A Novel Algorithm for Speaker Change Detection Based on Support Vector Machine

以支援向量機爲基礎之新穎語者切換偵測演算法

王駿發、林博川、王家慶、宋豪靜

wangjf@mail.ncku.edu.tw,tony@icwang.ee.ncku.edu.tw

國立成功大學電機研究所

摘要

對於不同的語者的切換，我們可以利用其不同之語音特徵來加以區別，本論文提出一個以支援向量機(support vector machine, SVM)爲基礎的新穎語者切換偵測演算法；我們定義一個「SVM 訓練分類錯誤率」來判斷語者之料之間的可分離性，藉此判斷是否爲同一個語者的聲音資料。實驗證明我們提出的演算法比貝氏訊息準則(Bayesian information criterion, BIC)演算法有更好的偵測能力並且有更低的誤報率(false-alarm rate)。同時對於兩秒鐘以下的語者變換也可以有效的加以判斷。

關鍵字: 支援向量機(support vector machine, SVM)、貝氏訊息準則(Bayesian information criterion, BIC)、語者切割(speaker segmentation)、語者切換點偵測(speaker change detection)。

I. 簡介

一般而言，對於不同的語者的切換，可利用其不同語音之特徵來加以區別，這種預處理的動作在廣播新聞分類、語音辨識、電話語音分類、自動字幕系統、自動會議記錄、語者識別、語者追蹤(speaker tracking)、語者聚類(speaker clustering)、口述語言資料檢索(spoken document retrieval, SDR)等都有很大的幫助。因此目前有相當多此類的語者分段(speaker segmentation)研究[1-2], [8], [6-12], [16-20], [22], [27], [29-32]。

同時，新聞廣播之聲訊信號源是目前常被研究的信號來源[8-9], [17], [20], [22], [24], [27], [33-37]，因爲其多樣性（包含純音樂、純語音、窄頻語音、具有背景環境或是噪音的語音…等）。而切換點的偵測主要是依照不同的語者、環境、channel等加以標記，最後將相同語者的訊號做群聚(clustering)與合併(merging)。相反地；如果只想取出純音樂的段落可以將非音樂段加以刪除。

對於一個語音串列(speech stream)，事前預蒐集聲學或是語者的模型不是很好的方法且有其困難性，因此不需要事先收集語者資料與任何模型或是訓練的偵測方法(unsupervised manner)是必須的[6], [11]。目前已經有許多研究從事於 unsupervised語者切割之研究 [1], [6], [8], [10-11], [24]。而這些方法主要分爲metric-based、model selection-based與energy-based三類:

A. Energy-based法

一般對話系統的行爲模式中，語者切換點之間通常有靜音段存在；利用能量的大小來判斷切換點[8], [11], [18-19]是很直覺也很簡單的方法。

B. Metric-based法

此方法在語音串列以平移的方式 [3-4], [8-9], [16-17], [27]，使用許多聲學距離：如 Kullback-Leibler distance (KL, KL2) [8], [28], generalized likelihood ratio (GLR) [10], [31], Mahalanobis distance 與 Bhattacharyya distance [9] 來評估兩個相鄰window的相似度；藉此產生

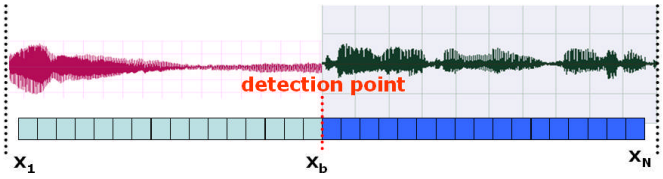
距離曲線(distance curve)；經由一個低頻濾波器濾除因爲雜訊造成的微小波動後[3], [9-11]；取出區域最大值的時間點做爲切換點的輸出。雖然可以有效的加速判斷，但也有許多缺點：(1)需要一個臨界值來選取區域最大值，無法確保所有的語音訊號都適用；(2)只利用鄰近兩個window蒐集的資料來判斷相似度；(3)若使用貝氏訊息準則爲基礎來做相似度判斷[3]；則爲了有充足的資訊量；使得window大小必須大於兩秒鐘；造成對於一秒鐘以下的切換點無法有效偵測[10]。

C. Model selection-based法

由 Schwarz[14]所提出的貝氏訊息準則爲一以模型複雜度加以懲罰(penalize)的可能性準則(likelihood criterion)，被廣泛使用在語者切換點偵測上[1-2], [4] , [6-7], [10], [12-13], [19-23]，對於一個模型 M_i 而言，BIC 定義如下

$$BIC(M_i) = \log P(D_1, D_2, \cdots, D_N \mid M_i) - \frac{1}{2} d_i \log N \text{.....(1)}$$

模型選擇的問題是要從候選模型 M_i， $i = 1,2,\cdots,m$ 中選擇一個來表示一個資料集合 $D = (D_1, D_2, \cdots, D_N)$，$d_i$ 是模型參數集合中獨立的參數個數，$P(D_1, D_2, \cdots, D_t \mid M_i)$ 是模型的最大資料概似(maximized data likelihood)，而 $\frac{1}{2} d_i \log N$ 項的相減動作是模型由 log-likelihood 用來懲罰(penalize)模型複雜度之用。



圖一、使用貝氏訊息準則做語者切換點偵測方法

而使用貝氏訊息準則做語者切換點偵測方法如圖一所示，定義 $X = x_i \in R^d, i = 1,2,\cdots,N$ 是只有一個語者切換點的語音串列；每個 frame 都有一組 cepstral 向量。假設在 $b \in (1, N)$ 這個 frame 有一個切換點，如果可以假設每個聲學 homogeneous 區塊都可以用 multivariate Gaussian process $X \sim N(\mu, \Sigma)$ 加以模型化。則語者切換點的偵測可以視爲一種介於以下巢狀模型(nested model)之模型選擇問題[11]：

$$\begin{aligned} & M_1 : X : x_1, x_2, \cdots, x_N \sim N(\mu, \Sigma) \\ & and \\ & M_2 : x_1, x_2, \cdots, x_b \sim N(\mu_1, \Sigma_1); \\ & x_{b+1}, x_{b+2}, \cdots, x_N \sim N(\mu_2, \Sigma_2) \text{.....(2)} \end{aligned}$$

其中，在 X 中假設所有的取樣都是獨立且類似一個高斯分佈，而 M_1 中假設前 b 個取樣也是一個高斯分佈，而 M_2 中假設最後的 *N-b* 個取樣則是另一個高斯分佈。而切換點 b 的判斷是利用兩個模型差值 $\Delta BIC(b)$ 的正負號來判斷：

$$\Delta BIC(b) = B\bar{I}C(M_2) - B\bar{I}C(M_1) \text{.....(3)}$$

$$= \frac{1}{2}(N \log\left|\hat{\Sigma}\right| - b \log\left|\hat{\Sigma}_1\right| - (N-b)\log\left|\hat{\Sigma}_2\right|) - \frac{1}{2}\lambda(d + \frac{1}{2}d(d+1))\log N .....(4)$$

其中 $\hat{\Sigma}$、$\hat{\Sigma}_1$ 與 $\hat{\Sigma}_2$ 是由相對應的資料所估算出的 ML covariance，λ 是懲罰係數(penalty factor)以補償少量取樣的情況，d 是 cepstral 參數的維度。根據貝氏訊息準則，如果 $\Delta BIC(b) > 0$ 則代表 b 是一個語者切換點，經由 MLE (Maximum Likelihood Estimation)，最終的 BIC 語者切換點判斷式如下:

$$\hat{b} = \arg \max_{1<b<N, \Delta BIC>0} \Delta BIC(b) .....(5)$$

然而Chen的方法[1]在運算量上是屬於二次複雜度(quadratic complexity)，無法實用在real time的系統之中。此外，BIC往往需要足夠的資訊蒐集量才足以判斷出不同語者的切換點，對於較短時間的語者片段無法有效的切割[10]。

在這一篇論文中，我們針對 unsupervised 語者切割提出一個以支援向量機爲基礎的新穎語者切換偵測演算法；定義一個 SVM 訓練分類錯誤率來判斷語者資料之間的可分離性，我們稱其爲「以 SVM 可分離性爲基礎(separability-based)之語者切換偵測演算法」；藉此判斷是否爲同一個語者的聲音資料。本論文的架構如下:在第 **II** 節之中我們對 SVM 演算法做一回顧並提出一個 separability-based 之語者切換偵測演算法，在第 **III** 節之中，SVM 與 BIC 之語者切換點偵測能力將做一系列的評估，而我們所提出的演算法將在第 **IV** 節之中做完整描述，第 **V** 節是實驗環境介紹與實驗結果，最後在第 **VI** 節中做總結。

