混合語言之語音的語言辨認

Language Identification on Code-Switching Speech

朱晴蕾 1, 呂道誠 2 , 呂仁園 1
1. 長庚大學資訊工程研究所
2. 長庚大學電機工程究所
E-mail: rylyu@mail.cgu.edu.tw, TEL:886-3-2218800ext5967

摘要

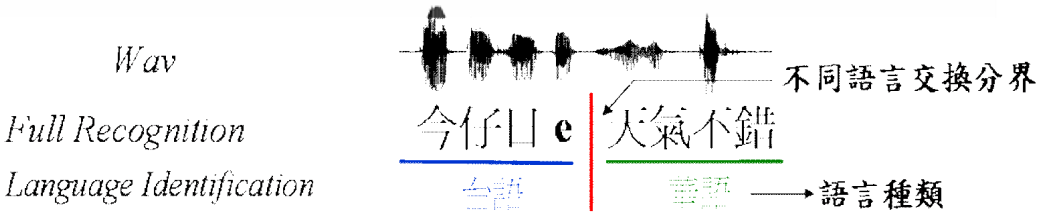
本論文主要針對台灣地區所出現的華語與台語的混合語言語音(Code-Switching Speech) ，研究其語音的語言辨識。藉由自動語音辨識技術得到語音音節序列，經語言詞典比對產生斷詞後，以字詞在不同語言中出現之頻率和字詞組合機率爲判別語言共同字詞準則，進而得到語音之語言種類辨識。最後以語言標籤和語言時間資訊兩種不同的評估方式，分別對語言辨別進行進一步正確率評估，約可達到 83.4%及 78%的語言辨識率。

一、緒論

傳統 LID 主要用於判別出一段語音是由何種語言所建構成，例如：「今天天氣不好」、“How are you？” …等，所判別的語音本身是僅由一種語言所組成。但由於多語的社會環境，因大量吸收外來語，及方言接受度的提高，導致現代語言結構規則發生改變，或爲因應不同需求或習慣，而擷取不同語言片段（/特徵）以創造新詞，使得越來越多不同語系的語詞夾雜應用於陳述上。在如今全球溝通交流普及的環境下，語言轉換（Code-Switching）技巧不僅在報章雜誌大眾傳媒上頻繁出現，也在日常生活上被一般大眾普遍使用。

所謂的混合語言語音（Code-Switching Speech）指的就是一段由 2 種或 2 種以上語言交替組合而成的語音。說話者從一種語言轉用另一種語言，轉換原因諸如遇到談話主題、對象或場合的改變，或是說話者雖有某事物的概念、卻僅能以某種語言形式表達時，語句便會在中途產生轉換，局部轉切至該語言形式，例如：「Star Bucks 的咖啡很好喝」爲華英 2 種語言的轉換用法，而「今天晚上有“夜市仔” 」，則爲地方方言與國語混合使用，爲華台 2 種語言的轉換用法。

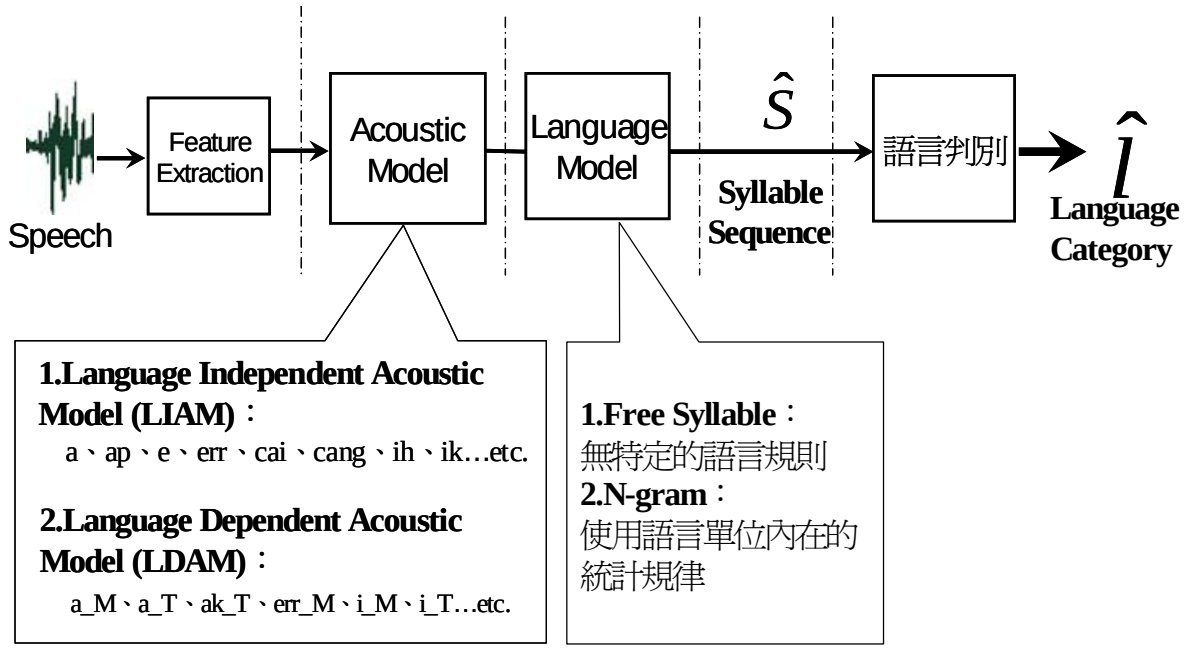
我們的研究目標，主要便是針對在台灣地區最普遍出現的語言交雜狀況—也就是以華語爲語言主幹，在語句中穿插其他種類語言的混合語言語音（Code-Switching Speech）者—爲研究對象。期望能找出一個較佳的策略，以區分一段未知語音中的不同語言的交換分界處，進而有效判斷出語言種類。如圖一所示，我們期望能找「今仔日 e」爲一個台語語音片段，而「天氣不錯」爲一個華語語音片段的結果。



