Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 27-42

可變速中文文字轉語音系統

Variable Speech Rate Mandarin Chinese Text-to-Speech System

江振宇*#、黃啓全*、王逸如*、余秀敏+、陳信宏*

Chen-Yu Chiang, Qi-Quan Huang, Yih-Ru Wang, Hsiu-Min Yu, and Sin-Horng Chen

摘要

本論文描述以隱藏式馬可夫模型爲基礎發展之「可變速中文文字轉語音系統」，訓練語料爲三種不同語速之平行語料，分別對三種語速訓練文脈相關隱藏式馬可夫模型，並利用給予不同語速模型權重值來內插調整語速。另外，從語料庫觀察發現到慢速語音之靜音停頓較多而快速語音較少，傳統以標點符號位置決定靜音停頓的簡單方法，在用於可變速語音合成是不適當的，因此本研究加入預估靜音停頓之機制，對於不同語速分別訓練靜音停頓預估決策樹，再利用調整權重值內插不同語速停頓決策樹機率的方法，達到不同語速下靜音停頓的預估。爲了評估本系統之效能，我們對系統進行客觀測試及主觀測試，在客觀測試中，評量靜音停頓預估之效能及量測合成語音和目標語音的誤差值；在主觀測試中，特別針對隱藏式馬可夫模型權重、靜音停頓決策樹權重以上兩組權重值的組合比較合成語音自然度，實驗結果顯示兩組權重值必須匹配才可合成出較自然的語音。期望以本論文提出方法建構之系統，較傳統單一語速之文字轉語音系統，更適合用於人機互動之中。

*國立交通大學電機工程學系, National Chiao Tung University, Hsinchu, Taiwan
E-mail: gene.cm91g@nctu.edu.tw; cchangwo@yahoo.com.tw; yrwang@cc.nctu.edu.tw; schen@mail.nctu.edu.tw
+中華大學語言中心, Language Center, Chung Hua University, Hsinchu, Taiwan
E-mail: Kuo@chu.edu.tw
#國立台北大學通訊工程學系, National Taipei University, New Taipei City, Taiwan
E-mail: cychiang@mail.ntpu.edu.tw

關鍵詞：文字轉語音系統、中文韻律、語速、停頓預估

Abstract

This paper presents an Hidden Markov Model (HMM)-based variable speech rate Mandarin Chinese text-to-speech (TTS) system. In this system, parameters of spectrum, fundametal frequency and state duration are generated by a context dependent HMM (CDHMM) whose model parameters are linear-interpolated from those of three CDHMMs trained by corpora in three different speech rates (SRs), i.e. fast, medium and slow. In addition, three decision tree (DT)-based pause break predictors trained by using the three SR corpora are used to interpolate the probabilities for inserting pause breaks. The performance of the proposed TTS system were evaluated by several objective and subjective tests. Experimental results suggested that coherence between interpolation weights for CDHMMs and DT-based pasue predictors is crutial for naturalness of the synthesis speech in variable SR. We believe that the proposed variable speech rate Mandarin Chinese TTS system is more suitable than conventional fixed SR TTS systems for applications of human-machine interaction.

Keywords: Text-to-Speech System, Mandarin Prosody, Speech Rate, Break Prediction

1. 緒論

1.1 研究背景、動機

文字轉語音技術在人機界面裡扮演著重要的角色，隨著大型語料庫(corpus-based)以及隱藏式馬可夫模型(HMM-based)爲基礎的文字轉語音技術興起，語音合成的品質較以往進步許多，在許多人機介面應用中已有不錯的表現，然而在不同的應用上會有不同說話速度語音的需求，以達到更有效的溝通。以電話語音訂票系統爲例，對本國籍的互動者來說，一般速度的合成語音可能過慢而浪費時間，這時快速語音就很適合此使用情形，但對於老人或外國人士來說，提供比一般速度稍慢的電話語音，才能讓他們有足夠時間反應聽懂內容，因此，在一些特定的人機互動情境下，傳統只做單一語速之語音合成系統便顯得不夠實用，因此開發不同語速的語音合成系統是一個值得深入探討的議題。

1.2 相關研究

1.2.1 語音合成方法

近期語音合成系統廣爲使用的合成方式主要有兩種，分別是大型語料庫(corpus-based) (Chou *et al*., 2002) 及隱藏式馬可夫模型(HMM-based approach) (Tokuda *et al*., 2000) 的語

音合成方法；大型語料庫合成法由錄製好的語料庫中，挑選適當的語音信號片段串接合成，因此可原音重現，有極佳的合成音質，但是如果要合成出不同特性的語音，如不同講話速度及多種情緒等應用，則須錄製大量的語料作爲挑選單元的基礎，然而欲收集不同特性之語料並不容易，因此，對於合成不同特性語音的應用，單元選取並不是一個適合的方法。

基於隱藏式馬可夫模型語音合成器是一種統計式參數語音合方法，是目前最爲廣泛採用的合成方法，它以文脈相關隱藏式馬可夫模型(Context-dependent HMMs, CDHMMs)來模擬不同語言參數或韻律架構下的聲學信號，從語料庫訓練得到頻譜模型(spectral parameter model)、基頻模型(F0 parameter model)及音長模型(duration model)。欲合成語音時，利用上述訓練好的三種模型，依據輸入文本的語言參數或預估之韻律標記找到適當 CDHMM 模型並串接之，再以特殊的演算法由串接之 CDHMM 參數產生 frame spectrum 及 frame F0 參數，最後將 spectrum 和 f0 參數輸入 MLSA 濾波器(Mel Log Spectrum Approximation filter) (Imai, 1983) 輸出合成出語音訊號。

