國立聯合大學

資訊管理學系碩士班

碩士論文計畫書

時序式深度學習於語音訊號中的心理狀態識別研究

A study on Mental State Recognition in Temporal Deep Learning Model in Speech Signals

研 究 生: 潘嶧德 撰

指導教授: 温敏淦 博士

中 華 民 國 一 0 八 年 三 月

摘要

口說在語言學習中扮演重要的角色，尤其在外語學習時，如何透過自動化系統進行學習者學習輔助，是缺乏一對一老師引導人力時極重要的課題。相較於目前語言學習系統以發音正確性，判斷學習者對語句的使用能力，本研究利用深度學習網路中的長短期記憶遞歸類神經網路(Long Short-Term Memory，LSTM； Recurrent Neural Networks，RNN)與時序式卷積網路模型(Temporal Convolutional Network，TCN)，先進行語音情緒識別，再分析外語學習者對引導學習復誦語彙的自信度，據以動態調整其演練同一語彙的週期，達到適性化學習的目的。

本研究將具時間序列特質的語音訊號，轉換成能有效描述其特質的梅爾倒頻譜係數(Mel Frequency Cepstral Coefficient，MFCC)結合原始語音訊號作為輸入，並由學習者自行標註的語音自信度，做為類神經網路模型訓練的資料集；而在辨識階段，同樣以 MFCC 做為語音前處理的轉換步驟，再以已訓練完成的模型進行識別。此一新穎的外語輔助學習模式，預期找出一可行的時序型辨識模型，可以提供學習系統引導學習者適性學習的重要資訊。

關鍵詞：自信度辨識、遞歸類神經網路、適性輔助學習

目錄

圖目錄

第一章　緒論

情緒在人類日常生活中有著相當重要的影響，可以幫助人們進行有效溝通，促進人際間的互動。在自動化及資訊化快速發展的世代，因為自動化機器大量介入產生的人機互動，讓機器理解反映人們的真實情境成了重要的議題。許多學者努力投入使機器擁有識別人類情緒能力的相關研究，形成情緒計算(Affective Computing)此一新興的研究領域，主要在使智慧系統能夠識別、感知、判斷和解釋人類的情緒(Poria, Cambria, Bajpai, &Hussain, 2017)，是一門跨計算機科學、心理學、社會科學和認知科學等學科的研究領域。情緒計算目前的發展包含了情感分析(Sentiment Analysis)與情緒識別(Emotion Recognition)兩種不同的研究議題。

情感分析主要在對樣本進行觀感的極性(polarity)分析，一般在正負向情感間衡量出一個情緒值(valence–arousal)以表達觀感狀態，而其最多的應用在分析文本中的情緒，例如 Yukun Ma 學者將句子分析成語意網路(SenticNet)後，利用深度學習中的長短期記憶模型(Long Short Term Memory Network,LSTM)，可以有效地分析句子的情感的分數(結論)(Ma, Peng, &Cambria, 2018)。

情緒識別是情緒計算中另一研究議題，主要根據樣本的特質，分類至外顯的情緒反應，常見的有 Paul Ekman 分類的六類情緒：憤怒、厭惡、恐懼、快樂、悲傷及驚訝，或更詳細分類的：趣味、蔑視、滿足、尷尬、興奮、內疚、成就感、放鬆、滿足、愉快、羞辱等十一種情緒(Ekman &Cordaro, 2011)。情緒識別經常應用於面部表情以及言語表達等媒介，以自動判斷識別人類情緒表現的類別，例如 D.Wang 利用 IAPS (International Affective Picture System)資料圖庫中的照片，以自適應神經模糊推理系統(Adaptive Neural Fuzzy Inference System ,ANFIS)模型經過訓練後，用以辨識影像中所表現的情緒屬於快樂、悲傷、憤怒或恐懼(D.Wang et al., 2018)。Tao 學者利用 IEMOCAP(Interactive Emotional Dyadic Motion Capture) 情緒語音庫中所收集的演員語音，訓練 LSTM 模型以辨識語音中情緒表現，其可分類為憤怒、中立、悲傷、快樂四種情緒(Tao &Liu,

2018)。相較於前述使用單一情緒媒介為分類對象的方法，也有學者提出整合多種資料類型的多模態情緒識別模型，以提昇分類的正確性，Poria 學者就是使用 MOUD (Multimodal Opinion Utterances Dataset) 話語數據集中的影片資料，將影片中的影像、聲音與文字整合成訓練的樣本，利用卷積神經網路(Convolutional Neural Networks, CNN)分類生氣、快樂、悲傷和中性四種情緒(Poria, Chaturvedi, Cambria, &Hussain, 2017)。

語音交談是一種人與人交流非常自然的媒介形式，在人機互動的應用設計上有強烈的需求，除了語言識別(Speech Recognition)，語音情緒辨識(Speech Emotion Recognition，SER)因為有助於提高機器對人們情緒意圖的理解，而逐漸受到研究者的重視，例如:網路教學、互動遊戲等系統，可以依照檢測到的情緒類別差異給予使用者不同的反饋(BSchuller, Rigoll, &Lang, 2004)。Williams 認為聲音中生氣的語音反應會相對響亮、快速且有強烈的高頻能量；悲傷的情緒則會產生緩慢、低音且很少有高頻能量的語音(Williams &Stevens, 1981)，因此想要辨識出語音中的情緒跟聲音的特徵有很大的關聯性。

進行語音訊號處理，主要的聲音特徵分別為韻律特徵(音調、能量、語速等)和頻譜特徵例如:線性預測編碼(Linear predictive coding, LPC)、梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCC)等(Hansen &Hasan, 2015)，從語音信號利用濾波器轉換為可使用的數學參數。但因為不清楚哪些語音特徵在區分情緒的效果最好，很容易因為表達者個人的說話風格而影響聲學辨識的結果。而在相關研究中裡最常被使用聲學特徵為梅爾倒頻譜係數(Mao, Dong, Huang, &Zhan, 2014)，梅爾倒頻譜係數是在頻域中以梅爾刻度劃分頻帶，將各頻帶的頻率結合在一起，作為能量強度，再將能量強度轉換成倒頻譜。以往學者會利用高斯混和模型(Gaussian Mixture Model, GMM)來作為聲學特徵提取，最近則利用深度學習來代替高斯混和模型在自動化的聲學特徵提取效果更好(Lei, Ferrer, McLaren, &Scheffer, 2014)。

語音情緒中的情緒除了有 Paul Ekman 定義的那十一種情緒，還有說話者的自信度也是影響表達的關鍵因素，但關於語音自信度辨識的實證研究相當少，表達者的自信度

對於其學習與表達有很大的影響，提升一個人的自信度，可以提昇學習的效果，也會提升其他人對其的評價，甚至會間接影響一個人的成功與否，所以是一個相當重要的因素。利用機器學習來處理語音訊號的方法已經越來越受到關注，並且已經成為進行語音辨識穩定且可行的方法(Maas &Le, 2012)，因此相當適合來處理語音自信心辨識的問題。

