應用機率式句法結構與隱含式語意索引於情緒語音合成之單元選取

陳俊甫 夏啟峻 吳宗憲

國立成功大學資訊工程學系
{cama, shiacj, chwu}@csie.ncku.edu.tw

摘要

在人機溝通介面中，語音逐漸扮演著重要的角色。然而，傳統電腦語音缺乏情緒特性，使得電腦與人的互動機能嚴重降低。因此，使電腦合成出帶有不同情緒特性的電腦語音是本文主要目的。在本論中，對於語料式情緒語音合成系統主要的問題，分為下列四項研究重點：1）根據不同情緒，設計一套平衡語料庫，並利用自動單元切割技術，生成基本合成單元；2）提出修正式可變長度單元機制，將機率式句法模型概念導入，決定單元長度與單元合適性；3）有別於一般聲學上的單元失真度計算，應用隱含式語意索引的概念，針對單元的語意失真度進行量度；4）最後，應用動態規劃與自動斷句預測，挑選出單元並合成情緒語音。在實驗中，首先針對中文斷句預測的正確率做比較；接著，對於語音合成的結果，觀察合成語音與實際語音在參數上的差距。並利用主觀式的評估方式，分別進行自然度 MOS 測試，情緒鑑定測試與理解度測試，本文提出之方法，在合成的自然度與情緒的表現上，皆有不錯的表現。

1. 緒論

現階段的語音合成技術已達成熟階段，但是關於帶有情緒特性的電腦語音合成技術，卻還處於起步的狀態。在語音合成方面國外最具代表性的研究機構為 AT&T[1]與微軟亞洲研究院，其相關的研究發展[2][3][4]皆有顯著的成果，微軟亞洲研究院更是投入大量的人力與物力來支持語音科技的相關研究；台灣在 80 年代開始中文語音方面的研究，如台大、清大、交大、成大、工研院電通所、交通部電信所、中研院等，都積極投入研究工作並累積了大量的研究成果。就中文而言，發音一般是以詞為基礎，音節僅能包含子母音相接連的連音變化方式，對於音韻的變化是比較不足的；另一方面，以詞為發音基礎的中文，語者情緒的變化也會呈現在詞的層級上。因此我們這套 Corpus-based 語音合成系統，是以詞（word）與音節（syllable）共同為最基本的合成單元，據此設計四種情緒的語料，以合成目標情緒語音。接著，利用兩階段語音切割機制，生成基本的合成單元。為了量測切割出來的語音片段的合適性，以其在對應的音節序列的統計模型的觀測機率，作為評量標準，來確定切割單元是否正確。由於中文語句是以詞為音韻基礎，本文提出一套修正式可變長度單元挑選機制，採用詞序列語音段為合成單元，再加上中文語音合成常用的基本單元－音節－為最小的合成單元，讓語音在串接時能有保有最原始的音韻與韻律節奏，達到情緒語音合成的目標。為決定候選詞序列單元的合適性，本文針對目前失真度定義上的不足與人類構句與發音時連音的型態，利用機率式句法結構（PCFG，Probabilistic Context Free Grammar）[5]，模擬最符合人類原始連音構句模式的單元挑選機制，並運用隱含式語意索引（LSI , Latent Semantic Indexing）量度合成單元在語意結構上的失真度。最後整合 1.）修正式可變長度單元挑選機制，2.）隱含式語意索引文法結構距離，3.）聲學參數失真度，計算出各種合成單元序列的失真度，再以動態規劃的方式，快速找出最佳的合成單元序列，合成出帶有情緒特性的電腦語音，系統流程如圖 1 所示：

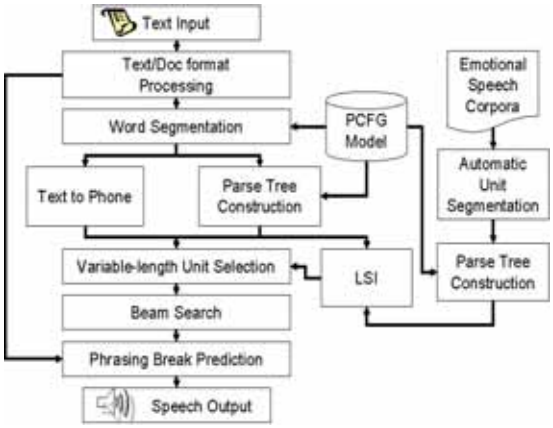

圖 1：情緒語音合成系統流程

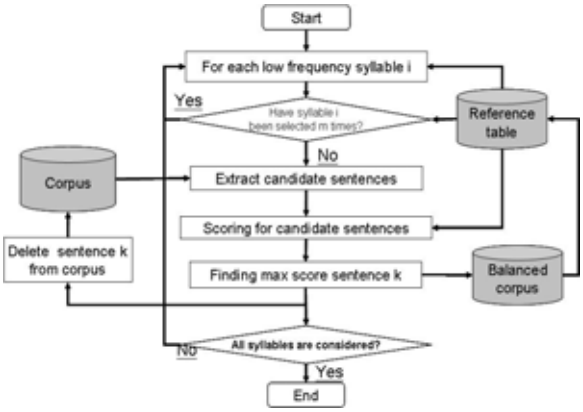

圖 2：平衡語料挑選流程圖

2. 平衡語料庫與合成單元之生成

2.1 平衡語料庫

根據研究人類情緒的文獻探討，本研究中結合 Russell[6]的 Dimensional Approach 與 Prototype Emotion Approach[7]的分類法，以四大類情緒為目標合成的語音之情緒：高興，生氣，悲傷，中性。為錄製收集語音合成系統所需之合成單元，必須先設計一套包含各發音單元及涵蓋大多數中文常用詞彙的文字平衡語料庫；從大量的文字語料（包括新聞、小說，短文等，合約兩萬字）中，根據從平衡條件所制定的計分原則加以挑選；由於用文字來表達情緒有不同程度的差別，本研究希望可以合成出明顯且強烈的電腦語音，為增加文句對情緒表達的程度[8][9][10]，用以下列準則修飾文句，使文句帶有部分情境及誘導目標情緒的特性，藉此增加情緒的表達；

(1) 插入可增強目標情緒的句子，例如：我好難過、太棒了…等；
(2) 加入可增強目標情緒的修飾辭，例如：非常、好、實在、耶…等；
(3) 保持句子的長度及音韻節奏來增強或保持語者的情緒狀態；
(4) 修改影響情緒表達的文句或是多餘的贅語。

語料的完整性與平衡特性是一套 Corpus-based 語音合成系統的基礎，本研究採用修正式可變長度單元挑選機制，所以在語料的設計上，需要包含較多的中文詞，所以本研究以常用詞為主要收集對象，根據詞頻計算每個候選文句的分數；此外，在需要收集到所有音節的考量情況下，音節的計分也被納入，據此我們定義了挑選平衡句的計分方式。計分條件（對第 j 個句子計分）

