**背景、目的及重要性**

　　語言（language）長久以來是人與人之間最自然且最方便的溝通方式，而語音（speech）指的是藉由口述來表達語言的內涵。隨著數位電子科技的蓬勃發展以及無線通訊與網際網路的創新普及，傳統的鍵盤、滑鼠已不能滿足現代人們的需求，而語音控制成為現在正蓬勃發展的前端技術之一。舉凡：1）語音輸入 2）聲紋辨識 3）聲音合成，在人類與機器間的溝通上，越來越被廣泛應用。在絕大部分的情況下，語音比肢體語言更能明確表達語者所要傳遞的訊息。當語音被正確地轉換成語言時，我們才能正確地理解語者所要傳達的內容或概念。因此，語音辨識技術對語言、語音之間的轉換扮演著相當重要的角色。本研究的資料來源為Google於Kaggle平台提供的多個英文單字的語音資料，該語音資料集中的每個單字皆為不同語者，其時長皆在一秒以內，且每個單字都包含兩千筆以上的資料。本研究著重於英文數字的語音辨識，故僅使用英文數字0到9作為資料集進行研究。

文獻回顧

卷積神經網路（Convolution Neural Network, CNN）在近代的深度學習模型當中，多被應用於圖像辨識，例如LeCun於1998年所發表的[1]，奠定了卷積神經網路的重要基礎；Alex所提出的多層卷積神經網路架構AlexNet[2]，奪得了2012年的ImageNet冠軍；以及考慮了殘差，進而被創造出來的ResNet[3]，更是終結了ImageNet這場行之有年的圖像辨識比賽。雖說如此，由於卷積神經網路的根本概念是將資料點p的資訊與其周圍k個資訊納入考量，此一特性若運用於時間序列資料上，亦會考慮時間數列在時間點t與前後n個時間點彼此之間的時間相關性，故使用卷積神經網路對語音資料進行處理的研究亦不在少數。在卷積神經網路尚未十分成熟時，便有[4]所提出的類卷積神經網路，運用於語音辨識上的實際例子，近幾年也有利用卷積神經網路對語音進行辨識的研究[5]。

由於考慮到時間前後的相關性，亦有多位研究者投入使用循環神經網路（Recurrent Neural Network, RNN），作為時間序列資料預測模型的核心技術。雖說循環神經網路能夠考慮時間序列資料的時間性，但缺點在於其記憶時間十分短暫，也因此長短期記憶網路（Long Short-Term Memory, LSTM）[6]，作為循環神經網路的改良版，更加廣泛的被應用於實務上。多層的神經網路所堆疊的架構，往往能帶來更為優良的結果，這在ImageNet比賽上已獲得了證明[2][8][3]，因此近代的長短期記憶網路，常會使用多層的長短期記憶網路架構[9]，而其往往也會有比較好的結果。此外，雙向的長短期記憶網路（Bidirectional LSTM, BiLSTM）[10]，亦為長短期記憶網路的重要架構，其於Seq2Seq上的應用[11]，一再地表明了同時使用向前與向後的時間序列資料，能使得模型更加準確。

若談論到自然語言處理（Natural Language Processing, NLP），不得不提到Encoder-Decoder模型[12]。Encoder-Decoder模型在提出後，迅速獲得高度關注，由於其可以將一不固定長度的時間數列資料投射至一固定長度的時間數列資料上[9]，且能透過梯度下降法（Gradient Descent）同時進行模型訓練，因此相對於以往的循環神經網路，Encoder-Decoder能使時間數列資料有更多發展性。隨後，Attention模型[13][14]隨著Encoder-Decoder模型的問世，因運而生。由Bahdanau與Luong先後提出的Attention模型架構，使自然語言處理獲得了重大的躍進。該架構先是使用了Encoder-Decoder模型作為基礎，再將Encoder的各個輸出的資訊集中彙整，並作為輸入的一部份使用於Decoder之中，解決了長短期記憶網路在過長的時間序列上，時間數列資料訊息消失問題。日後，Attention被應用於[15]，並且很好的對齊了語音資料與各個英文字母，結果十分顯著，再次檢證了Attention模型在Encoder-Decoder模型上的重要性。

　　語音辨識系統包含了兩個部份，第一部份是特徵擷取（Feature Extraction），第二是分類器的設計。如果特徵擷取得到的特徵向量可以保留重要的成分或可以帶出很高的鑑別力（Discriminability），如此分類器就可以使用較簡單的方法來作分類，自然地分類的結果也會比較精準。

