Computational Linguistics and Chinese Language Processing
Vol. 23, No. 1, June 2018, pp. 35-48

整合個人化磁振造影深度神經網路之演算法技術

Joint Modeling of Individual Neural Responses using a Deep Voting Fusion Network for Automatic Emotion Perception Decoding

謝宛庭*、李祈均*

Wan-Ting Hsieh and Chi-Chun Lee

摘要

不論在腦神經科學領域中，抑或是情感計算領域裡，理解聲音情感感知的潛在神經感知機制仍然是一個重要的研究方向。現今許多研究已顯示受試者中表現出的 fMRI 信號的大幅變動是由於個體差異的影響，即受試者間的變異性。然而，相對較少的研究開發用於處理此種特性的自動神經感知解碼任務中的建模技術。本研究中，我們通過學習在融合層應用的調整後的權重矩陣，提出了一種新的深度投票融合神經網絡結構的計算方法。該框架在四級聲音情緒狀態解碼任務中達到了 53.10%的 UAR 準確率，即相對於兩階段 SVM decision score 融合的改善 8.9%。此外我們嘗試音檔與 fMRI 資料的融合，藉由兩者資訊互通使得準確率可提升至 56.07%。我們的框架不僅證明了其處理個體差異的有效性，我們還加入音檔資訊，使得難過情緒分類結果大幅提升。

關鍵詞：個體差異、功能性磁振造影、聲音情緒認知、深度投票融合神經網路

Abstract

In the era underlying grouping life, affective computing and emotion recognition are closely bonding with daily life, and impose great impact on social ability. Understanding the individual differences is significant factor that should not be ignore in fMRI analysis while most of the brain studies on fMRI seldom truly deal

* 國立清華大學電機所
Department of Electrical Engineering, National Tsing Hua University
E-mail: stv7879694@gapp.nthu.edu.tw; cclee@ee.nthu.edu.tw

with it, we carry out a system considering individual variability to recognize the emotion to the vocal stimuli with BOLD signal. In our work, we propose a novel method using multimodal fusion in a voting DNN framework, where we utilize a mask on weight matrix of fusion layer to learn an individual-influenced weight matrix and realize voting in this network, and achieve 53.10% in UAR for a four-class emotion recognition task. Our analysis shows that the multimodal voting net is an effective neural network encoding individual differences and thus enhances the ability to emotion recognition. Further the join of audio feature also boosts the result to 56.07%.

Keywords: Individual Difference, fMRI, Vocal Emotion, Perception, Deep Voting Fusion Neural Net

1. 緒論 (Introduction)

現今功能性磁振造影的技術(functional Magnetic Resonance Imaging, fMRI)使得我們得以利用其擷取的血氧濃度相依對比訊號(Blood oxygen-level dependent, BOLD)觀察大腦中複雜的情緒認知機制(Zhou, Wang, Zou, Zhou & Qian, 2013; Fossati *et al*., 2003)。其中，血氧濃度相依對比訊號中具多樣性，即便是受到相同刺激的受試者也可能產生極大不同的血氧濃度相依對比訊號，而造成此現象的原因十分複雜，但大多與人體間本質上的差異有關，這也說明為何每個人都是一個與眾不同的個體。事實上神經科學相關的研究逐漸重視個體差異的議題，舉例來說，一些過去的研究顯示若僅僅將多個受試者的功能性磁振造影資料平均則會大幅地減少與腦袋結構有關的重要資訊(Van Horn, Grafton & Miller, 2008) (MacDonald, Nyberg, Sandblom, Fischer & Bäckman, 2008) (Kanai & Rees, 2011)；此外，Canli 等人也指出存在大量個體差異且不加以處理的資料也會造成較差的辨識結果以及錯誤的分析(Canli, Sivers, Whitfield, Gotlib & Gabrieli, 2002)。另外在一份延伸的研究中，Hamann 等人證實在腦區在處理情緒上更容易受到個體差異的影響(Hamann & Canli, 2004)。

在神經科學領域中，大多數的研究探討個體差異的方法為最常見的方式是列出每個人的分析結果再畫出相關性(correlation plot)來討論個體差異，例如 Dubois 等人透過在個人層面生成相關圖，驗證科學假設和收集的 fMRI 信號的可靠性(Dubois & Adolphs, 2016)；Parasuraman 等人使用類似的方法來觀察個體差異如何影響工作記憶和決策的認知過程(Parasuraman & Jiang, 2012)。這些方法揭示了考慮個體差異的重要性。然而，在開發用於從 fMRI 數據自動解碼人類情感感知的算法的背景下（例如，Wu, Chen, Liao, Kuo & Lee, 2017; Alba-Ferrara, Hausmann, Mitchell & Weis, 2011; Schirmer & Kotz, 2006)），很少建模技術將個體差異整合至演算法中。