自信度辨識可以應用在外語學習上，在全球化的世界中，外語能力算是相當重要的能力之一，在與外國人發展業務、外交、教學以及傳達新聞和訊息時，都需要利用對方的語言進行交流，而目前英文則為最多國家通用的語言，也是台灣人在學習外語的首選，在台灣從國小到大學都有安排英文學習的相關課程，擁有外語閱讀和口說的能力相當重要，但學習外語能力卻相當困難，需要花費大量的時間和精力來養成這項能力，在沒有老師或資源的狀況下，有些人就會選擇線上學習或者是使用 APP 軟體來輔助學習外語。

利用自信度辨識的技術對學習的輔助，除了告訴學習者複誦的是否正確，還可以知道學習者複誦的是否正確與是否有自信，進而決定使用者在學習的歷程上對於句子的頻率，當使用者越有自信且正確的複誦出單字，就可以降低該句子出現的頻率，利用適性化學習歷程可以幫助使用者學習，而這樣的學習方式，需要一個可以辨識語音中說話者的自信心框架。

本研究主要目的是利用時序型的模型來處理語音情緒識別問題，除了辨識語音情緒資料集，還自行收集學習者的語音與自信心標記，經過語音訊號前處理後，轉換成挑選的特徵加入訓練語音情緒模型，最後計算辨識率以驗證語音情緒模型的辨識效果，建立一個可辨識學習者自信心的模型，供未來可利用語音情緒辨識的模型，來輔助外語學習者的學習效果。

第二章　文獻探討

本研究藉由深度學習方法進行語音自信度辨識，並利用梅爾倒頻譜係與原始聲學特徵作為語音訊號，配合兩種與時序相關的模型，以建構語音情緒辨識模型。本章節將蒐集相關文獻並進行歸納與探討，內容包含自信度、語音情緒辨識、深度學習方法之研究。

第一節　自信度

在過去的文獻中有學者認為自信是一種人格維度，可以反映出一個人對於自己和環境的控制程度，並且依照個人過去的經驗已成功完成目標(Bearden &Teel, 1980)，或是認為自信是一種中介變量，利用自信度可以解釋一個人的個性和其發聲的傾向之間的關係(Chelminski &Coulter, 2007)，且會影響一個人的日常表現。且許多學習心理研究都指出自信心對於學習有重要影響，例如非母語的第二語言使用焦慮會影響學習語言的效果，有研究指出自信心可以激發和增強一個人的溝通慾望以及通過溝通實現目標的能力(Clément &Kruidenier, 1983)，自信心程度會影響第二語言的學習(Clément, Dörnyei, &Noels, 1994)。與同齡人相比自信心比較低的學習者，往往缺乏學習外語的動力(Clément, 1980)。

若能辨識出自信度，必能提升學習能力跟工作效率，但要辨識出一個人的自信度相當困難，相關的實證研究也相當稀少，從聲學分析研究發現增加聲音的響度、加快語速和不頻繁的短暫停頓可以增加表達的自信度，聲音中有較高的音高和比較大的能量波動和自信度有關(Scherer, London, &Wolf, 1973)，因此一個人的自信度是有可能從表達者的聲音中辨別的。Krajewski 學者曾做過語音自信度的分類器比較，透過五位專家投票來評估對表達者的自信心程度做為資料集，並訓練自適應增強(AdaBoost)的機器學習方法得到了 87.7%的辨識率(Krajewski, Batliner, &Kessel, 2010)。而現今深度學習方法已有大幅度的成長，在其他領域也有許多豐富的研究成果，語音自信度的辨識更是一個可行的議題。

第二節 語音情緒辨識

語音情緒辨識為透過語音訊號來分析出人的情緒分類，最早在 1999 年 Nicholson 學者認為:非人類訊息(Non-vorbal information)在人類交流中起著重要作用，除了用口語傳達的意思之外，所說的話語的方式也隱含了大量的信息，便想利用神經網路來進行語音辨識(Nicholson, Takahashi, &Nakatsu, 1999)，語音情緒識別逐漸成為一個熱門的議題。想要進行語音情緒識別，有兩個問題相當重要:用於辨識的模型和與情緒相關的語音特徵。

因此後來有許多學者提出了各種不同的方法與特徵來進行語音情緒辨識，最開始有許多學者會利用隱性馬可夫模型(Hidden Markov Model, HMM)作為分類器，HMM 作為分類器具有優於其他自身判別分類器的優點，不用進行幀長度的歸一化，並且可以通過使用狀態轉移機率來反映基本特徵的時間動態，可以通過添加速度和加速度數量來模擬短時時間的動態(Kwon, Chan, Hao, &Lee, 2003)。Nwe 學者利用 HMM 分類器與 LFPC 作為情緒特徵，可以得到 89%的辨識率(Nwe, Foo, &DeSilva, 2003)。BjörnSchuller 學者混和了神經網路和 HMM 模型，得到了 86.8%的辨識率比普通的 HMM 高了 9%的辨識率(BjörnSchuller, Rigoll, &Lang, 2003)。

因為支援向量機(Support Vector Machine, SVM)比起其他模型表現出很高的泛化能力，將輸入特徵向量變換為通常更高維度的特徵空間來解決非線性問題，通過在兩個類的邊界之間的分離平面的最佳放置來得到最好的分類結果(BSchuller et al., 2004)，並且在有限的訓練數據條件下可以具有非常好的分類性能。在各個不同的研究領域皆得到了不錯的成果，因此也有許多學者將其應用在語音情緒辨識的領域中。Kwon 利用原始語音訊號和 MFCC 進行高斯核 SVM 分類，在 SUSAS 和 AIBO 資料集分別得到 96.3%和 42.3%的辨識率(Kwon et al., 2003)。Pan 學者利用 MFCC 語音訊號進行 SVM 模型分類，在 Emo-db 語音情緒資料集中得到了 95.1%的辨識率(Pan, Shen, &Shen, 2012)。

近年因為硬體技術的大幅成長，深度神經網路(Deep Neural Networks, DNN)也跟著流行起來， DNN 是一種前饋神經網絡，其在輸入和輸出之間設計有多個隱藏層，讓它

能夠從原始特徵中學習高級特徵並有效地對數據進行分類(Han, Yu, &Tashev, 2014)。通過充分的訓練數據和適當的訓練策略，可以讓 DNN 在許多機器學習任務中表現得非常好，很適合處理語音情緒這樣複雜的問題。Mao 學者利用卷積神經網路配合稀疏自動編碼器來學習語音情緒中的局部不變特徵，對 Emo-DB 的語音情緒資料庫可以得到 88.3%辨識率，在吵雜的環境中也可以維持 78.3%正確率(Mao et al., 2014)。Fayek 學者利用卷積神經網路訓練 IEMOCAP 語音情緒數據庫得到了 64.78 %的正確率(Fayek, Lech, &Cavedon, 2017)。

在語音情緒識別的問題中取得聲學特徵是一個重要的因素，因為聲音容易受到表達者和表達的內容變化以及環境音等複雜的因素影響，如何取得穩定的聲學特徵仍然是一個困難的問題(Zeng, Pantic, Roisman, &Huang, 2009)。大部分相關研究採用低階的聲學特徵，即是直接從原始語音信號中提取特徵，例如音高、聲音強度、聲音持續時間等作為參數，再透過模型自動提取特徵，低階特徵中最有效也最常被使用的有線性預測分析(Linear predictive coding, LPC)和梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients, MFCC)兩種(Hansen &Hasan, 2015)。