流程圖

**研究方法**

　　本研究探討語音辨識系統的語音前處理、音素（phoneme）對齊和字詞（lexicon）辨識。所有辨識系統皆對前處理有相當大的倚重，語音更是如此。由於語音訊號（signal）會受到諸多因素影響，例如：語者聲調、說話語速、外在雜訊等等，使得語音的處理相對地複雜與困難。普遍上，語音的處理以著名的梅爾倒頻譜係數（Mel-frequency Cepstral Coefficients, 以下簡稱MFCC）作為主流的語音特徵擷取方法，該方法以人的聽覺感知頻率作為語音特徵擷取的對象，配合其一階與二階時間軸導數（Time Derivatives），得出39維的MFCC語音特徵，可以在語音辨識上得到不錯的效果。

**前處理**

本研究以MFCC作為語音前處理的主要方法，對所有語音資料皆進行了處理，而處理後的資料再進行主成分分析（Principal Component Analysis，以下簡稱PCA）降維處理；對於類神經網路（Neural Network），本研究並沒有進行PCA處理，而是直接以MFCC處理後的資料作為輸入，並進行辨識。

梅爾頻率倒頻譜係數

　　在語音訊號分析中，由於訊號在時域中變化快速，不容易看出訊號的特性，所以通常會將時域的訊號透過快速傅立葉轉換（Fast Fourier Transformation）投射到頻域中，利用頻譜來觀察語音訊號的特性。而倒頻譜係數就是將頻譜中的各個頻域成分取對數能量，再經過逆傅立葉轉換回到時域中的參數。其中，梅爾濾波器組（Mel-triangular filter banks）是根據人耳聽覺系統所產生出來的參數，因此被廣泛應用於語音辨識和語者辨認。

　　MFCC在頻域中使用梅爾刻度（Mel scale）劃分頻帶，將梅爾濾波器所計算出來的值，取對數能量，作為該聲音片段的能量強度，併入濾波器組得出的值，最後經過離散餘弦轉換（Discrete Cosine Transform, DCT）的運算，轉換成倒頻譜，得出MFCC的值。

　 MFCC主要有七個對語音訊號的運算過程，包括預強調（pre-emphasis）、音框化（framing）、窗化（windowing）、快速傅立葉轉換（FFT）、梅爾濾波器組（Mel-triangular filter banks）、對數能量（log energy）及離散餘弦轉換。

**預強調**

　　在發聲的過程中，聲帶和嘴唇產生高頻衰減的效應，所以我們聽到的聲音就像是經過了一個低通濾波器，利用一個參數對聲音訊號進行高頻補強，來補償高頻衰減的部分，此濾波器表示為

將時間點t的輸入訊號s[t]通過高通濾波器後，在時域表示為

其中α介於0.95和0.98之間，為預強調後的訊號。

**音框化**

　　由於語音訊號變化快速，而短時間內的語音訊號通常相對穩定，因此通常將數個語音訊號作為一個觀測單位，這個單位稱為音框（frame），再根據音框內的訊號進行分析，為了避免相鄰的音框變化過大，因此會讓相鄰的音框有一段重疊區域。本研究使用0.025秒的音框，每次移動0.01秒，得出99個時間段的MFCC特徵。

**窗化**

　　為了減少音框左右兩端信號的不連續，及保留語音訊號的完整性，會將每一個音框乘上一個視窗，本研究使用的是漢明窗（Hamming Window），其會凸顯音框中間的資訊，並縮減音框兩端的資訊，使得每個音框與下一個音框的資訊較為連續，漢明窗表示為

其中為音框內的訊號個數。

**快速傅立葉轉換**

　　由於語音訊號在時域（time domain）上變化快速，不容易看出訊號的特性，所以通常會將它轉到頻域（frequency domain）上觀察能量分布，不同的能量分布，就擁有不同的語音特性。所以在窗化後，每個音框經過快速傅立葉轉換來得到頻域上分布。