在此研究中，我們提出深度投票融合神經網路(Deep Voting Fusion Network, DVFN)幫助我們直接整合個體差異，以便自動地進行聲音情緒解碼。其中在此架構中，我們引入融合層用以學習個體的重要程度，接著我們將觀察到的個體權重重新置入深度投票融

合神經網路中，並進行微調(finetune)。此實驗中我們招募 18 個受試者，且每個受試者受到 251 句帶有不同情緒的語句音檔刺激，而這些情緒語句乃參考 the USC IEMOCAP 資料庫(Busso *et al.*, 2008)所設計。我們提出的架構使得在對人類 4 類情緒認知的實驗上有 53.10%的 UAR(Unweighted average recall)，與前人提出的方法，兩階段式投票技術(two-stage decision-level fusion technique (Wu *et al.*, 2017))相比，本實驗結果較其進步 8.9%。此外，我們加入音檔的的 fisher vector 與 DVFN 倒數第二層的結果融合，使得準確率更提升至 56.07%。

此研究對於解構大腦與情緒之間的關係有以下幾點貢獻：

1. 此研究提出一種新的投票融合方法用以整合個體差異。
2. 此投票融合方法對於由 fMRI 資料預測受試者所受的情緒刺激能有不錯的效果。
3. 此研究亦發現加入音檔資訊，使得難過情緒分類結果大幅提升。

接下來的第二部分將針對融合(Fusion)的方法進行文獻回顧；第三部分中我們將介紹功能性磁振造影的資料收集、情緒語句資料庫的準備方法、聲音特徵擷取，以及 DVFN 架構介紹；第四部分則包含我們整體實驗架構、結果與分析；而第五部份將總結這個實驗並提出未來的研究方向。

2. 文獻回顧 (Related Work)

在工程領域中有許多方法被用來整合多模態資料(Ayache, Quénot & Gensel, 2007)，例如 early fusion 利用特徵上的融合(fusion-level)，將不同的特徵連接(concatenate)，並輸入至一個模型中訓練；另一典型的作法為 late fusion，將各模態預測出來的決策分數 (decision score) 結合，再預測一次。此外，kernel fusion 亦為一種常見的做法，利用加權平均(weighted average)的方式結合來自不同模態的特徵值再做預測。

而在深度學習神經網路架構中，最知名的多模態融合方法(Ngiam *et al.*, 2011) 為利用跨模態的自編碼神經網路(cross modality deep autoencoder)訓練出富含多模態資訊的中間層(latent space representeation).

此研究整合傳統機器學習與神經網路的概念，將多個模態以 early fusion 的方式連接作為融合層，用以學習個體的重要程度，接著我們將觀察到的個體權重重新置入深度投票融合神經網路中，進行 finetune。

3. 研究方法 (Research Methodology)

3.1 情緒性聲音刺激設計與fMRI資料收集 (Vocal Emotion Stimuli Design and Collection)

我們參考 the USC IEMOCAP 資料庫(Busso *et al.*, 2008)設計情緒性聲音資料庫做為受試者進行功能性磁振造影時的刺激材料，此材料也曾用在觀察血氧濃度相依對比訊號與韻律特徵之間的關聯(Chen, Liao, Jan, Kuo & Lee, 2016)。此次使用的資料庫含 6 種不同的刺激分別為包含情緒正向、中性、負向以及激動程度高、中、低，而每個種類包含連續 5

分鐘的聲音語句作為受試者的刺激素材。此外這些情緒性聲音語句為 the USC IEMOCAP database 中來自同一表演者所說的話，總共是 251 句話。

3.1.1 情緒分類 (Emotion Classes)

The USC IEMOCAP database 中提供每句情緒語句的情緒標籤，在整個資料庫中共分成 8 種情緒包含難過、高興、興奮、驚訝、中性、生氣、痛苦及挫折。但由於我們只擷取裡面 251 句話，使得 8 類標籤的數量分布不均，且為了驗證本研究所提出的方法，我們參考前人的做法 (Wu *et al*., 2017)，將高興、驚訝、興奮融合成一類；而生氣、痛苦、挫折融合成另一類，因此原先的 8 類情緒標前輩融合成 4 類，如表 1 所式。

