# 用動態時間校正做特定種類鳥鳴偵測的方法

## An efficient approach to birdsong detection with Dynamic Time Warping

何佩蓁 [1*]、蔡芳升 [2]
[1] 台北市私立東山高級中學 學生
[2] 賽微科技股份有限公司

## 摘要

本論文研究如何在錄音資料中，有效率地偵測出特定種類鳥鳴發生的時間點和次數。在鳥類聲音的分析上，我們比較了幾種在時域和頻域上的特徵，包括過零率、短時能量、梅爾濾波器組(mel-filter banks)、以及梅爾倒頻譜係數(MFCC)。相較於人聲資訊集中於低頻，鳥聲於高頻有較明顯的能量分佈，因此我們也嘗試調整濾波器組的頻帶，以求更精細地描述鳥鳴特徵。在聲音偵測上，我們從錄音資料中選取一段較完整的鳥鳴段落作為參考樣本，用以比對所有錄音資料中的相似片段。我們採用動態時間校正演算法，以其對時間序列壓縮或延展不敏感的特性，避免因鳥鳴時間長短不一而影響相似度計算。本研究用 xeno-canto 網站下載台灣藍鵲資料進行實驗，我們也測試不同強度的噪音對偵測效果的影響。初步結果顯示，對未加額外噪音的原始音檔梅爾濾波器組對鳥鳴偵測有較好的效果，調整濾波器組的頻帶分佈在高噪音下對鳥鳴偵測有顯著的幫助，而在高強度的人聲噪音下倒頻譜係數則有較佳的表現。

**關鍵字**：鳥鳴偵測、動態時間校正、特徵向量、語音辨識

## Abstract

This paper presents an efficient approach to detect when and how many times a specific kind of birdsong has occurred. For birdsong signal analysis, we compared several features both in time and frequency domains, including Zero Crossing Rate, short-time energy, mel-filter banks, and MFCC. Compared to information of human voices which concentrates more in low frequency, birdsongs are more obvious in high frequency. Therefore, we also tried to adjust ranges of filter banks to better depict the features of birdsongs. On detecting specific sound pattern, we manually picked one relatively complete fragment of birdsong in the recording as reference sample to compare with other similar fragments. Although birdsongs from the same species sound similar, their durations might vary once in a while or individually. Thus, we adopted Dynamic Time Warping algorithm, which is insensitive to the extension and compression of time series, to alleviate the mismatch arisen from duration differences. We downloaded birdsong of Taiwan Blue Magpie from xeno-canto website for experiments. Noises in different levels were incorporated in the test data to evaluate the robustness of our approach. Preliminary experiments show that mel-filter banks give better results for the original recording, adjusting filter banks can noticeably improve birdsong spotting rate in noisy environments, and cepstral coefficients is more robust for data of high babble noise.

**Keywords:** Birdsong Detection, Dynamic Time Warping, Feature Vector

## (1) 介紹

透過無人自動錄音器材記錄並調查野生動物資源，已逐漸廣泛應用在野生動物的調查研究上[1]。若干研究亦以語音辨識技術對鳥鳴錄音資料進行處理，如[2]以梅爾倒頻譜係數、高斯混合模型、音高(pitch) 資訊等，用鳥鳴聲做鳥種類的判別。或許透過鳥鳴自動辨識鳥種類的準確率仍不夠理想，目前實務上野外錄音資料的處理仍以人工監聽為主。然而大量的錄音資料以人工 1:1 全時監聽並不符合效益，生態保育工作者多採取成本較低的取樣監聽，在[1]的統計中，以人工取樣監聽日出後與日落後各 15 分鐘加上日出後一小時開始每小時取樣 2 分鐘，即監聽約 5~6%的音檔，在春夏季和秋冬季分別可記錄到原始全部音檔的 73.5%和 55.2%的物種數；若再將音檔以編輯軟體檢視，目視掃瞄夜間時段頻譜，增加監聽量至全部音檔的 7%，前述物種涵蓋率可提升至 80.7%和 70.8%。