圖一、Code-Switching Speech

目前在自動語言辨識的課題上，大部份仍是以純單一語言構成之語音爲主，混合語言語音之研究則較爲少見。香港的 Joyce Y. C. CHAN 等人曾在 2004、2006 發表關於廣東話與英語混雜的混合語言語音研究[7][9]，分別嘗試使用 Knowledge Base 和 Data Driven 兩種不同的語言混合音素集合辨識語音序列，並以 Bi-phone 出現在廣東話中的機率做爲語言判別準則，其辨識結果約 69.62%、76.54%。而在國內，成功大學林俊憲、Chi-Jiun Shia、Chung-Hsien Wu 等人對於華語與英語混雜語音的相關研究[4-6]，使用 BIC 預先切割語音，再以加入隱含式語意索引(Latent Semantic Indexing) [2,3] 所訓練而成的高斯混合模型來辨別語音音序，以達到混合語言辨別的成效，其辨識結果約爲 74%。前者在語言辨識的架構較爲簡單且亦可得到一定語言辨識效果，故我們採取類似 Joyce Y. C. CHAN 等人的實驗系統架構，並於後端語言辨別處理部份則加以改進，期望能達到更高的語言辨識率。

我們系統架構流程如圖二所示：當一個語音進來時，先將其進行特徵的截取，經由聲學模型和語言模型兩部份，將得到一個 syllable 序列；接著將這個 syllable 序列給予語言判別區塊，以詞典斷詞、語言共同詞處理判斷，得到這整段語音的語言類別；最後再使用語言標籤和語言時間資訊爲兩種不同的評估方法，評鑑整體語言辨認正確率。



圖二、系統架構流程圖

本篇論文內容分別爲：第二節音節辨識，介紹所使用之聲學和語言模型，第三節語言判別策略，第四節實驗用語料、正確率評估方式和最後的結果討論。

二、音節辨識

在我們的語言辨認系統中，是基於自動語音辨識所得到的音節序列爲基礎，再做後續的語言辨識。而音節序列的正確與否將會影響後續語言辨識的正確性。

（一）聲學模型

在語言的標音上，我們使用福爾摩沙音標(Formosa Phonetic Alphabet, ForPA)。ForPA 音標是一個台灣地區三語(華台客)共用標音系統，在華語的音素有 37 個，而台語有 56 個，兩種語言音素聯集共有 63 個，而交集的有 32 個。

以 ForSDAT (Formosa Speech Database) 台灣地區多語言語音資料庫做爲訓練資料（Training Data），華語部份使用 ForSDAT 中 MD01 語料庫，寮由 100 人所錄製成的華語句型語料，總長度爲 11.3 小時。台語部份使用 ForSDAT 中 TW01 語料庫，爲由 100 人所錄製成的台語句型語料，總長度爲 11 小時。每個聲學模型皆以音節內左右相關的方式的三個連續音素爲單位，使用 3 個狀態的隱藏式馬爾可夫模型（Hidden Markov Model：HMM）。

在聲學模型方面，分爲語言獨立的聲學模型（Language Dependent Acoustic Model, LDAM）：在聲學模型中所有的音標沒有語言上的分別，如 a、ap、e、err、cai、cang、ih、ik…etc，整個聲學模型語言是由華語與台語共同訓練而成。另外，我們對於每種語言分別訓練其獨立的聲學模型，並在每個聲學模型中標記其所屬的語言類別，再將不同語言的聲學模型結合在一起，形成一個混合語言的聲學模型[11]，稱之爲語言相依之聲學模型（Language Independent Acoustic Model, LIAM）：在聲學模型中所有的音標皆帶有所屬語言的標籤，如 a_M、a_T、ak_T、err_M、i_M、i_T…etc.。