表1. 本文使用的情緒標籤分類方法
[Table 1. Summary of the original and the merged labels of the 251 utterances used in this work]

原類別	Valence 等級	Arousal 等級	數量	融合後的類別	數量
難過	負向	低	33	Class 1	33
高興	正向	高	12	Class 2	79
興奮	正向	高	64		
驚訝	正向	高	3		
中性	中性	中	69	Class 3	69
生氣	負向	高	19	Class 4	70
痛苦	負向	低	1		
挫折	負向	低	50		

3.1.2 功能性磁振造影資料蒐集與前處理 (fMRI Data Collection and Preprocessing)

我們總共招募 18 位年齡介在 20~35 歲如圖 1 所示、具有大學以上學歷的受試者。此實驗設計為 block design，原實驗目的為觀察受試著對於情緒正負向 (Valence 之正向、中性及負向) 以及情緒激動程度 (Arousal 之高、中、低) 的反應，因此每個 block 會聆聽 Valence 或 Arousal 中的一種程度之連續 5 分鐘的音檔，音檔內容即上段所敘述，為帶有情緒性的語句。而每段之間設計了 5 分鐘的休息，使受試者感官狀態回復至最初。

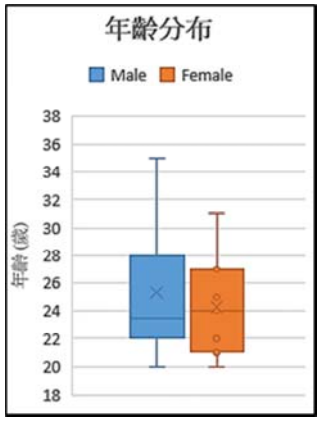

圖 1. 受試者年齡分布圖
[Figure 1. Age distribution of the subjects in this work.]

本次實驗中，我們更進一步的關注於每個 block 長達 5 分鐘的音檔裡每句話的確切情緒對於受試者的影響為何。我們使用的磁振造影掃描機台為 3T scanner (Prisma, Siemens, Germany)，TR=3(s)、體素大小(voxel size)為 3*3*3(mm^3)。接著我們利用 Data Processing Assistant for Resting-State fMRI (DPARSF)進行前處理(Yan & Zang, 2010)，此為一套基於 SPM 與 REST 開發的軟體，可由幾個簡單的按鍵自動的完成影像所有前處理的流程，包含：slice timing, realign, normalize, smooth。完成前處理後，我們將核磁共振影像內差為每秒一張以對應至情緒語句刺激的長度及語句之情緒。

3.1.3 音檔特徵擷取 (Acoustic Feature Extraction)

我們使用兩個步驟導出高維向量作為每個語句的聲音特徵：(1) 提取聲學低階描述(Low Level Description, LLD)，以及(2) 使用基於高斯混合模型（Gaussian Mixture Model, GMM）的費雪向量編碼(Fisher-vector encoding)。我們使用 Praat (Boersma, 2002)取出前 13 個 MFCC(Mel-scale 頻率倒譜係數)、音高、強度以及以 60 Hz 幀速率(framerate)提取的一階和二階特徵。我們將所有語句對中性情緒語句進行 z-score 標準化。由於每句話的長度不同，我們進一步採用 Gaussian Mixture Model- Fisher-vector encoding (GMM-FV)方法：費雪向量編碼通過首先訓練整體背景 GMM 並使用 Fisher 信息矩陣(Fisher Information Matrix, FIM)近似以進一步計算梯度向量來描述訓練的 GMM 參數所需的方向改變，即均值和方差，以獲得更好的擬合來操作關於感興趣的數據樣本，即每個話語的一系列 LLD。

3.2 利用卷積神經網路之特徵提取 (fMRI-CNN)

我們參考前人利用相同的資料庫所做的實驗與結果，發現顳葉區的腦袋資料用於聲音情緒認知辨識有最好的表現(Wu *et al*., 2017)，此外神經科學方面研究亦顯示顳葉區確實參與許多低階聲音情緒認知作業(Phillips, Drevets, Rauch & Lane, 2003; Holt *et al*., 2006; Schirmer & Kotz, 2006)，因此在本研究中，我們使用 Automated Anatomical Labeling (AAL)

的模板取出顳葉區的 fMRI 資料，並且利用與前人相同的手法，亦即卷積神經網路(Convolutional Neural Network, CNN)，來抽取每個受試者顳葉區的特徵向量。