雖然取樣監聽可大幅降低人工處理成本，但部分重要訊息也可能因此而遺漏。有別於[2]直接透過鳥鳴做自動鳥種類的判別，本研究試圖利用語音辨識技術在大量錄音資料中迅速找出特定鳥鳴發生的可能時間點位置，此資訊或可協助監聽人員提高工作效率，避免將人力資源浪費在錄音資料中不需監聽的部分。本論文後續章節組織如下：在第二章描述鳥鳴訊號的特徵分析，以及用動態時間校正自動偵測鳥鳴發生時間點的方法；第三章以台灣藍鵲鳥鳴資料進行實驗，除了一般錄音資料外，我們也加入若干不同強度的噪音，分析其對偵測效果的影響；最後是我們從實驗結果中提出的結論。

## (2) 方法

我們在時域和頻域上分析鳥鳴訊號的特徵，時域特徵包括短時能量(STE)和過零率(ZCR)，這兩種特徵也常用於有聲或無聲音的判別(voiced/unvoiced decision) [3]；頻域特徵使用語音辨識領域常用的梅爾濾波器組(mel-filter banks)和梅爾倒頻譜係數(MFCC)，為更精確描述鳥鳴特性，本研究嘗試對梅爾濾波器組進行一些變形。聲音偵測的方法則基於動態時間校正，我們從錄音資料中選取一段較完整的鳥鳴段落作為參考樣本，用以比對所有錄音資料中的相似片段。以下詳述我們對梅爾濾波器組頻帶分佈的修改、動態時間校正演算法、以及自動偵測鳥鳴的做法。

### 一、梅爾濾波器組和倒頻譜係數

語音辨識領域常用的頻域特徵包括梅爾濾波器組(mel-filter banks)和梅爾倒頻譜係數(MFCC)。首先將聲音訊號預強調、音框化、漢明窗，經過快速傅立葉轉換(FFT)後，通過梅爾濾波器組，即得到梅爾濾波器組特徵。梅爾倒頻譜係數是將梅爾濾波器組每個濾波器組的能量取對數，並進行離散餘弦轉換(DCT)至倒頻譜域，梅爾倒頻譜係數就是倒頻譜的幅度[4]。

梅爾濾波器組是根據人類耳蝸的構造所設計，人耳對於低頻的聲音相對較敏感，因此一般梅爾濾波器組在低頻時濾波器分布密集，越往高頻分布越稀疏，如圖 1(a)。然而從圖 2 台灣藍鵲的鳥鳴頻譜可觀察到能量集中在中、高頻，為了讓鳥鳴偵測更精確，我們也嘗試幾種不同分布的梅爾濾波器組加強關注特定頻率的訊號，包括將濾波器集中在高頻和平均分布，分別如圖 1(b)和(c)所示。

