西文語音辨識系統之設計研究

陳志堅
國立中山大學電機系副教授
chenc@mail.ee.nsysu.edu.tw

史世洲
國立中山大學電機系碩士生
chenc@mail.ee.nsysu.edu.tw

摘要

本論文探討西文語音辨識系統之設計與實作策略。吾人依照西語的發音規則，挑選出 242 個常用單音節之語音特徵，作為主要的訓練與辨識方式。將每個常用單音節，每輪每次以連續唸兩個相同單音的方式錄音，來作為西語語詞的特性依據。吾人錄製六輪共十二次的聲紋特性，來建立西文語音之訓練資料。系統採用梅爾頻率倒頻譜係數及線性預估倒頻譜係數，來萃取單音節語音之特徵參數；運用隱藏式馬可夫模型，作為單音之辨識模型。在 CPU 時脈為 1.6GHz 的 AMD Sempron 2800+ 之個人電腦與 Ubuntu 9.04 作業系統下，針對 4217 筆西文語詞，吾人約可達到 86%之正確辨識率，平均所需辨識時間約在 1.5 秒以內。而本系統所需的訓練時間約為兩小時。

關鍵詞：西文語音辨識、梅爾頻率倒頻譜係數、線性預估倒頻譜係數、隱藏式馬可夫模型

Abstract

This paper investigates the design and implementation strategies for a Spanish speech recognition system. It utilizes the speech features of the 242 common Spanish mono-syllables as the major training and recognition methodology. A training database of 12 utterances per mono-syllable is established by applying Spanish pronunciation rules. These 12 utterances are collected through reading 6 rounds of the same mono-syllables twice. Mel-frequency cepstral coefficients, linear predictive cepstral coefficients, and hidden Markov model are used as the two feature models and the recognition model respectively. Under the AMD Sempron Processor 2800+ with 1.6GHz clock rate personal computer and Ubuntu 9.04 operating system environment, a correct phrase recognition rate of 86% can be reached for a 4217 Spanish phrase database. The average computation time for each phrase is about 1.5 seconds, and the training time for the system is about two hours.

Keywords: Spanish speech recognition, Linear predictive cepstral coefficients, Mel-frequency cepstral coefficients, Hidden Markov model

1. 前言

西班牙語（Español）也稱卡斯蒂利亞語（Castellano），簡稱西語，是非洲聯盟、歐盟和聯合國的官方語言。根據美國桑莫語言學院(Summer Institute of Linguistics, SIL) 2009 年的統計[6]，全世界約有三億兩千九百萬人使用西文，排名第二名，僅次於中文。主要的使用地區是歐洲的西班牙與拉丁美洲的諸多國家，可見西文是相當重要的國際語言。

吾人將在論文中，建構一套西文語音辨識系統，對西文常用語詞資料庫進行分析，讓使用者輸入西文語音後，可以得知對應的“中文詞彙”，達到查詢及學習西語的效果。系統主要分成兩個設計：訓練模型和辨識系統。訓練模型部分，藉由收集的西語常用字資料庫，經過統計分析後，制定出一套單音模型，建構出資料庫調適化的系統。辨識系統部分，由使用者以西語輸入語料，當聲音訊號進入系統後，經過音節分析，萃取出特徵參數，再與單音模型進行比對、排序後，最後透過文字前後音節的交叉比對，得到最佳的辨識結果。

2. 西文語音學簡介

西班牙文共有 27 個字母，除了 a 到 z 外，還有 ñ ['ɛ nɪ a] 這個比較特殊的字母。表 1 與表 2 為西班牙語的母音與子音發音規則[2,3]。其中母音發音較特別的是：雙母音與三母音中，字母 ia、ie、io、iu、iai 與 iei 的發音會由 [ɪ] 變成 [j]。 而子音的發音規則與一般英文發音相差頗大，較特別的是字母 p、b、t、d、z 與 j 的發音，分別為: [b]、[v]、