（3.4）

將尤拉公式代入（3.4）式，可以得到

（3.5）

將（3.5）式的實部與虛部取平方後，相加得到頻域能量

**梅爾濾波器組**

　　人耳對不同頻率的感受程度，並非在所有頻域都是一樣的。一般來說，人耳對低頻有比較高的敏感度，也就是在低頻時可以分辨較細微的頻率差異，在1kHz頻率以下的臨界寬度約為100Hz，1kHz以上的臨界寬度則是成指數成長。因此，若是聲音在某個範圍變動下，人耳感覺不出差異，這個範圍就稱為臨界頻帶（critical band）。梅爾刻度（Mel scale）是根據人的聽覺系統所模擬出的線性轉換，在梅爾刻度中使用數個範圍不等的三角濾波器，梅爾刻度與頻率的關係如下：

**對數能量**

　　人耳對聲音大小的感知跟頻率同樣非呈直線關係，當聲音能量放大10倍時，人耳只會感覺放大1倍，而我們所得的語音訊號，可能因為語者的說話音量或背景的雜訊大小，影響語音訊號的數值變動。因為聲音的能量變動劇烈，通常會以對數的形式表示其能量大小，也就是將通過梅爾濾波器組的訊號平方後取對數值。

**離散餘弦轉換**

　　將通過梅爾濾波器組的訊號取對數值後，再做離散餘弦轉換到時域，這就是梅爾頻率倒頻譜係數，其中包含了13個倒頻譜係數和1個對數能量，DCT表示為

其中為梅爾頻率倒頻譜係數（MFCC），為經過梅爾濾波器組的訊號。

**主成分分析**

　　PCA能將維度間為相關的一群特徵向量用較少維度表示，且使得維度間變得彼此無關，同時能保有特徵向量的變異量，所以PCA可以找出特徵主要成分所在的基底向量。主成分分析的作法為使用所有訓練資料，為資料的總數，每個為維向量，來統計訓練資料的整體共變異矩陣（Covariance Matrix）：

其中為整體的平均向量，所以為維的共變異數矩陣。

對求特徵向量分解（Eigenvectors Decomposition），以特徵值（Eigenvalues）最大的前個特徵向量（Eigenvectors）當成基底矩陣的行向量，最後所有資料可以利用求得的轉換矩陣投影到新特徵空間。

**長短期記憶（Long Short-Term Memory, 以下簡稱LSTM）**

　　在針對處理有時間相關性的資料時，若利用一般的類神經網路，並不能擷取到時間前後的關聯。在語言溝通上，我們講述的一句話中都有上下文的關係存在，而我們若要知道一句話中所要表達的事，一般來說需要有前面時間點的資訊才有辦法完整的描述，並非單以句子中某單詞就能得知。而循環神經網路（Recurrent Neural Network, 以下簡稱RNN）則可以處理上述的時間序列資料，相較以往的類神經網路，RNN改變了其隱藏層的架構，使得輸入隱藏層的神經元中的資訊不僅僅包含輸入層的特徵，也包含了過去時間點的輸出結果，使其能擁有過去時間點的資訊，下圖為一個RNN的簡單架構示意圖：

由於基礎的RNN架構，會使得過去的資訊，在短時間內被新的資訊覆蓋，造成較早的資訊不斷地衰減，因此，難以獲取長時間的特徵。為了解決這個問題， LSTM架構被提出，並被運用在許多研究上，而LSTM是針對RNN隱藏層的部分進行改進，在隱藏層中各時間點傳遞資訊方面使用一個新架構的memory cell取而代之，其memory cell主要由Input gate、Forget gate、Output gate和一個cell所組成，如下圖所示：

Encoder-Decoder Attention模型

現存的諸多模型皆存在著Encoder-Decoder的架構，例如：Seq2Seq、語音辨識、機器翻譯等領域，都以Encoder-Decoder模型，進行相應的深度學習領域研究[]。Encoder可以視為利用時間數列資料的每一個時間點t-1進行預測，最後輸出一個t的結果，可表示成Encoder = （e0, e1, …, et-1），而其輸出結果為（ê1, ê2, …, êt），其中Encoder的輸入稱為Input，通常是欲預測的資料。而Decoder亦是利用時間數列資料的時間點T-1去預測T，可以表示成Decoder = （d0, d1, …, dT-1），其相對應輸出為（d̂1, d̂2, …, d̂T）。訓練完成後，Encoder的最終輸出êt即包含（e0, e1, …, et-1）序列的資訊，而其會作為輸入Decoder的一部份輸入，放入Decoder模型當中，以期得出Decoder的輸入在每一個時間點的輸出。本模型的Encoder輸入層為經過MFCC轉換過後的一秒英文數字檔案，為99個時間點的39維MFCC特徵，而Decoder的輸入為該音檔的音素（Phoneme）。以Decoder在每個時間點的輸出作為特徵，透過全連接層（Fully-connected）辨識該時間點的音素為何。Attention可視作嫁接Encoder與Decoder的橋樑，由於目前的RNN皆存在梯度消失以及梯度爆炸的問題，而Attention為一個十分簡單的全連接層架構，會對每一個時間點的RNN輸出進行特徵提取，並得出該輸出的注意力分配。使用RNN + Attention模型能夠過濾、提取出，對於輸出來說重要且具有時間相關性的特徵。

