The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 10-21

以多層感知器辨識情緒於國台客語料庫

Use Multilayer Perceptron To Recognize Emotion in Mandarin,Taiwanese and Hakka Database

詹佳憲 Chia-Hsien Chan
國立中山大學資訊工程學系
Department of Computer Science and Information Engineering
National Sun Yat-sen University
hihi4442@gmail.com

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

本研究為使用多層感知器 (Multilayer Perceptron, MLP) 基於聲學特徵參數的情緒辨識系統，此實驗使用一個新的台灣語言語料庫，此語料庫以仿照 EMO-DB 的方式錄製，包含台灣常見的三種語言，分別為國語、台語及客語。每種語言各由五男五女錄製而成，且事後以人工篩選的方式，將較不易分辨情緒的音檔刪除，本研究將使用 180 維的聲學特徵以多層感知器進行單一語言、跨語言及混合語言的實驗。在用單一語言作為訓練集時，國、台、客語分別得到 60%、48.9%、54.4% 的辨識率，而經過語者正規化與使用混合語料做為訓練集後國、台、客語分別得到最好 63.5%、53.1%、64.6% 的辨識率。

關鍵詞：情緒辨識、情緒辨識資料庫、多層感知器

1 緒論

近年來行動祕書日漸普及，包括我們熟知的 Siri，表示人機互動在未來也會越來越重要。如果電腦可以依照使用者當下的情緒去作出反應，例如智慧家電依照使用者情

緒去播放適合的音樂，或調整室內溫度等等。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。情緒的表達可以透過語意或語調，常常相同的一段話，在不同的語調下意思會大相逕庭，例如 “小心點” 可以是表示關心或威脅。此次實驗主要是透過語調來判斷語者的情緒，使用最近相當流行的神經網路辨識系統，結合以台灣本土語言錄製的語料庫來進行情緒辨識的研究。

在 1997 年時，Picard 等人 [1] 描述了情緒辨識的應用及重要性。比較廣為人知的情緒語料庫有德國柏林的 EMO-DB [2] 以及 FAU Aibo [3]，這兩種語料庫最大的差別為，EMO-DB 是由專業語者錄製而成，因此會有較為鮮明的情緒表現，Siqing Wu 等人 [4] 透過調變頻譜特徵與其他特徵組合進行分類可達到 91.6% 的辨識率。而 FAU Aibo 錄製的是孩童的自然對話，情緒較不鮮明，目前最好的辨識結果為 [5] 使用 Deep Belief Network (DBN) 與 Hidden Markov model (HMM) 分類的 45.6%。可看出兩個語料庫目前在辨識率上的差異是相當大的。由於兩個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。這次實驗使用 [6] 所錄製的語料庫，語者皆為台灣的大學生，使用的也是台灣人常用的國、台、客語。

情緒辨識分為一開始的訊號處理、特徵擷取，和之後的分類模型。[7] 歸納音訊特徵包含時域特徵如過零率、音高、能量，和頻域特徵如梅爾頻率倒譜係數 (MFCC) 等。常用的分類模型有支持向量機 (Support Vector Machine, SVM)、多層感知器 (Multilayer Perceptron, MLP)。此篇論文所使用的分類模型為 MLP，除了一般常見的輸入－隱藏層－輸出這樣的架構外，也會以多層的 MLP 來進行辨識。在 [8] 中 Iliou 等人以 MLP 和 SVM 對 EMO-DB 做情緒辨識，辨識率分別為 94% 與 80%，表示 MLP 在處理像 EMO-DB 這種資料量多的語料庫時辨識率會比傳統 SVM 好，這也是為何近年來深度學習開始被廣泛研究的原因。不過此次研究所使用的國台客語料庫資料量並不多，[6] 以 SVM 進行辨識後各語言的辨識率約是 60%。在資料量較少的情況下 MLP 是否能達到不輸 SVM 的辨識率，甚至進一步提升辨識率是本次研究的重點。本研究會嘗試使用語者正規化來消除每筆資料間不必要的差異性，或最大限度地增加訓練資料的數量。

2 研究方法

2.1 語料庫

此語料庫製作於 2013 年，每種語言各找五男五女錄製而成，接著再以十位測試者用人工辨識的方式，刪除辨識率低於 60% 的句子，表 1 為此語料庫三種語言各情緒的

句數。

表 1: 國台客語料庫各情緒句數

情緒	國	台	客	跨語言總數
生氣	58	73	62	193
無聊	66	55	59	180
噁心	46	55	54	155
害怕	57	61	64	182
開心	67	58	48	173
傷心	57	56	52	165
中性	87	64	62	213
總數	438	422	401	1261