當想要以現有模型去合成出不同特性的語音訊號，則可利用調整參數的方式達到目的，如內插(interpolation methods) (Yoshimura *et al.*, 2000)、調適(adaptation methods) (Tamura *et al.*, 2001)。跟單元挑選相反的，使用隱藏式馬可夫模型合成器，不需要大量目標的語料，只需要足夠的語料就能利用現有隱藏式馬可夫模型去合成出不同特性的語音信號。

1.2.2 不同語速韻律之研究

研究語音韻律的文獻雖然很多，但是討論相異語速的文獻卻很少，在 (Yu *et al*., 2007) 著作中，作者一開始先利用對話語音的說話速度較快，以及音高軌跡範圍較朗讀式語音爲窄的特性，將朗讀式語音利用 linear regression 的方式轉換成對話語音，另外，對話語音由於說話速度較快，音節的音高軌跡可能會因爲發音不完全導致軌跡不完整，相較於在朗讀式語音完整發音如呈現拋物線的音高軌跡，對話語音變成近似直線，利用相對於朗讀式語音音高軌跡不完整的特性，將朗讀式語音韻律轉換爲對話式語音。

在 (Li & Zu, 2008) 中，作者採用階層式韻律架構的觀念，採用三種不同語速之平行語料庫做分析，實驗對於語速的測量分爲兩類，一爲 speech rate (SR)，定義爲每秒鐘包含停頓時長(pause duration)的發音的音節個數；另一爲 articulation rate (AR)，定義爲每秒鐘的音節個數(不包含 pause duration)。實驗語料庫爲四個漢語文字段落，音節數分別爲 134、123、151 和 34，實驗語料有快、中、慢速的區別，分析了在不同語速下，不同韻律單元的 AR、SR 變化，發現改變說話速度對各韻律階層邊界的 silent pause 是非線性的，語速的快慢會影響基頻軌跡(F0)的平均，發現的現象是快速語料的音高比較高而慢速語料的音高比較低，且其音高軌跡的 dynamic range 比慢速語料小。此篇提出一些不錯的觀點，但是語料庫的資料量不夠大，導致其分析結果不夠一般性是比較可惜的地方。

在 (Tseng, 2008) 中，根據其所提出的階層式多短語韻律句群架構，使用線性回歸統計中的逐步回歸技術(step-wise regression technique)來估算語料，分析出三種不同中文語速之韻律詞、韻律短語和呼吸組層次的時長和音強 pattern，解析出不同語速下，各個層次韻律單元於時長和音強的貢獻，此實驗中平行語料庫之快速語料爲一位台灣男性播音員所發音，中速語料是由一位台灣女性播音員發音，而慢速語料則由北京女性播音員發音。此篇研究提出了不少新的發現，但因其語料庫不是由同一人發音，會導致有些影響實驗結果的因素沒考慮到。

上述這些文獻雖有探討到相異語速的韻律變化，但仍有幾項需要克服的因素，(Yu *et al.*, 2007) 的方法提供了 bottom-up 的方式分析，僅從音節層次討論音高軌跡會忽略到韻律結構上層的影響；至於 (Li & Zu, 2008) 和 (Tseng, 2008) 的階層式韻律架構則提供一個 top-down 的分析方式，對於底層之音節層次分析較缺乏，此外，傳統韻律階層的研究都需要人工事先標記韻律邊界，因此，在文獻 (Chiang *et al.*, 2009) 同時提供 bottom-up 和 top-down 的分析方式，盡可能從各個不同面向討論相異語速語音之韻律變化，採用的自動標記分析方法可以省時省力，同時還可兼顧採用大量語料做研究，分別對不同語速語料的訓練得到韻律模型，藉由分析不同語速之韻律模型參數，探討了不同語速的韻律特性，包含：（1）不同語速音節基頻軌跡之比較、（2）不同語速之 prosodic phrasing、（3）上層韻律單元的 patterns、以及（4）break 和語言參數的關係。此研究是近期對於不同語速韻律較大規模的研究，對於建構可變速語音合成提供了許多實用的資訊。

1.3 系統概述及研究方向

本研究是以 HMM-based 語音合成器爲基礎之「可變速中文文字轉語音系統」，系統架構如圖 1 所示，訓練語料爲一位女專業播音員所錄製的快、中、慢三種語速之平行語料庫，其文本爲中研院 Treebank 3.0 (Huang *et al.*, 2000) 選出之 348 篇短文。本研究先以這三種語速的語音資料庫，各自訓練出不同語速的 HMM-based 語音合成器(包含頻譜及音高 CDHMM 模型及 state duration 模型)，另外，爲了由輸入的文字或語言參數決定音節之間靜音停頓的存在與否，我們分別對三種語速的語音，以決策樹的方法由語言參數預估靜音停頓的插入。爲了達到可變速的語音合成，本研究以調整不同語速之靜音停頓決策樹模型以及 HMM-based 語音合成器參數之權重，可內插出不同語速之合成語音，探討不同語速之靜音停頓決策樹權重和 HMM-based 語音合成器權重關係，找到影響合成可變速語音品質的重要因素。