線性預測編碼的基本原理是假設目前的聲音取樣值，可以由前面的 M 個取樣值已線性組合來預測，時間序列的每個樣本可以近似為原先樣本的線性組合，如公式(2-1)所示(Spratling, 2017):

$$x(i) \approx r(i) = \sum_{j=1}^{n} y_j \, x(i-j) \qquad (2\text{-}1)$$

其中 r(i)是 x(i)的估計值，n 為模型的階數決定估算使用多少先前的樣本，可以使用它們來預測信號的未來樣本。還可以使用係數來估計遺失或已經被破壞的訊號樣本。因此，LPC 在信號插值(Signal Interpolation)、信號恢復和降噪方面都有應用，且可以有效的作為訊號傳輸，是強大的語音分析技術之一(Nica, Caruntu, Toderean, &Buza, 2006)。Pathak 學者將線性預測編碼特徵進行學習神經網路做情緒識別，線性預測編碼作為特徵的整體效率為 46%，比傳統的聲學特徵參數相比，線性預測編碼作為情緒分類的特徵是

更好的選擇，且發現快樂和憤怒的情緒辨識的效率最好(Pathak &Kulkarni, 2011)。Kim 學者基於線性預測編碼進行高斯混和模型與最大事後機率(Maximum a Posteriori ,MAP)的方法提取特徵，作為情緒分類特徵比一般的線性預測編碼特徵的正確率提高了 4.4%(Kim &Clements, 2015)。除了用在語音情緒辨識，在其他領域，例如 Belean 學者利用線性預測編碼進行語音處理，結合卡爾曼濾波(Kalman filter)可以在有噪音的環境下獲得準確的分類(Belean, 2013)。

梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients,MFCC)是提取聲譜特徵最普遍的方法，其利用以人耳感覺到等量的音高變化的梅爾刻度(Mel scale)為基準，作為頻域特徵被認為比時域特徵準確(Bou-Ghazale &Hansen, 2000)。Tao 學者將梅爾倒頻譜係數結合過零率(Zero-crossing rate, ZCR)、能量的特徵，訓練改良後的深度學習長短期記憶遞歸神經網路(Long Short Term Memory Network, LSTM)框架得到 58.7%的正確率 (Tao &Liu, 2018)。K.Wang 利用梅爾倒頻譜係數結合傅立葉參數(Fourier parameter, FP)，使用支持向量機(Support Vector Machine, SVM)和貝葉斯分類器(Bayes Classifier)，分析德國數據庫(EMODB)、中文數據庫(CASIA)和中國老年人情緒數據庫(EESDB)上使用梅爾頻率倒譜係數特徵分別提高 16.2%,6.8%和 16.6%的識別率(K.Wang, An, Li, Zhang, &Li, 2015)。

線性預測編碼與梅爾倒頻譜係數皆是較常用的聲學特徵，但有研究指出人類語音是非線性的輸入，因此利用線性預測的方式較不適合語音估計，且梅爾倒頻譜係數是基於人類聽覺系統的濾波器推導出來的，會相對比線性預測編碼有更好的結果(Dave, 2013)，Kamińska 學者利用 k-平均演算法(K-means)與支持向量機分類波蘭語音情緒資料庫，梅爾倒頻譜係數準確率分別以 77.92%與 83.95%高於線性預測編碼(Kamińska, Sapiński, &Anbarjafari, 2017)。Pan 學者將基頻、能量、過零率、線性預測編碼和梅爾頻率倒譜係數以組合的方式作為特徵，分別辨識柏林(EMO-DB)，日本和泰國(LINKS)的情緒數據庫，利用支持向量機進行情緒分類，最後基頻、能量和梅爾頻率倒譜係數的組合得到 89.80％、93.57％和 98.00％正確率最高(Pan et al., 2012)。其他亦有許多語音情緒辨識相

關研究採用梅爾倒頻譜係數做為特徵，因此本研究採用梅爾倒頻譜係數作為聲學特徵。

第三節 深度學習方法

近年人工智慧成為熱門的議題，機器學習就是實現人工智慧的其中一種研究領域，機器學習被 Arthur Samuel 學者定義為使電腦無須明確程式即可學習，程式一旦建立就能學習在程式以外進行的智能活動(Samuel, 1959)，機器學習的演算法只需要通過資料的訓練來學習，就可以自行建立自定義的程式來解決各種領域的每個問題(Sze, Chen, Yang, &Emer, 2017)。機器學習中的深度學習(Deep Learning)是目前許多人工智慧應用的基礎(Y., Y., &G., 2015)，深度學習之所以稱為深度，即是因為其神經網路多於三層隱藏層以上，隱藏層的數量越多可以學習更複雜和更抽象的高級特徵。

但人工智慧的議題以及深度學習的框架早已存在，深度學習在 1980 年由 Fukushima 提出(Fukushima, 1980)，雖然深度學習被證明是一個有效的的方法，但是當時的硬體設備的計算能力有限，想要訓練完成一個模型需要大量的時間，所以發展的較為緩慢。如今硬體技術已大幅成長，高效能圖形處理器的出現極大提高了計算速度(Cireşan, Meier, Gambardella, &Schmidhuber, 2010)，縮短了訓練時間，可以更快速的得到研究結果，推動了深度學習的發展。深度學習的突破應用於語音識別(Xiong et al., 2017)和圖像識別(He, Gkioxari, Dollar, &Girshick, 2017)皆有不錯的成果，近年 Google 的團隊利用深度神經網路開發的 Alpha Go 在圍棋擊敗了著名旗手李世乭(F. Y.Wang et al., 2016)，更是引起了大家的注意。

而深度學習的設計很適合處理複雜的非線性問題(Goodfellow, Bengio, &Courville, 2016)，因為其基於深度的神經網路，利用神經節點構成網路的架構，如圖 1 深度神經網路，類似於生物的神經元。每一個節點包含輸入值利用非線性激活函數加權計算後的值，透過全連階層的連接，每個節點可以代表大量的訊息，因此多個神經網路層的深度神經網路，能夠學習輸入和輸出之間的複雜關係。在深度學習中最廣泛被使用效果也最好的是卷積神經網路(Convolutional Neural Networks,CNN)和遞歸神經網路(Recurrent Neural Network,RNN)兩種類型的模型，也是本研究中會使用到的深度模型架構。