音節（Syllable）的出現頻率： $\left(\prod_{k=1}^{C\_S_j} fs_{j,k}\right)^{1/C\_S_j}$ (1)

以個別音節的頻率（Uni-gram）計算之。其中 $fs_{j,k}$ 表示第 j 個句的中的第 k 個音節的頻率，C_S_j 則表示第 j 個句子的音節個數。

詞（Word）的出現頻率： $\left(\prod_{k=1}^{C\_W_j} fw_{j,k}\right)^{1/C\_W_j}$ (2)

以每個詞的林率計算之。其中 $fw_{j,k}$ 表示第 j 個句的中的第 k 個詞的頻率，C_W_j 則表示第 j 個句子的詞數。

計分方式： $Score_j = \left(\prod_{k=1}^{C\_S_j} fs_{j,k}\right)^{1/C\_S_j} \times \left(\prod_{k=1}^{C\_W_j} fw_{j,k}\right)^{1/C\_W_j}$ (3)

結合以上兩種計分。

平衡語料挑選的流程圖如圖 2，整個架構分為兩層，外層用來保證將所有音節收入，內層迴圈用以將所有句子計分，挑選出分數最高的句子，每選出一句平衡句都必須要更新計分用的參考表，並且從原始語料將其刪除，終止條件為滿足所需的音節出現次數。

2.2 合成單元之生成

在本研究中，提出一套自動的語音單元切割與確認方法，來加快合成單元的標記作業以及提供驗證的方式。但是由於自動找邊界的結果，並非都是令人滿意的[11]，因此需要做一些調整，才能讓單元切割達到不錯的結果。因此，我們利用兩階段語音切割模組，第一階段利用隱藏式馬可夫模型進行初步預測斷點，第二階段利用語音參數的特性，根據觀察與測試利用規則將斷點調整在正確的位置上。最後，利用機率模型做比對，確認單元正確性。

1.隱藏式馬可夫模型：利用強迫路徑之隱藏式馬可夫模型進行語音切割，在參數的設定上，我們使用 26 維的參數，包括 12 階 MFCCs、12 階的△MFCCs、能量差（delta energy）以及能量差之差（delta delta energy）值，而在模型個數上，總共包括 150 個次音節模型。對於先前所錄製的聲音語料，根據其文字內容，利用已知音節型態（syllable type）所對應的隱藏式馬可夫模型，使用 Viterbi 演算法，搜尋每個對應模型狀態外轉的位置，據此找出每個音節的邊界位置。

2.斷點調整：經由觀察的結果，發現邊界切割錯誤主要有幾種：第一，子音開頭。可能受到前一個音節的母音結尾連音的影響，或是子音前的靜音部分的干擾，可能造成斷點錯誤。第二，母音結尾。母音結尾可能因為在頻譜上的落差太大，造成外轉提前發生，導致母音還未結束時就被切割下來。因此，本研究利用能量（energy）以及過零率（zero crossing rate）當作調整的觀測參數[11]，對於每一種不同的音節型態，設定多個規則來微調。

3.單元確認：根據前一步驟所述可將連續語音切成最基本的合成單元，然而並非每個單元在錄置的過程中，都是正確無誤的，本節的目的是要測試各語音切割結果跟相對應的音節型態是否一致，也就是測試此語音單元在對應的音節模型內的機率高低，如果機率高，表示此單元的結果是較為正確的，反之，則單元較不正確。

$$\begin{cases} P(X \mid \lambda_i) < threshold_i & \rightarrow \text{reject} \\ Otherwise & \rightarrow \text{accept} \end{cases} \tag{4}$$

其中，λ_i 代表相對於觀測資料 X 的音節模型，$threshold_i$ 代表此音節型態所對應的臨界值。

3. 修正式可變長度單元合成機制

3.1 可變長度單元合成機制

從一個大量的語料庫中挑選出合適的合成單元已經被證明確實有助於提升合成系統的品質[4][12]，而單元的型態包括音素（Phoneme）、雙音（Diphone）、半音節(Demi-Syllable)、音節(Syllable)、不定長度的單元(Non-Uniform Unit)等。就中文而言，如果能找到較長詞來當合成單元，當然是一個比較好的選擇，因為這樣的合成單元內，已經包含了本身的音韻，因此在串接的自然度上有一定的效果提升。過去，可變長度單元的挑選機制主要是以詞為基礎。對於每一個可能出現的詞或是音節，去搜尋所有可能的組合方式，找出一組最佳的詞序列。例如：

中 國 人 是 聰 明 的 民 族

就這個句子而言，他所可能衍生出來的可能組合性有很多：

(1)中國人 是 聰明 的 民族　　(2)中國人 是 聰明 ***的民族***

(3)中國人 是 聰明的民族　　(4)中國人 是 聰明的 民族

(5)中國人 ***是聰明*** 的民族　　(N)……

但是，其中有許多的組合是不符合中文音韻的組合，例如「***的民族***」 「***是聰明***」，而且若要搜尋所有可能的組合，所要耗費的時間跟空間複雜度太龐大。因此我們提出了一套新的可變單元長度挑選機制，主要考慮兩個觀點。第一，模擬人類構句的方式，根據中文發音的音韻與斷句，我們可以找到合適的合成單元。由於人類構句的方式，是先將單音節（syllable）組合成詞（word），再將多個詞組合成長詞或專有名詞，進一步組合成片語、句子，如圖 3 所示。

因此，我們可以根據這樣的想法，將不適合的組合性去除，並可根據不同階層上的組合方式，進行階層式的單元挑選。第二，除了聲學上的失真度之外，語意結構上的失真度也該被考量。根據中文語音學的觀點，同一個詞或是同一個音節，在不同的語句結構中，它們在聲學參數上的表現會不一樣，舉例來說：

(A). 例一：　漂亮的雙殺，化解了滿壘的危機　(44.3ms)

　　墾丁的風景還是一樣漂亮　(65.3ms)

這個例子中，同樣的詞，位在不同的詞性的長詞中，明顯的，兩個詞的音長是不同的。

(B). 例二：　在院子裡栽種了好多鮮豔的花　(39.1ms)

　　我的手藝真是越來越高超，花招也變多了　(28ms)

這個例子中，「花」作為不同詞性之用，在音高的變化上也有不同的結果。根據這兩個想法，本研究提出一個修正式可變長度單元挑選機制，利用機率式句法結構轉譯器（probabilistic syntactic parser），將中文句轉換成一個階層式樹狀語意結構，這棵樹上的每一個終端節點，代表的是一個詞。而每一個非終端節點，代表了一種可能的長詞組合，如圖 4 所示。這樣的作法有以下幾種優點：1.)可移除不適當的長詞組合；2.)利用樹狀結構，挑選出適合的合成單元；3.)可根據語意結構，量測單元間的語意失真度。