[d]、[ð]、[θ] 與 [h]。 字母 h 在西文中一般不發音 [none]。字母 c 一般唸 [g]，但後接 e 或 i，成為 ce 或 ci 時，則唸成 [sɛ] 與 [sɪ]。 字母 g 一般唸 [g]，但後接 e 或 i，成為 ge 或 gi 時，則唸成 [hɛ] 與 [hɪ]。

表 1 西語母音發音表

	字母	發音	字母	發音
單母音	a	[a]	i	[ɪ]
	e	[ɛ]	u	[u]
	o	[ɔ]		
雙母音	ai, ay	[aɪ]	au	[au]
	ei, ey	[ɛɪ]	eu	[ɛu]
	oi, oy	[ɔɪ]	ou	[ɔu]
	ia	[ja]	ie	[jɛ]
	io	[jo]	ua	[ua]
	ue	[uɛ]	uo	[uo]
	iu	[ju]	ui	[uɪ]
三母音	iai	[jaɪ]	iei	[jɛɪ]
	uai, uay	[uaɪ]	uei, uey	[uɛɪ]

表 2 西語子音發音表

字母	發音	字母	發音
p	[b]	s, x	[s]
b, v	[v]	y, ll	[j]
t	[d]	j	[h]
d	[ð]	ch	[ʧ]
h	[none]	m	[m]
c, g, k	[g]	n	[n]
ce, ci	[sɛ, sɪ]	ñ	[nɪ]
ge, gi	[hɛ, hɪ]	l	[l]
que, qui	[gɛ, gɪ]	r, rr	[r]
z	[θ, or s]	f	[f]

西班牙文為表音文字，見字即能讀，故須了解每個字母在語詞中的發音，根據在語詞中的位置以及子音與母音的搭配，每個音節的發音方法與一般羅馬拼音有所差異；故此，我們必須了解音節區分的方法和重音判斷的規則，才能正確發音。音節區分的方法根據母音與母音之間子音數量的不同而有所不同，當母音之間只有一個子音時，子音跟後面的母音一起發音，例如 amigo（朋友）分為 a、mi、go；如果是兩個子音，則第一個子音跟隨前面的母音一起發音，第二個子音跟後面的母音一起發音。要注意的是，西文在區分音節時，有 16 種由兩個子音的組合，亦視為一個子音，分別是 ch、ll、rr、pr、pl、br、bl、tr、tl、dr、cr、cl、gr、gl、fr 與 fl。例如 desgracia（物質）分為 des、gra、cia； 當母音與母音之間有三個子音，則第一個與第二個子音與前面的母音一起發音，第三個子音與後面的母音一起發音，例如 abstruso（深奧）分為 abs、tru、so。

重音判斷的規則在西文中相當重要，除了正確發音外，也用來分辨同字不同義的情形。其重音符號只會標示在母音上，如 á、é、í、ó、ú。主要的重音規則有三，首先，字尾是母音或 n、s 結尾，且語詞中無重音符號，重音在倒數第二音節，例如 amigo（朋友）[a 'mɪ gɔ]；其次，字尾是 n、s 以外的子音結尾，且語詞中無重音符號，重音在最後一個音節，例如 rapidez(速度) [rabɪ'ðɛs]。 最後，若語詞中標有重音符號，則重音在該音節上。

3.語音辨識系統的流程

本論文之語音辨識系統架構，如圖 1 所示，語音錄音之規格為每秒取 11025 個樣本，位元數為 16，音框長度為 256 點，前後音框重疊 128 點。辨識系統架構在時脈為 1.6 GHz 的 AMD Sempron 2800+ 之個人電腦與 Ubuntu 9.04 作業系統下。待測語音信號經音節分析判斷後，運用 20 維線性預估倒頻譜係數（Linear predicitive cepstral coefficients，LPCC）與包含有時間一次微分的 26 維梅爾倒頻譜係數（Mel-frequency cepstral coefficient，MFCC）[1, 4, 5]，進行單音節之雙特徵參數萃取，再與隱藏式馬可夫模型（Hidden Markov model，HMM）所訓練出的西語單音模型，運用維特比演算法(Viterbi Algorithm) [1, 4, 5]，來得到最接近的候選單音群。最後再經由西文語詞文字交叉比對，選出最合適的辨識結果。

