多媒體程式設計
音訊資料處理

Instructor: 馬豪尚

Librosa

- Librosa是專門用來分析聲音訊號的 Python 模組
- 提供音訊處理、時頻轉換處理、特徵擷取、繪製聲波圖形等等功能
- 安裝套件
  - pip install librosa
- 為了使 audioread 可以支援更多的聲音檔案格式，建議同時安裝 ffmpeg
  - apt install ffmpeg

Librosa讀取音訊

- data, sr = librosa.load(filename, sr, mono, offset, duration, dtype)
  - filename為檔案名稱
  - sr為指定的sample rate，預設是22050
  - mono為一個boolean值，True/False代表單/雙聲道，預設為True
  - offset表示要開始讀入的音訊位置，預設0.0
  - duration表示讀入的長度，預設None，表示全部音訊
  - dtype為指定回傳值音訊資料data的資料型態，預設為float32
  - 會回傳兩個值
    - 第一個為音訊本體資料(numpy array)
    - 第二個為音訊的sample rate

Librosa讀取音訊

- 儲存音訊的Numpy array為ndarray型態
- ndarray維度為(n,) 或 (…, n)
- 元素的值代表取樣點的振幅
  - 經過量化後的值

data單聲道(n,)

data雙聲道 (…, n)

Librosa範例音訊

- librosa 模組中有附帶範例的聲音檔案，可以做為開發與測試使用
- librosa.util.list_examples()
  - 可以列出所有範例聲音檔案的資訊
- librosa.example('brahms')
  - 下載檔名為brahms的音訊

```
AVAILABLE EXAMPLES
--------------------------------------------------------------------
brahms          Brahms - Hungarian Dance #5
choice          Admiral Bob - Choice (drum+bass)
fishin          Karissa Hobbs - Let's Go Fishin'
nutcracker      Tchaikovsky - Dance of the Sugar Plum Fairy
trumpet         Mihai Sorohan - Trumpet loop
vibeace         Kevin MacLeod - Vibe Ace
```