1.4 漢語多語速語料簡介

本研究所採用的實驗語料庫，是由一位專業的女性播音員讀稿之快速、中速及慢速之文本平行語料庫，此平行語料庫含有 348 個音檔，共有 48035 個音節，其語速及音高統計資訊如表 1 所示，其中 AR 與 SR 的定義同 1.2.2 節。語料庫的錄製順序是在第一梯次先錄中速語速，接下來才將其他兩種速度錄製完成，音檔均爲 20kHz 的取樣頻率及 16-bit 之 PCM 格式，語料庫的錄製文字爲 Sinica Treebank 語料庫中選出的短篇文章，主要內

容大多摘錄自新聞、網路文章、國小教科書等，由數個句子所組成的段落。所有音節的切割標記和基頻軌跡(F0)的偵測均先自動由 Hidden Markov Model Tool Kit(HTK) (Young *et al*., 2006) 和 WaveSurfer (Sjlander & Beskow, 2000) 完成，明顯的參數錯誤再以人工修正，平均每個語句(utterance)音節數爲 138，每個句子 10.37 個字，最短及最長分別爲 80 與 272 個音節。

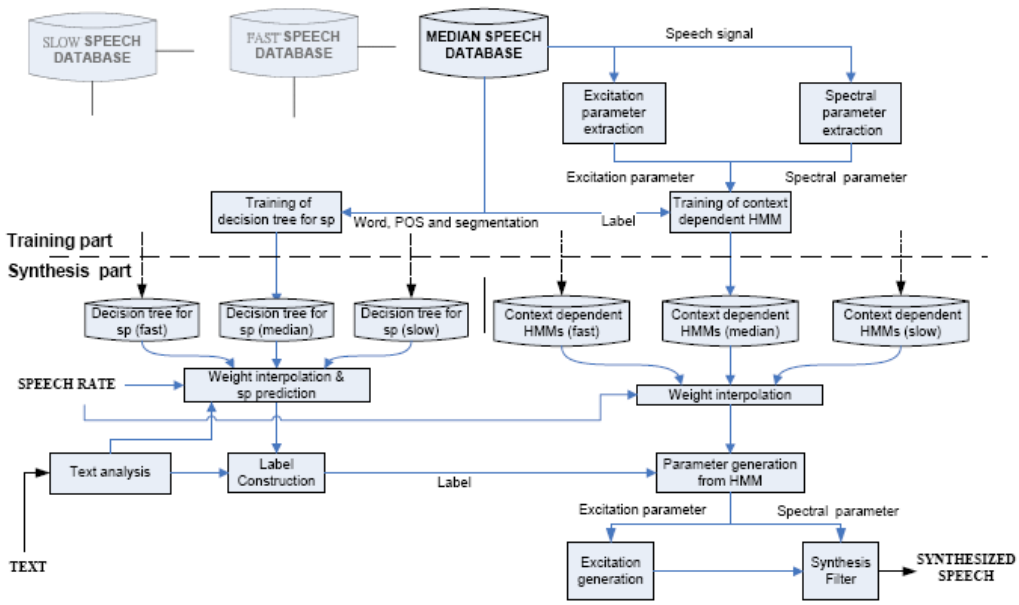

圖 1. 多語速文字轉語音系統之訓練及合成部份

表1. 平行語料庫的平均音長、SRs和ARs

語料庫類型	Fast	Median	Slow
每字平均音長(秒)	0.183	0.241	0.267
SR(syllables/sec)	4.48	3.01	2.47
SR 的變異數	0.082	0.040	0.044
AR(syllables/sec)	5.56	4.19	3.79
AR 的變異數	0.144	0.070	0.065
F0 的平均值(Hz)	201.38	195.88	195.594
F0 的變異數	2489.27	2559.20	2773.37

2. 文字轉語音系統之訓練

本系統是由三種語速的 CDHMM 和靜音停頓決策樹共同加權合成出可變速之語音，我們分別對三種不同語速各自訓練出其 CDHMM 和靜音停頓決策樹，詳細方法如下。

2.1 基於隱藏馬可夫模型之語音合成 (HMM-based Speech Synthesis)

我們將中文聲母、韻母、長靜音（SIL）以及短靜音(SP)模擬成五個狀態的 HMM 模型，也就是將他們模擬成最小的 HMM 訓練單元，對於每個最小單元給予文本標示紀錄其文脈相關資訊，利用由語料求取好的語音聲學參數和文本標示，訓練出文脈相關的頻譜及音高 CDHMM 模型及 state duration 模型。

2.1.1 聲學參數 (Spectral and excitation parameter extraction)

本研究中 CDHMM 模擬的聲學參數爲廣義梅爾倒頻譜係數(Mel-generalized cepstrum, MGC) (Tokuda *et al.*, 1994) 及基頻 (F0)。廣義梅爾倒頻譜係數可藉由調整其 γ 參數，將語音信號頻譜以 all pole (γ=-1)、Cepstrum (γ=0) 或是以廣義的 pole 和 zeros 一起表示 (γ≠-1,0)，亦可調整 α 參數以代表不同的 frequency wrapping，以方便考量人耳的聽覺效應。在本研究中，我們使用 SPTK (SPTK Working Group, 2009) 工具抽取 24 階廣義梅爾倒頻譜係數，設定 γ=0 以及 α=0.5，音檔取樣頻率爲 20kHz，所使用的分析音框爲 25ms（500 個資料點）的漢明窗（Hamming window），音框位移爲 5ms（100 個資料點）。另外，抽取基頻參數則使用 Wavesurfer 工具中的 ESPS 方法求取 (Sjlander and Beskow, 2000)，分析音框大小（window size）爲 7.5ms，而音框位移（window size）爲 5ms。

