# **Computer version applied to audio recognition**

Belinda , Skylol , Roon , Millie , Yee

National Tsing Hua University

No.101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 300044, R.O.C.

[millie1020667@gmail.com,a0912556613@gmail.com](mailto:millie1020667@gmail.com,a0912556613@gmail.com)

pme112030504@gapp.nthu.edu.tw

**Abstract**

*道路駕駛中常會遇到各種不同的路況，若使用駕駛輔助系統可以幫助我們在行駛的途中提高路況掌握度減少事故的發生，本專題希望製作一個聲音辨識來判別道路狀況並給予路況提示，分別蒐集警車、救護車、撞擊聲、正常道路聲音的四種情境，經過 audio data augmentation 提升模型訓練資料量，將音訊透過傅立葉轉換，從時域轉換到頻域，獲取頻域及聲波圖形，利用影像辨識的 CNN 模型判別圖形對應到的路況，進行分類，幫助駕駛掌握路況。*

# **Introduction**

當我們行駛在道路上，必須時刻關注來自四面八方的路況，然而，由於駕駛座的位置容易產生視線死角，將使我們難以全面觀察周圍的情況，這時即使我們聽見救護車的鳴笛聲，也難以確定其來自何處，進而增加了潛在的危險。

因此，若我們能搭配一個利用音訊處理技術的駕駛輔助系統，將能幫助駕駛更輕鬆的掌握路況，保障行車安全。

## **Why we choose audio**

目前的駕駛輔助系統主要是使用攝影機影像辨識，這使得系統只能作用在攝影機覆蓋範圍內操作。雖然增加鏡頭數量可以擴大視野，但也隨之增加了設備成本，因此，若是使用音訊處理技術，即為一個更加經濟實惠的選擇。此外，當我們在道路上行駛時，有時會遇到視線受阻礙的情況，此時聲音辨識可提供額外的信息，彌補視覺上的不足，提高我們對路況的了解。同時，聲音是一種能夠幫助我們提前感知危險情況的信息，因此可以幫助駕駛提前警覺到緊急情況。這種技術的使用有望在提高道路安全性的同時，降低花費成本，提供給我們更全面的路況感知。

# **Formatting your paper**

本專案將運用兩種主要技術實作道路駕駛輔助系統，蒐集道路聲音進行分析，選用辨識度較高的特殊車種、車禍撞擊聲與普通道路聲進行音訊特徵提取，經過傅立葉轉換後可產生頻域聲訊圖，利用CNN判斷音訊種類辨識音訊狀況，判斷道路狀況。

## **Margins and page numbering**

音頻信號是有規律的聲波信息載體，根據聲波的特徵，可把音頻信息分類為規則音頻和不規則聲音。規則音頻是一種連續變化的模擬信號，可用一條連續的曲線來表示，稱為聲波。聲音的三個要素是音調、音量和音色，聲波或正弦波有三個重要參數頻率、幅度和相位，決定了音頻信號的特徵，若先對音頻進行特徵提取，有助於提高辨識的準確度。

因此，若欲完成我們的目標，首先須蒐集音頻資訊，包含警車、救護車、撞擊聲和正常道路聲的數據，這些數據樣本應盡可能的涵蓋各種不同的情境。接下來，我們需對蒐集的音訊進行標記，標出每個樣本所對應的路況情境，用來訓練Supervised learning model。

音訊應該主要會從網路取得，道路實收偏困難，可以先決定要不要統一網路或是參雜少部分的實收，因為背景白噪音會有影響，音訊主要為疏密相間的縱波，透過取樣轉換為數位訊號，進而接續後面的處理。準備好所有辨識所需的資料後打散（random)，如果要分成測試、訓練練集、驗證集的話1：8：1，若是分成測試、訓練的話2：8。

## **Audio data augmentation**

若想要提高判斷的準確度，需要大量的資料來訓練模型，但如果全數的資料都由人工收取需要消耗大量的時間，所以運用data augmentation技術在資料集中產生更多樣本。

Data Augmentation是一種產生合成資料的方法，透過調整原始樣本中的因素來創建新樣本，獲得單一樣本的大量數據。這不僅有助於我們增加資料集的大小，而且還提供了單一樣本的多種變體，有助於避免模型過度擬合。聲波具有音調、響度、質量......等特徵，改變這些特徵可以實現噪音添加、時間平移、音調平移和時間拉伸的資料擴增。