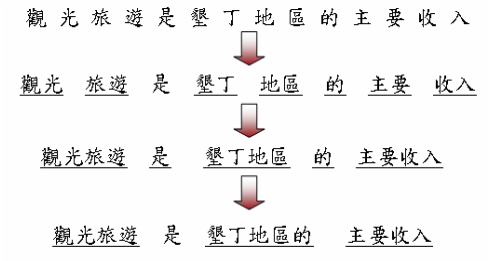

圖 3：人類構句過程模擬示意圖

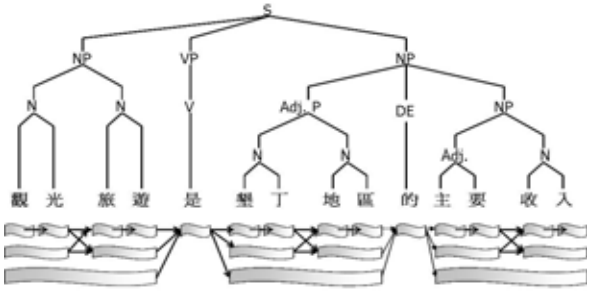

圖 4：中文文法樹範例

3.2 中文文法機率模型

首先，我們需要一個語意剖析器來處理中文文句，並建立對應的語意樹狀結構，本研究利用機率式句法結構（PCFG , Probabilistic Context Free Grammar）來對中文句進行剖析[5]。所謂的機率式句法結構是由句法結構（CFG, Context Free Grammar）衍生而來，以機率的觀點來看語言模型，更可藉由

賦予句法結構 CFG 的規則機率，使得機率式句法結構能夠更正確的模擬口述語言，使語意混淆度降低。機率式句法結構的觀念，與語音辨識中隱藏式馬可夫模型的概念類似，同樣的是想要找出若給定一個文法G，從起始符號N_0開始，產生一串詞序列$W_{1,T}=w_1,w_2\cdots w_T$的機率值：

$$P\left(S\overset{*}{\Rightarrow}W_{1,T}\,|\,G\right) \tag{5}$$

其中，箭號⇒表示衍生的意思，而箭號上方的星號*則表示所有衍生的路徑。這項機率值是由所有合法的衍生規則組合而成，每條規則的機率則是預先由訓練語料中估算求得。假設有一條規則是$A\to\alpha$，則此規則的機率求法為：

$$P\left(A\to\alpha_j\,|\,G\right)=\left[C\left(A\to\alpha_j\right)\right]\Big/\left[\sum_{i=1}^{m}C\left(A\to\alpha_i\right)\right] \tag{6}$$

其中，$C(\bullet)$代表的是每條規則出現的次數，m表示α_i的所有可能性，或所有由A衍生出來的規則個數。本研究採用中研院詞庫小組所定義的 Tree-Bank 文法規則以及相對應的機率為 PCFG 模組的原始模型，擷取部分作為本研究中的文法規則。在此我們導入 Chomsky Normal Form，目的是簡化說明 PCFG 模組以及本研究提出的文法結構距離量測。假設每個非終端項只能分為兩個非終端項的組合$N_i\to N_j+N_k$或是一個終端項（terminal term）$N_i\to w_l$，且其所有可能性的機率和為 1：

$$\sum_{j,k}P\left(N_i\to N_jN_k\,|\,G\right)+\sum_{l}P\left(N_i\to w_l\,|\,G\right)=1 \tag{7}$$

根據這套文法規則G，如圖 5，從起始符號N_0開始，推衍產生一串詞序列$W_{1,T}=w_1,w_2\cdots w_T$的機率值為：

$$P\left(N_0\overset{*}{\Rightarrow}w_1w_2\cdots w_T\,|\,G\right)=\sum_{i}\left(P\left(N_i\overset{*}{\Rightarrow}W_{m,n}\,|\,G\right)P\left(N_0\overset{*}{\Rightarrow}W_{1,m-1}N_iW_{n+1,T}\,|\,G\right)\right) \tag{8}$$

式(8)中，$P\left(N_i\overset{*}{\Rightarrow}W_{m,n}\,|\,G\right)$我們稱之為內部機率（Inside Probability），代表的是一個非終端項N_i被推成詞序列$W_{m,n}=w_m\cdots w_n$的機率值，我們將此機率值表示為$\beta_i\left(m,n\,|\,G\right)$。根據 Chomsky Normal Form 的表示式，一個非終端項只能被分為兩個非終端項的組合，以遞迴的寫法表示成：

$$\begin{aligned}P\left(N_i\overset{*}{\Rightarrow}W_{m,n}\,|\,G\right)&=\beta_i\left(m,n\,|\,G\right)=\sum_{j,k}\sum_{d=m}^{n-1}P\left(N_i\to N_jN_k\,|\,G\right)P\left(N_j\overset{*}{\Rightarrow}W_{m,d}\,|\,G\right)P\left(N_k\overset{*}{\Rightarrow}W_{d+1,n}\,|\,G\right)\\&=\sum_{j,k}\sum_{d=m}^{n-1}P\left(N_i\to N_jN_k\,|\,G\right)\beta_j\left(m,d\,|\,G\right)\beta_k\left(d+1,n\,|\,G\right)\end{aligned} \tag{9}$$

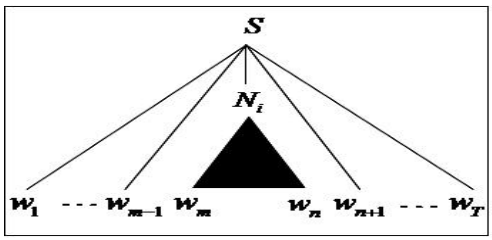

圖 5：機率式句法結構示意圖

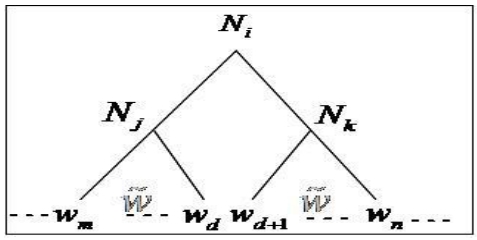

圖 6：單元內部機率

由於本研究只需要在建立樹狀結構的過程中，取分數最高的一棵樹，因此我們將式(9)改寫，在所有可以建出一棵樹狀結構的可能中，挑選出分數最高的當作輸出的機率值，如下表示：

$$\begin{aligned}\hat{\beta}_i\left(m,n\,|\,G\right)&=P\left(N_i\overset{\max}{\Rightarrow}W_{m,n}\,|\,G\right)=\max_{\substack{j,k\\ m\le d<n}}\left(P\left(N_i\to N_jN_k\,|\,G\right)\times P\left(N_j\overset{\max}{\Rightarrow}W_{m,d}\,|\,G\right)P\left(N_k\overset{\max}{\Rightarrow}W_{d+1,n}\,|\,G\right)\right)\\&=\max_{\substack{j,k\\ m\le d<n}}\left(P\left(N_i\to N_jN_k\,|\,G\right)\hat{\beta}_j\left(m,d\,|\,G\right)\hat{\beta}_k\left(d+1,n\,|\,G\right)\right)\end{aligned} \tag{10}$$

