The 2017 Conference on Computational Linguistics and Speech Processing
ROCLING 2017, pp. 254-263

Using Teacher-Student Model For Emotional Speech Recognition

蕭伯瑋 Po-Wei Hsiao
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
M053040067@student.nsysu.edu.tw

謝博丞 Po-Chen Hsieh
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
M053040031@student.nsysu.edu.tw

陳嘉平 Chia-Ping Chen
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
cpchen@mail.cse.nsysu.edu.tw

摘 要

本研究使用Teacher-student model藉由修改訓練資料的標籤來重新訓練靜態分類模型。研究中會使用偏斜强健性類神經網路做訓練及分類，網路在訓練時會加入與各類情緒資料筆數呈反比的權重，以解決資料不平衡的問題。資料前處理的部分則是對訓練資料和測試資料做語者正規化來消除各語者之間的差異性。上述方法使用FAU-Aibo 情緒語料庫來做評估，並與 Interspeech 2009 Emotion Challenge 分類子挑戰做辨識率的比較。在Interspeech 2009 Emotion Challenge 分類子挑戰中靜態模型的基準辨識率為38.2%，參賽者中最佳的辨識率為41.65%；而本實驗所得到的辨識率為46.0%。

關鍵字：情緒辨識，情緒語料庫，多層感知器，Teacher-Student Model

一、緒論

近年來機器學習發展迅速，應用到許多不同的層面，例如微軟所開發的智慧型個人助理Cortana，或是蘋果的Siri，意味著人工智慧的發展會越來越融入人類的生活。

過往人機介面都需要使用者主動操縱機器，現在即可藉由語音讓使用者跟電腦進行互動。雖然諸如上述的智慧型介面可以對於語音進行辨識，但是還無法對情緒進行精確分析。如果電腦可以依照使用者當下的情緒去作出反應，例如智慧系統依照使用者情緒與使用者進行互動。因此電腦在接收我們的指令時，除了字面上的意思外也應該考慮情緒的差異。例如以色列科技公司Beyond Verbal開發出可以根據對話偵測人類情緒、意圖與個性的軟體: Moodies Emotions Analytics。 從1980年代開始，已經有學者發現情緒上存在著普遍能辨識出的特徵，這些特徵與人類的發聲模式有關，開啓了使用語音情緒特徵進行分類的先河[1]。在1997 年時，Picard 等人[2] 描述了情緒辨識的應用及重要性。情緒語料庫例如：柏林情緒語料庫FAU-Aibo[3]、EMO-DB[4]、LCD情緒語料庫及波蘭情緒語料庫。LIN Chu-Hsuan 等人[5]整理常用於辨識的語音特徵，包括能量、音高、過零率、噪音諧音比和梅爾頻率倒譜係數等。本實驗所使用的情緒語料庫是FAU- Aibo，由於FAU-Aibo錄製的是孩童的自然對話，情緒較不鮮明，因此會影響辨識的準確度。因為此語料庫都是以德國人用德語錄製而成，除了語系不同外，德國人在情緒表現上可能也會和我們有所差異。

情緒辨識系統一開始會先對原始音檔做處理，之後會擷取聲學特徵並用訓練資料來訓練分類模型，最後使用訓練好的分類模型對其進行分類。常用的分類模型有支持向量機(Support Vector Machine, SVM)[6]、多層感知器(Multilayer Perceptron, MLP)[7]、高斯混合模型(Gaussian Mixture Model, GMM)[8]、隱藏式馬可夫模型(Hidden Markov Model, HMM)[9]、遞歸神經網路(Recurrent Neural Network, RNN)[10]。在 Interspeech 2009 Emotion Challenge[11] 分類子挑戰中靜態模型的基準辨識率為 38.2%，使用的分類器為SVM，前處理部分除了會對資料做標準化之外還會使用Synthetic Minority Oversampling TEchnique (SMOTE)對資料做平衡。

