行政院國家科學委員會專題研究計畫 期中進度報告

新一代語音科學與技術—由基礎到應用(2/3)
期中進度報告(精簡版)

計 畫 類 別 ： 整合型
計 畫 編 號 ： NSC 95-2218-E-002-027-
執 行 期 間 ： 95 年 06 月 01 日至 96 年 05 月 31 日
執 行 單 位 ： 國立臺灣大學電信工程學研究所

計 畫 主 持 人 ： 李琳山
共 同 主 持 人 ： 陳信宏、王小川、鄭秋豫、簡仁宗、簡立峰
王新民、王逸如、陳柏琳、洪志偉、廖元甫

處 理 方 式 ： 期中報告不提供公開查詢

中 華 民 國 96 年 05 月 31 日

行政院國家科學委員會專題研究計畫期中報告

新一代語音科學與技術-由基礎到應用（2/3）

New Generation Speech Science and Technologies-from Fundamentals to Applications (2/3)

計畫編號：NSC 94-2218-E-002-060-

執行期限：95 年6 月1 日至96 年9 月30 日

主持人：李琳山 國立台灣大學電信工程學研究所

鄭秋豫 中央研究院語言所

簡立峰 中央研究院

王新民 中央研究院

王小川 國立清華大學

陳信宏 國立交通大學

王逸如 國立交通大學

簡仁宗 國立成功大學

廖元甫 國立台北科技大學

陳柏琳 國立台灣師範大學

洪志偉 國立暨南國際大學

E-mail: lslee@gate.sinica.edu.tw, cytling@gate.sinica.edu.tw, lfchien@iis.sinica.edu.tw, whm@iis.sinica.edu.tw, hcwang@ee.nthu.edu.tw, schen@cc.nctu.edu.tw, yrwang@cc.nctu.edu.tw, jtchien@mail.ncku.edu.tw, yfliao@ntut.edu.tw, berlin@csie.ntnu.edu.tw, jwhung@ncnu.edu.tw

摘要

在本年度的計畫結案報告裡，我們分別敘述了本計畫在這一年來所進行的工作與諸多已完成的項目，主要包括了三個子計畫的具體研究成果、邀請日方學者來台參與跨國語音研討會、赴日本相關研究單位之參訪交流的內容、與日方共同合作研究所發表的論文、與計畫核心實驗室之發展等。以上諸多成果，除了在本報告描述外，也包含在計畫之網址、所列之論文、與中日聯合研討會的附件說明中。

I. 簡介

本計畫主要針對中文語音之諸多相關領域，研究新一代的科學及技術。整個計畫有兩大主軸，一是集合國內各研究機構的語音相關學者，包括中央研究院與七所大學的同仁，以期加強合作、整合團隊力量、提昇研究成果；另一主軸則是推動與日本之語音相關領域的主要研究機構的交流互訪， 其中以日本的ATR(Advanced Telecommunication Research Institute)為交流的最主要窗口，另外也包含了五個重點大學。除了短期的互邀參訪、主辦跨國語音研討會加強彼此交流學習之外，更具體且正執行中的交流方式為我方的教授帶領研究生赴日作較長時間的訪問，以達到更深入交流的目的。

本計畫主要的研究方向有三，分列於三個子計畫中。子計畫一強調語音之韻律、聲調及文句翻語音合成，希望建立語音的韻律及聲調變化模型並研究語音合成技術;子計畫二著重語音辨認之強健性技術，希望由訊號、模型及詞彙等各個層次研究提昇語音辨認之強健性的方法;子計畫三則以研發新一代語音系統為目標，希望為未來多媒體數位內容及無線通訊世界先期研發重要的語音系統。關於本計畫之更詳盡之介紹，可參考以下之網址：

http://sovideo.iis.sinica.edu.tw/SLG/ngsst/Index.htm

在本年度計畫結案報告裡，首先在第二節介紹本年度本計畫同仁前往日本各大學及研究單位參訪與深入交流的歷程，第三節、第四節與第五節分別敘述了三個

子計畫在本年度之具體研究成果，第六節介紹與日方合作所共同發表的論文內容。第七節是本計畫今年度所舉辦之跨國研討會的相關介紹，第八節則為本計畫國內同仁所成立之核心實驗室(core-lab)相關說明，第九節為精簡結論。

II. 同仁赴日參訪深入交流

本年度本計畫同仁赴日參訪深入交流的工作共分成兩個部份，一是與東京大學就韻律模型的合作研究，一是與 ATR 就相關領域作深入廣泛交流。此二部份分述於下。

2.1 與東京大學合作中文韻律模型研究

在中文語音韻律模型及聲調辨認研究方面，主要是與東京大學 Prof. Fujisaki 與 Prof. Hirose 合作，共實際安排兩次長期（一個月以上）赴日訪問研究，分別為：

（1）2006 年暑假六月底至九月初，由王逸如教授帶領博士班學生江振宇訪問日本東京大學，江振宇總共待在東大兩個半月，並實際與東京大學 Prof. Hirose 之博士班研究生王曉東，共同合作研究以 tone nucleus 與韻律模型為基礎之中文連續語音聲調辨認。此合作結果並已與日方聯名（co-author）投稿兩篇會議論文，包括一篇以潛在韻律模型為基礎之自動韻律狀態標記與聲調辨認，發表在 ICASSP' 2007 [1]，與一篇以 tone nucleus 為基礎之聲調辨認，發表在 IEICE-technology report [2]（詳細研究內容與成果於第 VI 節關於聯合發表之部分另外詳細敘述）。王逸如與江振宇在赴日訪問期間並順道拜訪東京工業大學 Prof. Furui 與 ATR Prof. Nakamura 各一天，互相做 presentation 與討論相關研究問題。

(2) 2007 年五月，由鄭秋豫教授研究助理蘇昭宇訪問研究東京大學 Prof. Hirose 一個月，主要研究題目為 Fujisaki model 對應階層性語流韻律架構 HPG（Hierarchical Framework of Discourse Prosody）在中文的分析與應用。

2.2 與 ATR 就相關領域作深入交流

陳柏琳教授於民國 95 年 8 月 1 日至 9 月 8 日間帶領李琳山教授博士學生許長文、宮崎益兩位同學赴日本 National Institute of Information and Communications Technology (NiCT)轄下 Advanced Telecommunication Research Institute(ATR)進行研究訪問，在這段期間與 ATR 的 Satoshi Nakamura 博士、Hideki Kashioka 博士、Xinhui Hu 博士以及其他研究人員透過演講與研討方式廣泛且深入地交流台日雙方在語音文件檢索與組織、語音強健處理等研究領域的研究成果，以及規劃將來在語音摘要研究的可能合作方式與題材。陳教授一行人並且在這段期間訪問了 University of Tokyo 的 Keikichi Hirose 和 Shigeki Sagayama 兩位教授、Tokyo Institute of Technology 的 Sadaoki Furui 教授、Kyoto University 的 Tatsuya Kawahara 教授以及 Nara Institute of Science and Technology 的 Kiyohiro Shikano 教授，進一步深入瞭解日方近年來在語音相關研究的進展。

III. 子計畫一：韻律、聲調及文句翻語音(Prosody, Tones and Text-to-Speech)

今年度本子計畫的研究方向與成果主要分為三個主題，包括（1）建立以潛在韻律模型為基礎之韻律狀態自動標記核心技術，預備將來應用此核心技術於語音辨認，合成與語者辨認，（2）中文語流階層式韻律架構模型，與（3）基於潛在韻律分析之語音與語者辨認，以下簡述各主題的研究內容。

3.1. 以潛在韻律模型為基礎之韻律狀態自動標記

在這方面的研究主要是提出以一聯合（joint）潛在韻律模型描述韻律標記、韻律特徵與語言特徵間的相互關係[3]，其完整數學模型如下式所示：

$$\begin{aligned}
\mathbf{B}^*,\mathbf{P}^* &= \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{B},\mathbf{P}|\mathbf{SP},\mathbf{PD},\mathbf{PE},\mathbf{L},\mathbf{T}) \\
&= \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{B},\mathbf{P},\mathbf{SP},\mathbf{PD},\mathbf{PE}|\mathbf{L},\mathbf{T}) \\
&= \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{SP},\mathbf{PD},\mathbf{PE}|\mathbf{B},\mathbf{P},\mathbf{L},\mathbf{T}) P(\mathbf{B},\mathbf{P}|\mathbf{L},\mathbf{T}) \\
&\approx \arg\max_{\mathbf{B},\mathbf{P}} P(\mathbf{SP}|\mathbf{B},\mathbf{P},\mathbf{T}) P(\mathbf{PD},\mathbf{PE}|\mathbf{B},\mathbf{L}^1) P(\mathbf{P}|\mathbf{B}) P(\mathbf{B}|\mathbf{L}^2)
\end{aligned}$$

其具有四個子模型，包含（1）音節基頻軌跡模型 $P(\mathbf{sp}_{k,n}|p_{k,n},B_{k,n-1},B_{k,n},t_{k,n-1},t_{k,n},t_{k,n+1})$，（2）停頓長度 $P(pd_{k,n},pe_{k,n}|B_{k,n},\mathbf{l}^1_{k,n})$，（3）韻律狀態轉換 $P(p_{k,n}|p_{k,n-1},B_{k,n-1})$ 與（4）停頓語法模型 $P(B_{k,n}|\mathbf{l}^2_{k,n})$。其中：

（1）音節基頻軌跡模型，假設音節基頻軌跡的分佈受前後文,聲調(t)，韻律狀態(p)與停頓(b)等潛在影響因素（affecting factor, $\boldsymbol{\beta}$）影響，如圖 1 所示：

並以下列加成性數學模型表示：

$$\begin{aligned}
\mathbf{sp}_{k,n} = \mathbf{sp}^r_{k,n} &+ \boldsymbol{\beta}_{t_{k,n}} + \boldsymbol{\beta}_{p_{k,n}} + \boldsymbol{\beta}^f_{B_{k,n-1},tp_{k,n-1}} \\
&+ \boldsymbol{\beta}^b_{B_{k,n},tp_{k,n}} + \boldsymbol{\mu}
\end{aligned}$$

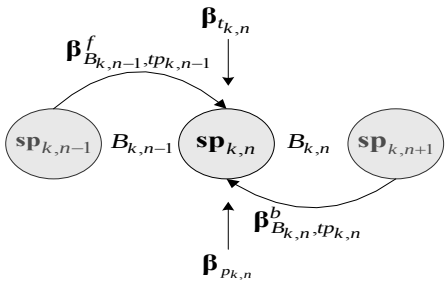


圖 1. The relationship of $\mathbf{sp}_{k,n}$ and its affecting factors.

（2） pause duration 則同時考慮 break (**B**)與 linguistic features（**l**）因素的影響，以下式表示：

$$P(pd_{k,n}, pe_{k,n} \mid B_{k,n}, \mathbf{l}^{1}_{k,n}) = g(pd_{k,n}; \alpha_{B_{k,n},\mathbf{l}^{1}_{k,n}}, \beta_{B_{k,n},\mathbf{l}^{1}_{k,n}})$$
$$N(pe_{k,n}; \mu_{B_{k,n},\mathbf{l}^{1}_{k,n}}, \sigma^{2}_{B_{k,n},\mathbf{l}^{1}_{k,n}})$$

（3,4）prosodic state transition probability $P(p_{k,n} \mid p_{k,n-1}, B_{k,n-1})$ 與 break-syntax model $P(B_{k,n} \mid \mathbf{l}^{2}_{k,n})$ 分別由統計或是由 decision-tree 學習而來。

此模型的訓練方式，由先初始此四個子模型，然後整合此四子模型成一完整的聯合模型，然後再以 expectation maximization（EM）方式，以 iteration 方式調整學習而來。

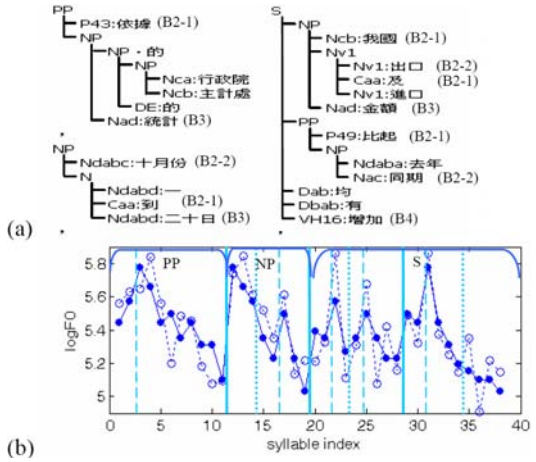


圖 2. An example of the automatic prosodic labeling. (a) syntactic trees with prosody tags and (b) syllable logF0 means: observed (open circle) and prosodic state+global mean(close circle). Solid/dash/dot lines represent B3/B2-1/B2-2 respectively.

