一個結合 SVM 與 Eigen-MLLR 新的多語者線上調適架構應用於泛在語音辨識系統

A New On-Line Multi-Speaker Adaptation Architecture Combining SVM with Eigen-MLLR for Ubiquitous Speech Recognition System

施伯宜　Po-Yi Shih
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
hanamigi@gmail.com

林苑寧　Yuan-Ning Lin
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
yukinaco@hotmail.com

王駿發　Jhing-Fa Wang
國立成功大學電機工程學系
Department of Electrical Engineering
National Cheng Kung University
wangjf@mail.ncku.edu.tw

摘要

本論文提出了一個以結合 SVM 和 Eigen-MLLR 爲基礎的線上多語者調適架構，應用於 ubiquitous 環境的語音辨識系統。語者獨立式的辨識系統相較於傳統的辨識系統有著更好的辨識效果，而語者調適方法便是其關鍵所在。本論文應用 SVM 和 Eigen-MLLR 的特性作爲調適技術的基礎，對於每個訓練語者的個別訓練語料做分類以及建立特徵參數向量空間。在語音辨識時，使用 SVM 找出測試語者所屬的類別，再找出類別相對應的 MLLR 特徵參數矩陣，並將其與非語者獨立模型結合成語者獨立模型。最後再利用辨識結果與原本的 MLLR matrix 和 Eigenspace 採取比重運算，並將運算結果更新原本的 MLLR matrix。相較於非語者獨立的辨識系統可以增加了 5~8%的辨識率。

Abstract

This work presents a novel architecture using SVM and Eigen-MLLR for rapid on-line multi-speaker adaptation in ubiquitous speech recognition. The recognition performance in speaker independent system is better than in conventional speaker dependence system, and the key point is speaker adaptation techniques. The adaptation approach is on the basis of

combine SVM and Eigen-MLLR, generating a classification model and building parameters vector-space for all speakers' individual training data. While in recognition, to find test speaker classification by SVM and look for MLLR parameters matrix correspond to speaker classification, then the MLLR parameters matrix and original acoustic model will integrate into speaker dependent model. Last, we estimate the adapted MLLR transformation matrix set by weighting function with recognition result, the present MLLR matrix, and Eigenspace. The estimate result will be used to update the MLLR matrices in adaptation phase. The experimental results show that the proposed method can improve 5% to 8% speech recognition accuracy with speaker adaptation.

關鍵詞：ubiquitous，語者調適，SVM，MLLR

Keywords: ubiquitous, speaker adaptation, SVM, MLLR,

一、緒論

（一）、研究動機

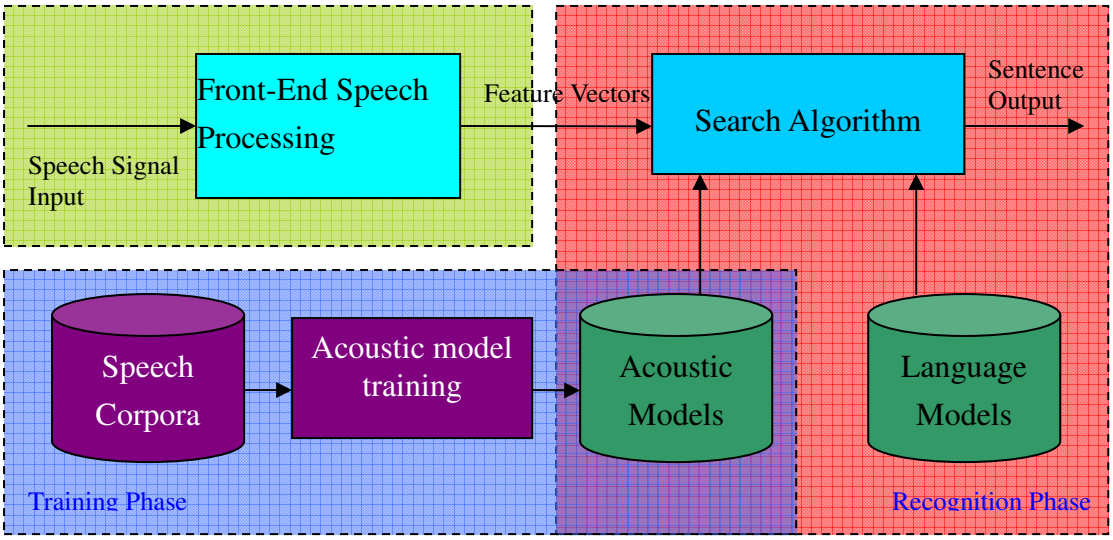
科技的發展始終來自於人類的需求，而以人爲本的生活應用科技一直是各方面研究的重點所在。在人的生活中，溝通，是人類這個社會生活中做重要的環節。人跟人之間的溝通模式之一，語言，創造出許多的生活價值，拉近人與人之間的距離，但在這個電腦化時代，資訊爆炸時期，語言是否可以在人跟電腦之間的溝通當中創造出另一種價值，因此語音科技便爲此提供有效的解決方法。

語音技術的發展已經有數十年，且累積了相當豐富的成果，對於這個時代的進步以及人們的需求也有相當的貢獻。在現有的語音技術中，能夠扮演人跟機器之間互動角色當屬語音辨識技術。語音辨識的研究當中，可以分爲聲音處理以及語言處理兩大方面，在聲音處理的研究上，語音模型是一切的根本，也可說是整個辨識系統的靈魂，聲音處理研究方面的重點之一。此外，語音辨識系統在應用方面也會碰到所謂強健性（robustness）的問題，一個系統如何在一般使用下，去學習適應環境以及新使用者的語音特性而來提升對此語者的語音辨識能力。而在這個課題中，語者調適（speaker adaptation）技術的研究變成爲一個重點課題。在傳統的語音辨識系統使用中，都是處於單一環境下和單一輸入，也就是只可以在特定地點使用辨識系統，但這對於現在的人類生活模式來說，這種模式已經無法滿足人們的生活需求。在這個 e 化的社會當中，所謂的數位生活（digital life）已經充斥在我們的生活之中，特別是在家庭中的數位生活應用，更希望可以隨時隨處都使用語音來控制、操作家中的各種電器生活用品，於是便發展出泛在語音辨識技術（Ubiquitous Speech Recognition Techniques），讓人們在家中的任何一個地方都可以透過語音辨識技術的應用來享受科技所帶來的生活便利。而語音辨識系統面對家中成員的不同，便更需要一套適合的且可以即時更新的多語者調適方法，讓語音辨識系統對家中每個成員的語音都有良好的辨識能力。