而 CNN 詳細架構如下：其中有 4 層卷積層、3 層池化層、3 層全連接層，及 1 層 softmax 輸出層以輸出對 4 類情緒的激活值，因此共為 11 層隱藏層。超參數的設定為：激活函數皆使用線性整流函數 (Rectified Linear Unit, ReLU)，優化器使用隨機梯度下降法(Stochastic Gradient Descent, SGD)，其參數 weight decay 設定為 0.000001、momentum 設為 0.9、learning rate 為 0.0001，此外 epoch 設為 20，且訓練資料的準確率可達 88%~95%當此深度卷積神經網路訓練完畢時，我們取出倒數第二層的隱藏層(500 個節點)作為功能性磁振造影的特徵。

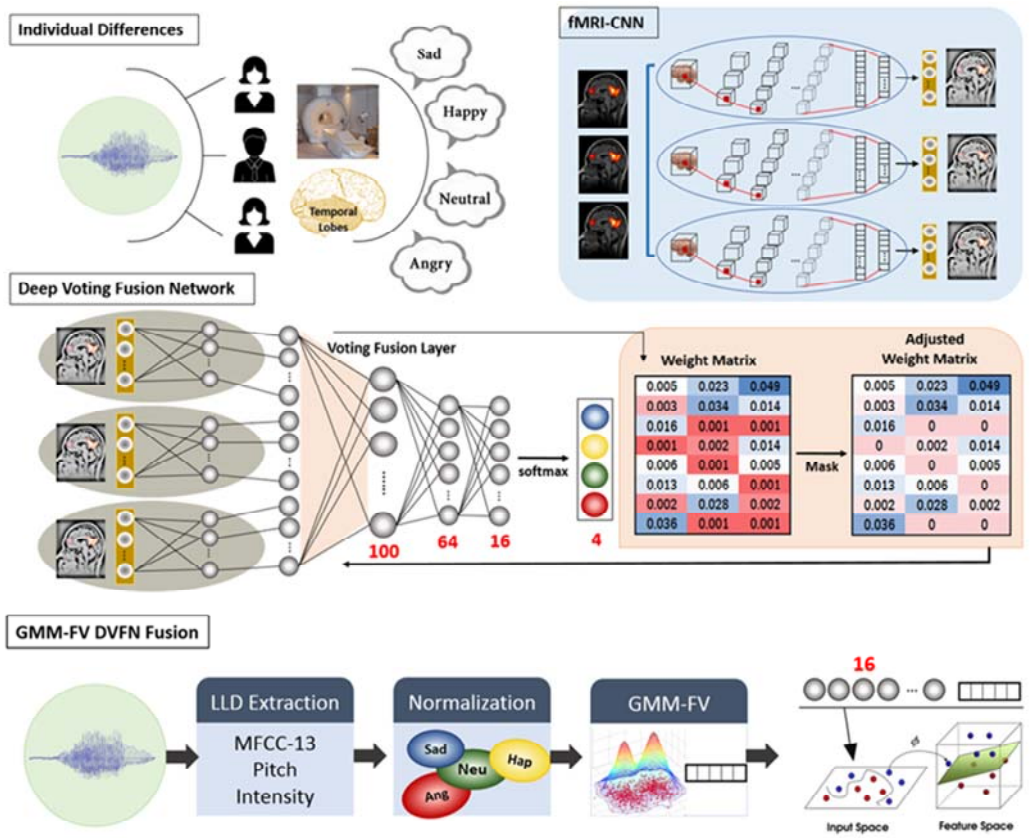

圖 2. 本文流程圖：上半部分為 fMRI 資料蒐集與特徵擷取；中間部分為 DVFN 模型架構；下半部分說明音檔特徵擷取及音檔與 DVFN 融合。

[Figure 2. The upper part of the schematic is our proposed deep voting fusion neural work in performing automatic 4-class vocal emotion decoding; While the lower part fusions the fMRI representation obtained from DVFN with the acoustic feature extracted by GMM-FV.]