此模型訓練完成後，則可以用來對測試語料作自動 prosodic tag 標記，以預備將來應用此核心技術於語音辨認，合成或語者辨認等多方面應用，下圖為一典型的自動標記範例，從此範例可看出此模型的有效性。

3.2. 中文語流階層式韻律架構模型

我們提出語流韻律結構必須以多短語為單位，並具有如圖 3 之階層式架構，包含由下而上的 lexical prosody (tone), syntactic prosody (intonation) 與 discourse prosody (cross-phrase semantic associations)等單元與結構，而且語流韻律受到由上而下的 discourse, syntactic 與 lexical information 等層層交互影響與節制[4，5]。

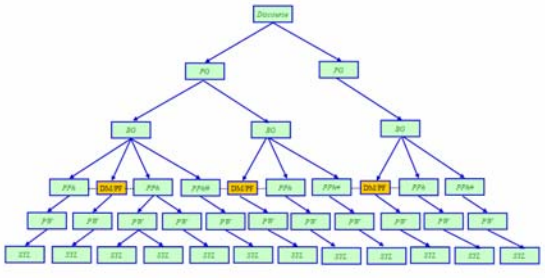


圖 3. A schematic representation of how PGs form spoken discourse.

我們進而對此大單位的次級單位、階層意義、單位邊界、多短語單位提出聲學證據，進而提出各級單位在基頻走勢、音節長度、能量分佈、單位邊界的終止式模版 (cadence templates)，並以此為基礎結合 Fujisaki model 建立一模組式數學模型，用以描述與預測韻律各種的變化。

此模型包含（1）F0 contour（使用 Fujisaki parameters)，（2）Pause，（3）Duration 與（4）Intensity 四個子模型，每個子模型並以 linear prediction 方式，由下而上在 HPG 的每一層次都建立一組線性預估模型，如圖 4 所示，主要是下層無法預估的誤差殘餘量，逐步由上層的線性預估模型吸收。

以下以 pitch contour 預估為典型範例，說明此模型的有效性，圖 5 為預估 pitch contour 時考慮或不考慮 prosodic phrase group (PG) 之間的相互關係，所合成之 pitch contour，由圖可看出考慮 PG 之間的相互關係時，所合成之 contour 可較接近原始軌跡。

此外，我們並整合 HPG 韻律模型與 TTS 系統，製作一中文氣象預報 TTS 合成展示系統，以展示 TTS 系統考慮 HPG 架構的必要性。

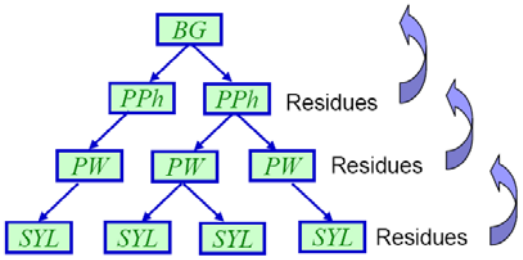
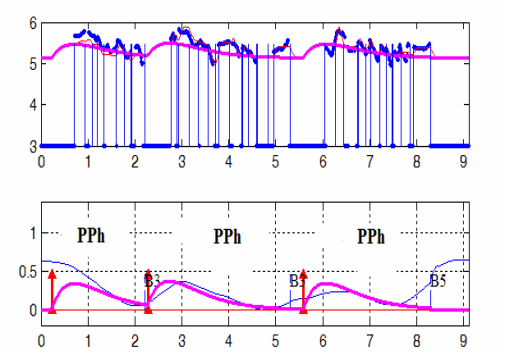
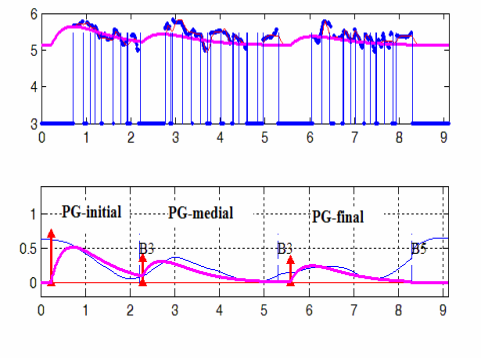


圖 4. A schematic representation of the Hierarchical linear model where residual between prediction and original values is regarded as the government from higher level instead of error.



(a)



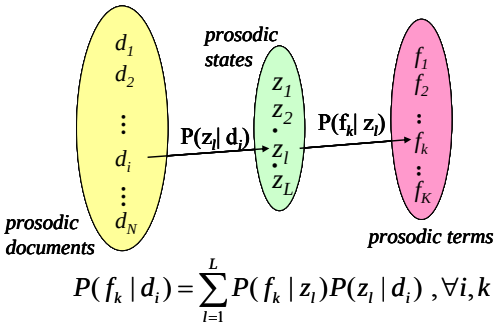
(b)

圖 5. Comparison of the synthesized pitch contour considering (a) without and (b) with PG-effect.

3.3. 基於潛在韻律分析之語音與語者辨認

在韻律模型應用方面，我們主要是以潛在韻律分析應用到語音與語者辨認方面，其方法是在傳統以頻譜為主之語音與語者辨認系統中，考慮如下式之頻譜與韻律之聯合模型如圖 6 所示，並進一步將韻律模型拆解成數個子模型，尤其是包含 tone 與 break 模型，每個子模型並用 probabilistic latent semantic analysis（PLSA）方式萃取與韻律相關之潛在因素（latent factor）之建立，並同時濾掉與韻律無關之潛在因素。

$$
\begin{aligned}
W^* &= \arg\max_W p(W|X,F) \\
&= \arg\max_W p(W)\,p(X,F|W) \\
&\approx \arg\max_W p(W)\,p(X|W)\,p(F|W) \\
&= \arg\max_W p(W)\,p(X|W)\,p(F|T,B)
\end{aligned}
$$



$$P(f_k|d_i) = \sum_{l=1}^{L} P(f_k|z_l)P(z_l|d_i)\,,\forall i,k$$

圖 6. The block diagram of the probabilistic latent semantic analysis (PLSA).

我們將此方法應用到大字彙連續語音辨認，自發性語音辨認，與語者辨認與確認上，分別處理語言模型，講話不流利現象（disfluency），與通道強健性上，皆有不錯之效果。

IV. 子計畫二：語音辨認之強健性 (Robustness in Speech Recognition)

本子計畫主要針對語音辨認中所遭遇的不匹配問題，在辨認系統中各個重要階段如訊號層面、參數層面、模型層面（包含語音模型和語言模型等）、最佳路徑蒐尋層面、以及系統層面，皆加以考量強健性的議題，並發展相對應的技術以提高語音辨認的效能與穩

定性。各層面的技術在語音辨認系統中之整合關係如圖 7 所示。在本年度裡，研究成果即依各層面加以敘述說明如下。

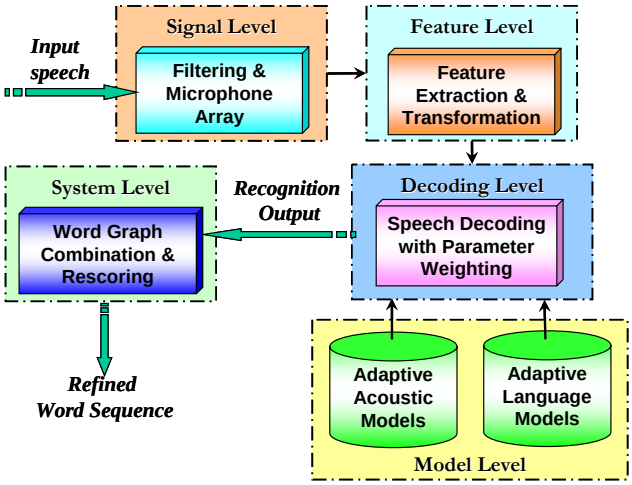


圖 7：語音辨識流程圖

4.1. 訊號補償與強化

本研究是以麥克風陣列為出發點來進行雜訊消除，首先使用功率頻譜相位，空間功率頻譜，訊號在時間軸的互相關性等三種方法來預估麥克風間的時間延遲，之後進行延遲相加之波束形成(Delay and Sum Beamforming)，希望能夠消除均值為零的背景雜訊。接著在麥克風陣列的系統中，以最小平方誤差(Least Square Error, LSE)為理論基礎的適應性濾波器來進行雜訊消除，並且在適應性濾波中參考雜訊的產生方式，再次利用 LSE 的觀念來進行改進。

經過以上的運算轉換成單一通道的訊號。並將其轉至頻域，使用遞迴平均最小值 (Minima Controlled Recursive Averaging, MCRA)預估法來估計剩餘雜訊的頻譜，此方法是以遞迴的方式來一步一步的估計雜訊的頻譜，並利用預估局部最小值的方式以及搭配類似訊噪比的參數及一些臨界值，來預估各個頻率是否含有語音，並根據上述判斷來預估雜訊。接著利用已經估計出來的雜訊頻譜來進行語音增強，最基本的方法即是頻譜刪減法(Spectral Subtraction) 但是這樣的操作在轉回時域的時候會產生音樂性雜訊(Musical Noise)。因此本研究便使用另外一種方法，稱為對數頻譜(Log-Spectral Amplitude)預估，它是以含雜訊的語音訊號(Noisy)為輸入，來估計一個增益函式(Gain Function)，使得輸入訊號乘上它，能夠得到乾淨的語音。原理則與估計雜訊能量類似，利用語音存在及語音不存在兩種條件，分別對各頻帶作動態的改變增益函式。當偵測出接近語音訊號，則調整一個較大的增益，反之，則調整成一個較小的增益。

最後再利用反傅立業轉換把訊號轉回時域，並使用重疊相加(Overlap and Add)演算法，重建訊號波形。

4.2. 強健性參數擷取

在本子計劃中，對於強健性參數擷取方面，共提出有兩項方法。首先為針對語音辨認的系統中改善特徵參數所使用之時間序列濾波器。在過去資料導向的時間序列濾波器藉由一些最佳化準則的使用已經被證實有能力提升語音辨識系統在雜訊環境下的辨識率。然而這些濾波器的設計通常是根據特徵參數在時域上的統計特性，而在本研究[6]中提出三種新的資料導向的時間序列濾波器，主要是使用特徵參數在調變頻域上的統計特性，包括受限之主軸成分分析法（Constrained - Principal Component Analysis，C-PCA）、受限之線性鑑別分析法（Constrained- Linear Discriminant Analysis，C-LDA ）以及受限之最大分類距離法（Constrained- Maximum Class Distance，C-MCD）；直接透過調變頻域上的統計特性來求得最佳化的時間濾波器序列之係數，同時我們也進一步將這些新技術與傳統倒頻譜正規化方法做更進一步的結合，包含倒頻譜平均與變異數正規化法（Cepstral Mean and Variance Normalization，CMVN）以及倒頻增益正規化法（Cepstral Gain Normalization，CGN）。研究中並採用國際通用的 Aurora 2.0 數字語音資料庫作為實驗之評估；由初步實驗結果顯示我們所提出之三種新時間序列濾波器技術，可以顯著提升語音辨識之正確率，不管在各種雜訊環境下皆有良好的表現。而當使用於梅爾倒頻譜特徵係數（Mel-Frequency Cepstral Coefficients，MFCC）時，顯示與倒頻譜平均與變異數正規化法和倒頻增益正規化法有加成的效果，得以進一步提升辨識的正確率。

一般而言，隨著雜訊的增加，雜訊語音與乾淨語音之間的統計量會有越來越顯著的差異。為了讓擷取出之參數其統計特性能接近於乾淨語音，在文獻當中可佐証，對倒頻譜參數之統計量作補償(Cepstral Mean Subtraction, CMS)或正規化(Cepstral Normalization)的動作是能有相當助益的。然而，由於倒頻譜平均數與變異數正規化方法僅考量到一階與二階統計量，因此在本研究中，也探討了高階統計量對正規化的影響，並提出倒頻譜高階動差正規化方法（Higher Order Cepstral Moment Normalization, HOCMN），使參數更具有強健性。在高階統計量的正規化過程當中，由於奇數統計量所對應的是有關“對稱性”的特性，因此我們希望其奇數動差函數值為 0；若為偶數之統計量時，我們便將其動差函數正規化成與正規化之常態分配所對應之統計量相同。在 Aurora 2.0 語料庫的實驗上，我們探討了不同的高階統計量所產生的影響，並

且驗證了所提出的倒頻譜高階動差正規化方法在環境不匹配的條件下，能得到更穩健之語音辨識效能。

4.3. 語音模型與語言模型

在傳統語音辨識系統中，語音模型和語言模型常被假設為互相獨立，分別去估測其對應之參數，再加以結合計算分數。然而語音特徵與語言特徵相互之間有著高度的相關性，在此我們提出以最大熵法則，於單一事後機率模型中同時考慮語音及語言特徵，使得語音和語言模型參數可於相同的訓練法則之下，同時達到最佳化，而其間的相依特性也將被結合入參數之中[7]。

最大熵法則的基本概念是-盡量完整的描述我們所觀察到的資訊，對於我們不知道的，將不對其做任何假設而使其保持均勻分佈。利用最大熵法則同於結合語音及語言模型，首先我們定義特徵函數(Feature Function, f)來反應我們希望擷取的特性，在此我們延續傳統隱藏式馬可夫模型與統計式 n-gram 語言模型，分別為其定義特徵函數，其中包含狀態初始特徵、狀態轉移特徵、混合權重特徵、K 階統計徵與 n-gram 特徵。利用定義之特徵函數，由訓練資料中擷取初期個別對應之統計量，並利用限制條件使所求的最大熵模型能符合該統計特性，在滿足此限制條件下，我們將模型的熵值最大化，最後求得整合式最大熵(Unified ME)模型如下:

$$P(W|X)=\sum_{S,L}\frac{\exp\sum_{i=1}^{F}\lambda_i^{\mathrm{AL}} f_i^{\mathrm{AL}}(W,X,S,L)}{\sum_{W'}\exp\sum_{i=1}^{F}\lambda_i^{\mathrm{AL}} f_i^{\mathrm{AL}}(W',X,S,L)}$$

