1 主題

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 目錄

3 研究背景

而我們之所以想以此為主題是因為隨著科技的演進，資訊產品越來越人性化，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。

研究目的:希望能透過不同的語音辨識技巧，找到提高語音辨識準確率的方法。

4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

5 語音資料

那在我們進到前處理之前，我們先簡單了解一下我們的資料，我們的資料來源是Google在kaggle這個網站上提供了多個英文單字的語音資料，資料的每種數字訓練資料都有超過兩千筆，皆為不同語者，且都是一個一個單字，例如one, two, three…

6 語音前處理(MFCC) 梅爾倒頻譜係數

在一段語音裡面，語音的音量、音高、清晰度都會影響辨識的準確率，所以需要在辨識前先做語音訊號的前處理。我們這邊以梅爾倒頻譜係數，也稱為MFCC來做處理。

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，所以我們要做預強調，來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，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

對頻譜進行平滑化，並消除諧波的作用，突顯原先語音的共振峰。（因此一段語音的音調或音高，是不會呈現在 MFCC 參數內，換句話說，以 MFCC 為特徵的語音辨識系統，我們所輸入的語音訊號，並不會因為語音音調的不同而影響其結果。）

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找出數個維度間的相關性，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。

8 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。

人工切割的優點是它可以切得很精確，但同時也耗費大量的時間跟人力，而且每個人對於切割單元邊界的認定是不一致的，那自動切割就完全解決了人工切割的缺點，但也因為它的切割方法有許多種，所以並不是每個都有很高的準確率。

9 自動切割方法

那我們可以看一下自動切割的方法

HMM-GMM是較為早期的東西，所以準確度較不那麼高，可能要搭配深度學習才能提高它的辨識率。

現在比較新的方法為RNN時間循環神經網絡和**Long Short-term Neural Network**

而我們目前還在評估要使用哪一種方法，或是是否能作結合。

10 語音辨識

我們在語音辨識主要使用深度學習的方法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。

我們用一個例子來說明一個簡單的 RNN：假設有一個人他吃飯的習慣固定，例如他第一天吃漢堡、第二天吃披薩，之後依序吃這些東西，在使用RNN的情況下，輸出「熱狗堡」送回神經網路判斷，就知道他隔天會吃「章魚小丸子」。所以我們可以知道RNN將前一個時間點的資料也納入考量，而語音資料的前後有一定程度的關聯，所以RNN在語音辨識的處理上是很重要的辨識方法。

11 預期結果

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。

12 參考文獻

近年來出現了許多企圖使用自動的方法對語料進行切割的相關研究，其中強迫對齊法是最熱門也最常被使用的一種自動切割方案。強迫對齊的做法類似語音辨認， 其以隱藏馬可夫模型 (Hidden Markov Model, HMM) 為基礎，利用先前訓練好的 HMM聲學模型強迫將聲學參數對應到不同model的每個HMM-state之中計算觀 察機率 (Observation probability) 與轉移機率 (Transition probability) ，最後得到 總合機率值最高的狀態序列 (State sequence) 即為自動切割的最終結果。但和辨 認不同的地方在於辨認只要求得到正確的文字序列 (Word sequence) ，而切割的 重點則在於各個切割單元起頭與結束的邊界位置，對特徵參數的 frame 與 model 中的 state 兩者的契合程度比辨認的要求更高，因此自動切割實作上比辨認更加

困難，直至目前為止強迫對齊法也還有相當大的改善空間，尤其如果需要處理的

對象是屬於自發性語料，則常因發音變異或音節合併等複雜的口語現象使得切割 單元之間的分界不明顯，自動切割的工作就又變得更加棘手了。

圖 2.1 為本研究所提出之二階段的流程圖。我們以 GMM-H