# 華台雙語發音變異性之語音辨識研究及 PDA 之應用

*呂道誠 [1, 3]，謝鴻文 [1]，李勇憲 [2]，劉仲英 [1]，許鈞南 [3]，江永進 [4]，呂仁園 [2]*

1．長庚大學電機工程研究所
2．長庚大學電資訊工程究所
3．中央研究院資訊所
4.清華大學統計研究所

E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967
daucheng@iis.sinica.edu.tw, TEL:886-2-27883799ext2104

**Abstract.** 本篇論文提出一種方法來有效的處理華台雙語同時存在於同一句話的語音辨識問題。主要的核心可分為三部分；一.聲學模型：此部分是用一個共同的標音系統，使相同的發音的標音在不同語言上能夠做語料的分享，而且在語音特徵擷取上也加上聲調的參數，以減少華字與音節間的混淆。二.發音模型：此部分是結合了以專家知識為主的發音辭典與實際上語料分析結果而成變異發音，前者是統計了的華台雙語辭典的華字對音節發音機率，找出一個華字在辭典上所有可能的發音；而後者是將音節的辨識結果做成發音對華字的混淆機率。第三部份是將華字直接嵌入在語言模型中，作為搜尋的節點。之後用唐詩300首的實驗，其針對目前台灣地區華台夾雜的語句，以及發音變異性的問題，都能確實降低一成五到兩成的漢字相對錯誤率。最後將此技術移植到PDA上，也做了相關的應用。

## 1 簡介

華語是目前世界上使用人口最多的一種語言之一，數目超過十億[1]以上，最主要的分佈是在中國大陸和台灣，然而中國大陸的國土廣大、地理位置的阻隔、或時間的演變、人口的遷徙與外來語的影響，使得華語產生了許多的變化，在不同的省分人們所講的話雖然都是華語但之間會有些明顯的差異。如北京話、上海話、廣東話、四川話、閩南話等等，以[2]來說，這些話彼此的關係是介於語言與方言之間，因為一種語言是一個國家或一群種族所說的共同話語，而方言是屬於地方性的語言，彼此之間的差異性並不大，很容易理解，然而上述所說的那五種話都屬於華語的分支，但變化又比方言複雜；以數字一到十做為例子，以上五種話的發音各不相同，相互間難理解，因此我們把這些話統稱為省話(因為大部分都是以省分為稱呼)。雖然這些話所發的音不盡相同，但還好，這些省話都有個共同的書寫系統與發音特性，其系統就是"漢字"，發音是以音節為單位，而每個音節都能對應到一個漢字。所以在做語音辨識的人就是在處理"音"與"字"的問題。

語音辨識的目的就是要把人所說的音轉化成文字，因此要做廣義華語的語音辨識，就有點像做多國語言的辨識了。就如之前所說，華語其實包含了所多的省話，有統一的漢字系統，但各個發音都不盡相同，以之前在做多語語音辨識的研究來說，[3]是先將語言的種類辨識出來，然後再用那一種的辨識引擎來辨識那個發音是哪一個字；而[4]是一次用多種語言的辨識引擎來做辨識，看哪一種於言的哪一個字機率最大，另一種[5]是用一個單一的標音方式來將多種語言的聲學模型作結合，也是一次將特定語言的字給辨識出來。而目前在台灣大部分的語音辨識研究都是以華語[6]為主閩南語[7]次之，客語是幾乎沒有，所以本篇論文就是要處理華語和台語雙語的語音辨識研究。

華台雙語除了其字體同樣是漢字以外，聲調也是其一特色。以語音學來說，華語有五種聲調，台語有7種聲調。而以音節的單位來說，中文不帶聲調音節有約400個，而帶調音節約1300個；以一萬三千個中文常用字來說，平均每33個字會對應到一個不帶調音節；而每10個字會對應到一個帶調音節。所以如果在做語音辨識不把聲調的特性也考慮在內的話，會造成嚴重的錯誤，如"睡覺"與"水餃"的混淆。而台語的情況更是嚴重。在[8][9]也證明了華語和粵語加入聲調的特徵會使得整體辨識率上升。因此既然語音辨識就是要把音轉成文字，那麼聲調的問題一定要處理。

對於一個漢字除了因為有不同的省話而產生不同的發音之外，個人的發音習慣、地域不同或上下文連音的發聲耦合(co-articulation)也會影響到一個漢字的發音，這種情況我們稱之為發音的變異性(pronunciation variation)。最明顯的幾個例子為：在台灣大部分的人都常常把捲舌音發成不捲捲舌