本論文主要分為四個部分：第一部分為緒論；第二部分為實驗架構及研究方法；這個部分會介紹實驗流程及資料處理的方法；第三部分會放上實驗結果；第四部份會歸納實驗結論。

二、 研究方法

（一）、 Teacher-Student Model

Teacher-student model 的訓練流程為：先使用訓練資料訓練分類模型，並以此模型對訓練資料做辨識分類，分類的結果會使每筆訓練資料皆會得到一個包含五類情緒事後機率的向量，再以這些向量中作為每筆訓練資料的新標籤，即為 teacher label，並重

新訓練分類模型。因此訓練資料的標籤將會由原本的 one-hot vector 表示法轉變為包含五類情緒事後機率的向量。此訓練過程即為 teacher-student training。訓練過程如圖一。

由於原始的標籤是由五位語言學者投票標記的，但不同學者對語音所要表達的情緒解讀可能不同，因此可能會出現特徵類似但被標記成不同類或是特徵不同但被標記為同類的情況。這種情況會不利於一般的 MLP 分辨語音的特徵，所以藉由將標記的方式由原本的 one-hot vector 改成以機率呈現的 teacher label 來使模型能夠更廣泛的考慮各類特徵並進行分類，進而提升辨識率。

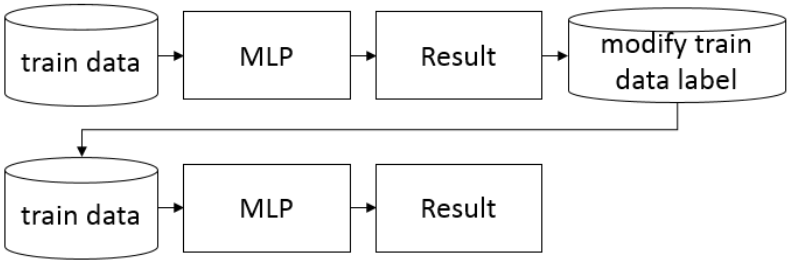

圖 一、 Teacher-Student Model 訓練流程圖

（二）、多層感知器(Multi-Layer Perceptron, MLP)

類神經網絡是一種模仿生物神經網絡(動物的中樞神經系統，特別是大腦)的結構和功能的數學模型或計算模型，用於對函數進行估計或近似。神經網絡由大量的人工神經元連結進行計算。大多數情況下人工神經網路能在外界訊息的基礎上改變內部結構，是一種自適應系統。現代神經網絡是一種非線性統計性數據建模工具。

多層感知器為一種前向結構的神經網路，使用 back propagation 作為學習的演算法，以監督式的方式進行學習，處理輸入與輸出之間的非線性映射關係。 Back propagation network 是由向前傳遞(forward pass)及向後傳遞(backward pass)兩部分所組成，向前傳遞是先將訓練資料放進網路中去執行，之後再計算出輸出值與目標值之間的誤差，而後向傳遞是根據誤差值去對權重進行調整，經過這樣多次的訓練之後，就能夠將網路的誤差值修正到極小的範圍內。 Back propagation network 的特性主要有：

I. 學習精確度高

II. 回想速度快

III. 可以處理非線性問題

在進行向前傳遞時:會使用sigmoid function(1) 作為做為激活函數，sigmoid function 會使得輸出值位於區間[0,1]函數如下。

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

會以Cross entropy(2)作為計算誤差的方法

$$-\sum_{k=1}^{K} target_k \times \log(predict_k) \tag{2}$$

加入權重之最佳化參數更新以下列公式做為表示

$$-r_{ik} \times \sum_{k=1}^{K} target_k \times \log(predict_k) \tag{3}$$

最後使用 Softmax function(4)將各類輸出結果轉換為機率。

$$\sigma(y_i) = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{4}$$

在進行向後傳遞時:計算輸出層(5)及隱藏層(6)的誤差值

$$\delta_k = (t_k - y_k) \times y_k \times (1 - y_k) \tag{5}$$

$$\delta_j = (\sum_{k=1}^{K} w_{jk} \times \delta_k) \times y_k \times (1 - y_k) \tag{6}$$