（二）語言模型

N-gram 的語言模型可以提供一種語言中其文字的序列規則，並以統計和機率的方式來呈現。以下是 N-gram 的表示式，W 爲 n 個 w 的集合，其中每個 w 代表一個字詞。假設第 n 個字詞的出現只與前面 N-1 個字詞相關，而與其他任何字詞都不相關，整句的機率就是各個字詞出現機率的乘積。這些機率可以通過直接從語料中統計 N 個字詞同時出現的次數得到。當 N 愈大時所需統計語料將愈多。

$$P(W) = P(w_1 w_2 w_3 \ldots w_n) = P(w_1) P(w_2 \mid w_1) \cdots P(w_n \mid w_1 w_2 \ldots w_{n-1})$$

$$= \prod_{i=1}^{n} P(w_i \mid w_{i-n-1} \ldots w_{i-2} w_{i-1})$$

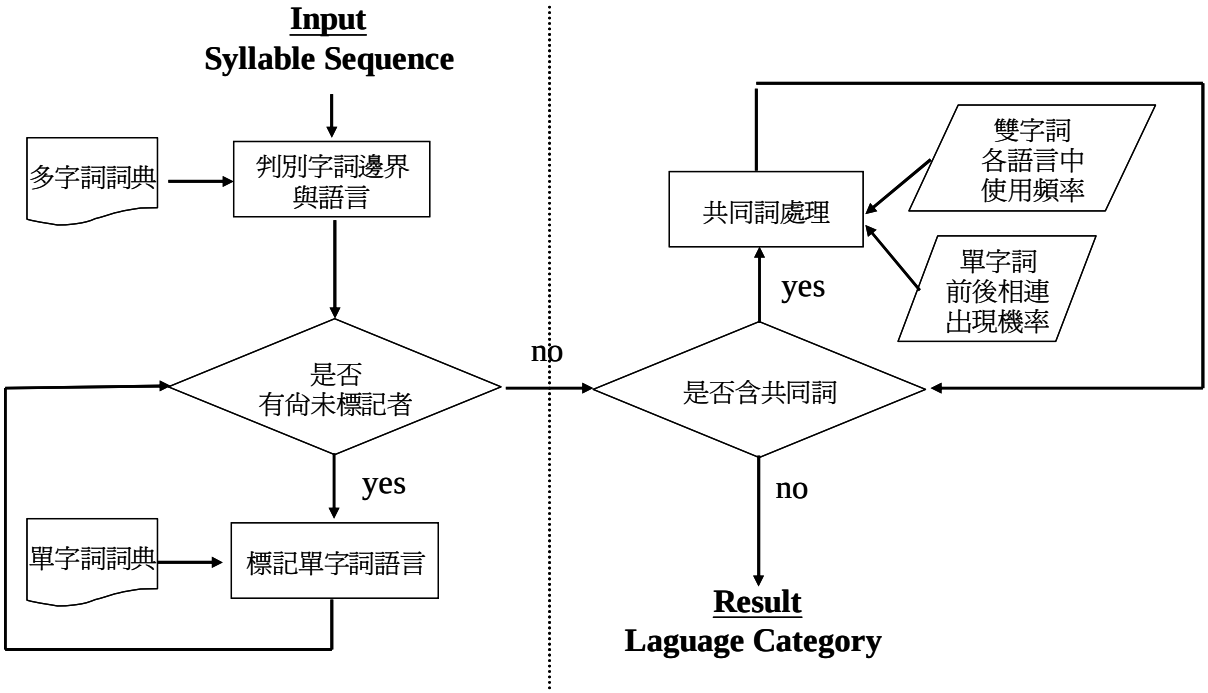
對正確的語言現象，字與字之間的共同出現機率較高，對一些較不符合語法者，字與字之間的共同出現機率較低。此機率的大小可以反映出一個語言的局部規律。例如：今天和今仔日有相同的意義，但“今+日”這種組合在華語中出現機率相對比在台語中出現機率高，而“今+仔+日”這種組合在台語中出現機率相對會比在華語中出現機率高。

依據現有已標音完成文字語料資料量，我們使用的 N-gram 語言模型上取 N=2，亦稱之爲 bi-gram，並以 Syllable 爲單位量。我們所使用來訓練 Bi-gram 機率的語料，在華語文章部份總句數約有 1 萬 7 千（17203）句，全部約 23 萬字（230236）左右；台語文章部份總句數約 9 千（9539）句，全部約 10 萬字（104324）左右。其中，所有文章中的句子皆僅由一種語言組成。

在 Bi-gram 的情況下，一個字詞出現的機率只會依據上個字詞而出現。Bi-gram 的機率表示法爲： $P(w_t \mid w_{t-1}) = \frac{C(w_{t-1}, w_t)}{C(w_{t-1})}$ ， w_{t-1}, w_t 表示時間上相連的兩個 Syllable，$C(w_{t-1}, w_t)$ 爲 w_{t-1}, w_t 在文章中共同出現的次數， $C(w_{t-1})$ 則爲 w_{t-1} 在文章中出現的次數。

三、語言判別

一段語音進來後經由聲學模型和語言模型後可以得到一 syllable sequence，將這個序列再進行語言判別。以流程圖圖三所示，當一個 syllable sequence 進來後，會先判別字詞邊界與語言，若序列中每個 syllable 皆已標記了可能語言種類後，再確認所標記的語言種類中是否有包含了語言間共有的字詞後，將這些共有字詞再進行第 2 次判別，最後將可得到這個序列的語言類別。這是我們的語言判別大略的步驟，詳細的內容在之後會逐一說明。



