基於 CNN+LSTM Model 之語音情緒識別
Speech Emotion Recognition Based on CNN+LSTM Model

Wei Mou[1], Pei-Hsuan Shen[1], Chu-Yun Chu[1], Yu-Cheng Chiu[1], Tsung-Hsien Yang[2] and Ming-Hsiang Su[1]
[1] Soochow University, Taiwan
[2] Telecommunication Laboratories Chunghwa Telecom Co. Ltd.
[1] weiswlight, claire1015069, tp973632, jean199925, huntfox.su@gmail.com
[2] yasamyang@cht.com.tw

摘要

由於智能對話助理服務的普及，語音情緒辨識已經變得越來越重要。在人與機器的溝通中，情緒辨識與情感分析能夠增強機器與人類的互動。本研究使用 CNN+LSTM 模型實作語音情緒辨識 (Speech Emotion Recognition, SER) 處理並進行預測。從實驗結果得知使用 CNN+LSTM 模型相對於使用傳統 NN 模型取得更好的效能。

Abstract

Due to the popularity of intelligent dialogue assistant services, speech emotion recognition has become more and more important. In the communication between humans and machines, emotion recognition and emotion analysis can enhance the interaction between machines and humans. This study uses the CNN+LSTM model to implement speech emotion recognition (SER) processing and prediction. From the experimental results, it is known that using the CNN+LSTM model achieves better performance than using the traditional NN model.

關鍵字：CNN、LSTM、情緒識別
Keywords: CNN, LSTM, Speech emotion recognition

1 Introduction

情緒定義為一種受到外在或內在刺激後引起的心理感受或反應 [1]。每種情緒都有其獨特的特徵：信號，生理和先前的事件。有別於「心情」的表現，「情緒」通常是起效快，持續時間短，發生率高的，因此也更能表現語者當下的反應。情緒的表現與分類，最早由 Tomkin 定義了八種情緒：驚訝、有趣、愉悅、憤怒、害怕、嫌惡、羞愧、痛苦 [2]。後續也有其他學者提出不同的分類方式，例如 Plutchik 以如同色輪一般的方式提出情緒輪的分類 [3]，如下圖一情緒輪所示，輪中接近的情緒是較為相似的，距離較遠之情緒則較無關聯性，而相對之情緒，如高興相對於悲傷則代表相反的情緒。不同的情緒如不同的顏色一般可互相混合而成。為了將情緒表現分類可視化，由 Posner、Russell 和 Peterson 學者於 2005 年提出 [4] 將其投射在一個能表現情緒相互關係二維空間中，並以表現出的愉悅程度(valence)以及激發程度(arousal)劃分成四個象限，又稱為情緒空間(Valence-Arousal space)，如 Figure 1 所示。

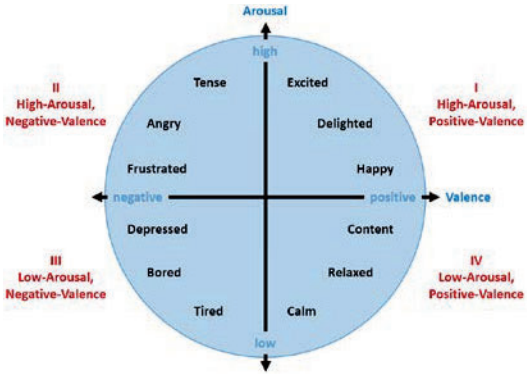


Figure 1: 情緒空間

如何自訊號中抽取利於辨識的情緒特徵，以及如何利用特徵正確辨識出情緒是重要的議題 [5-7]。常見語音訊號特徵如韻律特徵 (prosodic feature) 以及音訊頻譜特徵 (spectral features)。其中韻律特徵以音節及短語等口語斷點來計算該區段之音頻高低與聲音強度等 [5]，而音訊頻譜特徵以音框(frame) 作為音訊