3.3 深度投票融合神經網路架構 (DVFN)

我們提出的 DVFN 架構能幫助我們融合每個個體的特質以利認知神經影像所受的的情緒刺激。如圖 2 中所示，DVFN 架構裡共有五層隱藏層：第一層全連階層(dense layer)將從 CNN 萃取出的 500 維特徵向量濃縮至 100 維；第二層為合併層，目的為將每個人的 100 維向量連接再一起並於第三層中投票融合降至 64 維再降為 16 維。最後為一層 4 個節點的 softmax layer 以便做情緒認知辨識。訓練 DVFN 時，我們使用分類交叉熵(categorical crossentropy)作為損失函數、ReLU 作為激活函數、優化器選擇 Adam 且搭配 learning rate 使用 0.0001，最後 epoch 設為 10 次。

3.3.1 投票融合層 (Voting Fusion Layer)

我們引入投票融合層 f 幫助我們融合多個受試者的特質，其運作方式定義如下：

$$D_f = W_f \times D_2 \tag{1}$$

其中 D_f 是融合層 f 的輸出、W_f 為融合層 f 的權重矩陣，D_2 則為 DVFN 架構中第二層的輸出，亦即每人的特徵向量合併層。值得注意的是，我們移除此層的偏差項(bias)，也不使用激活函數，使得此層聚焦於學習每個人的特徵貢獻程度(W_f)。

3.3.2 深度投票融合神經網路 (DVFN)

DVFN 架構主要進行下述兩步驟：(1)從融合層取得 W_f 後，對此權重矩陣進行遮罩(masking)，(2)將調整後的權重矩陣 $W_{adjusted}$ 重新置入融合層作為初始矩陣，並對 DVFN 進行微調(finetuning)。如上述，W_f 代表個體的貢獻程度，我們考量個體對情緒認知的能力並希望能力佳者貢獻較多、反之則抑制，因此在步驟(1)中，我們對 W_f 作遮罩，並引入一個閾值(τ)：

$$\mathrm{f(w)} = \begin{cases} 1, & if\ |w| < \tau \\ 0, & if\ |w| < \tau \end{cases} \tag{2}$$

經過遮罩後的權重強調高貢獻度的節點，且將較小的值當作擾動，如此能更有效的反應出受試者情緒認知能力的優劣。接著在步驟(2)中，我們以 $W_{adjusted}$ 取代 W_f 並重新微調(finetune)整個 DVFN 架構。

3.4 聲音特徵與fMRI特徵融合 (GMMFV-DVFN SVM-Fusion)

我們從上述的 DVFN 架構中取出倒數第二層(16 個節點)的特徵，並將其與 GMM-based 的聲音特徵 fisher vector 相接，進行 early fusion。

4. 實驗設計與結果 (Experimental Setup and Results)

我們的實驗設計為利用 fMRI 影像預測 251 個語句分別為 4 類情緒中何種情緒。在驗證訓練好的架構時，我們採用 leave-one-utterance-out 的交叉驗證方法。我們的結果皆以

unweighted average recall (UAR)顯示。以下為一些我們用來比較的架構以及其敘述，其中 AVB 代表 average-based、INB 則為 individual-based 的縮寫：

- AVB average: 直接將所有人由 fMRI-CNN 萃取出的特徵向量平均起來再透過支持向量機 (Support Vector Machine, SVM)分類。
- INB Individual: 將所有人由 fMRI-CNN 萃取出的特徵向量個別透過 SVM 分類。
- INB SVM-Voting: 將所有人由 fMRI-CNN 萃取出的特徵向量個別透過 SVM 分類並利用 decision score 投票(Wu *et al*., 2017)。
- INB DNN-Fusion: 將所有人由 fMRI-CNN 萃取出的特徵向量放入 DVFN 架構中，然而不調整節點的權重。
- INB DNN-SVM-Voting: 將所有人由 fMRI-CNN 萃取出的特徵向量放入 DVFN 架構中，並取出倒數第二層利用 SVM 分類及 decision score 投票。
- INB DVFN DNN-Fusion: 本文段落二之(三)之架構。
- INB GMMFV-SVM: 對 251 個音檔進行 fisher vector 編碼，利用 SVM 分類。
- CNN-GMMFV SVM-Fusion: 將所有人由 fMRI-CNN 萃取出的特徵向量與音檔的 fisher vector 利用 Decision score 進行 late-fusion，再利用 SVM 分類。
- DVFN-GMMFV SVM-Fusion: 將所有人由 fMRI-CNN 萃取出的特徵向量放入 DVFN 架構中取出倒數第二層的特徵，並與音檔的 fisher vector 利用 Decision score 進行 late-fusion，再利用 SVM 分類。

4.1 fMRI情緒分類結果 (Emotion Classification Results)

表 2 為我們利用 fMRI 情緒辨識的全部結果，準確率皆以 4 類 UAR 表示。我們提出的 DVFN 架構表現最佳，準確率可達 53.1%，較之前研究的結果(Wu *et al*., 2017)，亦即 INB SVM-Voting，相對進步 8.9 %。從表 2 中可以發現，將個體的特徵個別考慮，亦即 INB 與 AVB 比較，結果就有小幅提升，也因此驗證了現今腦科學研究的趨勢：個體差異需要被考量。