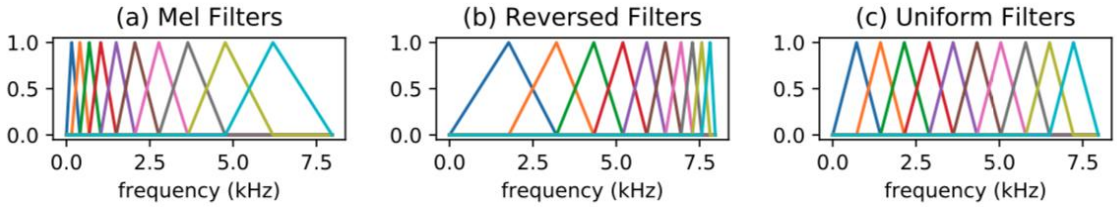


**圖 1 (a)梅爾濾波器組、(b)集中於高頻的濾波器組、(c)平均分佈的濾波器組**

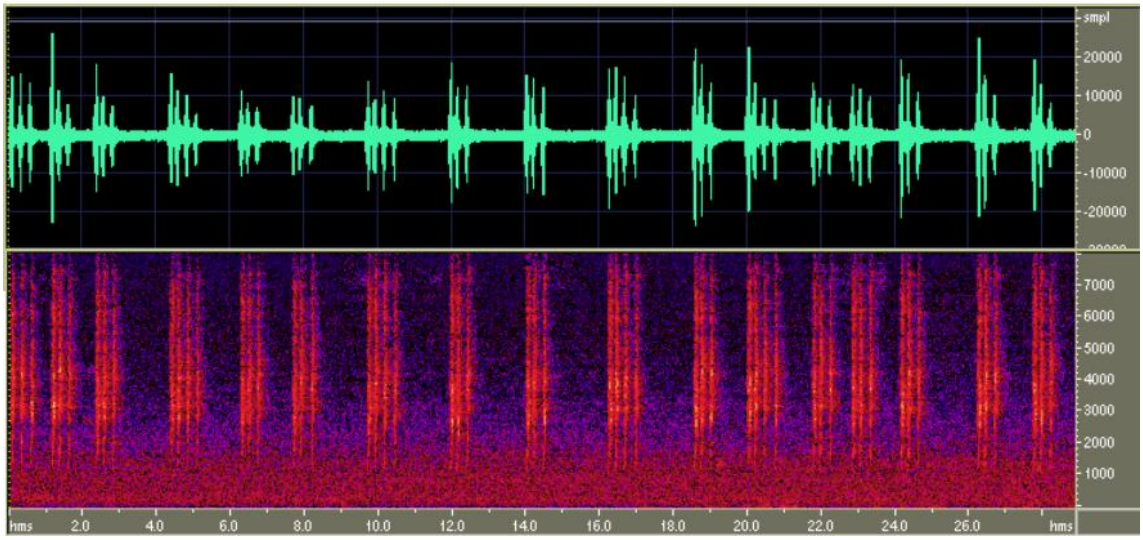


**圖 2 台灣藍鵲的鳥鳴頻譜 (取樣頻率 16kHz)**

二、動態時間校正

動態時間校正的演算法可用於計算兩個時間序列的相似程度，對於不同長度及節奏的時間序列，動態時間校正可以對時間序列進行壓縮或延展，使兩個序列的狀態盡可能一致，找到累積距離最小的路徑，求得最大的相似度[5]。假設參考序列 $C = c_0, c_1, \ldots, c_{m-1}$ 共 $m$ 個向量，測試序列 $Q = q_0, q_1, \ldots, q_{n-1}$ 共 $n$ 個向量，為了使兩個不同長度的時間序列對齊，我們需建構一個 $n \times m$ 的格狀平面，如圖 3 所示。

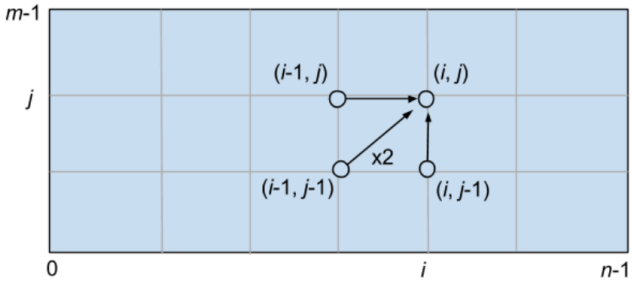


**圖 3 動態時間校正與其路徑行進限制**

節點 $(i, j)$ 表示 $q_i$ 和 $c_j$ 兩個特徵向量的對齊點，其訊號相似度以兩向量的距離 $d(q_i, c_j)$ 估算，距離越小則相似度越高。匹配路徑從左下角 $(0,0)$ 出發，到右上角 $(n-1, m-1)$ 結束。因為語音的前後次序不會因發音的快慢或高低而改變，我們限制

到達節點 $(i, j)$ 的前一個位置必為下方 $(i, j-1)$ 、左方 $(i-1, j)$ 、或在對角線左下方 $(i-1, j-1)$ 三個節點之一。

假設從起點 $(0,0)$ 到 $(i, j)$ 的所有路徑中的最小累積距離為 $a(i, j)$ ，到達 $(i, j)$ 的代價是從前述三個節點的最小累積距離加上 $(i, j)$ 位置的訊號相似度 $d(q_i, c_j)$ ，由於 $(i-1, j-1)$ 走對角線到 $(i, j)$ 相較於另兩個位置少了一步的距離，為公平地估算距離，此行進方向的成本要乘以 2，即 $2 \times d(q_i, c_j)$ ，如此不論走法為何，所有路徑皆為 $n+m-1$ 個特徵向量距離的總和。此外，為了使路徑盡量往對角線方向行進，我們對於往右和往上走的路徑加上一個懲罰值 *penalty*，以盡可能得到合理的相似度。綜上所述，路徑上每個節點的最小累積距離如下：

$$a(i, j) = \min \begin{Bmatrix} a(i, j-1) + d(q_i, c_j) + penalty \\ a(i-1, j) + d(q_i, c_j) + penalty \\ a(i-1, j-1) + 2 \times d(q_i, c_j) \end{Bmatrix}$$

路徑終點的 $a(n-1, m-1)$ 就是 $Q$ 和 $C$ 比對的最小累積距離。本論文圖表呈現此距離時會除以過程中累加 $d(q_i, c_j)$ 的次數，即除以 $(n+m-1)$ ，以排除 $Q$ 和 $C$ 時間序列長度對相似度估算的影響。

