# 從中文語料庫中自動選取連續國語語音特性平衡句的方法

王新民[1] 張元貞[2] 李琳山[1,2,3]

[1] 國立台灣大學電機工程學研究所
[2] 國立台灣大學資訊工程學研究所
[3] 中央研究院資訊科學研究所

摘 要：

本文提出一個能從中文語料庫中自動選取連續國語語音特性平衡句的方法，解決了以往必須由人工選造特性平衡句時所遭遇的費時費力的困難，這個方法除了可以有效的找到包含所有辨識單元（例如單音節）、足供訓練用的特性平衡句外，也可以自動加選句子使所選取的特性平衡句與來源語料庫擁有近似的辨識單元的統計分佈，故以它來對語音處理系統進行測試時，所得的結果更能夠反映出系統實際的效能。由於它是一個自動的系統，當應用領域或辨識單元改變時（例如改成聲母、韻母），在不更動原本架構下，只要重新定義新的辨識單元或語料庫，即可輕易的調適到新的應用領域(Domain)。在我們以國語1333個帶聲調的音節為辨識單元的實驗中，這套方法得到相當不錯的成果。

## 一、緒論：

中文連續語音特性平衡句是一組中文文句，它的用途在於能夠提供各種語音處理研究一套有效而完整的連續語音的訓練及測試語料。故所謂的連續語音特性平衡句就是指一組合乎文法和語意的條件，由語音特性差異很大的辨識單元所構成的相當少量的句子，卻能包含所有不同的辨識單元並有合理的統計分佈。由於希望句子的數量儘量減少，故應避免在句子中相同的辨識單元有太多重覆出現的現象[2]。

在過去，語音特性平衡句的產生，主要是由人工來選造[1][2]，也就是經由專家來選句或造句。人工選造語音特性平衡句所遭遇的問題是成本太高，且往往會造出一些較繞口的句子，加上難以兼顧其中辨識單元的統計分佈情形，因此模擬測試的結果並不能反映真實的情況。此外，只要應用領域（Domain）或辨識單元有所改變時（例如音節改成聲母韻母），往往必須重新選造一套適用的語音特性平衡句，此時將再度面臨成本太高的問題。以目前國內電信研究所已發展完成的一套由人工製訂的語音特性平衡句為例[2]，雖然這一套根據國語基本音節（407個不考慮聲調的音節）來選造的語音特性平衡句對於國內語音相關研究有相當的價值並已使用多年，然而它的應用範圍基本上是有所限制的，例如在聲調也必須一併考慮的應用上（407個變成一千三百多個音節），這套語音特性平衡句就不再適用了。基於對連續語音特性平衡句的需求及人工選造的缺點，且近年來國內在語料庫語言學的發展日漸成熟，因此，我們嘗試發展一套能從語料庫中自動選取連續語音特性平衡句的方法。這個方法基本上是一個兩階段式的架構，每個階段所選取的結果可以滿足不同的需求。其中第一階段處理的目的是希望找到一組總數量儘可能小而仍能包含語料庫中所有辨識單元的連續語音特性平衡句的集合。因此，第一階段選取的連續語音特性平衡句足以充當語音處理的訓練語料之用。至於第二階段處理的目的則是以第一階段的選取結果為基礎，從來源語料庫加選句子來使選取的連續語音特性平衡句能與來源語料庫有相近的辨識單元的統計分佈。因此，第二階段選取的連續語音特性平衡句將是極佳的測試語料，所得的模擬測試結果將能夠真實地反映出語音處理系統的實際效能。

這套方法主要的特點在於它具有普遍性(Generality)，不但適用於不同的應用領域(Domain)，而且因為無論將辨識單元定義成音節（帶聲調抑或不帶聲調）、次音節單位（聲母、韻母）、乃至考慮前後文相關(Context dependence)，只要提供語料庫，同樣的一套方法即可滿足

不同的需求。為使本文以下的內容說明簡單清楚，以下將以自動選取以國語1333個帶聲調的音節為辨識單元所需之連續語音特性平衡句為例，來詳細說明我們的方法和實驗的結果,當然這方法也可自動延伸到其他的辨識單元上。