式(8)中的$P\left(N_0\overset{*}{\Rightarrow}W_{1,m-1}N_jW_{n+1,T}\,|\,G\right)$，我們稱為外部機率（Outside Probability），代表的是由起始符號N_0推出詞序列$W_{1,m-1}=w_1\cdots w_{m-1}$與$W_{n+1,T}=w_{n+1}\cdots w_T$，且兩詞序列中夾著$N_j$的機率值，同樣的，把外部機率表示成$\alpha_j\left(m,n\,|\,G\right)$。由於非終端項$N_j$可能位於上一層非終端項$N_i$推導出的規則中的左項或右項。因此，將式子寫為所有可能的規則與詞斷點的機率和。

$$
\begin{aligned}
&P\left(N_0 \overset{*}{\Rightarrow} W_{1,m-1}N_jW_{n+1,T} \mid G\right) = \alpha_j(m,n \mid G) \\
&= \sum_{i,k}\left(\begin{array}{l}\sum_{d=n+1}^{T_q}\left(P\left(N_i \to N_jN_k \mid G\right)\times P\left(N_0 \overset{*}{\Rightarrow} W_{1,m-1}N_jW_{d+1,T} \mid G\right)P\left(N_k \overset{*}{\Rightarrow} W_{n+1,d}\right)\right) \\ +\sum_{d=1}^{m-1}\left(P\left(N_i \to N_kN_j \mid G\right)\times P\left(N_k \overset{*}{\Rightarrow} W_{d,m-1}\right)P\left(N_0 \overset{*}{\Rightarrow} W_{1,d-1}N_jW_{n+1,T} \mid G\right)\right)\end{array}\right) \\
&= \sum_{i,k}\left(\begin{array}{l}\sum_{d=n+1}^{T_q}\left(P\left(N_i \to N_jN_k \mid G\right)\alpha_i(m,d \mid G)\beta_k(n+1,d \mid G)\right) \\ +\sum_{d=1}^{m-1}\left(P\left(N_i \to N_kN_j \mid G\right)\beta_k(d,m-1 \mid G)\alpha_i(d,n \mid G)\right)\end{array}\right)
\end{aligned}
\tag{11}
$$

同樣的，由於我們只需要求取最高分樹的那棵樹狀結構，因此我們將式(11)改寫為：

$$
\begin{aligned}
&\hat{\alpha}_j(m,n \mid G) = P\left(N_0 \overset{\max}{\Rightarrow} W_{1,m-1}N_jW_{n+1,T} \mid G\right) \\
&= \max_{j,k}\left(\begin{array}{l}\max_{n+1\le d\le T_q}\left(P\left(N_i \to N_jN_k \mid G\right)\hat{\alpha}_i(m,d \mid G)\hat{\beta}_k(n+1,d \mid G)\right), \\ \max_{1\le d\le m-1}\left(P\left(N_i \to N_kN_j \mid G\right)\hat{\beta}_k(d,m-1 \mid G)\hat{\alpha}_i(d,n \mid G)\right)\end{array}\right)
\end{aligned}
\tag{12}
$$

由於本研究採用不固定長度的單元挑選機制，系統選用的候選合成單元不是音節而是詞序列，所以對於內部機率的剖析，須考慮所要的合成單元，此單元在剖西的過程中，不能再被切割。因此，我們需要求出一個由非終端項 N_i 推導出詞序列 $W_{m,n} = w_m \cdots w_n$ 且包含詞序列（合成單元）$\tilde{w}$ 的共同機率值，因此我們必須求得 $P\left(N_i \overset{*}{\Rightarrow} W_{m,n}, \tilde{w} \mid G\right)$，如圖六所示。

$$
\begin{aligned}
&P\left(N_i \overset{*}{\Rightarrow} W_{m,n}, \tilde{w} \mid G\right) = \gamma_i(m,n,\tilde{w} \mid G) \\
&= \sum_{j,k}\left(P\left(N_i \to N_jN_k \mid G\right)\times \sum_{d=m}^{n-1}\left(\begin{array}{l}\gamma_j(m,d,\tilde{w} \mid G)\beta_k(d+1,n \mid G)\delta(m,d,\tilde{w}) \\ +\beta_j(m,d \mid G)\gamma_k(d+1,n,\tilde{w} \mid G)\delta(d+1,n,\tilde{w})\end{array}\right)\right)
\end{aligned}
\tag{13}
$$

$$
\delta(m,n,\tilde{w}) = \begin{cases}1, & \text{if } \tilde{w} \text{ is a substring of } W_{m,n} \\ 0, & \text{otherwise}\end{cases}
\tag{14}
$$

同樣的，由於我們只需要求取最高分樹的那棵樹狀結構，因此我們將之改寫為：

$$
\begin{aligned}
&\hat{\gamma}_i(m,n,\tilde{w} \mid G) = P\left(N_i \overset{\max}{\Rightarrow} W_{m,n}, \tilde{w} \mid G\right) \\
&= \max_{\substack{j,k \\ m\le d<n}}\left(\begin{array}{l}P\left(N_i \to N_jN_k \mid G\right)\hat{\gamma}_j(m,d,\tilde{w} \mid G)\hat{\beta}_k(d+1,n \mid G)\delta(m,d,\tilde{w}), \\ P\left(N_i \to N_jN_k \mid G\right)\hat{\beta}_j(m,d \mid G)\hat{\gamma}_k(d+1,n,\tilde{w} \mid G)\delta(d+1,n,\tilde{w})\end{array}\right)
\end{aligned}
\tag{15}
$$

4. 文法結構距離與情緒語音合成

4.1 文法結構距離

在前一節提到，同樣的單元在不同的語意結構上，會有不同的表現，因此本研究設計了一套測量文法結構距離的方法，主要是根據機率式文法結構所產生出的語法樹，藉由隱含式語意索引，計算單元在不同語意結構上的差距。

4.1.1 文法結構樹向量化

由於每個句子可以由一棵文法結構樹來表示，而一棵樹只會由少數幾個規則所構成，會有稀疏資料（sparse data）的問題。因此，為了解決這個問題，並且求出單元在不同文法結構樹上的關係，因此我們採用資訊檢索中向量空間比對（Vector Space Model）的方法。將樹結構的比對，視為向量的比對。

將所有的文字語料轉換成規則向量，儲存在一個維度為 $R\times Q$ 的文法結構資訊矩陣 $\mathbf{\Phi}_{R,Q}$。其中 R 代

表整個 PCFG 模型G中文法規則的個數，Q代表語料庫中句子的個數。

$$\mathbf{\Phi}_{R\times Q}=\begin{bmatrix}\phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,Q}\\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,Q}\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{R,1} & \phi_{R,2} & \cdots & \phi_{R,Q}\end{bmatrix} \tag{16}$$

