Study for Implementing Chinese Text-to-Speech System on Internet

黃豐隆 賴彥愷 謝昇仲

國立聯合大學 資訊工程學系
苗栗市聯大 1 號
flhuang@nuu.edu.tw

摘要—本論文探討建置網際網路環境裏之中文語音合成系統。此合成系統具有四個模組，包含：文句分析、韻律訊息分析、合成單元選取與語音合成等，並建立相關之語料庫與語音庫，目前已完成線上測試階段。在本系統中，我們錄製中文單字之基本語音合成單元，以及不同時長之靜音檔以配合韻律參數作合成之用。為改善所錄製合成單元之能量差異，對所有單元進行能量正規化處理。由12位大學生作聽測實驗，進行合成語音品質之滿意度分析，包含能量與清晰度二項。實驗結果顯示，依10刻度之分數評量，線上輸出之合成語音音量正規化與清晰度滿意度分別為8.09與7.63。依合成之結果，我們已可在線上產生具有清晰、順暢之韻律訊息真人的合成語音。[1]

關鍵字：語音合成、韻律模組、串接合成法
連音變調，能量正規化。

Abstract: This paper aims at the implementation for Text-to-Speech (TTS) System on Internet. Our system is composed of four components as follows: Text analysis, Prosody prediction, Selection of speech units, Speech generation module and several corpora and speech database are generated. More than 4000 monosyllabic speech units of Chinese and several silences with various durations have been recorded as basic unit for speech synthesis. Twelve college students with frequent speaking capacity evaluated the synthesized speech for energy and speech quality. The results for the testing achieve 8.09 and 7.63. Based on the comprehensive evaluation, it is obvious that our system can provide real synthesis speech with frequent, prosodic and natural quality on Internet.

Keyword: Synthesis Speech, Prosody Module,
Concatenation, Tone Sandhi, Energy Normalization.

[1] 由國科會部份經費贊助，計畫 97-2815-C-239 -017 –E，特此致謝。

一、簡介

1.1 語音處理簡介

隨著資訊科技的發展，語音技術之應用日漸廣泛，資訊系統語音化服務已經成為一種趨勢。語音處理(Speech Processing)可分為二大類別，語音辨識（Speech Recognition）與語音合成（Speech Synthesis)[9][14]。前者係以語音(Speech)輸入，經辨識後輸出其對應之文字(Text)；反之，後者則以輸入一段文字或一篇文章，再輸出具有抑揚頓挫之正確人聲語音為目的[3][7]。

語音合成系統亦稱為文字轉語音（Text-to-Speech, TTS）系統，其應用十分廣泛，如人機介面設計、電子有聲書、語言翻譯機、104查號台、語音播報與多國語言(Multilingual)翻譯[1]等。TTS系統與網際網路結合後，可使用於網頁之語言與語音處理，進而可使用在Web系統多國語言之翻譯領域上。透過語音辨識與合成的技術，可經由電話語音查詢最新新聞與生活上各類的即時資訊，如各類的文件內容，含電子郵件、行事曆或網頁等內容均可由電腦說出來，我們可聽到具有清晰、順暢之韻律訊息且正確的自然語音 [6][15]。

1.2 語音合成技術之發展

在語音合成系統中，目前較常見語音之產生技術有二類:

1)波形拼接法(Formant Synthesis)：

同步疊加法(Pitch Synchronous Overlap and Add，PSOLA)在時域(Time Domain）上調整語音波形以產生合成語音[10]，此法可改變語音之音長(Duration)與音高(Pitch)韻律訊息，改良頻域(Frequency Domain)處理耗時的缺點。由於PSOLA可調整語音之音高(Pitch)，音高之變化正可以呈現語音之聲調，因此我們只需錄較少的合成單元，減少錄製成本與後製處理時間，惟合成之語音有部份之雜訊(Burst Noise)與較重的鼻

音，使合成語音比較不自然且機器合成的味道較重，且其清晰度亦較低，這些是PSOLA主要缺點。以中文而言，只需錄製中文408個第1聲之基本語音單元，可經由PSOLA調變其音高值產生另外2、3、4與輕聲語音，作為後續處理語音合成選用之基本單元。

2)串接合成法(Waveform Concatenation)：

主要是利用預先錄製好之合成單元(Synthesis Units)，存放在語音資料庫中，經選用之語音單元將其拼接起來，合成出所要的語音。通常合成單元要包含所有可能的發音種類，這些預錄的單元可以是音素(Phoneme)、雙音素與中文音節(Syllable)等。中文是一種具有聲調之語言，每一單音有不同聲調，需事先錄製，以便作為後續之合成使用。