（二）、研究方向

在語音辨識系統中，如圖一所示，需要一套聲學模型（acoustic model）來模擬各式的語音特性，而這語音特性同時也包含了語者的特性。每個聲音模型初期則是使用大量的訓

練語料（training corpus）經由統計的方式建立而成。爲了避免聲學模型在使用時環境因素跟訓練時的環境因素不同而造成便是效能大打折扣，如背景雜訊，所以便盡量收集許多不同環境因素下的語料，以維持聲音模型應有的準確性。但對聲音模型影響最大的並非只有環境因素，要隨時面對不同語者的語音訊號也是一大挑戰。在語音訊號中，因爲每個語者天生的物理特性差異，如共振腔發聲習慣、說話腔調，所以這也歸類爲一種環境因素。一個基本的聲音模型，便是透過多位訓練語者提供的語料經由統計的方式而建立的語者不特定模型（speaker independent model, SI model），對於每個語者的訊號模擬程度只能算是中等，不可能各方面適合每個人的語音特性。所以爲了對不同語者都提供有良好的辨識效能，系統必須配合語者所提供的語料來做適當的修正、調整，這個工作便稱爲語者調適（speaker adaptation）。在使用語者提供的語料對原本的聲音模型來做加以修正之後，這個聲音模型便稱爲語者特定模型（speaker dependent model, SD model）。但這又便會碰到許多延伸的問題：如何在短暫的時間內以及語者提供少數的語料下可以快速的修正聲音模型，使辨識系統辨識能力提升也達到好的服務品質，這便是目前語者調適技術需要克服的一個問題。



圖一、語音辨識系統流程圖

在許多語者調適研究當中，早期是以貝氏調適法（Bayesian adaptation）爲基礎，這個方法的的優點是可以將語音模型完全配合語者做很精確的調整，但缺點便是它對需求的語料相當的大，這並不能滿足語音辨識系統在快速調適上的要求；此外，若是在未知語料內容的非監督式調適（unsupervised adaptation ）的情形下，反而只會大幅降低整個模型的精確度。近年來最熱門的語者調適方法是使用最大相似度線性迴歸（Maximum Likelihood Linear Regression, MLLR），許多的研究實驗結果也都證實了此方法可以在快速調適以及非監督式調適上有良好的表現。然而在 MLLR 中，需要估測大量的參數，然而如果在語料稀少的情況下，有時也會發生將聲音模型調整的更差。在 [1] 所提出的使用 Eigenspace-MLLR 爲基礎的快速調適法中，便考慮到這方面的兩個特點，不僅可以在語料甚少的情形下保有參數估測值的可靠性，並且在非監督式調適的情形下表現良好。本論文便以此方法來加以延伸在新的語者調適架構中。SVM 系統近年來都被使用在解決關於分類（classification）、回歸（regression）、以及快速的偵測（novelty detection）這些相關的問題上，且應用層面廣泛，不只在關於機器學習（machine learning）上，也有許多應用於語者鑑別（speaker verification）。在 [2] [3] 的研究便使用以 MLLR 爲基

礎核心的 SVM 來實現語者辨識與鑑別。在本論文中便以 SVM 爲來做基礎的語者分類，使得可以找出在 MLLR 中相對應的語者估測參數，並且可以更新原本的語者估測參數。

（三）、研究主題與主要成果

本篇論文的主題主要是著重在兩大方向上，第一個是如何將 SVM 與 Eigen-MLLR 兩個不同性質的演算法結合使用，亦即如何將所有調適語者在有限的少量調適語料下同時使用 SVM 對語者做分類，以及 Eigen-MLLR 建立所有語料的特徵向量參數空間，並且將 SVM 中的個別分類與 Eigen-MLLR 中的特徵向量參數空間建立一個相對應的關係，這稱爲訓練階段（Training phase）。之後在每一位測試語者使用時，便會先透過 SVM 將語料找出相近的分類，並從分類中找出相對應的 MLLR 特徵參數，在使用此特徵參數與原本的聲音模型結合成語者特定模型，再進行語音辨識，這稱爲辨識階段（Recognition Phase）。第二個便是結合辨識結果，語者相對應的 MLLR 特徵參數以及特徵參數估測三者來使用比重取決，並且將結果來即時更新原本語者相對應的 MLLR 特徵參數，這稱爲調適階段（Adaptation phase）。根據實驗，我們結合使用了 SVM 以及 Eigen-MLLR 的方法，在對於泛在的語音辨識中辨識率提升確實有加成的作用，在整體效能上又向上提升。

在第二章中主要是介紹本論文中所主要使用的調適方法以及語者調適技術在實行及 SVM 的技術和應用簡介。第三章便會詳細的介紹研究主題的架構以及實行情況，第四張便會繼續介紹整個實驗環境和過程以及結果，並且會加以探討。第五章則是本論文報告的結論以及未來的相關研究遠景。

二、相關技術回顧

在這一章節將對語者調適的實行以及性質上作簡介，以及對於 SVM 和 MLLR 演算法做概要的敘述。語者調適法的角度可以從系統、語料以及本質等三方面探討，如表一所列。SVM 系統近年來都被廣泛的使用在分類（classification）、回歸（regression）、以及快速的偵測（novelty detection）相關的問題上。SVM [4] [7] 的所有相關技術也可以視爲是使用在分類和回歸的監督式學習演算法。MLLR 的基本意義是在於假設調適後的參數群組與現有的基準參數群組存在著一個線性迴歸函數的關係，而可以藉由使用最大相似度（Maximum Likelihood, ML）測法來求得現有參數群組間的函數關係。