訊號之抽取單位，抽取各種低階語音特徵(low-level descriptor, LLDs) 以及多個音框內低階語音特徵之統計資訊 [8]。這些特徵可能缺少對於情緒分析的客觀性 [9]，其忽略了訊號中隱含的情緒特徵，無法完整模擬人腦在做情緒辨識時所需的參考依據。近幾年，深度學習的研究日漸進步，其對於語音訊號之特徵抽取有重大的改進及貢獻，end-to-end 的特徵抽取方法，經由訓練網路層，找到輸入訊號與情緒目標之間的隱含關係，改善人為定義特徵不客觀的問題[10]，已有許多研究使用神經網路架構抽取音訊或頻譜 (spectrogram) 上之音訊特徵。Table 1 介紹使用神經網路架構的語音情緒辨識系統。

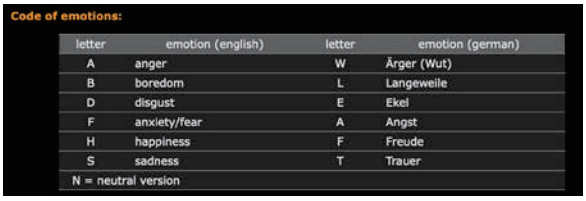
Table 1: 語音情緒辨識系統

Data	Feature Extraction	Corpus	Recognition
Spectrogram	1-layer 2D-CNN	Germany, English	LSTM [7]
Waveform	Auto-encoder	EMO-DB	SVM [11]
Waveform	2-layer 1D-CNN	eNTERFACE, MUSAN	BLSTM [12]
Spectrogram, waveform	Multi-layer 1D-CNN	Data from Cortana	CNN [13]
Waveform, Spectrogram	1-layer 1D-CNN, 1-layer 2D-CNN	NNIME	BLSTM [14]

由上述研究可以得知，現如今已有許多以神經網路進行情緒辨識之研究，他們多使用卷積神經網路 (convolution neural network, CNN) 進行特徵抽取，[7] 比較不同維度及層數的卷積層對辨識的影響，發現單層卷積層的效果較好，[7, 12, 14]皆使用長短期記憶模型(long short-term memory, LSTM)進行情緒特徵之分類，能有效處理訊號時序上的前後關係，以提升語音情緒之辨識效能。

2 Emotion Dataset

EMO-DB 資料集 [15] 是由柏林工業大學錄製的德語情感語音資料庫，由 10 位演員(5 男 5 女) 對 10 個語句 (5 長 5 短) 進行 7 種情感(中性/nertral、生氣/anger、害怕/fear、高興/joy、悲傷/sadness、厭惡/disgust、無聊/boredom，如 Figure 2 所示)的模擬得到，共包含 800 句語料，採樣頻率 48kHz (後壓縮到 16 kHz)，16 bit 量化。語料文本的選取遵從語意中性、無情感傾向的原則，且為日常口語化風格，無過多的書面語修飾。


Code of emotions:

letter	emotion (english)	letter	emotion (german)
A	anger	W	Ärger (Wut)
B	boredom	L	Langeweile
D	disgust	E	Ekel
F	anxiety/fear	A	Angst
H	happiness	F	Freude
S	sadness	T	Trauer

N = neutral version

Figure 2: 英/德語情緒對照表

語音的錄製在專業錄音室中完成，要求演員在演譯某個特定情感前，必須通過回憶自身真實經歷或體驗進行情緒的醞釀，來增強情緒的真實感。經過 20 個參與者 (10 男 10 女) 的聽辨實驗，得到 84.3% 的聽辨識別率。這個資料集經過聽辨測試後，保留 535 句(男性情感語句 233 句、女性情感句 302 句)。其中語句內容具有較高情感自由度 (包含日常生活用語的 5 個短句和 5 個長句)，但不包含某一特定情感傾向。每個檔案的命名意義如下：Position1-2 對應該人的編號、Position3-5 對應語音內容編號、Position6 對應情感編號（表一紅框處，因檔名中以德語單詞首字母標記，表一為英/德語之情緒詞語對照表）、Position7 若有兩種版本以上，則以 a, b, c 依序命名。