圖三、語言判別流程圖

（一）字典與詞典制作

首先，先介紹在一開始判別字詞邊界與語言這個步驟所用到的多字詞詞典與單字詞詞典。多字詞詞典部份，我們使用中研院的華語詞典與實驗室既有的台詞詞典混合成一

新的華台混合詞典，總詞數約 14 萬(143705)詞，並將每個多字詞後標記所屬語言種類。整個詞典將被分成三種語言類別；僅會在華語中出現的多字詞詞，標示爲 M，約 8 萬詞(88675)；僅會在台語出現的多字詞詞，標示爲 T，約 5 萬詞(53499)；最後一種爲兩種語言中皆可能出現者，標示爲*M/*T，約一千五百詞(1531)。

單字詞詞典的製作與多字詞詞典雷同，在單字詞詞典中將含有兩種語言所有可能出現的發音，總單字詞數爲 1010 個，並同樣分爲三種語言類別。僅在華語中使用的單字詞數爲 230 個，僅在台語中使用的單字詞數爲 580 個，華語與台語同時使用的單字詞數爲 200 個。

（二）判別字詞邊界與語言

我們實際觀察一些混合語言的語音文句，發現語言轉換時機似乎經常發生在詞（word）上，因此我們在這裡先假設語言轉換的時機在於 word 上，故 word 與 word 的邊界便可能爲一個語言交換的邊界處。在判別字詞邊界與語言這個步驟，便是將 syllable sequence 與詞典中詞句最長匹配比對，將比對到的部份做爲一詞與詞的分界點，並依詞典中每個詞後所帶的語言標記做爲其語言種類。例如：“uo zuei si huan cyu ia ci a chii dong si”「我最喜歡去（夜市仔）吃東西」這句話，若詞典中分別包含有“uo zuei si huan cyu”、“ia ci a”、“chii dong si” 三個詞，那判別的結果將會爲：uo zuei si huan cyu（M）|| ia ci a（T）|| chii dong si（M）|| 。

然而，於此方式的初步成果中，我們可以發現一些問題存在。

首先是在 syllable sequence 中的 syllable 組合有可能不存在於詞典中。以圖四爲例，最後的 dou hern tien（都很甜）這個 syllable 組合並不存在於詞典中，而這種問題我們將改以直接使用字典判斷單一 syllable 的語言種類來解決，故最後會得到 dou（M）||hern（M）|| tien（M）||這樣的結果。另外一個問題在於，有些字詞是同時存在於不同語言間共同使用，也就是原先分類爲*M/*T 的字詞，如例句中的 lai a（*M/*T）||，而這個部份，我們將以另一套方法來處理判斷。

最 近 的 梨仔 都 很 甜
zuei zin der (M)|| lai a (*M/*T)|| dou (M)|| hern (M)|| tien (M)||
初步斷詞結果 語言共同字詞 音節組合不存在於多字詞詞典中

圖四、判別字詞邊界與語言初步結果

（三）共同詞處理

在共同詞的處理上，會分爲多字詞與單字詞的兩種處理方式。在多字詞的處理上，我們使用中研究 CKIP 語料庫與實驗室語料庫所統計得到的在不同語言所出現的字詞與其出現次數表，利用詞在不同語言出現的相對頻率多寡來決定所屬語言。例如，共同發音字詞 ker ci，在台語中出現次數爲 149 次，而華語僅 79 次，分別除以所統計的華語、台語總詞數，取兩者相對頻率較高者，而在此我們可以發現得到的在台語詞中出現機率

較高，故將 ker ci 標記爲一個台語發音詞。

而在單字詞方面，由於我們使用了兩種不同的聲學模型，故將以兩種不同聲學模型所得到的 syllable sequence 分別做討論。

1、 Language Independent Acoustic Model（LIAM）：

若聲學模型爲 LIAM，則得到的字串中每個 syllable 本身即會帶有語言標籤。在這種 syllable sequence 中發現的語言共有單字詞，便直接以 syllable 本身所帶的語言標籤做爲語言依據。

2、Language Dependent Acoustic Model（LDAM）：

若爲另一種聲學模型 LDAM，所得到的字串中 syllable 將不帶有語言標籤。因本身沒有語言標籤，故我們無法以直接的方式得到語言種類。考慮到在我們假設轉換爲 word 的前提下，應不會以一個 syllable 做一次快速轉換，若前後兩個相鄰詞爲相同語言時，則此 syllable 應與前後兩相鄰詞爲同一語言。

以上爲第一種方法，但這個想法並無法完全解決所有語言共有單字詞的問題，若語言共有單字詞出現在前後兩相鄰詞爲不同語言類別間時，便會無法處理。爲此，我們提出了第 2 種方法，利用與前後兩相鄰單字詞同時出現的機率大小做爲判別依據。與 N-gram 的概念相同，當機率值相對較高時，表示愈有可能爲同種語言。在加入這個方法後，將可以成功處理所有語言共有單字詞。