## 三、自動鳥鳴偵測

要從一段長時間錄音檔偵測所有鳥鳴發生的時間點，首先我們從錄音資料中選取一段較完整鳥鳴段落，長度為 $m$ 的音訊作為參考樣本 $C$，接著從錄音資料時間 0 開始，每隔一小段時間取長度為 $n$ 的為測試資料 $Q$，$n$ 略大於 $m$，並用 DTW 計算 $Q$ 和 $C$ 的距離作為其相似度，重複上述做法直到整個音檔結束。圖 4 是以圖 2 前 11 秒的錄音資料進行測試，每隔 0.1 秒用 DTW 計算一次距離，畫出每個時點估算的相似度。藍線是各時間點與鳥鳴參考樣本 $C$ 的相似度，值越低越近似，其中每個相對低點都是鳥鳴可能發生的時間點。

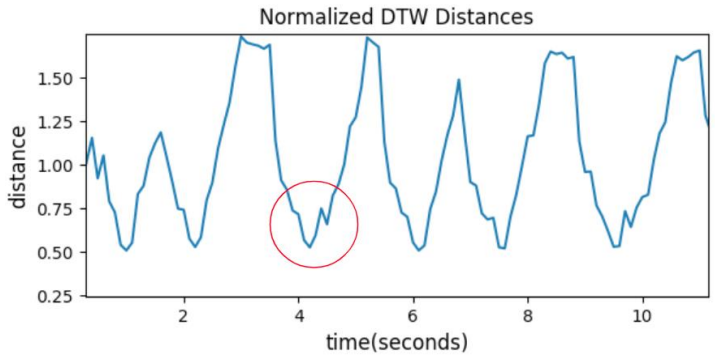


**圖 4 測試音檔各時間點與參考樣本 *C* 用 DTW 距離估算的相似度走勢**

我們將圖 4 的紅圈處放大如圖 5，在該處有兩個相對低點 A 和 B，但參照圖 2 在時間 4 至 5 秒的位置應該只有一次鳥鳴聲，且發生時間點為 A 所在位置。在此我們對每個相對低點加上一個門檻值δ為上限，即圖 5 的紅色虛線，可切下一個倒錐形藍色曲線段，限定每個被切下的倒錐形曲線段僅發生一次鳥鳴，可找

出其最小值位置 A，以此位置為判定該次鳥鳴發生的時間點，即可排除位置 B 的相對低點。若此門檻值δ訂得太低，B 點也會被判定為鳥鳴位置，而產生較多的誤接受率(false acceptance rate, FAR)，相反地訂得太高則產生過高的誤拒絕率(false rejection rate, FRR)。實際系統運行時應由人工判斷並選定合適的δ值，但為避免人為判斷不夠精確而影響實驗數據的判讀，本論文的實驗會測試各種可能的門檻值δ，在不超過某個預設的誤接受率(FAR)的情況下，取最低的誤拒絕率作為最佳的實驗結果。

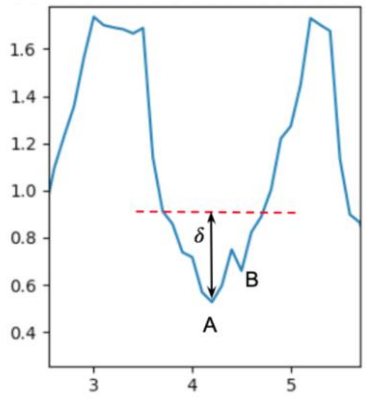


**圖 5 以門檻值δ限定鳥鳴發生的範圍**

**(3) 實驗**

本實驗的測試資料是從 xeno-canto 網站(https://www.xeno-canto.org)下載數個台灣藍鵲的鳥鳴資料串成約 3 分 22 秒的測試音檔，並人工標註鳥鳴發生的時間點作為比對偵測率的答案，共標示出 109 次鳥鳴，大部分鳥鳴發出 2 至 5 次聲響，持續時間約 0.5 至 1.2 秒。除了前述原始音檔，我們也加入兩種噪音觀察其偵測效果，分別為展場錄製的人聲噪音和車流不大的路旁行道樹下錄製蟬鳴的噪音，測試訊噪比(SNR)在 10dB、5dB、0dB、-5dB 下的偵測率。