2.2 分類方式

多層感知器 (Multilayer Perceptron, MLP) 是一種機器學習的演算法，其向前結構的人工神經網路能映射一組輸入向量到一組輸出向量。MLP 可被看作由多個節點構成的有向圖，每個節點可比做人類的神經元，每一神經元都帶有一個非線性的激活函數，例如 sigmoid function 或 hyperbolic tangent。並使用 Hinton 等人 [9] 所提到屬於監督式學習的反向傳播演算法來訓練 MLP，以梯度下降的方式來更新每個神經元的權重，(1) 為梯度下降的公式，θ 為權重、$\ell(\theta)$ 為 cost function。

$$\theta_j := \theta_j + \alpha \frac{\partial}{\partial \theta_j} \ell(\theta) \tag{1}$$

近年來由於深層學習的成功，MLP 算法又開始重新受到關注，本研究也會比較單層及深層 MLP 在辨識率上的差異性。我們使用 google 開發的機器學習工具 tensorflow [10] 建製單層 MLP 及具有兩層隱藏層的深層 MLP，且在每個隱藏層後加入 dropout [11] 以防止過擬合的狀況發生，一層網路的神經元數為 35 個，而兩層網路的神經元數分別為 60 及 15 個，在訓練前會先對訓練集與測試集做 normalize attributes，將每個特徵縮小範圍到 -1 與 1 之間，公式為 (2)，mean(X) 為平均值，max(X) 、min(X) 分別為最大值、最小值。圖 1 為單層 MLP 架構，圖 2 為深層 MLP 架構。

$$X' = \frac{X - mean(X)}{max(X) - min(X)} \tag{2}$$

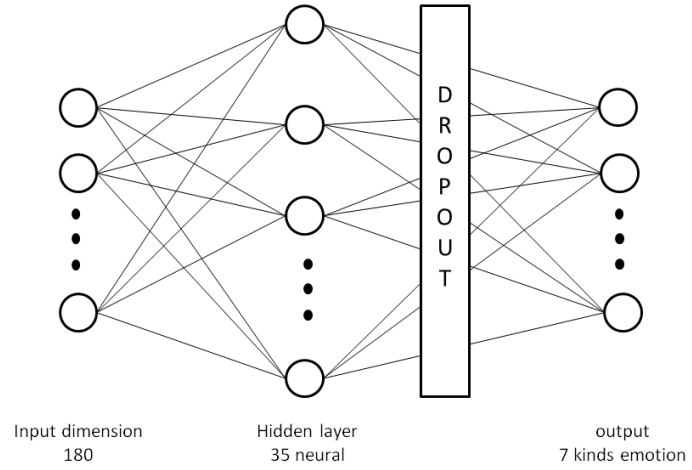

圖 1: 單層MLP

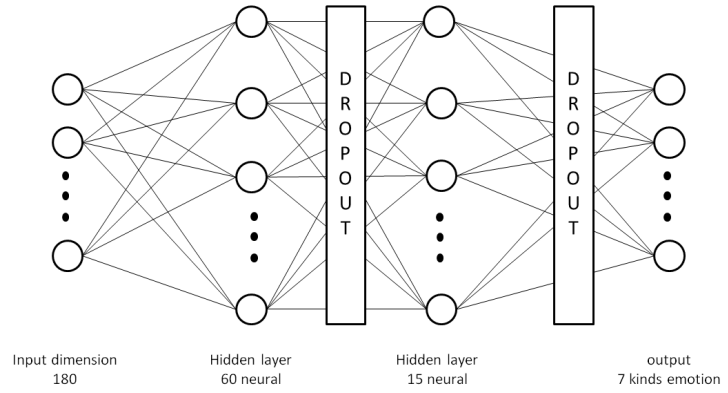

圖 2: 兩層MLP

2.3 語者正規化(Cross-speaker histogram equalization, CSHE)

語者正規化的目的為消除語者之間的差異性，只保留情緒的變異。圖 3 為語者正規化的流程，將多個訓練語者視為一個虛擬語者，如此我們能得到一個虛擬語者的資料分布，接著將每個語者分布皆轉換至虛擬語者的分布，使用的正規化方法為直方圖均衡法 (histogram equalization, HE) [6]。關於語者正規化後的分布差異可參考 [6]，本研究會比較有無語者正規化在辨識率上的差異。