2.1.2 文本標示 (label)

文本標示提供訓練 CDHMM 及 state duration 的文脈相關語言參數，或在合成時挑選適當的 CDHMM 及 state duration 模型。訓練 CDHMM 時依照文本標示提供的文脈相關資訊對聲學參數作訓練，文本標示的文脈相關參數會影響 HMM 單元本身的頻譜及韻律變化，也會影響 HMM 單元之間連接的狀況，如連音現象、詞首詞尾和句首句尾明顯的音高差異及音節伸長縮短。本系統使用的文脈資訊如表 2：

表2. 文脈相關語言參數

p_{n-1}, p_n, p_{n+1}	Previous(PRE)/current(CUR)/following(FOL) Initial/Final/SP
ST_{n-1}, ST_n, ST_{n+1}	Lexical tones of PRE/CUR/FOL syllable
PW_1 / PW_2	Syllable position in a lexical word (LW) (forward/backward)
PS_1 / PS_2	Syllable position in a sentence (forward/backward)
PM	Punctuation mark after the current syllable
$WL_{n-2}, WL_{n-1}, WL_n, WL_{n+1}, WL_{n+2}$	Lengths of PRE-PRE/PRE/CUR/FOL/FOL-FOL LWs in syllable
$WP_{n-2}, WP_{n-1}, WP_n, WP_{n+1}, WP_{n+2}$	POSs of PRE-PRE/PRE/CUR/FOL/FOL-FOL LWs
SL_{n-1}, SL_n, SL_{n+1}	Lengths of PRE/CUR/FOL sentences in syllable

由於我們將長靜音以及短靜音視爲 HMM 的訓練單元，長靜音就是在音檔開始和結束的靜音部份，而短靜音則定義爲語句中音節間大於 25ms 靜音停頓，所以在文本標示中，對於短靜音也給予文脈相關資訊，在訓練時也會學習到不同文脈相關資訊下的停頓長度。

2.1.3 隱藏馬可夫模型之訓練

文本標示的文脈相關資訊組合相當多，每一種組合都是個別的 CDHMM，在訓練語料不夠充足的情況下，多數組合的 CDHMM 訓練資料量過少，使得訓練出來的模型會不夠準確造成過度訓練（overfitting），因此本研究使用標準的 Tree-based CDHMM 訓練方法 (Zen *et al*., 2007; Yoshimura, 2002)，以決策樹搭配適當的問題集來分群作訓練，以語言學的知識爲基礎設計出合理的問題集，對於某些資料量較少的模型可以合併在一起訓練以增加訓練的資料，如此可訓練出較強健的模型。在合成時，輸入文本標示依據決策樹上每個節點的問題，可找出適當的 CDHMM 串接，進而產生聲學以及韻律參數。以下爲問題集的概述：

- ✧ 依據前一個、現在、後一個聲母或韻母的發音方法、發音位置、送氣不送氣以及清音濁音設定問題集。
- ✧ 依據前一個、現在、後一個音節聲調的調值特性作分類，設定問題集，如一聲和二聲以高調值(H)爲結尾、一聲和四聲以高調值爲開始、二聲和三聲以中(M)或低調(L)值開始。
- ✧ 考慮現在音節所在的詞長和詞的位置，將主要會影響韻律特性的位置和詞長合併，設定爲問題集，如現在音節是否在詞首或詞尾、詞長是否大於四字詞等。
- ✧ 考慮前後及現在詞的詞類，將中研院 46 類詞類依實詞虛詞、八大詞類及其他特殊詞類集合合併，產生問題集。
- ✧ 考慮現在音節所在的句長和句子的位置，將主要會影響韻律特性的位置和句長合併，設定爲問題集，如現在音節是否在句首或句尾、句長是否大於十個字等。

由上列問題集概述的考量，本研究所設定的問題集共約 2100 個左右。

2.2 基於決策樹之停頓預估

由於在訓練時把靜音停頓也視爲一個 CDHMM 來作訓練，其存在與否可由語音切割資訊來決定（短靜音定義爲語句中音節間大於 25ms 靜音停頓），靜音停頓的長度（state duration）可由標準的 Tree-based CDHMM 訓練後得到的決策樹依輸入的文本標示決定。但在合成時，靜音停頓存在與否，只能由文本標示的文脈相關語言參數資訊去預估。本研究分別對於不同語速的語料獨自訓練其靜音停頓決策樹模型，目標爲預估音節間是否有靜音停頓。由不同語速語料靜音停頓的觀察，發現快速語料的靜音停頓較中速少，而中速語料又比慢速少，利用這種語速語靜音停頓多寡的的關係，在合成時將決策樹對於每個音節間是否有靜音停頓求出機率值，即爲有靜音停頓的機率和沒有靜音停頓的機率，分別利

用快中慢的決策樹預估出三組機率，再利用權重值乘以相對應的機率值相加，以達到不同語速下預估靜音停頓的目的。

本研究只考慮詞和詞之間的靜音停頓，假設詞內音節間無靜音停頓，所以預估處理的單元爲詞邊界，所使用的文脈相關資訊如訓練 CDHMM 的文本標示一樣，但問題集只考慮表二之中詞以上的語言參數，而決策樹的分裂條件爲 maximum information gain。