其中 S、L 分別為狀態和混合成分序列，λ_i^{AL} 為此最大熵模型參數。

在此提出以最大熵法則同時結合語音及語言特徵，以估測出一整合的事後機率模型，語音和語言模型間的交互資訊也將被融入參數之中，同時利用最大熵法則，我們也可將更複雜的語音或語言特性，簡單並有效的集合到此整合性模型之中，以達到更強健的語音辨識效能。

4.4. 最佳路徑蒐尋

在語音辨識中，給定一組語音樣本的特徵向量，我們希望找出其對應的文字，然而該特徵向量中，不同維度對於辨識理應擁有不同的鑑別能力，因此我們利用熵(Entropy)值的計算對不同維度加以不同權重，以提高辨識的可靠性。

在本研究中[8]，對於某一個時間點的觀察樣本的某個維度，對於不同類別都將有一相似度分數，若這些分數彼此很相近，及其對應的熵值較大，反映出該時間點下，此維度對於辨識有較高的不確定性，難以對不同類別加以區別，因此我們給定一較小的權重值；相反的，若其對應的熵值較小，則給定較大的權重。如此動態的對不同維度加以權重調整，用以反應其鑑別能力，將可有效的的提升辨識的強健性與正確性。

4.5. 系統層面考量

在大詞彙連續語音辨識中，利用不同模型參數或是搜尋演算法，都會得到不同的辨識結果，而各種方法都有其優缺點，因此在本研究中我們提出一系統整合的方法，有效的將不同的系統的設定空間(Hypothesis Space)加以整合，其中我們利用多種不同的結合法則，包括 Consensus Score、Expected Phone Accuracy Score 以及 Time Frame Error 來擷取出一整合的辭圖(Word Graph)，並以設定空間之概念合併於系統當中，如此一來所有時間資訊都將被保留並使得重計分數(Rescoring)更有彈性[9]。

V. 子計畫三：新一代語音系統 (Next Generation Spoken Language Systems)

本子計畫執行的重點在發展新一代的語音系統技術，包括了語音文件自動轉寫、多媒體內涵的理解與組織、語音文件檢索、口語對話、以及分散式語音辨識等。在本年度裡，我們主要研究成果包含(1)分散式語音辨識、(2)自動語音文件摘要、(3)自動音素分段、(4)語者確認與分群、(5)動態語言模型調適、(6)鑑別式聲學模型訓練，分別於下面各次章節裡作介紹。

5.1 分散式語音辨識

分散式語音辨識(Distributed Speech Recognition)系統所面臨的問題為：隨身攜帶的手持設備面臨多變且無可預知的環境，環境雜訊(Environmental Noise)與壓縮帶來的信號失真(Quantization Distortion)以及無線通道傳輸所帶來的位元錯誤(Bit Errors)會互相加成起來，嚴重影響分散式語音辨識的效能。因此，探討如何減低碼本不匹配(Codebook Mismatch)、環境不匹配和無線通道傳輸錯誤(Transmission Errors)對辨識率影響的強健性研究便成為一個很重要的研究方向。本計畫針對訓練聲學模型與量化碼本的語音取得環境和實際進行語音辨識環境不匹配的問題，提出一套強健性的量化法，可以根據使用者手持設備的計算能力與頻寬限制，以及所處環境的背景訊噪比不同，動態調整資料傳輸率，使得辨識系統效能達到最佳化。這種新發展出來的以分佈統計為基礎的量化法(Histogram-based Quantization) [10]跳脫了以往傳統分割式向量量化法(Split Vector Quantization)使用固定碼本計算距離的概念，而是根據語音的分佈統計特性，動態訂出量化區間，完全解決了雜訊環境下碼本不匹配的問題，同時量化本身吸收了雜訊的干擾，對於極嘈雜環境和特性

不穩定(Non-stationary)的雜訊亦可有效處理，實驗結果顯示比起傳統以距離為基礎之量化法有大幅進步，對於環境雜訊和無線通道錯誤皆具強健性。

5.2 自動語音文件摘要

自動文件摘要(Automatic Document Summarization) 是一個已經被廣泛探討的主題。許多學者提出不同的方法嘗試找出文章中重要的片段並呈現。在語音文件的應用上，往往因為語音辨識錯誤，使得許多既有成功的摘要方法在語音文件上表現不如預期。目前已公認成功的的方法有詞頻反文件頻 (TF.IDF) 以及顯著度分數 (Significance Score)。本研究基於機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA) 提出了主題顯著度 (Topic Significance) 和主題亂度 (Topic Entropy) 等不同指標，有效的找出了語音文件中重要的片段，並且針對中文語音文件的特性提出了參考詞(Word)以外不同的語音單位，包括音節(Syllable)，字元 (Character) 以及雙音節片段 (Segments of Two Syllables)等[11][12]。另一方面，本研究也提出一個機率生成架構(Probabilistic Generative Framework)，將文句生成模型(Sentence Generative Model)與文句事前機率(Sentence Prior Probability)緊密地耦合，用於摘要之重要文句選取。其中，待摘要文件中每一文句被視為一個生成模型，藉以預測文件生成的機率。文句生成模型是透過隱藏式馬可夫模型(Hidden Markov Model, HMM)與關聯性模型(Relevance Model, RM)的結合，以解決文句模型參數估測時資料稀疏問題 [13]。

5.3 自動音素分段

本研究將「最小化音素錯誤」(Minimum Phone Error, MPE)訓練技術推廣到自動語音段落標記上。目前最廣為使用的自動音素分段方式是利用以最大化相似度(Maximum Likelihood, ML)法則訓練得到的音素 HMM 模型，透過維特比(Viterbi)演算法，在語音信號上根據人工標記的音素序列(Phoneme Sequence)進行強迫校準(Forced Alignment)。由於 ML 訓練法則與希望達成的音素邊界誤差最小的目標並不一致，傳統的 HMM 模型切音無法提供可靠的精確度。有鑒於此，本研究提出最小化邊界誤差(Minimum Boundary Error, MBE)訓練法來取代傳統的 ML 訓練方式。在強迫校準階段，需要找出最佳的邊界序列，使得邊界誤差愈小愈好。有鑒於此，本研究將減損函數定義為實際邊界誤差，在包含大量候選邊界序列的音素網格圖上進行基於 MBE 的音素強迫校準搜尋，並結合後處理時間持續模型來克服 HMM 的不足 [14][15]。

5.4 語者確認與分群

Likelihood Ratio($L(U)=\log p(U|\lambda)-\log p(U|\overline{\lambda})$)是目前最常被採用的語者確認方法，其中 U 是輸入語音，$p(U|\lambda)$ 是被宣稱使用者的模型輸出的likelihood，$p(U|\overline{\lambda})$ 則是非該使用者的模型輸出的likelihood。過去幾年，學者專家陸陸續續提出各種評估 $\overline{\lambda}$ 的方式，但沒有一種被公認對所有案例都是最有效的。這些方法可以歸納成為通式：$p(U|\overline{\lambda})=H(p(U|\lambda_1),...,p(U|\lambda_B))$，其中$H$是一個計算$B$個背景模型(Background Models) $\{\lambda_1, \lambda_2,..., \lambda_B\}$獲得的Likelihoods的一個轉換函數，例如算術平均、幾何平均、取最大值等。本研究將該通式表示成 $p(U|\overline{\lambda})=\sum_{i=1}^{B} w_i p(U|\lambda_i)$ 或是 $p(U|\overline{\lambda})=\prod_{i=1}^{B} p(U|\lambda_i)^{w_i}$，並利用最小化確認錯誤(Minimum Verification Error, MVE)訓練法來求得權重值(w_i)的解，在語者確認實驗上獲得相當不錯的辨識率提升 [16]。

另一方面，語者分群（或稱「非監督式的語者辨認」）的目的是要將屬於相同語者的語音區段集中至同一群中，而屬於不同語者的語音區段則被分開至不同群。語者分群的問題主要有三：(i)音段與音段之間的距離量測、(ii)分群演算法、(iii)最佳群數決定。傳統的作法是先選定一種距離量測，利用由下而上或是由上而下的階層式分群演算法建立分群樹，然後利用模型選擇法來決定最佳群數。本研究利用待分群音段間的整體關連特性取代單一音段對距離計算，可以克服短音段間距離估算的困難。有鑑於傳統的階層式分群法免不了會有錯誤傳遞(Error Propagation)的問題，而常用的基於貝氏資訊基準(Bayesian Information Criterion, BIC)的群數決定法也不是很可靠，他們進一步提出一個基於最小化 Rand Index 的分群法，此法最大的優點是將分群與最佳群數決定合而為一，因此不會發生錯誤傳遞，這個分群法可以利用基因演算法來實現 [17]。

5.5 動態語言模型調適

語音辨識在一些較複雜困難的課題如廣播及電視新聞自動轉寫，由於新聞的主題和語言內容的詞彙使用具多變性與時效性，會使得傳統 n-連(n-gram)語言模型往往很難準地確估測搜尋時候選詞的出現機率，因此語言模型調適格外顯得重要。過去學者嘗試使用潛藏語意分析(Latent Semantic Analysis, LSA)或者機率式潛藏語意分析(Probabilistic Latent Semantic Analysis, PLSA)等方法，以候選詞與其歷史詞序列(Word History)在向量(或機率)潛藏主題空間的相近關係來估測候選詞出現機率，達到調適 n-連語言模型的目的。本研究提出詞主題混合模型(Word Topical Mixture Model, WTMM)，放寬了 n-連語言模型的距離限制，同時直接以機率生成方式表示語言中詞與詞間在機率潛藏主題空間的關係。語音辨識時可透過組合歷史詞序列中詞所對應的詞主題混合模型，動態地求得候選

詞與其歷史詞序列在機率潛藏主題空間的相近關係，以估測候選詞出現機率。在中文大詞彙連續語音辨識的實驗結果顯示，所提出的詞主題混合模型較現有潛藏語意分析、機率式潛藏語意分析以及觸發對模型(Trigger-Based Language Model, TBLM)在語言模型調適上有較佳的表現 [18]。

5.6 聲學模型鑑別式訓練

聲學模型鑑別式訓練(Discriminative Training)旨在最小化訓練語句的分類錯誤，而不僅是去逼近聲學模型參數的真實密度分佈。因為鑑別式訓練同時考慮到聲學模型的分類錯誤資訊，所以其語音辨識效能會比傳統的最大化相似度訓練(Maximum Likelihood Training)來的好。另一方面，近年來以邊際基礎(Margin-Based)的模型估測法則在機器學習領域中已有高度的發展，例如：支撐向量機(Support Vector Machine, SVM)，其以邊際基礎的模型估測法則的中心思想，嘗試選出離決定邊界(Decision Boundary)較近訓練樣本(Training Samples)、利用最大邊際法則來訓練模型參數；最近幾年，以邊際基礎的聲學模型估測在語音辨識上已有不錯的成效。有別於傳統邊際基礎的聲學模型估測方法，本研究提出不同層次的語音訓練樣本選取方法應用在鑑別式模型訓練上，期望解決鑑別式訓練法則缺少推廣能力的問題。傳統邊際基礎的模型估測方法是建立在相似度定義域(Likelihood Domain)，而本研究則是嘗試基於詞錯誤率定義域(Word Error Rate Domain)來選取語句(Utterance)、音素(Phone)以及音框(Frame)等三層次的語音訓練樣本。在中文大詞彙連續語音辨識的實驗結果顯示所提出的方法能提升傳統最小化音素錯誤(Minimum Phone Error, MPE)及最大化相互資訊(Maximum Mutual Information, MMI)等兩種聲學模型鑑別式訓練的效能 [19]。

VI. 與日方合作完成之聯合發表(co-authored)的論文

本年度本計畫共完成兩篇與日方聯合發表之論文，包括一篇在 ICASSP' 2007 [1]，探討以聯合潛在韻律與 tone 模型為基礎之自動韻律狀態標記與 tone 辨認應用，與一篇 IEICE-technology report，研究以 tone nucleus 為基礎之聲調辨認 [2]，分述於下。

6.1. 以聯合潛在韻律與聲調模型為基礎之自動韻律狀態標記與聲調辨認

在此我們嘗試將潛在韻律模型應用到自動韻律狀態標記與聲調辨認，首先我們將觀察到之 pitch contour 以下式之線性加成公式表示成具有數個潛在影響因素（affecting factor）的公式：

$$\mathbf{x}_{k,n} = \mathbf{y}_{k,n} + \boldsymbol{\chi}_{t_{k,n}} + \boldsymbol{\gamma}_{p_{k,n}} + \boldsymbol{\mu}$$

其中 $\mathbf{x}_{k,n}$ 與 $\mathbf{y}_{k,n}$ 都是二維向量（mean and slope of log$F0$ contour of a syllable）分別用來表示在 utterance k 中第 n-th syllable 的原始（observed）與正規化後（normalized）的 pitch contours; $\boldsymbol{\mu}$, $\boldsymbol{\chi}_{t_{k,n}}$ and $\boldsymbol{\gamma}_{p_{k,n}}$ 則是語者，tone，$t_{k,n} \in \{1,2,3,4,5\}$ 與 prosodic state，$p_{k,n} \in \{1,2,\cdots,P\}$ 的影響因素。

我們將正規化後的參數 $\mathbf{y}_{k,n}$ 以 Gaussian distribution $N(\mathbf{y}_{k,n}; \mathbf{0}, \mathbf{R})$ 描述，則 $\mathbf{x}_{k,n}$ 可用下式描述：

$$P(\mathbf{x}_{k,n} \mid t_{k,n}, p_{k,n}, \lambda) = N(\mathbf{x}_{k,n}; \boldsymbol{\mu} + \boldsymbol{\chi}_{t_{k,n}} + \boldsymbol{\gamma}_{p_{k,n}}, \mathbf{R})$$

此外若進一步考慮韻律狀態轉移機率 $P(p_{k,n} | p_{k,n-1}, B_{k,n-1})$，則此聯合模型可以定義出一 likelihood function 如下：

$$L = \log\left\{\prod_{k=1}^{K}\left[\prod_{n=1}^{N_k} P(\mathbf{x}_{k,n} | t_{k,n}, p_{k,n}, \lambda) \cdot P(p_{k,1}) \prod_{n=2}^{N_k} P(p_{k,n} | p_{k,n-1}, B_{k,n-1})\right]\right\}$$

在此式中 K 表示 total number of utterances; 而 N_k 表示 total number of syllables in utterance k.