II. 應用 SVM 訓練錯誤率判斷之語者切換點偵測演算法

A. SVM 演算法簡介

支援向量機(support vector machine, SVM) 是一個新穎的統計學習方法；且近年來引起越來越多研究學者的注意[5][15][25-26][38-39]。SVM 是基於"結構風險最小化" (structural risk minimization, SRM)所構思的歸納原理[40]，主要是以歸納方式求取最小化邊界的錯誤爲目的；而不是以方均誤差最小化爲主。在許多應用之中；SVM 已被證實比傳統 learning machines 有更好的效果，且在分類問題[41]與回歸問題[42]上，被當作是有利的工具。例如在分類議題上；獨立字之數位手寫辨識[25],[43]、語者確認、人臉辨識、知識基礎分類器(knowledge-based classifier)[44]、文件分類(text categorization)[45-46]都是SVM的應用範疇。在迴歸估計(regression estimation)方面，SVM對於benchmark time series prediction tests [47-48]、financial forecasting [49-50]與Boston housing problem [51]都很有競爭力。

本論文使用到的 SVM 演算法部分，是將 l 筆訓練資料(每筆資料 $x_i \in R^N$)利用超平面(hyperplane)分類成爲 $-1,1$ 兩類，而 $y_i \in \{-1,1\}$ 是分類的標記。Hyperplane 的定義爲 $f_H(\mathrm{x}) = \mathrm{w}\cdot\mathrm{x} + b$ ，而最佳的 w 標記爲 $\overline{\mathrm{w}}$ ；求法如下:

$$\overline{\mathrm{w}} = \sum_{i=1}^{l} \overline{\alpha}_i y_i \mathrm{x}_i .....(6)$$

其中 $\overline{\alpha}_i$ 是求解 hyperplane 中導入的 Lagrange multipliers。求出最佳的 hyperplane 後；利用下列公式來加以將資料分類：

$$f_D(\mathrm{x}) = sign(f_H(\mathrm{x})) = sign(\sum_{i=1}^{l} \bar{\alpha}_i y_i \mathrm{x}_i \cdot \mathrm{x} + \bar{b}).....(7)$$

對於無法使用線性方程式將資料分類的情況，我們必須將樣本 x 轉換到高爲度特徵空間 Z 中來處理，即 $x \to \varphi(x) : R^N \to Z$ ，新的 hyperplane 爲：

$$f_H(\mathrm{x}) = \bar{\mathrm{w}} \cdot \mathrm{z} + \bar{b} = \sum_{i=1}^{l} \bar{\alpha}_i y_i K(\mathrm{x}_i, \mathrm{x}) + \bar{b}.....(8)$$

判斷函數爲

$$f_D(\mathrm{x}) = sign(f_H(\mathrm{x})) = sign(\sum_{i=1}^{l} \bar{\alpha}_i y_i K(\mathrm{x}_i \cdot \mathrm{x}) + \bar{b}).....(9)$$

其中 $K(*,*)$ 稱爲 kernel functions；常用的有以下四種：

Linear Kernel: $K(\mathrm{x}, \mathrm{y}) = \mathrm{x} \cdot \mathrm{y}.....(10)$

Polynomial: $K(\mathrm{x}, \mathrm{y}) = (\gamma \cdot \mathrm{x} \cdot \mathrm{y} + c)^d .....(11)$

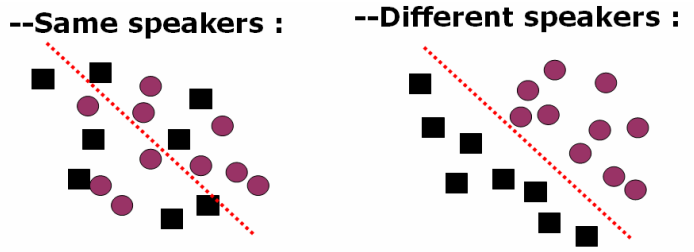
Gaussian Radial Basis Kernel: $K(\mathrm{x}, \mathrm{y}) = \exp(-\gamma \cdot |\mathrm{x} \cdot \mathrm{y}|^2).....(12)$

Sigmoidal Neural Network Kernel: $K(\mathrm{x}, y) = \tanh(\gamma \cdot \mathrm{x} \cdot \mathrm{y} + c)....(13)$

其中 γ 與 c 是常數，而 d 是維度。

B. 以 SVM 之可分離性爲基礎之語者切換偵測演算法

如圖二所示，對於不同的語者而言，可以利用 hyperplane 來將資料分成兩類(+1 與-1)：相反地，如果是同一個語者的語音特徵，hyperplane 將無法有效的將資料分成兩類。



圖二、使用 hyperplane 對相同/不同語者分類示意圖

因此我們可以利用 SVM 在訓練的過程中所找到的 hyperplane 用於計算訓練分類錯誤率(training misclassification rate)來作爲語者切換點的判斷，SVM 的 training misclassification rate 可以分爲兩種，分別是(-1)類誤判爲(+1)的比率與(+1)類誤判爲(-1)的比率；我們分別將其定義爲 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 。

對於同一個語者的語音特徵而言，由於系統所訓練出來的 hyperplane 無法有效的將資料分爲兩類； $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 都會相當高；相反地；對於不同語者的語音特徵而言，由於系統所訓練出來的 hyperplane 可以有效的將資料分爲兩類；因此 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 都會驅近於零，此判斷方式對於語者切換點的偵測能力(detectability)我們將於 III 中測試並評估。

以下詳細說明如何利用 SVM 的 training misclassification rate 來偵測語者切換點，由圖三所示:

Step 0: 將一個語音段 $X = \{x_i : i = 1, \cdots, N\}$ 的每個 frame 取出其語音特徵參數(13 階 MFCCs)。

Step 1:在虛線處(第 *k* 個 frame)假設有一個語者切換點(是不是真的切點還不知道)，立即將虛線左邊 window 的每個 frame 都全部分別標記成 -1 即 $X^- = \{x_i : i = 1, \cdots, k\} \in tag(-1)$；同時將虛線右邊 window 的每個 frame 都全部分別標記成 +1 即 $X^+ = \{x_i : i = k+1, \cdots, N\} \in tag(+1)$。

Step 2:利用 SVM 的訓練過程找出 hyperplane，嘗試將兩個 window 的語音段分開。

Step 3:利用剛才找到的 hyperplane 實際的來將所有 fraeme 分類並標記。假設(-1)類誤判爲(+1)的個數爲 *p*，(+1)類誤判爲(-1)的個數爲 *m*

Step 4:分別計算 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$，其中

$$mis^+(\hat{R}) = m \big/ N - k \quad ....(14)$$

$$mis^-(\hat{R}) = p / k \quad .....(15)$$

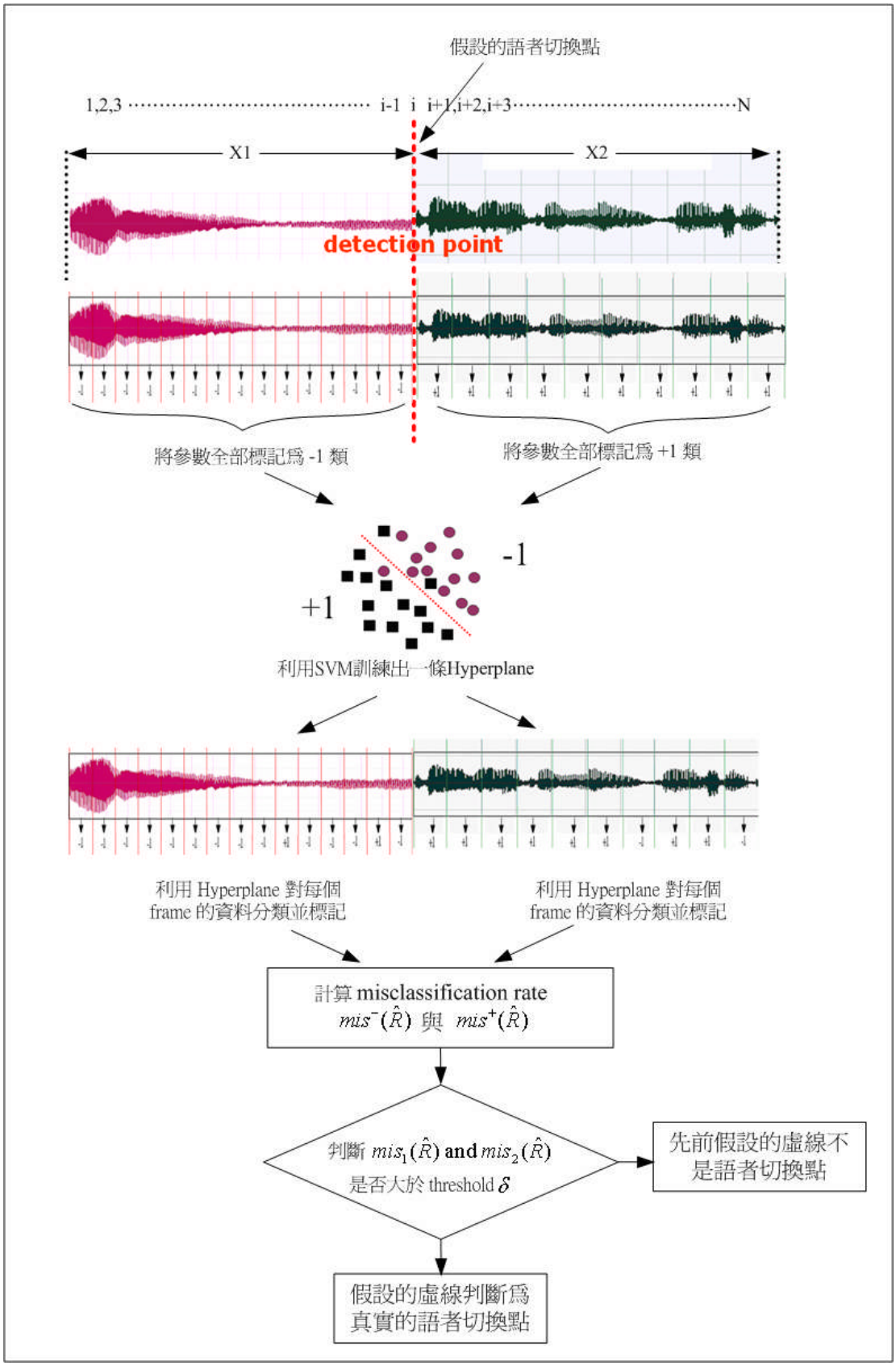
Step 5:判斷 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 使否都小於一個臨界值 δ；成立則判斷 Step 1 所假設的虛線爲真實的語者切換點，不成立則代表 Step 1 的假設錯誤。