相對於像過去電腦系統產生較單調與不自然語音，這種方法直接選用人們所發出的語音作為合成單元，語音訊號之處理比較單純，聽起來較具有人們講話的特性，語音輸出具有親和性，處理時間相對較短且效果好。相對而言，採用串接合成法時，需要錄製全部所需之語音單元，例如，中文408種單音節其5種聲調均需錄製，約2000多個語音(含男女生計有4000個合成單元)。此外，常用的中文詞約有數萬個，可以個別錄製，上述語音經切音後存入語音庫中，經由語音選取模組選取合成單元，合成時不作音高與時長處理。系統合成之語音不會有鼻音，具有真人發聲之自然與親切效果，因此目前語音合成系統大都採用串接合成。

近年來，上述之串接合成法已應用至不少中文合成系統，如：中興大學之語音系統[3]、微軟亞洲公司之木蘭(MULAN)雙語系統與大陸之訊飛中文語音系統[4]，合成效果不錯。因此，本系統之合成技術亦採用串接合成法。

1.3 中文之特性

中文語音含有聲調(Tone)特徵，計有408種不同基本音，每一基本音含有五種聲調變化，即：1、2、3、4聲與輕聲，且均為單音節(Mono-syllable)，計有近1400種不同聲調的語音。除了中文具有聲調外，其它如；客家（Hakka）語言之四縣腔有6種、海陸腔有7種聲調，此外台語(Taiwanese，閩南語)則有8種聲調。依趙元任5等級音高之分法，中文語音之聲調音高軌跡走勢參見圖1所示。

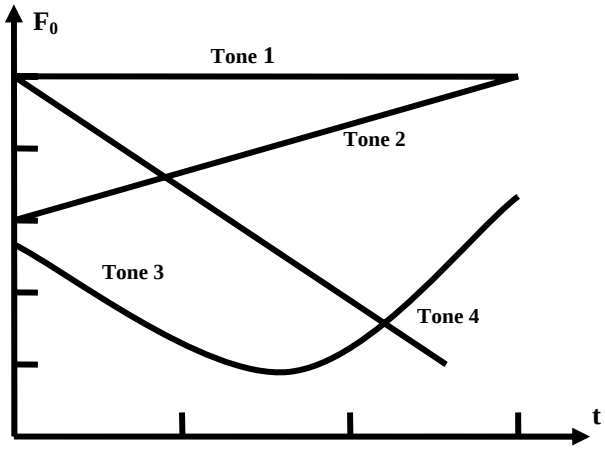

圖1: 中文4個聲調之音高軌跡變化

中文語義最基本的單位是中文詞(Word)而不是字(Character)。因此，另一中文的特性為斷詞(Word Segmentation)處理。英文文句之單字間已有空格(white space)，每一個詞彙之組成明確，中文字與字之間並無空格存在，因此，文句之正確語彙與語義需經電腦之斷詞處理方能獲得。不同的斷詞將有一樣的語義，斷詞結果亦是韻律訊息中停頓位置的重要參數，因此中文自然語言處理中需要斷詞處理。以文句「傳染病毒害正在擴大」為例，下列二種斷詞結果(E1與E2)，不同詞彙表示不同語義(底線表斷詞之中文詞)。例句(E3)是經由中研院中文斷詞系統[2]處理之結果，每一詞彙後均附上詞性(Part of Speech, POS)，這三句斷詞以E3結果為正確。此外，對人名、機構名稱、日期與量之語詞等，中文詞典不易包含各種可能之組合，斷詞之結果常需要進一步作構詞處理。

傳染 病毒 害 正 在 擴大 (E1)
傳染病 毒 害 正在 擴大 (E2)
傳染病(Na) 毒害(VC) 正在(D) 擴大(VC) (E3)

中文之斷詞常用的方法有三種，即：統計法[15]、規則法[9]與混合式斷詞。前者經由語料庫(Corpus)的資料來歸納語言特徵，運用數學模式依機率統計值來決定斷詞結果，而規則法通常需配合詞庫或辭典(Dictionary)作比對，較具代表的方法是「長詞優先法(Max. Matching Method)」，此項技術在於將語言學知識轉換為有效的規則。第三種則運用前二種方法，配合馬可夫機率模式(Hidden Markov Model, HMM)或樹庫(Treebank)作為斷詞依據。

本文第二節說明我們的語音合成系統與使用模組，第三節說明線上系統架構，第四節說明合成語

音品質之實驗結果與分析，最後一節為結論與未來的研究方向。

二、語音合成系統之建置

本研究的重點是，在網際網路(Internet)上建置一個文中之語音合成系統，使用者輸入正體字之文句後，輸出具有自然、順暢之中文語音，經由文句分析模組處理可產生中文文句之各類對映之標註符號(Transliteration)，如：漢語拼音、注音二式，並轉換出其對映之簡體字(Simplified Chinese)，可提供國際人士或國人學習中文之環境，作為線上學習(E-learning)平台。

