情感標籤。

二、Fearless Steps Challenge

(一)、數據集

為了評估本文所提出的模型性能，我們使用 Fearless Steps Challenge 所提供的美國宇航局阿波羅計劃的全程無線電通訊錄音資料庫，共有 100 個小時，包括火箭升空約佔

25 小時，登月約 50 小時，月球漫步約 25 小時。此外由於任務的不同，語料庫的語音活動密度在整個任務中常常變化，且語音數據的質量也常在 0 到 20dB (Signal-to-noise ratio , SNR)之間變化。Fearless Steps Challenge 為了確保能將數據公平地分配到的訓練，評估和開發子集中，根據噪聲水平與活動密度，對數據進行分類。

Fearless Steps Challenge 所提供的訓練子集，皆經人工轉寫逐字稿與標記情緒標籤。評估子集則只提供自動產生的逐字稿與情緒標籤。但測試子集則無提供任何情緒標籤也無逐字稿，因此本論文在測試資料時，需要對音檔先進行文字轉寫處理。由於 Fearless Steps Challenge 所提供得是太空中不同場景的語音記錄，總共提供了五個不同部分的頻道場景，Flight Director (FD)、Mission Operations Control Room (MOCR)、Guidance Navigation and Control (GNC)、Network Controller (NTWK)、Electrical, Environmental and Consumables Manager (EECOM)，表一提供不同事件的五個場景的時間分布表。

表一、Total Speech Durations per Channel and Event

	ECOM	**FD**	**GNC**	**MOCR**	**NTWK**	**Total**
Lift Off	2.1	1.2	1.3	0.8	3.9	9.3
Lunar Landing	3.7	1.3	4.0	0.9	4.4	14.3
Lunar Walking	3.9	1.1	3.0	1.4	2.8	12.2
Total	9.7	3.6	8.3	3.1	11.1	35.8

為了確保 Fearless Steps Challenge 數據公平性，在訓練資料和測試資料中由 Fearless Steps Challenge 來挑選 SNR 較為公平的資料，並根據靜音持續時間和語音持續時間來進行進一步挑選。表二為分別五個不同場景的 SNR 平均值和 SNR 的標準差。表上有分別五種不同的錄音場地，分別不同場景有分別不同的 SNR 標準差，其中 Mission Operations Control Room (MOCR)的標準差最高，但 Mission Operations Control Room (MOCR)在其中 SNR 平均值為最低，在這個錄音場景下的噪音動態範圍較為浮動。

表二、Signal to Noise Ratio Statistics (dB SNR) per channel for Dev Data

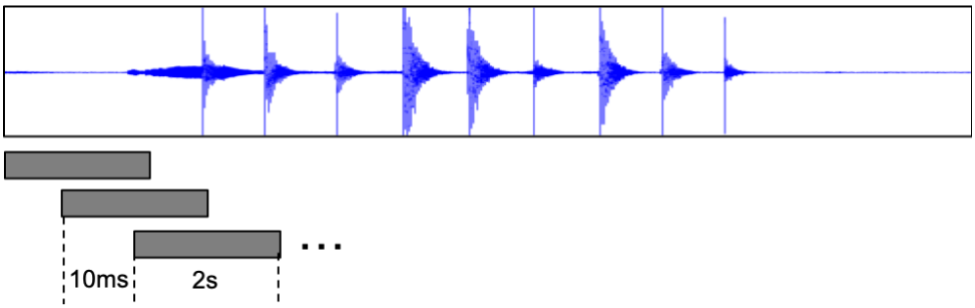
	ECOM	**FD**	**GNC**	**MOCR**	**NTWK**
SNR (Mean)	13.32	14.67	14.91	5.07	10.68
SNR (Std. Dev)	7.40	10.51	11.96	12.60	11.17

在訓練的過程中，由 Fearless Steps Challenge 來做 SNR 資料分群，Fearless Steps Challenge 在依據各個音檔的 SNR(Std Dev)數值去做消除雜訊的動作。