四、實驗評估

（一）語料庫設計

由於我們在現有的所有語料庫中，並沒有任何混合語言的語音資料。故在進行實驗前，我們將先行錄製建立一混合語言的語料庫做爲我們整個實驗的來源依據。

在錄製語音前，我們需要有混合語言的文句做爲錄音劇本。在混合語言的文句收集上，首先先由各式文章中，例如：電子新聞、部落格…等，尋找日常生活中可能使用到的混合語言文句來做爲我們的錄音劇本。由於台語爲一種方言（Dialect），在文字的呈現上，可能與華語使用文字上相同，或使用不常使用較爲特別的文字，較不易直接在一般文章中發現。因此，除了在各式文章中直接尋找華語與台語混合語言文句外，我們參考[8]的作法，我們將兩句由不同語言構成但意義完全相同的文句做一比對，可得到兩種語言在句法節構與用字上的相對情況，並將文句依相對情況切分成數個文字區塊。選取原文句中某文字區塊，替換成另一種語言文句中相對應的文字區塊，如此便可得到一混合語言文句。

在華語與台語混合語言文句上，我們收集了約 75 句。表一爲其中的一些例句。我們所錄製的華台混合語言語料庫，參與錄音的人數總數爲 10 人，約 750 句，平均每句長度爲 2 秒(0:02.391)，由麥克風及個人電腦的 32k 音效卡，在安靜無噪音的環境底下錄製 16KHz，16bits 的聲音訊號，並以 WAV 格式音檔儲存。

表一、華台語混合語言語音例句

Filename	Text	Transcription
MT_001	我最喜歡去（夜市仔）吃東西	uo3_zuei4_si3_huan1_cyu4_ia2_ci2_a4_chii 1_dong1_si1
MT_005	(歹勢！)我遲到了	painn4_se3_uo3_chii2_dau4_ler0

（二）、實驗結果

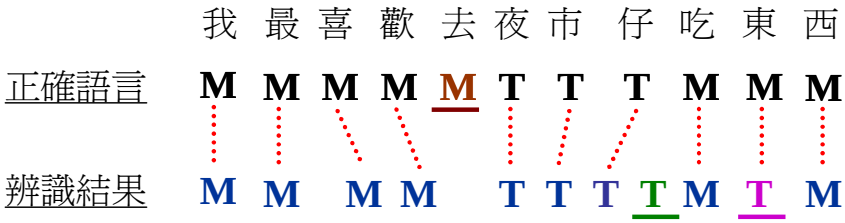
我們使用了兩種不同的聲學模型，以及兩種不同的語言模型來做比較，共分爲四種不同的實驗組合：語言獨立聲學模型與 Free Syllable 語言模型、語言獨立聲學模型與 Syllable Bi-gram 語言模型、語言相依聲學模型與 Free Syllable 語言模型、語言相依聲學模型與 Syllable Bi-gram 語言模型。將四種組合所得到的 Syllable Sequence 交予我們的語言判別策略做語言上的辨認，可得到四種不同的語言辨認結果。

在混合語言語音 LID 的實驗中，由於沒有語言轉換邊界的資訊，故無法直接得到如單一語言語音 LID 的語言區段。在混合語言語音 LID 所得到的結果會是以音節序列的方式做爲初始呈現，因此在這裡我們對原先在單一語言語音 LID 正確率計算上做了些改進。我們將辨識結果中相鄰爲相同語言的語言標籤合併，形成數個單一語言區塊，而每個單一語言區塊的語言標籤便爲我們計算正確率的單位量。這種單位的計算方式與在單一語言語音 LID 計算正確率的單位量是相同的，而做爲在混合語言語音 LID 的語言辨認正確率計算單位時，可同時顯示出語言轉換邊界的正確程度。另外，我們增加了另一種計算單位量，以組成單一語言區塊的基本單位－音節爲主，當完成語言辨識後一個音節即有一個對應的語言標籤，也就是最原始的結果單位。以這樣的計算單位，在語言辨認正確率評估上可同時顯示我們在音素辨認上的正確程度。

在評估語言辨認結果正確率上，我們同樣使用了兩種不同的評估方式。第一種是和單一語言語音 LID 正確率判別相同的評估方式，以語言標籤爲評估單位。另外我們在原本正確率的評估上再加入時間上的資訊，也就是第二種評估方式，以語言時間資訊做爲評估單位。

1、語言標籤評估法

第一種評估方法是以語言標籤來做爲評估單位。如圖五爲例來說，若有一句話：「我最喜歡去夜市仔吃東西」，正確答案的語言標爲：MMMMMTTTMMM，若我們辨識出來的結果爲 MMMMTTTTMTM，將正確答案與語言判別結果做比對，可以得到 4 種不同情況數值，分爲別：Hit：正確答案的語言標籤與語言判別結果的語言標籤完全相符。刪除錯誤（Deletion）：語言判別結果所缺少的語言標籤部份。插入錯誤（Insertion）：語言判別結果所多餘的語言標籤部份。替換錯誤（Substitution）：正確答案的語言標籤與語言判別結果的語言標籤不相符。



