第一章　緒論

1.1 語音合成簡介

在科技發達的現在，以語音作為資訊傳遞的媒介是最方便、最為人性化的，利用聲音當作介面來和機器溝通，不僅方便使用者操作，相較於傳統的文字介面也更來的親切，因此語音技術的發展是十分值得研究的。

語音技術又分為語音辨識與語音合成，語音合成是將電腦中的文字轉換成語音輸出（Text-to-Speech）的技術，是現今越來越普遍的應用。隨著高科技的蓬勃發展，目前大部分電腦皆具備良好的運算能力，若能搭配語音辨識與語音合成技術，人類便能達到直接地與電腦溝通的目標。

目前在語音合成的應用方面非常廣泛。對一般使用者而言，倘若我們可以提供一套語音合成系統，讓使用者只需藉由聽覺來獲取所需的資訊，如此便可以提高使用者的便利性；相對於視障朋友而言，電腦回應的訊息全轉為語音合成輸出，能使他們和正常人一般無障礙地使用電腦。另外，客服電話、翻譯機、電腦教學輔助系統、電腦有聲書、有聲玩具等也都可以配合語音合成的技術來提高使用產品的方便性。

1.2 語音合成器的架構

語音合成的技術中，文字轉語音（Text-To-Speech）合成系統包含四個部份：文句分析（Text Analyzer）、語音單元選取（Synthesis Units Selection）、韻律參數的產生（Prosody Parameter Generation）、與語音合成器（Speech Synthesizer）。如圖 1 所示，並分述如下：

（一）文句分析：將輸入的文字，經過分析之後，得到文字的相關資訊。包含斷詞與語言學上的特徵標記，例如聲調、聲母類別、韻母類別、詞性等等。

（二）語音單元選取：根據輸入的文句，從資料庫中選擇出欲合成的語音單元音檔。

（三）韻律參數的產生：利用文句分析得到的資訊，來估算最適合的韻律參數，包含音長（Duration）、基週頻率軌跡（Pitch Contour）、音量（Amplitude）等。這些韻律參數對於文字轉語音系統所合成出來的語音品質與自然度，有很大的影響。

（四）語音合成：利用估測的韻律參數，進行韻律的調整以合成出對應的語音訊號波形。對一般語音合成系統而言，合成基本單元可以有音素、雙音素、音節、詞、片語等。本論文則是以音節為基本合成單元。

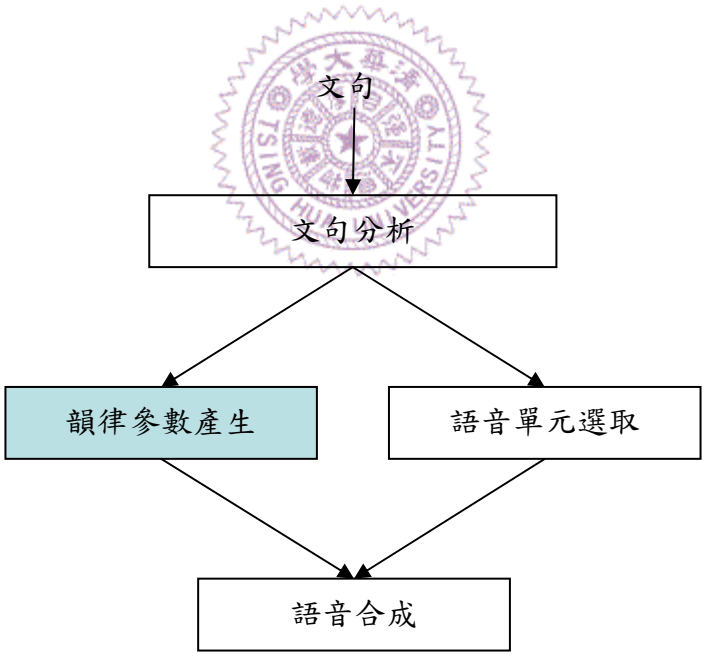

圖 1　語音合成器的架構

本論文的研究重點在於如何產生適當的韻律參數，在語料庫較小的條件下，能預測出最符合自然度的韻律參數，以期達到和一般大量語料庫為基礎的語音合成系統有同樣甚至更好的自然度。

1.3 相關研究簡介

語音合成器之韻律參數的產生，是語音合成中影響到合成音質與自然度，最為關鍵的一階段，可以分為規則法（Rule-Based）與統計法（Data-Driven）兩種取得方式。

（一）規則法：是經由專家對大量文句做語言學上的分析，定義條列式的規則，利用這些規則取得韻律參數，常用的方式是藉由決策樹（Decision Tree）來定義規則。中文的語音合成器使用規則法定義韻律參數的有：台灣大學李琳山教授[1]、長庚大學呂仁園教授的台語文字轉語音（語音合成）系統[2]等。