音，或台語的入聲音發不出來，我想這都是受到本身母語的影響；因為如果說話者本身是以台語為母語，則台語在語音學中並沒有捲舌的音素，因此在每次遇到捲舌音時，就已相近的不捲舌音來替代。相反的，如果本身是以華語為母語的說話者，因為在他成長的過程中，並沒有受到入聲音的訓練，因此不能正確的發出台語的入聲音。所以在這一方面的情形，我們也要考慮在內， [10]利用決策樹與發現規則的方法來做改進；而[11]是將音節的混淆程度利用機率的方法表現，來做語者的調適；在[12]更是發現其實並不是所有的發音都會有相同的規則，而是在部分的情況下某些發音才會改變，這些文章說明了發音變異性的問題是值得注意且必須重視解決的，這個問題會在未來會越來越多，因為語言的發音一直都在轉變。

以下為本論文的章節安排。首先在第二節將介紹同時處理華台雙語語音辨識所遇到的困難點，並針對這些問題我們提出一個有效的方法來解決。而在第三、四節中我們詳細的介紹國台雙語聲學模型以及發音模型的作法。而後第五節透過實驗測試來驗證此方法的優點。再來第六節並將此技術移植到PDA手機上做了一些相關的實際應用，最後是總結與未來的展望。

## 2 問題定義與解決方案

我們都知道，做雙語或多語的語音辨識比做單一語言來的困難；難就在難於兩種語言本身並沒有統一的發音標記方法、語言本身發音的不同，連帶的所用的文字不同、語料分配不平均的問題、要花許多的時間在瞭解語言，如文法與結構上的差異性大等等。然而目前全球化的速度越來越快，國與國的邊界越來越模糊，也造成語言和語言之間的相互影響越來越密切，一個人因為環境的音素同時會說多種語言的情形也越來越普遍。在台灣也不例外，有將近70%[13]以上的人口在台灣會同時使用華台雙語來作為日常生活中的交談，電視上的連續劇也常常會出現華台夾雜的對白，因此由以上的情形看來，在台灣多語的語音辨識也變的日漸重要了。

然而如果要同時處理華台雙語夾雜的語音，一些相關的問題必須要解決。在[14]提出用樹狀的辭典搜尋法能將以音節[15]為基礎的中文連續語音辨識改進為以漢字為基礎，而且效果快速且能提高辨識率，因此我們採用其特性做為依據，將原本兩階段辨識先辨識音節在轉成漢字的方式，改為直接用一階段的方法來辨識；然而此方式要同時處理兩種語言的話會遇到以下的問題：1. 因為我們不能限制說話者什麼時候講華語或台語，因此在語言模型的設計上必須要用開放式的架構來接受所有可能得情況，這樣來說，不但會增加搜尋空間，而且也會造成空間上不必要的浪費。2.在辭典上如何的有效整合兩種語言的發音？或如何將台語本身的南腔北調問題或是語者本身發音變異性的問題都反映在內？因為我們也不可能把所有可能的發音詞句都納入到發音辭典中，因為這樣也會導致發音的混淆。3.在語言模型中如何整合華台雙語本身的文法或詞結構不同的問題？

在這裡我們提出的方法是用一階段搜尋方式來做華台雙語大詞彙的語音辨識，其是利用華台雙語都是對應到同一漢字書寫系統的特性來解決這個問題。當然這個方法也可以擴大到整個以漢字為書寫系統的語言上，如上海話、廣東話、四川話等等。不管說話者說的是華語還是閩南語或是上海話，語音辨識引擎都是將音轉成漢字，如<圖一>所示。最右邊的語言模型上用樹狀結構的漢字當作搜尋的節點，一來可以加速說尋的速度，二來在做華台雙語的辨識上在不用考慮到語言的問題，因為最後的輸出就是漢字，而不是音節。關鍵就在於<圖一>中間的發音模型；此模型記載了一個漢字所有可能的發音，不管是辭典上有的或是發音變異而產生的，都有其相對應的機率，因此這樣的架構下才能讓使用者在一句話中可以任意的說出華語或台語，而不會增加語言模型的負擔。而聲學模型的架構是不變的，只是在這裡有考慮到華台雙語統一標記與聲調的部分。底下我們就將聲學模型與發音模型做詳細的解說。


<圖一>. 三層次的語音辨識示意圖

## 3 聲學模型

在華台雙語的辨識中，聲學模型要有效的整合華語和台語的語料，以即要考慮到避免在發音模型中造成一個漢字與對應發音的混淆，因此我們提出了兩大方向來處理聲學模型：

3.1 福爾摩沙標音系統(Formosa Phonetic Alphabet)