3. 多語速文字轉語音系統

3.1 Text Analysis

文字分析(Text analysis)是文字轉語音系統的第一級，傳統的國語斷詞器使用的是長詞優先及構詞規則，最著名的是中央研究院的中文斷詞系統。但自 2000 年起，由於 conditional random field (CRF)方法 (Lafferty *et al.*, 2001) 被提出，並有效的使用在自然語言處理中的各個問題，都被證實較傳統規則法或其他統計式方法爲佳 (Jiang *et al.*, 2006)。因此，本系統的 Text analysis 的斷詞、base-phrase chunker 及詞類標記部分，便是採用 CRF 的方法做爲核心，其系統架構如下圖 2 所示，其中包含了(1) symbol normalization、(2) word segmentation、(3) POS(part-of-speech) tagger、(4) Word construction、(5) base-phrase chunker 及(6) grapheme to phone 六部分。

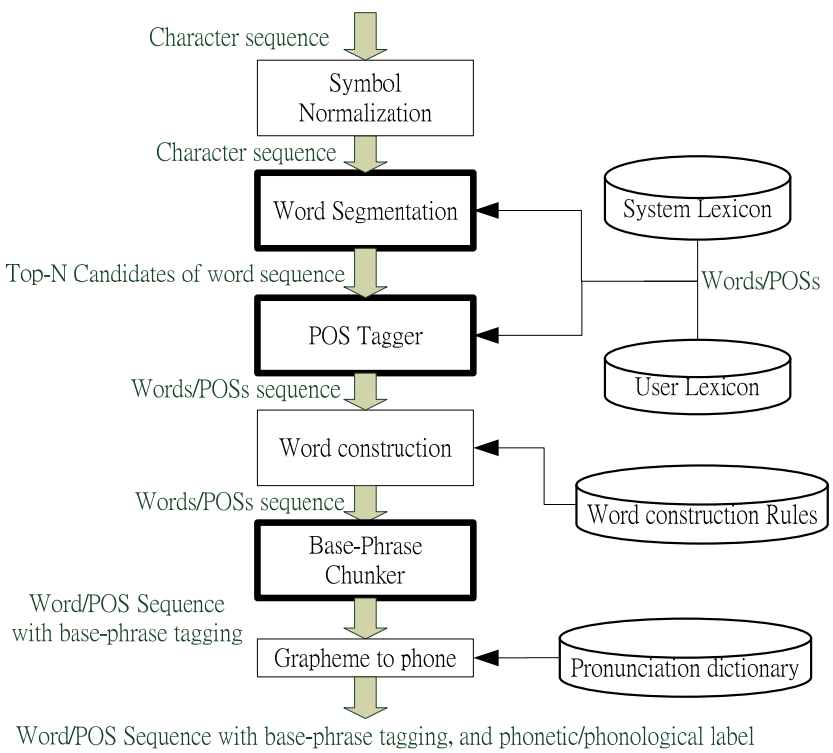

圖 2. Text analysis 之系統方塊圖。

1. Symbol normalization：在此級中將輸入的字串如有 ASCII 的部分，要轉換爲 BIG5，另外，有很多標點符號是屬於同一種標點符號類別，我們將這些同義異形的標點符號正規劃爲其中一種作爲代表。
2. Word segmentation：由於中文文章沒有標示詞的邊界，我們必須將詞的邊界識別出來以得到語音合成需要的語言參數，本系統是以 CRF 以每一個中文字做爲 input feature，要預估的目標爲每個中文字後的標示：{ B1, B2, B3, M, E, S}，其中 B1、B2、B3 分別表示該字位於一個詞的前三字位置，M 代表該字位於詞中第四個字元之後但非詞尾的位置，E 代表字位於詞尾，S 代表單字詞，另外我們也可以使用 user define 的外掛字典輔助斷詞。
3. POS tagger：利用 CRF 以詞、詞對應可能的 POS 爲 input feature，預估每個詞對應到的 POS。
4. Word construction：在這一級我們以規則法，將符合構詞規則的詞由前級斷詞和標示 POS 的結果來構成更具語法和語義的詞，這些詞包括定量複合詞、重複詞等等。
5. Base-Phrase chunker：在這一級我們利用斷出的詞和詞類，以 CRF 將一些基本語法片語標記出來，這些基本語法片語包含 VP：述詞詞組、NP：名詞詞組、GP：方位詞詞、PP：介詞詞組、AP/ADVP：形容詞詞組及副詞詞組。
6. Grapheme to phone：此級爲文字分析器的最後一級，將前級所斷出的詞以一個十二萬詞的發音字典標記上發音和聲調，另外我們也以規則法處理了一些常見的破音字，使其發音和聲調正確。

表 3 爲 word segmentation、POS tagging 以及 base-phrase chunker 效能的評估，其實驗語料的設定皆爲十分之九的訓練及十分之一的測試。由表所示的數據顯示，本文字分析的效能十分優良。

表 3. 文字分析器效能評估

實驗	實驗語料	accuracy	precision	recall	FB1
Word segmentation	Bakeoff-2004	98.30	95.95	96.79	96.37
POS tagging	中央研究院 漢語平衡語料庫	94.73	94.73	94.73	94.73
Base-phrase chunker	中研院 sinica treebank3.0	93.16	92.18	92.27	92.22