以下本論文的第二、三、四節將詳細介紹這套連續語音特性平衡句的選句方法，第五節則提出我們的實驗結果及相關討論，第六節是結論。

## 二、兩階段式的自動選句方法：

本文所提出的方法是一個兩階段式的選句架構。我們以國語的1333個帶聲調的音節為辨識單元為例，說明其基本構想 。首先第一階段是找出一組數量儘可能最小的連續語音特性平衡句集合，它能夠包含來源語料庫中所有的辨識單元（在這例子中也就是音節），所以這些句子唸出來的語音部分當做連續國語語音處理的訓練語料。但由於第一階段所選取的訓練語料它的辨識單元分佈情況與來源語料庫可能相差極大，所以未必適合做測試語料，因此第二階段的作用即是用來改進第一階段的選句結果，透過從來源語料庫中加選句子來增加連續語音特性平衡句集合與來源語料庫各辨識單元（音節）分佈情形的相似程度。當第二階段的選取結果達到我們要求的相似度標準時，所得到的連續語音特性平衡句集合唸出來的語音便可以做為連續國語語音處理的測試語料。

以下是此方法的進一步介紹，首先是根據我們對連續語音特性平衡句的需求，而有下列的選句原則：

1. 整套連續語音特性平衡句集合必須要涵蓋來源語料庫中所出現的所有辨識單元（也就是音節）。
2. 基於日常使用習慣及方便，長度為6至12字的句子優先考慮。
3. 在同一句子內的相異辨識單元（音節）數目愈多愈好，如此可以包含更多的不同辨識單元間的關連特性。
4. 在同一句子內相異的聲母、韻母數目也愈多愈好。
5. 第一階段中優先選取來源語料庫中低頻率辨識單元（音節）所屬的句子；而第二階段則優先選取來源語料庫中高頻率音節較多的句子，這樣才可達到上述的目標，其詳細原因及作法以下會進一步說明。

此兩階段式的選句方法，其流程如圖一所示，並詳述於以下兩節。

## 三、第一階段選句方法：

這一階段的輸入是來源語料庫，而輸出是一組連續語音特性平衡句集合，可以作為訓練語料之用。其詳細方法如下：

步驟1：統計來源語料庫的音節分佈，對每個音節給定一初始加權值，來代表其重要性，其加權值給法如下：

$$S[i] = N_T / n_T[i] \quad \text{for i} = 1 \text{ to } N_s$$

S[i]：第i個音節的加權值。
$N_T$：來源語料庫中音節總數。
$n_T[i]$：來源語料庫中第i個音節的總數。
$N_s$：國語帶聲調的音節個數(1333)。

也就是某個音節的加權值等於來源語料庫的音節總數除以這個音節在來源語料庫中出現的次數，意即出現頻率愈小的音節會有愈高的分數，這個特性有助於選句處理，以下會有進一步說明。此外，上述這些音節加權值將隨著平衡句逐一選出後而有異動。

步驟2：計算來源語料庫中所有尚未被選取句子的加權值，計算公式如下：

$$S_S = \frac{1}{L}\sum_{i=1}^{L} S[\text{syl}[i]] * W_{hs} * W_{hif} * W_L$$

$W_{hs} = 1 - 0.9 * (\text{句中同音字數目}/\text{句長(L)})$
$W_{hif} = 1 - 0.9 * (\text{句中相同聲韻母數}/\text{句中聲韻母總數})$
$S_S$：某一候選句的加權值。
L：某一候選句的音節總數。
syl[i]：某一候選句的第i個音節。
$W_{hs}$：某一候選句中同音字(Homonym)的加權係數。
$W_L$：某一候選句之句長加權係數。如果句長介於6至12之間，則$W_L$為1；否則為0.5。

它的作法是把候選句中每個字所對應的音節加權值總和，然後除以句子長度，所得的結果來當做候選句的加值；由於候選句中可能會出現同音字的情況，所以我們必須要乘上候選句中同音字的加權係數$W_{hs}$，$W_{hs}$的基本想法是要把具有同音字的候選句的重要性降低；但是如果某個句子全為同音字，而在來源語料庫中這個音節只出現在這個句子中，則依據我們的算法$W_{hs}=0.1$，最後還是會把這句選入平衡句集合中，但是如果在計算公式中少掉0.9這因子，則$W_{hs}=0$，使得該候選句完全沒有機會被選入，則此時不能夠滿足第一項選句原則：平衡句集合必須包含來源語料庫中所有辨識單元的條件。同理可得$W_{hif}$。至於$W_L$的使用則是為了滿足第二項選句原則。