由於聲學模型是透過語料所產生的，因此如何有效的利用語料使所產生的聲學模型更加強健是一個要考慮的問題。其二，本論文是做語音辨識，而不是語言辨識，不管語者是發台語"阿"的的音，還是華語"阿"的的音並不是重點，因此我們提出了一種標音系統能夠將目前在台灣主要的三種語言：華語、台語以及客語的發音都納入其中，稱為"福爾摩沙標音"系統[16]，簡稱ForPA，其有效的整合此三種語言的發音，一方面讓語料能夠充分的分享，另一方面也可讓音素的分佈更加均勻。因為用ForPA，在華語的音素有37個，而台語有56個，聯集共有63個，而交集的就有32個；相同的標音符號彼此分享語料，所以交集音素的部分在華台語裡面能過獲得雙倍的語料，就某種程度來說這是好的，因為相同的音素有更多的語料可以拿來訓練。在[17]證明了利用此標音方式將華台雙語所訓練出來的聲學模型在相同複雜度的語言模型下其辨識率比單一語言來的好。

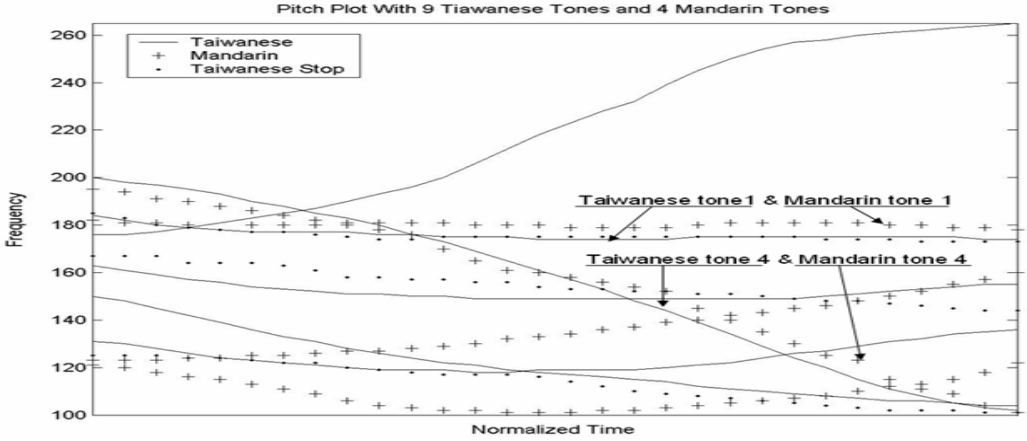
3.2 短時間聲調特徵擷取(short-time tonal feature vectors)

我們都知道，華語一個字的發音結構可拆成基本音結語與聲調，相同的基本音節結合不同的聲調其意義不相同，因此聲調也是一個特徵用來有效的區分字與字的分別。而一般目前在做語音辨識所用的特徵參數都是屬於短時間的梅爾倒頻係數(Mel-frequency cepstral coefficients)，因此為了要將聲調的係數結合梅爾倒頻係數，本論文所採用的聲調特徵取法是用短時間的正規化自動相關函數[18]所得到的。其研究顯示，大部分人的說話頻率介於65Hz到600Hz之間，為了使我們所取的聲調在一個音框中得到比較正確的頻率，通常在一個音框內希望能夠有三次以上的週期，所以在以16k取樣頻率的麥克風的語料在短時間上我們以40ms為一個聲調的音框。

由於聲調具有和諧的特性，為了避免所取的聲調數值變成原本頻率的倍數或一半，我們把自動相關函數所得到的頻率，在一個音框中最多設15個候選值，然後再利用動態規劃(Dynamic Programming)的方法將每一次的發音取得較佳的數值，這樣能夠有效的避免突然的雜訊，但缺點就是比較花時間。

另一個問題就是聲調只有出現在有週期訊號的音框中，如母音或韻母之類的部分，而子音或聲母的部分是沒有辦法得到聲調的數值。為了填補子音部分的聲調特徵參數，[19]提出了用指數函數的方式將聲調與聲調之間的空白連接起來，其中也做了五點平均，以平滑連接的部分，使聲調曲線看起來更加自然。

最後就是每個人的說話頻率不相同，訓練或是測試語料中有男生也有女生，為了使的聲調特徵參數沒有所謂的偏差(bias)，因此我們將其參數作了正規化(normalization)。由<圖二>可看出台語的一、二、三聲調的斜率幾乎相同，如果不作聲調的正規化，我們將不能容易的區分出不同人聲調的一、二、三分別。而正規化的方法有以動態時間為單位、每句話為單位、以男生女生類別為單位、或以每個人為單位，這些方法我們都有做實驗比較，結果是以每句話為單位的正規化效果最佳，因此我們將每句話求出一個最大值的聲調當作分母，而將一句話中每個音框的值都除以分母，而得到聲調特徵參數。




<圖二>. 華台雙語聲調示意圖 (由100人語料所統計而得聲調曲線)，可以看出台語聲調(實線)有三條線其斜率幾乎是相互平行的的。

## 4 雙語字典

在這個部分，將介紹連接聲學模型與語言模型之間的發音模型的作法。發音模型其實是一個漢字與音節的對應表，一邊是漢字，一邊是所有可能的音節，中間是個漢字可能發聲的機率，如<圖一>所示。我們的目的就是要找出一個漢字對華語和台語所有可能的發音，二來就是利用統計的方法來計算其相對的機率。依照華語發音的不同特性如：省話的不同或是腔調口音的影響，此發音模型將以辭典統計與實際語料發音的變異性這兩個方向來進行。