3.2 Weight Interpolation

爲了達成多語速合成，本系統具有兩組權重值，一組權重值爲調整預測靜音停頓決策樹的比重，調整此權重影響最大的是 SR，當權重值調成接近慢速，預估的靜音停頓會越來越多，利用決策樹對於每個音節間是否有靜音停頓求出機率值，即爲有靜音停頓的機率和沒有靜音停頓的機率，分別利用快中慢的決策樹預估出三組機率，在利用權重值乘以

相對應的機率值相加，以達到調整權重決定靜音停頓的目的，如下式：

$$sp_n^* = \arg\max_{sp_n} \sum_{i=1}^{3} w_i \times P_i(sp_n \mid L_n) \tag{1}$$

其中 i 爲決策樹的 index (i=1：慢，i=2：中，i=3：快)； w_i 爲第 i 個決策樹模型的權重值； $sp_n \in$ {靜音停頓, 非靜音停頓} 爲第 n 個詞後面的靜音停頓與否； L_n 爲文脈語言資訊； $P_i(sp_n \mid L_n)$ 爲經由文脈語言資訊 (L_n) 組成決策樹問題集後，由第 i 個決策樹結構裡，找尋到對應之葉節點下 (leaf node) 靜音停頓和非靜音停頓的機率值。

而第二組權重值影響著頻譜、音長及基頻，在語料分析中發現語速越快不僅音長變短，基頻也會隨著拉高，直接影響到隱藏式馬可夫模型的參數，很直觀地，當調整權重值越靠近快速語速語音之隱藏式馬可夫模型時，相對於僅使用慢速語音之隱藏式馬可夫模型，每個音節的音長會變短且音頻會提高。以不同權重值內差三種語速之模型參數方法如下式 (Yoshimura *et al*., 2000; Iwano *et al*., 2002)：

$$\boldsymbol{\mu} = \sum_{i=1}^{3} a_i \times \boldsymbol{\mu}_i \tag{2}$$

$$\mathbf{U} = \sum_{i=1}^{3} a_i^2 \times \mathbf{U}_i \tag{3}$$

其中 i 爲 CDHMM 模型的 index(i=1：慢，i=2：中，i=3：快)； a_i 爲第 i 個 CDHMM 模型的權重值， $\boldsymbol{\mu}_i$ 及 $\mathbf{U}_i$ 分別爲 CDHMM state 之 mean vector 及 covariance matrix。

第一組權重值影響靜音停頓的變化，而第二組權重值影響了音長、頻譜及基頻，在自然的語音訊號中，慢速語料靜音停頓較多，音節音長也會拉長，快速語料則相反。在給定權重值也需要按照語速的規則，當想要合成語速較快的語音訊號時，增加快速語速之靜音停頓決策樹的比重，使靜音停頓預估出的數量較少，只在適合的位置給定靜音停頓，同時，我們也調整隱藏式馬可夫模型權重，增加快速語料的比重，而可以產生出較短的音長及較高的基頻，這兩組權重值需要有正相關才會匹配，兩組不匹配的權重值會合成出不自然的語音訊號，因此在不同語速下兩組權重值的匹配是相當重要的，在之後的實驗會對這兩組權重值匹配作主觀測試的實驗。

3.3 Label Construction

欲合成的文字經由文本分析後，可得到對應文脈相關的語言參數資訊，使用之前以內插靜音停頓決策樹所預估之靜音停頓，放入文本標示(label)，最終產生的 label 具有欲合成文本中每個聲母、韻母及靜音停頓的文脈相關語言參數。

3.4 Parameter Generation from HMM

在 label construction 步驟後產生文本標示 (label)，依據文本標示使用三種語速之 CDHMM 模型、state duration 模型及 CDHMM 模型參數權重，由文本相關決策樹找到適當的模型，首先預估出每個聲母、韻母或靜音停頓的長度，再利用 maximum likelihood

法 (Tokuda *et al*., 2000) 產生每個音框的 logF0 及 MGC 頻譜參數。

3.5 Excitation Generation and Synthesis Filter

將上一步得到的每個音框之 logF0 和 MGC 頻譜參數輸入至 MSLA filter (Mel-Log Spectrum Approximation filter) (Imai, 1983)產生合成語音。

4. 實驗結果及討論

實驗語料已於 1.4 中介紹，對於每種語速取其約 328 個語句爲訓練語料，另 20 句爲測試語料，爲了評估本可變速漢語語音合成系統的效能，我們分別對合成語音進行客觀及主觀的測試。在客觀測試方面，我們量測了靜音停頓決策樹的預估正確性，另外，也量測了整個系統合成語音和目標語音的量化誤差。而在主觀測試方面，對系統兩組權重值匹配的狀況作主觀測試的實驗，合成音檔的展示請連結 http://140.113.144.71。

4.1 客觀測試

在第一個實驗中，我們對靜音停頓決策樹的效能進行評估，計算合成音檔和目標語句的靜音停頓預估的正確率及混淆程度，因爲只有單純三種語速的目標語句，沒有實際介於這三種語速的目標語句，所以只有對於三種不同語速目標語句的預估結果作觀察，以合成快速語音爲例，當測試快速語料時，我們調整快速的靜音停頓預估決策樹權重值爲 1，其他語速之權重爲 0，中慢速測試亦同。表 4 爲預估靜音停頓對於快中慢語速的結果。