（一）、語者調適的簡介

語者調適技術的目的在於利用語者所提供的有限語料，來改善辨識系統對於與使用者的辨識能力。表一爲現有調適法類別的分類。從系統的角度來看，當系統要進行語者調式工作時，須先獲得語者提供的語料，此稱爲訓練語料（training data）或是調適語料（adaptation data）。假如系統是以每收到一句語料便調適一次的話，稱爲循序調適法（sequential adaptation）；如果是一次獲得所有的語料再做調適的話，便稱爲批次調適法（batch adaptation）。如果從調適語料的角度來看，如果系統事先得知語料的內容，也就是清楚知道語料每一句的標音，系統及可以找出語音訊號和相對應的語音參數做精確的調整，這稱爲監督式調適法（supervised adaptation）；相反的，若不知道語料的內容，須先對語料辨認過後，才將辨認結果當作語料的內容來調適，則稱爲非監督式調適法

（unsupervised adaptation）。

辨識系統中原有的聲音模型（acoustic model）可以稱爲語者不特定模型（speaker independent model, SI model），又稱爲初始模型（initial model）。若以模型爲基礎的調適方式，則須先使用初始模型對語音訊號進行分析（切音），並找出每個音框（frame）所相對應最有可能的模型狀態，甚至還必須先進行一次辨認，因此，在辨識階段的搜尋演算法（Search Algorithm，如 Viterbi）中所獲得的結果也會與初始模型有相當大的關係。所以調適演算法的精確度也跟初使模型的切音有相當大的關係。

若回到語者調適的本質來看，又可分爲兩種基礎調適法：一是利用語音訊號的特徵向量爲基礎的調適法（feature-based adaptation），主要是以調整語音訊號的特徵向量，使得現有模型參數更能精確描述變化情形；一是調整模型的參數爲基礎的的調適法（model-based adaptation），使得可以更有效模擬語者的特性。近來相當多的研究顯示，以模型爲基礎的方法優於以特徵爲基礎的調適法，其在實作上的要求與系統複雜度也較爲簡單，而許多的熱門調適技術都是採用此性質的調適方式，如最大相似度線性迴歸（MLLR）、貝氏調適法（又稱爲最大事後機率法，Maximum a Posterior，MAP）。

表一、語者調適法的類別

語者調適法分類	調適法類別
以系統分類	循序調適法（sequential adaptation），批次調適法（batch adaptation）。
以語料分類	監督式調適法（supervised adaptation），非監督式調適法（unsupervised adaptation）。
以本質分類	特徵向量爲基礎的調適法（feature-based adaptation），模型參數爲基礎的調適法（model-based adaptation）

（二）、支援向量機（Support Vector Machine, SVM）

支援向量機（SVM）是目前經常使用來做爲分類（classification）或回歸（regression）的方法。當給予一群已經分類好的資料之後，支援向量機可以經由訓練（training）獲得一組模型（model）。之後，若有未分類的資料加入時，支援向量機便可以依據先前訓練出的模型去做預測（predict），並決定這筆資料所屬的分類。而因爲在建立起模型時，必須要有已經分類好的資料做爲訓練，所以是屬於監督式學習（supervised learing）的方法。支援向量機也是一種線性分類（Linear Classification）的方法，目的在於找出一個超平面（hyperplane）而可以將在特徵空間中（feature space）已經分類好的資料清楚的分開成不同的類別。

支援向量機最重要的特性便是確定模型參數符合最佳化的問題，特別是能把區域性的最佳化當成全域性的最佳化，因爲 SVM 使用 Lagrange multipliers 來探討延伸整個系統的最佳化問題。在參考文獻 [4] 表示，SVM 的基礎是一個利用核心函數（Kernel function）$K(\cdot,\cdot)$的總和來建立一個兩類別的分類器，其基本數學式如下：

$$f(\mathrm{x}) = \sum_{i=1}^{L} \gamma_i t_i K(\mathrm{x}, \mathrm{x}_i) + \xi \qquad (1)$$

t_i 表示爲理想的輸出值，$\sum_{i=1}^{L} \gamma_i t_i = 0$ ，並且 $\gamma_i > 0$ 。向量群 Xi 爲藉由最佳化處理從訓練集合中獲得的支援向量群（Support Vectors）。理想的輸出值藉著相關的支援向量是落在類別 1（其值爲＋1）或類別 2（其值爲－1）。對於分類來說，數值 $f(x)$所屬類別的決定是在於所定的標準（threshold）之上或之下。

（三）、特徵式最大相似度線性迴歸（Eigen-Maximum Likelihood Linear Regression, Eigen-MLLR）

首先我們從原始的 MLLR 演算法來看 [3] [5]，此方法的原理背後是透過假設改變後的參數會和原本的基本參數間唯一線性迴歸的函數關係：如下

$$Y = AX + B \tag{2}$$

若將此原理使用在語者調適中，語料的每一個特徵向量接代表著語者與因空間中的其中一個樣本，而我們也假設了待求參數和現有參數之間的函數關係，因此可以便可以使用最大相似度（Maximum Likelihood, ML）估測法來求得 A、B 的值，如下所示：

$$\begin{aligned} A &= \arg\max_{A} f(x \mid \theta, A, B) \\ B &= \arg\max_{B} f(x \mid \theta, A, B) \end{aligned} \tag{3}$$