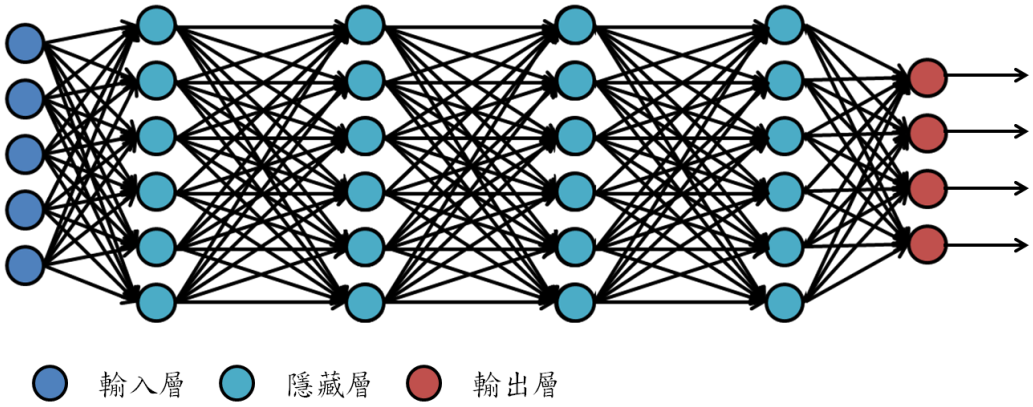

圖 1 深度神經網路

(一) 卷積神經網路(Convolutional Neural Networks,CNN)

卷積神經網路屬於深度學習的一種，雖然一般的深度學習即有很好的成效，但因為輸入資料的多維度造成運算量過大的問題，因此有學者提出可以自動抽取重要特徵的卷積神經網路。其中主要由多個卷積層和池化層(pooling layer)組成，透過這樣的結構設計使得神經網路可以輸入二維的資料結構，與其他深度學習結構相比，卷積神經網路在圖像和語音辨識能夠得到更好的結果，且使用的參數更少，訓練的時間也大幅縮短，所以算是深度學習中主要的網路架構。

在語音情緒辨識的相關研究中，亦常常使用卷積神經網路作為辨識模型，將語音訊號轉換成語頻譜圖，再利用卷積層自動提取特徵。例如 Mao 學者利用卷積神經網路配合稀疏自動編碼器來學習語音情緒中的局部不變特徵，對 Emo-DB 的語音情緒資料庫可以得到 88.3%辨識率，在吵雜的環境中也可以維持 78.3%正確率(Mao et al., 2014)。Fayek 學者利用卷積神經網路訓練 IEMOCAP 語音情緒數據庫得到了 64.78 %的正確率(Fayek, Lech, &Cavedon, 2017)。

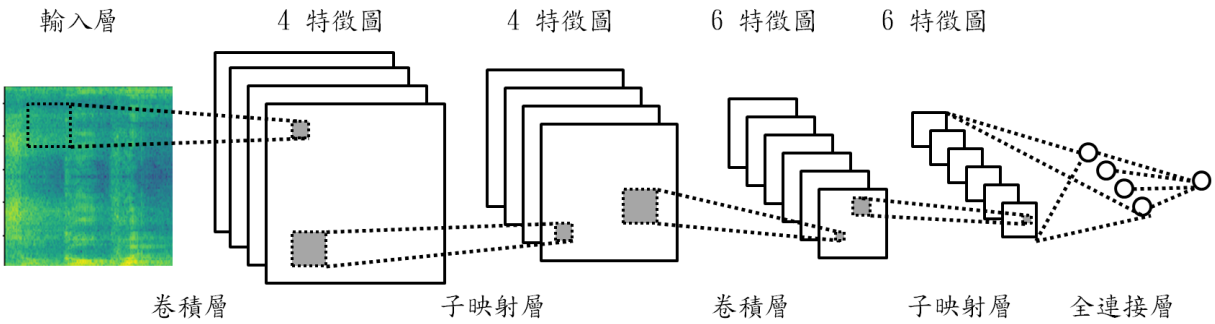

圖 2 卷積神經網路結構

(二) 遞歸神經網路(Recurrent Neural Network,RNN)

遞歸神經網路最早由 Elman 在 1990 年就已提出的時間序列的框架(Elman, 1990)，與一般的深度神經網路與卷積神經網路不同，一般的卷積神經網路的神經元只能向上一層傳播，資料的樣本在各個時刻獨立計算。而遞歸神經網路適合有順序的輸入資料，神經元的輸入可以在下一個時間點直接傳遞到自身計算，透過遞歸的方式在自身網路中傳遞。例如自然語言處理、影片、語音訊號多會比較重視前後順序的應用，多會採用遞歸神經網路處理。

但因為單純的 RNN 無法處理權重爆炸或梯度消失的問題（Vanishing gradient problem），為了解決這個問題 Hochreiter 學者在 1997 年提出了長短期記憶網路(Long Short Term Memory Network, LSTM)的改良模型(Hochreiter &Schmidhuber, 1997)，透過記憶功能來增加長期依賴(long-term dependency)，另外增加一條記憶分支，利用三個函數控制資料記憶，包含輸入閥、輸出閥、遺忘閥。改良後的 RNN 性能更好，因此 LSTM 也成為 RNN 的主要代表模型。

對於語音情緒辨識的相關研究中，因為情緒與時間先後順序有高度的相關，所以有許多相關研究亦常常使用遞歸神經網路。Tao 學者改良 RNN 提出 A-LSTM 模型在語音情緒識別，建立更好的取得時間前後關係的模型得到 58.7%的正確率(Tao &Liu, 2018)。Zhang 學者利用 LSTM 配合降噪自動編碼器，可以增加語音的特徵顯著性得到 69.1%的正確率比一般的模型性能更好(Zhang et al., 2016)。Lee 學者改良雙向長短期記憶的模型可以提升情緒辨識性得到 63.89%的正確度(Lee &Tashev, 2015)。