### 4.1 專家知識的方法(knowledge-base approach)

華語的字體是漢字；相同的漢字在不同的省話其發音是不完全相同的，而一種省話又有其他的方言存在，就是所謂的腔調或口音。以台灣的閩南話為例，就有分宜蘭腔、漳州腔、泉州腔、鹿港腔等等，這就是我們常說台語的南腔北調。此情形為一個漢字可以"合法"的對應到許多種的發音，在這裡所謂的"合法"是從各種辭典中找出一個漢字所有可能的發音；就以"長"這個字為例，其華語因為其意義的不同可發成/zhang3/、/chang2/的音，而台語可發為/dng5/、/dinnunn4/；此外台語本身因一個字在一個詞中的位置不同而產生的變調問題[20]也要考慮在內的話，"長"也可發成/dng1/ /dinnunn5/的音。所以光一個"長"字就有六種發音，這樣的發音就是由一些語言學專家所著的辭典統計得來的，而這種以辭典或專家知識為基礎的漢字發音我們稱為多種發音(multiple pronunciations)。

### 4.2 實際資料的方法(data-driven approach)

另外我們也考慮到一個漢字的發音，因個人的發音器官構造的問題、上下文發音的牽連、外來語或母語的影響，而造成原本要發的音變成以相似音來代替，最明顯的例子就是大部分的人在說"輕輕的"發音通常會以"親親的"發音來取代；原本/ng/的發音會變成/n/，我想大家都能體會。再來就是因說話速度的快慢，也會影響到發音。"這樣子"在說話速度快的時候常常會變成"降子"的發音，由三個發音變成兩個發音，這種情形稱之為發音的刪減。在[21]提出華語的發音是以音節為單位，因此大部分的發音變異性通常是取代，而較少刪減或插入的問題，因此在這裡只討論發音取代的問題。另一個現象就是目前台灣特有的華語發音，台灣國語。這種現象就是所謂的母語影響發音的問題，以ForPA來說，華語和台語的基本音素有交集的部分，但也有彼此特有的音素，因此如果從小就習慣以台語為說話的語言，則在發華語特有的音節時某些音就發不太出來，而以台語的相似音節代替，如"吃飯"就會發成"粗犯"，"阿扁"會變成"阿bieng3"，這就是由於台語本身沒有/ㄔ/與/ㄧㄢ/的聲母與韻母。而外來語的影響在台灣出現的比較少，這種其實也是母語影響的其中一個例子。

以上說了這些例子，也許是有規則，也許是沒有規則可循的，這些規則也是要靠觀察實際的語音整理統計而得的。但實際上我們又不可能用人工方式一句句的聽看看有沒有發音變化，因此在這裡我們用音節矩陣來統計出這些發音的變異性。方法是將評估的語料透過基礎的辨識引擎(baseline recognizer)，將帶聲調的音節辨識出來，之後再利用動態規劃(dynamic programming)的方法找出標準發音與辨識結果的發音做單音節文法(one-gram)的對位(alignment)。如此我們就可以得到以帶聲調音節為單位的相似音節矩陣，透過這個矩陣，再與原本的漢字作結合，我們就能得到實際上以語料庫為基礎的漢字發音變異機率。

有了多種發音(multiple pronunciations)與發音變異性(pronunciation variation)的漢字對音節的發音機率，將兩個機率用相同的權重將他們合併起來就成了本論文所用的發音模型。一方面可承接聲學模型，另一方面在語言模型中只要用漢字當搜尋的節點，就能處理以漢字為基礎的華語發音料，因為不管是什麼發音，都有個適當的發音機率對應到漢字，當然在這裡也要顧慮到一個漢字有太多的發音的時候，就會變成贅贅，反而造成辨識上的困擾，因此在實際上還是要做修剪，以達到最佳的辨識效果。

## 5 實驗

### 5.1實驗環境設定

#### 5.1.1語料

本論文所用的華台雙語訓練與測試語料如<表一>所示。訓練語料語料共有100人(50男，50女)，共將近22.5個小時的16k取樣頻率、辦公室環境的麥克風語料。評估語料是為了產生發音模型的另外20人語料。而測試語料是為了測試發音模型的10人的華台雙語語料，這三種語料是相互獨立的。因為本論文是為了測試一句話中同時存在華台雙語情形，因此我們用了唐詩300首做為我們的劇本。每個人各念了100句以台語和以華語為主的詩句，其中有50句是字正腔圓且單一語言的，如<表一>中的MtestR，其代表了以華語為主的標準測試語料。另外50句是使用者以華語或台語為主，但一句話中可以穿插夾雜任何其他另一種語言，自然而然的說出來，如"芙蓉帳暖度春宵"可能會發成這樣的音/(fu2 long2(沒捲舌) zang4(沒捲舌) nuan3) (do3 cun2 siau1)/，前面說華語後面說台語，或者可以非標準的發音來說，如"蓉"與"帳"都沒有發捲舌音，其完全看說話者的喜好或平常講話習慣；如<表一>中的TtestS，其代表了以台語為主的口語化測試語料。

