A Chinese Education Broadcast Emotion Corpus (中文教育廣播情緒語料庫)

Pin-Hsiu Lin, Hou-Chiang Tseng, Kuan-Yu Chen

National Taiwan University of Science and Technology
{torrance_lin, tsenghc, kychen}@mail.ntust.edu.tw

摘要

本研究提出了一個中文教育廣播之情緒語料庫，收錄台北酷課雲裡小學一年級至高中三年級各科目的教學影片、動畫片、幼兒學習影片、趣味科學節目等，共計 3,060 部影片。我們採用 Azure 語音辨識服務，產生影片中語音內容的轉寫文字；使用 Hume 語音情緒辨識服務，為影片內每個句子產生對應的情緒標籤。在實驗中，我們在中文教育廣播之情緒語料庫上進行一系列的實驗，並探討 8 組常用的語音特徵在語音情緒辨識的任務成效。這些實驗成果，為中文教育廣播之情緒語料庫奠定了重要的基礎，並可作為後續相關研究的參考。

Abstract

This study presents a Chinese Education Broadcast Emotion Corpus (CEBE), which includes a total of 3,060 videos from the Taipei City CooC Cloud platform, ranging from first grade in elementary school to twelfth grade in high school. The videos encompass teaching videos, animated clips, early childhood learning videos, and science programs. We utilized the Azure Speech Recognition service to generate transcriptions of the speech content in the videos and employed the Hume Speech Emotion Recognition service to assign corresponding emotion labels to each sentence in the videos. In our experiments, we conducted a series of tests on the CEBE corpus, exploring the effectiveness of eight commonly used speech features in speech emotion recognition. These experimental results establish an important foundation for the CEBE corpus and serve as a reference for future related research.

關鍵字：中文教育廣播、情緒語料庫、語音情緒辨識

Keywords: Chinese Education Broadcast, Speech Emotion Corpus, Speech Emotion Recognition

1 緒論

情緒辨識技術是人工智慧中人機互動領域裡一個重要的研究方向。傳統上，情緒是人類交流中不可或缺的一部分，我們會透過語言、表情和肢體動作表達和感知彼此的情緒。而隨著科技的日新月異，人工智慧不僅能夠理解和回應人類語言，還能夠識別和解讀人類情緒。它能透過分析臉部表情、語音語調、生理訊號等多種方式，將抽象的情緒轉化為可量化的數據。這一技術的發展帶來了眾多潛在應用，從智能客服和個人助理到心理健康監測和市場分析，都能看到情緒辨識的身影。

聲音中所蘊含的音調、音量、節奏等各種特徵，是情緒辨識中重要的資訊之一，可以用來識別說話者的情緒狀態（R. A. Calvo et al., 2010）。在對話的交流中，通常被區分為 Primary Channel 和 Secondary Channel（S. Casale et al., 2008）。前者與語法（Syntactic）、語意（Semantic）相關，傳達語言訊息；後者傳達副語言訊息，如語調（Tone）、情緒狀態（Emotional State）及手勢等。通過副語言訊息，可以更加的理解說話者當下的狀態，避免有誤解、判斷錯誤的情況發生。

隨著深度學習的演進，語音情緒辨識領域取得了顯著進展。然而，語音情緒辨識的訓練仍面臨諸多挑戰。除了資料集難以收集和需要大量人力時間進行標註外，不同文化、背景和個體之間的語音情緒表達方式存在差異。此外，如何保證語音情緒辨識系統的實時性和可靠性也是關鍵挑戰。

本研究建立了一個中文教育廣播情緒語料庫，涵蓋台北酷課雲的小學至高中教學影片、動畫、幼兒學習影片等，共收錄 3,060 部影片。我們使用 Azure 語音辨識服務轉寫影片語音內容（Microsoft, n.d.），並透過 Hume 語音情緒辨識服務為每個句子標記情緒（Hume AI, n.d.）。此外，我們更探討了 8 組常見的語音特徵在語音情緒辨識中的效果。這些實驗成果，為中文教育廣播之情緒語料庫奠定了重要的研究基礎，並可作為後續相關研究的參考。