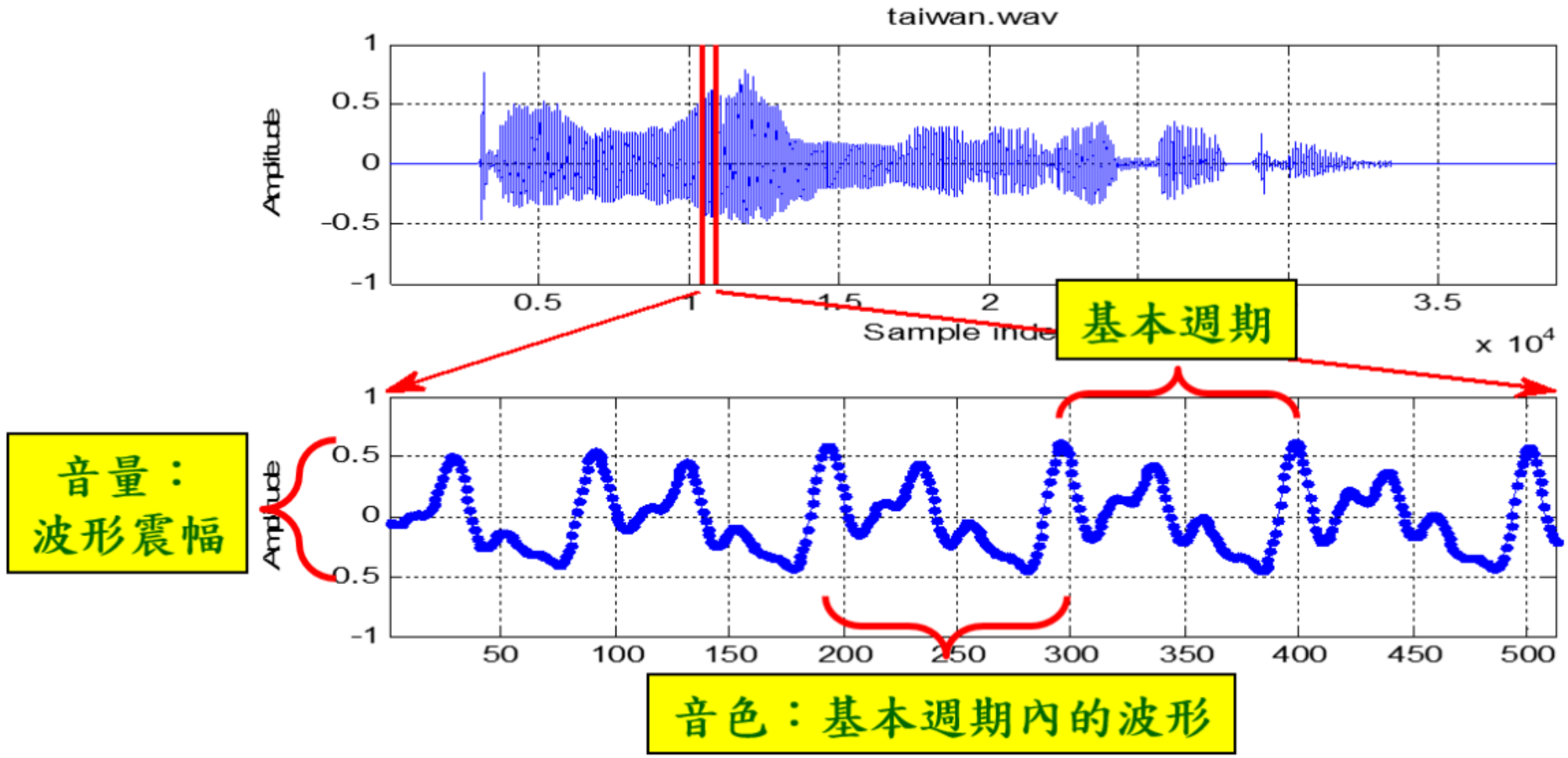
Librosa 繪製聲波圖

- 載入模組
  - import librosa.display
  - import matplotlib.pyplot as plt
- 用plt產生一個圖的物件
  - plt.figure()
- 搭配librosa.display將聲波輸入到物件上
  - librosa.display.waveshow(data, sr=sr)
- 顯示圖形
  - plt.show()