圖五、語言標籤做爲評估單位範例

在得到這四種不同的數值後，以現在最常使用的兩個衡量標準來評估我們的方法：精確率（Precision Rate）：在語言判別結果的語言標籤總數中，正確語言標籤所佔比率。N_H 爲語言判別結果中正確的語言標籤數(Hit 總數)，N_L 爲語言判別結果全部的語言標籤數(Hit 總數+Substitution 總數+Insertion 總數)。

$$precision = \frac{N_H}{N_L}$$

召回率（Recall Rate）：在正確答案中有被判別出來且爲正確的語言標籤佔正確答案語言標籤總數比率。N_H 爲正確答案中有被辨識出且爲正確的語言標籤數(Hit 總數)，N_C 爲正確答案中的語言標籤總數(Hit 總數+Substitution 總數+Deletion 總數)。

$$recall = \frac{N_H}{N_c}$$

通常，當有高召回率時，便很難能有高精確率；反之，有高精確率時，將難有高召回率。故我們另使用 F-Measure 做評估。F-Measure 是依據上面的精確率和召回率兩個衡量標準，加以綜合而成的另一個評估指標。

$$\frac{1}{F-measure} = \frac{1}{2}\cdot\left(\frac{1}{precision} + \frac{1}{recall}\right)$$

$$\Rightarrow \quad F-measure = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = 2\cdot\frac{precision\cdot recall}{precision + recall}$$

在 F-Measure 中，需要同時滿足精確率和召回率皆較高時才能得到高的數值，而這樣的方法可以在精確率和召回率上取得一個平衡。

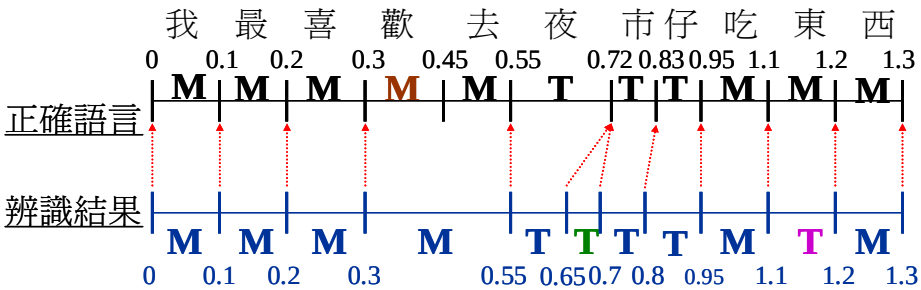
表二、以語言標籤做爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段爲語言計算單位		
	P	**R**	**F**	**P**	**R**	**F**
LIAM+Free Syllable	72.63%	68.55%	70.53%	58.67%	93.75%	72.17%
LDAM+Free Syllable	69.9%	65.7%	67.6%	69.11%	86.82%	76.96%
LIAM+Syllable Bi-gram	72.73%	68.71%	70.66%	82.26%	76.26%	79.14%
LDAM+Syllable Bi-gram	78.1%	68.51%	**73.02%**	82.35%	79.0%	**80.67%**

表二爲我們以語言標籤做爲評估單位的實驗結果，在這種評估單位下，我們可以發現在 LDAM+Syllabe Bi-gram 的實驗組合並以合併後的語言區塊爲計算單位的情況下能得到最高的正確率。

2、語言時間資訊評估法

第二種評估方式是使用語言時間資訊來做爲評估的單位，語言的時間資訊所指是每個語言標在語音中的起始時間、結束時間與持續時間長度(Duration)。這種評估方式，是參考[10]在辨認語音中不同語者(Speaker)的正確率計算方法，由於在語音中有一個以上不同語者交替出現與語音中有不同語言交替出現相類似，故我們使用[10]的評估方式來評估我們的正確率。首先，將語言判別結果的語言時間邊界與正確答案語言時間邊界做對位（Alignment）的動作，以判別結果語言區塊時間佔正確答案語言區塊時間比例爲主要對位上的準則。以圖六爲例，在完成對位後再比較兩者的語言標籤，同樣可以得到 Hit、Deletion、Insertion、Substitution 4 種數值，但在這邊數值的計算方式是取用時間爲單位來表示。



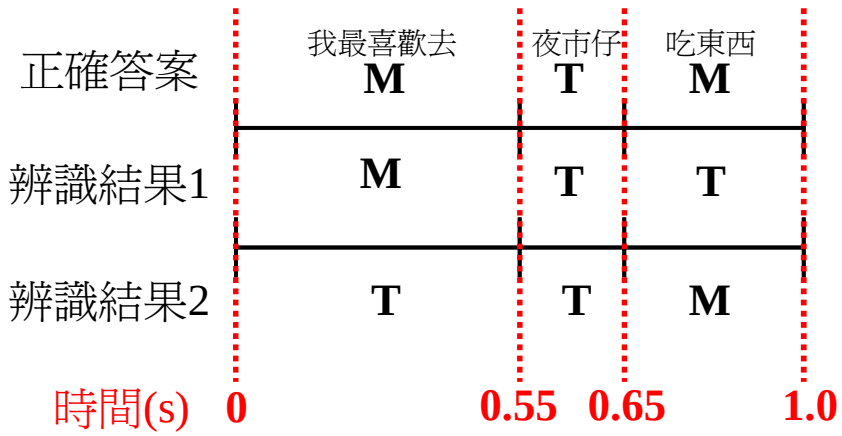
圖六、以語言時間資訊爲評估單位

同樣的我們計算其精確率、召回率與 F 測度，在這個方法下精確度爲在所有語言判別結果的語言標籤中，是正確語言標籤者的時間長度占辨識結果總時間長度的比率。而

回現率也變更爲在所有正確的語言標籤中，有被辨識出且爲正確語言標籤者的時間長度占正確答案總時間長度的比率。因此，若當每個語言區間時間長度相同時，其實會退化回成以語言標籤爲評估單位元同樣的方式。