在這裡 θ 指的是所有語音參數模型的集合，x 則表示調適語料的觀測值。

在許多的語者調適法當中，最大相似度線性迴歸的方法被廣泛的應用在快速語者調適，例如它只需要些許的語料便可以對模型參數做調適。在最大相似度線性迴歸中，非語者獨立的模型參數可以根據一個或多個仿射轉換方式（affine transformations）來達到調適的目的。最大相似度線性迴歸調適法是使用仿射轉換方式來調適高斯混合模型（Gaussian mixture model, GMM）中所有混合元件的平均值（mean），而同一個仿射轉換法則可以提供所有的混合元件共享，表示法如下：

$$\hat{\mu}_i = \mathrm{A}\mu_i + \mathrm{b} \quad \forall i \tag{4}$$

μ_i 表示在 GMM 中還沒被調適過的平均值，而 $\hat{\mu}_i$ 表示已經調適過的平均值。

在眾多的資料量當中，混合的元件可以被歸類成多個類別，且不同的仿射轉換方式可以在不同的類別中被共用，延伸(4)的表示法：

$$\hat{\mu}_i = \mathrm{A}_1\mu_i + \mathrm{b}_1 \quad \forall \hat{\mu}_i \in \mathrm{class}_{1,} \tag{5}$$

$$\hat{\mu}_i = \mathrm{A}_2\mu_i + \mathrm{b}_2 \quad \forall \hat{\mu}_i \in \mathrm{class}_{2,} \tag{6}$$

在單一和多個類別的情況下的轉換方式是透過選擇最大的相似度來決定的。由於在這個調適方法中是利用到語料共享的觀念，當有一模型在沒有任何的調適語料情形下也可以藉助同一類別中有語料的模型來求出 A, b 值，所以在每次調適時，只要選到適當的類別，則所有的模型參數都可以調整到，這也是最大相似度線性迴歸法可以應用於快速調適語者的主要原因。

特徵式-MLLR [8] [10]爲一改良式的 MLLR，其目的是以特徵向量空間計算 MLLR 回歸矩陣（Eigen-MLLR），主要是採取了以向量空間爲基礎之語者調適技術以及傳統 MLLR

的優點。向量空間爲基礎之語者調適技術的優點，就是將所有訓練語者的聲學參數向量化，並利用這些向量撐出一片可信賴的聲學向量空間。此後，在執行語者調適時，所做的工作便只是找出測試者在這聲學空間上的最可能的位置，因此所處的位置一旦被決定，測試者最後整套的聲學參數也就呼之欲出了。其向量空間的建立步驟如下：

Step1.模型參數向量化：首先，我們先將每位訓練語者的語者特定模型參數向量化，即是對於整套 SD model 裡所有高斯混合的平均值向量，像火車一節一節地串接起來。如此，我們便可得到一個高維度的向量，亦即若高斯混合的平均值向量之維度爲 n，一套 SD 模型中有 M 個高斯混合，則此新向量的維度 D 便是 $M \times n$。

Step2.將模型參數向量組成矩陣：接下來，有了這些由所有 SD 模型參數所組成的向量後，扣除向量中各維度的平均值，隨後再將這些向量以 row vector 的姿態組成一個矩陣 Z，若語者的個數爲 K，則 Z 的大小即爲 $K \times D$。

Step3.計算空間基底 最後再計算相關矩陣 $Z^T \cdot Z$(correlation matrix ,Z^T 爲 Z 的轉置矩陣)的特徵向量(eigenvector ,維度亦等於 D)後便大功告成。這些求得的特徵向量即做爲向量空間的基底(basis)，在語者調適中成爲一具代表的特徵，所以求得的空間便稱做特徵向量空間(Eigenspace)。

當 Eigen-MLLR 在實現時，建構特徵向量空間的取材已由模型參數改爲 MLLR 回歸矩陣(還包含偏移向量 b)，所以我們可以從中發現，除了矩陣 A 或著是向量 b 中的參數各自有不同的意義外，兩者的參數在物理意義上更是截然不同。因此在這情形下，所有相異的參數在事前做正規化的處理就愈顯得格外的需要。Eigen-MLLR 相較於傳統的方式，在有適當的訓練語料下，是更加提高了矩陣估測值的準確性。

三、所提出的演算法架構

在本段中我們將會介紹一些實驗背景所採用的方法及設定，並且敘述架構中每個部份。

（一）、實驗相關背景

在建立語音辨識系統的基本聲學模型，又稱爲語者不特定模型或是初始模型，我們是使用台灣口音中文語料庫(Mandarin Across Taiwan, MAT-400)來做爲訓練語料庫，裡面包含了約 400 個語者所錄製的共 5,000 個不同的語音檔，以及 77,324 個詞語和 5,353 個句子。其詳細內容如下表：

表二、台灣口音中文語料庫簡介

Databases	Number of Files	Prompting Item Numbers	Speaking Style	Description
MATDB-1	3600	1-9	spontaneous	Short answering statements
MATDB-2	2000	10-14	read	Numbers pronounced in five different ways
MATDB-3	4800	15-26	read	Mandarin syllables
MATDB-4	12000	27-56	read	Words of 2 to 4 syllables
MATDB-5	4000	57-66	read	Phonetically balanced sentences