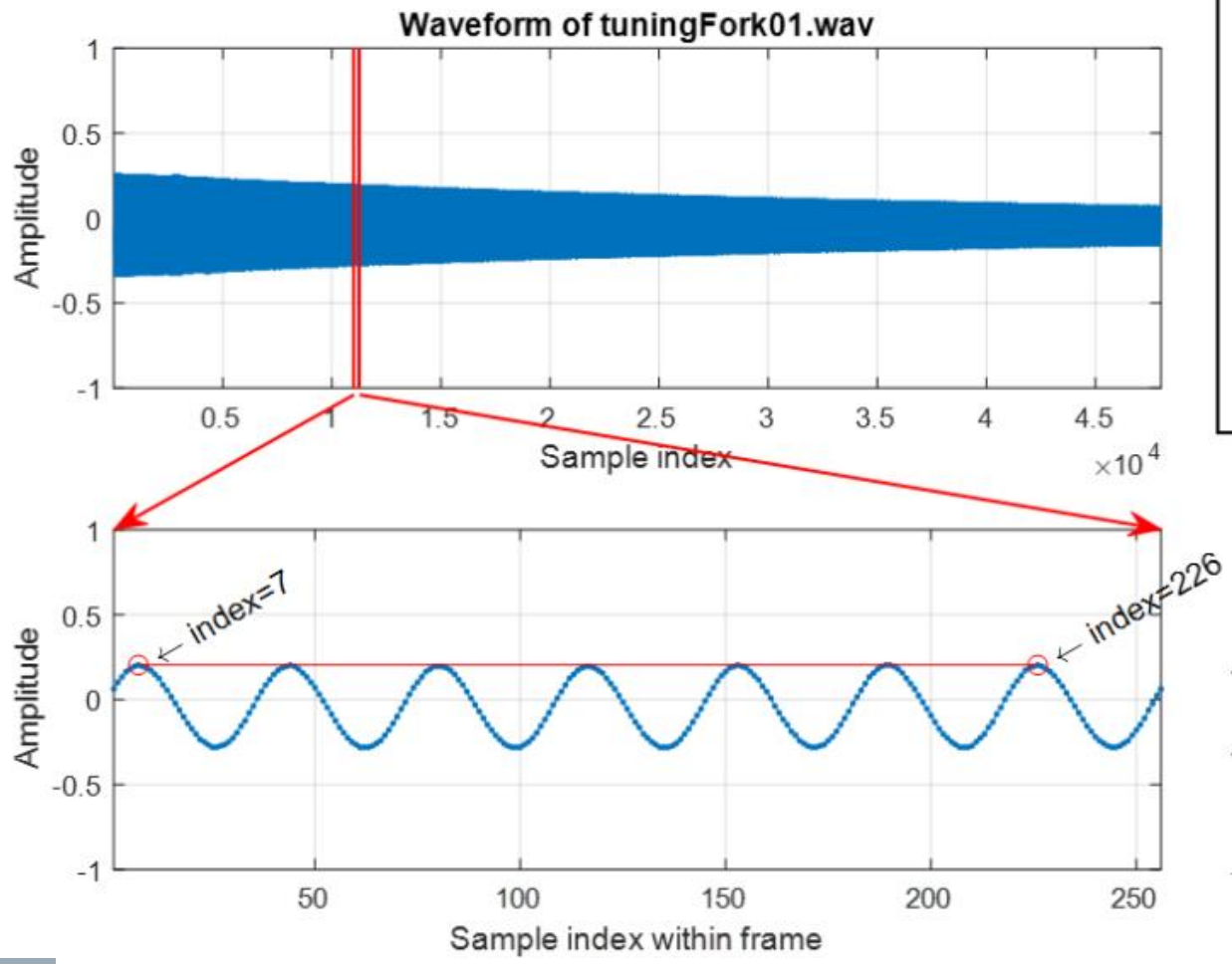
音訊特徵

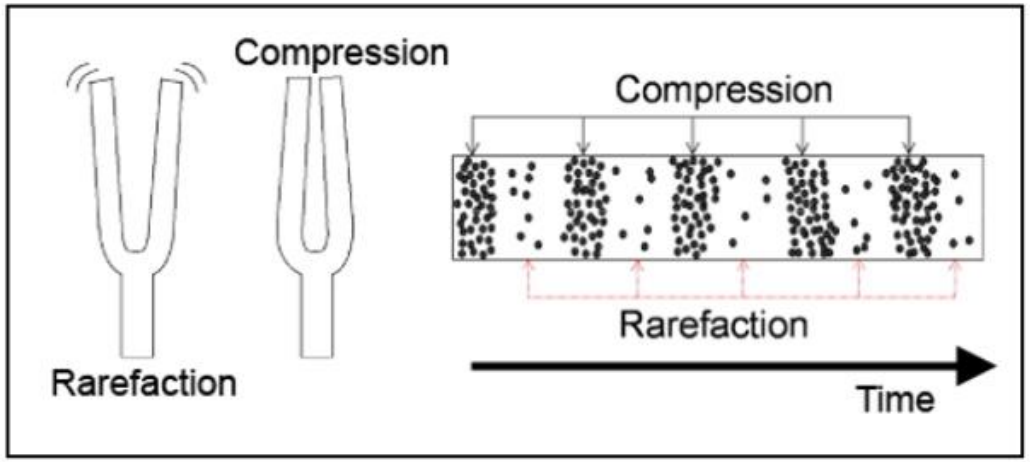
- 音高（Pitch）：代表聲音的高低，可由基本頻率來類比
- 音色：代表音訊的內容，例如「ㄚ」和「ㄛ」的發音方式不同，就會產生不同的音色
- 在音訊資料裡，量化(quantization)是把震幅類比值變成離散值，儲存的都是震幅的資訊
- 想分辨不同的音，需要靠音高和音色來分辨
  - 訊號的頻率，也就是聲音震動的頻率
  - 其代表的是音調的高低，頻率越高，音調就越高
  - 使用音訊在不同頻率的能量分布，來代表音色