矩陣中每個元素$\phi_{r,q}$代表著第r條規則在第q個句子S_q中所佔的重要性。因此本研究中定義$\phi_{r,q}$的估計法如下：

$$\phi_{r,q}=(1-\varepsilon_r)P\left(Rule\ r:N_i\rightarrow N_jN_k,W_{1,T},\tilde{w}\,|\,G\right) \tag{17}$$

其中，等號右側第二項代表的是該條規則佔該句語法結構的比重，該項可以寫為：

$$P\left(Rule\ r:N_i\rightarrow N_jN_k,W_{1,T},\tilde{w}\,|\,G\right)=C\left(N_i\rightarrow N_jN_k,W_{1,T},\tilde{w}\right)\Big/\sum_{a,b,c}C\left(N_a\rightarrow N_bN_c,W_{1,T},\tilde{w}\right) \tag{18}$$

而第一項是用來度量該條規則在語料中的鑑別性是否足夠，當作矩陣中該元素的權重，利用量度文字亂度（Entropy）的方法，量度某條規則在該語料中是否具有鑑別性：

$$\varepsilon_r=-\frac{1}{\log Q}\sum_{q=1}^{Q}\left(\frac{C\left(N_i\rightarrow N_jN_k,W_{1,T_q}^{(q)}\right)}{\sum_{a=1}^{Q}C\left(N_i\rightarrow N_jN_k,W_{1,T_a}^{(a)}\right)}\log\frac{C\left(N_i\rightarrow N_jN_k,W_{1,T_q}^{(q)}\right)}{\sum_{a=1}^{Q}C\left(N_i\rightarrow N_jN_k,W_{1,T_a}^{(a)}\right)}\right) \tag{19}$$

其中$W_{1,T_q}^{(q)}=w_1^{(q)}\cdots w_{T_q}^{(q)}$表示語料庫中第$q$個句子，$T_q$表示該句的長度，而$C\left(N_i\rightarrow N_jN_k,W_{1,T_q}^{(q)}\right)$則表示文法規則$N_i\rightarrow N_jN_k$出現在第$q$個句子的次數。

4.1.2 中文文法結構距離

由於語意樹結構矩陣十分的龐大，在計算上也非常耗時，本研究導入資訊檢索上的隱含式語意索引技術（LSI, Latent Semantic Indexing），不僅可以找出規則間的隱含關係，更可達至大幅降低向量維度的目標，隱含式語意索引是由奇異值分解後，由奇異值矩陣上決定要保留的變異比例，藉此決定所需的維度，再將所有的向量透過轉換矩陣，投射到教低維度且較有鑑別能力的空間上，且可以有效保留住規則與語意樹的關係。數值運算如下所示，本研究中所則保留 98%的變異量：

$$\mathbf{\Phi}_{R\times Q}=\begin{bmatrix}\phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,Q}\\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,Q}\\ \vdots & \vdots & \ddots & \vdots\\ \phi_{R,1} & \phi_{R,2} & \cdots & \phi_{R,Q}\end{bmatrix}=\mathbf{T}_{R\times n}\mathbf{S}_{n\times n}\left(\mathbf{D}_{Q\times n}\right)^{\mathrm{T}} \tag{20}$$

$$\text{where } n=\min(R,Q)$$

$$\tilde{\mathbf{\Phi}}_{R\times Q}=\mathbf{T}_{R\times d}\mathbf{S}_{d\times d}\left(\mathbf{D}_{Q\times d}\right)^{\mathrm{T}}\ \text{ where } d<n,\ d=\min_k\left(\sum_{i=1}^{k}\lambda_i\Big/\sum_{i=1}^{n}\lambda_i\right)>98\% \tag{21}$$

經過奇異值分解後，我們可以利用$\mathbf{T}_{R\times d}$矩陣，將兩個句子的文法結構向量投射到較低維度的向量空間做比對，假設要合成的目標語句是$\mathbf{x}$，而包含的所需的合成單元$\tilde{w}$的候選語句為$\mathbf{y}$，則利用上述方法，定義文法結構距離：

$$SyntacticCost(\mathbf{x}^{(\tilde{w})},\mathbf{y}_q^{(\tilde{w})})=-\log\left(\hat{\gamma}_0\left(1,T_q,q,\tilde{w}\,|\,G\right)\times\frac{\left(\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{x}^{(\tilde{w})}\right)\bullet\left(\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{y}_q^{(\tilde{w})}\right)}{\left\|\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{x}^{(\tilde{w})}\right\|\times\left\|\left(\mathbf{T}_{R\times d}\right)^{\mathrm{T}}\times\mathbf{y}_q^{(\tilde{w})}\right\|}\right) \tag{22}$$

4.2 情緒語音合成

根據前幾張的介紹，本研究提出一套基於語意結構的可變長度單元挑選機制，決定了合成單元的選擇基準，而且考慮單元在不同語意結構上的關係，定義出語意失真度，本節將定義其他聲學上的失真度，並利用中文斷句預測，在長句中，加入自然且合理的停頓，以符合中文發音該有的韻律跟節奏。

4.2.1 聲學失真度

（一） 頻譜斜度

藉由量測連續兩個合成單元間在各頻譜間的不連續性，計算音節間失真度。首先，將合成單元語音在串接點的前後三個框架做 256 點的 FFT（Fast Fourier Transform）轉換，轉成各頻譜的能量。接著，將轉換出來的頻譜範圍，分成 k 個頻帶。針對各個不同的頻帶，利用線性迴歸（Linear Regression）的

方法，在連續的三個框架間，求出一條迴歸曲線；最後，量測連續兩個合成單元在每個頻帶上，其迴歸曲線的斜率差值。

$$SD(u_n, u_{n+1}) = \sum_{i=1}^{k} w(i)[\Delta u_{n+1}(i) - \Delta u_n(i)]^2 \tag{23}$$

（二） 音高與能量

連續的兩個合成單元，利用 Autocorrelation 的方法，求取其平均基週，藉由量測此兩單元間的基週差異，訂定一音高失真度的量測。同樣的，利用計算兩單元的漢明能量，訂定一能量失真度。在這兩個失真度之間，取一個權重，取其總和，定義為音韻失真度。

$$PD(u_n, u_{n+1}) = w_{Fo} C_{F0}(u_n, u_{n+1}) + w_{ene} C_{ene}(u_n, u_{n+1}) \tag{24}$$

4.2.2 整句元網格最佳路徑搜尋

根據上述的失真度定義，我們可以將本研究中的失真度[13]，分為以下兩種：

音節失真度：利用單元與單元在不同的語意結構造成其發音語音韻的不同，並利用機率式句法結構與隱含式語意索引，定義出語意失真度，作為音節失真度。

$$C_C = w_{SD} SD(u_n, u_{n+1}) + w_{PD} PD(u_n, u_{n+1}) \tag{25}$$

音節間失真度：利用語音在聲學上連續的特性，分別利用頻譜斜率、音高與能量，量測連續兩個合成單元在這些參數上的差異，定義為音節間的失真度。

$$C_S = SyntacticCost(\mathbf{x}^{(\tilde{w})}, \mathbf{y}_q^{(\tilde{w})}) \tag{26}$$