2.1合成系統模組與資料庫

我們所建置的中文語音合成系統包含下列四個模組：

A)文句分析(Text Analysis)：

B)韻律預估(Prosody Prediction)：

C) 選取合成單元 (Selection of speech Units) ：

D)語音合成(Speech Generation)：

參見圖2所示，輸入文句後依續經各模組之分析號、處理，最後產生合成之語音。

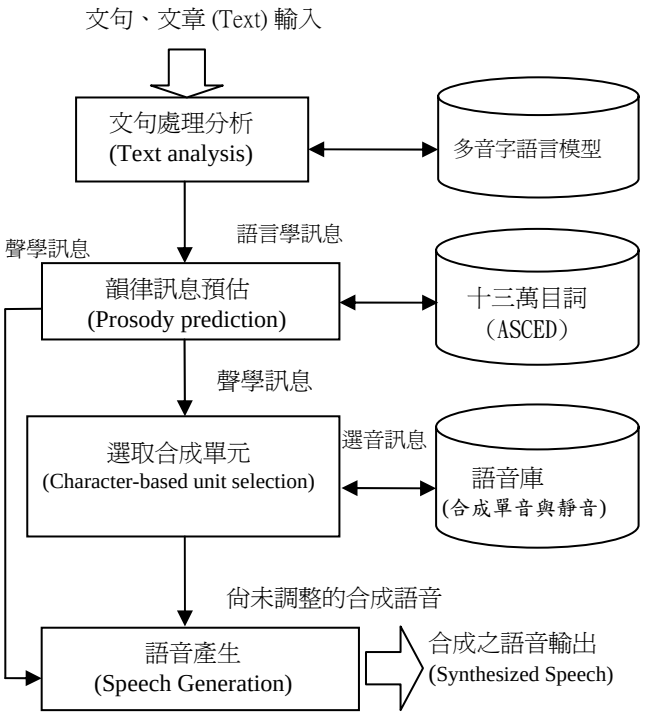

圖 2: 中文 TTS 系統架構

2.2 各模組功能簡述

對於TTS系統而言，所輸入的文字或文章，文句並無任何聲學的特性，如說話的聲調、停頓、發音長短等韻律資訊，只有語言學(Linguistics)資訊。在合成系統中，需依據文句內所包含之語彙與上下文作為預測相關聲學的資訊，如此方能產生自然的合成語音。

1)**文句分析(Text Analysis)**：

中文的常用與次常用字約有近1萬個字，其中一個音可能有多種對映的中文字(一音多字)，而一個字亦可能有二種以上的發音(一字多音)，此即「**中文多音字**」。如：中文字「**中**」具有二種不同發音:「ㄓㄨㄥ」與「ㄓㄨㄥˋ」，中文字「**的**」則有三種發音:「ㄉㄧˊ」、「ㄉㄧˋ」與「ㄉㄜ˙」。經統計約有1000個中文之多音字，一個字至多有8種不同的發音，這是語言岐異(Ambiguity)現象之一。如何有效解決多音字岐異問題是自然語言處理中十份重要的，因此在文句分析模組中，首先要能決定文句中每一個字之正確發音。

解決語義(Sense)岐異的技術，可歸納為規則法(Rule-based)與統計法(Statistical method)，後者需收集大量之語料，經統計方式建立語料庫(Corpus)。本論文即採用統計法之語言模型(Language Model)，經訓練大量語料後，建置多音字語料庫作為預測依據。

我們已發表的論文[11]提出組合式策略(Unify Approach)，可有效預測中文多音字之發音類別。組合式策略包含語言模型(Language Model)與投票計分法(Voting Scheme)，並經二階段預測信心度，決定不同的替代法(Alternatives)以組合式方法提升預測值。實驗之外部語料測試正確率列後達95%。

此外，文句中包含各種特殊的非文字符號，如／，＋，－與＊等，每一符號可多種可能的含意，亦需轉換成適當之語意文字，例如：字串「2008/08/10」應轉為中文「**貳零零捌年捌月拾日**」，「3/4拍」轉為中文「**四分之三拍**」。

2)**韻律訊息預估 (Prosody Prediction)**：

中文文句之組成，由小至大可分為：字、詞、片語與韻律段(Prosody Segment)等單元，中文基本的語意單元為詞(Word)，經斷詞處理後可將文句分割出中文詞，再經構詞將相關詞合伴成較大的語意單元，可稱為韻律段。經統計方法可預估出合成

語音的音長(Duration)、音高(Pitch)、音量(Energy)等聲學(Acoustics)參數。此外，合成語音還需具有清晰度(Intelligibility)，語音清晰度是指說話者要表達的意思能夠被聽話者瞭解的程度，清晰度可定義為聽測者所得到正確訊息的程度，含合成語音之正確聲調、鼻音的程度等因素之判斷。

表1:文句中常見之點號與其停頓時長

	標點符號	停頓時長(sec)
句末點號	句號(。)	0.7
	問號(?)	0.7
	驚嘆號(!)	0.7
句內點號	冒號(:)	0.6
	分號(;)	0.5
	逗號(，)	0.35
	頓號(、)	0.2

文句中的標點符號是輔助文字記錄語言的符號，用以表示停頓、語氣以及詞語的性質和作用。常用的標點符號有16種，分點號和標號兩大類。表1列出文句中常見之點號這些標號依其在文句的位置而有不同之停頓時長，表中係以一般人每分鐘講話150字左右為例，如果說話速度加快的話，此表內之數值相對需要縮短，這些時長值將作為合成時之靜音停頓參數。