| | 訓練語料 100 人 | | 評估語料 20 人 | | 測試語料 10 人 | | | |
|---|---|---|---|---|---|---|---|---|
| 語料編號 | Mtrain | Ttrain | Mevl | Tevl | MtestR | TtestR | MtestS | TtestS |
| 人數 | 100 | | 10 | 10 | 10 | | | |
| 句數 | 43078 | 46086 | 1000 | 1000 | 250 | 250 | 250 | 250 |
| 時間（小時） | 11.3 | 11.2 | 0.28 | 0.28 | 0.14 | 0.14 | 0.13 | 0.14 |
| 每句平均音節數 | 2.7 | 1.9 | 2.5 | 2.6 | 5.9 | 5.9 | 5.9 | 5.9 |

<表一>. 實驗用的華台雙語訓練、評估與測試語料一覽表

5.1.2**聲學模型**

本實驗是以隱藏式馬可夫(Hidden Markov)模型為主的聲學模型中，每個聲學模型以聲母與右相關帶聲調韻母為單位，其狀態數目分別為3個與4個。相同標音的聲韻母分享彼此的語料。特徵參數共有42個，包含了12個梅爾倒頻係數，一個以對數為單位的正規化能量，再加上聲調的係數；之後再取一階與二階差分係數當作本實驗的語音特徵參數。所使用的工具為[22]

5.1.3**發音模型**

在發音模型中，首先針對多語發音的問題，我們從本實驗室的華台客福爾摩沙辭典中統計出一個漢字所有可能的發音機率，而建立了第一個發音辭典為K-Lexicon。之後再用基礎的華台雙語辨識引擎(辨識率：華語64%,台語也是64%)來辨識另外的10人的評估測試語料，建立了漢字發音變異性的混淆矩陣。再從中統計其發音變異機率，而建立了重實驗或語料中而得的發音變異辭典D-Lexicon。

而在語言模型中，我們就用以樹狀結構為基礎的詞彙搜尋網路，其詞彙量為3223，每個詞彙的節點以漢字為單位。因此不用為了語言的轉換而另外的設計語言模型，這樣在不增加語言模型複雜度的情形下，有效的解決多語發音與發音變異性的問題。

## 5.2 **實驗結果**

本次的實驗室為了測試發音模型在華台雙語中的重要程度，因此我們設計了一套以唐詩300首詞句的測試語料，用了兩套的發音模型，其一就是只用K-Lexicon的多語發音模型，其二是將K-Lexicon結合D-Lexicon的發音變異辭典，在辭典中每個漢字所對應的發音總和為1.0，因此以K-Lexicon來說，平均每個漢字有2.5個發音，而D-Lexicon來說，有1.3個發音，整合之後(K+D)-Lexicon發音辭典，平均有2.7個發音，這是因為我們將發音機率小於0.1以下的都刪除，之後平均的加回到剩下的發音上，以減少一個漢字因太多的發音而造成辨識上的混淆。

實驗的結果列於<表二>，針對唐詩300首的測試語料，我們有用了兩種發音模型，分別是K-Lexicon與(K+D)-Lexicon。整體而言(除了華語的標準發音MtestR)，漢字的錯誤率，使用第二種發音模型比第一種來的好。尤其是用在口語式的語料上(MtestS, TtestS)，有20.1%與15.1%相對錯誤下降率。而華語的標準發音語料為何其相對錯誤率反而是上升的呢？原因應該是因為其語料本身的發音就比較正確，沒有過多的發音變異性，以就是一個漢字就對應到一個標準華語發音，但在發音模型上反而有發音變異性的機率存在(一個漢字對應到2.7個發音)，這樣的發音機率會干擾到辨識，而造成漢字的辨識率不升反降。

| 語料編號 | MtestR | TtestR | MtestS | TtestS |
|---|---|---|---|---|
| 以 K-Lexicon 的 漢字錯誤率[%] | 5.9 | 30.5 | 3.9 | 39.8 |
| 以(K+D)-Lexicon 的 4 漢字錯誤率[%] | 6.1 | 28.2 | 3.1 | 30.7 |
| 相對的漢字錯誤減少率[%] | -3.4 | 11.2 | 20.1 | 15.1 |

