

使用簡易音高週期浮現演算法(EAPA)及多層認知元網路(MLP)之台語聲調辨識

林威伯¹、陳志宇¹、傅振宏¹、江永進²、呂仁園¹

1 長庚大學電機工程研究所

2 清華大學統計研究所

Email: rylyu@mail.cgu.edu.tw

Tel: 886-3-3283016#5677 Fax: 886-3-3288026

摘要















SIFT 是一個廣泛地被運用在尋找語音信號音高高週期的方法，但是 SIFT 卻有兩個缺點：執行速度緩慢及必須利用一個不可靠的試誤方法來進行平滑化。我們提出一個只需利用低通濾波器的新方法---EAPA。特別是在後處理：有聲/無聲判定及平滑化問題上，EAPA 特別具有優勢。在辨認方法上，我們建議使用利用倒傳遞(Back Propagation)訓練的多層次認知元(MLP)來克服這個困難。利用長度正規化幫助吸收音高週期的變異(Variation)以減低錯誤率；利用音高正規化可以讓我們實現多語者辨認聲調。最後，我們利用以上製作了一個可以在 win95/98 平台上執行的多語者台語聲調系統。

1. 簡介

在台灣，台語（或稱閩南語、河洛話）是超過 75% 的人所使用的母語，而對台語的語音處理研究是這幾年才漸漸受到重視。在推動本土語文教學的政策下，讓電腦「講台語嘛會通」，是一件很有意義的事。在一些台語音的先驅研究中，聲調通常暫不被考慮。在多音節詞的辨認裏，不考慮聲調時，同音詞的數目不多，故問題不大。但在碰到單字詞時，若再不考慮聲調，則同音詞的問題就變得十分嚴重。所以我們期望增加提供聲調的資訊來解決同音詞的問題。

如同華語，台語是單音節性具有聲調的語言。傳統上，每一個台語的單音節可以分成：聲母（initial），韻母（final）和聲調（tone）三個部分。台語一共有 18 個聲母（包含一個空聲母）94 個韻母（含入聲韻母），以及 7 個聲調。聲母也就是音節開頭的子音。韻母就比較複雜，它由 1 到 3 個母音組成。特別的是台語有由/p/、/t/、/k/、/h/作結尾的所謂「入聲韻母」。聲調則是指語音基本頻率(Fundamental Frequency, F0)的高昇低降。它和聲母、韻母一樣具有辨義的作用。傳統語音學把台語聲調分成一共有七種。它們分別被定義為：(1)高平調，如「東」；(2)中平調，如「洞」；(3)低降調，如「棟」；(4)高降調，如「黨」；(5)低升調，如「同」；(6)高束調，如「獨」；(7)低束調，如「督」；。表一可以清楚地看到每個聲調的基本頻率趨勢。其中，含前五種聲調（聲調一—聲調五）的音節其實與含後兩種的聲調（聲調六、七）的音節之音素組合不同。故在語音辨認研究裏，前五種聲調可與後兩種聲調分開考慮。本論文即採這種觀點。

<表一>台語七種聲調之調名、調號、波形、基頻及其通用拼音[8]表示法

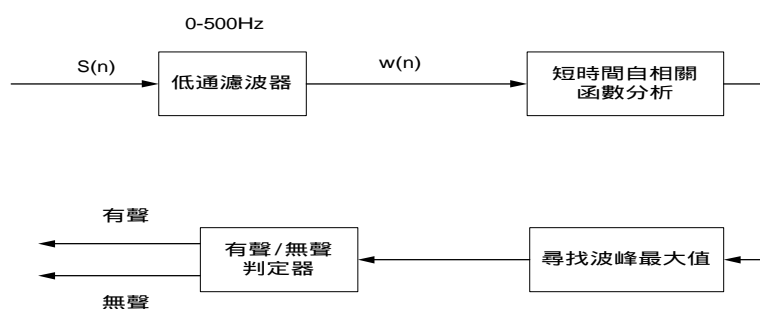
| 漢字 | 東 | 洞 | 棟 | 黨 | 同 | 獨 | 督 |
|--------|---|---|---|---|--|---|---|
| 調名 | 高平調 | 中平調 | 低降調 | 高降調 | 低升調 | 高束調 | 中束調 |
| 調號 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| 波形 |  |  |  |  |  |  |  |
| 基本頻率趨勢 |  |  |  |  |  |  |  |
| 通用拼音 | dong1 | dong2 | dong3 | dong4 | dong5 | dok6 | dok7 |