步驟3：選取加權值最高的句子( Stc )加入連續語音特性平衡句集合中。

步驟4：檢查平衡句集合中是否已包含來源語料庫的所有音節，如果是，則結束第一階段選句過程，進入第二階段選句過程，否則繼續執行步驟5。

步驟5：將Stc內每個字所對應的音節加權歸零。

```
for i = 1 to L of Stc
   S[syl[i]] = 0;
```

syl[i]：Stc的第i個音節。

重覆步驟2到步驟5直到連續語音特性平衡句集合中包含來源語料庫中所有出現的音節為止。

在這個階段的初始加權值給法是來源語料庫中的音節總數除以這個音節在來源語料庫中出現的個數。意即在來源語料庫中出現愈少的音節，其

加權值會愈高。所以含有這些音節的候選句子，它的加權值會相對地較其它的候選句子為高，因此會先被選進連續語音特性平衡句集合中。其原因可以下例進一步說明：

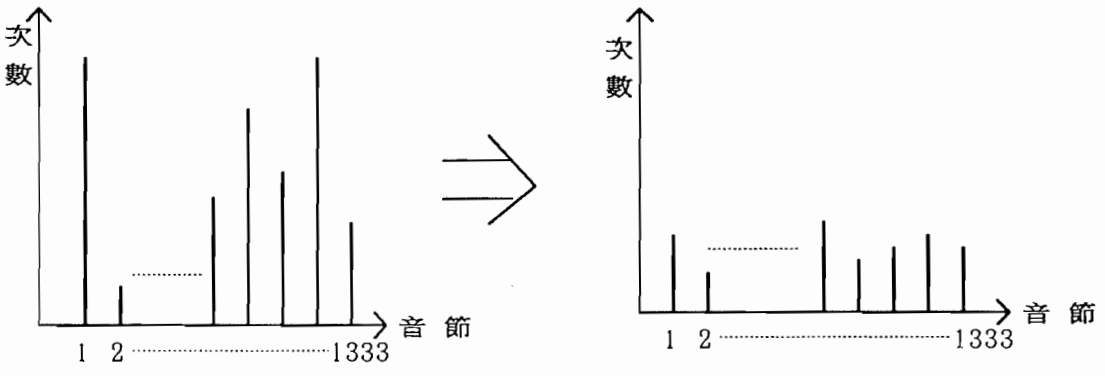
(1) 我 讀 一 本 書

ㄨㄛˇ ㄉㄨˊ ㄧˋ ㄅㄣˇ ㄕㄨ

(2) 我 看 一 本 書

ㄨㄛˇ ㄎㄢˋ ㄧˋ ㄅㄣˇ ㄕㄨ

假設(1)(2)兩句是來源語料庫中的兩個句子，如果ㄉㄨˊ這個音節在來源語料庫中只在句(1)中出現，而ㄎㄢˋ這個音節有較高的出現頻率，從加權值來看，會先選取句(1)到連續語音特性平衡句集合中，因此，後續的選句過程就無須將句(2)再選入，但是如果我們先選取句(2)時，因為ㄉㄨˊ只在句(1)中出現，所以我們必須將句(1)也選入連續語音特性平衡句集合中，這樣的話ㄨㄛˇ、ㄧˋ、ㄅㄣˇ、ㄕㄨ就多選一次，不符合我們希望以較少量字數來句含語料庫中所有音節的要求。此外，在我們把某一候選句選入連續語音特性平衡句集合之後，則將這個句子中所出現音節的加權值歸零，在重覆選句的過程中，只要某一句子的加權值不為0，就表示它含有連續語音特性平衡句集合目前所沒有的音節。因此，當最高分的候選句的加權值是0時，就表示已經找到所有的音節。而這個階段希望達到的理想狀態如下圖所示：