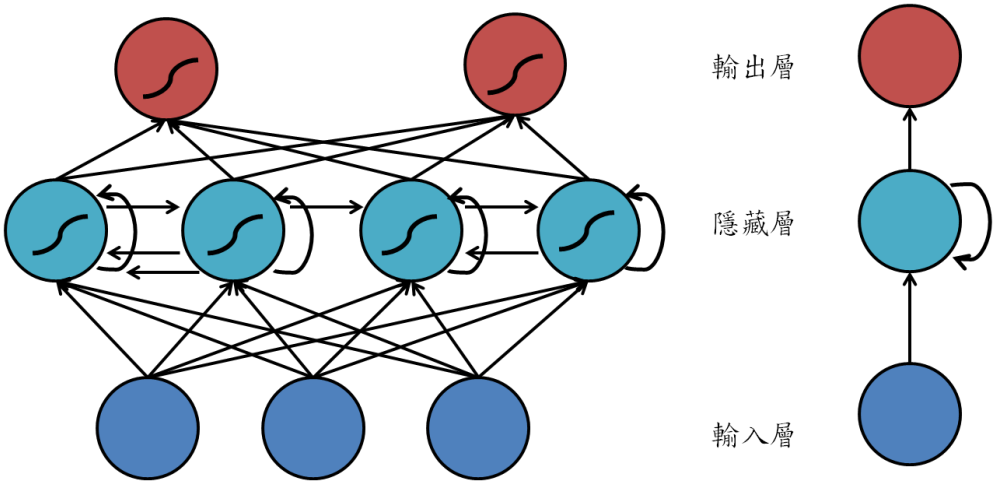

圖 3 遞歸神經網路結構

第三章　研究設計

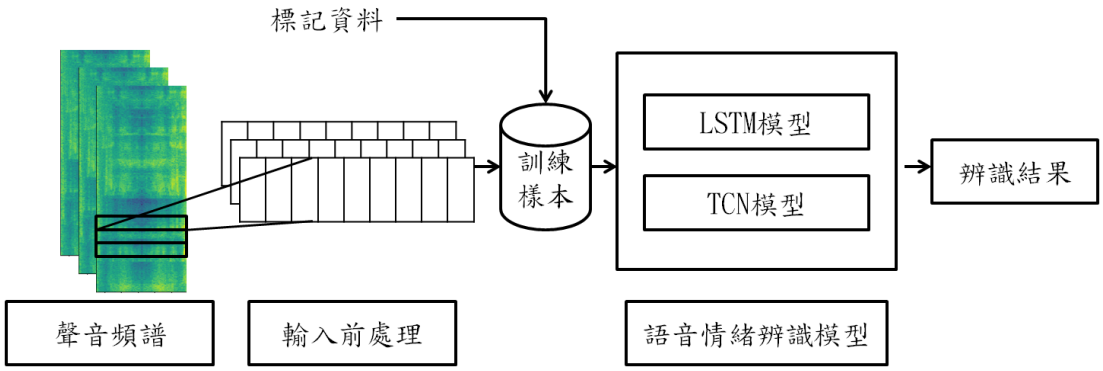

圖 4 模型訓練流程

研究設計將收集的語音進行前處理轉換成語音訊號，切割成一個一個固定大小的訊號框，再與標記資料一同作為訓練樣本加入長短期記憶神經網路與時序式卷積神經網路進行自信心辨識模型訓練，最後取得辨識結果以評估辨識率，如圖 4 所示。

第一節　聲音訊號前處理

將一般人所能聽到的聲音轉換成電腦可以使用語音訊號，總共提取 115 個特徵，每一個音框提取:基頻(Fundamental Frequency , F0)5 個、聲音的能量 5 個和 105 個 Mel 倒頻譜系數(MFCC)作為語音訊號，這些參數的組合被證明可以得有效的分辨率(Pan et al., 2012)，例如:聲音的能量在當人對所說的話有自信時會較高、基頻聲音的穩定性在沒自信的人時會不穩定等，因此選擇這些係數作為情緒辨識特徵。

聲音可以分解為多個正弦波，而基頻就是週期波形中最低的頻率，縮寫為 f0，表示從零開始計數的最低頻率。通常以對數尺度來計算，而不是使用線性尺度，以匹配人類聽覺系統的分辨率。

能量即為聲音的響度，它可以簡單地從時間窗口內的語音樣本計算得出，如(3-1)所示(Pan et al., 2012):

$$E_v = \sum_{n=1}^{N} s_n^2 \qquad (3\text{-}1)$$

梅爾倒頻譜係數(Mel-Frequency Cepstral Coefficients,MFCC)基於人耳聽覺的特徵，其利用以人耳感覺到等量的音高變化的梅爾刻度(Mel scale)為基準。快速傅立葉變換(Fast Fourier Transform, FFT)算法將每個幀(frame)樣本從時域轉換到頻域，如(3-2)中所述:

$$S[k] = \sum_{n=1}^{N-1} s[n].e^{\frac{j2\pi nk}{N}}, 0 \le k \le N-1 \quad (3\text{-}2)$$

梅爾濾波器組由重疊的三角濾波器組成，用以得到梅爾刻度，截止頻率由兩個相鄰濾波器的中心頻率決定，能得到每一個濾波器輸出的對數能量，對數具有將乘法變為加法的效果，如(3-3)中所述:

$$F[k] = \log\left(\sum_{k=1}^{N-1} |\tilde{x}[j]|^2 H_m[k]\right), 0 \le m \le M \quad (3\text{-}3)$$

最後，計算對數濾波器組輸出的能量進行離散餘弦變換(Discrete Cosine Transform, DCT)以得到梅爾倒頻譜係數，如(3-4)中所述:

$$C[n] = \sum_{n=1}^{M} F[m]\cos\left(\frac{\pi n(m-1)}{2M}\right), 0 \le n \le M \quad (3\text{-}4)$$

C[n]為語音訊號的梅爾倒頻譜係數，本研究總共使用 115 個特徵，每一個音框提取:基頻 5 個、聲音的能量 5 個和 105 個 Mel 倒頻譜系數(MFCC)作為語音情緒辨識的語音訊號輸入。

第二節 語音情緒辨識

本研究將使用長短期記憶神經網路與時序式卷積神經網路並應用於語音情緒辨識。進而提出一個語音情緒辨識模型以協助學習外語時的自信度分析，並利用分析結果配合學習，進而提升學習效率。本章將介紹長短期記憶神經網路與時序式卷積神經網路的方法。

(一) 長短期記憶網路模型(Long Short Term Memory Network, LSTM)

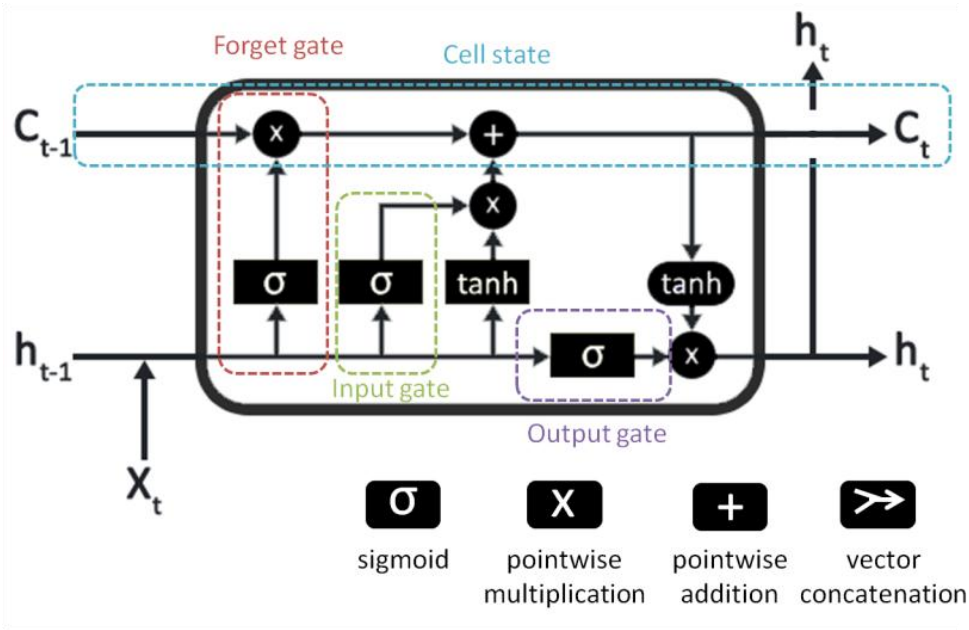

圖 5 長短期記憶層內部工作圖

遞歸神經網路以 DNN 為基礎，傳統的 DNN 僅能以個別的資料作為訓練輸入，但對於語言、情緒的前後輸入有時間效果且互相有關聯性的資料，則無法有效的訓練，這個問題可以透過遞歸神經網路(RNN)模型來解決，本研究採用遞歸神經網路(RNN)中流行的長期的短期記憶(LSTM)網路(Yang, Tao, Wen, Li, &Chao, 2015)，因為其適合處理和預測時間序列間隔較長的問題，與本研究的語音情緒辨識需求相符，LSTM 的公式如下:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{3-5}$$