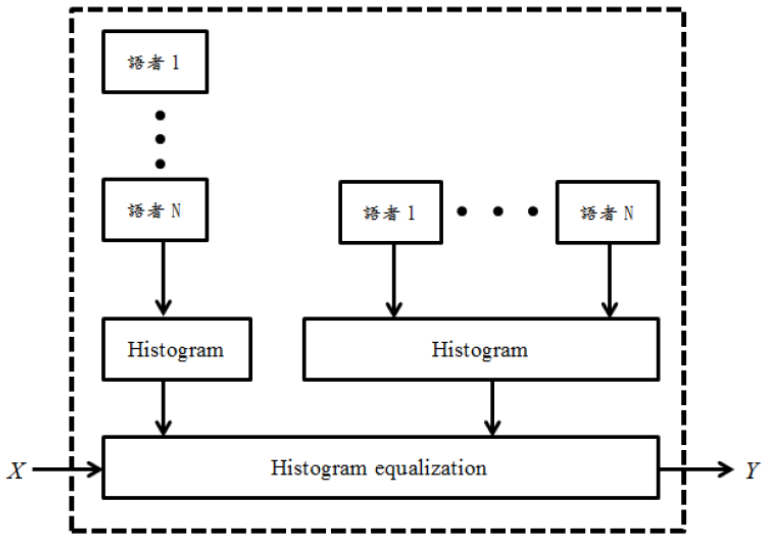

圖 3: 語者正規化流程

3 實驗結果

本研究使用 [6] 中的 180 維聲學特徵，其中包含 15 個低階參數，分別為 13 個 MFCC、1 個音高 (pitch)、1 個過零率 (zero-crossing rate,ZCR)。這 15 個低階參數加上其一階係數差 (delta) 再乘以 6 個泛函就是本研究所使用的 180 維特徵。每組實驗都會比較一層 MLP 與兩層 MLP 在有無做 CSHE 且使用不同訓練資料後辨識率的差異。RAW 表示訓練集和測試集皆未做 CSHE。每次 CSHE 實驗都是以測試資料語言的所有語者作為虛擬語者，再將其他語者分佈轉換至該虛擬語者的分佈，以消除語者間的差異性。例如當測試資料為國語時，則以國語的語者作為虛擬語者並將其他語言的語者轉換至其分佈。

3.1 單一語料實驗

此實驗是採 leave one speaker out 的方式進行，輪流將其中一位語者的資料做為測試集，其他語者作為訓練集，例如我們有十位語者 〈a1,a2,...,a10〉，第一次訓練以 a1 語者的資料作測試集，其他 a2 ～ a10 語者當訓練集，反覆上述步驟直到所有語者資料都被當作測試集為止，最後將每位語者分別測試時正確辨識的句子數加總，除以測試集語言的總句數即為該次實驗的辨識率。表 2 為實驗結果。由結果可看到，一層的辨識率全都優於兩層，此外國語跟客語的辨識率較為接近且都優於台語，而做完 CSHE 之後台語的辨識率下降，但國語跟客語都有所提升。

表 2: 單一語料實驗結果

	RAW		CSHE	
	MLP層數		MLP層數	
	一層	兩層	一層	兩層
國語	60.0	39.7	60.0	44.7
台語	48.9	26.6	37.7	18.2
客語	54.4	30.4	61.3	39.7

3.2 混合語料實驗

此實驗在訓練集加入更多的語料，要測試的語言訓練與測試一樣是採 leave one speaker out，但會在訓練集中加上其他語言的語料，且其他語言的語料不做 leave one speaker out，例如我們要測試國語的辨識率，第一次訓練以國語 a1 語者做測試集，a2 ～ a10 語者再加上其他語言的全部語料做訓練集，反覆上述直到國語所有語者的資料都當過測試集為止，辨識率的計算方法同單一語料。每個表的第一列皆為 baseline，用於比較實驗結果。

實驗結果為表 3、表 4、表 5。從結果可看出一層的結果大多會比兩層還要好，而做完 CSHE 後，一層跟兩層的辨識率會較為接近。當使用最多訓練資料的狀況下，辨識結果都會優於 baseline，尤其是客語兩層網路的辨識率 63.8% 會相當接近一層網路的 64.3% (表5) 。 分開來看的話，一層網路使用原始資料在資料量變多後每種語言的辨識率都是下降的，但經過 CSHE 後則都有上升。兩層網路使用原始資料且增加資料量後是台語跟客語有些微的提升，經過 CSHE 後再增加資料量，辨識率都有所提升，表 4 與表 5 中，台語跟客語在使用最多訓練資料的情況下，辨識率有將近兩倍的提升。

表 3: 國語做測試集

	RAW		CSHE	
	MLP層數		MLP層數	
訓練集	一層	兩層	一層	兩層
國	60.0	39.7	60.0	44.7
國+台	52.1	39.0	62.8	58.0
國+客	51.4	42.0	63.0	59.6
國+台+客	50.9	35.6	63.5	56.4