3 Convolution Neural Network

卷積神經網路 (Convolutional Neural Network, CNN) 由一個或多個卷積層和池化層(pooling layer) 組成。CNN 最開始的概念是經由 Hubel 等學者在生物領域上的研究而啟發 [16]，而後在 1982 年由 Fukushima 等人將神經網路的架構提出 [17]。之後 1995 年的 B. Lo 等人 [18] 與 1998 年的 Y. Lecun 等人 [19] 在神經網路的架構中加入卷積層(convolution layer)、池化層(pooling layer)等逐漸完善成現在的卷積神經網路。基本的卷積神經網路包含卷積層以卷積的方式取得局部資訊並透過激化函數做為特徵、池化層將由卷積層而來的數值進行採樣做為代表值，而最後全連接(full connection)至目標輸出。本研究使用一維單層自適應卷積神經網路架構進行聲音特徵抽取，如 Figure 3 所示。

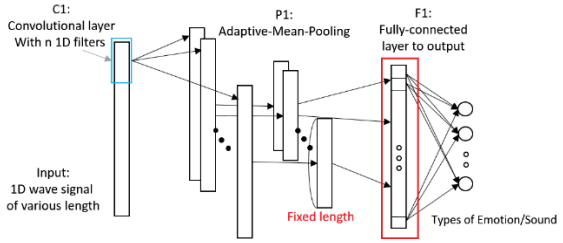


Figure 3: 一維單層自適應卷積神經網路架構

4 Long Short-Term Memory

長短期記憶神經網路 (long short-term memory, LSTM) 為一種遞歸神經網路 (recurrent neural network, RNN) 的變形。是為了解決傳統遞歸神經網路在損失函數從輸出層進行反向傳播時，可能造成梯度消失的問題，使得網路停在區域最佳解而難以學習節點間的連接關係。於是 Hochreiter 等人 [20] 提出長短期記憶單元構成的神經網路，可藉由記憶單元的特殊結構，學習到輸入間隔較長的資訊彼此的相互關係。

如 Figure 4 所示，長短期記憶單元的結構包含關鍵的細胞狀態 C_t 於圖片上方的水平線，而其中包含三個主要的閘(gate)，分別為遺忘閘、輸入閘、輸出閘，用以保護、控制細胞狀態，讓資訊選擇性的通過，而輸出皆會經過 sigmoid function(σ)使值介於 0 到 1 之間，用以描述通過的量。0 表示完全不通過、1 表示完全通過。

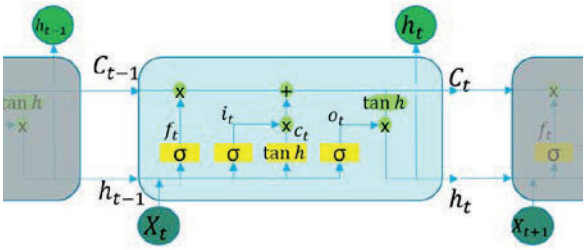


Figure 4: 長短期記憶網路

在 Figure 4 中，X_t 先經過遺忘閘，透過加權矩陣W_f和激活函數σ對上一時間的輸出與當下輸入的運算，來控制細胞狀態丟棄的資訊量。接著透過輸入閘來決定細胞狀態應該取得的新資訊，此部分有兩個步驟，第一步先由 sigmoid function 決定要更新的值。第二步通過 tanh function 決定細胞狀態的候選值c_t。得到遺忘閘與輸入閘的運算結果後，來對細胞狀態進行更新。其中前項以遺忘閘結果和舊的細胞狀態相乘決定要遺忘的資訊，後項由輸入閘結果與候選值相乘決定細胞狀態的新資訊，將兩者相加後即代表更新後的新細胞狀態。最後決定要輸出的值。由輸出閘決定要輸出多少資訊，再將細胞狀態透過 tanh function 與輸出閘結果相乘，計算出此細胞的輸出值。