此外，投票神經網路，亦即投票融合層的引入，能使得神經網路學習時能共同的學習個體融合的權重，因此能更有效的提升情緒分類的結果。從表 2 中可以觀察利用神經網路投票的效果(INB DNN-Fusion)較 SVM 分類時利用 decision score 投票(INB SVM-Voting)佳；最後再加上調整權重並再次訓練 DVFN 的效果(INB DVFN DNN-Voting)則更甚於沒有調整權重時(INB DNN-Fusion)，準確率可高出 6.9%。這是因為調整權重使得神經網路能專注於貢獻度高的受試者，使他們的特徵有效的提升情緒辨識的結果。

表 2. 使用 DVFN 以及其餘融合方法的四類情緒辨識結果。準確率皆以 UAR(%) 呈現
[Table 2. It presents the 4-class emotion classification results of our proposed deep voting fusion neural network and other fusion techniques. The accuracy is measured in UAR (%).]

4-Class	***AVB: Average***	***INB: Individual***	***INB SVM-Voting***	***INB DNN-Fusion***	***INB DNN-SVM-Fusion***	***INB DVFN DNN-Voting***
Class 1	15.15	12.24	15.15	15.15	15.15	24.24
Class 2	72.15	77.43	84.81	89.87	87.34	87.34
Class 3	44.93	46.41	55.07	55.07	57.97	56.52
Class 4	37.14	40.87	40.00	38.57	47.14	44.29
UAR	42.34	44.31	48.75	49.67	51.90	**53.10**

4.1.1 閾值分析 (Threshold Analysis)

本研究中，我們還列出不同閾值下的準確率，而閾值的調整範圍為 0.001 到 0.01，並使用 Leave One person Out Cross Validation (LOOCV) 的方式進行，結果如圖 3 所示。當閾值設為 0.002 時，對於 4 類情緒預測的準確率最佳。總體來看我們發現當閾值越小(例如≦0.003)準確率看起來較高，因此若使將融合層權重中較多的值調整為 0 則效果會較好。

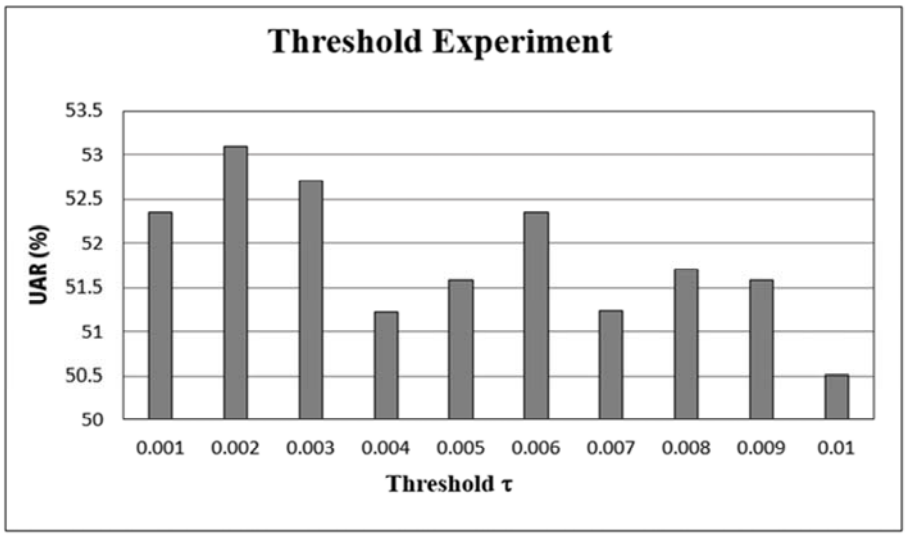

圖 3. 不同閾值 τ 下的四類情緒辨識結果。準確率皆以 UAR(%)表示
[Figure 3. The 4-class motion classification performance measured in UAR (%) as a function of threshold (τ) value]

4.2 音檔與fMRI融合結果 (Multimodal Fusion Results)

表 3 為音檔與 fMRI 融合並情緒辨識的全部結果。首先可以看到當僅利用音檔取出的 GMMF 特徵經過 SVM 對四類情緒分類(*INB GMMFV-SVM*)便可有不錯的結果(53.83 %)，且第一類的準確率高達 63.63%；相較而言利用 fMRI 資料分類時(*INB SVM-Voting*)雖準確率不差(48.75%)，但第一類的分類結果僅有 15.15%，即便在表 2 中不同實驗方法下，第一類分類最高也僅達 24.24%。由此可知音檔的特徵較易辨識難過的情緒，反觀人腦受難過情緒刺激下的變化較難辨別，因此希望藉由音檔與 fMRI 融合的方式提升第一類分類的結果，進而使整體準確率提高。