表 4: 台語做測試集

	RAW		CSHE	
	MLP層數		MLP層數	
訓練集	一層	兩層	一層	兩層
台	48.9	26.6	37.7	18.2
台+國	44.5	32.0	52.5	40.8
台+客	48.8	30.3	52.0	34.8
台+國+客	46.4	32.0	53.1	51.4

表 5: 客語做測試集

	RAW		CSHE	
	MLP層數		MLP層數	
訓練集	一層	兩層	一層	兩層
客語	54.4	30.4	61.3	39.7
客+國	46.1	39.4	61.3	61.6
客+台	46.6	34.4	62.3	59.4
客+台+國	45.9	34.7	64.6	63.8

3.3 跨語料實驗

此實驗使用跟測試集不同語言的語料做訓練，例如測試集為國語，則訓練集就不使用國語，本次實驗不做 leave one speaker out，直接以該語言所有語料做測試，以和測試集不同語言的語料做訓練。，辨識率即是正確句數除以測試集總句數。每個表的第一列皆為 baseline，用於比較實驗結果。

實驗結果為表 6、表 7、表 8，可看出一層的結果大多都比兩層的還好。但在訓練資料使用兩種語言的語料時，辨識率會較為接近，且做完 CSHE 後一層跟兩層網路的辨識率會只有大約 2% 的差距。一層網路的部分，每種語言若使用不同語言的語料做為訓練集，不管是否有做 CSHE，辨識率都會比單一語料實驗的結果差，只有表 7 在經過 CSHE 且使用國 + 客語訓練時有約 10% 的提升，但有 CSHE 結果會比使用原始資料要來的好。兩層網路的部分，三種語言在做完 CSHE 且使用兩種語言語料作訓練後辨識率都會比單一語料實驗還要好，表示資料量的多寡對兩層網路的影響較大，但若沒在訓練集中加入和測試集同種語言的語料，辨識率還是會不太理想。

表 6: 國語做測試集

	RAW		CSHE	
	MLP層數		MLP層數	
訓練集	一層	兩層	一層	兩層
國	60.0	39.7	60.0	44.7
台	53.0	27.6	55.7	42.9
客	50.5	22.4	51.1	34.2
台+客	55.7	40.9	57.5	58.7

表 7: 台語做測試集

	RAW		CSHE	
	MLP層數		MLP層數	
訓練集	一層	兩層	一層	兩層
台	48.9	26.6	37.7	18.2
國	41.0	26.6	43.1	32.7
客	41.7	27.5	43.6	29.1
國+客	46.4	45.0	48.6	46.2

表 8: 客語做測試集

	RAW		CSHE	
	MLP層數		MLP層數	
訓練集	一層	兩層	一層	兩層
客	54.4	30.4	61.3	39.7
國	48.4	35.2	49.0	39.5
台	44.1	27.2	47.3	38.0
台+國	50.4	50.9	53.5	51.0

4 結論

比較三種實驗結果發現，混合語料在加入了其他語料做訓練且經過 CSHE 消除語者間的差異後，所有的辨識率都會高於只使用單一語料訓練的辨識結果。但在跨語料的部分，若使用不

同語言做測試，辨識率幾乎都是下降的，只有在以另兩種語言訓練時部分才有些微上升，且即使是做了 CSHE 也不能有效提升辨識率。表示在資料量沒增加的情況下，用不同語言的資料訓練通常只會降低辨識率，而即使增加了一些資料量，若沒有包含相同語言的語料，除了資料量提升不夠多外，只用與測試資料不同語言的語料訓練對辨識率的提升也沒有幫助。從這些結果可看出台灣人在情緒的表現上，不同語言間雖仍存在差異性但是不大，因此若欲辨識的語言資料量不足，適量加入其他種台灣語言的語料並消除語者間的差異，是提升辨識率的有效方法。

此外兩層的網路在訓練資料量變多的狀況下，不管是跨語料還是混合語料，辨識率常會有非常顯著的提升，甚至可以到將近兩倍。而這三種語料在不同訓練集的訓練下，國語跟客語的辨識率最高都可以達到 60% 以上，而台語最高也有 53% 。此外總結來看每種類型的實驗，辨識率最高通常是出現在資料量最多且有做 CSHE 的狀況下因此我們可以認為，增加資料量且做 CSHE 可以相當有效的提升辨識率。