2. 聲調特徵擷取

聲調特徵擷取----也就是尋找所謂的基頻走勢(Fundamental Frequency Contour)。目前最有名的聲調特徵擷取方法是 SIFT (Simple Inverse Filtering Tracking)演算法[1]。是由 Markel 先生在 1973 年所提出的。它是結合線性預測碼與自相關函數來尋找基本頻率趨勢。一般認為 SIFT 演算法是偵測音高週期的利器。然而，SIFT 仍有其缺點：在對較高音高週期語者（如兒童）的處理，就沒有一般來得好。此外，因為在判定有聲或無聲時，所使用的門檻值(threshold)也許會將有聲區的音高週期值誤判為無效而捨棄。因此，我們還需要一個平滑化(smoothing)的動作，將被誤砍的音高週期值還原回來。

針對 SIFT 的麻煩的後處理過程，我們提出簡易音高週期浮現演算法(Easy Appearing Pitch Algorithm, EAPA)來改善。EAPA 是將 SIFT 加以簡化。但在音高週期偵測上，卻可以提供更好的特性。圖一是 EAPA 的方塊圖。

<圖一>EAPA 方塊圖



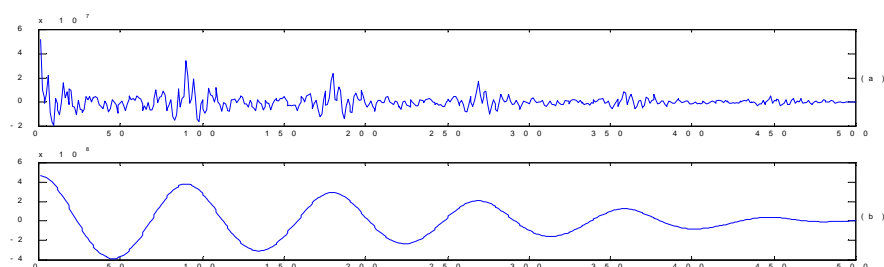
一般來說，男性語者的基本頻率約在 80-200Hz 之間，而女性語者約在 150-350Hz 間。這些基本頻率資料除了當尋找波峰最大值的邊界以避免抓錯音高週期和節省搜尋

時間外，還可以對音高週期偵測做出進一步的貢獻。我們利用一個截止頻率為 500Hz 的低通濾波器將輸入語音信號 $s(n)$ 濾出後所得的低通信號 $w(n)$ 。我們可以發現， $w(n)$ 也很像正弦波，而且這些似正弦波與音高週期同相。跟 SIFT 利用線性預測碼的殘餘信號(residual signal) 比較，可以明顯看出，這些低通信號比殘餘信號能更清楚地描述音高週期特性：不管原始語音信號多麼複雜，EAPA 都將一個音高週期濾成只存在一個波峰似正弦波；而 SIFT 利用線性預測碼所轉出的殘餘信號，還會可能有 1-3 個次波峰。接下來的作法如同 SIFT，先利用自相關函數定位，尋找波峰最大值，最後通過有聲/無聲判定器。但在有聲/無聲判定器中的作法與 SIFT 有差異。EAPA 的作法是很自然的，若定義自相關函數如下：

$$f(k) = \sum_n s[n]s[n+k], \text{其中 } s[n] \text{ 為語音訊號}$$

則對類週期(pseudo-periodic)訊號而言，存在一個 p 使得 $f(k) \approx f(k+p)$ ， p 就是所謂的音高週期，而且 p 應該界在我們所設定的邊界範圍內。假如語音信號不具週期性的話，則 p 會趨近無限大。因為我們在邊界範圍內尋找波峰最大值，所以 p 會趨近於我們所設定的上邊界。事實上，一般男性語者的基本頻率值鮮少有超過 200Hz。所以，如果 p 值的倒數也就是所謂的基本頻率趨近 250Hz 的話，就暗示著這段語音信號可能是無週期性的。因此，在這裡得到的基本頻率值就是無效的，應該要捨棄。對 EAPA 而言，有聲/無聲的判定是很簡單而自然的，完全不需要靠外在資訊如能量等的幫助。在實作中，還有一點需要注意的，就是要捨棄掉幾個靠近 250Hz 的值。在我們的系統設定中，我們在短時間自相關函數所採用的音框是 40ms，每次移動 8ms，也就是移動 0.2 個音框。在不規則信號尚未完全離開音框的情況下，所得到的基本頻率值也不是很可靠的，應該捨棄。在上述的實驗環境下，我們建議大概需要捨棄 2-3 個靠近 250Hz 基本頻率值。EAPA 在有聲/無聲判定中不使用能量門檻值，就順便大大地簡化平滑化問題，因為不需要考慮還原誤砍值的關係，所以只需要考慮誤抓倍基本頻率或半基本頻率的問題。事實上，連倍半基本頻率這個問題也都是很罕見的，原因是經由 EAPA 所取得的信號在一個音高週期內只會有一個波峰，週期性相當明顯，再經自相關函數的定位後，倍頻或半頻被抓到的機會是很小的。另外，還有一個很重要的優點，就是這個方法具有強健性(Robustness)。不管是其他語音或是其他語者，沒有週期性的語音趨近上邊界是很自然的事，不必像能量門檻值一樣需要依照情形做刻意的調整。所以說，EAPA 在偵測音高週期的後處理是非常強大的。