（二）統計法：則是使用統計的方式，針對訓練文句集的韻律參數做統計訓練，得到統計模型，再依據此統計模型，對未知韻律參數的文句，預測機率上最可能的韻律參數值。在中文語音合成器上目前有幾種作法：

i. 隱藏式馬可夫模型（Hidden Marcov Model, HMM），例如：台灣科技大學古鴻炎教授的產生豐富音色之國語音節信號合成方法[3]。

ii. 類神經網路，例如：交通大學陳信宏教授的盲用電腦之國語單詞輸入及語音輸出系統是使用遞迴類神經網路 RNN 模型[4]、台灣科技大學古鴻炎教授的 ANN 模型[5]等。

iii. 基於 EM（Expectation Maximum）演算法的韻律參數模型設計，例如：交通大學陳信宏教授近幾年的研究[6][7]。

1.4 本論文的研究方法

本實驗室發展的語音合成系統，原先是以現今中文語音合成器發展的主流—

—大量語料庫（Corpus Based）為設計。我們發現大量語料庫有幾個缺點：

- 所需語料多，語料收集不易
- 單元選取的不一致，造成接合片段的不連續

於是我們改以承載式語料庫（Carrier Sentence）[8]設計，以減輕收集大量語料庫的負擔。然而，由於承載式語料庫並沒有非常足夠的合成單元可供選取，在串接每個合成單元輸出之前，改變每個單元的韻律參數經常是必要的。於是，架構於本實驗室目前發展的承載式語料庫語音合成系統，並搭配一個適當的韻律產生器，將能兼顧語音的合成品質與自然度。我們的中文語音合成系統流程與架構圖如下圖 2。

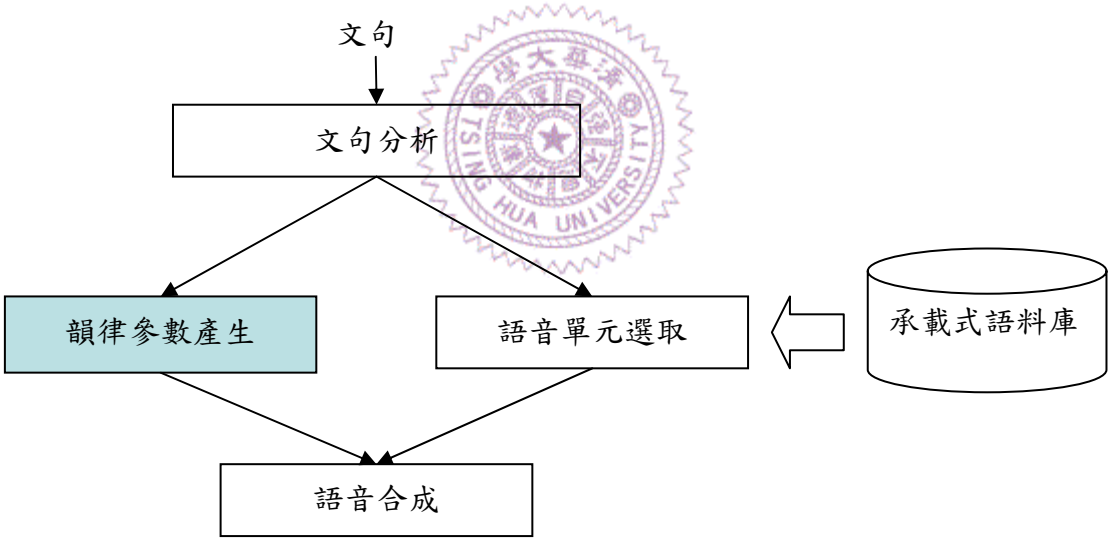

圖 2　本論文研究的中文語音合成系統架構

在本論文中，我們探討了幾種改善韻律參數值的方法。除了利用常見的類神經網路的倒傳遞網路（Back Propagation Network, BPN）[9]來預測韻律參數值，同時也比較了線性迴歸（Linear Regression）[10]與支撐向量機（Support Vector Machine）[11]，並與台科大古鴻炎教授和交通大學陳信宏教授的類神經網路的作法做比較。

我們訓練的韻律參數包含音長（Duration）、基週頻率軌跡（Pitch Contour）、與音量大小（Amplitude）。至於韻律參數中停頓長（Pause）的處理，則維持使用原先系統的規則法定義。這主要是因為，即使是同一個文句，同一個語者，停頓長的變動性仍然很大，所以我們採用規則法定義停頓長模型。

本論文所採用的合成單元為國語單音節，合成方式是在時域上處理，利用基頻同步累加法（Pitch Synchronous Overlap and Add, PSOLA）[12]來調整基週頻率，用波形相似性疊加法（Overlap-add Technique Based on Waveform Similarity, WSOLA）[13]來調整音長。

1.5 章節概要

本論文第二章是研究的相關背景介紹，包含本實驗室發展之語音合成系統的語料庫與合成方式簡介，以及三種常見的迴歸統計方法，如線性迴歸法、類神經網路、與支撐向量機。第三章將介紹如何使用上述三種迴歸方法建構韻律參數模型，並詳述說明這些模型的參數使用。第四章則是展示各種韻律產生器的誤差值評估以及聽測實驗的結果。第五章提出本文的結論以及未來展望。