語音訊號的特徵參數中採用 13 維的梅爾倒頻譜系數(MFCC)及一階和二階的回歸係數 Δ (delta)，設定如表三：

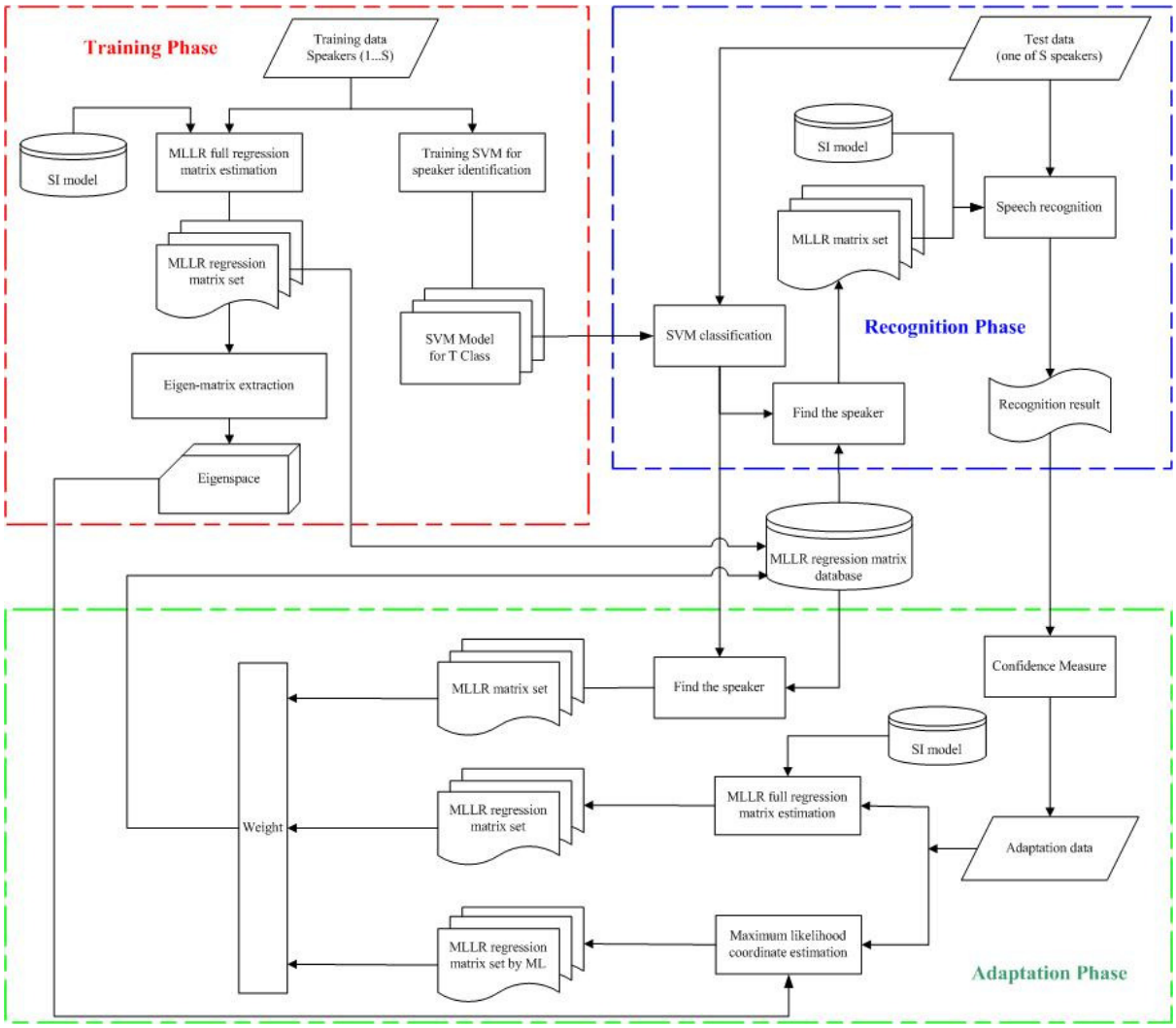
表三、本論文中所使用的參數擷取設定

取樣頻率	8 kHz
預強濾波器	$1\text{-}0.97z^{-1}$
分析視窗	Hamming Window
視窗長度	20ms
音框平移	10ms
梅爾濾波器個數	23
特徵向量參數	13 MFCC + Δ + ΔΔ & log energy + Δ + ΔΔ

在聲學模型架構中，則是採用最通用的連續密度馬可夫模型(continues density hidden Markov models, CDHMM)，採用由左至右(left-to-right)的型態，也就是狀態轉移上只允許從抹一狀態跳至鄰近的下一狀態或是停留在原本的狀態上。而在模型單位的選取，則採用音節右相關聯音素模型，每個音素模型包含 5 個狀態，3 個音素、1 個靜音、1 個短暫停。整個聲學模型是利用劍橋大學所提供的 HTK [6] 工具來建立。

（二）、演算法架構

整個演算法的架構可分爲三個階段，訓練階段（Training Phase），辨識階段（Recognition Phase）和調適階段（Adaptation Phase）。當系統一開始預設時，會先使用訓練語料來建立初始的模型，當建立之後，便會在每次作辨識的時候再針對個別語者的語音模型做調適。整個流程的框架結構如下圖：