調整隱藏層到輸出層(7)及輸入層到隱藏層(8)的權重，x_i :輸入層的神經元，z_j :隱藏層的神經元，y_k :輸出層的神經元 η：learning rate , m：momentum

$$w_{jk} = w_{jk} + \eta \times \delta_k \times z_j + m \times \Delta w_{previous} \tag{7}$$

$$w_{ij} = w_{ij} + \eta \times \delta_j \times x_i + m \times \Delta w_{previous} \tag{8}$$

資料前處理的部份，由於不同語者產生的聲音會有所差異，因此本實驗會使用語者正規化的方法來消除此差異性，並只保留情緒的變異。語者正規化會將多個實際語者轉換為一個虛擬語者，如此一來我們就能夠得到一個虛擬語者的資料分布，接下來將每個實際語者都轉換成虛擬語者的分布，正規化的方法為直方圖均衡法(Histogram Equalization, HE)[12]。語者正規化的流程如圖二。

為了處理各類訓練資料不平衡的問題，對於每一類別，本研究參考[13]引入一個類別權重 r_{ik} 來調整參數更新。其中，r_{ik} 為該類別之訓練資料數與整體資料數之相對頻率的倒數，與該筆資料所屬類別之總資料數成反比(9)。加入類別權重進行訓練的MLP模型即為Skewness-robust MLP。

$$r_{ik} = \frac{N}{N_k} \propto \frac{1}{N_k} \tag{9}$$

各類別的權重如表一所示。

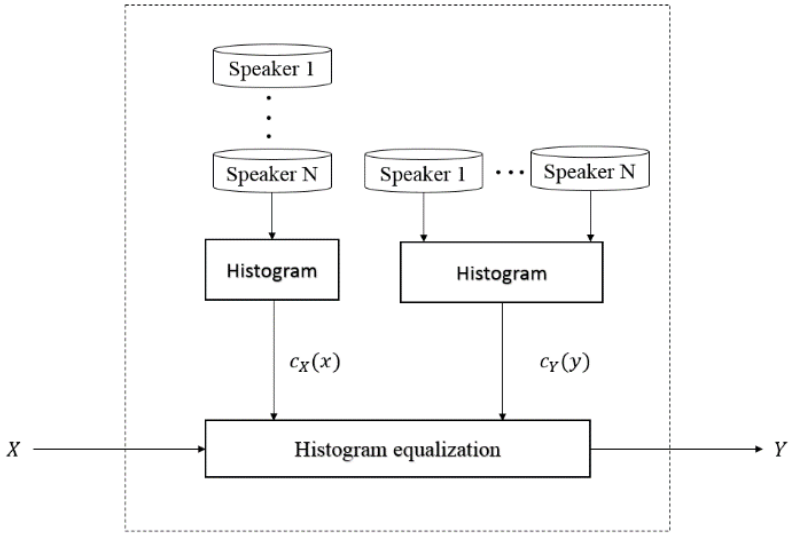

圖 二、 語者正規化流程

表 一、 類別權重

	Angry	Emphatic	Neutral	Positive	Rest
Weight	1.1	0.5	0.2	1.5	1.4

三、 實驗結果

FAU-Aibo情緒語料庫為德國研究員Stefan Steidl根據 51 名 10-13 孩童與 Sony 的機器狗 Aibo 互動 9 小時所產生的語音檔，透過近距離麥克風將孩子們自然發出的聲音所產生的情緒記錄下來，用人為手動的方法將音訊檔切割成較小的片段，其中訓練資料包含來自同一個學校的 26 名孩童，而測試資料的 25 名孩童是來自於另一所學校。語料庫的情緒標記工作由 5 名專業的語言學者共同完成，共分為11類情緒，分別為：歡樂(Joyful)、驚訝(Surprised)、强調(Emphatic)、無奈(Helpless)、敏感(Touchy)、憤怒(Angry)、媽媽語(Motherese)、無聊(Bored)、譴責(Reprimanding)、中性(Neutral)與正向 (Positive)。本實驗依照 Interspeech 2009 Emotion Challenge的情感識別挑戰，選出憤怒(Angry)、强調(Emphatic)、中性(Neutral)、正面(Positive)、其餘(Rest)五類情緒。此五類的資料量如表二所示。