2 文獻

在語音情緒辨識（Speech Emotion Recognition, SER）領域，過去的研究主要集中於情緒分類的問題。這些研究通常採用各種機器學習和深度學習方法，利用聲學特徵來對語音樣本進行情緒分類。

早期的研究多使用 LSTM 或 CNN 作為深度學習的架構，後期也出現了兩者一起使用的模型。如 Jianfeng Zhao 等人（2019）構建了兩個網路模型：一個是一維 CNN-LSTM，用於從原始音頻片段中學習特徵；另一個是二維 CNN-LSTM，用於從對數梅爾頻譜（Log-mel Spectrogram）中學習特徵。該實驗使用 Berlin Emotional Database（Burkhardt et al., 2005）和 IEMOCAP（Busso, C. et al., 2008）這兩種資料集做驗證。在 Speaker-dependent 的條件下，一維 CNN-LSTM 和二維 CNN-LSTM 分別在 Berlin Emotional Database 達到了 92.34% 和 95.33% 的平均準確率；而在 IEMOCAP 上一維 CNN-LSTM 和二維 CNN- LSTM 則分別達到了 67.92% 和 89.16% 的平均準確率。研究結果表明，結合 CNN 和 LSTM 的深度學習網路能夠從語音數據中學習到豐富的情緒特徵，並顯著提高語音情緒辨識的準確性。

隨著注意力機制（Attention Mechanism）的崛起，有效的解決了 LSTM 或 CNN 因為應付較長的序列導致微分時梯度消失、梯度爆炸的問題，也因為可平行運算的模型設計，大大的減緩計算時間隨序列長度成正比增加的問題。Y. Wang 等人（2021）使用了在語音辨識（Automatic Speech Recognition, ASR）領域中性能最優秀的自監督式預訓練模型 Wav2Vec 2.0 和 HuBERT，探討模型的部分微調（僅調整 Transformer 層）及全數微調（同時調整 CNN 和 Transformer 層），在 SER、Speaker Verification（SV）和 Spoken Language Understanding（SLU）任務的成效。在語音情緒辨識的研究中，該論文以 IEMOCAP 資料集進行實驗。結果顯示微調模型顯著地優於未微調模型；而部分微調的 HuBERT 大型模型表現最佳，表明這種方法在語音情緒辨識任務中特別有效。

相較於較制式的情緒種類，Russell, James 和 Mehrabian, Albert（1977）提出了 AVD 模型，它是一種捕捉人類情緒體驗豐富性的強大工具。通過將情緒分解為這三種維度：Arousal（激活度，從平靜到興奮，描述情緒的強度或激烈程度）、Valence（愉悅度，從不愉快到愉快，描述情緒的正面或負面程度）及 Dominance（支配度，從被動到主動，描述感受到的控制或影響力程度）。研究人員和實踐者可以更深入地了解情緒的性質和動態，從而促進更好的情緒理解和調節。

此外，傳統的方法依賴於靜態描述（例如統計函數或通用背景模型），但這些方法無法有效地捕捉語音表達中的動態時間變化。近期的深度學習方法可以直接提取句子級別中對應的特徵，但由於每個語音語句的長度不同，因此常見的方式會先將語句進行裁剪，或在最後填補零的方式，將所有語句的長度對齊，但這些方式都可能會損失有用的資訊。有鑒於此，W. -C. Lin and C. Busso（2023）提出了一種新的動態區塊分割法，以事先定義區塊個數的方式，動態地將每個語音語句切割成固定個數的區塊。爾後，將每個區塊分別進行特徵的擷取，最後再進行整合。透過這樣的技術，期望可以完整地、不破壞語音內容完整性地進行語音資訊的特徵擷取，進而提升情緒辨識的效能。