<圖三>SIFT 與 EAPA 之自相關函數比較



以下，我們將分別就波形特徵 有聲/無聲判定及平滑化問題等三方面與 SIFT 比較。

- (1) 波形特徵：從圖三可以清楚地看到，上圖是一般經 SIFT 短時間自相關分析的結果，我們發現有許多不必要的雜訊，而且這些雜訊如果大一點，就有可能會造成誤抓倍半音高週期或是最大值前後的誤差。所以這些雜訊就是造成錯誤的主因。下圖則是由 EAPA 經過自相關函數所成的信號，比起 SIFT，「乾淨」了許多，相對的就減低很多錯誤發生的可能，對找尋波峰最大值有很正面的幫助。
- (2) 有聲/無聲判定：對 SIFT 來說，選擇一個適當能量門檻值的一件很重要的事，如果選取得太高會造成取得的基本頻率趨勢變得『坑坑洞洞』的鋸齒狀而使得後面的平滑化問題更加棘手；以及失去部分的基本頻率值，造成一些較短的語音信號特別是像是入聲調的音，無法擷取到足夠的特徵。如果能量門檻值選取的過低，就會失去處理的意義。而偏偏語音信號的能量是會隨語者或是語者心情而改變。也就是說，對一段語音要取得一個無過與不及的能量門檻值並不困難；但對一群語音要找到一個適當的門檻值，卻是十分困難的事。這就暗示著，將所取得的基本頻率趨勢做平滑化處理是無可避免的。甚至我們可以這麼說，能量門檻值問題將平滑化問題惡化了。所以我們認為使用能量門檻值不是一個好方法。在 EAPA 要判定有聲/無聲是一件十分簡單而自然的事，我們已經在前文做了詳細的描述。因為不需要利用外加的能量門檻做判定，使得 EAPA 對有聲/無聲的判定更具強健性。除此之外，也因為不會對正確的基本頻率造成『坑坑洞洞』的破壞而使得平滑化問題變得簡單易解。EAPA 在有聲/無聲判定的改進，是極具意義的。
- (3) 平滑化問題：在 SIFT 演算中，使用能量門檻值的第二個缺點是使得平滑化問題變得更加棘手，而平滑基本頻率趨勢一直扮演了一個重要的角色。首先在一段基本頻率趨勢中，或多或少會遇到遭到誤砍的基本頻率，要還原這些值並不如想像中的那麼容易，因為在起始的地方，很難決定哪一個才是正確的基本頻率值，尤其是能量門檻值取得過高造成大量的基本頻率被誤砍或是與倍半基本頻率一起出現時。根據 Sunberg 先生的說法，基本頻率的最高變化率，不應該超過 $1\%/ms$ [6]。傳統的作法是當變化率超出 $1\%/ms$ 時，這個地方很可能就發生了倍半基本頻率或是誤砍的現象，必須加以平滑化以還原成原來正確的基本頻率趨勢。使用這個方法是存在有一點風險的，就是在某些場合，雖然很罕見但確實存在有變化率超過 $1\%/ms$ 的地方，而造成錯誤的情形產生。此外，還有前向追蹤法(forward track)和後向追蹤法(back track)進行平滑化的方法，這是一種考量前後音框來決定本身音框要在哪一個範圍內找波峰最大值，使得全體誤差量最小的方法。效果比較好，但是如何保證找到一個好的初始值，使得這方法有一點複雜。EAPA 則是大大地簡化平滑化問題。由於不會出現誤砍的情況，所以只需考慮倍半基本頻率出現的問題就足夠了。事實上，連倍半基本頻率的出現是很罕見的。也就是說，我們只要將捨棄靠近上邊界和其鄰近的頻率值，我們就可以得到很漂亮的基本頻率趨勢，而幾乎不必利用到平滑化。所以對 EAPA 而言，平滑化是簡單而花費有限的。這是另一個用 EAPA 取代 SIFT 的重要理由。