中文是一種具有聲調(Tonal)變化之語音，這是重要的中文韻律特性。聲調的特性尤其顯現在母音的音高軌跡，參見圖1中第1聲到第4聲之音高軌跡形狀與其相對位置。經由實驗繪出中文語音音高(Pitch)變化，發現注音符號上的聲調符號，和實際上的音高軌跡(Pitch Contour)類似。其中，聲調最高者為第1聲，第4聲的音長平均最短。輕聲的音高軌跡原則上是受到前後語音聲調的影響，音節大小通常比其他四聲短，音量也較小，最常見的是中文「**的**(ㄉㄜ˙)」這個輕聲字。

3）**選取合成單元（Selection of Basic Speech Units）**：

依文句分析後資訊，自語音庫中選取正確之合成單元，提供後續合成處理。如前所言，我們已錄製約2000個中文字之語音合成單元，以及不同時長之靜音檔。依據前一模組所得之資訊，自語音庫中選取基本的合成單元，含中文語音檔與不同時長之停頓音檔，以輸出具有韻律訊息之中文合成語音。

4）**語音產生**(Speech Generation)：

利用已經預估好的韻律參數進行韻律的調整，最後輸出合成的語音。本系統可以調整音量大小與語者講話之速度(即快慢)。語音之產生使用前述之「串接合成法」，依據韻律參數，調整語音訊號，再將所選取之合成單元作串接之合併，輸出以人們語音為基礎之合成語音。

2.3 中文連音變調處理

中文另一個特殊之處是連音變調(Tone Sandhi)的現象，當二個三聲調的中文字連在一起時，基於人們發音之結構，唸法之發聲將會產生變化。基本上，可分為連續二個字與連續三個字(含三字以上)為三聲調之情形。三聲字因為有下降再上揚的音調轉折(見圖 1)，則音高分佈值可記為 21134。三聲字都依其原本的音長和調值唸出，必然會影響說話的流暢與協調，因此產生了變調的情形。

以下列出簡單幾項規則：

A) 2 個三聲字連在一起：將前一個三聲字變調，讀成二聲。例如:「**保**」險、「**永**」遠、「**冷**」暖、「**海**」島與「**總**」統等，合成時將第一個均字轉為第二聲。

B) **3 個三聲聲調相連**:若是「**雙-單**」詞語結構，即 (AB) + C 的構詞，前兩個三聲字結合成一個詞語，則前兩個三聲字變讀為二聲，形成 ／／∨(223)唸法。例如：狗尾草、老鼠屎、選舉法、手寫體、水彩筆、總統府等。若是「**單-雙**」詞語結構，也就是 A＋（BC）的構詞，前後個三聲字結合成一個詞語，則將第 2 個三聲字變讀為二聲即可，形成 ∨／∨(323)唸法。例如：蔣**總統**、馬**總統**或想**洗澡**，這三個詞均應唸成∨／∨(323)聲調。

至於，如出現前二字、後二字均可形成中文詞的情況時，我們則分別查詞典(中研院之八萬目詞，ASCED)中其對映之詞頻(Frequency)，例如，**小雨傘**，分別查典中「**小雨**」與「**雨傘**」之詞頻大小。前二字之詞頻大於(>=)後二字之詞頻時，依據上面(AB) + C 構詞之唸法(223)。反之，視為 A＋（BC）構詞，為(323)唸法。經由此規則方法，大致上可以解決中文第 3 聲之連音變調問題了。

然而，當三個以上(多個三聲字)之三聲字連在一起、組成一個句子時，變調的方式則涉及句法資訊的影響就比較複雜，惟還是以中文詞為單位再運用前述之轉調規則。例如，下列文句:

「米老鼠想洗澡」 (E4)

轉調前的聲調(下列全部為 3 聲調):

ㄇㄧˇ ㄌㄠˇ ㄕㄨˇ ㄒㄧㄤˇ ㄒㄧˇ ㄗㄠˇ

轉調後的聲調(劃底線者已轉換為 2 聲):

ㄇㄧˇ ㄌㄠˊ ㄕㄨˇ ㄒㄧㄤˊ ㄒㄧˊ ㄗㄠˇ

2.4 語音能量之正規化

語音合成單元需仰賴專業人士錄製，所錄製語音可能因人為情緒、操作錄音設備與外在環境等因素產生不同效果，將使合成之輸出音效降低品質。以音量為例，個別的語音可能因錄製產生特別的差異(特別大或特別小)，本系統使用能量正規化處理(Energy Normalization)以改善能量落差造成的影響。由於中文語音具有不同之聲調，為保存不同聲調之間的差異性，我們分別對5個聲調處理。首先，對每一個語音求出其音量之平均值如下:

$$\overline{E}_{i,j} = \frac{\sum_{k=1}^{\gamma} |S_{i,j}(n_k)|}{\gamma} \quad (1)$$