<表二> 唐詩300首測試語料用於K-Lexicon 與 (K+D)-Lexicon 的漢字錯誤率

# 6 PDA的應用

從上一節的結果知道，這樣的方法確實能解決部分的問題，因此我們把這樣的技術移植到PDA上，再做實際上的應用，而我們也成功的將此方法移植到XDA II[23]與HP H5550[24]的機子上，底下是其應用與實際上所遇到的問題和解決的辦法。

## 6.1 應用

### 6.1.1語音搜尋mp3播放機：

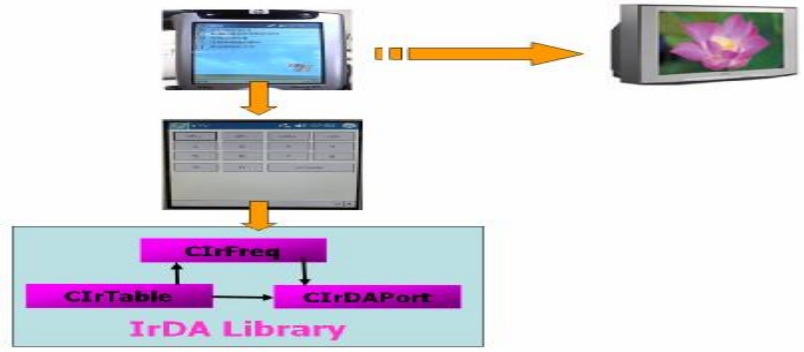
主要的功能就是讓使用者減少搜尋MP3的時間，畢竟有時自己喜愛的歌一多，要找起來的費時許多，在這分秒必爭的時代裏，這將帶給人們不少的便利，所以操作介面以簡易操作、方便使用為主，輸入一段語音聲波後，經多語辨識引擎後，直接以Window Media Player播放之。如<圖三>所示。


<圖三> 語音搜尋mp3播放機

### 6.1.2聲控資訊家電：

IrDA是通過紅外線進行數位信號交換的技術，IrDA數據傳輸技術被推薦使用在高速、短距離，點對點的無線數據傳輸場合，如：掌上電腦、數位相機等。自1994年以後，IrDA數據傳輸技術已經在超過30億電子的產品得到使用，包括PC、筆記型電腦、掌上電腦、印表機、數位照相機、行動電話、PDA等設備。

由於要控制紅外線與電視台名稱頻率對照表，以及各家廠商的紅外線頻率設定，我們以三個class來操作它們，以達到系統的強健性。如<圖四>所示。




<圖四> 聲控電視示意圖

由於各家廠牌紅外線規格不同，下面只舉了Sony廠牌的頻率規格，如<表三>所示。

| Brand | Length | Type | HeadP | HeadS | 1Pulse | 1Space | 0Pulse | 0Space | Space |
|---|---|---|---|---|---|---|---|---|---|
| Sony | 15 | Pulse | 2200 | 550 | 1100 | 550 | 550 | 550 | 23000 |
| P.S：All numbers are time in us (micro seconds) | | | | | | | | | |

<表三> Sony紅外線規格

6.1.3**手機人名撥號：**

手機人名播號其實主要是要減少使用者尋找的時間而且可以Hands Free與Eyes Free，若你在開車時，可以增加很多方便的地方。這個主要只是將連絡簿與名字跟PDA的連絡人做個連接，我們以兩個class來實作完成它。如<圖五>所示。


<圖五> 手機撥號示意圖

## 6.2 問題與解決方法

在整個將PC之辨識核心移植至PDA上，遇到許多程式方面的問題，條列如下：

問題1：WinCE並沒有支援在PC上的string.h這個檔案，我們在PDA上處理字串時，並沒有在PC上來得便利。

解決方法：我們採用傳統c語言中char的方式或是MFC提供的CString與TCHAR來處理我們的大量字串。

問題2：在我們處理漢字多音的對照表時，如果單純用字元指標去處理文字會有問題發生，常常在列印或顯示時，出現一堆莫名奇怪的記憶體暫存資料。

解決方法：使用字元陣列(給定大小值)並給它初始值，便解決這個難以查覺的問題。

問題3：在建立搜尋網路的地方，WinCE所支援之Standard Template Library(STL)的函式庫不是很齊全，所以讓我們在建立網路的過程中，部份演算法及函式都不能使用。

解決方法：儘量利用有提供之演算法與函式來達到我們的要求。

問題4：在PDA上，由於沒有支援ifstream.h、iostream.h、ofstream.h等檔案，在讀寫檔時也是非常不方便。

解決方法：我們用傳統c語音的FILE以及MFC提供之CFile來處理檔案讀寫的功能。

問題5：在聲音處理部份，雖然在PDA與PC上都是控制最底層的WaveInOpen、WaveInStart等函式，但由於PC版的是由Borlnad C++ Bbuilder 6.0所製作完成的，跟PDA的Embeded Visual C++ 4.0是不同，所以聲音處理上，我們也花了許多時間下去研究。