表4. 不同語速下預估靜音停頓的結果，XX*代表預測爲靜音停頓或非靜音停頓（以百分比表示），Total 爲 Non-SP 或 SP 的總個數。

慢	Inside				Outside		
	Non-SP*	SP*	Total		Non-SP*	SP*	Total
Non-SP	90.05	9.95	28108	Non-SP	89.66	10.34	1885
SP	30.19	69.81	20486	SP	33.57	66.43	1415
中	Inside				Outside		
	Non-SP*	SP*	Total		Non-SP*	SP*	Total
Non-SP	92.77	7.23	29119	Non-SP	91.55	8.45	1977
SP	37.81	62.19	19314	SP	39.61	60.39	1323
快	Inside				Outside		
	Non-SP*	SP*	Total		Non-SP*	SP*	Total
Non-SP	96.34	3.66	35380	Non-SP	94.83	5.17	2496
SP	49.5	50.5	11613	SP	52.74	47.26	804

由實驗結果發現，對於快速合成語音預估靜音停頓的結果是最差的，錯誤大多是在預測目標語句有靜音停頓的部份，主要原因可能是因爲快速語料裡音節間的靜音停頓較少，所以造成了決策樹學習到音節間無靜音停頓的機率較大，在預測結果也是偏向沒有靜音停頓，另外可能的原因，是考慮到快速語料不論是 AR 和 SR 變化都是最大的，語句和語句間語速有較大的差異，因爲語速和靜音停頓的多寡有關係，語速的差異潛在會造成快速語料靜音停頓預估上的困難。在慢速語料上雖然在非靜音停頓預測上略輸快速語料，但在有靜音停頓預測上比快速語料準得多，可能是因爲語者於朗讀慢速語料時，會將詞或韻律詞的結構清楚念出，所以在語料上產生較一致性的靜音停頓，較容易從語言參數學習到規則，因此準確度比快速要高的多。

第二個客觀測試，我們分別測量合成和目標語句其基頻、停頓靜音的長度以及音節的長度的誤差，使用均方根誤差（Root Mean Square Error, RMSE）用來評估誤差值，因語料庫只有三種語速，所以在測量快速語料的 RMSE 時，預測靜音停頓決策樹的權重和隱藏馬可夫式模型的權重，均設定快速權重值爲 1，其他權重設爲 0，中慢速語料也使用同樣的方法測量，表 5 爲實驗結果。

由整體來看 Inside test 的 RMSE 都低於 Outside test 這是因爲過度訓練（overfitting）的關係，經觀察發現靜音停頓音長的預測不論 Outside test 和 Inside test 在語速快的 RMSE 均爲最低，由於快速語速在靜音停頓的音長並不長，就算靜音停頓沒有正確預估出來，誤差也不會太大，而慢速的靜音停頓就不一樣，靜音停頓音長較長，沒有正確預估到靜音停頓誤差就會較大，我們觀察音節音長的 RMSE 也看到同樣的結果，在語速快的音節音長 RMSE 均爲最低，因爲快速語料音節音長都較短，計算誤差也不會太大。

表 5. 快中慢語料作測試之 RMSE 值

測試項目	語速		
	Fast	Median	Slow
Inside F0 (Hz)	36.28	34.38	35.21
Outside F0 (Hz)	42.66	42.78	45.23
Inside sp duration (ms)	44.97	64.19	84.17
Outside sp duration (ms)	56.55	60.02	85.55
Inside syllable duration (ms)	37.53	41.44	44.19
Outside syllable duration (ms)	39.23	42.66	47.08

4.2 主觀測試

主觀測試目的爲測試系統兩組權重值不同的組合，以主觀測試判別合成語音的自然度，對於快中慢兩組權重值設爲：1-0-0、0-1-0、0-0-1、0-0.5-0.5（x-x-x 中的 x 順序代表慢速、中速、以及快速權重值），因爲考慮的組合數量過多，而且慢速跟中速語料依據 SR、

AR 以及基頻的統計差異並不大，因此不考慮 0.5-0.5-0 這個權重值組合，所以本實驗只有 16 種靜音停頓-CDHMM 權重的組合。

每一個合成文本均爲 outside test 的語句，一個文本依據兩組權重值變化會產生 16 種不同語速變化的合成音檔，各分爲四組作測試，以同樣隱藏馬可夫模型的權重值爲同一組，目的爲固定一組權重值，觀察不同權重預測靜音停頓的匹配程度。主觀測試中語音自然度的評分爲五分制，分數爲一至五，一爲最不自然，五爲最自然，總共對 6 人作主觀測試，每個測試者由九句文本中選聽兩句文本的語句，其中一句文本與另一個測試者重複，因爲每文本有 16 種語速權重組合，所以每個人聽 32 句測試語句，整個測試語句共有 192 句，測試結果如表 6 所示。

表6. 主觀測試的平均值±一個標準差，x-x-x中的x順序代表慢、中、快權重值

預測靜音停頓權重值	隱藏馬可夫權重值			
	1-0-0	0-1-0	0-0.5-0.5	0-0-1
1-0-0	2.33±0.61	3.08±1.36	2.79±0.98	2.21±0.70
0-1-0	2.54±0.88	3.38±0.96	3.25±0.391	2.21±0.52
0-0.5-0.5	2.67±0.60	3.08±0.99	3.67±0.79	2.54±1.43
0-0-1	2.83±0.88	2.88±1.00	3.71±0.93	3.25±1.66