圖二、語者調適演算法架構圖

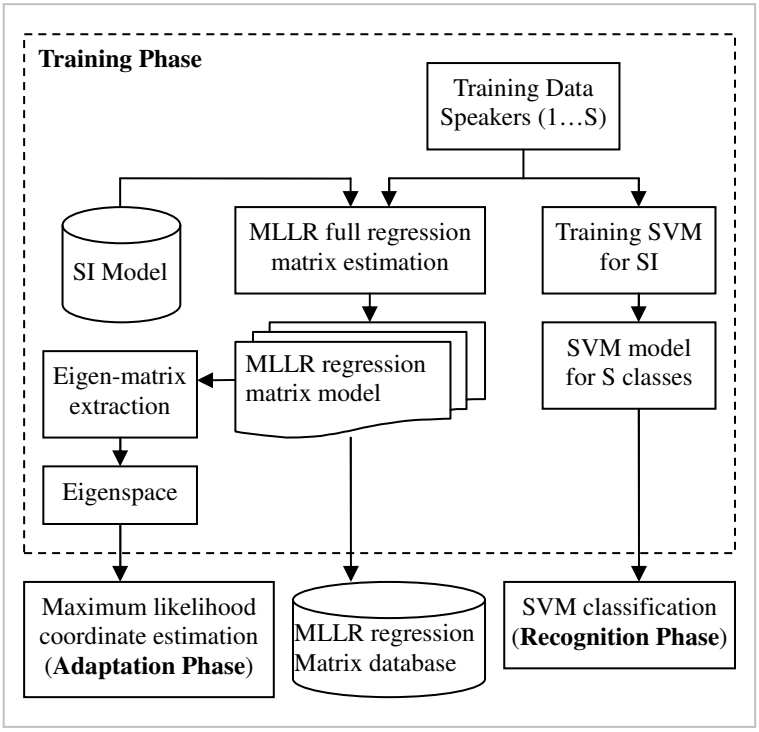
我們在這針對個別的階段來做詳細介紹。

訓練階段（Training Phase）：

在訓練階段，首先必須讓所有的測試語者都說一定量的語句來當做調適訓練語料，使用的方式是監督式調適學習。假設共有 S 個訓練語者，對每一個語者必須使用訓練語料和非語者獨立模型參數計算總共 C 個的傳統最大相似度線性迴歸完全回歸矩陣（MLLR full regression matrices），也對每一位語者用其全部的語料做一次最大相似度線性迴歸的調適計算以取得回歸矩陣，並且在訓練語料中找出足夠可以描述語者特徵的數量。而對於每一個訓練語者來說，所有 C 個的最大相似度線性迴歸迴歸矩陣可以當成單一語者特有的矩陣集合。接著從 S 個語者特有的矩陣集合當中攫取出 S 個根本構成要素，這便稱爲特徵矩陣（eigen-matrices），最後我們便取 S 個特徵矩陣。在這個階段中，傳統最大相似度線性迴歸完全迴歸矩陣的計算和以特徵空間爲基礎迴歸矩陣估測是分開始實現的。另一方面，這階段同時將調適訓練語料使用支援向量機（SVM）來做語料分類，共有 S 個類別，並且在建立語料分類的同時也建立起跟迴歸矩陣之間的相關性，使得每一分類都有其相對應的特徵矩

陣，而這相對應關係會是在辨識時最重要的。

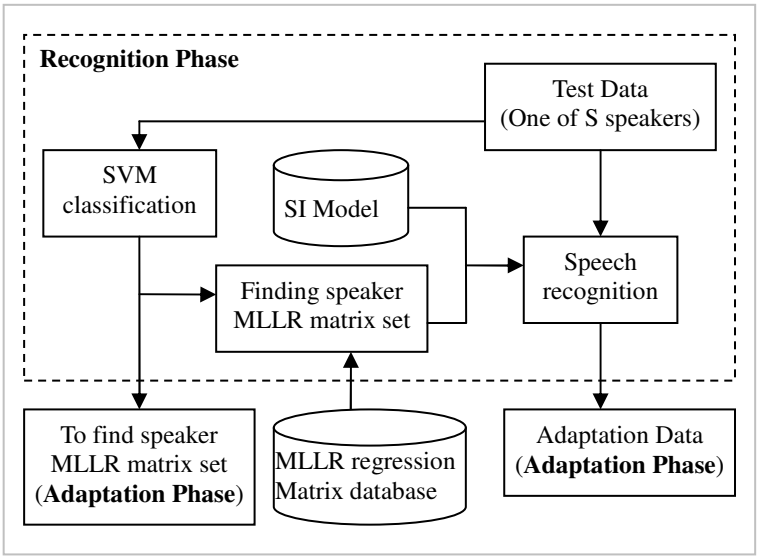
在有別於傳統的最大相似度線性回歸，我們是利用建立一個迴歸矩陣特徵空間來尋找特定語者的最大相似度線性回歸回歸矩陣，而不需要透過語者的語料求得迴歸矩陣中的每個參數，只需要利用語料找出回歸矩陣最有可能在特徵向量空間中的所在位置。也由於這個特徵空間是使用足夠的語料來估測得到的 MLLR 回歸矩陣，所以能替回歸矩陣估測提供相當足夠的資訊。如圖三所示。



圖三、訓練階段架構圖

辨識階段（Recognition Phase）：

圖四爲辨識階段的流程圖。在辨識階段中，從提供訓練語料的語者中挑選一個做測試，當語者說話經由語音特徵擷取後，便會傳送至 SVM 中做分類，若發現與第 n 個類別最相近，便從最大相似度線性回歸回歸矩陣集合中挑出相對應第 n 個迴歸矩陣，再將此迴歸矩陣跟原本的非語者特定模型結合成語者特定模型，再以這個模型進行語音辨識，然後輸出結果。這個辨識結果同時也提供給在調適階段中更新回歸模型的一個主要參考。



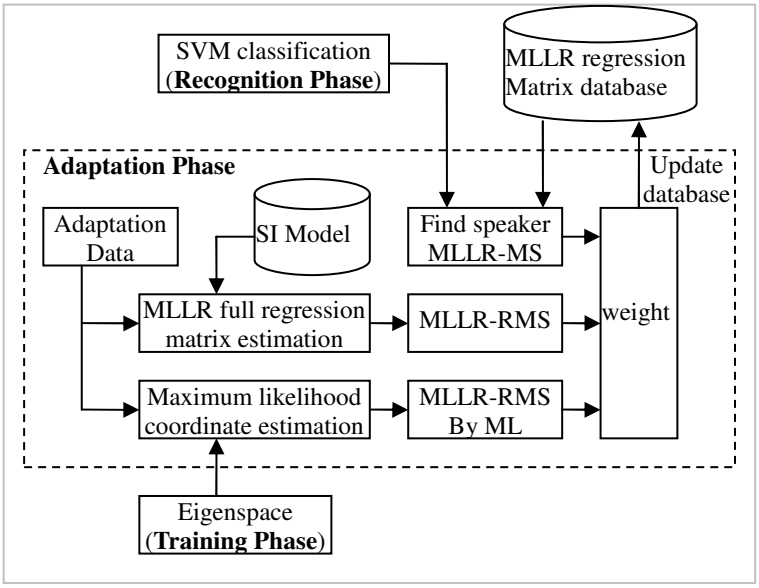
圖四、辨識階段架構圖

調適階段（Adaptation Phase）：

在經過辨識階段取得辨識結果後，會將辨識結果當成爲調適語料（adaptation data），對於每一個測試語者所提供的調適語料會使用最大相似度估測（Maximum Likelihood estimate, ML）方法來將語者定位在特徵空間中的回歸矩陣。最大相似度估測和最大相似度線性回歸會使用調適語料來個別做一個新的估測。個別估測後所獲得的結果將會和利用 SVM 所找出對應語者的 MLLR 回歸矩陣，三者做一個 weighting 的運算，並將獲得的結果對原本語者的 MLLR 回歸矩陣作更新。於是我們修改了 [10] 中的 Equation (5) 增加了一個調適信心度的估算。下列爲一個更新值的運算。