解決方法：將錄放音整個改寫成另一種新的版本，這種格式的錄放音對於移植性有更大的空間。

問題6：最後在做辨識部份的地方，由於資料結構太複雜，故我們盡量使用有STL支援的演算法，但前面有說過，演算法功能的不足，使得我們在做用上與操作上也花了很多心思在改變它處理的方式。

解決方法：改變其比大小與排序的運算子(operator)與函式指標(function pointer)，以達成我們想要的功能。

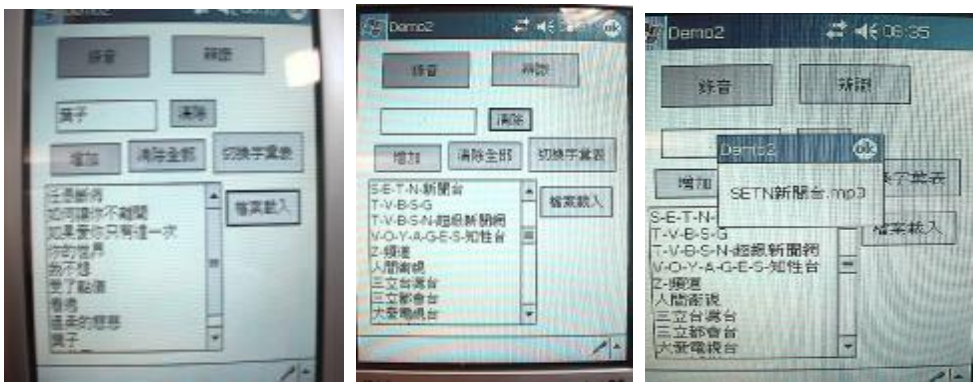
## 6.3 結果

PDA實驗結果列於<表四>，所用的聲學模型與PC上用的相同；而測試語料是用PDA所錄製的， 共有442句，分為歌名、電視台、人名與歌名四大類。而每一類用三種語言來做測試。如第二行的歌名類測試語料種共錄製了32句，其中有10句是用國語發音；20句是用台語發音；而2 句是用英語發音。每句的平均時間長度約在兩秒左右。而關鍵詞彙量就是所謂的語言模型詞彙量。

辨識率平均來說都有不錯的結果。在小於100句關鍵詞彙量的辨識結果都有九成以上，而在兩百句的詞彙量也有8成8。對於人名來說，在辨識上本來就比較難，所以辨識只到84%。辨識時間上，平均每秒的語料所辨識的時間也會因為詞彙量的變大而變慢，這是可以理解的。之後我們再仔細的分析辨識時間成分，大約平均有三分之一的時間花在語音檔轉特徵參數的轉換上，而剩下的三分之二時間花在計算機率與維特比搜尋上。

| 種類 | 測試語料<br>(國語,台語,英語) | 平均每句<br>秒數 | 關鍵詞彙量 | 辨識率 | 平均每秒語音<br>辨識的秒數 |
|---|---|---|---|---|---|
| 歌名 | 32(10,20,2) | 2.344 | 32 | 96.88% | 16.31 |
| 電視台 | 79(30,25,24) | 2.431 | 79 | 94.93% | 23.57 |
| 人名 | 103(68,20,15) | 1.569 | 103 | 84.47% | 35.02 |
| 歌名 | 228(102,75,51) | 2.087 | 228 | 88.60% | 90.92 |

<表四> PDA上的辨識結果




<圖六> Mp3操作介面圖、電視台名稱載入以及辨識結果

## 7 結論與未來展望

本篇論文提出了一個在以華台雙語為基礎的語料，利用整合式發音模型，有效解決多語發音與發音變異性的問題。實驗也證明了此方法的確適合用於雙語或多語的語料，同時也能補償因發音不標準而產生的辨識錯誤問題。其二也將此技術成功的用於PAD上。相關應用也說明了此方法能夠很簡單的加入一些英文的詞彙，做有效的辨識，而不必再重新訓練英文語料。

對於未來的目標我們很希望能夠將個別的單詞句辨識強化成連續多詞彙的辨識，這樣才能真正的利用此技術用於實際上應用。第二對於發音模型的研究，未來希望能找出更好的方法來統計一個漢字的適當發音數目，以瞭解發音的個數與辨識率之間的關係，因為我想太多不必要的發音或單一發音，對辨識率都不好。第三就是對於PAD上的研究，希望能夠更徹底一點，已解決目前辨識數度慢的問題。目前我們的語音長度為3.5秒，從一開始錄音進去，經過整個辨識核心後，平均辨識率約為40秒，速度慢雖是它的缺點，但是辨識結果幾近八、九成是正確的。在速度上，也許硬體會加快或記憶體加大，這都有助於解決速度上方面的缺失。至於為何會那麼慢？跟float point 及fixed point運算這也是關係很大。

## 參考


[1] http://www.sinica.edu.tw/ioe/staff/c9-1-28.htm

[2] W. H. Tsai, “Automatic Identification and Indexing of Chinese Multilingual Spoken Messages,” Ph.D. Dissertation, department of Communication Engineering, National Chiao Tung University, Hsinchu, Taiwan, ROC, 2001.