測試音檔取樣率為 16kHz，我們使用若干種不同的特徵向量進行實驗：時域特徵包括短時能量(STE)和過零率(ZCR)，及其一階和二階的差分，分別組成 3 維度的向量；頻域特徵包括如圖 1(a)(b)(c)所示的 40 維度梅爾濾波器組(mel filters)、較集中於高頻濾波器組、平均分佈濾波器組，以及用這些濾波器組產生的 16 維度的梅爾倒頻譜係數(MFCC)。以上實驗每秒鐘的音檔皆抽取出 100 個特徵向量。

用 DTW 計算相似度時，每次時間點取出的測試資料 $Q$ 長度 $n$ 為參考樣本 $C$ 長度 $m$ 的 1.2 倍。為方便設定實驗參數，計算向量距離時會除以向量維度，最終的 DTW 總距離再除以 $(n+m-1)$。除此之外，為了得到較合理的匹配路徑，我們讓 DTW 路徑往右和往上走時加上些許的懲罰值，以增加其往對角線方向行進的機會。

本論文計算正確率的時間容許條件為 0.5 秒，即偵測到的鳥鳴發生時間與人工標註答案必須在 0.5 秒內才算正確。未偵測到的鳥鳴佔所有測試資料鳥鳴次數的比率即誤拒絕率(FRR)，亦可視為辨識錯誤率，本論文即以此錯誤率為偵測效果的評估指標，越低越好。一般而言錯誤率與誤接受率(FAR)呈負相關，即可容許的 FAR 越高，錯誤率就會越低，而能偵測出越多的鳥鳴發生。本論文的實驗皆以 FAR 須 10%以下為前提，測試各種可能的鳥鳴相似度門檻值δ，取最低的錯誤率為偵測結果數據。

首先測試時域特徵的過零率和短時能量，表 1 和表 2 分別為其在蟬鳴加車噪和人聲噪音環境下的錯誤率。過零率的效果忽好忽壞，不甚穩定，而短時能量隨著噪音的增大，辨識效果大幅衰退。基本上，時域特徵在鳥鳴偵測上的效果皆不理想。

**表 1 時域特徵在蟬鳴加車噪環境下辨識的錯誤率**

| 蟬鳴加車噪 | 原始音檔 | SNR 10dB | SNR 5dB | SNR 0dB | SNR -5dB |
|---|---|---|---|---|---|
| 過零率 | 41.3% | 55.0% | 23.9% | 17.4% | 28.4% |
| 短時能量 | 20.2% | 40.4% | 63.3% | 85.3% | 88.1% |

**表 2 時域特徵在人聲噪音環境下辨識的錯誤率**

| 人聲噪音 | 原始音檔 | SNR 10dB | SNR 5dB | SNR 0dB | SNR -5dB |
|---|---|---|---|---|---|
| 過零率 | 41.3% | 52.3% | 60.6% | 57.8% | 53.2% |
| 短時能量 | 20.2% | 30.3% | 51.4% | 86.2% | 89.9% |

接著測試頻域特徵，其在蟬鳴加車噪的環境下效果如表 3 所示。整體而言濾波器組的辨識效果優於倒頻譜係數，各種濾波器對原始音檔都有不錯的表現，偵測率可達九成，其中最佳者為梅爾濾波器，錯誤率僅 8.3%，而加了噪音的測試資料則以平均分佈的濾波器表現最穩定。相較於傳統重視低頻資訊的梅爾濾波器組，平均分佈的濾波器或許也能精確描述鳥鳴訊號的特徵，且對鳥鳴偵測有更佳的抗噪音效果。

**表 3 頻域特徵在蟬鳴加車噪環境下辨識的錯誤率**

| 蟬鳴加車噪 | 原始音檔 | SNR 10dB | SNR 5dB | SNR 0dB | SNR -5dB |
|---|---|---|---|---|---|
| 梅爾濾波器 | **8.3%** | 12.8% | 12.8% | 26.6% | 35.8% |
| 集中高頻濾波器 | 11.0% | 11.0% | 11.9% | 41.3% | 38.5% |
| 平均分佈濾波器 | 11.9% | **10.1%** | **12.8%** | 19.3% | **16.5%** |
| 梅爾倒頻譜 | 21.1% | 25.7% | 24.8% | 37.6% | 50.5% |
| 倒頻譜－集中高頻 | 33.9% | 26.6% | 19.3% | **16.5%** | 22.9% |
| 倒頻譜－平均分佈 | 23.9% | 24.8% | 22.9% | 26.6% | 27.5% |