$$f_t = \sigma(W_f \cdot [X_t, h_{t-1}] + b_f) \quad (1)$$
$$i_t = \sigma(W_i \cdot [X_t, h_{t-1}] + b_i) \quad (2)$$
$$c_t = tanh(W_c \cdot [X_t, h_{t-1}] + b_c) \quad (3)$$
$$C_t = f_t \times C_{t-1} + i_t \times c_t \quad (4)$$
$$o_t = \sigma(W_o \cdot [X_t, h_{t-1}] + b_o) \quad (5)$$
$$h_t = o_t \times tanh(C_t) \quad (6)$$

5 Experiment Settings and Results

本研究將數據利用 Root Mean Square normalization 進行數據歸一化，接著計算與分類各情緒中檔案個數。接著按照情緒分類，將 535 筆資料建立 20%為訓練集、64% 為測試集與 16% 的驗證集。，因為音檔轉換後的資料長度不一，最長 143652，最短為 19608，因此需要進行填補資料。於是我們將經歸一化的音頻數據進行切割，將一個音頻數據根據固定長度 (16000) 切割成數筆資料，最後不足的部分，進行補 0。最後可以每一個音頻可以形成 2 維資料以作為 CNN+LSTM 模型的輸入。

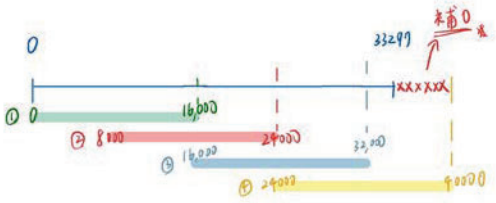


Figure 5: 音檔切割示意圖

本研究所提出的情緒識別模型是由四層 CNN、一層 LSTM 加上全連接層所建構而成。CNN 中含有 一維卷積層、批量標準化層、激活函數層、最大池化層。其中激活函數使用 Elu function(exponential linear unit)。對比 ReLu，Elu 可讓負的輸入值也有輸出值，相對較穩健。而 LSTM 中激活函數為 tanh function，編譯層

中使用優化方式為 SGD，其參數 momentum 用於 SGD 在相關方面上前進，抑制震盪，nesterov = True 表使用 Nesterov 動量。於訓練集使用 EarlyStopping 於出現過擬合或模型指標無明顯改進時提前中止訓練;利用 ModelCheckpoint 於每個訓練期後保存模型。實驗結果顯示以七種情緒進行識別，其情緒識別正確率為 57.83%，若改以四種情緒共 339 筆資料進行訓練及預測，則情緒識別正確率為 83%。

此外，本研究亦使用傳統類神經網路 (neural network, NN) 進行情緒識別，輸入資料為 1 維資料 (原本 2 維資料藉由 flatten function 轉換為 1 維資料)。藉由不同的測試資料比例，可以看出傳統 NN 模型在測試資料集比例為 20% 時，得到最佳的辨識正確率，七種情緒辨識正確率為 53.30%，而四種情緒辨識正確率為 77.90%。

	全部（七種）情緒		四種情緒	
test_size = 0.2	test	0.533	test	0.779
	random	0.477	random	0.765
test_size = 0.3	test	0.497	test	0.716
	random	0.491	random	0.676
test_size = 0.4	test	0.495	test	0.699
	random	0.519	random	0.706

Figure 6: 音檔切割示意圖

最後本研究僅使用 CNN 跟 LSTM 模型進行情緒識別，實驗結果顯示以七種情緒進行識別，其情緒識別正確率 CNN 模型為 45.80%，而 LSTM 模型情緒識別正確率為 50.50%。

6 Discussion

經過不同模型測試後，本研究發現在僅四種情緒的訓練及測試情況下，情緒辨識正確率明顯較高於七種情緒辨識正確率，我們認為可能原因為以下兩點：四種情緒間差異性較大及全部筆數較少無法有足夠的樣本進行訓練。若能再增加較多的樣本進行訓練及測試，應能提升情緒辨識正確率。