III. SVM 與 BIC 之語者切換點偵測能力評估

我們做了以下實驗來評估 SVM 與 BIC 兩個演算法對於語者切換點之偵測能力[1]，實驗中；我們將 BIC 的懲罰係數(penalty factor) λ 設定爲 1[1]。

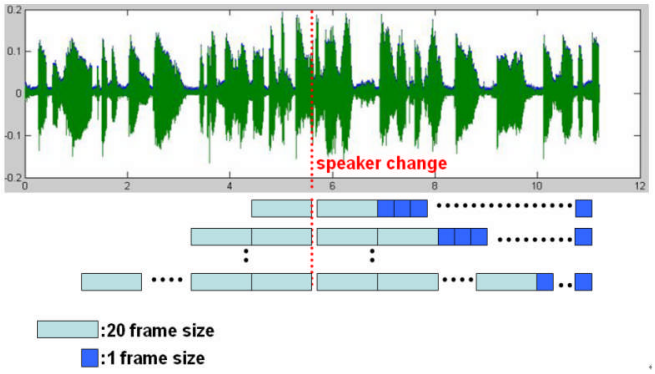
A. 使用左右不同尺寸之 window 來偵測一個已知的語者切換點

本實驗使用了一個 20 秒鐘的語音段(含有兩位語者) 來做爲實驗對象，每位語者分別說 10 秒鐘的語音，由圖四(a)所示，虛線處代表實際的語者切換點。我們從切點處左右挑選出不同的 window 大小來做實驗。實驗中我們先固定切點左邊 window 的大小初值爲 M 爲 20 個 frame；同時定義一個右邊 window 擴增動作來測試，此擴增動作如圖四(a)所示:切點右邊 window 從 M 個 frame 開始，每次增加一個 frame，一直增加到 350 個 frame 爲止。圖(b)與(c)的 X 軸代表切點右邊 window 擴增的大小，而每次挑出的 window 都分別將 ΔBIC 值與 SVM 的兩個 training misclassification rate 即 $mis^-(\hat{R})$ 與 $mis^+(\hat{R})$ 畫在 Y 軸。當切點右邊的 window 每做完一次擴增後，我們將切點左邊 window 的大小 M+10 並重複上述的擴增動作，一直增加到 M=100 爲止，並以不同的顏色來代表。由於 ΔBIC 對於較接近 window 邊界的切點或是 window 收集量太少的情況下判斷力會因爲統計的距離不足而變差[3],[11]，但是如果增加 window 資料收集量將會面臨一個不可避免的問題：那就是 miss detection rate 也將提高[3]。因此這個實驗主要評估兩個驗算法在資料收集量不同下的切點偵測能力。

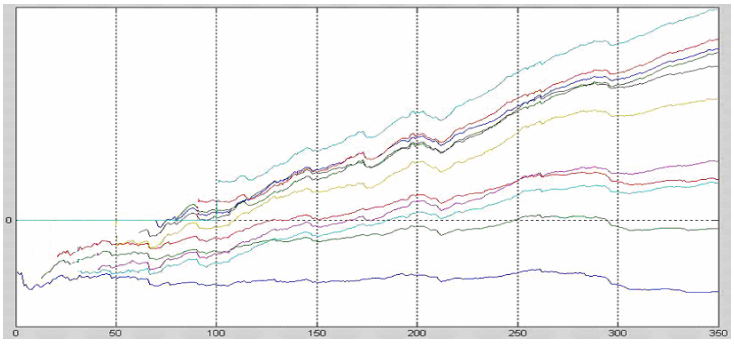


圖三、利用 SVM 的 training misclassification rate 來偵測語者切換點

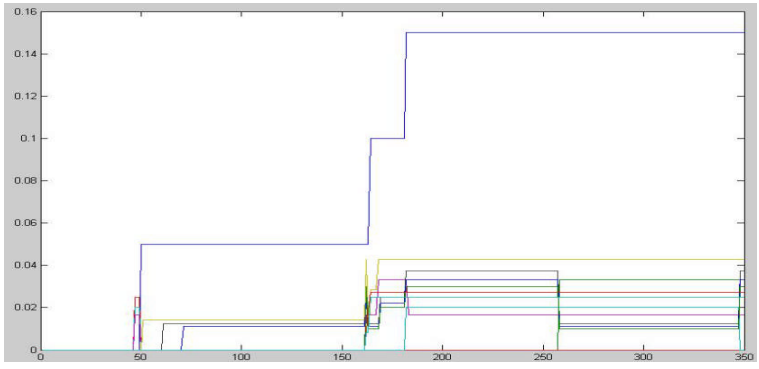
由圖四(b)我們可以發現 BIC 演算法必須要求切點左邊的 window 大小 M 必須大於 90 個 frame 以上(大於一秒鐘)才能使得 ΔBIC 大於零。而 SVM 如圖無論 window 大小如何，都可以得極低的 training misclassification rate；這樣的分佈意味著 SVM 即使在資料收集量極低的情況下仍然可以有良好的判斷力。



(a)



(b)



(c)

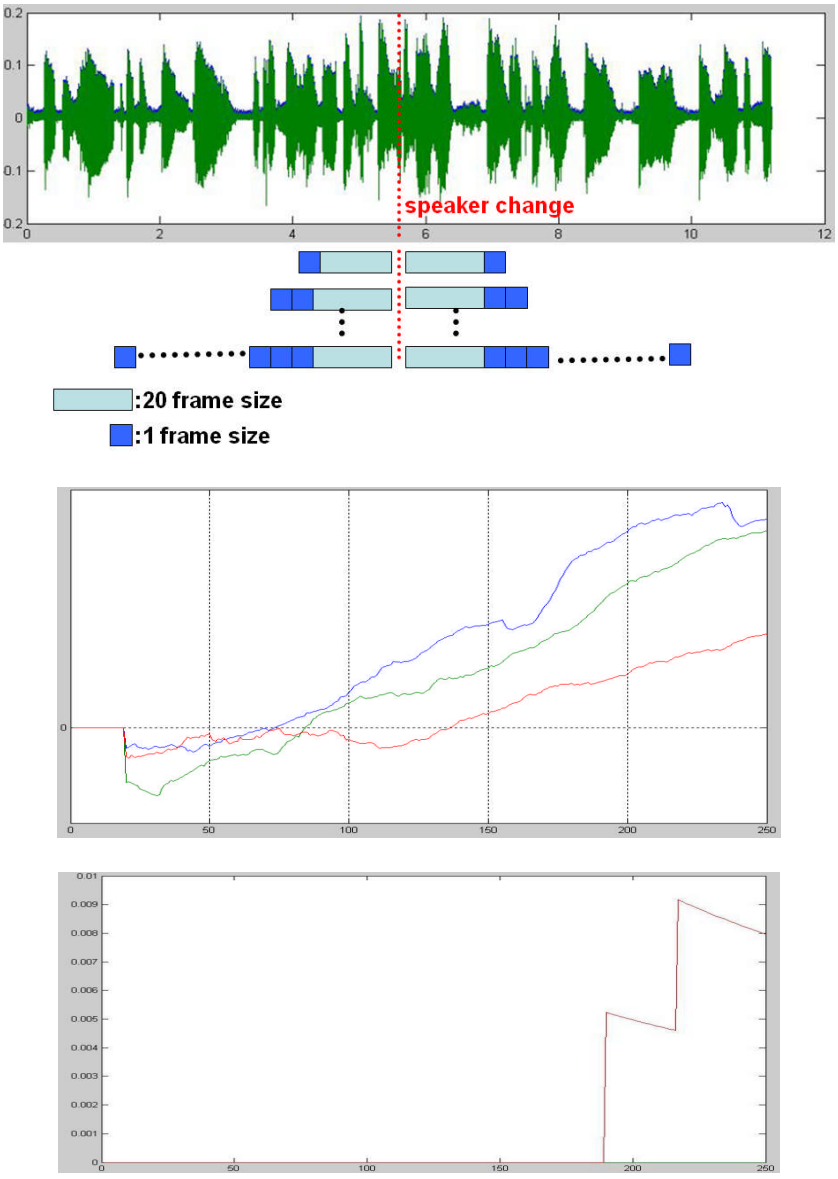
圖四、使用左右不同尺寸之 window 來偵測一個已知的語者切換點

B. 使用左右相同尺寸之 window 來偵測已知的語者切換點

本實驗使用了三個 20 秒鐘的語音串(每個語音串含有兩位語者)來作爲實驗對象，每位語者分別說 10 秒鐘的語音，由圖五所示，虛線處代表實際的語者切換點。我們從切點處左右挑選出相同的 window 大小來做實驗，X 軸代表切點左右兩邊 window 從 20 個 frame 開始，每次切點兩邊的 window 增加一個 frame，一直增加到 250 個 frame 爲止。Y 軸表示每次挑出的 window 所對應的 ΔBIC 值與 SVM 的 training misclassification rate。