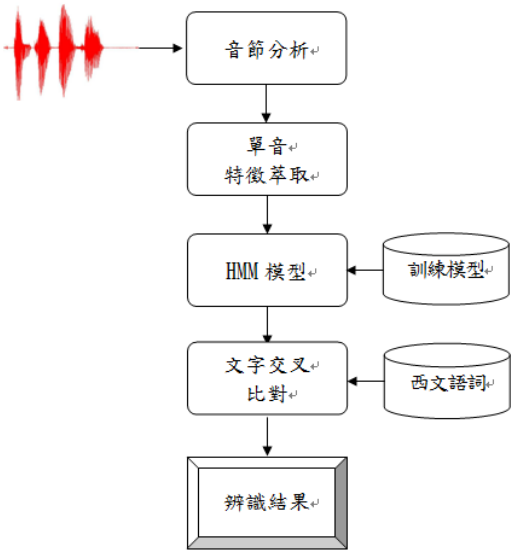

圖 1 辨識系統流程圖

4. 單音訓練與系統辨識率

論文中吾人蒐集並錄製二至五音節共 4217 筆之西文語詞[表 3]，來評量不同的單音訓練方式，所能達到的辨識效能。資料庫內容包含食、衣、住、行、工作、運動、環境與健康等方面之語詞。透過了解辨識率與單音訓練次數的關係、每次錄製不同個數單音對辨識率的影響、以及每輪每次錄兩個單音的條件下，不同訓練次數對辨識率的影響，並經由模擬系統辨識率的高低與單音訓練時間的多寡之實驗結果，來進行分析與討論，吾人得以選取相對有效的單音訓練方式。

表 3 西文常用語詞資料庫統計表

資料量 / 字詞數	模擬語詞數量
2 音節	1235
3 音節	1450
4 音節	887
5 音節	645
合計	4217

4.1 辨識率與單音訓練次數之關係

吾人錄製共 14 輪之西文 242 個常用單音節。錄製一輪須錄完所有 242 個單音語料，才會接著錄下一輪。而單音模型之訓練次數從 4 次開始，所以 6 次訓練的語料將會包含前 4 次的單音語料，依此類推至共 14 次之訓練語料。實驗的目的在了解單音訓練次數與辨識率之關係。由表 4 中，吾人獲知系統之辨識率，在單音訓練次數達到 12 次後，趨於飽和。

表 4 正確辨識率與單音訓練次數之關係

	2 音節 (1235) 辨識率	3 音節 (1450) 辨識率	4 音節 (887) 辨識率	5 音節 (645) 辨識率	平均 (4217) 辨識率
4	77.33 %	82.41 %	88.39 %	90.85 %	83.47 %
6	78.14 %	83.10 %	88.73 %	91.16 %	84.06 %
8	79.03 %	83.45 %	88.95 %	91.47 %	84.54 %
10	79.68 %	83.72 %	89.40 %	91.63 %	84.94 %
12	79.84 %	83.93 %	89.74 %	91.94 %	**85.18%**
14	79.76 %	83.86 %	89.63 %	91.94 %	85.11 %

4.2 每次錄製不同個數單音的影響

在發音過程中，因為習慣的不同，有人說話的速度比較快，有人發某些音則是較特殊。由於上述因素，不同語者在說話時，其字與字之間的發音關聯性會明顯不同。而上小節(4.1)的單音訓練方法，每次只錄一個單音，並無字和字之間的關聯性。有鑑於此，單音模型的訓練，希望能涵蓋相同單音間彼此之關聯性。表 5 為固定個別單音總訓練次數為 12 次後，每次唸 1 到 4 個之相同單音，所能達到的辨識率。觀察表 5，得知每輪每次唸 2 個以上之相同單音，能有效地提升辨識率，其中又以每輪每次唸 2 個相同單音(共唸六輪)之辨識率為最佳。故此，採取每輪每次錄 2 個單音作為訓練單音模型的方式。最後以此方法做訓練，觀察訓練次數不同對此方法的影響。