$$\hat{\mathrm{W}}_c = \xi_{conf} \cdot \left(\frac{\lambda \cdot \mathrm{W}_c^{\,EIGEN} + \sum_{m=1}^{M} \sum_{n=1}^{N_c} \gamma_n(m) \mathrm{W}_c}{\lambda + \sum_{m=1}^{S} \sum_{n=1}^{N} \gamma_n(m)} \right) + (1 - \xi_{conf}) \cdot \mathrm{W}_c^{\,present} \tag{7}$$

M 表示爲特徵向的數目，n 表示爲在 N_c 中的一個混合元件，$r_n(m)$表示在時間爲 t 時的觀測機率，$\mathrm{W}_c^{\,present}$，$\mathrm{W}_c^{\,EIGEN}$，$\mathrm{W}_c^{\,estimate}$ 則分別爲類別 c 目前的回歸矩陣，由 ML 估測出來的特徵空間矩陣，以及由 MLLR full regression 估測出來的回歸矩陣。$\hat{\mathrm{W}}_c$ 爲更新過後的回歸矩陣。ξ_{conf} 爲信心比重，信心比重則是依照辨識結果而得的。增加信心比重參數是避免當發生辨識結果產生錯誤而接受，也就是對的語音辨識成錯的，或錯的語音辨識成對的，來導致整個調適模型越來越差。整個調適過程如圖五。



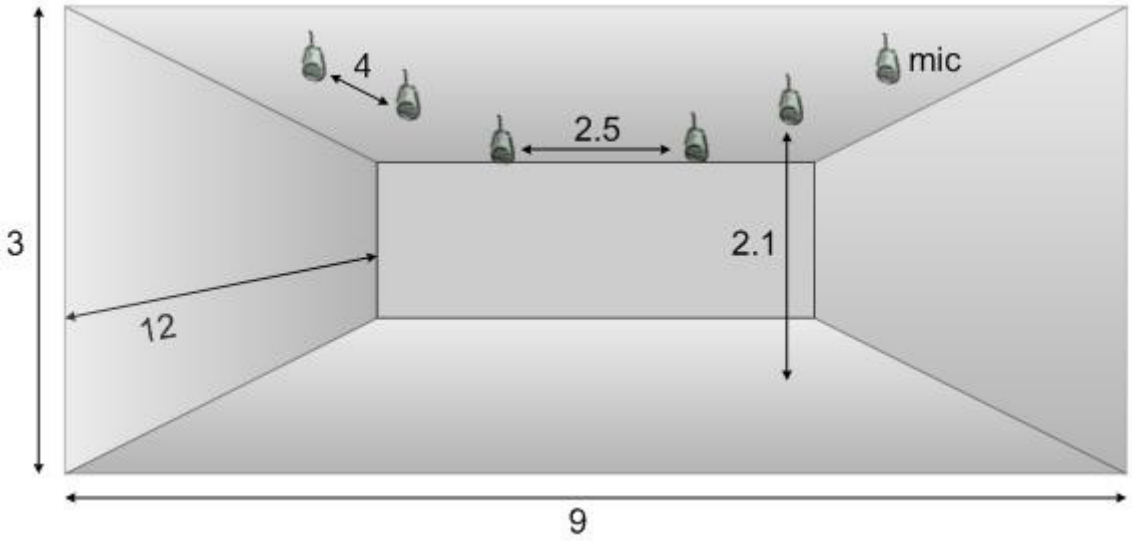
圖五、調適階段架構圖

四、實驗結果

我們實現實驗的結果是以台灣口音中文語料 MAT-400 爲基礎的語音模型。調適語料則是以日常生活用語大約 7~10 個字，而在測試時則以人名或是日常生活用語爲主。語音特徵的攫取設定則定義在表三。我們使用 MLLR 調適製造出擴張每一個語句的基底 SVM 特徵，對於高斯混合模型的每一個具體的類別都要做調適。我們便使用 HTK[6] 來建立上述的工作並且產生一遞迴式 MLLR 來建立轉換式。

對於已經提供訓練語料的目標語者，每句話都會用一個 SVM 特徵向量來代表，而選擇 SVM 的類別則會透過目標語者的特徵向量和原有的 SVM 特徵向量類別來作一比重選擇。

在我們的實驗中，使用了泛在語音辨識系統（ubiquitous speech recognition system）來測試，總共使用了 6 支全向性麥克風佈在房間的天花板，其收音範圍可以涵蓋了整個房間內部，如圖六。由於其排列方式並非傳統的麥克風陣列方式，無法使用傳統的雜訊消除法來達到較好的效果，所以便將 6 支麥克風使用多通道混音（multi-channel mixer）的方式成單一輸入，再使用子空間式語音增強演算法 [9]（Subspace Speech Enhancement, Using SNR and Auditory Masking Aware Technique）在語音訊號的前處理來消除語音雜訊，再開始取語音特徵。



圖六、泛在麥克風陣列示意圖

實驗過程總共使用了 10 個人爲提供訓練語料以及調適的語者，每個語者的訓練語句都爲 15 句。所採取的辨識率計算以正確率爲主，皆以百分比表示，計算方式如下：

$$\text{辨識正確率}=\frac{\text{辨識正確語句句數}}{\text{全部語句句數}}\times 100\% \tag{8}$$

表四爲在尚未有任何語者調適方法的泛在語音辨識系統下所測試的結果，可以發現初始聲音模型確實是對任何一個語者都提供相同的辨識準確效能，總平均準確率爲 85.7%。

表四、未有任何語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	130	124	126	133	129	126	127	131	134	125	1285
錯誤語句	20	26	24	17	21	24	23	19	16	25	215
辨識正確率(%)	86.6	82.6	84	88.7	86	84	84.7	87.3	89.3	83.3	85.7