人聲噪音相對蟬鳴在頻譜上有更大的變化，一般而言對語音辨識效果的影響也較大。頻域特徵在人聲噪音環境下的辨識效果列於表 4，整體而言，濾波器組的辨識效果在 SNR 為 5dB 以上的中低度噪音時優於倒頻譜係數，但在 SNR 為 -5dB 的高強度噪音下，濾波器組完全不堪使用，錯誤率高達六成，此時倒頻譜係數則有較佳的表現，尤其以集中高頻的濾波器所產生的倒頻譜係數效果最好，在 SNR 為 0dB 和在-5dB 時的錯誤率分別僅 15.6%和 23.9%。

本實驗從 xeno-canto 所下載的原始音檔是由一般民眾自發性上傳提供，大多有參雜一些背景噪音，且使用的錄音設備不一，不同的通道效應也讓這些背景噪音產生更多的差異。在表 3 和表 4 可觀察到加了些許測試噪音讓 SNR 為 10dB，有時辨識效果反而會比原始音檔好，我們推測可能這些外加的噪音讓原始

音檔內的噪音模糊化，彼此差異變小，相對地對辨識造成的影響也隨之降低。但隨著外加噪音強度提高，逐漸影響原有鳥鳴訊號，辨識效果也就跟著反轉向下。

**表 4 頻域特徵在人聲噪音環境下辨識的錯誤率**

| 人聲噪音 | 原始音檔 | SNR 10dB | SNR 5dB | SNR 0dB | SNR -5dB |
|---|---|---|---|---|---|
| 梅爾濾波器 | 8.3% | **9.2%** | 16.5% | 33.0% | 63.3% |
| 集中高頻濾波器 | 11.0% | 11.0% | 12.8% | 18.3% | 59.6% |
| 平均分佈濾波器 | 11.9% | 11.0% | **11.9%** | 16.5% | 64.2% |
| 梅爾倒頻譜 | 21.1% | 18.3% | 14.7% | 23.9% | 36.7% |
| 倒頻譜－集中高頻 | 33.9% | 20.2% | 17.4% | **15.6%** | **23.9%** |
| 倒頻譜－平均分佈 | 23.9% | 20.2% | 23.9% | 25.7% | 41.3% |

## (4) 結論

本論文以台灣藍鵲的鳥鳴為測試標的，描述如何使用動態時間校正演算法有效率地偵測出錄音資料中鳥鳴發生的時間點。我們實驗若干聲音特徵的分析方法，發現調整濾波器組的頻帶分佈，相較於傳統重視低頻資訊的梅爾濾波器，或許更能精確地描述鳥鳴特徵，在噪音的環境下能得到更好的偵測效果。而在高強度的人聲噪音下，倒頻譜係數的偵測效果則優於一般濾波器組。本論文實驗用的台灣藍鵲錄音資料量體不夠大，還需要更大量的資料、經過更多的實驗來驗證我們提出的方法的可靠性。雖然如此，我們相信本論文的做法是合理且應該有效果的，繼續深入研究，或許在未來能協助生態保育工作人員節省錄音資料監聽的時間，而發揮更大的工作效率。

## (5) 參考文獻

1. 姜博仁, 蔡哲民, 蔡世超, 吳禎祺, 鄭蕙如 (2015) 錄音技術應用於野生動物調查之應用與評估, 臺灣林業 41:4 2015.08 頁 33-38

2. 廖偉恩, 黎欣捷, 蔡偉和 (2011) 應用語音辨識技術於鳥鳴聲辨識, The 23rd Conference on Computational Linguistics and Speech Processing (ROCLING 2011)

3. Bachu R., Kopparthi S., Adapa B., Barkana B. (2010) Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. In: Elleithy K. (eds) Advanced Techniques in Computing Sciences and Software Engineering. Springer, Dordrecht

4. Haytham Fayek (2016) “Speech Processing for Machine Learning: Filter banks, Mel-Frequency Cepstral Coefficients (MFCCs) and What's In-Between”, https://haythamfayek.com/2016/04/21/speech-processing-for-machine-learning.html

5. H. Sakoe , S. Chiba (1978) “Dynamic programming algorithm optimization for spoken word recognition”, IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume: 26, Issue: 1, 1978.