3 中文教育廣播之情緒語料庫

我們提出了一個中文教育廣播情緒語料庫（A Chinese Education Broadcast Emotion Corpus, CEBE），這是針對教育領域專門設計的語料庫，旨在促進情緒辨識與教育科技的結合。我們統合了台北酷課雲小學一年級至高中三年級各科目教學影片、動畫片、幼兒學習影片、趣味科學節目等，共計 3,060 部的影片，涵蓋了廣泛的教育資源和年齡層。透過分析教育影片中的情緒表達，研究者可以探索教

Tag	Amount	Tag	Amount	Tag	Amount
Interest (O)	122,500	Sadness (S)	1,354	Negative Surprise (U)	97
Concentration (O)	73,093	Sympathy (O)	955	Horror (F)	90
Calmness (N)	38,640	Disgust (D)	561	Love (O)	77
Determination (O)	22,210	Positive Surprise (U)	534	Adoration (O)	44
Excitement (H)	13,830	Awkwardness (O)	526	Embarrassment (O)	33
Contemplation (O)	8,776	Satisfaction (H)	522	Relief (O)	20
Joy (H)	8,554	Tiredness (O)	470	Awe (O)	17
Amusement (H)	5,557	Nostalgia (O)	400	Desire (O)	11
Realization (O)	5,231	Pride (O)	315	Triumph (H)	4
Admiration (O)	5,163	Anxiety (F)	280	Guilt (O)	3
Anger (A)	4,106	Craving (O)	267	Shame (O)	3
Confusion (O)	3,710	Pain (S)	215	Envy (O)	2
Doubt (O)	3,634	Empathic Pain (S)	199	Ecstasy (H)	1
Distress (S)	2,966	Boredom (O)	180	Contentment (H)	0
Aesthetic Appreciation (O)	1,870	Contempt (C)	168	Entrancement (O)	0
Disappointment (S)	1,511	Fear (F)	145	Romance (O)	0

表 1. CEBE 資料集情緒種類和數量表，按照數量由大到小排序。由於類別數量繁多，因此我們根據語意相似性，重新將情緒類別歸納為 Angry（A）、Sad（S）、Happy（H）、Surprise（U）、Fear（F）、Disgust（D）、Contempt（C）、Neutral（N）、Other（O）與 No agreement（X），共十種。

師的情緒如何影響學生的學習效果，以及學生在不同情境下的情緒反應，抑或是學生在進行翻轉教育時，在台上的表達能力是否良好。此外，CEBE 可用於訓練情緒辨識模型，這些模型可以應用於各種教育科技中，例如智能教室、虛擬教練、情緒回饋系統等。

在資料標註方面，我們首先使用 Azure 取得這些影片的逐字稿，精確地將影片中的語音內容轉換為文本。接著，我們使用 Hume Python SDK 對每部影片中每段句子的音檔進行情緒分析，產生 48 種情緒維度。這些情緒維度代表了人們根據語音變化能夠區分的情緒含義，它是根據模型對韻律的分析來判斷的機率。換句話說，當某個音檔的「開心」維度得分最高時，這表明該音檔最有可能被人們解讀為表達開心的情緒。因此，我們選擇最高分的情緒維度作為該句子的情感標記結果。最後將文本、分數及最終結果合併成一個資料表。

為證明 Hume Python SDK 標註的可靠性，Alan S. Cowen 和 Dacher Keltner（2021）的研究使用大量自然資料和機器學習技術來自動進行情感標記，成功識別多達 34 種情感。透過 fMRI 驗證，發現這些自動標記的情感與大腦的神經活動模式高度一致，且能跨文化、跨模式準確預測情感反應。結果顯示，機器學習自動標記能捕捉到比傳統模型更複雜的情感表徵，證明其在情感研究中的有效性。