由圖五顯示 ΔBIC 值必須要 window 左右兩邊分別收集到大約 137 個 frame(大於一秒鐘)以

上才能保證三個音檔的切點都被偵測出來，這意味著如果不同的語者段落若小於一秒以下將無法被偵測。而 SVM 的效果就相當理想，當 window 收集的資料量在很小的情況下，training misclassification rate 都等於零；亦即 SVM 幾乎都可以百分之百的將資料分成兩類(兩個語者)，也就是 SVM 即使在資料收集量極低的情況下仍然可以有良好的判斷力。唯有其中一個測試語音；當資料量收到大於 190 frames (1.5 秒)時，會有少部分資料呈現誤分類(misclassified)情形，進而造成 training misclassification rate 稍微上升。



圖五、使用左右相同尺寸之 window 來偵測已知的語者切換點

C. 以固定 window 大小掃描之結果分佈來評估 SVM 與 BIC 之易測性

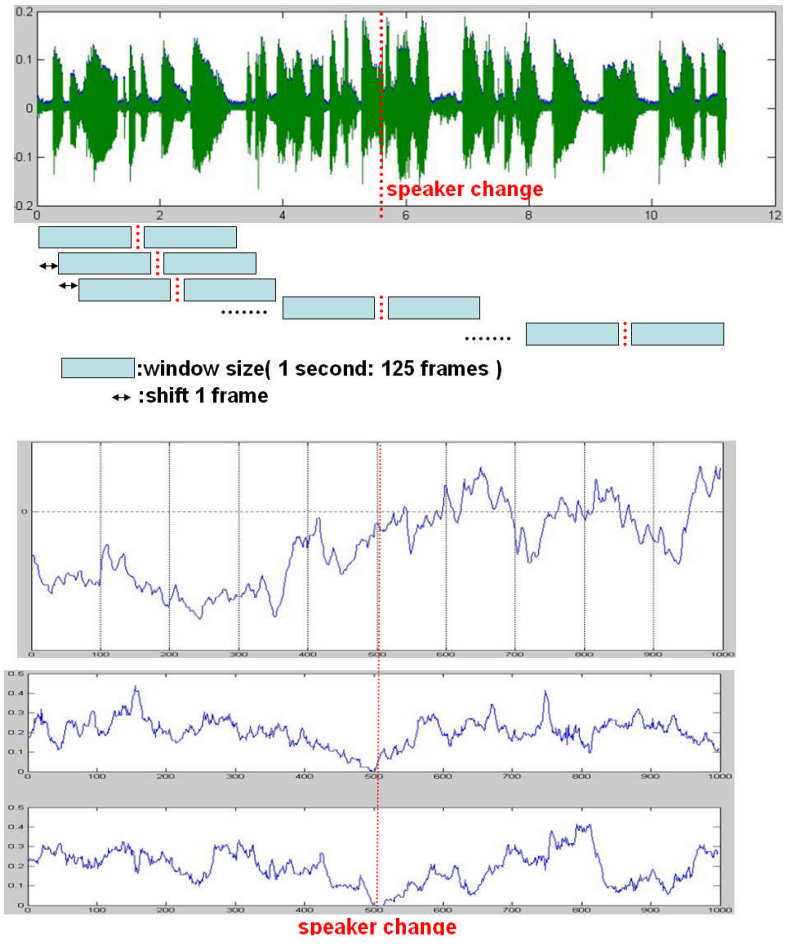
Metric-based 法[3-4], [8-9], [16-17], [27]以假設語者切換點的位置(虛線處)左右相鄰之固定大小 window 來計算 ΔBIC 值，利用 sliding window 的方式掃描出一條 ΔBIC 值曲線，藉由 local peak 來判斷切換點。而這樣的掃描方式，我們恰巧可以套用在 II.***B*** 所提出的判斷方式上；即以 SVM 演算法的 training misclassification rate 曲線來判斷切換點。使用這種方式來偵測切換點的另一個理由是：這種掃描方法可以避免 error broadcasting (先前 windows 的切點判斷錯誤將影響下一個

window 的判斷)並且比 windows 擴張法(expanding scheme)[1-2], [6-7], [11], [21-22]節省運算量[3], [11]。

如圖六所示，同樣以 III.***A*** 的語音段來作爲實驗的對象；我們使用預估切點左右兩邊都是 125 frame (一秒鐘)的 window 去掃描，並且每次右移(shift)一個 frame 直到右邊 window 到達語音終點爲止。

圖六也同時顯示 ΔBIC 值與 SVM 演算法的 training misclassification rate，我們可以看出在實際的語者切換點附近 ΔBIC 的值並不是最高，甚至沒有大於零，因此這是一個 miss detection 的情形。此外也有許多 ΔBIC 值大於零(false-alarm)的情形發生；儘管使用 local peaks 或是一個低頻濾波器來處理這個曲線[9-10]，仍然會有太多的假切點會被判斷成可能的切點(candidate)。

反觀 SVM 演算法的兩個 training misclassification rate；即公式(14)與(15)中的 $mis^{-}(\hat{R})$ 與 $mis^{+}(\hat{R})$ 都在切點附近明顯的下降並趨近於零，我們可以很容易的找出一個有效的 threshold 來判斷，亦即當 $mis^{-}(\hat{R})$ 與 $mis^{+}(\hat{R})$ 同時都低於此 threshold (如 0.05) 時，將其判斷成可能的切點。上述的判斷機制將比[9]更爲簡單有效，並且沒有任何 false-alarm 發生，就語者切點的易測性而言；SVM 演算法有絕對的優勢。



圖六、以固定 window 大小掃描之結果分佈來評估 SVM 與 BIC 之易測性

D. 低於兩秒以下之語者切換點偵測能力評估

由於 BIC 在資料的蒐集上必須有一定長度，因此在 metric-based 的系統上大部分在 window 的設計上都是以預測切換點左右兩秒鐘爲主，而擴增法的系統則都是以兩秒鐘 window 開始擴增。因此在以往的文獻中，對於低於兩秒以下之語者切換點都無法有效的判斷[1-4], [8-10], [16], [32]。爲了測試 SVM 的 training misclassification rate 在低於兩秒鐘語者切換點判斷上的強健性程度，我們設計了兩個實驗來加以測試。

如圖七所示，我們錄製一個由三位語者發聲的語音，每兩秒鐘即切換到另一個語者，由實驗結果發現 ΔBIC 仍然會有 false-alarm，而 SVM 的 training misclassification rate 依然相當準確，易測性也相當高。

如圖八所示，在第二個實驗中；同樣由三位語者發聲；但是第二位語者只有錄製一秒鐘，由實驗結果發現 BIC 無法判斷出任何的語者切換點(所有 ΔBIC 都小於零)，而 SVM 的 training misclassification rate 依然相當準確；易測性也相當高。

藉由上述的實驗評估；我們發現 SVM 具有以下優勢：

1. 可以用更少的資料收集量來判斷切點；對更短的語者切點做偵測。亦即對於 duration 在兩秒鐘以下的短語音段落[11]，有更好的判斷力與偵測能力，因此 miss detection rate 可以降低[3]。
2. 易測性高，不需要額外使用一個低頻濾波器[3][9-11]。
3. 相較於 ΔBIC 值，對於一個不可避免的 heuristic threshold 而言，SVM 的 training misclassification rate 之分佈很明顯，因此 threshold 大小很容易選擇。
4. False-alarm 機率極低。