此聯合模型之訓練則以上式之 likelihood funciton，以遞迴式 expectation and maximization（EM）演算法訓練，估出各影響參數之數值。

表 1、2 為訓練完成後，模型自動找出之 tone 與 prosodic state 的個別 affecting factor 與每個 prosodic state 所代表的意義，皆符合語言學（linguistic）的意義。

表 1. The learned affecting factors of the five tones.

Tone	1	2	3	4	5
Mean (LogF0)	0.17	-0.09	-0.21	0.09	-0.13
Slope	0.00	0.01	-0.03	-0.03	-0.01

表 2. The learned affecting factors of the 16 prosody states.

state	1	2	3	4	5	6	7	8
LogF0	-0.85	-0.58	-0.41	-0.29	-0.2	-0.14	-0.09	-0.04
state	9	10	11	12	13	14	15	16
LogF0	-0.01	0.04	0.08	0.14	0.21	0.29	0.40	0.55

測試時則以圖 8 之交錯（iteration）方式，逐步自動找出韻律狀態標記與辨認出每個音節的聲調來。

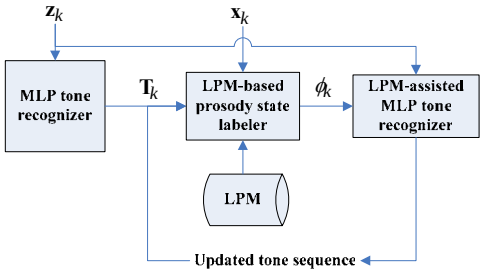


圖 8. The proposed block diagram of the online LPM for simultaneous prosody state detection and tone recognition.

圖 9 則為使用此聯合模型，在 TCC300 中文語料庫上做自動 prosodic state 標記與 tone 辨認的結果，由圖中可見自動韻律狀態標記的結果亦皆符合語言學（linguistic）的意義，此外由表 3、4 的聲調辨認結果比較，顯示考慮 tone 與 prosodic state 之聯合模型，可以比傳統 MLP 方式，得到比較好的 tone 辨認結果。

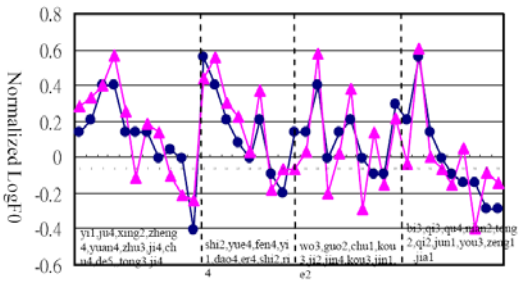


圖 9. A typical result of the LPM-based automatic labeling of prosody states of an input utterance (triangles: original logF0 means of syllables normalized by the speaker affecting factor, circles: pitch means of the corresponding prosody state sequence, dashed lines: correct prosodic phrase boundaries).

表 3. Confusion matrix of the conventional MLP tone recognizer

Ans\Rec	tone 1	tone 2	tone 3	tone 4	tone 5	Total
tone 1	88.17%	5.22%	0.75%	5.12%	0.75%	
tone 2	8.24%	84.89%	2.66%	2.75%	1.46%	
tone 3	5.19%	8.90%	56.00%	26.82%	3.09%	
tone 4	3.35%	1.92%	3.41%	90.20%	1.12%	
tone 5	4.80%	12.00%	7.20%	21.20%	54.80%	80.86%

表 4. Confusion matrix of the LPM-assisted MLP tone recognizer

Ans\Rec	tone 1	tone 2	tone 3	tone 4	tone 5	Total
tone 1	90.19%	5.01%	0.75%	3.52%	0.53%	
tone 2	6.61%	86.78%	2.40%	1.80%	2.40%	
tone 3	3.96%	9.52%	61.80%	19.90%	4.82%	
tone 4	4.22%	1.49%	4.03%	88.72%	1.55%	
tone 5	5.60%	12.00%	7.20%	13.60%	61.60%	82.55%

6.2. 以 tone nucleus 為基礎之 tone 辨認

此部分的研究主要是假設 tone pattern 中只有如圖 10 所示的中心 tone nucleus 部分，才是不變的部分，除此外的部分皆是前後 tone 連結時，因前後關係（context）產生的自由變化，因此若我們可以單獨萃取出 tone nucleus 部分，應可去除 context 的影響，有效提升 tone 的辨認率。

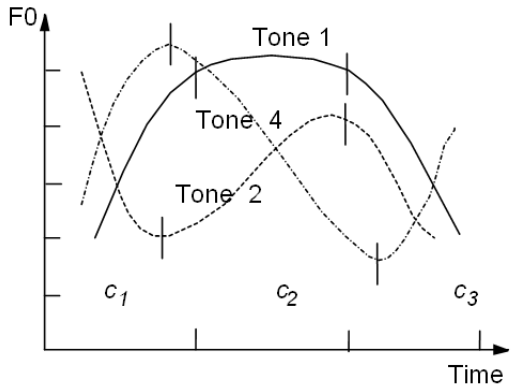


圖 10. Illustration of tonal F0 contours with possible articulatory transitions for Tones 1, 2 and 4. The left and right vertical sticks in each contour correspond to the possible tone onset and offset locations, and the three F0 segments delimited by tone onset and offset are onset courses, tone nuclei and offset courses, depicted as c_1, c_2 and c_3.

圖 11 為我們所提出之以 maximum likelihood 與 Viterbi search 為基礎之 tone nucleus 萃取方法。

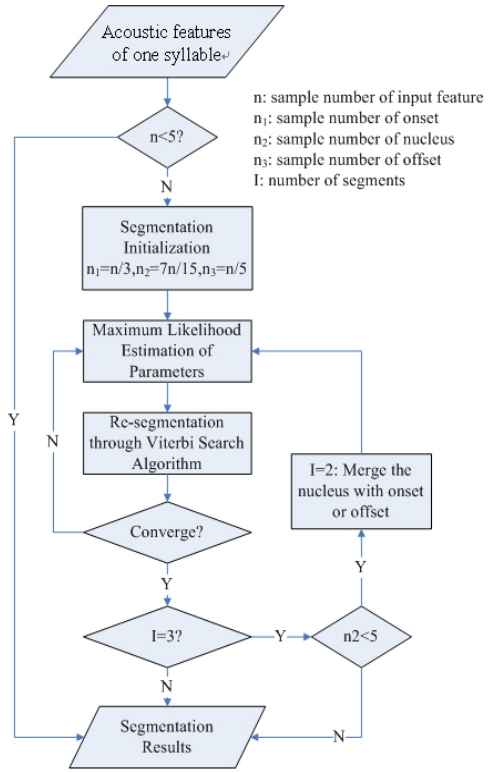


圖 11. Illustration of tone nucleus extraction algorithm.

藉由此自動切割所求得之 tone nucleus 則進一步求取特徵參數後分別與以 multiple layer perceptron（MLP）與 hidden Markov model（HMM）方式辨認，並與傳統方式做比較。圖 12 為我們所提出方法的與傳統特徵參數在香港中文大學中文語料庫上做 tone 辨認的效能比較，可以看出以 tone nucleus 方式求取特徵參數可以有較低的錯誤率。

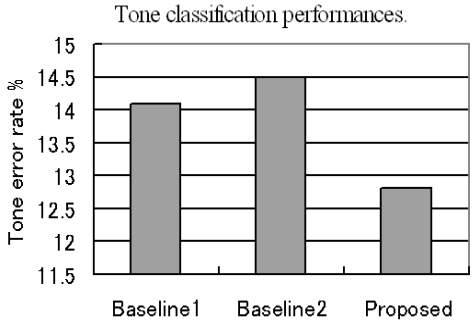


圖 12. Tone recognition performances in error rates.

VI.中日跨國研討會
(Taiwan-Japan Joint Workshop)

今年本計畫於三月底邀請日本十位從事語音研究的專家學者來台舉辦中日跨國研討會，以與國內語音相關研究人士進行學術交流。此次中日跨國研討會由本計畫主辦、中華民國計算語言學會及台灣大學電機系協辦，於三月 29、30 日兩天於台灣大學博理館國際會議廳舉辦，會中共安排了 2 個 keynote speech、12 個 invited talk、poster/demo session 及座談會各一個。研討會共有 165 名國內學界及業界人士報名參加，其中業界參與人士包括國內各大研究機構及從事語音研究之公司，如中華電信研究所、工業技術研究院、台達電等 50 人。

此次研討會所邀請的日方專家學者計有：東京大學的 Hiroya Fujisaki 教授、東京理工大學的 Sadaoki Furui 教授、國家資訊研究所的 Shuichi Itahashi 教授、東京大學的 Keikichi Hirose 教授、京都大學的 Tatsuya Kawahara 教授、NTT 通信實驗室的 Shoji Makino 博士、ATR 暨 NICT 的 Satoshi Nakamura 博士、東京大學的 Shigeki Sagayama 教授、奈良科技研究所的 Kiyohiro Shikano 教授、早稻田大學的 Yoshinori Sagisaka 教授，每位都是在國際上享有盛名之語音相關研究專家學者。在台灣學者方面，除有本計畫中三個子計畫的報告之外，還邀請了成功大學吳宗憲教授代表本計畫以外的國內學者於會中發表專題演講。

會中專題演講之題目包含：(1)日本學界之研究成果；如：東工大 Furui 教授之 Question and Answering system、東京大學 Hirose 教授之 Synthesizing Fundamental Frequency Contours in the Framework of Generation Process Model for Flexible Control of Prosody、京都大學 Kawahara 教授之 Automatic Detection of Sentence and Clause Units from Spontaneous Speech，這些內容讓聽眾能瞭解日本學界語音方面在 QA、spontaneous speech recognition 及 prosody 方面的成果，而東京大學 Fujisaki 教授所講授的” Fujisaki Model” 是從事語音合成研究人員之重課題、早稻田大學 Sagisaka 教授之 Prosody Generation for Communicative Speech Synthesis 是極具前瞻性的新方向、東京大學 Sagayama 教授之 Seeking New Directions Toward Speech, Music, and Acoustic Signal Processing 是聲學模型在音樂信號分析之應用、國家資訊研究所 Itahashi 教授之 On the Activities of Oriental COCOSDA and NII Speech Resources Consortium 則是讓聽者充分瞭解亞洲地區語料庫之蒐集狀況，(2)日本業界的研究成果及其發展之實用系統，如：ATR 暨 NICT Nakamura 博士之 Corpus-based Speech Translation、NTT 通信實驗室 Makino 博士之 Audio

Source Separation based on Independent Component Analysis、奈良科技研究所 Shikano 教授之 New Speech Media Applied to Universal Communication，其中 Makino 博士之演講內容亦受邀於今年的 ICASSP 中做 tutorial 演講， (3)在台灣學者方面，除有本計畫三個子計畫的報告之外，成功大學吳宗憲教授之 Generation of Speaking Styles and Emotional Expressions in Text-to-Speech Synthesis 則涵蓋了情緒語音及語者習性在語音合成之研究成果。所以會中的 14 個專題演講可以說涵蓋了各個語音相關研究領域之重要課題。

會中除中日雙方的專家學者之精彩演講之外，並於三月 30 日上午安排 poster/demo session，共有 17 個 posters 及 8 個系統展示，以讓中日雙方專家學者可以更進一步的進行學術交流。三月 30 日下午也安排了座談會，座談會中也熱烈討論了如何讓雙方更進一步進行合作，尤其是台灣學生赴日進修的事宜。

VIII. 語料及工具核心實驗室之建立 (Corpus and Tool Core Lab)

本計畫成立Corpus and Tool Core Laboratory (CTCL)的目的之一是整合國內各單位在語音相關領域現有的語料、工具、模型及系統等資源，提供國內學術界與研究機構參考使用，讓更多學者及研究單位參與語音研究，以提升我國語音相關研究的水準及在國際學術界的影響力。本計畫持續進行國內各單位語料庫資訊收集的工作，亦針對國外相關資源，如Linguistic Data Consortium (LDC)的中文語料庫進行整理，相關資料均已在計畫網站公告。另外，本計畫亦將各單位可以對外分享的模型及技術條列整理，而各單位開發的雛形展示系統亦在本計畫網站統一展示。本計畫網站的目標是作為中文語音處理研究的入口網站，希望國內外對中文語音處理領域有興趣的團體或個人透過本網站可以獲取各種有用的資訊。