此外，我們進行了嚴格的資料清理以確保語料庫的品質和一致性，包括移除重複的影片、長度低於兩秒的語句、與資料庫不符或不存在的影片等。最終，我們將語料庫切分為 328,844 個句子，總長約 372 小時又 10 分鐘。CEBE 資料集所包含的情緒種類和數量如表 1 所示。由於中文教育廣播情緒語料庫涵蓋了豐富的情感表達形式，並提供大量、充足的資料量，將能促進語音情緒辨識模型的研究與發展。

以 emotionResult_100_0_0_0.wav 為例，該音檔是來自編號 100 部影片第 0 大段第 0 小段中的第一句，語者說的話為”台北市線上教學影片”。該音檔經過 Hume Python SDK 得到的 48 種情緒維度中由最高的前三名依次為 Calmness, Boredom 和 Concentration，分數分別為 0.5396, 0.3309 及 0.3285。我們選擇了得分最高了 Calmness 作為此音檔的標註結果。

Layer	Channels	Kernel	Stride	Dimension	Activation
Input	N/A	N/A	N/A	$m \times d$	N/A
Permute	N/A	N/A	N/A	$d \times m$	N/A
CNN-block	128	(1,3)	1	depends	ReLU
CNN-block	64	(1,3)	1	depends	ReLU
CNN-block	32	(1,3)	1	depends	ReLU
Flatten	N/A	N/A	N/A	depends	N/A
Linear	N/A	N/A	N/A	$1 \times b$	ReLU

表 2. 動態區塊分割模塊的模型架構與參數。

4 實驗設定

4.1 聲學特徵

openSMILE 工具（F. Eyben et al., 2010）是一套經常被使用於聲學特徵提取的工具，並且他提供許多公用的語音特徵集。在本研究中，我們使用 openSMILE 工具進行語音特徵的抽取，並採用 INTERSPEECH 國際會議在 2009~2013 年，每年所舉辦的各項語音內涵比賽之特徵集 [1]（Björn Schuller et al., 2009, Schuller et. al., 2010, Björn Schuller et al., 2011, Björn Schuller et al., 2012, B. Schuller et al., 2013），以及常見的語音情緒辨識特徵集 emobase、emoLarge 和 emobase2010 等共八種特徵集，為每個語音語句進行語音特徵的抽取。

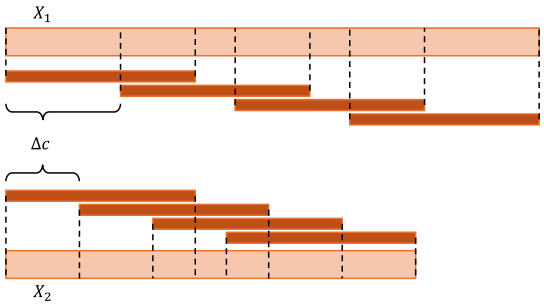

圖 1. X_1與X_2為兩個不同長度的語音語句。當我們將區塊個數參數C設定為5，且固定區塊長度的參數w後，語音語句X_1與X_2都被切割成C個相同長度的區塊。切割時，因為X_1較長，所以X_1的移動時間長度Δc較X_2的長。

4.2 語音情緒辨識模型

由於先前的研究指出，動態區塊分割法可以有效地解決語音語句長度不一的問題，且在語音情緒辨識的任務中能夠取得良好的任務成效（W. -C. Lin and C. Busso, 2023）。因此，本研究以此為基礎，建立一套簡單而有效的語音情緒辨識模型。我們的語音情緒辨識模型有四個模塊：特徵提取、動態區塊分割、區塊層級的特徵表示、句子層級的時間聚合。

特徵提取模塊是用來從語音信號中提取幀（Frame）級別的聲學特徵，如頻譜圖特徵（Spectrogram）等的低階特徵。因此，當給定一個語音訊號X後，特徵提取模塊可以將其表示成一系列的低階特徵向量$F = \{f_1, \dots, f_T\}$。