[3] P. Dalsgaard, O. Andersen, H. Hesselager, B. Petek "Language-Identification using Language-Dependent Phonemes and Language-Independent Speech Units", in Proceedings of the International Conference on Spoken Language Processing, Philadelphia, USA, October 1996.

[4] A Study of Multilingual Speech Recognition, F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, SRI International, 1997

[5] Waibel, A. (2000) Multilinguality in Speech and Spoken Language Systems. Geutner, P.; Tomokiyo, L.M.; Schultz, T.; Woszczyna, M. Proceedings of the IEEE: Special Issue on Spoken Language Processing, Vol.: 88 Issue: 8, pp. 1297 -1313

[6] Bo-ren Bai, Berlin Chen, Hsin-min Wang, Lee-feng Chien, and Lin-shan Lee, "Large-Vocabulary Chinese Text/Speech Information Retrieval Using Mandarin Speech Queries," in Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP98), Singapore, Dec. 1998, pp. 284-289.

[7] Ren-yuan Lyu. Chi-yu Chen, Yuang-chin Chiang and Min-shung Liang (2000) Bi-lingual Mandarin/Taiwanese(Min-nan), Large Vocabulary, Continuous Speech Recognition System Based on the Yong-yong Phonetic Alphabet., ICSLP2000,Oct. 2000, Beijing, China

[8] Hank, Huang C.H., Frank Seide ''Pitch Tracking and Tone Features for Mandarin Speech Recognition''. In Proc. ICASSP, 2000

[9] Tan Lee, Wai Lau, Y. W. Wong and P.C. Ching, "Using tone Information In Cantonese Continuous Speech Recognition," ACM Transactions on Asian Language Information Processing, Vol. 1, pp. 83 - 102, 2002

[10]Mirjam Wester, "Pronunciation Modeling for ASR-knowledge-based and Data-driven Methods," Journal of Computer Speech and Language 17(2003), pp. 69-85, 2003

[11]Lee, Kyung-Tak / Melnar, Lynette / Talley, Jim (2002): "Symbolic speaker adaptation for pronunciation modeling", In PMLA-2002, 24-29.

[12]Liu, Yi and Pascale Fung, “Partial change accent models for accented Mandarin speech recognition.” In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, U.S. Virgin Islands, December, 2003.

[13]http://teach.erdc.cyc.edu.tw/~chinese/newfile9.html

[14]Berlin Chen, Hsin-min Wang, and Lin-shan Lee, "Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues," the 7th European Conference on Speech Communication and Technology (Eurospeech 2001), Demark, September 2001.

[15]Hsin-min Wang, "Experiments in Syllable-based Retrieval of Broadcast News Speech in Mandarin Chinese," Speech Communication, 32(1-2), pp. 49-60, Sept. 2000.

[16]Liang M.S., R.Y. Lyu, Y.C. Chiang "An efficient algorithm to select phonetically balanced scripts for constructing corpus" NLP-KE, Beijing 2003

[17]Dau-Cheng Lyu, Bo-hou Yang, Min-Siong Liang, Ren-Yuan Lyu, Chun-Nan Hsu "Speaker Independent Acoustic Modeling for Large Vocabulary Bi-lingual Taiwanese/Mandarin Continuous Speech Recognition", SST, Melbourne, 2002

[18]Paul Boersma ''Accurate Short-Term analysis of the Fundamental Frequency and the Harmonics-To-Noise Rate of A sampled Sound'', 1993.

[19]Dau-Cheng Lyu, et al, "Large Vocabulary Taiwanese (Min-nan) Speech Recognition Using Tone Features and Statistical Pronunciation Modeling" In Proc. EuroSpeech, Switzerland, 2003.

[20]R. Y. Lyu, Z. H. Fu, Y. C. Chiang, H. M. Liu "A Taiwanese(Min-nan) Text-to-Speech(TTS) system Based on Automatically Generated Synthetic Units", the 6th International Conference on Spoken Language Processing (ICSLP2000), Oct. 2000, Beijing, China

[21]Mingkuan Liu, Bo Xu, Taiyi Huang, Yonggang Deng,Chengrong Li, "Mandrain Accent Adaptation Based on Contest-Independent/Context-Dependent Pronunciation Modeling," In Proc. ICASSP,

2000
[22] Steve Yang et al. Hidden Markov Model Toolkit V3.1, Cambridge University Engineering Department, 2002
[23]http://www.my-xda.com/
[24]http://www.search4hardware.com/10/p_10_246_HP_h5550.html