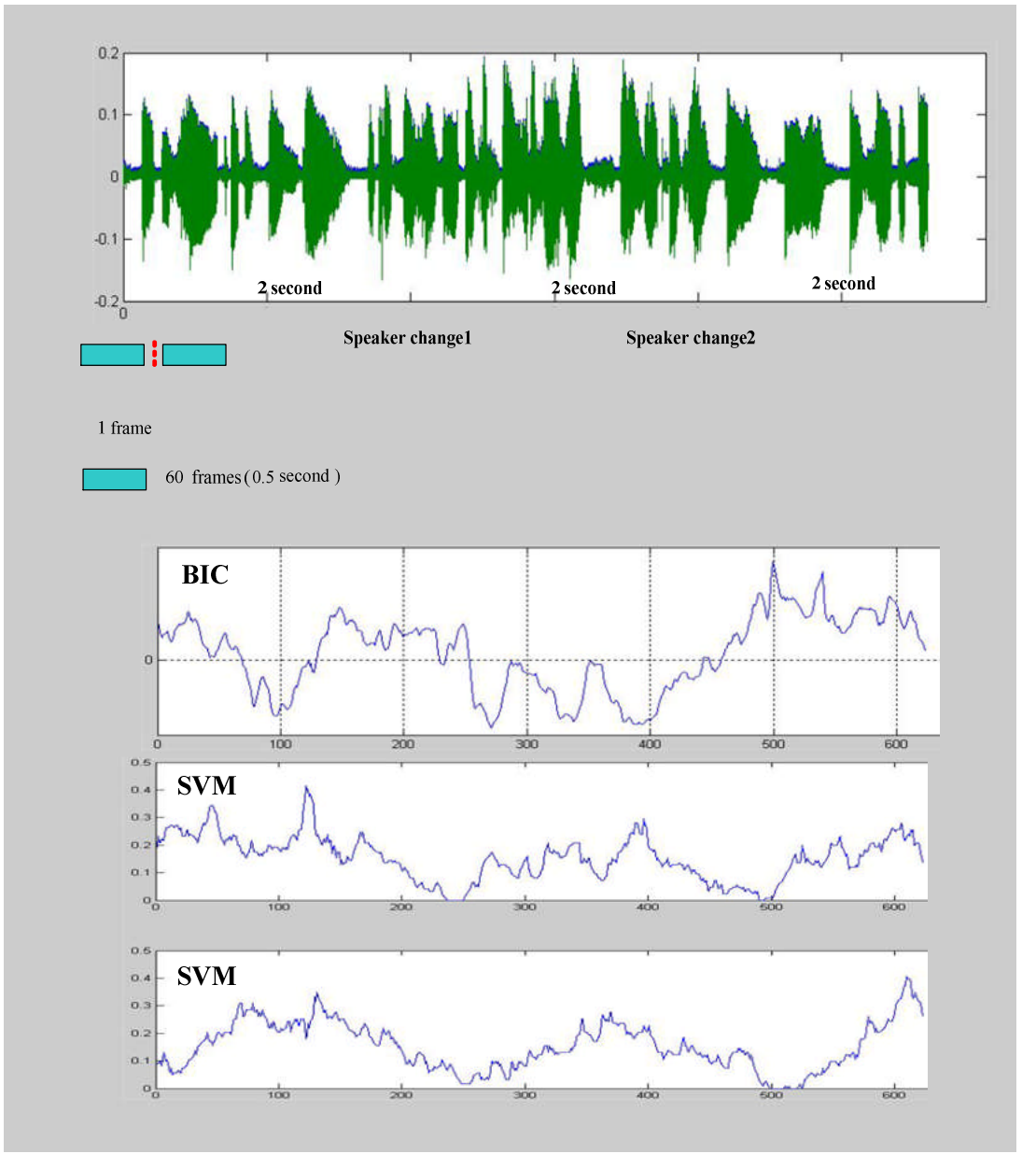
E. 使用相同語者之語音串評估滑動窗與 training misclassification rate 之 threshold 大小

如圖九所示，爲了要找出一個適合於 SVM 的 sliding window 大小與 training misclassification rate 之 threshold，我們使用四個語音串(以不同顏色區分)來做評估，每個語音串都只有一位語者。利用 III.*B* 中；使用切點左右相同尺寸之 window 來掃描。由 SVM 的 training misclassification rate 來觀察； $mis^{-}(\hat{R})$ 與 $mis^{+}(\hat{R})$ 在 0.05 (紫色虛線)以上將會穩定的爬升，意味著當 window 收集資料越多，SVM 對於同一個語者的語音將越無法有效的找出一條 hyperplane 將資料分類，進而導致 training misclassification rate 越來越大。因此我們使用 0.05 來當作 training misclassification rate 之 threshold，當 training misclassification rate 大於 0.05 則表示左右兩邊的 window 是同一個語者的聲音，因此原先假設的切換點無效。相反地，misclassification rate 小於 0.05 表示 SVM 找到一條有效的 hyperplane 將資料分類，因此原先假設的切換點正確。

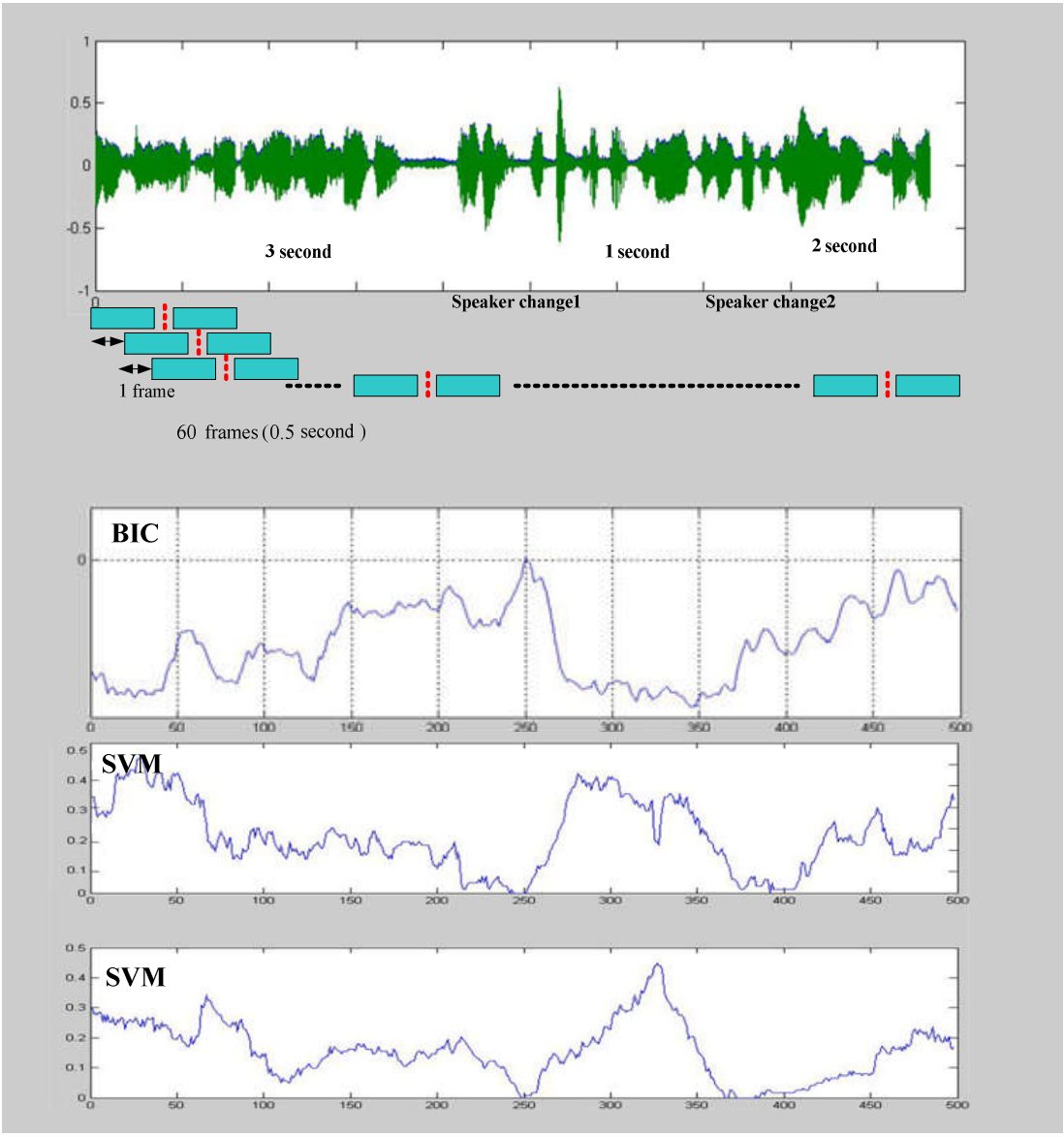
此外，當 sliding window 大於 70 個 frame 時，所有語音串的 training misclassification rate 都會大於 0.05 並且穩定的爬升。我們可以說：當切點兩邊相鄰的 window 收集資料都大於 70 個 frame (左右兩邊共 140 個 frame) 時，SVM 將可以有效地判斷出左右兩邊的 window 是同一個語者的聲音。