接著，在動態區塊分割模塊裡，則是要將不定長度的低階特徵向量，轉換成固定個數的高階特徵向量。我們首先定義一個區塊個數參數C，與區塊長度的參數w。每個語音語句不論長度，都會被切割成C個長度w為區塊，且切割時的移動時間長度是Δc：

$$\Delta c = \frac{T-w}{C-1} \quad (1)$$

由式子(1)可知，當區塊長度w固定時，若區塊數量C增加，移動時間長度Δc就會縮小，導致區塊之間的重疊增加。切割的示意圖如圖 1 所示。在切過後，我們使用三層的一維卷積神經網路（1D CNN）來處理每一個區塊內的低階特徵向量。最後，為每一個區塊產生一個高階的特徵向量。詳細的模型參數如表 2 所示。因此，通過動態區塊分割模塊，語音訊號X的低階特徵向量$\{f_1, \dots, f_T\}$，將轉換為C個高階特徵向量$H = \{h_1, \dots, h_C\} \in \mathbb{R}^{C \times b}$。這種動態區塊分割方法特別適合處理長度不定的時

[1] https://audeering.github.io/opensmile/get-started.html#default-feature-sets

間序列資料。由於切割過後片段的大小是固定的，這不僅簡化了深度學習模型的設計，也使得模型在處理不同長度的語句時更加一致；其次，這一過程可以並行化處理，從而允許同時計算多個區塊的特徵表示，大幅提升計算效率。

最後，句子層級的時間聚合模塊則是用來總結區塊層級的特徵，有效地統合整個語句的情緒訊息，從而產生最佳的語音情緒辨識結果。我們探討了平均向量法、權重式平均向量法、注意力法與自注意力法等四種模型架構。

- 平均向量法：將所有區塊的高階特徵向量進行平均，得到最終的語句級別表示向量z：

$$z = \frac{1}{C}\sum_{t=1}^{C} h_t \tag{2}$$

- 權重式平均向量法：在平均向量法中，每個區塊的權重皆是一致的。然而，在一段語句中，情緒的展現可能僅出現在某些片段中而已。因此，在整合區塊的向量表示法時，每個區塊應該有不同的權重。我們首先利用一個簡單的前饋神經網路（Feedforward Neural Network）搭配激活函數Sigmoid，為每個區塊進行權重的計算：

$$g_t = \sigma(W^a \cdot h_t + b) \tag{3}$$

其中，W^a為前饋神經網路的參數，$\sigma(\cdot)$表示Sigmoid 函數，b則為偏移量（Bias）。接著，我們將每個區塊的向量表示法乘上對應的權重後相加，作為最後句子層次的向量表示法：

$$z = \sum_{t=1}^{C} g_t h_t \tag{4}$$

- 注意力法：雖然權重式平均向量法已經考慮了每個區塊應有不同的權重，但權重的計算方式卻是每個區塊獨自運算，而沒有考慮全域（也就是所有區塊）的資訊。有鑒於此，參考先前的研究（T. Luong et al., 2015），我們使用注意力法，藉由考慮全域的資訊，為每個區塊計算權重。首先，我們將C個高階特徵向量輸入一層遞迴式神經網路（Recurrent Neural Network, RNN），為每個高階特徵向量轉換為具有相鄰資訊的向量表示法：

$$H^{RNN} = RNN(H) \tag{5}$$

然後，我們利用具有相鄰資訊的向量表示法來為每個區塊計算注意力權重：

$$s_t = (h_t^{RNN})^{\top} W^b h_C^{RNN} \tag{6}$$

$$[\alpha_1, \dots, \alpha_t, \dots \alpha_C] = softmax([s_1, \dots, s_t, \dots s_C]) \tag{7}$$