Encoder

Encoder分為三大部分，全連接層、ResNet層[]，以及LSTM層。全連接層先對每個時點的39維特徵進行線性投射至高維度，突顯該時點的特徵。經由三個ResNet層將鄰近的時間也納入考量，此處的ResNet將模型輸入經由三層卷積神經網路層進行特徵擷取，且將輸入與第三層卷積神經網路的輸出結果相加，再透過ReLU輸出，可以有效避免模型的梯度消失，而為避免梯度爆炸問題，在所以卷積神經網路層的後面，都加上了一層Batch Normalization。將第一與第二層的ResNet每三個時間點的特徵進行連接，使時間長度t壓縮成t/3，也就是每個新的時間t/3，皆包含3個時間長度的資訊。時間的壓縮，有助於減少模型於LSTM層的運算時間，如此一來便可以加速模型訓練的進行。而最後一層的ResNet輸出，會作為Encoder的LSTM層的每一個時間點的輸入（e0, e1, …, et-1）得到（ê1, ê2, …, êt）。

Attention

Attention可以簡單分為兩個全連接層，與一個輸出，稱為Context Vector（C）。依據[][]所述，將Encoder的輸出（ê1, ê2, …, êt-1）稱作query（Q），而êt稱為value（V）。透過兩個不同的全連接層f與g，分別對query與value進行轉換過後，以element-wise的方式相加，最後透過tanh進行轉換，得出第一層的結果（O）。接著再次通過一層僅有一個結果的全連接層h，便可得出Encoder輸出所形成的score（S）。而對score取softmax轉換後的結果，為該Attention的weights（W）。將weights與value以element-wise的方式相乘並加總過後，稱為Context Vector。

Decoder

Decoder主要為一層LSTM，其輸入為音素序列。音素序列首先會先通過一個Embedding層，將每個音素投射至指定維度的空間，使得音素不會因為音素的編號數字影響輸出結果，音素序列（）前後會加入序列開始（Start of sequence）與序列結束（End of sequence）的信號，亦會同時透過Embedding層進行轉換，稱經過Embedding的音素序列為音素向量（）。接著將音素向量與Decoder的開始信號與Attention模型所輸出的C連接起來，作為LSTM層的輸入值，對開始信號之後的每一個時間步進行預測，直到預測的結果出現序列結束信號，又或者達到設定的預測個數上限時，才會停止。最後，將該預測的結果通過全連接層進行分類，進而得到每個時間點所預測的音素。

模型訓練

訓練Encoder-Decoder模型時，共讀取20000筆語音資料，每種類別皆讀取2000筆，並預留1000筆語音資料作為驗證集，進行訓練。每筆語音資料皆進行過MFCC轉換，並伴隨著相對應的標籤與音素。模型訓練是以四筆語音資料作為一個批次（Batch）進行，在一次Epoch中，總共有4750個批次，而模型每次的Epoch約需耗時15分鐘。經過多次反覆的測試過後，最終決定採用Adam優化器[]，並將模型的學習率（Learning rate）設定為0.0001，Epoch最高上限訂為20次，終止條件為訓練集的損失小於0.02，進行模型訓練。在聲音模型（Acoustic Model）的訓練完成之後，以該結果進一步訓練字詞模型（Lexicon Model），並以字詞模型所輸出的結果與其他模型進行比較。

已7頁

摘要 1頁（中英文）

文獻回顧 1頁

流程圖 2頁

研究結果　結論 1

參考文獻

摘要

背景

資料集介紹

文獻回顧

研究方法

-前處理

-CNN模型

-CNN

-模型結構

-LSTM模型

-LSTM

-模型結構

-EN-DE模型

-ENCODER

-ATTENTION

-DECODER

-模型結構

結果

結論

參考文獻

附錄