其中 $\tilde{w}$ 定義為單元 u_n 相對應的中文描述。根據這兩個失真度的定義，我們便可得到下面的式子：

$$\hat{u}_{1:N} = \arg\min_{u_{1:N}} \left(C_S(u_0, u_0') + C_C(u_0, u_1) + C_S(u_1, u_1') + C_C(u_1, u_2) + \cdots + C_C(u_{N-1}, u_N) + C_S(u_N, u_N') \right) \tag{27}$$

因此，我們利用動態規劃演算法，在一連處的候選單元序列中，求得一個失真度總和最小的合成單元序列。但是，由於當語料大或是句子太長，會導致搜尋的空間過大，使的時間複雜度太高，因此我們利用 Beam Search，來限定路徑，減少搜尋時間。

4.3 中文音韻詞組預測

對於中文發音中韻律與節奏的產生，停頓佔了扮演了一個很重要的角色，他不只可以避免語意上的扭曲（semantic ambiguity），更可以增加中文句音韻的效果。由於中文是一種單音節詞的語言，通常在一個單詞的音節間，不會有停頓的產生，因此如何在這些詞與詞之間，找出停頓的位置與長度，便是這節的重點。首先，我們可以將音韻詞組預測的預測，視為一種自動學習的問題，也就是說，如何從小量的資料訓練中，找出預測音韻詞組的位置跟長度的。本研究利用分類與回歸樹（CART，Classification and Regression Tree）的方法來達成。我們設計一套問題集，包含了關於詞性，詞長，以及詞性對的的相關問題，利用根據一小量的訓練資料，自動訓練出一棵決策樹，其中，每個非終端節點代表的是一個問題，而每個葉節點才代表著一種停頓類型，如圖 7 所示。本研究中，我們間停頓的種類分為三類：A.)沒有停頓（No break）：在詞與詞間沒有停頓、B.)次停頓（Minor break）：詞與詞之間有一個小的無聲區、C.)主要停頓（Major break）：詞與詞之間有一長停頓。當一個測試資料進來，根據這棵決策樹，我們就可以判斷在兩個詞之間，是否有停頓的產生以及停頓是屬於何種停頓。

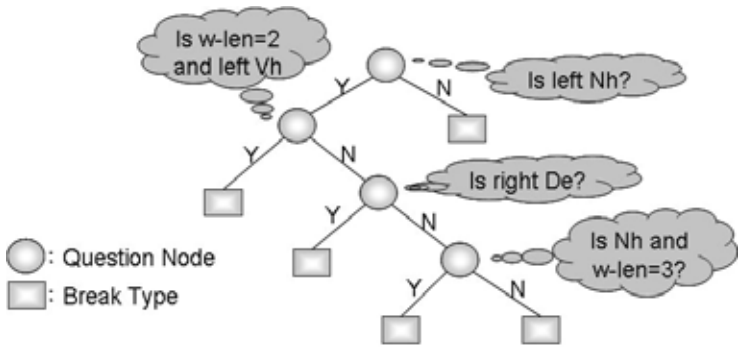

圖 7：中文音韻詞組預測 CART 示意圖

5.實驗結果與討論

5.1 中文音韻詞組預測

本研究中，人工標記了 400 句中文句，其中三百句當作訓練語料，一百句為測試語料。在一百句的測試語料中，總共有 574 個音韻斷點，每種斷點的預測結果如下表所示。由表 1 可以發現，在大部分的音韻斷點中，多數是屬於不在詞之間插入停頓，而主要停頓的次數也是較少的。而利用本方法所得到的正確率，平均都可達到 80%以上。

表 1：中文斷句預測結果

	No break	Minor break	Major break
No break	243	12	6
Minor break	24	156	8
Major break	18	15	65

5.2 合成語音參數比照

為了觀察合成語音與原始語音在平均基週、平均能量以及音節長度之差別，依據本研究中所訂定的四種情緒，錄製語料外的情緒語句，並利用本研究之合成系統，合出相同中文文句之語音，相互比較，圖 8、圖 9 分別是悲傷與生氣對照圖。我們可以發現，在基週、能量、音長上，兩條曲線在大部分的音節是相近的。但可以發現，音長的曲線上，有某些音節例如生氣句的第 13、14 個音節"可以"，高興句的第 17、18 音節"終於"，其音長與能量原始語音有差距，主要是因為沒找到對應的長詞，因此都以單音節的方式合成，所以會造上差異。

5.3 主觀式評估與聽覺實驗

（一）自然度評估（Naturalness Evaluation Test）：本研究採用平均鑑定分數（Mean Opinion Scores , MOS）作為評估之標準，這種評估方式將合成語音輸出的自然度與情緒表達度分為優良（Excellent），良好（Good），尚可（Fair），差（Poor），極差（Unsatisfactory）五個等級，分別給予 5 至 1 不等的分數。測試人員在聽過合成的語音後，以所感覺到的自然度與情緒表現度評分。測試是由合成系統根據基本合成單元長度與語意失真度的使用與否，合成同樣的中文句，做對照實驗。對於每種情緒，合成十個句子，選擇十位大學及研究生（8 為男性，2 位女性），聆聽並根據自己所感受的語音自然度打分數，最後取一個平均。實驗中，比較三套系統（A）、（B）、（C），在合成語音自然度上的差異。（A）系統是利用單一音節為合成單元之合成系統、（B）系統為可變單元長度，但沒有加入語意失真度、（C）系統為本研究所提之系統。由表 2 結果可瞭解，利用本研究所提出的方法，進行單元的挑選，在自然度的表現上，相較於利用單音節的方式，所合成的語音，有相當大改進，在挑選過失真度上，若加入語意失真度，會使的挑選出的語句，在中文音韻上，更符合目標句所要表達的。

（二）可理解度評估（Intelligibility Evaluation Test）：本實驗的目的，是希望探討利用本實驗提出的方法所合成的語音，在可理解度上，是否達到實用的階段，並做相關比較。實驗部分，要求受測者，將所聽到的中文結果，以聽寫的方式寫出來，計算與原始文字的異同，計算其聽寫正確率。同樣的，用前一節所提到的（A）、（B）及本研究中所實作之系統，分別進行實驗。對於每個系統，四種情緒各產生十個句子，讓受測者聽寫。每個受測者平均聽寫了 1632 個音節。由圖 10 可以看出，雖然三套系統，平均都有不錯的理解度：（A）83%，（B）89.5%，（C）96.5%，但是本系統之方法，仍較一般可變單元長度之方法高。這結果顯示，本系統在可理解度以及實用性上是足夠的。