三、基於多模式之情緒檢測系統

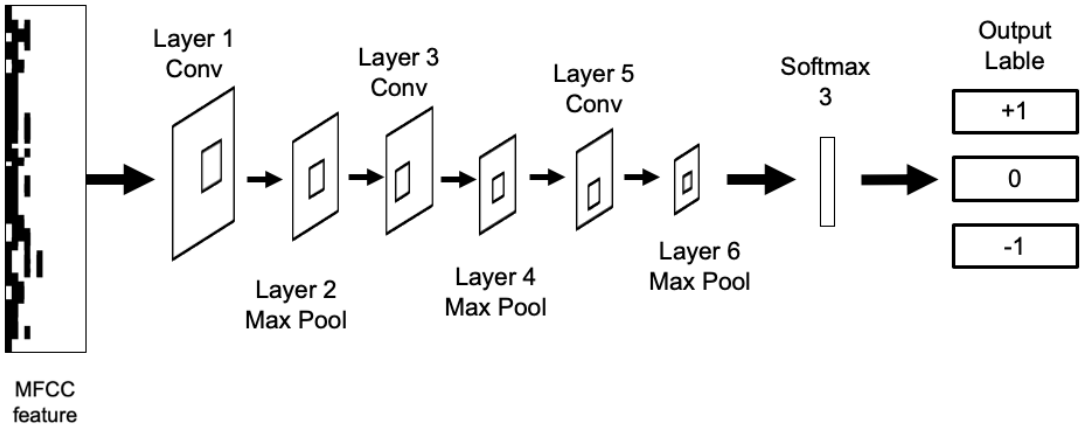
(一)、卷積神經網路聲學情緒模型架構

本篇論文提出的方法的第一階段，先對輸入語音信號執行取音框與求取語音信號的頻譜。其中我們使用 2 秒的窗口大小與 10ms 的音框位移，來獲得足夠可訓練資料。然後將訊號轉換至頻域，在此處以 Mel-frequency 三角形濾波器組過濾頻譜，轉成 Mel-frequency filterbank 參數，再獲得最終的 MFCCs。在本文中，我們還使用 20 個濾波器組和 40-MFCC 進行特徵提取，再將 MFCCs 矩陣輸入 CNN 中。



圖二、Sliding Windows for Sentiment Detection 示意圖

在我們的例子中，CNN 扮演一個從語音訊號頻譜中，提取聲學情緒特徵參數的重要作用。由圖三中可以看出，本論文將 MFCCs 作為 2D 放是作爲輸入，輸入緊接六層 CNN 的基本層數，如圖三所示，CNN 具有[INPUT-CONV-RELU-POOL-CONV-RELUPOOL]的基本架構。CNN 輸入的大小為 40 * 32，為了盡可能保留 Fearless Steps Challenge 提供的信息，我們為每個情緒資料利用 Sliding Windows 窗口採樣訓練數據。最後，將完整的 CNN 架構音頻識別的部分加入混合神經網路模型架構。