其中，W^b為一個可訓練的模型參數。藉由注意力權重α_t加權對應的相鄰資訊向量，產生一個統合資訊向量v：

$$v = \sum_{t=1}^{C} \alpha_t \, h_t^{RNN} \tag{8}$$

最後，我們串接統合資訊向量v與最後一個區塊的相鄰資訊向量h_C^{RNN}，通過一個前饋神經網路與激活函數$tanh(\cdot)$來生成最終的句子層級特徵表示：

$$z = tanh\,(W^e[v; h_C^{RNN}]) \tag{9}$$

W^e是前饋神經網路的模型參數。

- 自注意力法是基於現在流行的多頭自注意力（Multi-head Self-attention）結構的模型。他不僅結合了權重式平均向量法與注意力法考慮到為每個區塊賦予不同權重的優點，更進一步地改善在注意力法中，產生相鄰資訊的向量表示法時，僅考慮單向資訊的缺點。此外，自注意力還有可平行化的優點，可以加速整個計算過程。再者，由於多頭的模型設計，可以讓模型自動地從不同面向加權不同區塊的權重。在實際的運算上，我們將C個高階特徵向量送入多頭自注意力模型進行運算：

$$Head_j = selfatt(HW_j^Q, HW_j^K, HW_j^V) \tag{10}$$

其中，$Head_j$表示第j組自注意力運算結果，$selfatt(\cdot,\cdot,\cdot)$是自注意力運算函式，而$W_j^Q$、$W_j^K$和$W_j^V$是自注意力運算中的參數：

MRR	Feature Set							
	IS2009	IS2010	IS2011	IS2012	IS2013	emobase	emoLarge	emobase2010
平均向量法	0.8338	**0.8555**	0.8376	0.8437	0.8526	0.8293	0.8307	0.8267
權重式平均向量法	0.8312	0.8312	0.8378	**0.8478**	0.8321	0.8332	0.8289	0.8468
注意力法	0.8467	0.8325	0.7995	0.8433	0.8451	0.8096	0.8209	**0.8523**
自注意力法	0.8176	0.8375	0.7878	0.8337	0.8013	0.7981	0.7981	**0.8401**

表 3. CEBE 資料集在使用平均倒數排名評估下的實驗結果。

Macro F1-Score	Feature Set							
	IS2009	IS2010	IS2011	IS2012	IS2013	emobase	emoLarge	emobase2010
平均向量法	0.1449	0.1456	0.1473	0.1491	0.1480	0.1327	0.1443	0.0958
權重式平均向量法	0.1455	0.1433	0.1476	0.1491	0.1466	0.1403	0.1451	0.0958
注意力法	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958
自注意力法	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958	0.0958

表 4. CEBE 資料集在使用 Macro Average F1-score 評估下的實驗結果。

Weighted F1-Score	Feature Set							
	IS2009	IS2010	IS2011	IS2012	IS2013	emobase	emoLarge	emobase2010
平均向量法	0.6970	0.6980	0.6994	0.7010	0.7000	0.6847	0.6968	0.6538
權重式平均向量法	0.6972	0.6960	0.6994	0.7007	0.6992	0.6925	0.6962	0.6538
注意力法	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538
自注意力法	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538	0.6538

表 5. CEBE 資料集在使用 Weighted Average F1-score 評估下的實驗結果。

$$selfatt(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (11)$$

d表示向量的維度。我們將J組自注意力運算結果併合後，透過一個簡單的轉換，形成最終的向量表示法$\widetilde{H} \in \mathbb{R}^{C\times b}$：

$$\widetilde{H} = Concat(Head_1, \dots, Head_J)W^O \quad (12)$$

W^O為簡單轉換的參數矩陣。最後，我們將每個向量表示法相加後取平均，作為最後句子層次的向量表示法：

$$z = \frac{1}{C}\sum_{t=1}^{C}\widetilde{H}_t \quad (13)$$

藉由平均向量法、權重式平均向量法、注意力法或自注意力法等四種模型架構，我們可以得到句子層次的向量表示法z。基於這個句子層次的表示法，我們採用兩層的前饋神經網路，進行最終的語音情緒辨識。