（三）情緒鑑定評估（Emotion Identification Test）：本實驗利用本研究提出之系統，針對四種情緒，各合成十個語音範例，隨機播放，讓受測者決定聽到的語音是何種情緒，由此判定系統在情緒的表現的程度。實驗由受測者，分別聆聽生氣、快樂、悲傷、中性語音，但是先不告知所聽的合成語音是何種情緒，讓受測者依據自己的聽覺，判斷並記錄，表 3 是情緒句子範例。圖 11 顯示情緒鑑定評估在各情緒的正確率：高興 83%，中性 70%，悲傷 93%，生氣 92%。由圖可以看出，中性與高興情緒被誤判的機率明顯的比生氣與悲傷高，主要是因為錄音員在錄製後兩組語料時，在情緒表達上較為強烈，因此被誤判的機率，相對降低，除此之外，也由於本研究所提出的方法，在自然度與可理解度上的提高，因此情緒鑑定的結果也較佳。

表 2：自然度實驗結果

項目	自然度		
系統	(A)	(B)	(C)
快樂	3.2	3.5	4.1
中性	2.7	3.25	3.6
悲傷	3.01	3.2	3.85
生氣	2.85	3.15	3.7

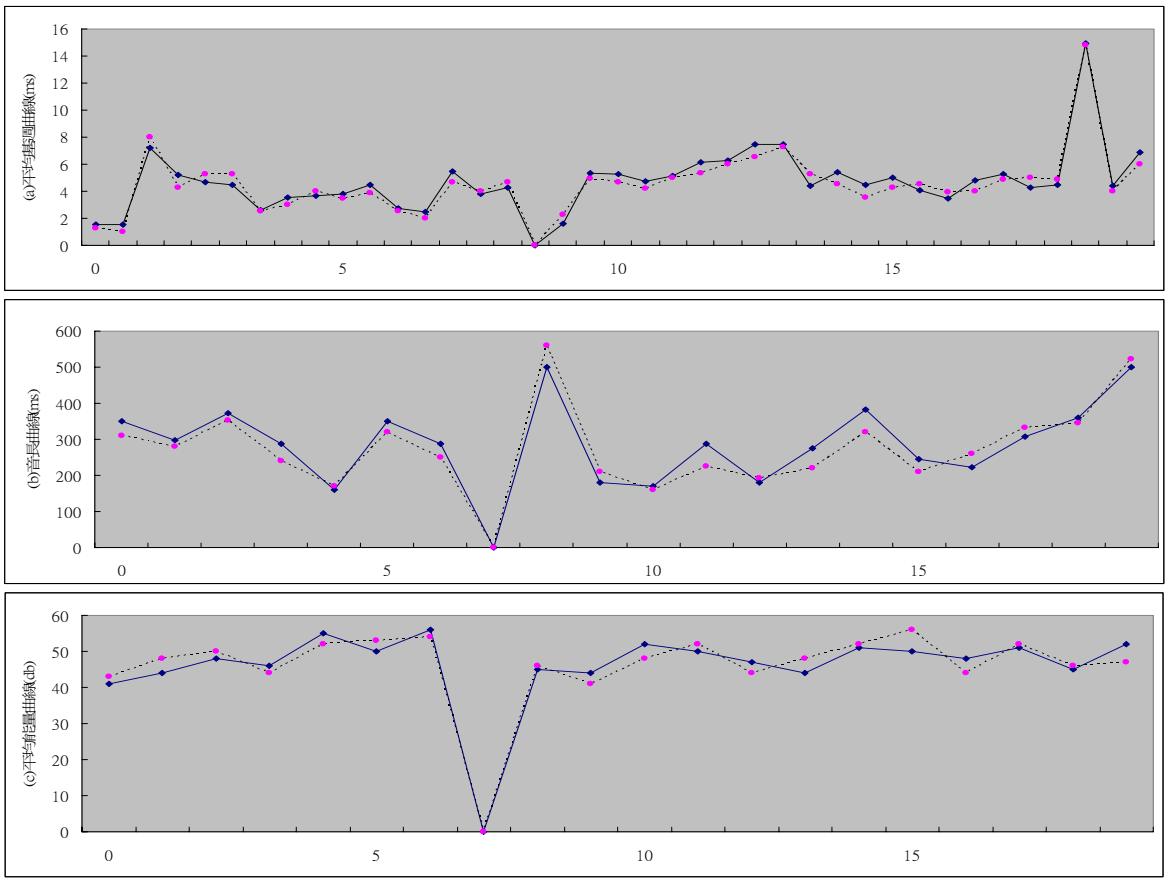

圖 8 快樂對照語句：我今天真的好高興，因為我的著作終於問世了。

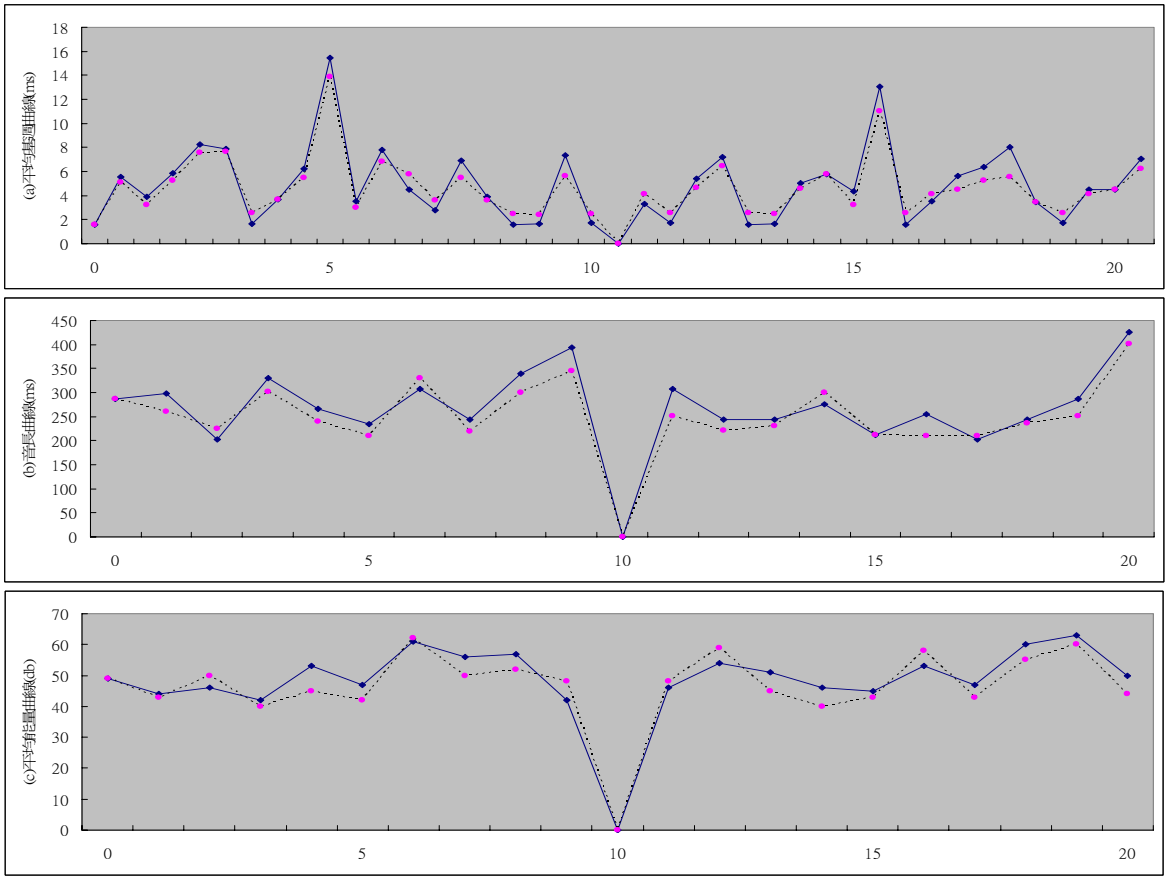

圖 9 生氣對照語句：這次旅行業者太過分了，居然可以不顧旅客安全。

表 3：情緒鑑定之例句

編號	情緒範例句
001	我今天真的好高興，我的書有不錯的銷售成績。
002	因為他的過世，我現在跟家屬一起哭泣擁抱。
003	這次的期末考再考不好，妳就完蛋了。
004	政府為了打擊犯罪，成立了聯合執行小組，顯示對於犯罪打擊不遺餘力。

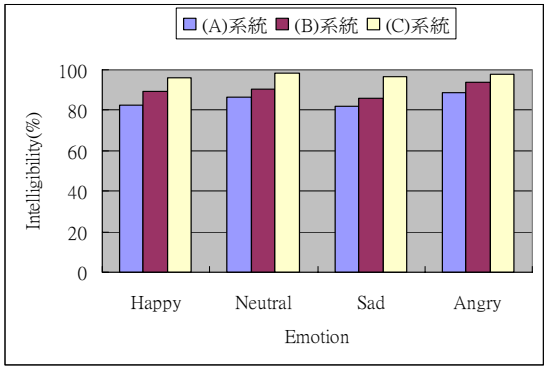

圖 10：理解度實驗結果直方圖

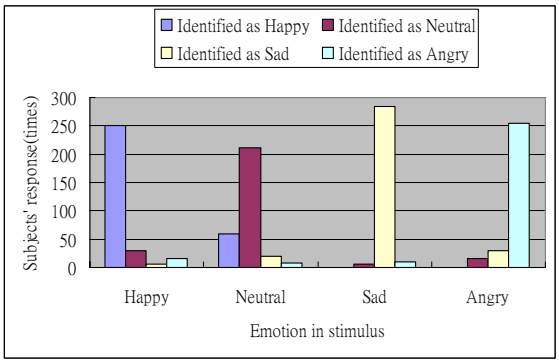

圖 11：情緒鑑定評估結果直方圖

6. 結論與未來展望

本論文中，提出了一個新的情緒語音合成系統的架構，此架構中，利用一套平衡語料挑選機制，設計並產生一套最小但包含足夠音節及常用詞資訊之情緒語料。除此之外，單元挑選方法中，不需要使用音韻模型數字化的去預測句子的音韻參數，相反的，將人類發音、構句的特性，與中文語音音韻、重音等現象，利用機率式句法結構將句子解構成一個樹狀結構，並利用樹狀結構上階層式的關係，選擇適當、合理的合成單元，在這個過程中，保留了原始語音的音韻特性。進一步，運用隱含式語意索引，計算出合成單元之間的語意失真度，挑選最適合的單元。

透過實驗評估，合成語音品質有不錯的表現，但仍有下列問題有待改進：1.)在語音的切割上，情緒語音的切割結果，相較於中性語音，有較差的斷點位置，主要是因為我們利用隱藏式馬可夫模型進行斷點切割時，並未考慮情緒因素。2.)語料式合成系統，若能收集足夠的語料，期能有較好的合成表現。3.) 在機率式句法結構模型中，會出現新詞問題（OOV）與文法規則不足（OOR）的問題，需要提出一套自動修正的方法，才能避免類似的問題。

參考文獻

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, “The AT&T Next-Generation TTS System,” in *Proc. of ICSLP’98*, Sydney, Australia, pp. 931-934,1998

[2] Jon Rong Wei Yi, *Corpus-Based Unit Selection for Natural-Sounding Speech Synthesis*, Ph.D. thesis, Massachusetts Institute of Technology, 2003

[3] T. Dutoit, *Text, Speech and Language Technology. vol.3: An Introduction to Text-to-Speech Synthesis.*, Kluwer Academic Publishers, Dordrecht, 1997