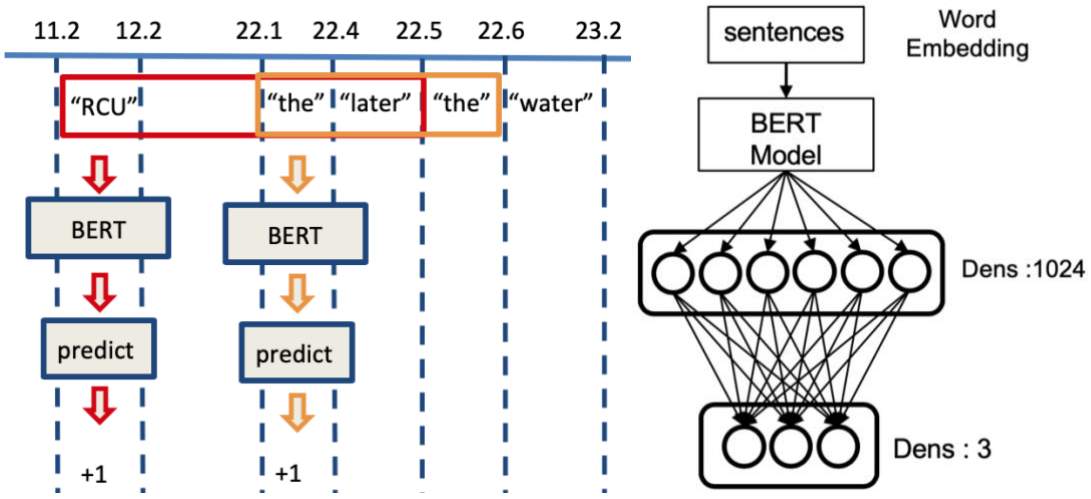


圖三、CNN Architecture for Sentiment Detection

(二)、BERT 神經網路語意情緒模型架構

我們使用 Google 的 BERT 模型。輸入的部分是情緒句子的句向量$[v1, v2, v3 \ldots]$，BERT 與其他模型不同的是，採用了一種簡單的方法，即隨機屏蔽（masking）部分輸入 token，然後只預測那些被屏蔽的 token。將這個過程稱為(masked LM, MLM)，他在訓練雙向語言模型時把少量的詞彙替換成 Mask。

本論文為了輸入較多跟情緒相依性的特徵，一樣採用 Sliding Windows 的方式對文字進行每幀的數據採樣，作法如圖四所示，在當前單字往前取 12 個單詞往後曲 12 個單詞總共 25 個單詞作為 BERT 句子的輸入，對句子做單一個詞的位移來取得下一句，得到句子之後對句子做 Word Embedding 輸入進 BERT Model 如圖四，最後串接 Dens 層連接 Softmax 進行分類。



圖四、BERT 輸入文字採樣示意圖及 BERT 串接 LSTM 示意圖

(三)、混合神經網路情緒模型架構

在長時訓練及辨識的文字和語音由 Fearless Steps Challenge 所提供的美國宇航局阿波羅計劃的全程無線電資料庫裡，我們使用混合模型將語音及文字進行同步訓練。

特徵級融合是最常見和直接的方式，其中所有提取的特徵直接連接成單個高維特徵向量。然後，可以用這種高維特徵向量訓練單個分類器用於情緒識別。大量先前的作品[15-19]證明了情感識別任務中特徵級融合的表現。但是，因為它以直接的方式合併音頻和文字特徵，所以特徵級融合不能模擬複雜的關係。 具體地，每個輸入模態用情緒分類器獨立建模，然後將這些識別結果與某些代數規則組合，例如：”Max”、”Min”、”Sum”等。因此，在情感識別中採用了決策融合。然而，決策層融合無法捕捉不同模態之間的相互關聯，因為這些模態被認為是獨立的。因此，決策級融合不符合人類情緒特徵的特性。

模型級融合作為特徵級融合和決策級融之間的折衷，也被用於情感識別最佳解決方法。該方法旨在獲得音頻和文字模態的聯合特徵表示。其實現主要取決於所使用的融合模型。例如，[4]採用(Hidden Markov Model , MFHMM)來實現模型級融合。 [8]採用誤差半耦合馬爾可夫模型融合以進行情感識別。對於神經網絡，通過首先連接對應於多個輸入模態的神經網絡的不同隱藏層的特徵表示來執行模型級融合。然後，添加額外的隱藏層以從連接的特徵學習聯合特徵表示。現有的模型級融合方法仍然不能有效地模擬音頻和文字模態之間的高度非線性相關性。

四、情緒偵測分類實驗

(一)訓練與測試語料

長時訓練及辨識的文字和語音由 Fearless Steps Challenge 所提供的美國宇航局阿波羅計劃的全程無線電資料庫，包括 100 個小時。選擇的阿波羅 11 號任務主要分為三個階段：（i）升空、（ii）登月、（iii）月球行走。為任務系統開發提供了 80 小時的音頻。在這 80 個小時內，提供了 20 小時的經過人工驗證的答案。對於剩餘的 60 小時音頻，提供 Baseline 系統生成的輸出答案，另外一組 20 小時將發布用於開放測試。

表三、 Fearless Steps Challenge 資料統計與比較

	Fearless Steps Challenge					
	Train			Dev		
	NEUTRAL	POSITIVE	NEGATIVE	NEUTRAL	POSITIVE	NEGATIVE
Avger time	0:00:30	0:00:13	0:00:24	0:00:02	0:00:01	0:00:02
Max time	0:11:10	0:09:22	0:11:10	0:09:03	0:00:16	0:09:03
Min time	0:00:0.4	0:00:0.11	0:00:0.4	0:00:0.17	0:00:0.14	0:00:0.17
#Total time	14:42:44	4:58:34	4:38:38	2:56:00	0:42:30	0:23:03
Count	1724	1327	685	4646	1912	492