Table 2: 情緒識別正確率

Model	7 種情緒	4 種情緒
NN	53.30%	77.90%
CNN	45.80%	-
LSTM	50.50%	-
CNN+LSTM	**57.83%**	**83.00%**

7 Conclusion

本研究使用 CNN+LSTM 模型實作語音情緒辨識 (Speech Emotion Recognition, SER) 處理並進行預測。從實驗結果得知使用 CNN+LSTM 模型相對於使用傳統 NN 模型取得更好的效能。

未來可能嘗試的改進方法為將其他關於情緒辨識之開放資料與此 EMO-DB 資料共同進行訓練，可供訓練樣本增加可能會使準確度提升；另外，在資料前處理的部分對音訊嘗試進行更妥善的數據前處理方式以及不同的轉換方式測試，如傅立葉轉換等。

References

[1] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4): 169-200, 1992.

[2] Paul Ekman, Wallace V. Friesen, and Ronald C. Simons. 1985. Is the startle reaction an emotion? *Journal of personality and social psychology*, 49(5): 1416.

[3] Robert Plutchik. 1980. *A general psychoevolutionary theory of emotion*, Chapter 1 in Theories of emotion: Elsevier, pages 3-33.

[4] Jonathan Posner, James A. Russell, and Bradley S. Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3): 715-734.

[5] K. Sreenivasa Rao, Shashidhar G. Koolagudi, and Ramu Reddy Vempada. 2013. Emotion recognition from speech using global and local prosodic features. *International journal of speech technology*, 16(2): 143-160.

[6] Houwei Cao, Štefan Beňuš, Ruben C. Gur, Ragini Verma, and Ani Nenkova. 2014. Prosodic cues for emotion: analysis with discrete characterization of intonation. *Speech prosody*, 130-134.

[7] Namrata Anand and Prateek Verma. 2015. Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data. In *Technical Report*: Stanford University.

[8] Tzinis, Efthymios, and Alexandras Potamianos. 2017. Segment-based speech emotion recognition using recurrent neural networks. In *Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pages 190-195. https://doi.org/10.1109/ACII.2017.8273599.

[9] Lianzhang Zhu, Leiming Chen, Dehai Zhao, Jiehan Zhou, and Weishan Zhang. 2017. Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors*, 17(7): 1694.

[10] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proceedings of 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pages 5200-5204. https://doi.org/10.1109/ICASSP.2016.7472669.

[11] Jun Deng, Sascha Frühholz, Zixing Zhang, and Björn Schuller. 2017. Recognizing emotions from whispered speech based on acoustic feature transfer learning. *IEEE Access*, 5:5235-5246.

[12] Che-Wei Huang, and Shrikanth Shri Narayana. 2017. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *Proceedings of 2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, pages 583-588. https://doi.org/10.1109/ICME.2017.8019296.

[13] Kim, Suyoun, and Michael L. Seltzer. 2018. Towards language-universal end-to-end speech recognition. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 4914-4918. https://doi.org/10.1109/ICASSP.2018.8462201.

[14] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su, and Yi-Hsuan Chen. 2019. Speech emotion recognition using deep neural network considering verbal and nonverbal speech sounds. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pages 5866-5870. https://doi.org/10.1109/ICASSP.2019.8682283.

[15] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Proceedings of Ninth European conference on speech communication and technology*.

[16] Hubel, David H., and Torsten N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1): 106-154.

[17] Fukushima, Kunihiko, and Sei Miyake. 1982. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In *Proceedings of Competition and Cooperation in Neural Nets*, Springer Berlin Heidelberg, pages 267-285. https://doi.org/10.1007/978-3-642-46466-9_18.

[18] Shih-Chung B. Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T. Freedman, and Seong K.Mun. 1995. Artificial convolution neural network for medical image pattern recognition. *Neural Networks*, 8(7): 1201-1214.

[19] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, 86(11): 2278-2324. https://doi.org/10.1109/5.726791.

[20] Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8): 1735-1780.