音訊特徵


taiwan.wav
Amplitude
Sample Index
x 10^4
基本週期
音量：
波形震幅
音色：基本週期內的波形

觀察聲音頻率和音高



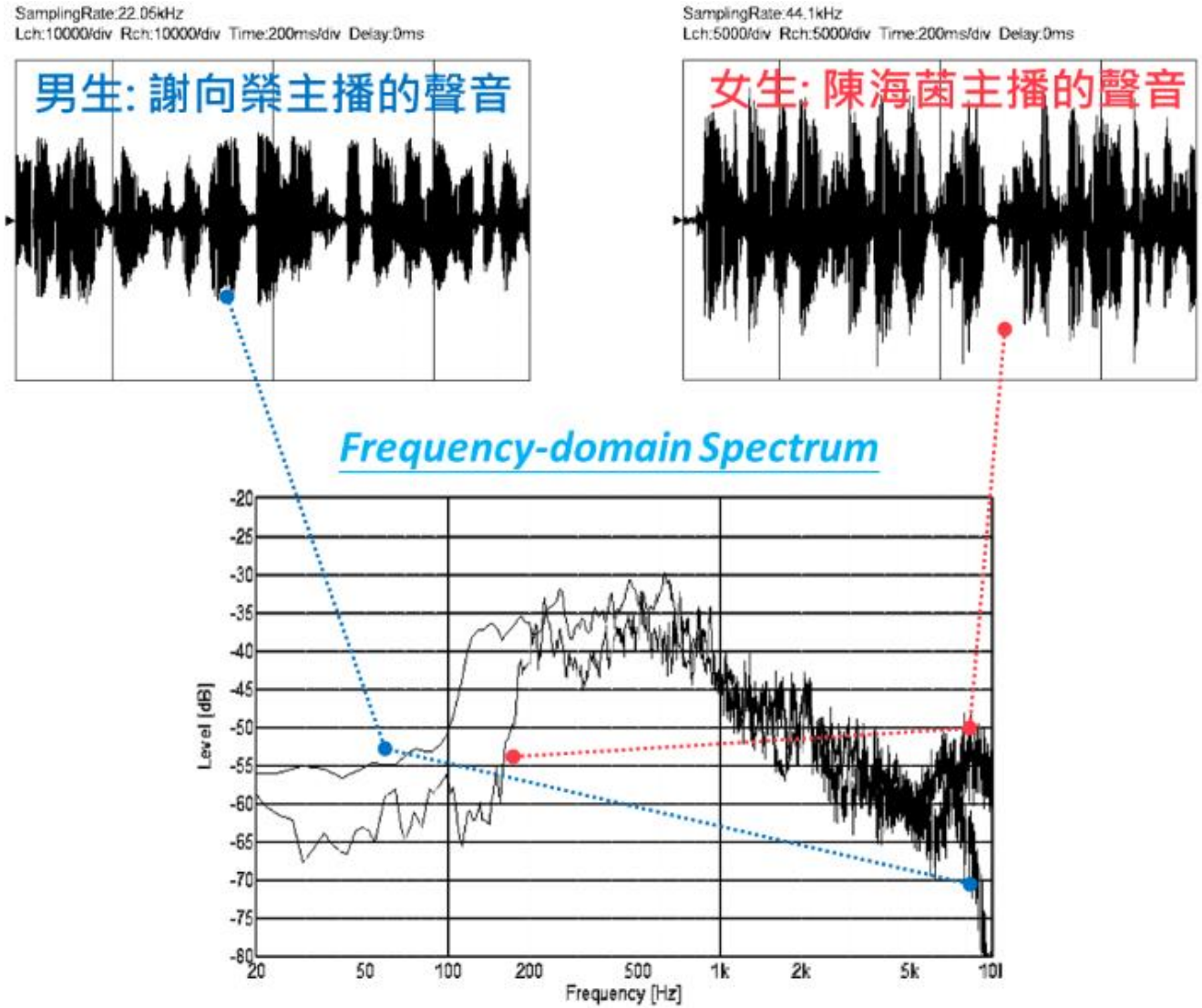


$$fp = (226 - 7)/6 = 36.50 \text{ points}$$

$$ff = 16000/fp = 438.36\ Hz$$

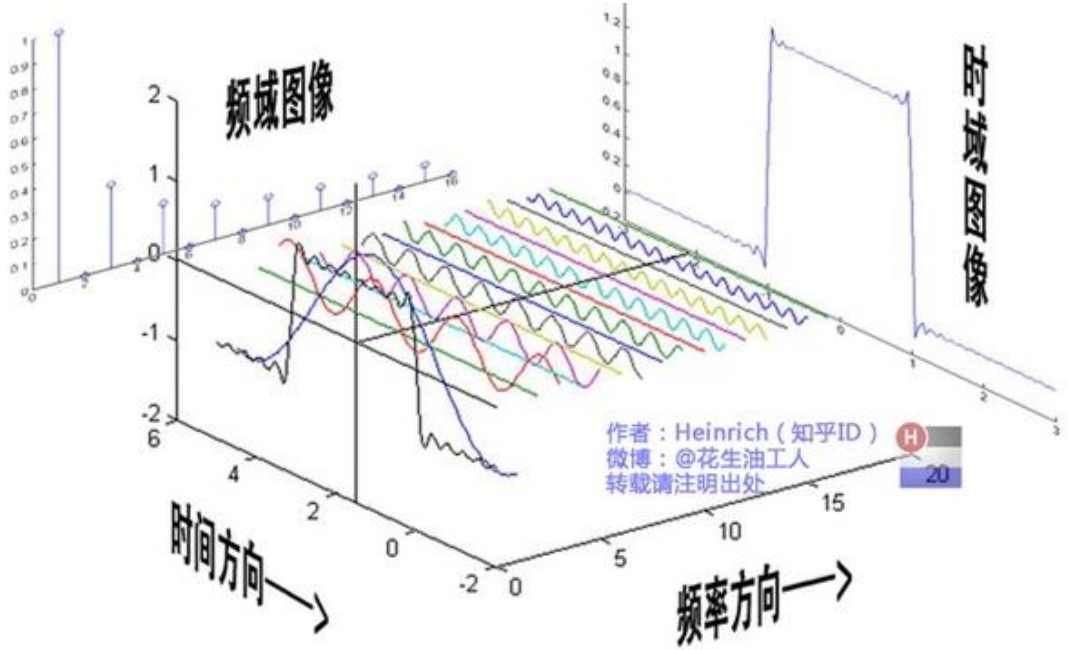
$$pitch = 69 + 12 * \log_2\left(\frac{ff}{440}\right) = 68.94\ semitone$$

將震幅轉成頻率


SamplingRate:22.05kHz
Lch:10000/div Rch:10000/div Time:200ms/div Delay:0ms
男生: 謝向榮主播的聲音
SamplingRate:44.1kHz
Lch:5000/div Rch:5000/div Time:200ms/div Delay:0ms
女生: 陳海茵主播的聲音
Frequency-domain Spectrum
Level [dB]
-20
-25
-30
-35
-40
-45
-50
-55
-60
-65
-70
-75
-80
20
50
100
200
500
1k
2k
5k
10k
Frequency [Hz]

訊號處理

- 傅立葉轉換
  - 傅立葉告訴我們，任何週期函數，都可以看作是不同振幅，不同相位弦波的疊加。
  - 傅立葉轉換可以將時域上的週期函數，轉成頻域上的能量函數
  - 在時域上處理訊號很困難



Librosa傅立葉轉換

- dataFFT = librosa.stft(data, n_fft=2048)
  - 用來計算短時距傅立葉轉換（STFT，Short-time Fourier transform
  - 輸入要計算的音訊資料物件data
  - n_fft為傅立葉轉換的音訊框長度
  - 回傳一個計算完的音訊資料物件

Librosa傅立葉轉換

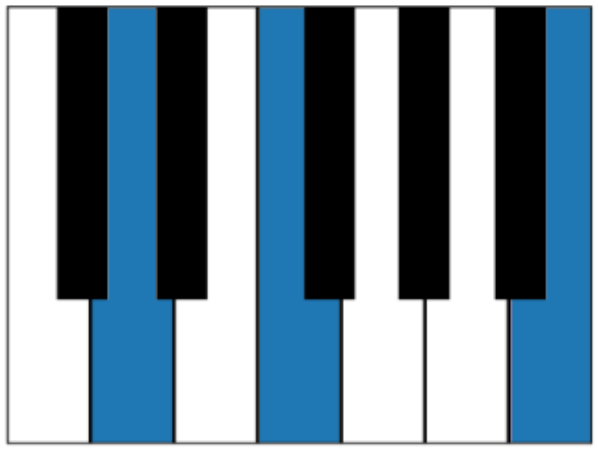
- 每一個變換之後的值是一個複數，為a+bj的形式
- FFT得到的複數的絕對值就是對應的頻譜震幅
  - dataFFTAbs = np.abs(dataFFT)
  - 用這個dataFFTAbs即可畫出來聲音的頻譜圖
- FFT得到的複數的角度值就是對應的相位
  - dataFFTAng = np.angle(dataFFT)
- 取出譜震幅和相位的函數
  - S, phase = librosa.magphase(librosa.stft(y=y))
    - 第一個回傳值是譜震幅
    - 第二個回傳值是相位