(二)評估指標

Fearless Steps Challenge 比賽規則如下，音檔實際判斷正確時間長度單位為 10ms，在參考答案當中只有偵測到和參考答案範圍內一樣才給予得分如圖五，若判斷超出參考得分範圍則不扣分也不予計分只計算真實得分數，每個得分區域將計算每 10ms 幀的真實相同答案的數值（標籤上的最低分辨率）。

圖五、得分範圍參考圖

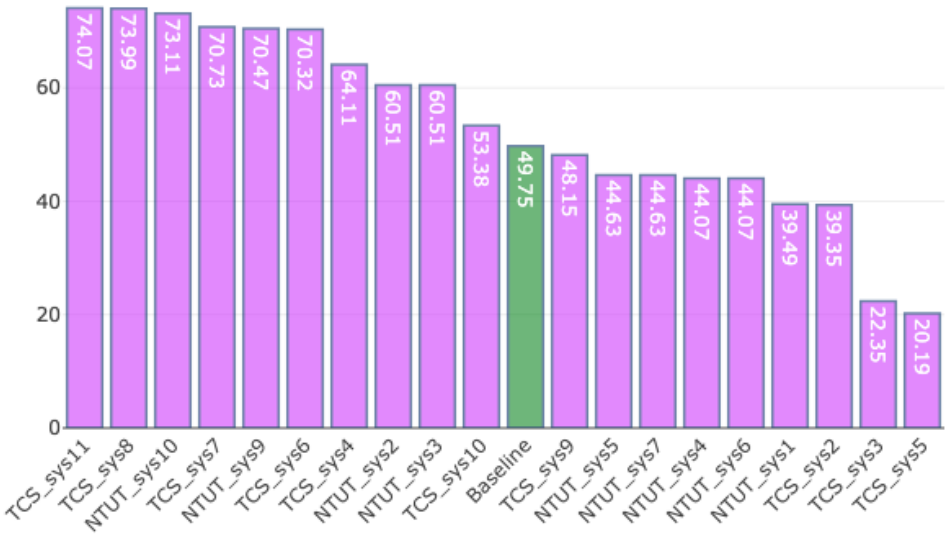
評分公式如下，$tatal\ TP\ time$ 為 System Detected 的真實得分的總時間和，$annotated\ total\ speech\ time$ 為 Reference annotation 的參考答案時間總和，$tatal\ TP\ time$ 除以 $annotated\ total\ speech\ time$再乘以百分比為最後，Fearless Steps Challenge 比賽排名的參考依據。

$$Acc_{sent} = \frac{total\ TP\ time}{annotated\ total\ speech\ time}$$

五、實驗語競賽結果

本實驗使用 Fearless Steps Challenge 資料庫行訓練，Fearless Steps Challenge 資料庫分成了 Train data。本實驗將 Train data 資料庫的每份音檔利用 Sliding Windows 的方式切出訓練資料，窗口大小為 2s 每次位移 10ms 進行 Train data 的資料採樣。以下先單獨對各個部分進行實驗，分為 CNN 架構的音頻部分和 BERT 文字部分別進行討論，再討論多模式情緒偵測模型。

此外，圖六為我們提交至 Fearless Steps Challenge 官方，經過官方評測後的成績排名結果，我們總共提交了 10 個不同設定的系統，在以下實驗中會逐步說明。



圖六、Fearless Steps Challenge 官方排名總表

實驗一，聲學與文字模式情緒偵測

1. 聲學 CNN 模型

多模式模型音頻前及處理部分單獨進行討論， Fearless Steps Challenge 的答案共分為三種 positive、neutral、negative，而在評估指標內還有包含 Non-Sentiment 的部分，因此在訓練同時將測試集 Non-Sentiment 的部分使用 Sliding Windows 進行數據採集。

在音頻測試中可以看到，因資料庫音檔雜訊過多且在大部分音檔當中的對話情緒起伏並不明顯，所以造成 positive、negative 的準確率偏低，但在 neutral、silence 的部分以圖七混淆矩陣來看 silence 的準確率最高，因此在單音頻測試模型下成效較為顯著，但