$$i_t^{(l)} = \sigma(W_{xi}X_t^{(l-1)} + W_{hi}X_{t-1}^{(l)} + W_{ci}X_{t-1}^{(l)} + b_i^{(l)}) \tag{3-6}$$

$$f_t^{(l)} = \sigma(W_{xf}X_t^{(l-1)} + W_{hf}X_{t-1}^{(l)} + W_{cf}X_{t-1}^{(l)} + b_i^{(l)}) \tag{3-7}$$

$$c_t^{(l)} = f_t c_{t-1}^{(l)} + i_t \tanh(W_{Xc}X_t^{(l-1)} + W_{hc}X_{t-1}^{(l)} + b_i^{(l)}) \tag{3-8}$$

$$o_t^{(l)} = \sigma(W_{Xo}X_t^{(l-1)} + W_{ho}X_{t-1}^{(l)} + W_{co}X_t^{(l)} + b_o^{(l)}) \tag{3-9}$$

$$h_t^{(l)} = o_t^{(l)} \tanh(c_t^{(l)}) \tag{3-10}$$

在公式(3-5)其中σ是激活函數，以及 i,f,o 和 c 變數分別代表輸入門、忘記門、輸出門和激活向量，它們全部與向量 h 大小相同。從細胞到各門的權重矩陣是對角線，使得每個門(gate)向量中的元素都是細胞相同元素的輸入(Graves, Mohamed, &Hinton, 2013)，在此 LSTM-RNN 模型使用的 LSTM 層節點皆為 100 個，第一個 LSTM 層會回傳所有序列結果，第二個 LSTM 則只回傳最後一個序列的結果。

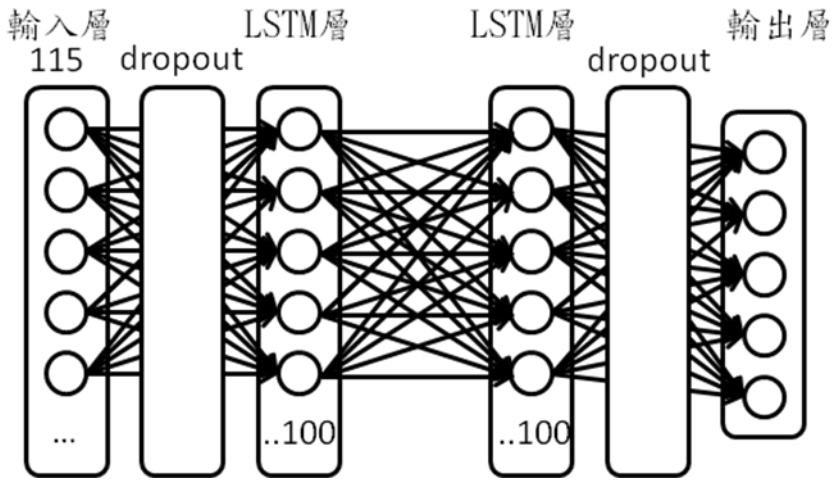

圖 6 長短期記憶網路模型架構

本研究使用的長短期記憶網路語音情緒辨識模型，包含一個輸入層 115 個特徵、兩個 LSTM 層各包含 100 個節點，第一個 LSTM 層會回傳全部結果，第二個僅回傳最後一個結果，在中間設置 dropout 為採樣率為 0.2，和一個輸出層線性輸出包含 5 個節點，分別對應 1 星到 5 星，如圖 5 顯示了長短期記憶網路的模型架構圖。

(二) 時序式卷積網路模型(Temporal Convolutional Network, TCN)

時序式卷積網路模型為在一般卷積神經網路中改良為因果卷積並加入擴張卷積和殘餘連接等元素，讓卷積神經網路也可以處理時序型的資料，Bai 比較了 MNIST、JSB Chorales 與 PennTreebank (PTB)等多種不同資料分類問題，比較了 LSTM、GRU(Gated Recurrent Unit)和 RNN 三種不同時序型的深度學習模型，都得到時序式卷積神經網路比遞歸神經網路正確率更高、更少的記憶體空間(Bai, Kolter, &Koltun, 2018)，且可以得到

更有效的長期記憶。時序卷積神經網路有幾個特性，其中的卷積層由因果卷積層組成，會依照資料間的因果關係卷積，每次卷積的時間點不會含有未來的輸入資料，僅包含現在與過去的資料。且因果卷積層網路的輸入與輸出為相同的長度，為了使用深度的網路增加了擴張卷積(Dilated Convolutions)和殘差層(Residual Layers)的組合。

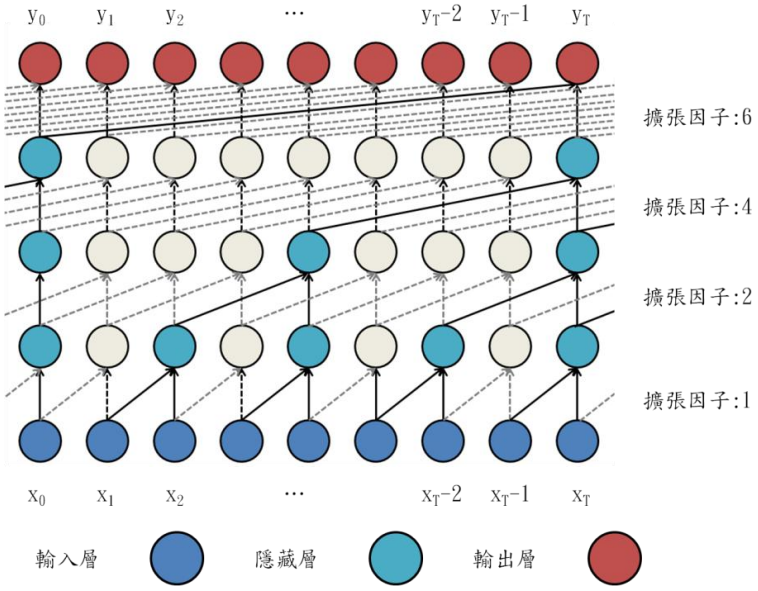

圖 7 因果卷積層內部架構

擴張卷積的公式如公式(3-11)所示(Bai et al., 2018):

$$F(s) = (X * d\ f)(T) = \sum_{i=0}^{k-1} f(i) \cdot X_s - d \cdot i \tag{3-11}$$

其中 d 為擴張係數，k 為濾波器的大小，$X_s - d \cdot i$表示過去的方向。在每一個濾波器之間加入一個固定的間隔，擴張卷積的設計可以讓輸出包含更廣泛的輸入，從而有效地擴展卷積網路的感受野(Receptive Field)。