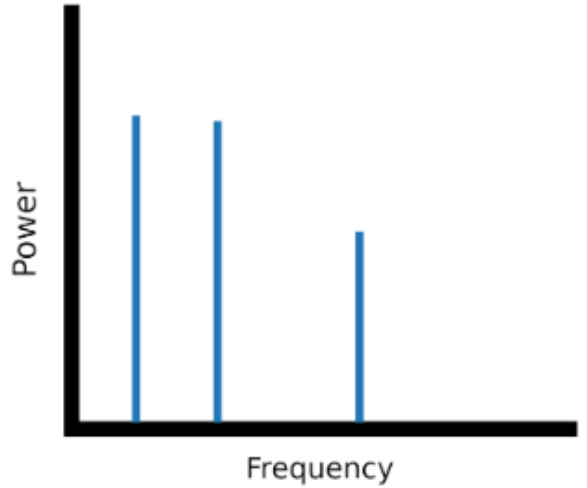
Librosa繪製頻譜圖(spectrogram)

- 頻譜圖的意義就表示這個聲音在不同時間時候的頻率分佈是三維的資訊
- 先將頻譜震幅轉成dB
  - librosa.amplitude_to_db(dataFFTAbs)
- 繪製頻譜
  - librosa.display.specshow(data, sr, x_axis, y_axis)
    - data為要繪製的聲音頻譜資料
    - sr為sample rate
    - x_axis為圖上的x軸名稱
    - y_axis為圖上的y軸名稱
- 可以加入一個顏色對照表，提醒顏色代表不同的值
  - plt.colorbar()

聲音在頻域上的表現

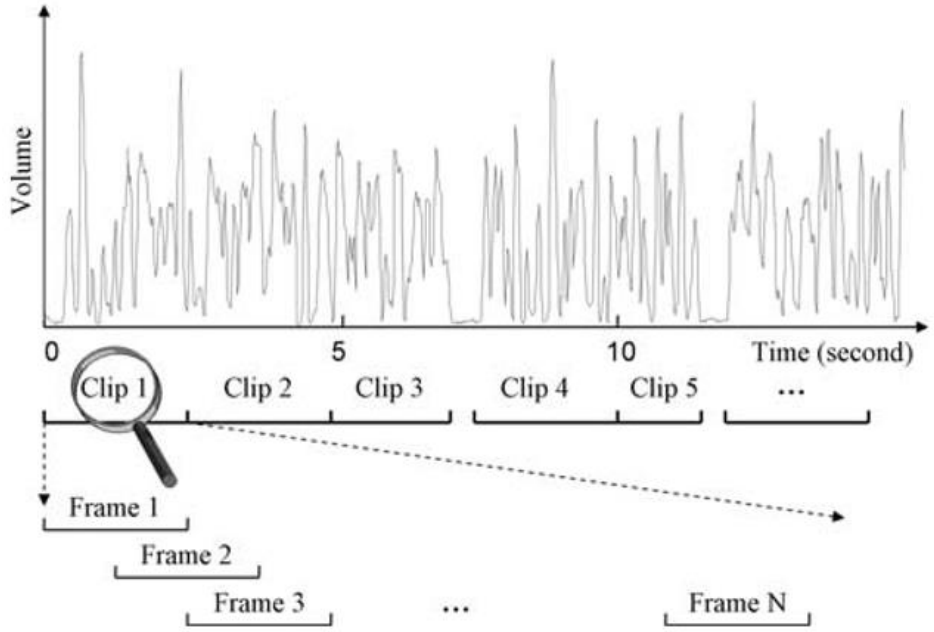
- 很多音訊的特徵都會表現在頻域上
- 例如鋼琴上的音，不同的音會有著不同的頻率和譜振幅