在 neutral 的判斷還是有部分些許不準確。

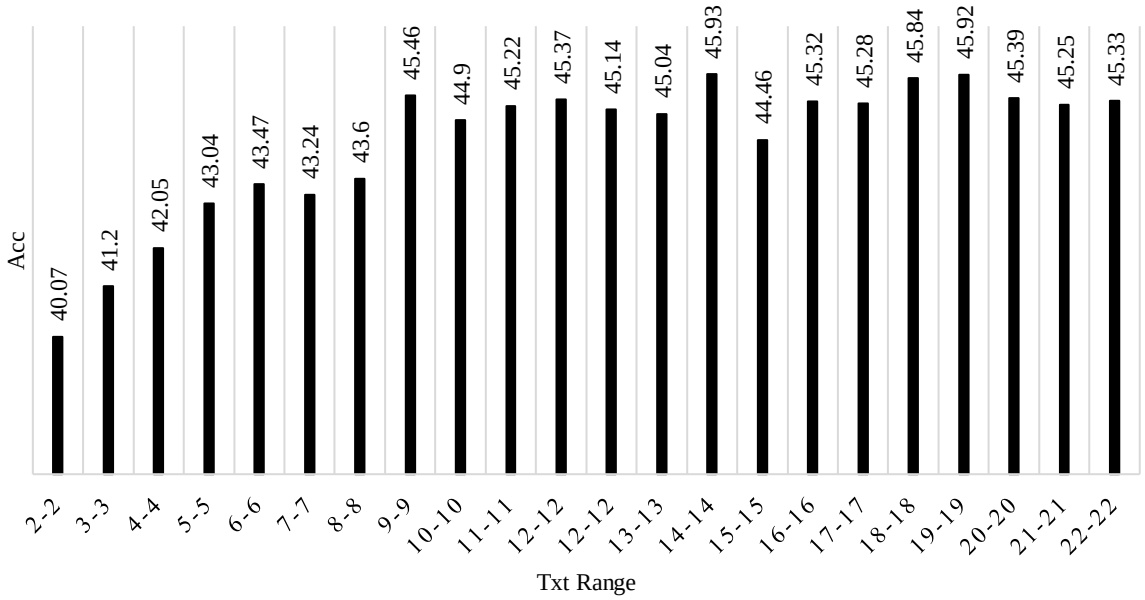
本論文將音頻單獨使用 CNN 神經網路模型進行單獨資料庫訓練，模型如圖三，在 Fearless Steps Challenge 官方網站準確率為 44.07%參考排名如表五，因此單音頻測試對於 silence 和 neutral 偵測有一定的準確度，但離最高準確率還是需要靠文字的輔助下達成。

2. 文字 BERT

由於 Fearless Steps Challenge 測試集並無提供音頻的文字，因此在使用測試集辨識時將音頻使用語音辨識(Automatic Speech Recognition, ASR)進行辨識，但語音辨識只能獲得單詞起始時間和結束時間，所以在文字預處理本論文使用 Sliding Windows 方式，在要辨識單詞時間底下往前往後取一定範圍的單詞量組成句向量，輸入如圖四所示 BERT 模型內進行單文字測試。

在測試文字中發現，文字採樣範圍不同時會有不同準確率，當採樣文字採樣範圍達到往前往後 14 個字時之後準確率趨近於穩定，如表四所示，在採樣範圍從 2 至 8 個字時明顯採樣特徵不足因此造成準確率沒有明顯提升，因此將採樣範圍提升從 9 至 22 個字進行測試，在 8 至 9 個字時準確率有明顯提升，由此實驗可證實當文字採樣範圍會對於情緒識別準確率有一定的成效。