來源語料庫
音節分佈情況

連續語音特性平衡句
音節分佈情況

## 四、第二階段選句方法：

第二階段的選句目的是要趨近來源語料庫的音節統計分佈情形，因此在衡量相似程度方面，採用下面的作法，我們把來源語料庫和連續語音特性平衡句集合的音節分佈情況視為兩個向量，分別是：

$$\vec{V}_1 = (n_T[1], n_T[2], \ldots n_T[i] \ldots, n_T[N_S])$$

代表來源語料庫音節分佈情形
其中，$n_T[i]$表示來源語料庫中第i個音節的總數

$$\vec{V}_2 = (n_b[1], n_b[2], \ldots n_b[i] \ldots, n_b[N_S])$$

代表連續語音特性平衡句音節分佈情形
其中，$n_b[i]$表示連續語音特性平衡句集合中第i個音節的總數

因此，我們可將二者的音節統計分佈相似度定義如下：

$$\text{統計分佈相似度} = \frac{\vec{V}_1 \bullet \vec{V}_2}{|\vec{V}_1||\vec{V}_2|} = \cos\theta$$

亦即以此兩向量的正規化(Normalized)內積值作為統計分佈的相似度，這個值也相當於兩個向量夾角的餘弦值。顯然的，當$\vec{V}_1 = k\vec{V}_2$時，$\cos\theta = 1$，亦即二者完全相同。

這一階段的輸入是來源語料庫，及第一階段所選出的語音特性平衡句，而輸出是具備較理想統計分佈的連續語音特性平衡句集合（測試語料）。

步驟1：根據來源語料庫及第一階段選入之連續語音特性平衡句集合的音節分佈，對每個音節給定一初始加權值，來代表其重要性，方法如下：

```
for i = 1 to Ns
{
  S[i] = Constant;
  Sd[i] = S[i] / nT[i] ;
  S[i] = S[i]- Sd[i]* nb[i];
}
```

$N_s$ ：國語帶聲調的音節個數(1333)。

$n_T[i]$ ：來源語料庫中第i個音節的總數。

$n_b[i]$ ：連續語音特性平衡句集合中第i個音節的總數。

$S_d[i]$ ：第i個音節的減分加權值。

起初，所有的音節都給定同樣的加權值，$S_d[i]$則表示某個音節在每次被選入連續語音特性平衡句所要減掉的分數。 比如說對每個音節給定的初始加權值都是1000，而某個音節在來源語料庫中出現10次，則其減分加權值為1000/10=100分。如果這個音節在第一階段中已經被選入3次，則這個音節真正的初使加權值為1000-100×3=700分。接下來解釋為何這樣的加權值給法可以反映出音節在來源語料庫的分佈情況，比如說另一個音節在來源語料庫中出現50次，已經選入5次，則其初始加權值為1000- (1000/50)×5=900分，因此從音節的加權值來看，我們便可以發現這個音節在這個階段的選句過程中較前一音節重要。透過這樣的加權值給法，即可達到第二階段的選句目的。

步驟2：計算來源語料庫中所有尚未被選取句子的加權值。細節同第一階段的步驟2。

步驟3：選取加權值最高的句子(Stc)，看它在加入平衡句集合後是否能夠改進統計分佈相似度，如果不行則將目前加權值最高的句子，其加權值歸零，重覆步驟3。如果選入的句子(Stc)可以改進統計分佈相似度，則將Stc加入連續語音特性平衡句集合中，並進入步驟4。

步驟4：更新新選入句所包含音節之加權值。

for i = 1 to L of Stc
$S[syl[i]] = S[syl[i]] - S_d[syl[i]]$;

重覆步驟2至步驟4直到達成我們要求的語料相似度標準為止。

綜上所述，我們歸納出這個方法的一些特點。兩個階段的選句結果都各有其用途，而我們只要更換來源語料庫或重新定義辨識單元，就可以將它轉換到不同的應用領域(Domain)上。再者，因為使用的運算並不複雜，所以在執行速度上是可以接受的。此外，因為資料來源是語料庫，所以選出的語音特性平衡句有一定的品質。