IV. 語者切換點偵測驗算法

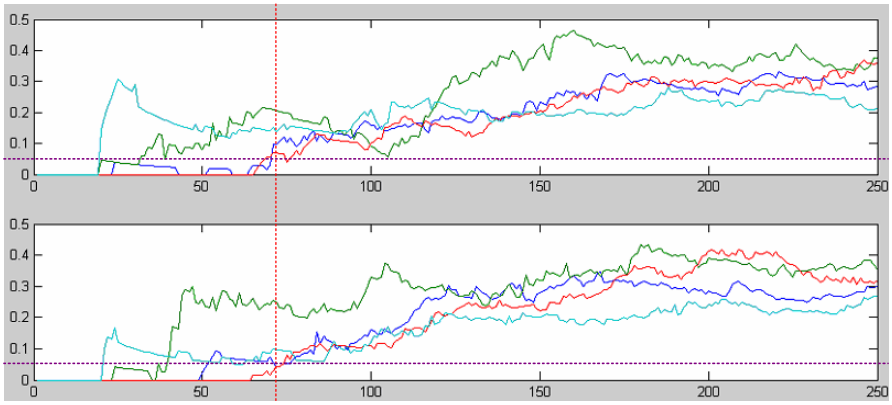
我們所提出的演算法主要分爲三個步驟：首先，第一步驟我們將可能的語音切點(potential speaker change point)偵測出來，這個步驟是整各演算法的關鍵步驟，由於接下來的第二步驟：確認處理(confirmation process)完全仰賴第一步驟的偵測品質[11]。在第三步驟中我們將相鄰且是相同語者的語音段加以合併(merging)。



圖七、低於兩秒以下之語者切換點偵測能力評估一



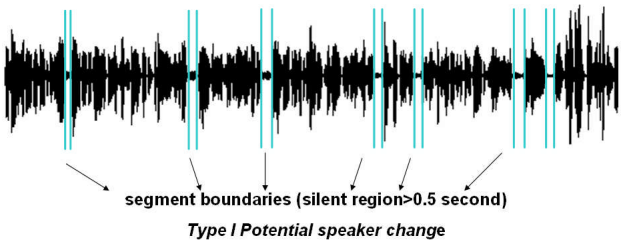
圖八、低於兩秒以下之語者切換點偵測能力評估二



圖九、使用相同語者之語音串評估滑動窗與 training misclassification rate 之 threshold 大小

A. 步驟一之第一類可能語者切換點(位於靜音段邊緣的切換點)之偵測

對於位於靜音段邊緣的切換點，我們使用最直覺的 energy-base 判斷方式來做偵測，而靜音段的 duration 要多久以上才有可能是一個語者切換點? 我們可以使用一個經驗值的 threshold 來判斷。如圖十所示，是一個使用 0.5 秒爲 threshold 的靜音段來判斷可能切換點的切割結果。



圖十、第一類可能語者切換點(位於靜音段邊緣的切換點)之偵測

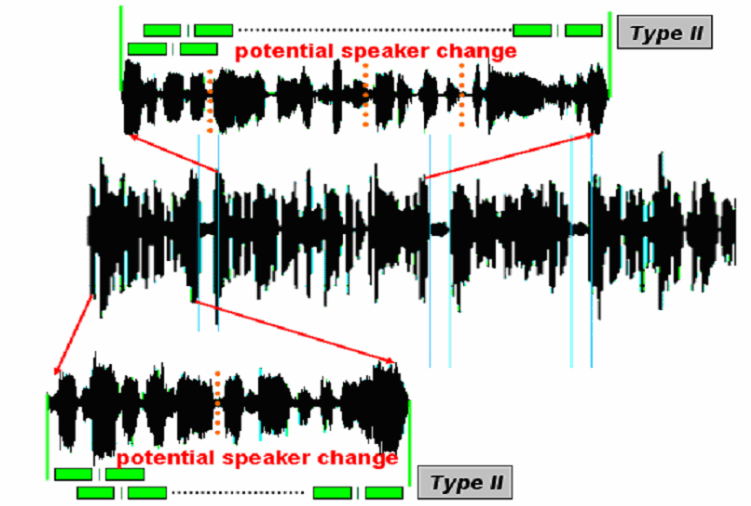
B. 步驟一之第二類可能語者切換點(非靜音語音段的切換點)偵測

如圖十一所示，針對此類非靜音語音段的語者切換點，我們使用 II.***B*** 所提到的應用 SVM 訓練錯誤率判斷之語者切換點偵測演算法；每次右移一個 frame 來做掃描，藉此來判斷出此類可能的切換點。在 III.***E*** 中，我們曾對所提出演算法的 window size 做評估，其結論是 70 個 frame 是一個不錯的資料收集量，但對可能的語者切換點而言；我們使用較小的 60 個 frame 來當作 sliding window 的大小，這樣設定的原因是我們寧願提高 false-alarm 將所有可能的切換點都找出來，藉此降低 miss detection rate。而那些多出來的 false-alarm 可以使用「切換點確認」與「音段合併」等步驟有效的加以排除[3], [12-13]。而 60 個 frame (左右兩邊共 120 個 frame)之 window size 可以更有效的判斷一秒鐘左右(每秒 125 個 frame)的切換點。相較於其他系統[1-4], [8-10], [16], [32]只能判斷兩秒鐘以上的語者切換點而言，有更多的優勢。另一個優點是：對於每次 shift 一個 frame 的 sliding 機制而言，僅用 60 個 frame 的大小可以大幅的降低運算量。

C. 步驟二:切換點確認

從 **IV.*A*** 與 **IV.*B*** 所偵測出來的可能切換點只是初步的推測，在這個步驟中，我們取出那些可能切點左右兩邊各 1.5 秒鐘的 window 重複地使用 **II.*B*** 提出的方法來做判斷，企圖以更多的資料收集量來確認先前在 **IV.*A*** 與 **IV.*B*** 找出的切換點，同時刪除 false-alarm 的切換點，詳細情形如圖十二所示。

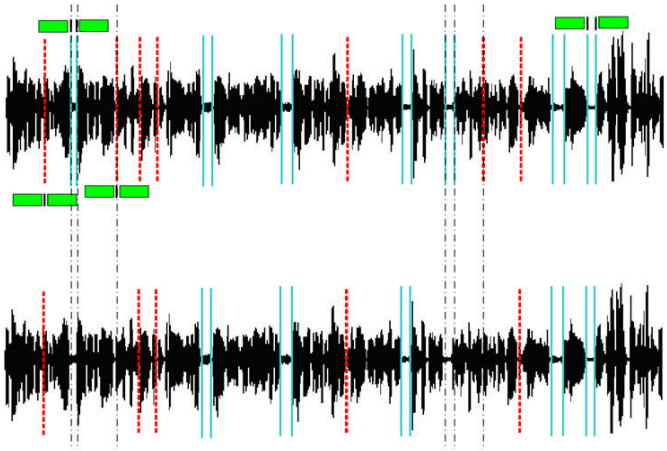
由於靜音段的資料對於判斷切換點沒有幫助甚至會影響 training misclassification rate 的判斷，因此對於第一類靜音的切換點而言(藍色實線)；我們取出切音點左右兩邊非靜音的 1.5 秒來當做 window。對於第二類非靜音語音段的語者切換點(紅色虛線)而言，直接取出可能切點左右兩邊各 1.5 秒鐘的 window 即可。黑色虛線部份即爲 false-alarm 的切換點刪除效果。



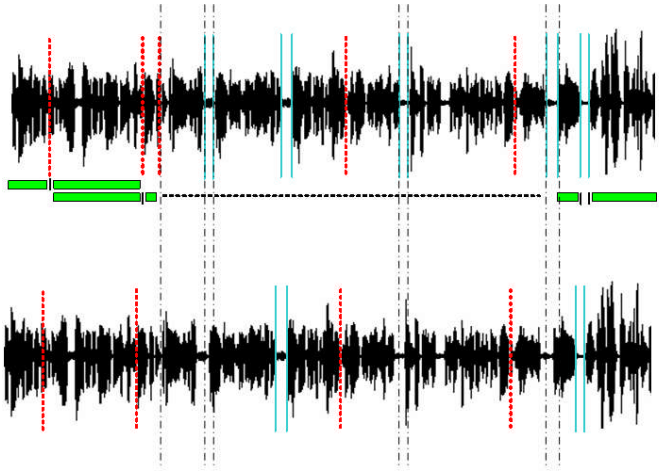
圖十一、第二類可能語者切換點(非靜音語音段的切換點)之偵測

D. 步驟三:相同語者音段合併

最後一個步驟中我們將先前所確認後的語音段落加以合併，採用左右不固定 window 大小的方式；直接將兩個鄰近的語音段落使用 **II.*B*** 提出的方法再次做切換點偵測，藉此將同一個語者的語音段落加以合併，詳細情形如圖十三所示。當語音段太多的時候，這個步驟的動作可以重複兩次或是三次，以求效果。