當將 CNN 抽取的特徵與音檔 GMMFV 融合時(*CNN-GMMFV SVM-Fusion*)，準確率較沒融合時佳，可達 55.25%，尤其對於第一類而言，音檔確實大幅的幫助 fMRI 資料，使其準確率達 36.63%，且第四類的結果也因此種相互資訊的流通而使 UAR 提升至 50%。此外當使用 DVFN 此種考量個體差異的模型所擷取出的 fMRI 特徵與 GMMFV 結合時(DVFN-GMMFV SVM-Fusion)，第一類的預測效果又更加進步至 42.42 %，且總準確率也較所有結果高，達到 56.07 %。

表3. 音檔與fMRI 資料融合方法的四類情緒辨識結果。準確率皆以UAR(%)呈現。[Table 3. It provides a summary of our recognition results using the fusion of audio and fMRI, and the accuracy is measured in UAR (%).]

4-Class	***INB GMMFV-SVM***	***INB SVM-Voting***	***CNN-GMMFV SVM-Fusion***	***DVFN-GMMFV SVM-Fusion***
Class 1	63.63	15.15	36.63	**42.42**
Class 2	49.37	84.81	76.08	74.68
Class 3	60.87	55.07	58.26	57.97
Class 4	41.43	40.00	50.00	49.22
UAR	53.83	48.75	55.25	**56.07**

5. 結論與未來研究 (Conclusion and Future Work)

存在於人體神經反應的個體差異在情緒認知方面以及其他高認知功能的腦區造成影響，因此對這方面的研究而言，如何處理個體差異是一項挑戰，需要以更複雜的方式模擬這項機制。我們提出創新的多人融合投票架構偵測情緒刺激，此一演算法全面的考量個人特質，透過融合層及其權重矩陣觀察個體對於預測情緒的重要程度，並藉由此深度神經網路自動化的辨識情緒刺激。利用此架構預測 4 類情緒時，我們的準確率(Unweighted average recall, UAR)可達 53.10%，此外我們發現當將權重矩陣中越多的值調整為 0 能提升準確率。且若再加入聲音資訊，準確率又可提升至 56.07%，且第一類的預測結果大幅提升 75%。

未來有許多方向可以繼續延伸，其一為利用時序架構例如遞迴神經網路(recurrent

neural network, RNN)或長短期記憶神經網路(long-short term memory neural network, LSTM) 模擬腦訊號並預測情緒 (Li, Song, Zhang, Hou & Hu, 2017; Soleymani, Asghari-Esfeden, Fu & Pantic, 2016)，這是因為 fMRI 影像具時間資訊，可以一併放入這些專門編碼時間訊息的神經網路中以獲得更有效的特徵。其二，我們正努力延伸此概念在研究腦區與特定情緒的關聯性，期望能夠找出受情緒影響的腦區，並觀察較容易辨識情緒的受試者與不易辨識的受試者間的腦區特徵與差別。

參考文獻 (References)

ALBA-FERRARA, L., HAUSMANN, M., MITCHELL, R. L., & WEIS, S. (2011). The neural correlates of emotional prosody comprehension: disentangling simple from complex emotion. *PLoS One*, *6*(12), e28701. doi: 10.1371/journal.pone.0028701

AYACHE, S., QUÉNOT, G., & GENSEL, J. (2007). Classifier fusion for SVM-based multimedia semantic indexing. In *Proceedings of Proceedings of the 29th European conference on IR research*, 494-504. doi: 10.1007/978-3-540-71496-5_44

BOERSMA, P. (2002). Praat, a system for doing phonetics by computer. *Glot international*, *5*(9-10), 341-345.