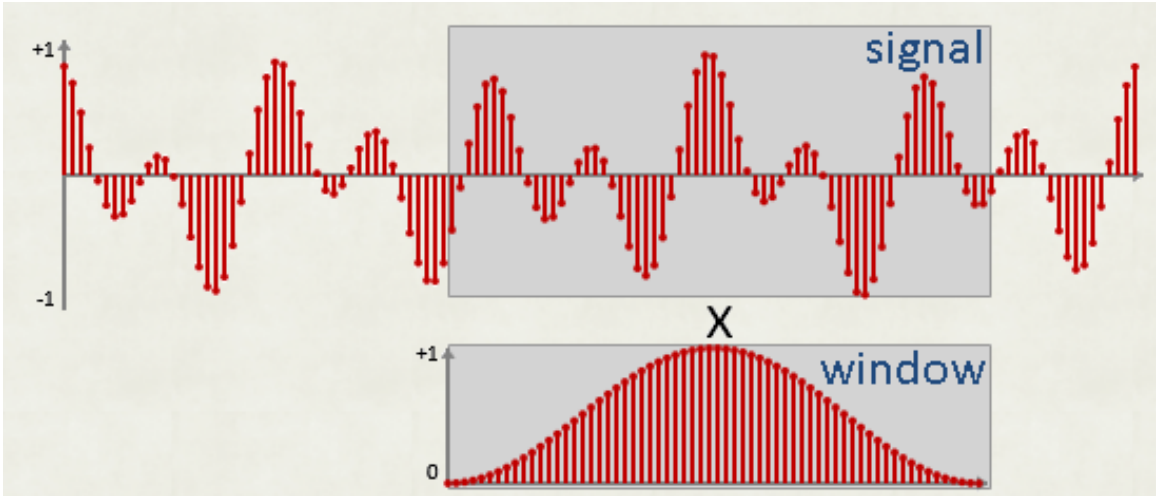
Librosa特徵擷取

- 音框（Frame）化
  - 在做特徵擷取前，將 N 個取樣點集合成一個觀測單位，稱為音框
  - 為了避免相鄰兩音框的變化過大，會**讓兩相鄰音框之間有一段重疊**
  - 假設所用的音訊的 sample rate 為 16 KHz 且音框長度為 256 ，對應的時間長度就是 256/16000*1000 = 16 ms



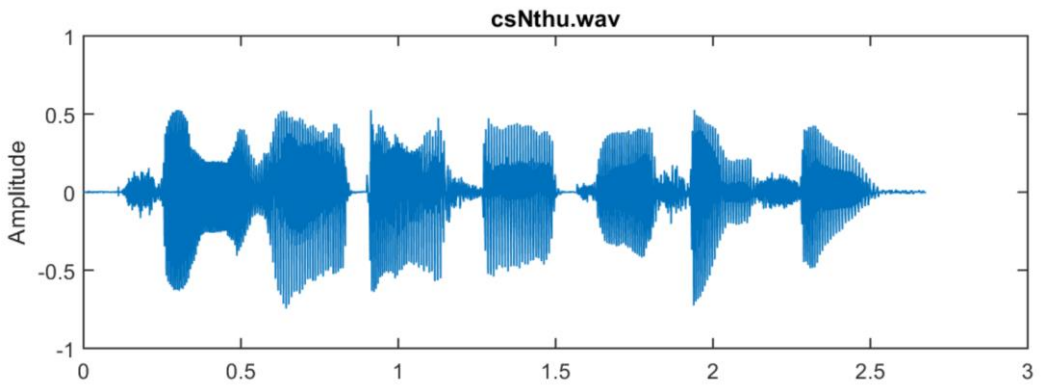
Librosa特徵擷取

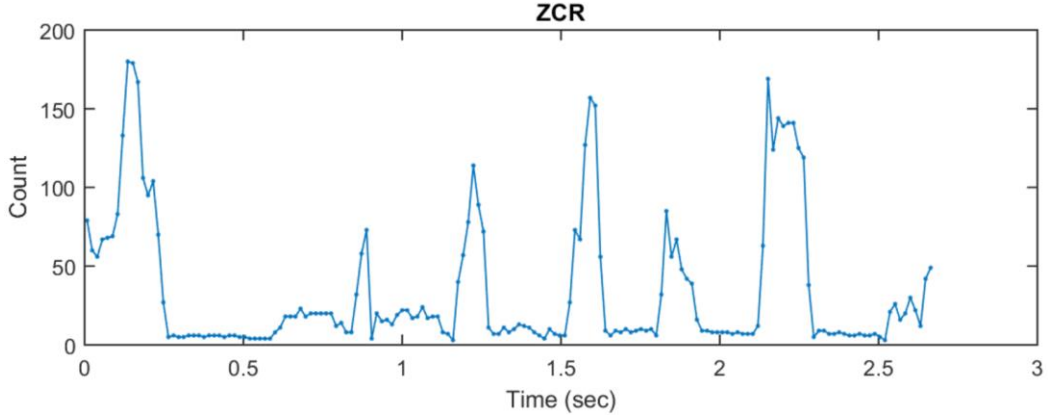
- 乘上窗函數
  - 原本完整的聲音波形，被框(frame)截斷，若截斷的地方不是完整的一個週期的音訊，對傅立葉轉換會造成影響
  - 將框內的音訊乘上一個窗函數(中間高兩側低，數值從0-1之間)
  - 將框的兩端的訊號漸漸減弱，減少影響