CTCL的另一項重要工作是藉著公開的講習會提供國內各研究單位互相交流學習的機會，我們已經在95年1月19至20日、95年8月17至18日及96年2月8日，在交通大學舉辦過三次講習會。第一次講習會由五位博士班研究生分別就Language Identification、Pronunciation Variation Analysis and Modeling、Discriminative Training、Maximum Entropy Language Modeling及Text-to-speech Synthesis五個主題做深入淺出的介紹。第二次講習會共安排七位博士班及甫畢業碩士班研究生報告，主題包括Spoken Document Retrieval、Speaker Adaptation、Noisy Speech Recognition、Speaker Verification、Speech Event Detection及Confidence Measure。第三次講習會則安排六位博士班及甫畢業碩士班研究生報告，主題包括Robust Features for Speech Recognition、Fujisaki Model for Prosody Analysis、Duration and Pause Modeling for Prosody Analysis、Latent Prosody Modeling及Spoken Document Summarization。三次講習會都相當成功，每次均吸引國內從事語音相關研究的學者、專家、研究生約90人參加，所有講義均已上傳至計畫網站供各界下載。第四次講習會預計在96年暑期舉辦。

IX. 結論

在本計畫執行的第二年間，如本報告所述，我們相繼進行了諸多方向的工作，包括三大主題深入整合的研究、同仁赴日的交流參訪、舉辦中日跨國研討會、國內同仁彼此的研究合作與發表論文、技術、語料與工具資料庫的整合、及針對各研究單位之學生研究人員舉辦重要研究主題的講習與訓練等，未來我們將在以上這些方向持續推動，以逐步完成本計畫的各項目標。

X. 參考文獻

[1] Chen-Yu Chiang, Xiao-Dong Wang, Yuan-Fu Liao, Yih-Ru Wang, Sin-Horng Chen, Keikichi Hirose, “LATENT PROSODY MODEL OF CONTINUOUS MANDARIN SPEECH”, ICASSP’2007.

[2] Xiao-Dong Wang, Jin-Song Zhang, Keikichi Hirose, Nobuaki Minematsu, Chen-Yu Chiang, Yih-Ru Wang, Yuan-Fu Liao, “ Tone Recognition of Continuous Mandarin Speech Based on Tone Nucleus Model and MLP Nucleus Network, ”, IEICE technical report. Speech, Nagoya, Japan, Vol. 106, No. 443, pp. 107-112, SP2006-103, Dec., 2006.

[3] Yuan-Fu Liao, Zi-He Chen and Yau-Tarng Juang, “Latent Prosody Analysis for Robust Speaker Identification”, to appear in IEEE Trans. on Speech, Audio and Language Proc. July, 2007.

[4] Tseng Chiu-yu, Pin Shao-huang, Lee Yeh-lin, Wang Hsin-min and Chen Yong-cheng, “ Fluent speech prosody: framework and modeling”, *Speech Communication, Special Issue on Quantitative Prosody Modelling for Natural Speech Description and Generation.* Vol. 46, issues 3-4, pp. 284-309, 2005.

[5] Tseng Chiu-yu, “Recognizing Mandarin Chinese Fluent Speech Using Prosody Information—An Initial Investigation”, *The 3rd International Conference on Speech Prosody 2006.* Dresden, Germany, 2006.

[6] Jeih-weih Hung, “Optimization of Temporal Filters in the Modulation Frequency Domain via Constrained

Linear Discriminant Analysis (C-LDA) For Constructing Robust Features In Speech Recognition", in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2007), Hawaii, USA, April 2007.

[7] Chuang-Hua Chueh and Jen-Tzung Chien, "Maximum Entropy Modeling of Acoustic and Linguistic Features", in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2006), Toulouse, France, May 2006.

[8] Yi Chen, Chia-yu Wan and Lin-shan Lee, "Entropy-based Feature Parameter Weighting for Robust Speech Recognition," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2006), Toulouse, France, May 2006.

[9] I-Fan Chen, Lin-shan Lee, "A New Framework for System Combination Based on Integrated Hypothesis Space," ICSLP 2006, pp. 533-536.

[10] Chia-yu Wan, Yi Chen and Lin-shan Lee, "Three-Stage Error Concealment for Distributed Speech Recognition (DSR) with Histogram-Based Quantization (HQ) under Noisy Environment" in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2007), Hawaii, USA, April 2007.

[11] Sheng-yi Kong and Lin-shan Lee, "Improved spoken document summarization using probabilistic latent semantic analysis (PLSA)," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2006), Toulouse, France, May 2006.

[12] Sheng-yi Kong and Lin-shan Lee, "Improved summarization of chinese spoken documents by probabilistic latent semantic analysis (PLSA) with further analysis and integrated scoring," in Proc. IEEE Workshop on Spoken Language Technology (SLT 2006), Aruba, Dec 2006.

[13] Yi-Ting Chen, Suhan Yu, Hsin-min Wang, Berlin Chen, "Extractive Chinese Spoken Document Summarization Using Probabilistic Ranking Models," in Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP 2006), Singapore, December 2006.

[14] Jen-Wei Kuo and Hsin-min Wang, "A Minimum Boundary Error Framework for Automatic Phonetic Segmentation," in Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP2006), Singapore, December 2006.

[15] Jen-Wei Kuo, Hung-Yi Lo, and Hsin-min Wang, "Improved HMM/SVM Methods for Automatic Phoneme Segmentation," submitted to Interspeech2007-EUROSPEECH.

[16] Yi-Hsiang Chao, Wei-Ho Tsai, Hsin-min Wang, Ruei-Chuan Chang, "Improved Methods For Characterizing The Alternative Hypothesis Using Minimum Verification Error Training For LLR-Based Speaker Verification," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2007), Hawaii, USA, April 2007.

[17] Wei-Ho Tsai and Hsin-min Wang, "Speaker Clustering Based on Minimum Rand Index," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2007), Hawaii, USA, April 2007.

[18] Hsuan-Sheng Chiu and Berlin Chen, "Word Topical Mixture Models for Dynamic Language Model Adaptation," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal processing (ICASSP2007), Hawaii, USA, April 2007.

[19] Shih-Hung Liu, Fang-Hui Chu, Shih-Hsiang Lin, Berlin Chen, "Investigating Data Selection for Minimum Phone Error Training of Acoustic Models," in Proc. IEEE International Conference on Multimedia & Expo (ICME 2007), Beijing, China, July 2007.

中日跨國研討會(Taiwan-Japan Joint Workshop)會議報告

今年本績優團隊國際合作計畫於三月底邀請日本十位從事語音研究的專家學者來台舉辦中日跨國研討會，以與國內之語音相關研究人士進行學術交流。此次研討會所邀請的日方專家學者計有：東京大學的Hiroya Fujisaki教授、東京理工大學的Sadaoki Furui教授、國家資訊研究所的Shuichi Itahashi教授、東京大學的Keikichi Hirose教授、京都大學的Tatsuya Kawahara教授、NTT通信實驗室的Shoji Makino博士、ATR暨NICT的Satoshi Nakamura博士、東京大學的Shigeki Sagayama教授、奈良科技研究所的Kiyohiro Shikano教授、早稻田大學的Yoshinori Sagisaka教授，每位都是在國際上享有盛名之語音相關研究專家學者。在台灣學者方面，除有本計畫中三個子計畫的報告之外，還邀請了成功大學吳宗憲教授代表本計畫以外的國內學者於會中發表專題演講。會中共安排了 2 個keynote speech、12 個invited talk、poster/demo session及座談會各一個。會中的 14 個專題演講可以說涵蓋了各個語音相關研究領域之重要課題。

此次中日跨國研討會由本計劃主辦、中華民國計算語言學會及台灣大學電機系協辦，於三月 29、30 日兩天於台灣大學博理館國際會議廳舉辦，研討會共有 165 名國內學界及業界人士報名參加，其中業界參與人士包括國內各大研究機構及從事語音研究之公司，如中華電信研究所、工業技術研究院、台達電等 50 人。

會中除中日雙方的專家學者之精彩演講之外，並於三月 30 日上午安排 poster/demo session，共有 17 個 posters 及 8 個系統展示，以讓中日雙方專家學者可以更進一步的進行學術交流。三月 30 日下午也安排了座談會，座談會中也熱烈討論了如何讓雙方更進一步進行合作，尤其是台灣學生赴日進修的事宜。

以下分別將此次中日跨國研討會之
(1) 研討會議程
(2) 研討會演講、poster 及展示之摘要
(3) 研討會活動照片
詳列於後。

1. 中日跨國研討會(Taiwan-Japan Joint Workshop)議程

March 29 (Thu)

08:30-09:00	Registration		
09:00-09:10	Opening	Lin-shan Lee, National Taiwan University	
09:10-10:10	Keynote Speech: An Overview of Question Answering in Tokyo Institute of Technology	Sadaoki Furui, Tokyo Institute of Technology	Chair: Lin-shan Lee, National Taiwan University
10:10-	Break		

10:30			
10:30-11:20	Prosody Generation for Communicative Speech Synthesis	Yoshinori Sagisaka, Waseda University	Chair: Sin-Horng Chen, National Chiao Tung University
11:20-12:10	Seeking New Directions Toward Speech, Music, and Acoustic Signal Processing	Shigeki Sagayama, University of Tokyo	
12:10-13:10	Lunch		
13:10-14:00	Synthesizing Fundamental Frequency Contours in the Framework of Generation Process Model for Flexible Control of Prosody	Keikichi Hirose, University of Tokyo	Chair: Jen-Tzung Chien, National Cheng Kung University
14:00-14:50	Sub-Project (I): Prosody, Tones and Text-to-speech Synthesis	Sin-Horng Chen, National Chiao Tung University Chiu-yu Tseng, Academia Sinica Yuan-Fu Liao, National Taipei University of Technology	
14:50-15:20	Break		
15:20-16:10	Corpus-based Speech Translation	Satoshi Nakamura, ATR and NICT	Chair: Yuan-Fu Liao, National Taipei University of Technology
16:10-17:00	New Speech Media Applied to Universal Communication	Kiyohiro Shikano, Nara Institute of Science and Technology	
17:00-17:50	Sub-Project (II): Robustness in Speech Recognition	Jen-Tzung Chien, National Cheng Kung University	

March 30 (Fri)

09:00-10:00	Keynote Speech: Modeling the Control Process of Fundamental Frequency of Speech with Applications to Analysis and Representation of the Tone Systems of Some Tone Languages	Hiroya Fujisaki, University of Tokyo	Chair: Chiu-yu Tseng, Academia Sinica
10:00-10:50	On the Activities of Oriental COCOSDA and NII Speech Resources Consortium	Shuichi Itahashi, National Institute of Informatics	
10:50-11:10	Break		

11:10-12:10	Posters and Demos	Taiwanese presenters	
12:10-13:10	Lunch		
13:10-14:00	Automatic Detection of Sentence and Clause Units from Spontaneous Speech	Tatsuya Kawahara, Kyoto University	Chair: Hsiao-Chuan Wang, National Tsing Hua University
14:00-14:50	Generation of Speaking Styles and Emotional Expressions in Text-to-Speech Synthesis	Chung-Hsien Wu, National Cheng Kung University	
14:50-15:20	Break		
15:20-16:10	Audio Source Separation based on Independent Component Analysis	Shoji Makino, NTT Communication Science Laboratories	Chair: Chung-Hsien Wu, National Cheng Kung University
16:10-17:00	Sub-Project (III): New Generation Spoken Language Systems	Hsin-min Wang, Academia Sinica Lin-shan Lee, National Taiwan University	
17:00-17:50	Panel Discussion: What's Next?	Moderators: Hiroya Fujisaki, University of Tokyo, Lin-shan Lee, National Taiwan University	

2. 中日跨國研討會(Taiwan-Japan Joint Workshop)演講、海報及展示之摘要

Keynote Speeches

An Overview of Question Answering in Tokyo Institute of Technology

Sadaoki Furui (Tokyo Institute of Technology)

This talk presents an overview of the data-driven and non-linguistic approach to open-domain factoid question answering (QA) that has been developed over the past 4 years at Tokyo Institute of Technology and which culminated last year in our participation in three international evaluations of QA technology at TREC (for English), CLEF (for Spanish and French) and NTCIR (for Japanese). In TREC2005 we placed 11th out of a total of 30 participants in the factoid QA task and in TREC2006 we came 9th out of 27 participants with an accuracy of 25.1%. While our performance in the official CLEF QA tracks was poor due to a large number of unsupported answers, the performance on an informal "real-time" Spanish QA exercise was one of the best.

Modeling the Control Process of Fundamental Frequency of Speech with Applications to Analysis and Representation of the Tone Systems of Some Tone Languages

Hiroya Fujisaki (Professor Emeritus, University of Tokyo)

This paper first presents the physiological and physical properties of the vocal fold and the laryngeal structure involving intrinsic laryngeal muscles that are mainly responsible for the generation of F0 contours with global components and positive local components. It then shows the mechanism involving extrinsic laryngeal muscles for generating negative local components that are commonly found in F0 contours of tone languages, Based on

these evidences, a command-response model is derived that is capable of generating F0 contours of tone languages with an extremely high accuracy. Experimental results are shown to support the validity of the model for several languages including Mandarin, Cantonese, Thai, Vietnames, etc. Apart from it usefulness in phonetic studies and techonological applications, the model leads to a novel way of representing the phonological structure of the tone system of a tone language in terms of timing and polarity of tone commands.