4.3 評估指標

在情緒辨識的任務中，我們首先使用平均倒數排名（Mean Reciprocal Rank, MRR），來呈現情緒辨識結果中，正確答案平均出現的位置。此外，我們也使用 Macro Average F1-score（Macro F1-Score）和 Weighted Average F1-score（Weighted F1-Score）作為評估指標。這兩者是常見的分類模型評估方法。Macro F1-Score 是對所有類別的 F1-score 進行平均，每個類別的 F1-score 在計算中具有相同的權重，因此不會因類別不平衡而偏向於較大的類別。Weighted F1-Score 則根據每個類別的樣本數量進行加權平均，更能反映模型在實際應用場景中的總體性能。總結來說，Macro F1-Score 更關注模型在不同類別上的一致性表現，而 Weighted F1-Score 則更關注模型在實際應用中的整體效能。

4.4 訓練與模型參數設置

我們使用 PyTorch 版本 24.05 實現本次的任務。每個情緒指標都視為獨立的任務，個別建立

單獨的模型。CEBE 語料庫長度最長為 11 秒，根據公式將每個音檔切成 11 塊，每一個區塊長度為 1 秒。模型訓練時，使用 Adam Optimizer，Batch 大小設定為 128，訓練週期（Epoch）設定為 100。我們採用 Focal Loss（T. -Y. Lin et al., 2020）作為損失函數，並根據發展集（Development Set）上的損失作為提前停止（Early Stopping）的判斷標準來保存最佳模型。此外，由於 CEBE 資料集情緒類別繁多，但每個類別所包含的數量相當不平均，因此我們將原先的 48 種情緒類別根據語意相似性重新歸類為十種情緒類別，如表 1 所示。

5 實驗結果與分析

使用 CEBE 資料集來進行語音情緒辨識的實驗結果如表 3~5 所示。在實驗中，我們比較了 8 種常見的聲學情緒特徵，並搭配本研究採用的 4 種句子層級之向量表示法模型。藉由比較表 3~5，有許多值得探討的結果。首先，從表 3 來看，不同的句子層級之向量表示法模型搭配不同的語音特徵集，都可以穩定地達到 0.8 左右的平均倒數排名分數。這顯示，正確的情緒類別大概都是預測結果中，機率第一或第二高的。但當我們更仔細地觀察表 3 的結果可以發現，不同的句子層級之向量表示法模型，需要搭配不同的特徵集，才能獲得最佳的結果：平均向量法＋IS2010、權重式平均向量法+IS2012、注意力法+emobase2010、自注意力法+emobase2010。這個結果使我們驚訝地發現，較複雜的特徵集並非一定能帶來較好的辨識結果，反而簡單的特徵集（IS2010）搭配平均向量法，就可以獲得很好的成績；此外，我們也觀察到，注意力法與自注意力法，並沒有在所有情況下都比平均向量法和權重式平均向量法獲得更好的結果。這可能是因為動態區塊分割模塊提供了一個很好的語音切割處理，不僅讓不同長度的語音句子可以有相同數量的區塊，也讓使得後續的模型僅需簡單的架構，就可以獲得良好的任務成效！當我們參考表 4 與表 5 的結果，發現在使用 F1-score 的計算下，平均向量法與權重式平均向量法更凸顯了與注意力法和自注意力法的差異。並且，因為 Recall 值較低的關係，所以 Macro F1-Score 的分數都很低。

6 結論與未來展望

本研究提出一個全新的中文教育廣播情緒語料庫（CEBE），為語音情緒識別領域帶來了新的挑戰與研究方向。雖然從我們的實驗中可以發現，平均倒數排名分數已可獲得不錯的結果，但在 Macro F1-Score 的評分上，結果則不理想。這些結果顯示，CEBE 資料集提供了一個大量且嶄新的平台，並且揭示了教育類語音資料在情緒識別任務中的獨特挑戰。教育語音在情感表達上的特殊性，應被仔細的思考與研究。