$$precision = \frac{\sum_{i=1}^{N_H} \sigma_i}{\sum_{i=1}^{N_L} \sigma_i} = \frac{N_H * \sigma}{N_L * \sigma} = \frac{N_H}{N_L}$$

依語言時間資訊爲評估單位時，可以較清楚瞭解判別錯誤的部份對整句語音的影響程度。以圖七「我最喜歡去夜市仔吃東西」爲例，使用單一語言片段做計算單位，可分爲三個單一語言片段 MTM。若有二種不同的語言辨識結果，第一種辨識結果爲 TTM，而第二種辨識結果爲 MTT 時，以語言標籤爲單位的正確率計算方式上，兩種結果皆爲三個語言標籤中有一個語言標籤爲錯誤，所得到的正確率會相同，皆爲約 67%；若以語言標籤時間資訊爲單位時，第一種結果錯誤的語言長度爲 0.35 秒，而第二種結果錯誤的語言長度爲 0.55 秒，此時便可分出第二種辨識結果其實較第一種辨識結果好。



圖七、兩種評估方法比較範例

表三、以語言時間資訊爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段爲語言計算單位		
	P	**A**	**F**	**P**	**A**	**F**
LIAM+Free Syllable	75.3%	68.68%	71.84%	60.2%	52.78%	56.25%
LDAM+Free Syllable	77.1%	68.87%	72.75%	68.8%	58.05%	62.97%
LIAM+Syllable Bi-gram	84.1%	70.76%	76.8%	79%	60.57%	68.57%
LDAM+Syllable Bi-gram	84.6%	71.76%	**77.3%**	79.1%	61.67%	**69.31%**

表三爲我們以在以語言時間資訊爲評估單位所得到的結果。將兩種不同評估單位的結果做相互對應，在以音節爲計算單位時，使用時間資訊爲評估單位的情況下能得到的正確率 81.4%，而以語言標籤爲評估單位的方式爲 73%，故可估測在語言辨識錯誤的音節，可能多數爲持續時間較短的音節。

在以合併的語言區塊爲計算單位時，我們可以發現使用語言標籤來做爲評估單位的情況下可得到的語言辨識結果約 80.7%。但在使用語言時間資訊來做爲評估單位的方式中，我們可以發現以語言區塊爲計算單位的情況可得到約 74.8%的語言辨識率。這表示我們辨識正確的語言片段佔所有的語言片段約八成左右，但正確語言的時間佔總體時間僅約七成，故可估測在爲正確語言標籤的語言區段時間中，含有部份的錯誤時間段。

（三）、**Word** 與 **Syllable** 混合之 **Bi-gram**

在我們的假設中，混合語言語音中語言轉換可能較常發生於 word 上，故若能得知 word 與 word 間之相應機率，對混合語言語音的語言辨識應能有所幫助。但由於在各種語言中，word 的數量皆相當多，若想以 word 爲單位訓練出其 Bi-gram 機率，則所需的訓練語料資料量將會相當龐大。因此我們折衷計算部份較常出現的 word 與 syllable 間的 Bi-gram 機率。由於我們既有的已標音完成的語料庫資料不足，故我們需要重新收集並製作新的語料庫。首先，收集大量文章文字，我們由新聞稿、鄉土文學、散文、生活小品等文章中，收集結成約 10 萬句總字數約 100 萬字的文章，但這些文章皆無標音。接著，我們將這些文章中的文字做標音，由於我們期望能得到類似 code-switching 此類較貼近我們的主題的文句，故我們先以約 2 萬較常用台語詞（僅雙字詞以上）爲主，將有出現在文章中的詞均先標記爲台語的標音，其餘尚未標音的文字再以同樣方式，以約 7 萬華語字詞爲做標音。

在整個標音過程中，真正有使用到的 word 個數約 22731；其中華語字詞數爲 15881 個，台語字詞數爲 6765 個，而華語與台語共同字詞數 85 個。另外，在標音的同時，我們計算所使用到的每個詞所出現的次數，並重新分別製作新的華台語字詞出現頻率表，以做爲之後語言共同字詞處理時使用。

而在標音完成後便可將文章轉爲一個華台語混合的音標序列，並使用完成標音的混合語言音標序列來做爲 word 與 syllable 混合 bi-gram 語言模型的訓練語料。最後將系統原本的 syllable Bi-gram 語言模型替換成新制作的 word 與 syllable 混合 Bi-gram 語言模型，再以與之前相同的方法，進行語言辨認的實驗。