Invited Speakers from Japan and Their Lectures

Prosody Generation for Communicative Speech Synthesis

Yoshinori Sagisaka (Waseda University)

A corpus-based prosody modeling is proposed aiming at F0 control for communicative speech synthesis. For input information in communicative speech synthesis, input word attributes are employed. Two experimental results are introduced to show the possibility of communicative speech prosody control from input word attributes. The results showed that F0 height, F0 dynamic patterns and duration could be consistently controlled by the word attributes. The positive-negative characteristics can be controlled by F0 height, while confident-doubtful, allowable-unacceptable characteristics were reflected in F0 dynamic patterns and duration. Through these analyses, the correlations between perceptual impressions on output speech and input word attributes are turned to be useful for F0 characteristics prediction from input word attributes.

Seeking New Directions Toward Speech, Music, and Acoustic Signal Processing

Shigeki Sagayama (University of Tokyo)

One of most important issues in speech recognition include robust speech recognition in noisy and reverberant environments. We present a model-based approach to this issue studied at the Department of IPC (Information Physics and Computing), Graduate School of Information Science and Technology, The University of Tokyo. Another issue is the usability of spoken dialog systems. We conducted "Galatea Project" to provide with a license-free open-source toolkit for building anthropomorphic spoken-dialog agent with face animation. On the other hand, speech technology has been highly developed and ready to be applied to other interesting areas. Particularly, music contains rich research areas to which major concepts in speech recognition can apply including spectral analysis, HMM, CSR, training algorithm, language models, and other related items. Harmonically constrained GMM is successfully utilized in multipitch analysis for automatic audio-to-MIDI conversion. HMM is used in rhythm analysis in combination of language model of rhythm vocabulary and unsupervised training for iteratively finding the fluctuation tempo. Ergodic HMM is suitable to represent the chord transition for finding the music tonality (key). Dynamic Programming is useful in automatic counterpoint to find a melody matching to the given melody with music theoretic constraints. Automatic fingering decision, automatic accompaniment and automatic harmonization to the given melody and automatic music composition for a given text are also well formulated with speech technology such as HMM. Statistical estimation algorithms such as EM algorithm is also useful in acoustic signal processing using microphone arrays especially in presence of background noise. The CSR framework best matches online recognition of handwritten Kanji and math expressions exploiting good combination of HMM and language models.

Synthesizing Fundamental Frequency Contours in the Framework of Generation Process Model for Flexible Control of Prosody

Keikichi Hirose (*University of Tokyo*)

Introduction of firm criteria when searching optimum parameters has lead corpus-based methods to a great success in various applications. Currently, almost all speech synthesis systems adopt a scheme based on corpus-based selection and concatenation of waveform (or acoustic parameter) segments. Although synthetic speech from these systems sounds quite natural, it still includes "strange" prosody. The major reason is that prosodic features such as fundamental frequency (F0) are processed in a similar way as segmental features are: frame-by-frame manner. This may cause sudden F0 undulations, which are not characteristic of human speech. Prosodic features cover a wider time span than segmental features, and should be treated differently. Taking these into account, we have developed a corpus-based method of synthesizing F0 contours in the framework of the generation process model (henceforth F0 model) and realized speech synthesis in reading and dialogue styles, and with neutral and several basic emotions. By predicting the model commands instead of frame-by-frame F0 values, sudden undulations in F0 movements do not occur in the generated F0 contours, still keeping acceptable

speech quality even if the prediction is done incorrectly. Also, it is rather easy to analyze the prosodic controls obtained by statistical methods and to modify the generated F0 contours according to our knowledge obtained through observations of natural utterances. My talk will include not only Japanese speech synthesis but also Chinese one, where arrangement of training corpus with F0 model parameters comes difficult. A hybrid of rule and corpus based methods is introduced to cope with the problem.

Corpus-based Speech Translation

Satoshi Nakamura (ATR and NICT)

Speech translation system is a pipe dream for human-being. Authors' group has been studying on speech translation technology consist of speech recognition, machine translation, and speech synthesis from 1986. Problems are to develop technologies which make a system work in real situation, namely, it satisfies the requirements such as coverage of all of daily travel conversations, portability of the topic and domain, possibility to add new topics, adoptability to new language pairs, and robustness to noisy acoustic conditions. The technologies had not reached to the level yet around year of 2000. Challenge of our group is to shift a paradigm from rule-based approach to corpus-based one, to develop a mechanism to realize topic portability, new language pair adoptability, and robustness to real acoustic environments. In this presentation the state of the art Japanese-English, and Japanese-Chinese speech translation system which translates travel utterances into target language will be introduced.

New Speech Media Applied to Universal Communication

Kiyohiro Shikano (Nara Institute of Science and Technology)

We found a new quiet speech media, Non-Audible Murmur (NAM), and a new noise robust blind source separation algorithm, SIMO-ICA. These two inventions can enlarge the speech communication areas to further quiet and noisy environments, and for pre-school children to aged people including speech or hearing handicapped people.

NAM is an extremely small voice, which is audible only for a speaker himself but inaudible for a nearby listener. NAM is detected by NAM microphone. NAM makes it possible to communicate in quiet environments without speaking, which is called quiet telephone. NAM is also applied to speech recognition, which is called quiet speech recognition. In this talk, NAM is applied to speech handicapped aid using our voice conversion technique.

SIMO-ICA has the blind source separation capability without distortion by the single-input multiple-output constraint. SIMO-ICA realizes hands-free communication in noisy environments. In order to implement realtime communication, we combined SIMO-ICA with the binary masking successfully.

We have been developed and operating speech guidance systems in real environments more than four years, based on large vocabulary speech recognition (Julius). Takemaru-kun agent based speech guidance system has been used in Ikoma city communication center these four years. The number of daily inputs is about 600. Takemaru-kun can discriminate noises, laughter, cough, children and adults. Takemaru-kun is positively accepted by Ikoma citizen, especially children. We have been also operating two types of speech guidance system in the noisy environment at the nearby railway station. They are a robot type (Kita-robo) and an agent type (Kita-chan). These two speech guidance systems are accepted by the railway company and wide range of users.

On the Activities of Oriental COCOSDA and NII Speech Resources Consortium

Shuichi Itahashi (National Institute of Informatics)

The talk is composed of two parts: the first part introduces the activities of Oriental COCOSDA and the second part introduces the NII-Speech Resources Consortium.

The purpose of Oriental COCOSDA is to exchange ideas, to share information and to discuss regional matters on creation, utilization, dissemination of spoken language corpora of oriental languages and also on the assessment methods of speech recognition/synthesis systems as well as to promote speech research on oriental languages. A series of Oriental COCOSDA Workshop has been held annually since the preparatory meeting held in 1997. The Oriental COCOSDA is managed by a convener, three advisory members, and 26 representatives from ten regions in Asia. We note that speech research has become popular gradually in Oriental countries.

The National Institute of Informatics (NII) in Japan has decided to initiate the Speech Resources Consortium

(SRC) with the goal of creating future value in information media, particularly speech media. SRC aims at collection, distribution, investigation, research, and standardization of electronic data and software tools that are necessary for the development of science, education, and industry concerning speech. The consortium will contribute to the development of the information society through these activities. Based on these activities, it facilitates the distribution, promotion, and publicity of speech resources. The talk also explains the background of SRC's establishment and the outline of speech corpora development in Japan.

Automatic Detection of Sentence and Clause Units from Spontaneous Speech

Tatsuya Kawahara (Kyoto University)

Detection of sentence and clause units in spoken documents such as lectures and meetings is vital in improving the readability and post-processing such as translation and summarization. This talk first presents two different approaches based on statistical language model (SLM) and support vector machines (SVM) for sentence boundary detection. The SVM-based approach realizes more effective and robust performance, achieving the best F-measure for the CSJ (Corpus of Spontaneous Japanese). Then, for more robust detection of clause units, a novel cascaded chunking strategy is presented. The method incorporates local syntactic dependency, and realizes even better performance in ASR outputs.

Audio Source Separation based on Independent Component Analysis

Shoji Makino (NTT Communication Science Laboratories)

This paper introduces the blind source separation (BSS) of convolutive mixtures of acoustic signals, especially speech. A statistical and computational technique, called independent component analysis (ICA), is examined. By achieving nonlinear decorrelation, nonstationary decorrelation, or time-delayed decorrelation, we can find source signals only from observed mixed signals. Particular attention is paid to the physical interpretation of BSS from the acoustical signal processing point of view. Frequency-domain BSS is shown to be equivalent to two sets of frequency domain adaptive microphone arrays, i.e., adaptive beamformers (ABFs). Although BSS can reduce reverberant sounds to some extent in the same way as ABF, it mainly removes the sounds from the jammer direction. This is why BSS has difficulties with long reverberation in the real world. If sources are not "independent," the dependence results in bias noise when obtaining the correct separation filter coefficients. Therefore, the performance of BSS is limited by that of ABF. Although BSS is upper bounded by ABF, BSS has a strong advantage over ABF. BSS can be regarded as an intelligent version of ABF in the sense that it can adapt without any information on the array manifold or the target direction, and sources can be simultaneously active in BSS.

Project Presentations by Taiwanese Speakers

Sub-Project (I): Prosody, Tones and Text-to-speech Synthesis

Sin-Horng Chen (National Chiao Tung University), Chiu-yu Tseng (Academia Sinica), and Yuan-Fu Liao (National Taipei University of Technology)

The main concerns of this subproject are the prosody modeling of Mandarin speech and its applications. Three recent studies are reported in this presentation. They are: (1) An automatic prosody labeling method, (2) Hierarchical modeling of different prosody styles and boundaries, and (3) Applications of prosody modeling - explicit and latent prosodic modeling for speech/speaker recognition. They are summarized as follows.

(1) An automatic prosody labeling method: The new automatic prosody labeling method first introduces four models to describe the relationships of the prosody tags to be labeled (i.e., inter-syllable break type and syllable prosodic state), the prosodic features (i.e., syllable pitch contour and inter-syllable pause duration), and linguistic features of various levels. It then employs a sequential optimization procedure to estimate parameters of these four models and find all prosody labels. Experimental results on the Sinica Tree-Bank corpus showed that most prosodic tags labeled were meaningful and the estimated parameters of these four models matched well with our a priori knowledge about Mandarin prosody.

(2) Hierarchical modeling of different prosody styles and boundaries: The top-down HPG framework (Tseng, 2004, 2005, 2006) found with quantitative evidences that hierarchical constrains from higher level discourse information that forms the necessary cross-phrase semantic cohesion is key to fluent speech prosody. Mandarin prosody is in fact layered and cumulative contributions from lexical (tone), syntactic (intonation) to semantic (discourse) prosody. Cross-phrase intonation modifications are systematic and predictable instead of random

variations. Two further questions are addressed: 1. Could one cross-phrase prosody base form such as HPG generate different prosody styles? 2. Could we explain why post PW (Prosodic Word) boundary breaks varies most in pause duration? Or whether perceived boundary breaks postulated by HPG help understand boundary properties? We found that: 1. Systematic but significantly distinct distribution patterns are found across prosodic layers, thus proving how varied distribution patterns of layered contributions on one base form, i.e., the HPG framework, would be sufficient to generate various output prosody styles; 2. Significant contrasts of acoustic properties of the immediate neighboring prosodic units with or without in-between pauses form the percept of boundaries thus explaining why in fluent speech boundary breaks involve more than simply pause duration, why extremely short pauses are consistently labeled as B3's by humans, and finally why B3's are now predictable.

Sub-Project (II): Robustness in Speech Recognition

Jen-Tzung Chien (National Cheng Kung University)

In this subproject, we propose several works to resolve the robustness issues in the whole procedure of speech recognition. These works are grouped into different hierarchical levels, including signal level, feature level, model level, decoding level and system level. In signal level, we present noise reduction methods based on the running spectrum band-pass filtering and the improved microphone array processing using Griffiths-Jim beamformer. In feature level, we develop different learning criteria for temporal filtering in cepstral domain. High-order normalization of cepstral features is presented for speech recognition. In model level, we exploit maximum entropy (ME) acoustic and linguistic modeling where the mutual dependencies between HMM and n-gram parameters can be properly merged into a unified ME framework. We also present combination method of MLLR for model adaptation. Further, in decoding level, we present an entropy-based weighting of decision function, which comes up with a HMM rescoring scheme. Finally, in system level, we improve LVCSR performance by combining word graphs in hypothesis space that are obtained by different systems and setups. With these approaches in hierarchical levels, we are able to build a robust speech recognition system in many applications.

Sub-Project (III): New Generation Spoken Language Systems

Hsin-min Wang (Academia Sinica) and Lin-shan Lee (National Taiwan University)

As the information technology environment evolved from PC-based era to a new Post PC era with multi-media content and wireless technologies, a whole new generation of spoken language systems will become important. The multi-media network content can be indexed, browsed, and retrieved by its speech information. The users can interact with the network or the search engine via spoken and multimodal dialogues. Therefore the new generation spoken language systems will be featured by information extraction and retrieval, spoken document understanding and organization, spoken and multimodal dialogues, and distributed speech recognition. In this presentation, we will first highlight some of our research achievements in the above research areas, then demonstrate a prototype system, and finally indicate our future directions.