其中，i 為中文之聲調代號(i = 1 表示第 1 聲，i =2 表示第 2 聲，依此類推)，$S_{i,j}$ 表示中文第 i 聲調中第 j 個語音檔，n_k 為語音 $S_{i,j}$ 語音之第 k 點取樣值，γ 為此音檔取樣量化之點數，$\overline{E}_{i,j}$ 表示的語音能量之平均值。

$$\overline{AE}_i(n) = \frac{\sum_{j=1}^{N_i} \overline{E}_{i,j}}{N_i} \quad (2)$$

其中，$\overline{AS}_i$ 為第 i 聲調的能量標準值，N_i 為中文語音中聲調 i 之語音個數，本系統中 N_i = 408。

$$\tilde{S}_{i,j}(n) = S_{i,j}(n)\frac{\overline{AE}_i}{\overline{E}_{i,j}} \quad (3)$$

$S_{i,j}(n)$ 為原始語音信號，$\tilde{S}_{i,j}(n)$ 為標準化後之語音信號。使用 10 句中文之例句作為能量正規化之測試，測試句之平均字數為 16.5 字，其中含有標點符號，參見附錄 1。

2.5 系統資料庫說明

本系統架構中，包含有三個相關資料庫，內容說明如下:

語音庫(Speech Database)

系統語音庫包含所錄製中文語音之單字音作為基本合成單元，計有408種基本聲調之語音檔，每一基本音檔均有1、2、3、4與輕聲等5種聲調(Tone)，含男女聲與不同時長靜音檔，總計超過4000個合成之基本語音單元(Basic Speech Unit)，作為後續選用之合成單元。這些合成單元錄製格式為: 44.1kHz、16 bits，儲存成Windows PCM格式(wav檔)。

多音字語料庫(Corpora for Polyphone)

為能解決中文多音字，本系統使用了語言模型(Language Models, LM)與詞典比對為基本方法，提出投票計分法(Voting Scheme)作為預測多音字之發音類別。相關訓練之語料經收集、分類訓練後，建置語料庫，作為統計之依據。

中研院詞典(Sinica Dictionary)

在自然語言處理過程裏，如字轉音或斷詞等工作，辭典扮演非常重要的角色，好的辭典可以提高系統正確率。在中文部份，辭典來源有中研院8萬目詞為基礎，再從中研院平衡語料庫抽取出未包含在8萬目詞內的中文詞，構成約13 萬(129669)目詞辭典。8萬目詞(ASCED) 內每一詞目含有中文語、詞性(POS)、詞頻(Frequency)等資訊，由平衡語料庫抽出者則有詞性但無詞頻，中文詞之詞長從1字至10字，其中以2字詞與3字詞分別有71429 (55.1%)與34700 (22.8%)，合計佔全部詞數之77.9%，見圖3。

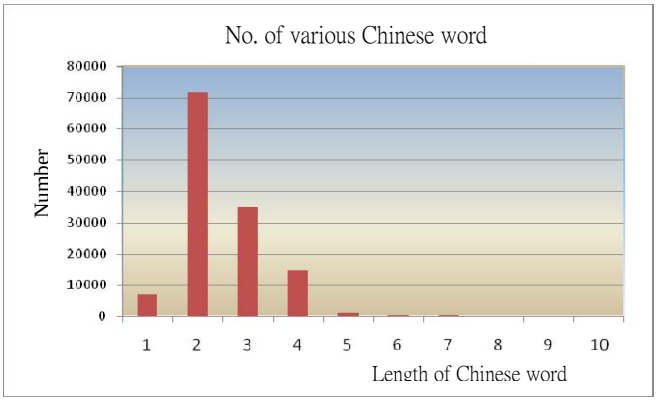

圖: 中文詞典所含不同字數之詞數。

三、系統建置

3.1系統架構

本系統開發環境說明如下；作業環境為 Windows XP SP2, 安裝 Apache Web Server v2.2.4，資料庫系統使用 MySQL Database v5.0.45，資料庫管理平台為 phpMyAdmin v2.10.2，網路開發程式採用