我們在混合語言語音實驗中，以音節 Bi-gram 的語言模型目前可得到約 74.8%的辨識率。在我們假設中，詞可能爲混合語言語音轉換的單位，且猜測 word-base 的語言模型應能具有更佳的語言鑑別率，因此我們製作了以詞與字節混合 Bi-gram 語言模型，取代原本的音節 Bi-gram 語言模型，並進行語言辨識實驗。在正確率計算上，一樣使用兩種計算單位和兩種評估方式評測整體的語言辨識率。最後以目前擁有最佳語言辨認結果的音節 Bi-gram 架構做爲新語言模型的實驗對照組，比較詞與音節的 Bi-gram 語言模型對於語言辨識上是否有益，而實驗的結果如表四與表五所示。

表四、以語言標籤爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段 爲語言計算單位		
	P	**R**	**F**	**P**	**R**	**F**
LDAM+Syllable Bi-gram	78.1%	68.5%	73.0%	82.4%	79.0%	80.7%
LDAM+Word&Syl Bi-gram	86.4%	63.3%	**73.0%**	88.4%	78.9%	**83.4%**

表五、以語言時間資訊爲評估單位正確率

	以 **Syllable** 爲語言計算單位			以單一語言片段 爲語言計算單位		
	P	**R**	**F**	**P**	**R**	**F**
LDAM+Syllable Bi-gram	78.8%	84.4%	81.5%	69.2%	79.0%	73.8%
LDAM+Word&Syl Bi-gram	78.2%	86.7%	**82.2%**	72.8%	84.1%	**78.0%**

由實驗統計可發現，無論以語言標籤爲評估單位或以語言時間資訊爲評估單位，詞與音節混合的 Bi-gram 方法所得到的結果皆比音節 Bi-gram 進步了約 3%左右。由此可知，以 word-base 的 Bi-gram 語言模型的確比音節 Bi-gram 的語言模型更具有語言鑑別能力。

五、結論

在我們的語言辨認系統中，由於是基於自動語音辨識所得到的音節序列爲基礎，音節序列的正確與否將會影響後續語言辨識的正確性。故我們使用了不同的聲學模型和語言模型做爲音節辨識上正確率提升的嘗試。由實驗結果得知，在特徵參數擷取中，直接使用 39 維 MFCC 在我們的系統能有較好的語音識別表現。在聲學模型上，以語言獨立的聲學模型能得到較好的結果，而在語言模型上，使用 Bi-gram 可表現出語言的規律性，有助於語言的鑑別，且以詞與音節混合 Bi-gram 又可比使用音節 Bi-gram 得到更高的語言鑑別度。而在評估方面，以語言時間資訊爲評估方式在語言辨識正確率的計算上，比以語言標籤爲單位的評估方式多考慮了時間因素，可得到更加精確的語言辨識率，使得評估結果上可更加客觀有力。

參考文獻

[1] 林奇嶽、王小川,”自動語言辨認簡介”,計算語言學通訊第十七卷第四期,2006

[2] Haizhou Li,Bin Ma,Chin-Hui Lee,”A Vector Space Modeling Approach to Spoken Language Identification”,IEEE TRANSACTIONS ON AUDIO,SPEECH,AND LANGUAGE PROCESSING,VOL.15,NO.1,JANUARY 2007,P271-P284

[3] Sheng Gao,Bin Ma,Haizhou Li,Chin-Hui Lee,”A Text Categorization Approach to

Automatic Language Identification",INTERSPEECH 2005

[4] Chi-Jiun Shia, Yu-Hsien Chiu, Jia-Hsieh and Chung-Hsien Wu,"Language Boundary Detection and Identification of Mixed-Language Speech Based on MAP Estimation",ICASSP 2004

[5] Chung-Hsien Wu, Senior Member, IEEE, Yu-Hsien Chiu, Chi-Jiun Shia, and Chun-Yu Lin,"Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs",IEEE Transactions on audio, speech, and language processing, Vol.14, No.1, January 2006, p266-p276

[6] 林俊憲，＜應用隱含式語意索引與語言模型於中英夾雜語音之語言鑑別＞，國立成功大學，碩士論文，民國 91 年

[7] Joyce Y. C. Chan, P.C. Ching, Tan Lee and Helen M. Meng,"Detection of Language Boundary in Code-switching utterances by Bi-phone Probabilities", ISCSLP, 2004

[8] Joyce Y.C. Chan, P. C. Ching and Tan Lee,"Development of a Cantonese-English Code-mixing Speech Corpus", in Proc. of Eurospeech 2005,pp.1533-1536, Lisbon,2005

[9] Joyce Y.C. Chan, P.C. Ching, Tan Lee and Houwei Cao ,"Automatic speech recognition of Cantonese-English code-mixing utterance",INTERSPEECH 2006

[10] Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain, Member, IEEE,"Multistage Speaker Diarization of Broadcast News", IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, 2006

[11] Dau-Cheng Lyu, Ren-Yuan Lyu, Yuang-chin Chiang and Chun-Nan Hsu,"Language Identification by Using Syllable-based Duration Classification on Code-switching Speech", In Proceedings of Lecture Notes in Artificial Intelligence, Kent Ridge, 2006