BUSSO, C., BULUT, M., LEE, C. C., KAZEMZADEH, A., MOWER, E., KIM, S., … NARAYANAN, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, *42*(4), 335-359. doi: 10.1007/s10579-008-9076-6

CANLI, T., SIVERS, H., WHITFIELD, S. L., GOTLIB, I. H., & GABRIELI, J. D. E. (2002). Amygdala response to happy faces as a function of extraversion. *Science*, *296*(5576), 2191. doi : 10.1126/science.1068749

CHEN, H.-Y., LIAO, Y.-H., JAN, H.-T., KUO, L.-W., & LEE, C.-C. (2016). A Gaussian mixture regression approach toward modeling the affective dynamics between acoustically-derived vocal arousal score (VC-AS) and internal brain fMRI bold signal response. In *Proceedings of 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5775-5779. doi: 10.1109/ICASSP.2016.7472784

DUBOIS, J. & ADOLPHS, R. (2016). Building a science of individual differences from fMRI. *Trends in cognitive sciences*, *20*(6), 425-443. doi: 10.1016/j.tics.2016.03.014

FOSSATI, P., HEVENOR, S. J., GRAHAM, S. J., GRADY, C., KEIGHTLEY, M. L., CRAIK, F., …MAYBERG, H. (2003). In search of the emotional self: an fMRI study using positive and negative emotional words. *American Journal of Psychiatry*, *160*(11), 1938-1945. doi: 10.1176/appi.ajp.160.11.1938

HAMANN, S. & CANLI, T. (2004). Individual differences in emotion processing. *Current opinion in neurobiology*, *14*(2), 233-238. doi: 10.1016/j.conb.2004.03.010

HOLT, D. J., KUNKEL, L., WEISS, A. P., GOFF, D. C., WRIGHT, C. I., SHIN, L. M., …HECKERS, S. (2006). Increased medial temporal lobe activation during the

passive viewing of emotional and neutral facial expressions in schizophrenia. *Schizophrenia research*, *82*(2-3), 153-162. doi: 10.1016/j.schres.2005.09.021

VAN HORN, J. D., GRAFTON, S. T., & MILLER, M. B. (2008). Individual variability in brain activity: a nuisance or an opportunity? *Brain imaging and behavior*, *2*(4), 327-334. doi: 10.1007/s11682-008-9049-9

KANAI, R. & REES, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience, 12*(4), 231-242. doi: 10.1038/nrn3000

LI, X., SONG, D., ZHANG, P., HOU, Y., & HU, B. (2017). Deep fusion of multi-channel neurophysiological signal for emotion recognition and monitoring. *International Journal of Data Mining and Bioinformatics*, *18*(1), 1-27. doi: 10.1504/IJDMB.2017.086097

MACDONALD, S. W., NYBERG, L., SANDBLOM, J., FISCHER, H., & BÄCKMAN, L. (2008). Increased response-time variability is associated with reduced inferior parietal activation during episodic recognition in aging. *Journal of Cognitive Neuroscience*, *20*(5), 779-786. doi: 10.1162/jocn.2008.20502

NGIAM, J., KHOSLA, A., KIM, M., NAM, J., LEE, H., & NG, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 689-696.

PARASURAMAN, R. & JIANG, Y. (2012). Individual differences in cognition, affect, and performance: Behavioral, neuroimaging, and molecular genetic approaches. *Neuroimage*, *59*(1), 70-82. doi: 10.1016/j.neuroimage.2011.04.040

PHILLIPS, M. L., DREVETS, W. C., RAUCH, S. L., & LANE, R. (2003). Neurobiology of emotion perception I: the neural basis of normal emotion perception. *Biological psychiatry*, *54*(5), 504-514. doi: 10.1016/s0006-3223(03)00168-9

SCHIRMER, A. & KOTZ, S. A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in cognitive sciences*, *10*(1), 24-30. doi: 10.1016/j.tics.2005.11.009

SOLEYMANI, M., ASGHARI-ESFEDEN, S., FU, Y., & PANTIC, M. (2016). Analysis of EEG signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, *7*(1), 17-28. doi: 10.1109/TAFFC.2015.2436926

WU, Y., CHEN, H.-Y., LIAO, Y.-H., KUO, L.-W., & LEE, C.-C. (2017). Modeling perceivers neural-responses using lobe-dependent convolutional neural network to improve speech emotion recognition. In *Proc. of Interspeech 2017*, 3261-3265. doi: 10.21437/Interspeech.2017-562

YAN, C.-G. & ZANG, Y.-F. (2010). DPARSF: a MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. *Frontiers in systems neuroscience*, *4*, 13. doi: 10.3389/fnsys.2010.00013

ZHOU, T., WANG, H., ZOU, L., ZHOU, R., & QIAN, N. (2013). A Study of Neural Mechanism in Emotion Regulation by Simultaneous Recording of EEG and fMRI Based on ICA. In: Guo C., Hou ZG., Zeng Z. (eds), *Advances in Neural Networks – ISNN 2013*.

Lecture Notes in Computer Science, (pp. 44-51). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-39065-4_6