PHP5。我們所建置之線上語音合成系統架構，如圖 4 所示。

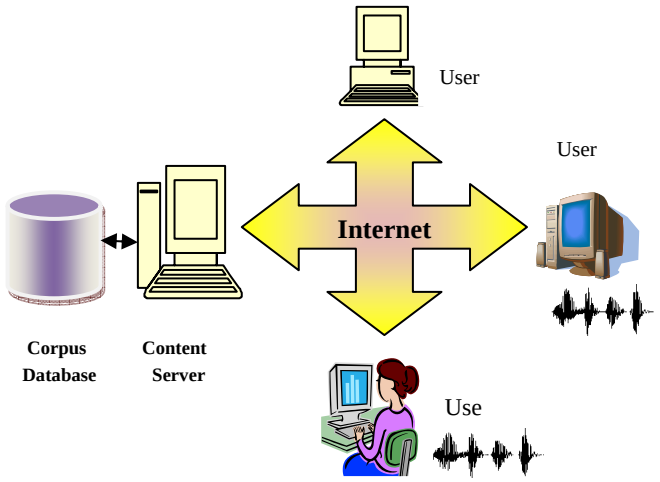

圖 4: 線上中文語音合成系統架構

3.2 合成系統之輸出

使用者可經由 Browser 瀏覽本系統，輸入中文文句，可選擇男女語音、調整音量與說話速度，按送出鈕後，經由後端主機處理，再傳回合成之語音給使用者。本系統已建置完成[8]，圖 5 為文句「聯大線上中文語音合成」合成語音(女聲)之波形。除了合成語音外，畫面上還提供文句所對映之各種拼音轉換與對映之簡體字。

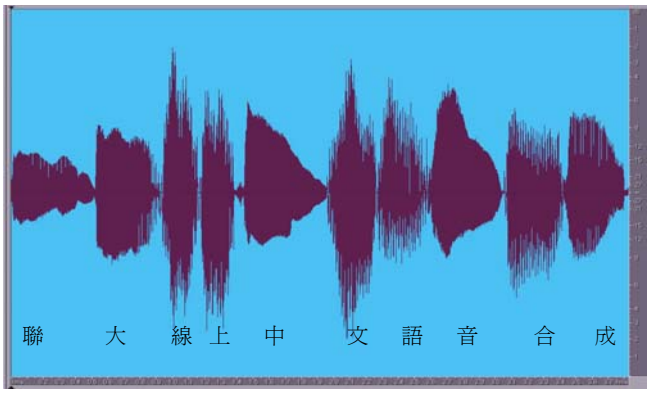

圖5: 「聯大線上中文語音合成」合成語音波形。

四、語音品質測試與分析

有關語音品質的評量，可歸納為兩類，即客觀評定和主觀評定。客觀評定方法用客觀測量的手段來評價語音編碼的質量，常用的方法有信號噪音比(Signal-Noise Ratio, SNR)，簡稱信噪比，還有加權信噪比、平均分段信號噪音比等，以度量均方誤差為主。此法特點是計算簡單、不需經由人的參與，惟不易反映人類對語音質之真實感受。另一方面，主觀評定法則較符合人類聽話時對語音質量的感受，目前較被廣泛應用。

主觀評分常採用主觀評定等級(Subjective Opinion Scale)，亦稱平均評定分數(Mean Opinion Score，MOS)。一般 MOS 之得分採用五級評分標準，其方法是，由聽者在相同環境中試聽並給予評分，再作統計平均分數。為使分數更能反應測試者的感受，我們將分數刻度更細分為 1~10 個等級，作為評量相關語音品質之依據。

4.1 語音能量正規化分析

語音韻律訊息中聲音之大小是語音是否清晰、自然之因素之一，首先我們對合成語音能量作正規化處理。如 2.4 節所述，為使錄音後合成單元之能量調整至較理想之情況，本系統以相同聲調之所有語音進行正規化處理，計算出各個聲調之平均值，依此作為調整能量之依據。意即:比能量平均值大之語音將調降，反之將調升，圖 6 為正規化處理前後能量調整情形，圖 7 為處理前後之波形變化。

附錄一列出測試用 10 個中文例句，原始語音與正規化後之語音經由 12 位大學生聽測實驗評定，採用滿分 10 分之計分，比較二者之滿意度，結果如圖 8 所示。原始與處理後合成語音之平均滿意度分別為 7.51 與 8.09，提升 0.58 分，可見我們使用之能量正規化處理方法對音量滿意度有明顯改善效果。

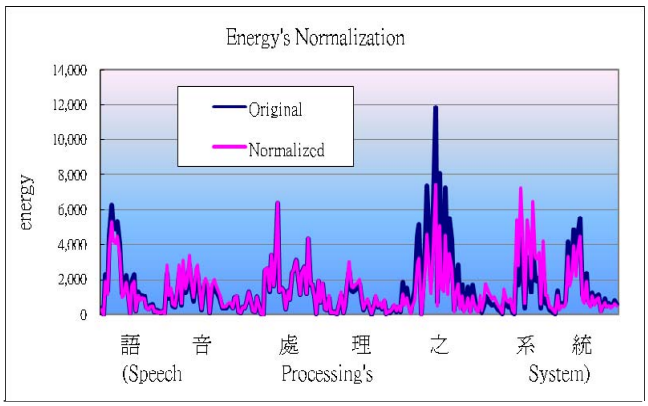

圖 6: 音量正規化之結果

經過能量正規化處理，再統計處理前後各聲調所有語音之能量變化，男聲與女聲各聲調語音之平均值參見表 2，統計結果顯示原始音以第 1 聲最高、輕聲最低，男女聲均如此。至於語音之時長(Duration)，依結果分析，男女聲語音之時長均為第 2 聲最長、輕聲最短。