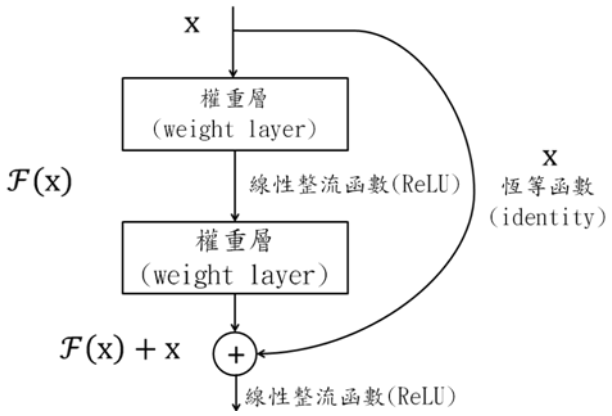

圖 8 殘差層設計

殘差層的公式如公式(3-12)所示(Wu, Zhong, &Liu, 2017):

$$o = \mathrm{RelU}(x + \mathcal{F}(x)) \quad (3\text{-}12)$$

殘差塊多出一個分支將 F 計算的結果加上輸入的 x，利用殘差塊的機制可以讓學習對恆等映射(Identity Mapping)修改而不是整個轉換，這已經被證明對比較深的神經網路有幫助。

本研究使用的時序式卷積網路語音情緒辨識模型，包含一個輸入層 115 個特徵，卷積層中使用的內核大小為 7(以 1, 2, 4, 8, 16, 32, 64 為擴張間隔)、過濾器的大小為 25，在中間設置 dropout 為採樣率為 0.05，和一個輸出層線性輸出包含 5 個節點，分別對應 1 星到 5 星。

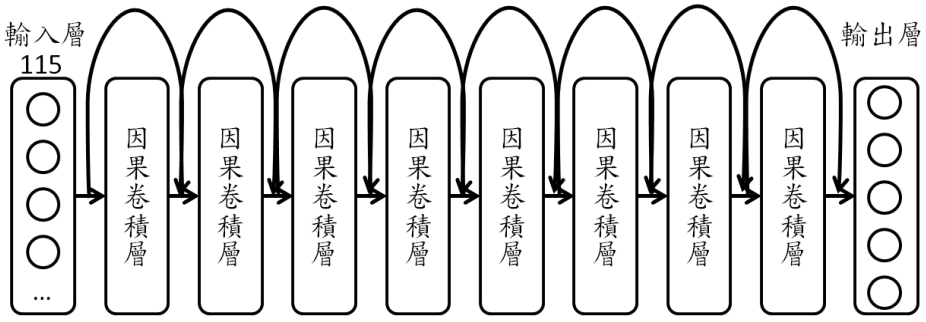

圖 9 時序式卷積網路模型架構

第四章　實驗流程

此實驗主要目的是驗證語音情緒辨識模型的可行性，先利用語音情緒辨識常用的IEMOCAP 語音情緒資料集驗證正確率，並自行收集學習者的語音與自信度標記，經過語音訊號前處理後，轉換成挑選的特徵加入訓練語音情緒模型，最後計算辨識率以驗證語音情緒模型的辨識效果，建立一個可辨識學習者自信度的模型，供未來可利用語音自情緒辨識的模型，來輔助外語學習者的學習效果。

第一節　資料集

(一) 語音情緒資料集

IEMOCAP(Interactive Emotional Dyadic Motion Capture database)由南加州大學（University of Southern California ,USC）的語音分析和解釋實驗室(Speech Analysis and Interpretation Laboratory, SAIL)收集。IEMOCAP 資料集包含十個演員，分為五組每一組為兩個演員的對話，其中包括影像的臉部和頭部的標記(Busso et al., 2008)。但本研究僅使用該資料集中的語音資料，IEMOCAP 屬於表演的引發型情緒，分為快樂、憤怒、悲傷、沮喪和中性五種情緒，資料集包含了大約 12 小時的數據，每一段語音持續時間為 2-5 秒，IEMOCAP 經常被使用於語音情緒辨識的研究中，但仍有許多可以改善的地方，作為本研究的情緒資料集。

(二) 語音自信度資料集

本研究邀請 145 受測者參與測試，為了避免其他因素影響聲學資訊的辨識度，分別在年齡、教育程度等進行變因控制：邀請男性 25 位與女性受測者 98 位，可分開驗證是否性別會影響測試結果；參與者為大學學生其教育程度一致；平均年齡為 20 歲且控制在正負 2 歲。

參與者利用手機作為媒介進行錄音。因為大部分台灣人學習外語主要選擇英文，所以選擇英文做為測試語言，本研究播放英文短句 25 句，其中避免因為英文難度影響結

果，挑選的 25 個短句由簡單到難的單字皆有，先確認參與者是否認識該單字並進行測驗，並請參與者唸出該單字，參與者也可以選擇聆聽語音。最後依照參與者的作答狀態，評估其對唸英文的自信度，衡量自信度由評級 1(較無自信)到 5(非常有自信)的範圍進行評定，本實驗收集 3,625 個受測者複誦的單字錄音，以及對其自信度的評估值，作為本研究的自信度資料集。

第二節 聲音樣本前處理

本研究使用語音情緒資料集以及自信度資料集，為了保留一句話的完整性與所有特徵，前處理不去除無聲音的部分，並從聲音中擷取 115 個聲學特徵，包含基頻 5 個、聲音的能量 5 個和 105 個 Mel 倒頻譜系數(MFCC)，以 16000 樣本率(sample rate)，將語音訊號進行音框(frame)切割，音框視窗長度為 25ms 切割，計算特徵參數，並在移動音框時將視窗重疊(overlap) 10ms，每一個聲音樣本包含 115 個特徵，取兩秒作為模型訓練資料輸入。

第三節 模型訓練

本研究利用 Google 公司所提供的開源軟體 tensorflow 與開源的 keras 類神經網路庫，建立長短期記憶神經網路與時序式卷積神經網路模型，模型設置採用完全連接的類神經網路，輸入層使用前處理提取的 115 個係數。在訓練期間，迭代(epoch)設為 10000 次，輸出層為 5 個節點，分別為情緒對應的快樂、憤怒、悲傷、沮喪和中性五種情緒和自信度的從無自信到有自信，以最高分的分類為該語音的分類結果。

第四節 辨識率評估

在本研究中，為了驗證自信度辨識的辨識率，我們將 IEMOCAP 情緒語音資料集和收集的自信度語音資料加入模型訓練，採用留一驗證(leave-one-out Cross Validation)，從所有的資料樣本分為 5 份，每次抽 20%的資料作為測試資料，進行 5 次試驗，每次採用第 N 份做為測試資料，N 為第幾次實驗，其他則作為訓練資料，分別加入長短期記憶神經網路與時序式卷積神經網路模型進行模型訓練，再將該次選取的測試資料來測試模

型，留一驗證的方法已經被證明能夠避免模型過度擬合(overfitting)，且適合小量數據集的資料測試。最後使用準確率(Accuracy)評估模型的成效，準確率之定義為成功預測數/總預測次數。

第五章　預期成果

本研究利用長短期記憶神經網路與時序式卷積神經網路兩種類神經網路來分析語音自信度，使用學習者學習外語的口說進行研究，設計出可以辨識使用者對自己自信度評估的模型。在前處理階段將聲學訊號轉換成 MFCC 語音值與取得原始訊號，用以訓練辨識的長短期記憶神經網路與時序式卷積神經網路模型，並評估兩種模型何者更適合用於語音自信度評估。本研究貢獻主要包含三大部分，分別為標示自信度語音資料集建置、聲學訊號前處理、及長短期記憶神經網路與時序式卷積神經網路語音情緒辨識模型的訓練建置。未來可進一步以不同的語言、對象及短語探討，進一步觀察類神經模型的適應性，預期能獲得準確的語音情緒辨識的模型，以提供適性輔助外語學習系統更有效的學習引導模式建立。

參考文獻

Bai, S., Kolter, J. Z., &Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *ArXiv Preprint ArXiv:1803.01271*.