Typical Work in Taiwan outside the Project

Generation of Speaking Styles and Emotional Expressions in Text-to-Speech Synthesis

Chung-Hsien Wu (National Cheng Kung University)

With the progress in speech synthesis techniques, emotion, speaking style, and individuality become the important characteristics for the synthesized speech. Recent research on text-to-speech synthesis has focused on the generation of emotional expressiveness and various speaking styles in synthesized speech. In this talk, first, generation of spoken text with idiolect, speaker-specific usage of words and phrases, for an individual will be discussed. This talk will describe how to convert a written text to the spoken text embedded with the idiolects for text-to-speech (TTS) conversion. In speech generation, corpus-based unit selection approach for emotional speech synthesis will be presented. To have a variety of emotion expressions, current corpus-based methods require a large collection of speech recording for each style. This talk will present a voice conversion method using a small-sized parallel corpus, to convert the source speaker's speech to the target speaker's speech. Finally, some evaluation results and synthesized speech samples will be given.

Posters and Demos by Taiwanese Presenters

Posters:

P1: From One Base Form (HPG) to Multiple Prosody Outputs - Systematic but Dynamic Higher-Level Contributions and Distributions

Chiu-yu Tseng and Zhao-yu Su (Academia Sinica)

We proved from earlier corpus analyses of Mandarin (Tseng et al, 2004; 2005; 2006) that fluent speech prosody contains higher level discourse information above IU that could not be overlooked. Our Hierarchical framework of Prosodic Phrase Grouping (HPG) specifies how layered contributions cumulatively make up cross-phrase output prosody, and accounted for how individual phrase intonations adjust systematically to yield fluent speech prosody. Modular acoustic simulation models correlating to HPG were also constructed for speech synthesis applications. To test whether distribution of contribution from each prosodic layer is fixed or dynamic within and across speakers and speaking styles, and whether the HPG framework serves as a base form on which various styles of fluent speech prosody could be built, we further analyzed read speech of different literary styles from rigid (classical poetry) to semi-rigid (Chinese classics) surface prosodic formats into three categories, namely, 1.regular 2.semi-regular and 3.irregular. The results showed that the distribution of contribution patterns from the PPh (Prosodic Phrase) level to the PG (Prosodic Phrase) level varied systematically with respect to different literary style. The more regular the style is, the bigger the planning templates are used; and the more contributions from higher level information to output prosody. Systematic but significantly distinct distribution patterns are found across prosodic layers, thus proving how varied distribution patterns of layered contributions on one base form, i.e., the HPG framework, would be sufficient to generate various output prosody styles.

P2: Boundaries and Boundary Properties in Mandarin Fluent Speech Prosody - the HPG Perspectives and Implications

Chiu-yu Tseng and Chun-Hsiang Chang (Academia Sinica)

The aim of the present study is to investigate whether (1.) acoustic properties in relation to boundary breaks are also boundary cues in Mandarin fluent speech, (2.) the HPG (Hierarchical Prosody Group) perspectives (Tseng et al, 2004; 2005; 2006) could account for boundary variations found in speech data, and (3.) the results obtained bear any technological implications. Note in our fluent speech corpora COSPRO (Tseng et al, 2005), the annotation design stressed perceived boundary breaks between prosodic units. The rationale being that hand annotated outcomes would enabled us to study the inconsistencies/gaps between human speech perception and the physical data. We found as expected relatively wide distribution of pause durations for all labeled breaks, most notably B3's (from 0 to over 300 msec), and were therefore interested why some very short pauses were consistently judged as a B3 across transcribers. We further analyzed duration and intensity of the immediate neighboring prosodic units PW (Prosodic Word) of B3's in detail, compared them with B2's; and found different corresponding patterns in these acoustic parameters. Thus it became clear that to the human ear, B2's and B3's are distinguished from each other not by the duration of the following pauses alone, but by the duration and intensity patterns of the preceding PW as well. Accordingly, we included factors of duration and intensity to revise and fine-tuned the linear regression model used, and recalculated contribution predictions from the Prosodic Word (PW) layer to final prosody output under our HPG framework (Tseng et all, 2005, 2006). Our results are not only consistent with the actual distribution of breaks, but also improve the total correlation by 3%-10%. Our HPG framework could now better account for layered-and-cumulative contributions that make up output prosody, and further explain why in fluent speech boundary breaks involve more than simply pause duration, why acoustic properties of the immediate neighboring prosodic units are to be included, why extremely short pauses are consistently labeled as B3's by humans, and finally why B3's are now predictable. We believe these significant findings could be directly applied to automatic annotation as well as speech recognition.

P3: Latent Prosody Model of Continuous Mandarin Speech

Chen-Yu Chiang (National Chiao Tung University), Xiao-Dong Wang (University of Tokyo), Yuan-Fu Liao (National Taipei University of Technology), Yih-Ru Wang (National Chiao Tung University), Sin-Horng Chen (National Chiao Tung University), and Keikichi Hirose (University of Tokyo)

The major difficulty of prosody modeling and automatic tone recognition of continuous Mandarin speech is the complex interaction of tones and prosody/intonation on F0 contours. In this study, we propose a latent prosody model (LPM) aiming to jointly model the affections of tone and prosody state on F0. The main purposes are twofold including (1) automatic prosody state labeling and (2) improving tone recognition accuracy. The basic idea is to introduce latent prosody state variables into an additive statistic model of F0 which already considers the affecting factors of tone and speaker. Experiments on the Tree-Bank corpus showed that LPM not only gave meaningful prosody state labeling results but also improved the average tone recognition rate from 80.86% of a multi-layer perceptron (MLP) baseline to 82.55%.

P4: An Automatic Prosody Labeling Method for Mandarin Speech

Chen-Yu Chiang (National Chiao Tung University), Yuan-Fu Liao (National Taipei University of Technology), Yih-Ru Wang (National Chiao Tung University), Sin-Horng Chen (National Chiao Tung University)

Traditionally, prosody modeling is performed using a well-annotated speech corpus with break indices properly labeled manually in advance. However, the labeling process is always labor-intensive and may suffer from the drawback of inconsistency especially if the database is large. In this study, a new automatic prosody labeling method is proposed. It first introduces four models to describe the relationships of the prosody tags to be labeled (i.e., inter-syllable break type and syllable prosodic state), the prosodic features (i.e., syllable pitch contour and inter-syllable pause duration), and linguistic features of various levels. It then employs a sequential optimization procedure to estimate parameters of these four models and find all prosody labels. Experimental results on the Sinica Tree-Bank corpus showed that most prosodic tags labeled were meaningful and the estimated parameters of these four models matched well with our a priori knowledge about Mandarin prosody.

P5: Distributed Speech Recognition (DSR) with Histogram-based Quantization (HQ) and Error Concealment Techniques

Chia-yu Wan and Lin-shan Lee (National Taiwan University)

In this paper, a three-stage error concealment (EC) framework based on the recently proposed Histogram-based Quantization (HQ) for Distributed Speech Recognition (DSR) is proposed, in which noisy input speech is assumed and both the transmission errors and environmental noise are considered jointly. The first stage detects the erroneous feature parameters at both the frame and subvector levels. The second stage then reconstructs the detected erroneous subvectors by MAP estimation, considering the prior speech source statistics, the channel transition probability, and the reliability of the received subvectors. The third stage then considers the uncertainty of the estimated vectors during Viterbi decoding. At each stage, the error concealment (EC) techniques properly exploit the inherent robust nature of Histogram-based Quantization (HQ). Extensive experiments with AURORA 2.0 testing environment and GPRS simulation indicated the proposed framework is able to offer significantly improved performance against a wide variety of environmental noise and transmission error conditions.

P6: Pronunciation Modeling for Spontaneous Speech Recognition Using Latent Pronunciation Analysis (LPA) and Prior Knowledge

Che-Kuang Lin and Lin-Shan Lee (National Taiwan University)

In this paper, we propose a new framework for pronunciation modeling, in which the search algorithm tries to focus primarily on the clearly-pronounced portion of speech, while deemphasizing the observations of the slurred portion. This is based on the prior analysis that the pronunciation variation has to do with the predictability and the importance of the words in the spoken utterances, which may be estimated to some extent. We define a set of pronunciation-related features and develop a Latent Pronunciation Analysis (LPA) to estimate the "latent pronunciation states" in the speech. The LPA probabilities, pronunciation-related features and another set of prior knowledge obtained from two distance measures between phonemes are integrated in a SVM classifier to produce a "pronunciation variation indicator" for each frame, based on which the Viterbi decoding was performed. Very encouraging initial results on Mandarin spontaneous speech were obtained in preliminary experiments.

P7: Speech Enhancement based on Microphone Array and Two-cycle Noise Estimation

Ji-Wen Yang, Hsiao-Chuan Wang (National Tsing Hua University)

This paper presents a noise reduction method based on the beamforming technique on microphone array and the two-cycle noise estimation in post-processor. The microphone array provides the function of adaptive

beamforming so that the interference is suppressed. Following the microphone array, a post-processor performs the speech enhancement to further reduce the additive noise. Techniques of minima controlled recursive averaging (MCRA) and log-spectral amplitude (LSA) estimation are employed for noise estimation and speech enhancement. The MCRA process is repeated twice to perform a two-cycle noise estimation. The experiment shows that this two-stage process is effective in removing the additive noise.

P8: Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environments

Chung-fu Tai and Jehi-weih Hung (National Chi Nan University)

The energy parameter has been widely used as an extension to the basic features of mel-frequency cepstral coefficients (MFCCs) to improve the recognition accuracy in speech recognition. In this paper, a simple and effective approach for energy normalization for silence (non-speech) portions in an utterance is proposed. This approach, named as silence energy normalization (SEN), sets the log-energy of the non-speech frames to be a small constant, while keeps that of other frames unchanged. In the experiments conducted on AURORA2 database, we showed that SEN provides an averaged word error rate reduction of 34.9% and 44.6% for Test Sets A and B when compared with the baseline processing. It was also shown that SEN outperforms similar approaches like energy subtraction (ES) and feature vector selection (FVS). Finally, we showed that SEN can be integrated with cepstral mean and variance normalization (CMVN), to achieve further improved recognition performance.

P9: A Reference Model Weighting-based Method for Robust Speech Recognition

Yuan-Fu Liao (National Taipei University of technology)

In this poster a reference model weighting (RMW) method is proposed for fast hidden Markov model (HMM) adaptation which aims to use only one input test utterance to online estimate the characteristic of the unknown test noisy environment. The idea of RMW is to first collect a set of reference HMMs in the training phase to represent the space of noisy environments, and then synthesize a suitable HMM for the unknown test noisy environment by interpolating the set of reference HMMs. Noisy environment mismatch can hence be efficiently compensated. The proposed method was evaluated on the multi-condition training task of Aurora2 corpus. Experimental results showed that the proposed RMW approach outperformed both the histogram equalization (HEQ) method and the method proposed in European Telecommunications Standards Institute (ETSI) distributed speech recognition (DSR) standard ES 202 212.

P10: Latent Prosody Analysis for Robust Speaker Identification

Yuan-Fu Liao (National Taipei University of technology)

Handsets that are not seen in the training phase (unseen handsets) are significant sources of performance degradation for speaker identification (SID) applications in the telecommunication environment. In this poster, a novel latent prosody analysis (LPA) approach to automatically extract the most discriminative prosody cues for assisting in conventional spectral feature-based SID is proposed. The concept of the LPA approach is to transform the SID problem into a full-text document retrieval-like task via (1) prosodic contour tokenization, (2) latent prosody analysis, and (3) speaker retrieval. Experimental results of the phonetically balanced, read-speech, handset-TIMIT (HTIMIT) database demonstrated that the proposed method of fusing the LPA prosodic feature-based SID systems with maximum likelihood a priori handset knowledge interpolation (ML-AKI) spectral feature-based SID outperformed both the pitch and energy Gaussian mixture model (Pitch-GMM) and the bigram of the prosodic state (Bigram) counterparts for both cases of counting all and only unseen handsets.

P11: Factor Analyzed Subspace Modeling and Selection

Chuan-Wei Ting and Jen-Tzung Chien (National Cheng Kung University)

We will present a novel subspace modeling and selection approach for noisy speech recognition. In subspace modeling, we develop factor analysis (FA) representation of noisy speech, which is a generalization of signal subspace (SS) representation. Using FA, noisy speech is represented through the extracted common factors, factor loading matrix and specific factors. The observation space of noisy speech is accordingly partitioned into a principal subspace containing speech and noise and a minor subspace containing residual speech and residual noise. We minimize the energies of speech distortion in principal subspace as well as minor subspace so as to estimate clean speech with residual information. More attractively, we explore optimal subspace selection via solving hypothesis test problems. We test the equivalence of eigenvalues in minor subspace to select subspace dimension. To fulfill FA spirit, we also examine the hypothesis of uncorrelated specific factors/residual speech.

Subspace can be partitioned according to a consistent confidence towards rejecting null hypothesis. Optimal solutions are realized by likelihood ratio tests, which come up with the approximated chi-square distributions as test statistics.