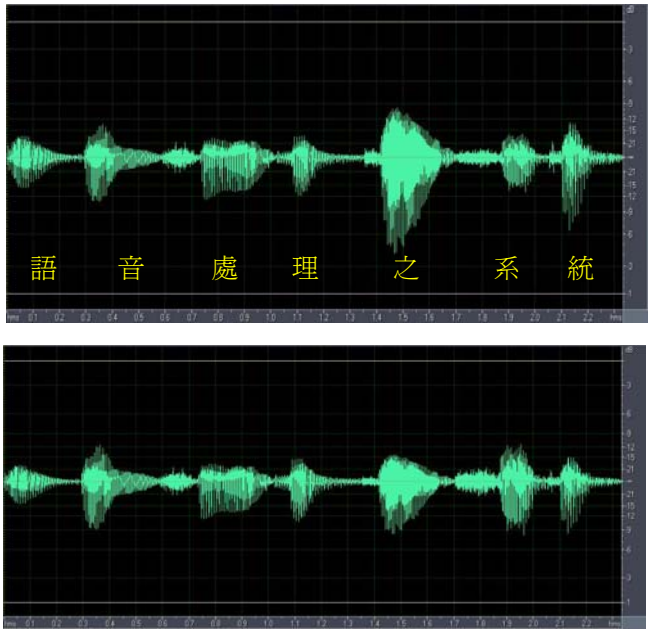

圖 7: 能量正規化處理前(上)後之波形變化。

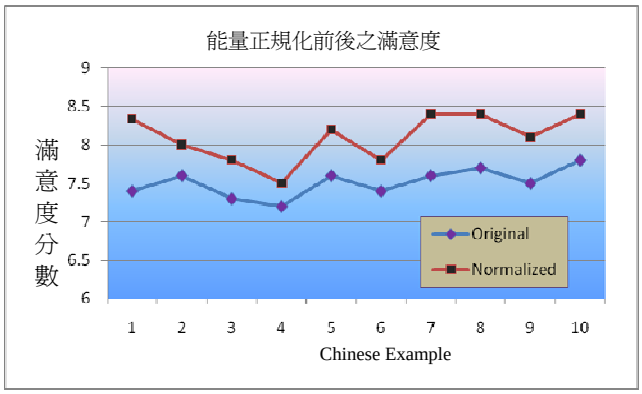

圖 8: 語音能量正規化處理前後之滿意度比較。

表 2 : 男女錄製語音之能量與時長平均值。

	Male		Female	
	energy	duration	energy	duration
tone 1	1736.2	0.375	1548.2	0.389
tone 2	1414.2	0.392	1206.8	0.399
tone 3	1021.7	0.288	1042.0	0.344
tone 4	1627.9	0.315	1470.5	0.372
tone 5	1000.9	0.279	1064.6	0.292

4.2 語音清晰度之分析

語音之韻律訊息除需要作能量正規化處理外，另一重要的因素為語音之清晰度與鼻音現象，合成之語音是否具有良好清晰度與易於辨識語音內容的效果呢?

同樣以前述 10 個中文例句作為判別，同樣由 12 位大學生作清晰度實驗，測試結果參見圖 9，清晰度整體滿意度為 7.63。由能量與清晰度之聽測結果可知，我們系統之語音輸出已具網際網路上真人之語音，合成品質已達到相當程度的自然度與順暢性。

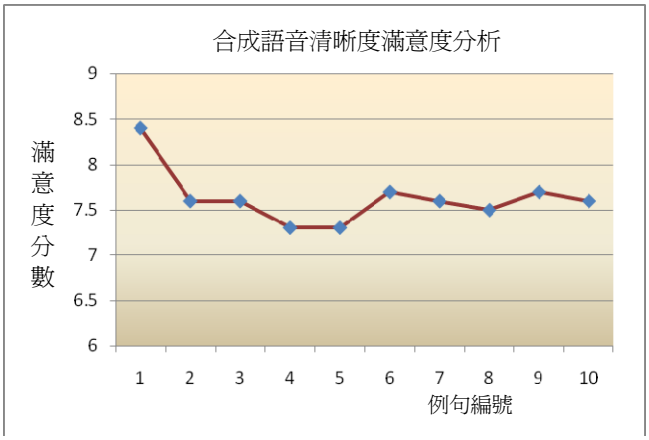

圖 9 : 合成語音清晰度之滿意度統計

4.3合成系統提供之資訊與應用

文轉語音系統中，除中文語音合成之功能外，經由文句分析處理，並提供正確中文文句所對映之標註符號。本系統之輸出介面還包含有各種中文之注音與拼音資訊，如：注音符號、通用拼音、漢語拼音與注音二式(MPS-II)，並轉換出其簡體字，以及調整語音能量與說話速度之選項，可以產生實際需要合成輸出。上述中文之拼音與注音資訊進一步可以提供國際人士或學生認識中文網際網路環境上數位學習系統。使用畫面參見圖10所示。