本實驗根據[11]中所設定的 基準聲學特徵並使用 openSMILE 工具擷取，包含 16 個低階參數(Low-Level descriptors, LLDs)及 12 個泛函(Functionals)。16 個低階參數分別是：梅爾頻率倒譜系數 (Mel-Frequency Cepstral Coefficients，MFCCs)(1-12維) 、均方根

表 二、 FAU-Aibo情緒語料庫

	Angry	Emphatic	Neutral	Positive	Rest
Train	881	2093	5590	674	721
Test	611	1508	5377	215	546

能量(RMS energy)、過零率(Zero Crossing Rate, ZCR)、諧音噪音比(Harmonics-to-Noise Ratio, HNR)、音高頻率(Pitch Frequency)，加上每個低階參數的一階係數差(Delta)。12個泛函(Functionals)為：平均值、標準差(standard deviation)、峰度(Kurtosis)和偏移態(Skewness)、最大最小值、相對位置(Relative Position)、範圍(Range)以及另外兩個線性迴歸係數(Linear Regression Coefficients)及其均方差(Mean Square Error, MSE)。因此，對於每一個低階參數，經過一階係數差計算再經由12個泛函計算後，最後得到的特徵集包含了 16×2×12=384 維特徵參數。

在訓練網路之前，會先將訓練資料正規化到 [0,1] 以降低原始資料間的差異性。實驗中所使用的 MLP 架構為一層 30 個神經元的隱藏層如圖三。本實驗會比較有無使用 teacher-student training 的差異，第一組實驗以 FAU-Aibo 訓練資料對偏斜强健性類神經網路訓練 600 個 epoch，使用的參數如表三，於 FAU-Aibo 測試資料得到 44.6% 的辨識率。第二組實驗使用同樣的參數以及訓練資料對偏斜强健性類神經網路做 teacher-student training，再以此模型對 FAU-Aibo 測試資料做分類且得到 46% 的辨識率，實驗結果如表四。各類情緒分類情形如表五、表六。本研究使用 teacher-student model 所得到的辨識率(46%)比基準辨識率(38.2%)高出約8%，此外，根據 Interspeech 2009 Emotion Challenge 參賽者所得到的多組實驗結果[14]中，最佳的結果為Marcel Kockmann 等人[15]所獲得的的41.65%。

四、 結論

根據表四所得到的實驗結果顯示，使用 teacher-student model之後，辨識率能夠從44.6%提升到46%。因此可得知原本用人為標記的方式存在一些問題，可能會造成 MLP 在學習時，無法針對資料的特徵值進行學習，而在使用 teacher label 改變原本的標籤後，有助於提升 MLP 對 FAU-Aibo 情緒語料庫的辨識率。有鑑於對資料的標籤做修改能夠提升辨識率，因此在未來的研究中，我們想進一步的去研究資料標籤的標記方法，若能以其他方式結合 teacher-student training ，或許能在更短的時間內對大量的

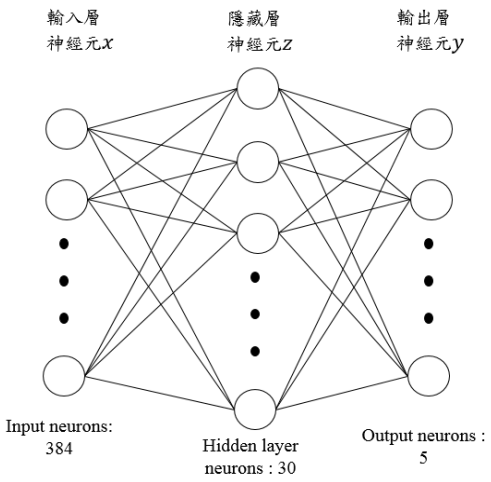

圖 三、 MLP架構圖

表 三、 實驗參數

Hyperparameter	Value
Mini-batch	100
Learning rate	0.4
Learning rate decay	0.0005
Momentum	0.5
Optimizer	Stochastic gradient descent
Loss function	Cross-entropy
Epoch	600