圖十二、切換點確認



圖十三、相同語者音段合併

V. 實驗結果

A. 實驗環境介紹

我們使用的參數爲 13 階的的 MFCC 參數，分別對兩個語音串做實驗；平均每個語音串有 30 分鐘的語音；並且含有六位語者交互說話。

針對語者切換點偵測系統而言；一般有兩種錯誤，對於系統無法找到實際換點位置的錯誤稱爲 detection error，另一種錯誤指的是系統找到的切換點位置不是或是無法對應到實際的切換點位置，稱之爲 false-alarm 或是 segment insertion。我們也可以用資訊檢索的 precision 與 recall 來表示上述的兩個錯誤，其定義如下:

$$\Pr ecision = \frac{number\ of\ correctly\ found\ boundaries}{total\ numer\ of\ boundaries} .....(16)$$

$$\mathrm{Re} call = \frac{number\ of\ correctly\ found\ boundaries}{numer\ of\ hypothesized\ boundaries} .....(17)$$

另外我們也使用 F-measure [13]來對 precision 及 recall 進行綜合評估，針對對等參數(neutral parameterization)而言；我們給予 precision 與 recall 相同的權重，因此 F-score 的定義如下:

$$Fscore = \frac{2 \times \Pr ecision \times \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call} .....(18)$$

我們定義一個放寬時間(tolerance) Δt 來做判斷，對於一個真實切換點位置在 t_0 的切換點而言，如果偵測出來的切換點發生在 $t_0 - \Delta t < t_0 < t_0 + \Delta t$ 之間；都視爲正確的偵測，在我們的實驗中我們設定 $\Delta t = 0.5$ 。此外；這樣的判斷也突顯演算法的強健性，對於系統偵測出來的切換點；我們僅容忍偏移量在左右 0.5 秒以內，才算是正確的偵測，並非其他系統所採用的：左右偏移 1 秒鐘[3]都算偵測正確的評估方式。

B. 相關參數設定實驗

對於偵測第二類語者切換點所需要的相關參數設定；在 **IV.*B*** 我們已經決定了 sliding window 大小爲 60 個 frame，以及在 **III.*E*** 中，我們決定了 SVM 的 training misclassification rate 爲 0.05。

而針對切換點確認(**c**onfirmation)與合併(merging)；我們提出四種不同的組合來實驗，分別是 CS_MS、CS_MB、CB_MS 與 CB_MB，藉此觀察 SVM 與 BIC 的效能。其縮寫定義如下:

- **CS**: **C**onfirmation by **S**VM.
- **CB**: **C**onfirmation by **B**IC.
- **MS**: **M**erging by **S**VM.
- **MB**: **M**erging by **B**IC.

針對以上 BIC 的懲罰係數(penalty factor)與 SVM 的 training misclassification rat 臨界值之設定；我們採用 10 個 10 秒鐘的語音串來做實驗；每個語音串都只有一個語者切換點。從這 10 個小測試中求得一個最適當的參數設定。最後決定的參數設定如表一所示：

表一、BIC 的懲罰係數(penalty factor)與 SVM 的 training misclassification rat 臨界值之設定

Algorithm	SVM based	BIC based
Parameter	Threshold of training misclassification rate	Penalty factor
Confirmation	0.1	1.4
Merging	0.05	1.4

在 SVM 的 **c**onfirmation 過程中我們採用較高的 training misclassification rate 臨界值來判斷；藉此放寬 threshold。最後的 merging 步驟再使用較低的 training misclassification rate 臨界值來判

斷，這樣的設定主要是希望獲的更高的 recall rate，因爲對於一個切換點偵測系統而言；確實偵測到語者切換點是最主要的目的，因此我們對 recall rate 的要求更勝於 precision rate。

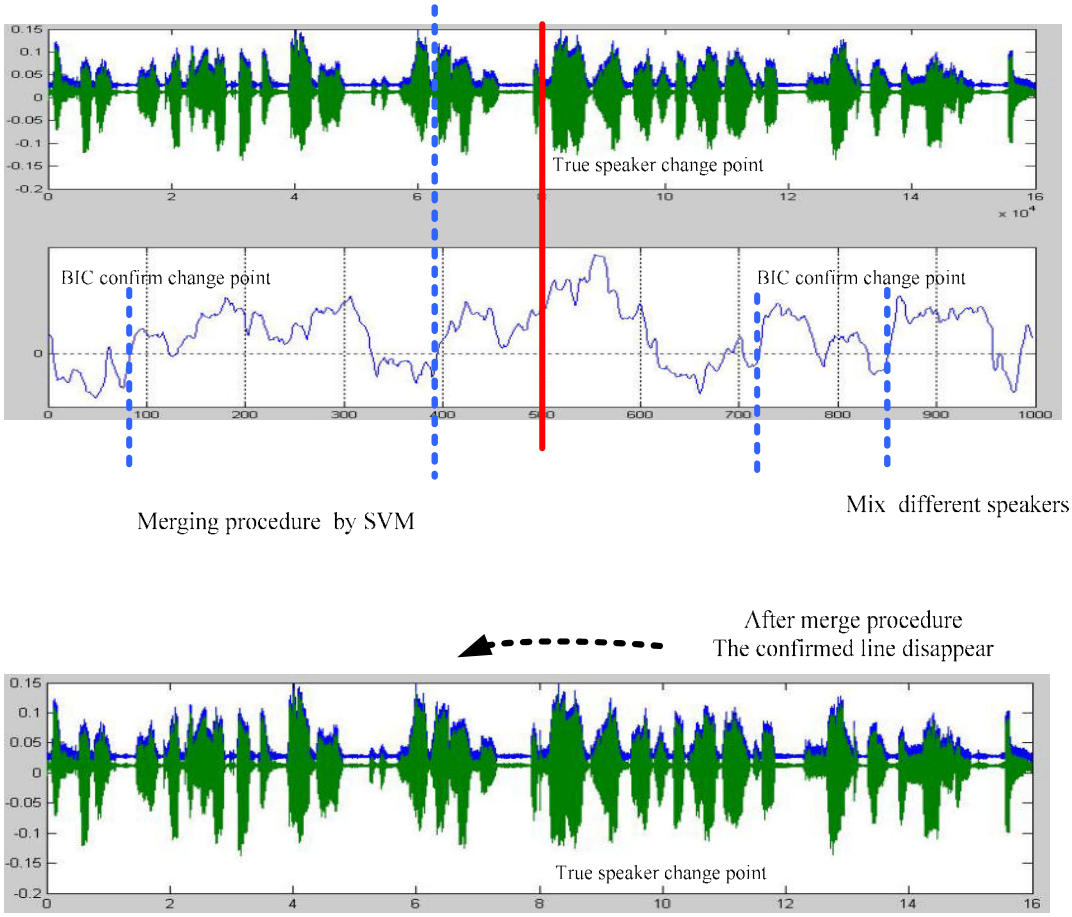
C. 整體實驗結果比較與討論

整個實驗的結果整理如表二所示，其中 CS_MS (Confirmation by SVM & Merging by SVM) 有最好的偵測效果，我們推測是因爲在 **IV.*B*** 之中；針對第二類可能的切換點使用較小的 window 去掃描的原因；再加上 training misclassification rate 的強健性，所有可能的切換點都可以被找到的原因。

表二、實驗結果整理

Data		Precision	Recall	F-score
Chinese	CS_MS	0.76	0.90	0.82
	CS_MB	0.79	0.81	0.79
	CB_MB	0.57	0.77	0.65
	CB_MS	0.53	0.75	0.62
English	CS_MS	0.70	0.92	0.79
	CS_MB	0.72	0.80	0.75
	CB_MB	0.44	0.85	0.57
	CB_MS	0.57	0.62	0.59

在 CB_MS(Confirmation by BIC & Merging by SVM)實驗中；BIC 的確認點與實際的語者切換點有偏移的現象，如圖十四所示：實際的切換點爲紅色實線，而 BIC 的 confirmation 結果爲藍色虛線；這樣子的確認結果在 SVM 的 merging 之步驟中會將紅色橢圓虛線視爲同一語者的聲音，但其實已經混雜兩個語者的語音資料。這樣的結果會導致 training misclassification rate 上升，進而造成 SVM 的誤判；recall rate 因此最差。



圖十四、CB_MS(Confirmation by BIC & Merging by SVM)實驗分析

VI. 結論

在這一篇論文之中我們提出了以SVM可分離性爲基礎(separability-based)之語者切換偵測演算法，這種新穎的語者切換點判斷機制，不同於以往採用的演算法，而是使用 SVM 的 training misclassification rate 與 metric-based 的掃描方式加以結合。在偵測能力的評估上，我們亦做了許多實驗來證實 SVM 演算法的強健性。此外，本方法也同樣擁有像 BIC 不需要事先訓練語者資料或是建立任何模型的優勢。由實驗發現 CS_MS (Confirmation by SVM & Merging by SVM)有最好的偵測效果，證明 SVM separability-based 法應用在語者切換點偵測上之優越性。

參考文獻

[1]. S. Chen and P. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian information criterion, “in Proc. DARPA Broadcast News Transcription Understanding Workshop, Feb. 1998, pp. 127-132.