常見的音頻特徵

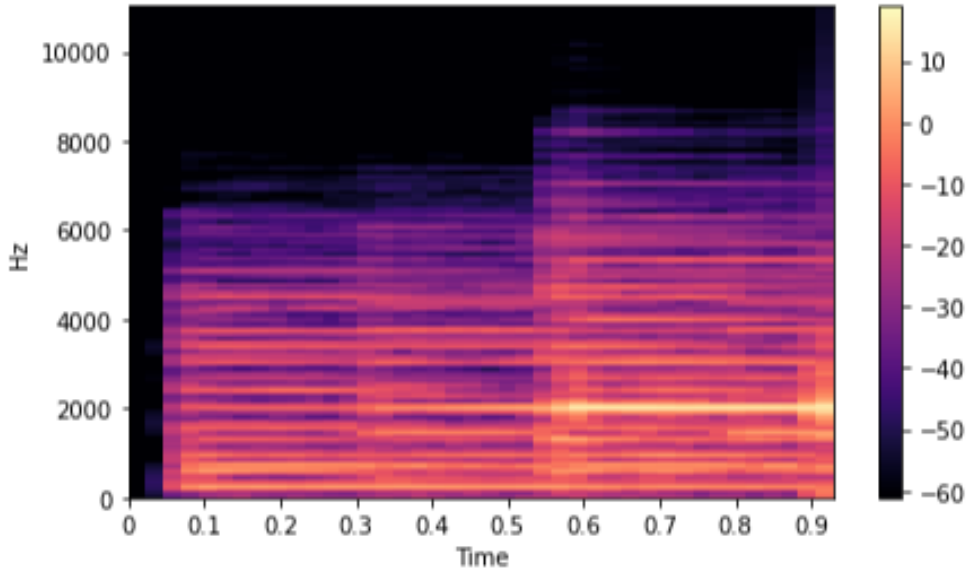
- 過零率(zero crossing rate, ZCR)
  - ZCR是指在每幀(frame)資料中，信號通過零點(正變成負或負變成正)的次數，此特徵在語音辨識和音訊信息檢索領域廣泛應用，是金屬聲音和搖滾樂的關鍵特徵





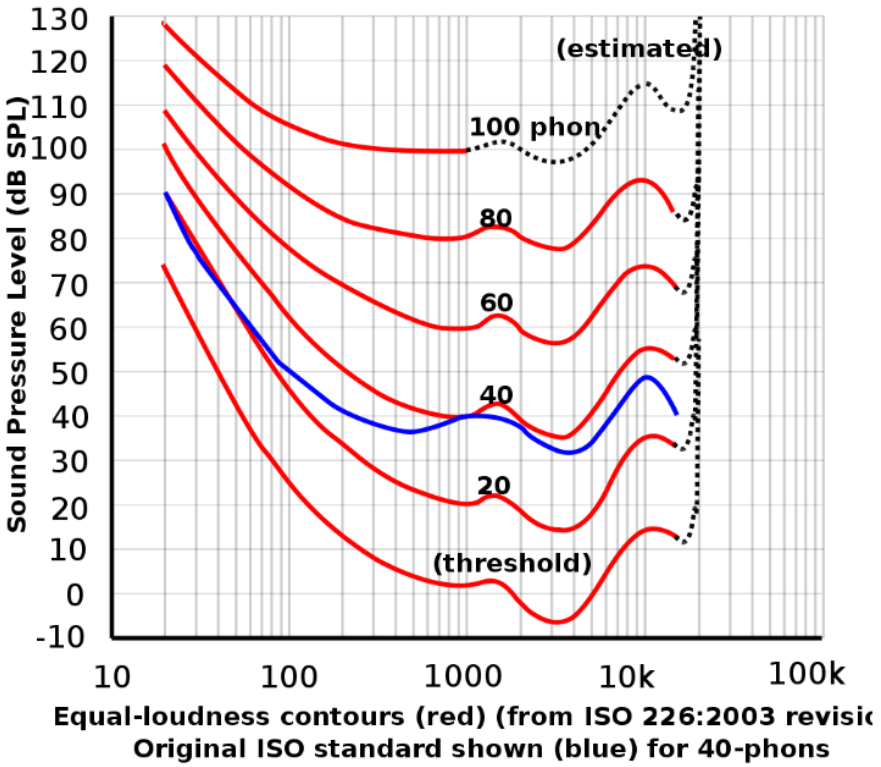
常見的音頻特徵

- 頻譜中心(spectral centroid)
  - 頻譜中心代表聲音的"質心"，又可稱為頻譜一階距，其數值越小，代表越多的頻譜能量集中在低頻範圍內



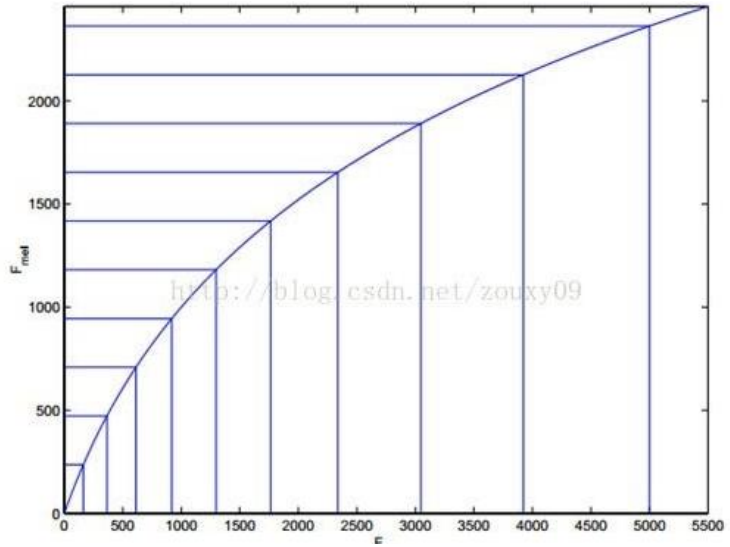
常見的音頻特徵

- 人類聽覺天生對不同頻率有不同的反應，簡單以頻率變化敏感度來說，低頻>高頻，但響度(聽覺上的大小聲)的敏感度卻反過來是高頻>低頻
  - 1kHz以下，越低的頻率要越大聲才能聽起來有同等響度。
  - 1kHz~2 KHz間人耳對音量的敏感度會稍差些
  - 2kHz～5kHz之間為人耳最敏感的區域，而且人耳在低音量時比起高音量時對此區域敏感
  - 6kHz以上，人耳的敏感度會逐步下降，但比起低頻率來說，音量大小對人耳低頻的敏感度影響高於5kHz以上頻率。