接著我們以傳統的 MAP 語者調適方式來使用於系統，表五爲使用 MAP 語者調適後的結果。可以發現對於每個語者做調適之後，辨識系統對於每個人的準確率平均都有 2%~3%的準確率提升，總平均準確率爲 88.2%。

表五、使用 MAP 語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	135	126	131	136	134	129	132	135	137	128	1323
錯誤語句	15	24	19	14	16	21	18	15	13	22	177
辨識正確率(%)	90	84	87.3	90.7	89.3	86	88	90	91.3	85.3	88.2

表六則使用傳統 MLLR 語者調適法使用於辨識系統的辨識結果，相較於 MAP 調適法，有著更高的辨識率，會造成這樣的結果最有可能的原因是 15 句的調適語料對 MAP 來

說是非常的少量，而使得 MAP 真正的優點根本還來不及發揮，相較於 MLLR，反而能在少量語料時就顯示出相當出色的表現，比未調適的正確率多出 5%~8%，比 MAP 多出 3%~5%的辨識正確率。

表六、使用 MLLR 語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	139	136	137	142	142	135	142	141	142	138	1394
錯誤語句	11	14	13	8	8	15	8	9	8	12	106
辨識正確率(%)	92.6	90.6	91.3	94.6	94.6	90	94.6	94	94.6	92	92.9

表七的實驗結果則爲本論文所提出應用特徵式 MLLR 與 SVM 的語者調適架構。與 MAP 調適法來比較，由於藉著 MLLR 的特性關係，還是可以藉著少量語料而達到明顯的效果，且利用 SVM 來直接選擇相對應的特徵參數矩陣，少掉了重新由語料計算參數的運算量，也比傳統的 MLLR 平均提升了 1%左右辨識率，相較於未調適的辨識系統更提升了 8%平均辨識率，相較於 MAP 平均提升了 4%~5%辨識率。

表七、本論文所提出的語者調適法 150 句生活用語測試

測試語者	1	2	3	4	5	6	7	8	9	10	**總結果**
正確語句	142	137	139	144	141	138	140	144	145	138	1408
錯誤語句	8	13	11	6	9	12	10	6	5	12	92
辨識正確率(%)	94.7	91.3	92.7	96	94	92	93.3	96	96.7	92	93.9

五、結論

在語音辨識系統中，語者調適的工作是整體辨識效能的一個很重要的環節，而且對於在多人使用的環境下，語者調適技術在快速調適以及非監督式調適的情況下就更必須加強，這兩方面也會是今後語者調適技術發展的重點。一個調適技術不僅要能夠達成快速調適，且必須能夠在只有少量語料的情形下將現有原始的聲學模型調整至更適合當下語者的狀態，特別是對於在一個空間中的固定成員更是有所需要，如家庭成員。在本論文提出的架構中，利用特徵式最大相似度線性回歸（Eigen-MLLR）所建立的多語者特徵向量空間結合支援向量機（SVM）的分類來達成快速多語者調適。測試語者的語句經過 SVM 分類完畢之後，便會在 MLLR 回歸矩陣群中找出 SVM 類別相對應的回歸矩陣並且與初始模型結合成語者特定模型，再進行語音辨識。然後將語音辨識的結果利用 MLLR 回歸矩陣估測（MLLR regression matrix estimate）以及最大相似度估測（Maximum Likelihood estimate）三者來使用比重取決重新計算，並且將結果來即時更新測試語者相對應的 MLLR 回歸矩陣參數。在本論文中也發現，若可以再加強語音增強處理的演算法降低更多雜訊以及提升訊號強度，則對整個語者調適和辨識效能再進一步提升。在未來的語音辨識環境中，希望能夠增加更多的麥克風，使的能達到寬容度更高的泛在語音使用環境，也可以隨時加入新的語者讓系統可以自我更新以及作調適，也希望這項技術能結合其他的應用到更廣泛的層面，讓人們生活可以藉著數位化更便利。

參考文獻

[1] K. Chen et al, “Fast speaker adaptation using eigenspace-based maximum likelihood linear regression,” in Proc. ICSLP, Beijing, Oct. 2000.

[2] P. C. Woodland, “Speaker Adaptation: Techniques and Challenges”, *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, pp.85-90, 2000.

[3] Brian Kan-Wing Mak, and Roger Wend-Huu Hsiao, “Kernel Eigenspace-Based MLLR Adaptation”, *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, Mar 2007.

[4] Nello Cristianini and John Shawe-Taylor, *Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[5] M.J.F.Gales and P.C.Woodland, “Mean and variance adaptation within the MLLR framework”, *Computer Speech and Language*, vol. 10, no. 4, pp. 249-264, 1996.

[6] S.Young et al, “The HTK book,” http://htk.eng.cam.ac.uk.

[7] Zahi N. Karam and William M. Campbell, “A multi-class MLLR kernel for SVM speaker recognition,” in ICASSP 2008.

[8] Nick J.C.Wang, Wei-Ho Tsai and Lin-shan Lee, "Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification," 2001 European Conference on Speech Communication and Technology, Aalborg, Denmark, Sept 2001

[9] Wang Jia-Ching, Lee Hsiao-Ping, Wang Jhing-Fa, and Yang Chung-Hsien, “Critical Band Subspace-Based Speech Enhancement Using SNR and Auditory Masking Aware Technique”, IEICE Transactions on Information and Systems. vol 90; number 7, pages 1055-1062, July 2007

[10] K. T. Chen, W. W. Liau, H. M. Wang, and L. S. Lee, “Fast speaker adaptation using eigenspace-based maximum-likelihood linear regression,” in Proc. ICSLP, 2000, vol. 3, pp. 742–745.