[4] W. J. Wang, W. N. Campbell, N. Iwahashi and Y. Sagisaka, “Tree-based Unit Selection for English Speech Synthesis,” in *Proc. of ICASSP’93*, Minneapolis, MN, vol.2, pp. 191-194, Apr. 1993

[5] X. Huang, A. Acero and H. W. Hon, *Spoken Language Processing*, pp. 133-190, Prentice Hall, 2001

[6] J.A. Russell, “Measures of Emotion,” in R. Plutchik and H. Kellerman (Eds.), *Emotion Theory, Research, and Experience*. pp. 83-111, Academic Press, N.Y., 1989

[7] A. Iida, “A Study on Corpus-based Speech Synthesis with Emotion,” Doctor of Media and Governance thesis, Graduate School of Media and Governance, Keio University, Sep. 2002

[8] R. Carlson, G. Granstrom and L. Nord, “Experiments with Emotive Speech, Acted Utterances and Synthesized Replicas,” *Speech Communication*, vol. 2, pp.347-355, 1992

[9] N. Frijda, *The emotions*, Cambridge University Press, N.Y., 1986

[10] L. K. Guerrero, P. A. Andersen and M. R. Trost, “Communication and Emotion: Basic Concepts and Approaches,” in P. A. Andersen and L. K. Guerrero (Eds.), *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*, pp. 3-27. Academic Press, San Diego, 1998

[11] C. C. Kuo, C. S. Kuo, J. H. Chen and S. C. Chang, “Automatic Speech Segmentation and Verification for Concatenative Synthesis,” *in Proc. of Eurospeech’03*, Geneva, Switzerland, 2003

[12] M. Chu, H. Peng, H. Y. Yang and E. Chang, “Selecting Non-uniform Units from a Very Large Corpus for Concatenative Speech Synthesizer,” *in Proc. of ICASSP’01*, vol. 2, pp.785-788, Salt Lake City, Utah, U.S.A., 2001

[13] C. H. Wu and J. H. Chen, “Automatic Generation of Synthesis Units and Prosodic Information for Chinese Concatenative Synthesis,” *Speech Communication*, vol.35, pp.219-237, 2001