在未來，我們將更進一步地檢視、清理與再確認 CEBE 資料集的標籤正確性；探討更多現今卓越的語音情緒模型在 CEBE 資料集上的成效；探究教育語音與其他不同種類的語音內容之情緒表達的差異；著眼於為教育類語音情緒辨識提出一套新穎且有效的模型方法與訓練技術，從而提升教育類語音情緒辨識的準確性和泛化能力。

Acknowledgement

This work was supported by the National Science and Technology Council of Taiwan under Grants NSTC 113-2410-H-011-001, NSTC 112-2628-E-011-008-MY3 and NSTC 113-2640-B-002-005. This project was financially supported by the “Empower Vocational Education Research Center” of the National Taiwan University of Science and Technology (NTUST) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. We thank the National Center for High-performance Computing of the National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

Alan S. Cowen, Dacher Keltner. Semantic Space Theory: A Computational Approach to Emotion. In *Cognitive Sciences, Volume 25, Issue 2*, Pages 124-136. 2021.

Björn Schuller, S. Steidl, and Anton Batliner. The Interspeech 2009 emotion challenge. In *Interspeech (2009)*, ISCA, Brighton, UK. 2009.

Björn Schuller, Anton Batliner, Stefan Steidl, Florian Schiel, Jarek Krajewski. The INTERSPEECH 2011 Speaker State Challenge. In *Proc. INTERSPEECH 2011*, ISCA, Florence, Italy, pp. 3201-3204, 28.-31.08. 2011.

Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, Benjamin Weiss. The INTERSPEECH 2012 Speaker Trait Challenge. In *Proc. INTERSPEECH 2012*, ISCA, Portland, OR, USA, 09.-13.09. 2012.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, Samuel Kim. The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proc. Interspeech 2013*, pp. 148–152. Aug. 2013.

Burkhardt, Felix & Paeschke, Astrid & Rolfes, M. & Sendlmeier, Walter & Weiss, Benjamin. A database of German emotional speech. In *9th European Conference on Speech Communication and Technology*. 5. 1517-1520. 2005.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. In *Journal of Language Resources and Evaluation,* vol. 42, no. 4, pp. 335-359. December 2008.

F. Eyben, M. Wollmer, and B. Schuller. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proc. ACM Int. Conf. Multimedia*, pp. 1459–1462. Oct. 2010.

Galen Andrew and Jianfeng Gao. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40. 2007.

Hume AI. Hume AI. https://www.hume.ai/. Accessed 10 Aug. 2024.

Jianfeng Zhao, Xia Mao, Lijiang Chen. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. In *Biomedical Signal Processing and Control*, Volume 47, Pages 312-323. 2019.

Microsoft. Speech AI Services. Azure. https://azure.microsoft.com/en-us/products/ai-services/ai-speech. Accessed 10 Aug. 2024.

R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. In *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37. 2010.

Russell, James & Mehrabian, Albert. Evidence for a Three-Factor Theory of Emotions. In *Journal of Research in Personality*. 11. 273-294. 1977.

S. Casale, A. Russo, G. Scebba and S. Serrano. Speech Emotion Classification Using Machine Learning Algorithms. In *IEEE International Conference on Semantic Computing*. Santa Clara, CA, 158-165. 2008.

Schuller, Björn & Steidl, Stefan & Batliner, Anton & Burkhardt, Felix & Devillers, Laurence & Müller, Christian & Narayanan, Shrikanth. The INTERSPEECH 2010 paralinguistic challenge. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*. 2794-2797. 2010.

T. Luong, H. Pham, and C. Manning. Effective approaches to attention-based neural machine translation. In *Proc. Conf. Empirical Methods Nat. Lang. Process.*, pp. 1412–1421. Sep. 2015.

T. -Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár. Focal Loss for Dense Object Detection. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318-327. 1 Feb. 2020.

W. -C. Lin and C. Busso. Chunk-Level Speech Emotion Recognition: A General Framework of Sequence-to-One Dynamic Temporal Modeling. In *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215-1227. 1 April-June 2023.

Y. Wang, A. Boumadane, and A. Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*. 2021.