## 五、實驗結果

在初步的舉例實驗中，我們選用1333個帶聲調的音節為辨識單元，而來源語料庫是中國時報一個月份的報紙，它含有107419個句子、822829個字元。實驗結果如表1所示，其中效益的定義如下：

$$\text{效益} = \frac{\text{目前累積音節數}}{\text{語料庫音節總數}} \times 100\%$$

| 句　　數 | 角　　度 | 餘　弦　值 | 累積音節數 | 效　　益 |
|---|---|---|---|---|
| 366 | 24.990 | 0.9064 | 2790 | 0.34% |
| 400 | 19.777 | 0.9410 | 3022 | 0.37% |
| 450 | 14.515 | 0.9681 | 3377 | 0.41% |
| 500 | 11.433 | 0.9802 | 3723 | 0.45% |
| 550 | 9.295 | 0.9869 | 4067 | 0.49% |
| 600 | 7.808 | 0.9907 | 4405 | 0.54% |
| 650 | 6.742 | 0.9931 | 4744 | 0.58% |
| 700 | 5.817 | 0.9949 | 5093 | 0.62% |
| 750 | 5.171 | 0.9959 | 5477 | 0.67% |

表 1 ：選句結果統計表

表 1 中的第一列數據為第一階段的選句結果，其餘部份則為第二階段的選句結果，從上表可以明顯地看出只須選到366句(2790音節)，即可涵蓋來源語料庫中所有出現的音節。附錄A提供部份平衡句以供參

考。平均而言，每個音節只出現大約2.5次。此外，僅就第一階段選句的結果而言，已經和來源語料庫有相當的相似程度，因為常用的音節總是會在選每一音節時被夾帶選入，故總數一定較多。當我們進行第二階段的選句時，隨著句數的增加，相似度也隨著迅速提高，當選到750句時，來源語料庫及連續語音特性平衡句集合分佈向量的夾角約等於5度，而其句數遠小於來源語料庫中的句子數目，我們認為這是一個相當不錯的結果。

## 六、結論：

在本篇論文中，我們提出了一個能從語料庫中自動選取連續語音特性平衡句的方法，它不僅能夠幫助我們產生訓練語料，而且也可以產生測試語料。就方法的層次而言，我們提供了相當合理的說明，其實用性及正確性也從實際執行所得的結果得到證實。

## 參考文獻：


〔1〕 "IEEE Recommended Practice for Speech Measurement", IEEE Transactions on Audio and Electroacoustics, Vol. AU-17, No. 3, pp. 225-246, Sep, 1969.