表四、BERT 模型各種文字採樣範圍正確率



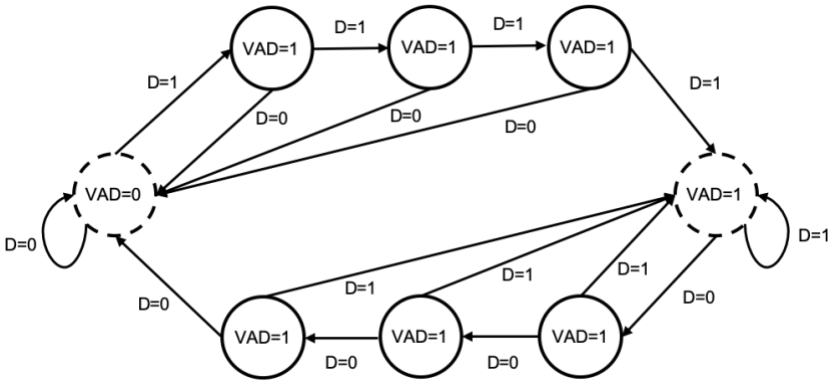
本論文將文字單獨使用 BERT 神經網路模型進行單獨資料庫訓練，模型如圖四，在 Fearless Steps Challenge 官方網站準確率為 44.63%參考排名如表三，因此使用單文字識別情緒時如果有相關情緒字眼出現時則會對準確率照成一定的影響，但某些場景下無法純粹依靠單文字測試，因此本論文使用多模式神經網路模型將兩者模型混合。

實驗二，多模式情緒偵測

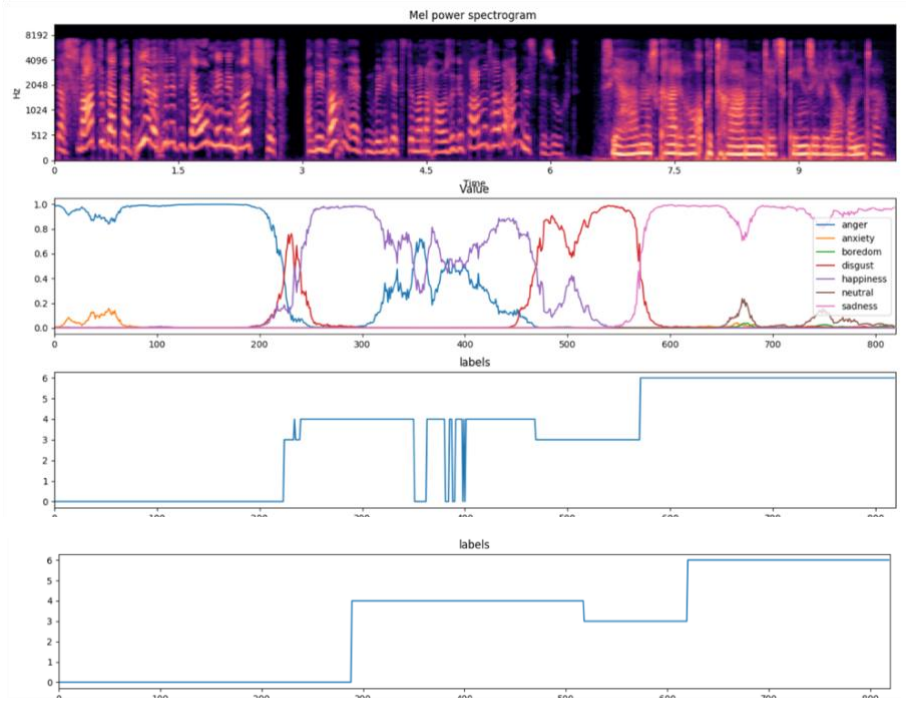
在多模式實驗當中本論文使用音頻模型和文字模型進行情緒識別，交叉測試發現音頻測試當中發現 positive 和 negative 的情緒類別較為不準確，文字測試部分也發現文字採樣範圍對於實驗結果有一定影響，音頻無法識別的 positive 和 negative 利用多模式模型，由文字模型來識別 positive 和 negative 的相關情緒字眼，以利於提升模型準確率，音頻測試當中 silence 和 neutral 的準確度也有一定成效，因此也可輔助多模型識別 Non-Sentiment 的切割位置準確度和 neutral 的正確率，所以本論文使用混合神經網路模型架構來提升模型準確度。