表 四、 MLP、Teacher-Student Model 實驗結果

	Recall
Skewness-robust MLP	44.6%
Teacher-student model	46.0%

資料做標記或修改，可能會遇到的挑戰包含標記的結果是否具有足夠的可靠性，以及使用此標籤進行訓練時，對於神經網路訓練過程的影響。希望能夠結合相關的資料標記方法來讓 MLP 對於五類 FAU-Aibo 情緒語料庫的辨識率提升。

表 五、 Skewness-robust MLP 分類結果混淆矩陣

	Angry	Emphatic	Neutral	Positive	Rest	Recall
Angry	300	131	70	30	80	49.1%
Emphatic	218	778	281	52	179	51.6%
Neutral	528	900	2209	666	1074	41.1%
Positive	11	10	29	116	49	54.0%
Rest	300	79	121	104	150	27.5%
					Avg.recall	44.6%

表 六、 Teacher-student model 分類結果混淆矩陣

	Angry	Emphatic	Neutral	Positive	Rest	Recall
Angry	329	110	72	37	63	53.8%
Emphatic	265	776	278	92	97	51.5%
Neutral	630	948	2085	1073	641	38.8%
Positive	8	7	32	141	27	65.6%
Rest	80	86	117	151	112	20.5%
					Avg.recall	46.0%

參考文獻

[1] Van Bezooijen, Renée, Stanley A. Otto, and Thomas A. Heenan. "Recognition of vocal expressions of emotion: A three-nation study to identify universal characteristics." Journal of Cross-Cultural Psychology 14.4 (1983): 387-406.

[2] Picard, Rosalind W., and Roalind Picard. Affective computing. Vol. 252. Cambridge: MIT press, 1997.

[3] S. Steidl, “Automatic classification of emotion related user states in spontaneous children’sspeech,” PhD thesis, University of Erlangen-Nuremberg, 2009.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, “A database of german emotional speech.” in Interspeech, vol. 5, 2005, pp. 1517–1520.

[5] LIN Chu-Hsuan, CHEN, Yen-Sheng, “結合非線性動態特徵之語音情緒辨識 (Speech Emotion Recognition via Nonlinear Dynamical Features)”[In Chinese], in ROCLING 2015.

[6] Hu, Hao, Ming-Xing Xu, and Wei Wu. ”GMM supervector basedSVM with spectral features for speech emotion recognition.”Acoustics, Speech and Signal Processing, 2007. ICASSP 2007.IEEE International Conference on. Vol. 4. IEEE, 2007.

[7] Kamaruddin, Norhaslinda, and Abdul Wahab. ”Emulating humancognitive approach for speech emotion using MLP and Gen-SofNN.” Information and Communication Technology for theMuslim World (ICT4M), 2013 5th International Conference on.IEEE, 2013.

[8] Cheng, Xianglin, and Qiong Duan. ”Speech emotion recognition using gaussian mixture model.” The 2nd International Conference on Computer Application and System Modeling. 2012. APA

[9] Metallinou, Angeliki, Athanasios Katsamanis, and Shrikanth Narayanan. ”A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs.” Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012.

[10] Hopfield, John J. "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the national academy of sciences 79.8 (1982): 2554-2558.

[11] Schuller, Björn W., Stefan Steidl, and Anton Batliner. ”The INTERSPEECH 2009 emotion challenge.” Interspeech. Vol. 2009. 2009

[12] B.-C. Chiou, “Cross-lingual automatic speech emotion recognition,”Master’s thesis, National Sun Yat-sen University, 2014

[13] P.-Y.Shih, Skewness-Robust Neural Networks with Application to Speech Emotion Recognition, Master’s thesis, National Sun Yat-sen University, 2016.

[14] Schuller, Björn, et al. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge." Speech Communication 53.9 (2011): 1062-1087.

[15] Kockmann, Marcel, Lukáš Burget, and Jan Černocký. "Brno university of technology system for interspeech 2009 emotion challenge." Tenth Annual Conference of the International Speech Communication Association. 2009.