表 5 每次訓練不同個數單音之正確辨識率

	2 音節 (1235) 辨識率	3 音節 (1450) 辨識率	4 音節 (887) 辨識率	5 音節 (645) 辨識率	平均 (4217) 辨識率
1	79.84 %	83.93 %	89.74 %	91.94 %	85.18 %
2	81.05 %	84.48 %	90.42 %	92.71 %	**85.99%**
3	80.81 %	84.28 %	90.08 %	92.25 %	85.70 %
4	80.89 %	84.14 %	89.97 %	92.40 %	85.68 %

4.3 每輪每次錄兩個單音的條件下，不同訓練次數對辨識率的影響

在上小節（4.2）實驗的單音訓練上，每輪每次唸 2 個相同單音(共唸六輪)之辨識率為最佳。因此吾人希望了解每輪每次錄兩個單音的條件下，不同訓練次數對辨識率的影響。表 6 為其實驗結果。訓練次數超過 12 次，系統正確辨識率，並未顯著提升。由實驗吾人獲知，依據辨識率的高低與單音訓練時間的多寡之權衡，訓練 12 次為最佳之次數。

表 6 訓練次數與辨識率的關係

	2 音節 (1235) 辨識率	3 音節 (1450) 辨識率	4 音節 (887) 辨識率	5 音節 (645) 辨識率	平均 (4217) 辨識率
4	80.08 %	84.07 %	89.97 %	92.09 %	85.37 %
8	80.64 %	84.21 %	90.19 %	92.25 %	85.65%
12	81.05 %	84.48 %	90.42 %	92.71 %	**85.99%**
14	80.89 %	84.34 %	90.30 %	92.40 %	85.81 %

5. 西文語音辨識系統之實作效能

吾人收集西文生活中常用的語詞，包含食、衣、住、行、工作、運動、環境與健康等方面之語詞，總共數量為 4217 筆，經由本論文所設計的單音訓練方式來實作西文常用語詞辨識系統，系統模擬結果如下:

表 7 西文常用語詞之系統正確辨識率

資料量 / 字詞數	模擬語詞數量	辨識率
2 音節	1235	81.05 %
3 音節	1450	84.48 %
4 音節	887	90.42 %
5 音節	645	92.71 %
合計	4217	85.99 %

6. 結論

本論文探討的西文語音辨識系統，是以單音節為主要的辨識方式。由實驗結果得知，增加單音訓練次數，會提升辨識率；但當訓練次數超過 12 次時，辨識率提升的效果將漸趨平緩。為了在語料的訓練次數、訓練效能與訓練時間上取得平衡，本辨識系統的單音模型訓練方式，是將全部的 242 個單音訓練 12 次，共六輪。每輪每次以連續唸兩個相同單音的方式錄製。本辨識系統實作於 CPU 時脈為 1.6 GHz 的 AMD Sempron Processor 2800+ 之個人電腦與 Ubuntu 9.04 之作業系統環境下，針對 4217 個西文語詞，吾人約可以達到 86% 之正確辨識率，平均所需辨識時間約在 1.5 秒以內。而本系統所需的訓練時間約為兩小時。

參考文獻

[1] 王小川，***語音訊號處理***，全華圖書出版社，2004。

[2] 林娟娟，***西班牙文發音法***，冠唐出版社，2006。

[3] 何仕凡，***西班牙語發音一學就會***，三思堂出版社，2002。

[4] 賴昭榮，***中文語音辨識系統降低訓練量之策略研究—以地址系統與二、三、四字詞系統為例***，國立中山大學電機工程研究所碩士論文，2008

[5] 謝文廣，***中文語音辨識系統增進辨識率之策略研究-以地址系統與二、三、四字詞為例***，國立中山大學電機工程研究所碩士論文，2009

[6] M.Paul Lewis, ***Ethnologue: Languages of the World***, 16th Edition, SIL, 2009 http://www.ethnologue.com/ethno_docs/distribution.asp?by=size