基於以上理由，我們採取 EAPA 進行聲調特徵擷取，並於後文中進行實驗驗證。

3. 單音節聲調辨認與實驗結果

我們分別採取隱藏式馬可夫模型(Hidden Markov Model, HMM)[3] 多層次認知元(Multi-layer perceptron, MLP)[4]兩種方法進行台語聲調辨認並驗證。在隱藏式馬可夫模型的實驗中，我們採用七個使用左到右模型來表示台語七個聲調，每個模型三個狀態，每個狀態兩個混和高斯機率分佈，每個單音節除標頭外，共輸入四個參數，依次為：基本頻率軌跡(F0 Contour)，基本頻率軌跡差分(Delta F0 Contour)，能量(Energy)以及能量差分(Delta Energy)。在多層次認知元實驗中，我們採用三層每層五個認知元的網路來進行辨認，每個單音節只用一種參數：固定長度的基本頻率軌跡。實驗結果如下：

<表二>辨認台語單音節聲調結果

| 特徵擷取方法 | 辨識方法 | 辨識率(前一名/前二名辨認率) |
|--------------|------|-----------------|
| SIFT 未經平滑化處理 | HMM | 22.85% |
| SIFT 經平滑化處理 | HMM | 88.62% |
| EAPA 未經平滑化處理 | HMM | 92.92% |
| EAPA 經平滑化處理 | HMM | 93.54% |
| SIFT 經平滑化處理 | MLP | 97.54% / 100% |
| EAPA 經平滑化處理 | MLP | 99.72 % / 100% |

由<表二>我們可以看到平滑化處理對 SIFT 的重要性與必要性。使用 SIFT 沒有經過平滑化處理所擷取聲調特徵的辨識率為 22.85%，幾乎已經跟隨機亂猜測沒什麼差別。這很有可能是我們能量門檻值取得過高，造成大量的基本頻率被誤砍的關係。經過平滑化之後，辨識率上升到 88.62%，但還不能算理想。EAPA 加上 HMM 的辨識率在未經平滑化處理以及經平滑化處理的辨認率分別是 92.92%及 93.54%。不論何者，都比 SIFT 加上 HMM 的辨識率好。可見 EAPA 的確是比 SIFT 更佳的聲調特徵擷取方法。除此之外值得注意的是，EAPA 在經過平滑化處理與未經平滑化處理的辨識率差別不大，可以看出，平滑化的效果就不是那麼明顯。也就是說，對 EAPA 而言，其抽出的聲調特徵，需要平滑化的程度較輕。這支持了我們在前文的說法。

利用 MLP 時，不管是 SIFT 還是 EAPA，都比在 HMM 時表現得好。特別是 SIFT，提升了約 9%的辨識率，所以我們認為，MLP 比 HMM 更適合辨認聲調，在所有 SIFT 與 EAPA 比較的實驗項，EAPA 也比 SIFT 適合辨認聲調。所以，以下的實驗。都將以 EAPA 加上 MLP 來進行辨認。

4. 連續語音聲調辨認與實驗結果

我們把連續語音的聲調辨認建立在單音節模型上。所以，我們首先要做的就是

單音節從連續語音中切出來。我們採取利用基於 HMM 的連續音節辨認的方法找單音節的邊界，再進行聲調辨識。目前本實驗室已建立有大詞彙台語辨識系統[]，我們將利用此系統，協助進行台語連續語音聲調辨識。

<表三>台語連續語音聲調辨識結果

| 連續語音形式 | (前一/前二)辨認率 |
|--------|---------------|
| 詞 | 91.88%/98.68% |
| 句 | 87.56%/97.09 |

表三所列之辨認率比表二所列稍低，檢討原因有：(1) 連續語音的聲調會受到鄰近語音聲調的影響；(2) 台語本身的特性，即部分聲調三在連續語音中會有跟聲調一二類似的「平調」基頻頻率趨勢；(3) 基本頻率軌跡在長時間下，本身的偏移現象。

基頻走勢是時變的。同樣的一個音，由同一個語者在不同的時間來唸，所得到的基頻走勢幾乎是不一樣的。造成這樣的原因很多，如語者的生理狀態----想睡覺、剛睡醒，或是心情的改變----生氣、興奮等等。在多語者聲調辨識時這個問題就更加嚴重，比如說語者的性別、年齡、個人的語言習慣 等等因素。所謂的個人語言習慣就是指基頻走勢本身的高度或是一些像有的人的「平調」的基頻走勢的尾部會稍稍往上揚等等的問題。