1. **噪音添加**

向樣本添加白噪聲，白噪音是按規則間隔分佈的隨機樣本，平均值為 0，標準差為 1。

1. **時移（time shifting）**

在這裡，我們將波形移動了sample\_rate／10，波可以沿時間軸向右移動給定的值。

1. **時間拉伸**

在不影響音調的情況下改變聲音速度／持續時間的過程，採用波動樣本和拉伸因子作為輸入。

1. **變調**

使用音高縮放來實現，改變音調而不影響其速度的過程，需要波形樣本、取樣率和音調必須移動的步數，步數大約界在 -5 到 5 之間的步數非常有利。

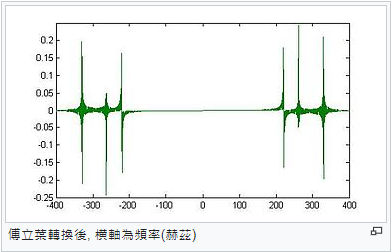
## **STFT（Short-time Fourier Transform）**

短時傅立葉變換用於決定隨時間變化訊號的局部正弦頻率和相位，將長時間訊號分成數個較短的等長訊號，然後再分別計算每段的傅立葉轉換。通常用來描繪頻域與時域上的變化，為時頻分析中的重要工具。傅立葉轉換只提供了有哪些頻率成份的資訊，卻沒有提供時間資訊，而短時傅立葉轉換則清楚的提供這兩種資訊，這種時頻分析的方法有利於頻率會隨著時間改變的訊號，如音樂訊號和語音訊號等分析。

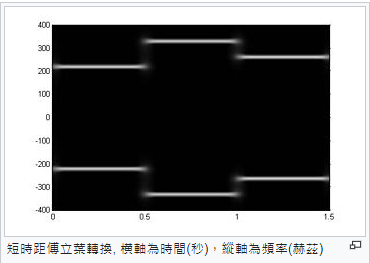
先乘上一段時間不為零的窗函數再進行一維的傅立葉變換，窗函數沿著時間軸挪移，得到一系列的傅立葉變換結果展開成二維表象，用數學公式可表達成以下式子：



其中的 是窗函數， 是待變換的函數，是傅立葉轉換的結果，根據上述公式可以發現當時間變換時，窗型函數也會跟著時間一起移動， 則會留下窗型函數對訊號的傅立葉轉換。有別於以往的一般傅立葉轉換，短時距傅立葉轉換加入了時間這個參數，這讓根據時間而產生頻率變化的音樂訊號在經過轉換後，得到時間-頻率的關係圖



圖一



圖二

圖1是經過傅立葉轉換得到的例子，橫軸是頻率縱，軸是峰值；圖2是經過短時距傅立葉轉換的例子，橫軸是時間，縱軸是頻率（圖片來源：維基百科）

對於聲音而言，峰值大小代表聲音的大小，要分辨出聲音的種類則需要收集聲音的頻率，因此在生訊這個領域方面，短時距傅立葉轉換的功用會比一般傅立葉轉換來的優秀，在什麼時間有著什麼樣的聲音頻率對於我們這次的專案有著非常大的關聯，因此我們選擇使用短時距傅立葉轉換來作為我們提取聲音特徵的技術。

## **CNN（Convolutional Neural Network）**

CNN圖像辨識由卷積層（convolution layers）、池化層（pooling layers）、全連接層（fully connected layers）組成，透過卷積層讀取局部特徵，接著再由池化層降低參數數量，最終由全連接層輸出結果。

CNN主要具有兩大特點，其中之一為權值共享，在一般的圖像內有許多的特徵是相同的，權值共享可以讓幾個相同神經元組成的卷積核去學這個特徵，透過滑動窗口（卷積核）對整張圖片進行卷積，進而達到節省參數的效果，大大減少了所需處理的數據。

而另一大特點則是能夠保留圖像的位置資訊，圖片中的像素與其鄰近的像素會有一定的關聯度，所以在池化層中我們可以取得每個區域為重要的資訊；在池化層和全連接層間，通過 Flatten 的步驟，把多維的資訊拉成一維的向量，雖然會大量失去特徵之間的空間資訊，但全連接層會關注特徵間的關聯性，並預測出原始我們希望預測的目標。在模型訓練中，我們也會使用像是激活函數、損失函數等數學工具來更優化模型的訓練情形。

# **Final**

需要完成的技術

1. 蒐集數據，四種數據分別包含100個聲音資料
2. Audio Data argumentation
3. Short-time Fourier Transform
4. CNN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Collectingdata | Augmentation | STFT | CNN |
| 10/9 |  |  |  |  |
| 10/16 |  |  |  |  |
| 10/23 |  |  |  |  |
| 10/30 |  |  |  |  |
| 11/6 |  |  |  |  |
| 11/13 |  |  |  |  |
| 11/20 |  |  |  |  |
| 11/27 |  |  |  |  |
| 12/4 |  |  |  |  |
| 12/11 |  |  |  |  |

# **References**

1. https://kknews.cc/zh-tw/code/44bveyg.html
2. https://medium.com/@keur.plkar/audio-data-augmentation-in-python-a91600613e47
3. https://medium.com/ai-academy-taiwan/speech-recognition-using-neural-network-d1af6f482c9b