表 9 為三種語言在各實驗最佳的辨識結果統整。而表 10、表 11、表 12 呈現每種語言各類情緒在最好的辨識率下的分類情形，也就是混合語料有做 CSHE 實驗的詳細分類狀況。在此以混淆矩陣表示，橫列為所屬情緒，直行為被歸分類為何種情緒，例如表 10 的第一列生氣類，共有 40 句被正確分類為生氣、4 句分為噁心、11 句分為開心、3 句分為中性。為了以此混淆矩陣評估各類情緒的辨識難易度，在此以計算每類情緒的 precision 與 recall 來做比較，precision 的算法為該情緒被正確辨識的句數除以被分類為該情緒的句數。recall 的算法為該情緒被正確辨識的句數除以該情緒總句數。從各類情緒的 recall 及 precision 可以看到，害怕的 recall 及 precision 普遍較高，比較容易辨識。而開心跟傷心的 recall 及 precision 普遍較低，較難以辨識。

表 9: 三種實驗結果比較

	單一語料	混合語料	跨語料
國	60.0	63.5	58.7
台	48.9	53.1	48.6
客	61.3	64.6	53.5

表 10: 混合語料國語做測試分類結果

情緒/分類結果	生氣	無聊	噁心	害怕	開心	傷心	中性	總數	recall
生氣	40	0	4	0	11	0	3	58	0.68
無聊	0	49	5	0	2	5	5	66	0.74
噁心	3	1	25	1	7	3	6	46	0.54
害怕	1	2	0	42	5	4	3	57	0.73
開心	21	1	4	5	32	1	3	67	0.47
傷心	0	7	12	1	3	31	3	57	0.54
中性	14	3	3	2	3	3	59	87	0.67
precision	0.51	0.78	0.47	0.82	0.51	0.66	0.72		
總句數								438	

表 11: 混合語料台語做測試分類結果

情緒/分類結果	生氣	無聊	噁心	害怕	開心	傷心	中性	總數	recall
生氣	58	0	5	0	4	0	6	73	0.79
無聊	2	23	3	0	1	13	13	55	0.42
噁心	12	2	29	2	6	4	0	55	0.53
害怕	2	1	2	43	7	2	4	61	0.70
開心	13	0	7	6	29	3	0	58	0.50
傷心	5	8	4	5	1	27	6	56	0.48
中性	19	7	8	3	8	4	15	64	0.23
precision	0.52	0.56	0.50	0.73	0.51	0.52	0.34		
總句數								422	

表 12: 混合語料客語做測試分類結果

情緒/分類結果	生氣	無聊	噁心	害怕	開心	傷心	中性	總數	recall
生氣	51	1	1	2	3	0	4	62	0.82
無聊	0	40	2	1	1	11	4	59	0.68
噁心	5	4	34	2	3	5	1	54	0.63
害怕	4	0	2	49	6	2	1	64	0.77
開心	14	0	5	3	22	1	3	48	0.46
傷心	1	5	6	6	6	24	4	52	0.46
中性	6	1	4	3	6	3	39	62	0.63
precision	0.63	0.78	0.63	0.74	0.47	0.46	0.70		
總句數								401	

綜觀以上，MLP 的辨識表現在台灣本土語料庫上是還不錯的，只要使用適當的網路結構和訓練資料組合，都能達到五、六成的辨識率。且從混合語料的實驗結果可以發現，在資料量越多的情況下，兩層網路的辨識率會越接近一層網路，可見造成兩層網路辨識率較差的主因是因為訓練資料不夠多所導致。本研究所使用的國台客語料庫，資料量只有約一千多筆，相信日後若資料量更多，辨識的結果會更好，且使用兩層網路的辨識率可能會優於使用一層網路。

參考文獻

[1] R. W. Picard and R. Picard, "Affective computing" . MIT press Cambridge, 1997, vol. 252.

[2] 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.

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

[4] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech communication, vol. 53, no. 5, pp. 768–785, 2011.

[5] D. Le and E. M. Provost, “Emotion recognition from spontaneous speech using hidden markov models with deep belief networks,” in Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. IEEE, 2013, pp. 216–221.

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

[7] 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.

[8] T. Iliou and C.-N. Anagnostopoulos, “Svm-mlp-pnn classifiers on speech emotion recognition field-a comparative study,” in Digital Telecommunications (ICDT), 2010 Fifth International Conference on. IEEE, 2010, pp. 1–6.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” DTIC Document, Tech. Rep., 1985.

[10] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro,G. S. Corrado, A. Davis, J. Dean, M. Devin et al., “Tensorflow:Large-scale machine learning on heterogeneous distributed systems,”arXiv preprint arXiv:1603.04467, 2016.

[11] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting.” Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.