資料庫分為三類 negative，neutral，positive，進行這三類的辨別。因情緒變化動態較慢而我們所使用的 Sliding Windows 的辨識方式讓結果輸出的變化太大，所以在連續辨認時設置了 state machine 的輸出機制，在連續輸出一定數量的答案才會確定輸出否則會繼續輸出前一個答案。例如：圖八在未使用 state machine 時輸出答案不穩定會一直跳動，但在加上 state machine 後可以看到答案輸出趨近於穩定。

state machine 狀態圖如圖七，預設輸出為 0 當 D 連續輸出三次轉態為 1 時 VAD 才會判斷輸出為 1，當D轉態出現中斷或是小於 3 次時回道原始狀態的 VAD 值，反之則將狀態轉為轉態數值。也就是說，當輸出第二次出現不一樣的數值時先放入暫存器，然而繼續輸出相同數值，直到連續得到相同轉態數值，才確定轉態。這可以使本論文模型輸出趨近於穩定。



圖七、state machine 狀態圖



圖八、state machine 前後比較

在 state machine 的幫助下，本實驗使用圖五的多模式神經網路架構，將文字以及聲音使用 Sliding Windows 的方式切出訓練資料，窗口大小為 2s 每次位移 10ms 進行訓練。在 128 訓練次數後，正確率達到 60.51%，

在第二次實驗下，修改 state machine 的暫存器個數來讓情緒浮動的範圍不會變化的太快，將 state machine 暫存器修改為 15 個最本論文最高正確率，正確率來到了 73.11% 為 Fearless Steps Challenge 比賽中 Sentiment Detection 項目的 Rank 3 排名，

實驗三：提交至 Fearless Steps Challenge 官方之系統差異說明

以下分別說明在 Fearless Steps Challenge 官方網站總排名中，不同 NTUT_sys 系統的做法與設定差異：

1. NTUT_sys1 使用 google 語音辨識將音檔切割為 15 秒一個單位音檔不重疊，進行單音頻測試神經網路模型如圖三，Fearless Steps Challenge 官方正確率為 39.49%
2. NTUT_sys2 使用 google 語音辨識將音檔切割為 15 秒一個單位音檔不重疊，進行多模試神經網路模型如圖一，Fearless Steps Challenge 官方正確率為 60.51%
3. NTUT_sys3 為 NTUT_sys2 的重複卻認正確率，因此在回傳一次給 Fearless Steps Challenge 官方卻認正確率為 60.51%
4. NTUT_sys4 使用 Sliding Windows 的方式進行驗證將音框設為 2s 位移時間為 10ms，進行單音頻測試神經網路模型如圖三，Fearless Steps Challenge 官方正確率為 44.07%
5. NTUT_sys5 使用 Sliding Windows 的方式進行驗證將文字採樣範圍調整為前後 14 個字，進行單文字測試神經網路模型如圖四，Fearless Steps Challenge 官方正確率為 44.63%
6. NTUT_sys6 為 NTUT_ sys4 的重複卻認正確率，因此在回傳一次給 Fearless Steps Challenge 官方卻認正確率為 44.07%
7. NTUT_sys7 為 NTUT_ sys5 的重複卻認正確率，因此在回傳一次給 Fearless Steps Challenge 官方卻認正確率為 44.63%
8. NTUT_sys8 使用 Fearless Steps Challenge Train data 所算出的答案進行回傳，因此不列在官方排名中
9. NTUT_sys9 使用 Sliding Windows 的方式進行驗證，多模式神經網路進行識別，state machine 暫存器設為 3 個，Fearless Steps Challenge 官方正確率為 70.47%
10. NTUT_sys10 使用 Sliding Windows 的方式進行驗證，多模式神經網路進行識別，state machine 暫存器設為 15 個，Fearless Steps Challenge 官方正確率為 73.11%

六、結論

在本論文中，我們提出了基於 CNN 與 BERT 的多模式情緒識別神經網路架構，融合聲學與語意情緒特徵參數，用以偵測語音訊號中傳達的情緒狀態，以強化系統的情緒

狀態偵測效能。並以 state machine 減緩輸出跳動的情況，有效的解決輸出時產生的不穩定性，提升準確度。最後，由正式比賽結果發現，我們的系統的情緒狀態偵測正確率達到 73.11%，在所有隊伍提交中的 20 個結果中，排第三名，不但超越主辦單位提供的基準參考系統（49.75%），並只差第一名（74.07）不到 1%。