由主觀實驗發現，兩組權重值必須有正相關的關係，合成出的語音才會自然，當兩組權重值不相匹配的時候，合成出的語音大多不自然，因此通常在表六對角線附近會有最大的自然度，但當靜音停頓決策樹權重值爲 0-0-1 和隱藏馬可夫權重值爲 1-0-0 是比較令人訝異的結果，猜測在隱藏馬可夫權重值爲 1-0-0 時語速很慢，造成測試者聽得厭煩，由表六固定隱藏馬可夫權重值 1-0-0 觀察，發現無論靜音停頓決策樹權重如何調整，受測者所給予的自然度都偏低，在這種權重值組合下，受測者覺得厭煩分數都給的較低。

5. 結論與未來方向

本系統爲可變速中文文字轉語音，經由權重值調整所合成出的語音，合成出來的語音基本上尚佳，在主觀實驗中兩組權重值皆調爲 0-0.5-0.5 所合成出的語音自然度也是令人滿意的，其語速介於中速及快速之間，系統可預測出適當的靜音停頓、頻譜及其他韻律參數，達到合成出不同語速的自然語音。由客觀實驗發現快速的靜音停頓預估結果較差，未來的研究會以慢速爲基準預估其他語速的靜音停頓，因爲在慢速時讀稿人會完整分析詞和韻律詞結構後念出語句，考慮相對靜音停頓的變化，如某些音節間或詞間不管在慢速還是快速都需要靜音停頓，而有些靜音停頓在快速時反而消失，考慮這些相對的變化再進一步進行靜音停頓預估是需要的。

靜音停頓決策樹是由語言參數預估詞之間靜音停頓的出現與否，雖然本系統所預估的靜音停頓結果尚佳，但以這個系統所預估出來靜音停頓特性並沒有考慮實際靜音停頓

的長度，是這個預估靜音停頓系統的重大盲點，要改進此靜音停頓預估系統，必須分析靜音停頓長度隨語速變化的特性，依照這些特性設計出更適合的預估系統。

Reference

Chou, F.-C., Tseng, C.-Y., & Lee, L.-S. (2002). A Set of Corpus-Based Text to Speech Synthesis Technologies for Mandarin Chinese. *IEEE Trans. on Speech and Audoio Processing*, 10(7) , 481-494.

Chiang, C.-Y, Tang, C.-C., Yu, H.-M., Wang, Y.-R., & Chen, S.-H. (2009). An Investigation on the Mandarin Prosody of a Parallel Multi-Speaking Rate Speech Corpus. In *Proc. of Oriental COCOSDA 2009*, 148-153.

Huang, C.-R., Chen, K.-J., Chen, F.-Y., Gao, Z.-M., & Chen, K.-Y. (2000). Sinica Treebank: Design criteria, annotation guidelines, and pn-line interface. In *Proc. of the Second Chinese Language Processing Workshop 2000*, 29-37.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., & Tokuda, K. (2007) The HMM-based speech synthesis system version 2.0. In *Proc. 6th ISCA Workshop Speech Synth.*, 294-299.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. of ICASSP*, 93-96.

Iwano, K., Yamada, M., Togawa, T., & Furui, S. (2002). Speech-rate-variable HMM-based Japanese TTS system. In *Proc. of IEEE TTS Workshop 2002*, 219-222.

Jiang, W., Guan, Y., & Wang, X.-L. (2006) Conditional Random Fields Based Label Sequence and Information Feedback. *Lecture Notes in Computer Science of Natural Language Processing and Expert Systems*, (4114), 677-689.

Li, A.-J., & Zu, Y.-Q. (2008). Speaking Rate Effects on Discourse Prosody in Standard Chinese. In *Proc. of the Speech Prosody2008*, 449-452.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of ICML*, 282-289.

Sjlander, K. & Beskow, J. (2000). Wavesurfer - an open source speech tool. In *Proc. of the ICSLP 2000*, 4, 464-467.

SPTK Working Group. (2009). Reference Manual for Speech Signal Processing Toolkit Ver 3.3. available at http://sp-tk.sourceforge.net/

Tokuda, K., Kobayashi, T., Masuko, T. & Imai, S. (1994). Mel- generalized cepstral analysis - A unified approach to speech spectral estimation. In *Proc. of ICSLP'94*, 1043-1046

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-Based speech synthesis. In *Proc. of ICASSP*, 1315-1318.

Tamura, M., Masuko, T., Tokuda, K., & Kobayashi, T. (2001). Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proc of ICASSP*, 805-808.

Tseng, C.-Y. (2008). Corpus Phonetic Investigations of Discourse Prosody and Higher Level Information. Language and Linguistics, Institute of Linguistics, 9(3).

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., & Kitamura, T. (2000). Speaker interpolation for HMM-based speech synthesis system. *J. Acoust. Soc. Jpn.* (E), 21(4), 199-206.

Yoshimura, T. (2002) Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based text-to-speech systems. *Ph.D thesis, Nagoya Institute of Technology*.

Yu, J., Huang, L.-X., Tao, J.-H., & Wang, X. (2007). Modeling Incompletion Phenomenon in Mandarin Dialog Prosody. *In Proc. of the Interspeech2007*, 462-465.

Young, S. J., Evermann, G., Gales, M. J. F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., & Woodland, P. C. (2006). The HTK Book, version 3.4. Cambridge University Engineering Department, Cambridge, UK.

中央研究院中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/, last visit 2009/09/09