〔2〕 余秀敏、劉繼誼，"國語語音特性平衡句的建立"，電信研究季刊，第19卷，第1期，民國78年3月。

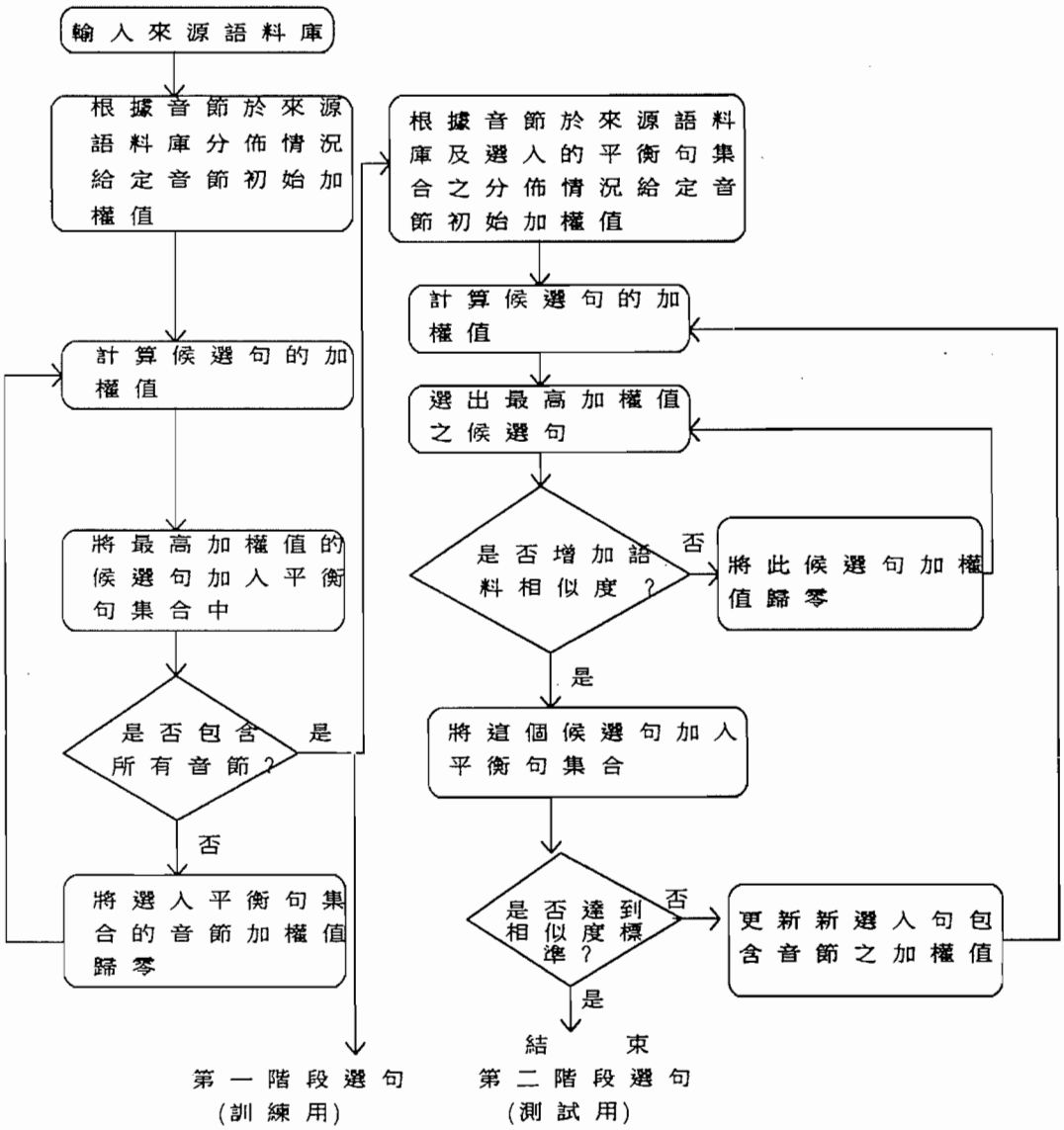

輸入來源語料庫
根據音節於來源語料庫分佈情況給定音節初始加權值
計算候選句的加權值
將最高加權值的候選句加入平衡句集合中
是否包含所有音節？
是
否
將選入平衡句集合的音節加權值歸零
第一階段選句
(訓練用)
根據音節於來源語料庫及選入的平衡句集合之分佈情況給定音節初始加權值
計算候選句的加權值
選出最高加權值之候選句
是否增加語料相似度？
否
將此候選句加權值歸零
是
將這個候選句加入平衡句集合
是否達到相似度標準？
否
更新新選入句包含音節之加權值
是
結束
第二階段選句
(測試用)

圖 一 ：兩階段式選句法

# 附　錄　A：連續國語語音特性平衡句之部份例句

1. 必須撙節開支
2. 希望別再出摟子
3. 小傢伙靦腆地答
4. 鞋襪應大小合適而且通風
5. 雨天則泥濘不堪
6. 終日在院內踱步
7. 為什麼非得提前授課不可
8. 涉嫌拐誘無知少年離家出走
9. 令人有被捉弄的感覺
10. 捐款救助非洲娃娃
11. 咱們等著瞧
12. 拿著臉盆裝水
13. 所謂覆巢之下無完卵啊
14. 無不目眩神搖
15. 儘管粥少僧多
16. 讓歹徒屢試不爽的得逞
17. 藉以表達慰勞之意
18. 嘉南大圳換上了新面貌
19. 不啻是一個奢侈的夢想
20. 債券可能成為未來投資新寵
21. 引起各界揣測不已
22. 這個理由簡直是荒謬可笑
23. 實在令人納悶
24. 肥胖不見得是福
25. 沒有人能忍受這種窩囊氣
26. 怎麼做都不免挨罵
27. 歡迎共襄盛舉
28. 難道醜人真要作怪嗎
29. 民眾若想趁機撈一筆
30. 醞釀杯葛行動
31. 旗幟飄飄十分美麗
32. 我敢拍胸脯保證
33. 慈濟不講深奧的佛理
34. 時報再領風騷
35. 台灣居民生活富裕