由於聲調相互的關係的相對而非絕對。我們嘗試地找出這個語者的發聲時的最高和最低頻率，再將所有的音對應到 0 到 1 之間，這樣我們就可以將所有的音放在同一個水平上進行比較。這個實驗所使用的認知元網路，是純粹由男性語者 A 的語料所訓練的模型。

<表四>多語者台語聲調辨識結果

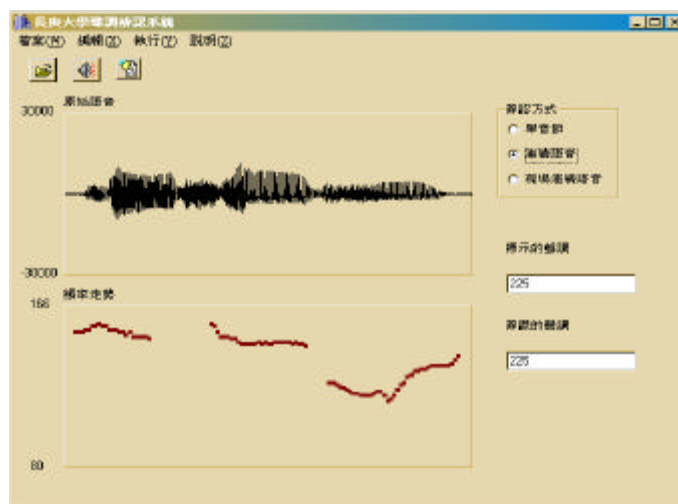
| 多語者形式 | (前一名/前二名)辨認率 |
|-------------|---------------|
| 男性語者 B 單音節 | 92.23%/96.24% |
| 女性語者 C 單音節 | 88.21%/94.99% |
| 男性語者 B 連續語音 | 81.49%/95.60% |

表四所列之辨識率差強人意，男性語者 B 部分因為其基本頻率範圍較小，一點點波動，就會造成很大的影響，使得聲調一二三的鑑別度下降是主因。女性語者 C 則是一些的個人語言習慣，平調尾部會上揚，造成一點辨識上的誤差。

5. 實作系統

我們綜合前面所述，製作了一個可以在 win95/98 平台上執行的台語聲調辨認系統。這個系統除了提供兩種辨認和現場語音畫基本頻率趨勢的功能外，還有放音儲存的功能，對台語連續語音的變調研究，提供了一個方便的環境。圖四就是我們所製作的系統

<圖四>



6. 結論

在這篇論文中，我們介紹了一個新的聲調特徵擷取方法，EAPA。並與目前最有名的 SIFT 演算法比較。同時我們也介紹了兩種辨認的方法——HMM 及 MLP。在一連串的實驗中，我們證明 EAPA 與 MLP 之組合，頗適於台語聲調辨認。除此之外，我們也對多語者的聲調辨認進行研究，但很可惜的是受限於語料不足，我們得到的只是一個很初步的結果。

7. REFERENCE

- [1] J.D. markel, "the SIFT algorithm for Fundamental Frequency Estimation." IEEE trans. On audio and Electroacoustics, Vol AU-20, No 5, pp.367-377, December 1972
- [2] Ren-yuan Lyu, Yung-jin Chiang, Ren-jou Fang, Wen-ping Hsieh, "A Large-Vocabulary Taiwanese (Min-nan) Speech Recognition System Based on Inter-syllabic Initial-Final Modeling and Lexicon-Tree Search", **ROCLING XI Conference**, Aug. 1998, Hsinchu
- [3] Wu-Ji Yang Jyh-Chyang Lee, Yuen-Chin Chang and Hsiao-Chuan Wang, "Hidden Markov Model for Mandarin Lexical Tone Recognition", IEEE Trans. on ASSP, Vol.36, No7 July 1988, pp 988-992
- [4] Pao-Chung Chang, San-wei Sun and Sin-Horng Chen, "Mandarin Tone Recognition by Multi-Layer Perceptron", ICASSP-90, pp.517-520
- [5] L.R.Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc.IEEE, 77(2) p257-286, Feb, 1989
- [6] A.M. Kondoz "Digital speech coding for low bit rate communications system" 1995
- [7] M.T Hogan, H.B Demuth, Mark Beale, "Neural network design" 1995
- [8] 余伯泉、徐兆泉、吳長能, "台灣語通用拼音". 1999, 台北：南天書局。