Acknowledgements

This work was partly supported by Slovak Research and Development Agency under contract no. APVV SK-TW-2017-0005, APVV-15-0517, APVV-15-0731, partly Cultural and educational grant agency from project KEGA 009TUKE-4/2019 and partly Scientific grant agency by realization of research project VEGA 1/0511/17 both financed by the Ministry of Education, Science, Research and Sport of the Slovak Republic and partly by the Taiwan Ministry of Science and Technology MOST-SRDA contract No. 107-2911-I-027-501, 108-2911-I-027-501, 107-2221-E-027-102, 107-3011-F-027-003 and 108-2221-E-027-067 and partly supported by Telecommunication Laboratories, Chunghwa Telecom, Taoyuan Taiwan contract No. TL-108-D301.

參考文獻

[1]. Y. Wang, L. Guan, An investigation of speech-based human emotion recognition, pp. 15-18, 2004.

[2]. Z. Huang，M。Dong，Q。Mao，Y。Zhan，Speech Emotion Recognition Using CNN，pp.801-804,2014。

[3]. Y. Wang, L. Guan, "Recognizing human emotional state from audiovisual signals", IEEE Trans. Multimedia, vol. 10, no. 5, pp. 936-946, Aug. 2008.

[4]. Z. Zeng, J. Tu, B. M. Pianfetti, T. S. Huang, "Audio–visual affective expression recognition through multistream fused HMM", IEEE Trans. Multimedia, vol. 10, no. 4, pp. 570-577, Jun. 2008.

[5]. M. Mansoorizadeh, N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech", Multimedia Tools Appl., vol. 49, no. 2, pp. 277-297, 2010.

[6]. M. Glodek et al., "Multiple classifier systems for the classification of audio-visual emotional states" in Affective Computing and Intelligent Interaction, Berlin, Germany:Springer, vol. 6975, pp. 359-368, 2011.

[7]. M. Soleymani, M. Pantic, T. Pun, "Multimodal emotion recognition in response to videos",

IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 211-223, Apr./Jun. 2012.

[8]. J.-C. Lin, C.-H. Wu, W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition", IEEE Trans. Multimedia, vol. 14, no. 1, pp. 142-156, Feb. 2012.

[9]. J. Wagner, E. Andre, F. Lingenfelser, J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data", IEEE Trans. Affect. Comput., vol. 2, no. 4, pp. 206-218, Oct. 2011.

[10]. A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification", IEEE Trans. Affect. Comput., vol. 3, no. 2, pp. 184-198, Apr./Jun. 2012.

[11]. D. Gharavian, M. Bejani, M. Sheikhan, "Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks", Multimedia Tools Appl., vol. 76, no. 2, pp. 2331-2352, 2017.

[12]. S. Zhalehpour, O. Onder, Z. Akhtar, C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states", IEEE Trans. Affect. Comput..

[13]. R. R. Sarvestani, R. Boostani, "FF-SKPCCA: Kernel probabilistic canonical correlation analysis", Appl. Intell., vol. 46, no. 2, pp. 438-454, 2017.

[14]. M. Bejani, D. Gharavian, N. M. Charkari, "Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks", Neural Comput. Appl., vol. 24, no. 2, pp. 399-412, 2014.

[15]. Y. Wang, L. Guan, "Recognizing human emotional state from audiovisual signals", IEEE Trans. Multimedia, vol. 10, no. 5, pp. 936-946, Aug. 2008.

[16]. M. Mansoorizadeh, N. M. Charkari, "Multimodal information fusion application to human emotion recognition from face and speech", Multimedia Tools Appl., vol. 49, no. 2, pp. 277-297, 2010.

[17]. Y. Wang, L. Guan, A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition", IEEE Trans. Multimedia, vol. 14, no. 3, pp. 597-607, Jun. 2012.

[18]. B. Schuller, R. Müller, B. Hörnler, A. Höthker, H. Konosu, G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations", Proc. 9th Int. Conf. Multimodal Interfaces (ICMI), pp. 30-37, 2007.

[19]. C. Busso et al., "Analysis of emotion recognition using facial expressions speech and multimodal information", Proc. 6th Int. Conf. Multimodal Interfaces (ICMI), pp. 205-211, 2004.