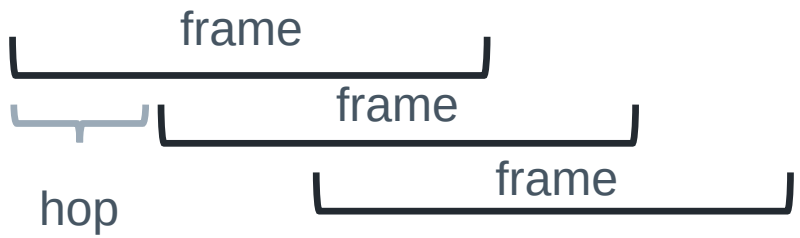
常見的音頻特徵

- 梅爾頻率倒譜系數(MFCC)
  - 梅爾頻率代表一般人耳對於頻率的感受度
- 取得MFCC的分析流程可以大致分為
  - 先對語音進行預處理，切割成frame和設定window
  - 對每一個window，透過傅立葉轉換得到對應的頻譜
  - 將上面的頻譜透過梅爾濾波器組得到梅爾頻譜
  - 在梅爾頻譜上面進行倒譜分析
    - 取對數，做逆變換，實際逆變換一般是做DCT離散餘弦變換，然後取DCT後的第2個到第13個係數作為MFCC係數



Librosa特徵擷取

- Librosa直接提供許多特徵擷取的函式
- 計算過零率
  - librosa.feature.zero_crossing_rate(data, frame_length=2048, hop_length=512)
    - data為輸入的音訊資料(聲波震幅)
    - frame_length為frame的長度，會以設定的長度為單位切割音訊成frame
    - hop_length為每個frame之間移動的距離(樣本數)，一般設定1/4個frame
    - 回傳值為一個numpy，儲存每個frame的過零率



Librosa特徵擷取

- 計算頻譜中心
  - librosa.feature.spectral_centroid(y, sr, S, n_fft=2048, hop_length=512)
    - 可以輸入原始的聲波震幅資料，也可以輸入轉成頻率的頻譜震度資料
    - y為音訊資料、sr為sample rate
    - S為頻譜音訊資料

Example

輸入為一般震幅的音訊

```
y, sr = librosa.load('trumpet.wav'))
librosa.feature.spectral_centroid(y=y, sr=sr)
```

輸入為頻譜震度的音訊

```
S, phase = librosa.magphase(librosa.stft(y=y))
librosa.feature.spectral_centroid(S=S)
```

Librosa特徵擷取

- 將音訊做MFCC特徵擷取
  - librosa.feature.mfcc(y, sr, S, n_mfcc, n_fft, hop_length)
    - 可以輸入原始的聲波震幅資料，也可以輸入轉成頻率的頻譜震度資料
    - y為音訊資料、sr為sample rate
    - S為頻譜音訊資料
    - n_mfcc為回傳的mfcc特徵數量
    - n_fft為傅立葉轉換的frame長度
    - hop_length為相鄰frame移動的長度

Librosa特徵擷取

- MFCC 特徵擷取 Example
  - 使用預設值且輸入為音訊波形震幅資料
    - y, sr = librosa.load(librosa.example('libri1'))
    - librosa.feature.mfcc(y=y, sr=sr)
  - 設定不同的hop length，得到不同數量的frame的MFCC特徵
    - librosa.feature.mfcc(y=y, sr=sr, hop_length=1024)
  - 設定不同數量的回傳特徵，得到不同數量的MFCC特徵
    - librosa.feature.mfcc(y=y, sr=sr, n_mfcc=40)

Librosa 節奏特徵擷取

- 取得音訊節奏(tempo)
  - tempo, beat_frames = librosa.beat.beat_track(y=y, sr=sr, hop_length=512)
    - y為音訊資料、sr為sample rate
    - 會回傳節奏頻率(tempo)和節拍出現的 frame 編號(beat_frames)
    - 每一個 frame 的長度是由hop_length所指定
- 將節拍出現的frame編號轉成時間
  - librosa.frames_to_time(beat_frames, sr=sr)
    - 根據sample rate和去計算的秒數
    - 回傳節拍出現的秒數np array

練習

- 使用自己輸入的音訊或下載Librosa的範例音訊
- 將音訊的波形和頻譜繪製出來
- 取出音訊的特徵
  - zero crossing rate
  - spectral centroid
  - MFCC(13個特徵)
- 根據每一個音框去儲存這些特徵, 存成一個文字檔
  - 一個音框的特徵[zero crossing rate, spectral centroid, MFCC1, …, MFCC13]總共15個維度
  - 一個音訊的特徵為[15]*frame數量的維度