[2]. M. Cettolo and M. Vescovi, “Efficient audio segmentation algorithms based on the BIC,” Proceedings of ICASSP2003.A Sequential Metric-based Audio Segmentation

[3]. Shih-Sian Cheng and Hsin-min Wang "METRIC-SEQDAC: A Hybrid Approach for Audio Segmentation," Proc. International Conference on Spoken Language Processing (ICSLP2004), Jeju Island, Korea 2004.

[4]. S. S. Cheng and H. M. Wang, “A sequential metric-based audio segmentation method via the

Bayesian Information Criterion," Proceedings of Eurospeech 2003.

[5]. Jhing-Fa Wang Taiwan - Jia-Ching Wang, Tze-Hsuan Huang and Cheng-Shu Hsu, "Home Environmental Sound Recognition Based on MPEG-7 Features," 46th IEEE International Midwest symposium on Circuits and systems, 2003.

[6]. B. W. Zhou, and John H. L. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian Information Criterion," Proceedings of ICSLP 2000.

[7]. A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian Information Criterion," Proceedings of Eurospeech 1999.

[8]. M. Siegler, U. Jain, B.Raj, and R. Stern, "Automatic segmentation, classification and clustering of broadcast news audio, " in Proc. DARPA Speech Recognition Workshop, Feb, 1997,pp. 97-99

[9]. J.W. Hung, H.M. Wang, and L.S. Lee. Automatic Metricbased speech segmentation for broadcast news via principal component analysis. Proceeding of ICSLP'2000.

[10]. P. Delacourt, C. J. Welkens, DISTBIC: A Speaker-based segmentation for Audio Data Indexing, Speech Communication, v. 32, pp 111-126, 2000.

[11]. Bowen Zhou and John H. L. Hansen, "Efficient Audio Stream Segmentation via the Combined T2 Statistic and Bayesian Information Criterion", IEEE Transactions On Speech And Audio Processing, Vol.13, No.4, July 2005.

[12]. Meinedo, H., Neto, J.A., "Audio Segmentation, Classification and Clustering in a Broadcast News Task", Proc. ICASSP'2003 - Hong Kong, China, Apr. 2003.

[13]. X. Zhong, M. Clements, and S. Lim, "Acoustic change detection and segment clustering of two-way telephone conversation," Proceedings of Eurospeech2003.

[14]. G. Schwarz, "Estimating the dimension of a model," The Annals of Statistics, vol. 6, no. 2, pp.461–464, 1978.

[15]. Cheng-Shu Hsu, "Home Environmental Audio Classifier Based on SVM and MPEG-7 Audio Low-level Descriptors", Master Thesis, NCKU, Taiwan, 2002.

[16]. H. Beigi and S. Maes, ``Speaker, channel and environment change detection", Proceedings of the World Congress on Automation, 1998.

[17]. Luis Perez-Freire and Carmen García-Mateo, "A Multimedia Approach For Audio Segmentation In Tv Broadcast News", ICASSP 2004.

[18]. F. Kubala et al., ``The 1996 BBN Byblos Hub-4 transcription system", Proceedings of the Speech Recognition Workshop, pp 90-93, 1997.

[19]. S. Chen et al., ``IBM's LVCSR System for Transcription of Broadcast News Used in the 1997 Hub4 English Evaluation", Proceedings of the Speech Recognition Workshop, 1998.

[20]. M. Harris, X. Aubert, R. Haeb-Umbach, and P. Beyerlein, "A study of broadcast news audio stream segmentation and segment clustering,"in Proc. EUROSPEECH, Budapest, Hungary, 1999, vol. 3, pp. 1027–1030.

[21]. P. Sivakumaran, J. Fortuna, and A. M. Ariyaeeinia, "On the use of the Bayesian Information Criterion in multiple speaker detection," in Proc. EUROSPEECH, Aalborg, Denmark, 2001, vol.

2, pp. 795–798.

[22]. M. Cettolo, “Segmentation, classification and clustering of an Italian broadcast news corpus,” in Proc. of the 6th RIAO-Content-Based Multimedia Information Access - conference, Paris, France, 2000.

[23]. A. Raftery, "Bayesian Model Selection in Social Research", Tech. Reports, Dept. of Stat., Univ. of Washington, 1994.

[24]. R. Bakis et al., "Transcription of broadcast news shows with the IBM large vocabulary speech recognition system ", Proceedings of the Speech Recognition Workshop, pp 67-72, 1997.

[25]. C. Cortes and V. Vapnik, “Support vector networks,”Machine Learning, vol. 20, pp. 273-297, 1995.

[26]. V. Vapnik, The Nature of Statistical Learning Theory. New York: Springer, 1995.

[27]. Laurent Couvreur and Jean-Marc Boite, "Speaker Tracking in Broadcast Audio Material in the Framework of the THISL Project", Proc. of European Speech Communication Association (ESCA) European Tutorial and Research Workshop (ETRW) on Accessing Information in Spoken Audio, 1999.

[28]. M. Basseville, "Distance Measures for Signal Processing and Pattern Recognition", Signal Processing, Vol. 18, 1989.

[29]. P. Woodland and al., “The development of the 1996 HTK broadcast news transcription system,” in DARPA speech recognition workshop, 1997.

[30]. J. Johnson and P. Woodland, “Speaker clustering using direct maximisation of the MLLR-adapted likelihood,” in ICSLP98, 1998.

[31]. H. Gish and N. Schmidt, “Text-independent speaker identification,” IEEE signal processing magazine, oct. 1994.

[32]. P. Delacourt and C. J. Wellekens, “Audio data indexing: use of second-order statistics for speaker-based segmentation,” in ICMCS, 1999.

[33]. S. Chen, E. Eide, M. Gales, R. Gopinath, D. Kanevsky, and P. Olsen, "Recent improvements to IBM’s speech recognition system for automatic transcription of broadcast news,” in Proc. DARPA Broadcast News Transcription Workshop, 1999.

[34]. J. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system,” Speech Commun., vol. 37, no. 1–2, pp. 89–108, 2002. [33] Recent improvements to IBM’s speech recognition system for automatic transcription of broadcast news

[35]. T. Hain, S. E. Johnson, A. Tuerk, P. C.Woodland, and S. J. Young, “Segment generation and clustering in the HTK broadcast news transcription system,” in Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, VA, 1998, pp. 133–137.

[36]. S. Wegmann, P. Zhan, and L. Gillick, “Progress in broadcast news transcription at Dragon systems,” in Proc. IEEE ICASSP-99: Inter. Conf. Acoust., Speech, Signal Process., May 1999, 1912.

[37]. P. Zhan, S.Wegmann, and L. Gillick, “Dragon systems’ 1998 broadcast news transcription

system for Mandarin," in Proc. DARPA Broadcast News Transcription Workshop, 1998.

[38]. V. Vapnik, Statistical Learning Theory. New York: Wiley, 1998

[39]. B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A. Smola, "Input space vs. feature space in kernel-based methods," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1000-1017, 1999

[40]. V. Vapnik, Estimation of Dependences Based on Empirical Data. Springer-Verlag, 1982

[41]. C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121-167, 1998.

[42]. Smola and B. Schölkopf, "A tutorial on support vector regression," Tech. Rep. NC2-TR-1998-030, Neural and Computational Learning II, 1998

[43]. J. C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector learning machines," in Advances in Neural Information Processing Systems 9 (M. Mozer, M. Jordan, and T. Petsche, eds.), pp. 375-381, Cambridge, MA: MIT Press, 1997.

[44]. G. Fung, O. L. Mangasarian, and J. Shavlik, "Knowledge-based support vector machine classifiers," in Advances in Neural Information Processing, 2002.

[45]. T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in Proceedings of ECML-98, 10th European Conference on Machine Learning (C. Nédellec and C. Rouveirol, eds.), (Chemnitz, DE), pp. 137-142, Springer Verlag, Heidelberg, DE, 1998.

[46]. K. Crammer and Y. Singer, "On the learnability and design of output codes for multiclass problems," in Computational Learning Theory, pp. 35-46, 2000

[47]. K.-R. Müller, A. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, and V. Vapnik, "Predicting time series with support vector machines," in Articial Neural Networks - ICANN'97 (W. Gerstner, A. Germond, M. Hasler, and J.-D. Nicoud, eds.), pp. 999-1004, 1997.

[48]. S. Mukherjee, E. Osuna, and F. Girosi, "Nonlinear prediction of chaotic time series using support vector machines," in 1997 IEEE Workshop on Neural Networks for Signal Processing, pp. 511-519, 1997.

[49]. F. E. H. Tay and L. Cao, "Application of support vector machines in financial time series forecasting," Omega, vol. 29, pp. 309-317, 2001.

[50]. L. J. Cao, K. S. Chua, and L. K. Guan, "c-ascending support vector machines for financial time series forecasting," in 2003 International Conference on Computational Intelligence for Financial Engineering (CIFEr2003), (Hong Kong), pp. 317-323, 2003.

[51]. H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," in Advances in Neural Information Processing Systems, vol. 9, p. 155, The MIT Press, 1997.