P12: Maximum Entropy Acoustic and Linguistic Modeling

Chuang-Hua Chueh and Jen-Tzung Chien (National Cheng Kung University)

Traditionally, speech recognition system was established assuming that acoustic and linguistic information sources were independent. Parameters of hidden Markov model (HMM) and n-gram were separately estimated for maximum a posteriori (MAP) classification. However, the classification performance is limited because the speech features and the lexical words are inherently correlated in natural language. In this study, we relax the assumption of model independence and achieve sophisticated acoustic and linguistic modeling for speech recognition. Our idea is to merge acoustic evidence into linguistic parameters or linguistic evidence into acoustic parameters based on maximum entropy (ME) principle. The discriminative ME (DME) models are exploited by inducing features from competing sentences. Furthermore, we directly build a unified ME (UME) model for word posterior probability where model dependence is not only embedded in acoustic parameters but also linguistic parameters. The scaling factor between HMM and n-gram in MAP decoding is implicitly planted. This framework is generalizable to high order feature statistics and word regularities.

P13: Word Topical Mixture Models for Dynamic Language Model Adaptation

Hsuan-Sheng Chiu and Berlin Chen (National Taiwan Normal University)

This research considers dynamic language model adaptation for Mandarin broadcast news recognition. A word topical mixture model (TMM) is proposed to explore the co-occurrence relationship between words, as well as the long-span latent topical information, for language model adaptation. The search history is modeled as a composite word TMM model for predicting the decoded word. The underlying characteristics and different kinds of model structures were extensively investigated, while the performance of word TMM was analyzed and verified by comparison with the conventional probabilistic latent semantic analysis-based language model (PLSALM) and trigger-based language model (TBLM) adaptation approaches. The large vocabulary continuous speech recognition (LVCSR) experiments were conducted on the Mandarin broadcast news collected in Taiwan. Very promising results in perplexity as well as character error rate reductions were initially obtained.

P14: Speaker-and-environment Change Detection in Broadcast News using Maximum Divergence Common Component GMM

Yih-Ru Wang (National Chiao Tung University)

In this paper, the supervised maximum-divergence common component GMM (MD-CCGMM) model was used to the speaker-and-environment change detection in broadcast news signal. In order to discriminate the speaker-and-environment change in broadcast news, the MD-CCGMM signal model will maximize the likelihood of CCGMM signal modeling and the divergence measure of different audio signal segments simultaneously. Performance of the MD-CCGMM model was examined using a four-hour TV broadcast news database. A result of 16.0% Equal Error Rate (EER) was achieved by using the divergence measure of CCGMM model. When using supervised MD-CCGMM model, 14.6% Equal Error Rate can be achieved.

P15: Improved Methods for Characterizing the Alternative Hypothesis Using Minimum Verification Error Training for LLR-Based Speaker Verification

Yi-Hsiang Chao (Academia Sinica), Wei-Ho Tsai (National Taipei University of Technology), and Hsin-min Wang (Academia Sinica)

Speaker verification based on the log-likelihood ratio (LLR) is essentially a task of modeling and testing two hypotheses: the null hypothesis and the alternative hypothesis. Since the alternative hypothesis involves unknown imposters, it is usually hard to characterize a priori. In this paper, we propose a framework to better characterize the alternative hypothesis with the goal of optimally separating client speakers from imposters. The proposed framework is built on either a weighted arithmetic combination or a weighted geometric combination of useful information extracted from a set of pre-trained anti-speaker models. The parameters associated with the combinations are then optimized using Minimum Verification Error training such that both the false acceptance probability and the false rejection probability are minimized. Our experiment results show that the proposed framework outperforms conventional LLR-based approaches.

P16: A Minimum Boundary Error Framework for Automatic Phoneme Segmentation

Jen-Wei Kuo and Hsin-min Wang (Academia Sinica)

This paper presents a novel framework for HMM-based automatic phoneme segmentation that improves the accuracy of placing phoneme boundaries in speech utterances. In the framework, both training and segmentation approaches are proposed according to the minimum boundary error (MBE) criterion, which tries to minimize the expected boundary errors over a set of possible phoneme alignments. This framework is inspired by the recently proposed minimum phone error (MPE) training approach and the minimum Bayes risk decoding algorithm for automatic speech recognition. To evaluate the proposed MBE framework, we conduct automatic phoneme segmentation experiments on the TIMIT acoustic-phonetic continuous speech corpus. MBE segmentation with MBE-trained models can identify 80.53% of human-labeled phoneme boundaries within a tolerance of 10 ms, compared to 71.10% identified by conventional ML segmentation with ML-trained models. Moreover, by using the MBE framework, only 7.15% of automatically labeled phoneme boundaries have errors larger than 20 ms.

P17: Phonetic Transcription Using Speech Recognition Technique Considering Variations in Pronunciation

Min-Siong Liang (Chang Gung University, Taiwan), Ren-Yuan Lyu (Chang Gung University, Taiwan), and Yuang-Chin Chiang (National Tsing Hua University)

We propose a new approach for performing phonetic transcription of speech and text that combines automatic speech recognition (ASR) and grapheme -to- phoneme (G2P) techniques. By augmenting the text with speech and using automatic speech recognition with a sausage searching net constructed from multiple text pronunciations corresponding to human speech utterance, we are able to reduce the effort for phonetic transcription. By using a multiple pronunciation lexicon, a transcription error rate of 12.74% was achieved. Further improvement can be achieved by adapting the pronunciation lexicon with pronunciation variation (PV) rules and an error rate reduction of 17.11% could be achieved.

Demos:

D1: A Mandarin TTS System for Weather Forecast with Integrated HPG Prosody Model

Yong-cheng Chen, Hsin-min Wang, and Chiu-yu Tseng (Academia Sinica)

HPG (Hierarchical Prosody Group) is essential to characterize the prosody for Mandarin fluent speech. Evidence of HPG prosody model has been found both in adjustments of F0 contours and temporal allocations within and across phrases. We demonstrate a Mandarin TTS system for weather forecast with an integrated HPG prosody model for duration, F0, intensity, and break predictions. The database consists of a weather-forecast corpus with 7062 syllables and 1292*3 syllable-tokens chopped off specially designed three-phrase carrier sentences. Since temporal allocations and rhythmic structure in speech flow are carefully dealt with, the TTS system is capable of converting long paragraph text input into natural synthesized speech output.

D2: Mandarin Singing Voice Synthesis with Improved Harmonic-plus-noise Model

Hung-Yan Gu and Huang-Liang Liao (National Taiwan University of Science and Technology)

Here, we study to improve and extend the harmonic plus noise model (HNM) in order to synthesize more natural and clearer Mandarin singing voice. A few improvements are made. For examples, the determination of the maximum voiced frequency (MVF) is now more robust, and the estimation of the parameter values of the harmonics are now more precise. In addition, the model is extended to provide more abilities. For examples, synchronization of phase delay is added to synthesize more natural voice, phoneme duration determination is made with ADSR concept to promote fluency, and vibrato and portamento processing are added to synthesize more expressive singing voice. With the improvements and extensions mentioned, a real-time Mandarin singing voice system has been built now.

D3: SoVideo - A Mandarin Chinese Broadcast News Retrieval System

Hsin-min Wang, Shih-Sian Cheng, Yong-cheng Chen, and Yi-Ting Chen (Academia Sinica)

SoVideo is a Mandarin Chinese broadcast news retrieval system. The system is based on technologies such as large vocabulary continuous speech recognition for Mandarin Chinese, automatic story segmentation, and information retrieval. Currently, the database consists of more than 400 hours of broadcast news, which yielded 10,343 stories by automatic story segmentation.

D4: SoMusic - A Query-by-Singing Music Retrieval System for Karaoke Applications

Hung-Ming Yu (Academia Sinica), Yu-Ren Chien (Academia Sinica), Wei-Ho Tsai (National Taipei University of Technology), and Hsin-min Wang (Academia Sinica)

SoMusic is a query-by-singing music system for karaoke applications. Unlike regular CD music, where the stereo sound involves two audio channels that usually sound the same, karaoke music encompasses two distinct channels in each track: one is a mixture of the lead vocals and background accompaniment, and the other consists of accompaniment only. Although the two audio channels are distinct, the accompaniments in the two channels often resemble each other. We exploit this characteristic to (i) infer the background accompaniment for the lead vocals from the accompaniment-only channel, so that the main melody underlying the lead vocals can be extracted more effectively; and (ii) detect phrase onsets based on the Bayesian Information Criterion (BIC) to predict the onset points of a song where a user's sung query may begin, so that the similarity between the melodies of the query and the song can be examined more efficiently. To further refine extraction of the main melody, we propose correcting potential errors in the estimated sung notes by exploiting a composition characteristic of popular songs whereby the sung notes within a verse or chorus section usually vary no more than two octaves. In addition, to facilitate an efficient and accurate search of a large music database, we employ multiple-pass Dynamic Time Warping (DTW) combined with multiple-level data abstraction (MLDA) to compare the similarities of melodies. Currently, the database consists of 1,071 popular songs.

D5: An Experimental Language Learning System for Pronunciation and Prosody Assessment based on Tutor's Utterance

Chieh-Chih Chang, Chi-Wei Hung, Ke-Shiu Chen, Shi-Han Chen, Hsu-Chih Wu, and Chih-Chung Kuo (Industrial Technology Research Institute)

This is an experimental system for pronunciation and prosody assessment, which features: (a) template-based: assessment based on a demonstrated utterance by a language tutor. (b) Both pronunciation of each phoneme and prosody of each syllable are assessed and scored. (c) The prosody of a learner's utterance can be modified according to the demonstrated utterance (i.e., imitating the tutor's prosody). (d) Either the tutor's or the learner's utterance can be played back in coordination with a speech animated and image-based talking head.

The system is designed as a test-bed for pronunciation learning technology. All phonetic and prosodic parameters are extracted and estimated from both the tutor's and the learner's utterances. An assortment of distance measures and assessment algorithms can be designed and integrated into the system for evaluation. All the values of parameters and results can be easily looked up and displayed for validation.

D6: Web Information Integration and Crowdsourcing for Term Translation and Categorization

Jian-cheng Wu, Tracy Lin, and Jason S. Chang (National Tsing Hua University)

We currently are developing TermMine, a method system for learning to find transliterations of proper nouns on the Web. The purpose is to improve coverage of translation tools and dictionaries in the area of English-Chinese translation as well as transliterations, with application to both a human user and to machine translation or cross-language information retrieval. We use the web as a corpus, and use query-expansion and information retrieval techniques to automatically find desired translation/transliterations embedded in web pages along with their contexts. Though our vocabulary coverage is already a level better than the standard machine translation tools and hand-compiled dictionaries of today, it is still less than ideal. We propose a project building on TermMine and making its results available online as a publicly available vertical search tool upgraded as our method improves, and in addition, combining our tool with collaborative, user-selected and user-provided translations. We have three primary goals. The first goal is to provide up-to-date, highly-accurate, broad-coverage bilingual terminology information through either TermMine augmented with existing sources (bilingual WordNet, NICT Terminology Database), or failing that, through user provided information including terms, term translations, and categories. Second, we then use the large amounts of user data and user feedback as a real-world benchmark to refine and improve our vertical search in ways not possible with traditional approach to bilingual lexicography or statistical machine translation. Third, because TermMine exploits the optimized query expansion to find the term in translation or definition-like contexts, we propose to use these transliterated terms and contexts in conjunction with an ontology built from Wikipedia categories, for terminology management and document clustering.

D7: Network Communication Platform using Distributed Speech Recognition (DSR) frontend, VoIP, Text Chat & Galatea Face Simulation

Yuan-Fu Liao (National Taipei University of technology)

We currently are developing a network communication platform using distributed speech recognition (DSR) frontend, UDP-based VOIP, TCP/IP-based text chat and Galatea face simulation module. In fact, this is the first step to take the advantage of the open-source Galatea toolkit to develop a Mandarin spoken dialog system over the internet.

D8: Speaker Verification-based Security System

Yuan-Fu Liao (National Taipei University of technology)

We recently developed a speaker verification-based security system. It combines the technologies of keyword spotting and speaker verification to give an authorized user a convenient way to quickly (in about 10 seconds) pass security controls.

3. 中日跨國研討會(Taiwan-Japan Joint Workshop)活動照片

- 2007/3/28 參觀活動

新竹地區參訪--參訪工研院電通所

新竹地區參訪--參訪聯發科技公司

新竹地區參訪

新竹地區參訪 – 聯電無塵室

● 2007/3/29-30 研討會會場

中日研討會 Opening – 李琳山教授

中日研討會之 Keynote speech -- Professor Hiroya Fujisaki

中日研討會的 keynote speech -- Professor Sadaoki Furui 的演講

Invited talk -- Professor Keikichi Hirose

Invited talk -- Professor Kiyohiro Shikano

Invited talk -- Professor Yoshinori Sagisaka

Invited talk -- Professor Shoji Makino

Invited talk -- Professor Tatsuya Kawahara

Invited talk -- Professor Shuichi Itahashi

Invited talk -- Professor Shigeki Sagayama

Invited talk -- Dr. Satoshi Nakamura


Panel discussion: What's next?
- Moderators: Professor Hiroya Fujisaki and Professor Lin-shan Lee

日本來賓與李琳山教授合照

會場聽眾

coffee break

午餐

2007/3/30 晚宴

2007/3/31 參觀故宮

其餘詳細資料請見 NeGSST 計畫網頁 http://sovideo.iis.sinica.edu.tw/NeGSST/workshop2007.htm