Bearden, W. O., &Teel, J. E. (1980). An Investigation of Personal Influences on Consumer Complaining. *Journal of Retailing*.

Belean, B. (2013). Comparison of formant detection methods used in speech processing applications. In *AIP Conference Proceedings*.

Bou-Ghazale, S. E., &Hansen, J. H. L. (2000). A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech under Stress. *IEEE Transactions on Speech and Audio Processing*.

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*.

Chelminski, P., &Coulter, R. A. (2007). The Effects of Cultural Individualism and Self-Confidence on Propensity to Voice: From Theory to Measurement to Practice. *Journal of International Marketing*.

Cireşan, D. C., Meier, U., Gambardella, L. M., &Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*.

Clément, R. (1980). Ethnicity, contact and communicative competence in a second language. *Social Psychology and Language*.

Clément, R., Dörnyei, Z., &Noels, K. A. (1994). Motivation, Self-confidence, and Group Cohesion in the Foreign Language Classroom. *Language Learning*.

Clément, R., &Kruidenier, B. G. (1983). ORIENTATIONS IN SECOND LANGUAGE ACQUISITION: I. THE EFFECTS OF ETHNIC TY, MILIEU, AND TARGET LANGUAGE ON THEIR EMERGENCE. *Language Learning*.

Dave, N. (2013). Feature Extraction Methods LPC , PLP and MFCC In Speech Recognition. *International Journal for Advance Research in Engineering and Technology*.

Ekman, P., &Cordaro, D. (2011). What is meant by calling emotions basic. *Emotion Review*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*.

Fayek, H. M., Lech, M., &Cavedon, L. (2017). Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks*.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*.

Goodfellow, I. J., Bengio, Y., &Courville, A. (2016). Regularization for Deep Learning. In *Deep Learning*.

Graves, A., Mohamed, A. R., &Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Han, K., Yu, D., &Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Hansen, J. H. L., &Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*.

He, K., Gkioxari, G., Dollar, P., &Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*.

Hochreiter, S., &Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*.

Kamińska, D., Sapiński, T., &Anbarjafari, G. (2017). Efficiency of chosen speech descriptors in relation to emotion recognition. *Eurasip Journal on Audio, Speech, and Music Processing*.

Kim, J. C., &Clements, M. A. (2015). Formant-based feature extraction for emotion

classification from speech. In *2015 38th International Conference on Telecommunications and Signal Processing, TSP 2015*.

Krajewski, J., Batliner, A., &Kessel, S. (2010). Comparing multiple classifiers for speech-based detection of self-confidence - A pilot study. In *Proceedings - International Conference on Pattern Recognition*.

Kwon, O., Chan, K., Hao, J., &Lee, T. (2003). Emotion Recognition by Speech Signals. *Conference: 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland*.

Lee, J., &Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Lei, Y., Ferrer, L., McLaren, M., &Scheffer, N. (2014). A deep neural network speaker verification system targeting microphone speech. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Ma, Y., Peng, H., &Cambria, E. (2018). Targeted Aspect-Based Sentiment Analysis via Embedding Commonsense Knowledge into an Attentive LSTM. *AAAI-2018*.

Maas, A., &Le, Q.V. (2012). Recurrent Neural Networks for Noise Reduction in Robust ASR. *Interspeech*.

Mao, Q., Dong, M., Huang, Z., &Zhan, Y. (2014). Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*.

Nica, A., Caruntu, A., Toderean, G., &Buza, O. (2006). Analysis and synthesis of vowels using MATLAB. In *2006 IEEE International Conference on Automation, Quality and Testing, Robotics, AQTR*.

Nicholson, J., Takahashi, K., &Nakatsu, R. (1999). Emotion recognition in speech using

neural networks. In *ICONIP 1999, 6th International Conference on Neural Information Processing - Proceedings*.

Nwe, T. L., Foo, S. W., &DeSilva, L. C. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*.

Pan, Y., Shen, P., &Shen, L. (2012). Speech emotion recognition using support vector machine. *International Journal of Smart Home*.

Pathak, S., &Kulkarni, A. (2011). Recognizing emotions from speech. In *ICECT 2011 - 2011 3rd International Conference on Electronics Computer Technology*.

Poria, S., Cambria, E., Bajpai, R., &Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*.

Poria, S., Chaturvedi, I., Cambria, E., &Hussain, A. (2017). Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Proceedings - IEEE International Conference on Data Mining, ICDM*.

Samuel, A. L. (1959). Some Studies in Machine Learning. *IBM Journal of Research and Development*.

Scherer, K. R., London, H., &Wolf, J. J. (1973). The voice of confidence: Paralinguistic cues and audience evaluation. *Journal of Research in Personality*.

Schuller, B., Rigoll, G., &Lang, M. (2003). Hidden Markov model-based speech emotion recognition. In *Proceedings - IEEE International Conference on Multimedia and Expo*.

Schuller, B., Rigoll, G., &Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, p. I-577).

Spratling, M. W. (2017). A review of predictive coding algorithms. *Brain and Cognition*.

Sze, V., Chen, Y. H., Yang, T. J., &Emer, J. S. (2017). Efficient Processing of Deep Neural

Networks: A Tutorial and Survey. *Proceedings of the IEEE*.

Tao, F., &Liu, G. (2018). Advanced LSTM: A Study about Better Time Dependency Modeling in Emotion Recognition. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*.

Wang, D., He, T., Li, Z., Cao, L., Dey, N., Ashour, A. S., …Shi, F. (2018). Image feature-based affective retrieval employing improved parameter and structure identification of adaptive neuro-fuzzy inference system. *Neural Computing and Applications*.

Wang, F. Y., Zhang, J. J., Zheng, X., Wang, X., Yuan, Y., Dai, X., …Yang, L. (2016). Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA Journal of Automatica Sinica*.

Wang, K., An, N., Li, B. N., Zhang, Y., &Li, L. (2015). Speech emotion recognition using Fourier parameters. *IEEE Transactions on Affective Computing*.

Williams, C. E., &Stevens, K. N. (1981). Vocal correlates of emotional states. *Speech Evaluation in Psychiatry*.

Wu, S., Zhong, S., &Liu, Y. (2017). Deep residual learning for image steganalysis. *Multimedia Tools and Applications*.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., …Zweig, G. (2017). Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio Speech and Language Processing*.

Y., L., Y., B., &G., H. (2015). Deep learning. *Nature*.

Yang, M., Tao, J., Wen, Z., Li, Y., &Chao, L. (2015). Long Short Term Memory Recurrent Neural Network based Multimodal Dimensional Emotion Recognition.

Zeng, Z., Pantic, M., Roisman, G. I., &Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence.

Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., &Schuller, B. (2016). Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.