五、結論

本文主要探討網際網路上建置中文語音合成系統之相關，本系統包含有四個模組：文句分析、律韻訊息分析、合成單元選取與語音合成，並建立相關之語料庫與語音庫。我們錄製中文單字之基本語音合成單元與不同時長之靜音檔，配合韻律參數作為合成之用。為改善所錄製合成單元之能量差異，對所有單元進行能量正規化處理，以提升輸出的語音品質。

由12位熟悉中文之大學生進行10分等級之聽測實驗，進行合成語音品質之滿意度分析，包含能量與清晰度二項。聽測實驗之統計結果顯示，合成語

音之音量正規化與清晰度分數為8.09與7.63，顯示本系統之輸出具有真人語音的效果，品質已達到一定程度之自然、清晰與順暢性。

此外，本系統亦提供文句分析功能，如中文對映之標註符號，擴充合成系統之應用，可作為認識中文之數位學習系統。未來，我們將進一步研究下列主題:

1)提升中文多音字之正確率。
2)多種語言間之翻譯。
3)中文正體字與簡體字系統之轉換。
4)韻律訊息預估與語音串接技術之改善。

圖10: 本系統之輸入網頁與其輸出結果。
上: 輸入畫面，下:語音與文字標注之輸出。

參考文獻

[1] 中研院中文斷詞系統:

http://ckipsvr.iis.sinica.edu.tw/

[2] 悠揚語音合成系統，香港中文大學人機通訊實驗室。

[3]余明興、張唐瑜、許燦煌、蔡育和, 2005, 使用韻律階層及大量詞彙的中文文轉音系統, ROCLING 2005.

[4]劉士弘, 朱芳輝, 陳柏琳, 2007, 改善以最小化音素錯誤為基礎的鑑別式聲學模型訓練於中文連續語音辨識之研究, ROCLING 2007.

[5]鄭秋豫 蘇昭宇, 2007, 從不同韻律格式驗證階層式韻律架構並兼論對語音科技的應用 , ROCLING 2007.

[6] 臺灣本土語言互譯及語音合成系統, 國立臺灣大學資訊工程學研究所自然語言處理實驗室. http://nlg.csie.ntu.edu.tw/systems/TWLLMT/index.html

[7]訊飛語音系統，安徽科大網頁:

http://voicebook.iflytek.com/

[8] 聯大中文語音合成系統:
http://203.64.183.226/public2/style_sentence-input7-shuhua.html

[9] Chen K. J. And S. H. Liu, Word Identification for Mandarin Chinese Sentences, Proceeding of COLING-92, 14th Int. Conf. On Computational Linguistics, pp. 101-107, 1992.

[10] Chu M., Peng H., Yang H. Y. and Chang E., " Selecting Non-Uniform Units from A Very Large Corpus for Concatenative Speech Synthesizer ", Proceedings of ICASSP 2001, IEEE, Volume 2, pp.785 - 788, Salt Lake City.

[11] Feng-Long Huang, 2008, Disambiguating Effectively Chinese Polyphonic Ambiguity Based on Unify Approach, IEEE International Conference in Machine Learning and Cybernetics, (ICMLC) 2008 12-15, Jul., KunMing Mainland, pp. 3242-3246.

[12] Hung-Yan Gu, Yen-zuo Zhou, 2007, An HNM Based Method for Synthesizing Mandarin Syllable Signal, ROCLING 2007.

[13] Sin-Horng Chen, Shaw-Hwa Hwang, and Yih-Ru Wang, 1996, A Mandarin Text-to-Speech System, Computational Linguistics and Chinese Language Processing, Vol. 1, No. 1, pp. 83-94.

[14] Shih-Hsiang Lin, Yao-Ming Yeh, Berlin Chen, A Comparative Study of Histogram Equalization (HEQ) for Robust Speech Recognition, International Journal of Computational Linguistics and Chinese Language Processing, Vol. 12, No. 2, pp. 217-238, June 2007.

[15] Sporat R. and C. Shih, A Statistical Method for Finding Word Boundaries in Chinese Text, Computer Processing of Chinese and Oriental Languages, Vol. 4, No. 4, pp. 336-351, 1990.

附錄一： 中文合成語音之測試用例句。

下列文句合成之語音用於能量正規化與清晰度語音品質測試，計有10句，含標點符號在內，最長與最短文句之字數為27與13，測試例句之平均字數為16.5。

例句 1: 澳門的街頭巷尾都充滿了南歐風情。

例句 2: 有歷史文化意涵的建築，值得細細品味。

例句 3: 這是語音測試，音量之調整。

例句 4: 王建民成為台灣之光，是洋基王牌之一。

例句 5: 語音結構和文法，都有相似的規則。

例句 6: 《萬王之王》是國內第一款線上遊戲。

例句 7: 觀察周遭的景物，去包容變遷的環境。

例句 8: 「藍色」經研究結果，有幫助於人類的睡眠。

例句 9: 開車族最煩人的問題就是找停車位。

例句 10: 台灣這次所看到的滿月，景象將非常壯觀、令人嘆為觀止啊！