1

我們是第11組，我們的主題是「英文數字0-9的語音辨識」

我們的組員有統計四…...

2 研究動機

而我們之所以想以此為主題是因為

隨著科技的演進，資訊產品不斷朝著人性化的方向發展，而語言正是人類溝通最自然的工具，良好的語音辨識系統能提供更方便的使用介面，所以這是我們以此作為我們的主題的原因

應該有人會好奇語音辨識的研究已經這麼多人做過且已經這麼成熟，像是ios的SIRI和Google的語音助理都已經被廣泛運用，我們為什麼還要再做？

這是因為

3研究方法

我們的方法分為三個部分，包含語音前處理、語音切割及語音辨識

那我們將在以下依序介紹

4 語音資料蒐集

Google在kaggle上提供了多個英文單字的語音資料，其中包含多個不同人所講出的英文數字0-9，例如one, two, three…

5 語音前處理(MFCC)

預強調(pre-emphasis):補償語音信號高頻損失的部分

由於人類發出的語音信號，會受到聲帶和嘴唇的影響，導致信號隨著頻率增高而約有6dB/oct的衰減，因此當語音信號在進入系統後，會先經過高通濾波器，補償語音信號高頻損失的部分。

音框化(frame):讓相鄰兩音框有一段重疊區域

音框就是把N個取樣點集合成一個觀測單位。而為了避免相鄰兩音框的變化過大，所以會讓相鄰兩音框之間有一段重疊區域。

窗化(window):減少音框左右兩端的不連續

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漢明視窗(hamming window)

快速傅立葉轉換(fast fourier transform):時域轉頻域

由於訊號在時域（Time domain）上的變化通常很難看出訊號的特性，所以通常將它轉換成頻域（Frequency domain）上的能量分佈來觀察，不同的能量分佈，就能代表不同語音的特性。

三角帶通濾波器(triangular bandpass filters):消除諧波的作用

如圖

對數轉換(log transform):增加一個音框的音量

因為音框的音量(也就是它的能量)，也是語音的重要特徵，所以將它加進去，變為其中一個語音特徵。

離散餘弦轉換(discrete cosine transform):降低語音特徵的維度

語音特徵的維度總共有十三維，我們藉由離散餘弦轉換去如個維度間的相關性，並降低語音特徵的維度。

逆傅立葉轉換(Fourier Inverse Transform):將頻域轉為時域

將前面去掉不要的頻率的資料，再轉為時域，得到我們真正想看的資料變化，再去做下一步的分析。

6 語音切割

人工切割 vs 自動切割

語音切割主要有分為兩種，人工切割與自動切割。

人工切割是仰賴人眼，一個一個標注語音檔案；而自動切割則是靠程式去切割語音檔。但由於人工切割的方式會消耗大量人力與時間，所以在這裡我們所使用的是自動切割。

7 自動切割方法

自動切割又可以用好幾種方法去切割語音訊息，我們這裡可能會採取的方法有k-means、Dynamic Time Warping、Hidden Markov Model、Recurrent Neural Network及Long Short-term Neural Network。

他們都有各自的優缺點，所以我們想嘗試融合幾個以上的方法，取其擅長的部份來幫助我們接下來的切割作業，看是否能提高我們語音辨識的準確率。

8 辨識

深度學習的方法 CNN, RNN

我們在語音辨識主要使用深度學習的方法Convolution Neural Network及Recurrent Neural Network。

CNN最初使用在電腦視覺作業上。但由於語音的前後有一定程度的相關，也就是上一秒所說的事情與下一秒不會毫無關聯，所以語音資料亦可以使用CNN進行辨識。

而RNN的架構與CNN幾乎相同，但RNN會將前一個時間點的資料也納入模型考量，而語音正是這樣的資料。所以RNN在語音辨識的處理上，是一個很重要的辨識方法。

9 預期結果 手機號碼辨識、驗證碼語音辨識

綜合以上方法，我們預期的研究結果為，當我們對著電腦講英文數字0~9，電腦可以準確地將我們所講的內容用文字顯示出來，且我們可以擴大將此應用到手機號碼辨識及驗證碼語音辨識上面，方便使用者透過講話的方式就能夠達成以上兩種辨識。

10 參考文獻