The 2019 Conference on Computational Linguistics and Speech Processing
ROCLING 2019, pp. 368-377

應用文脈分析於中英夾雜語音合成系統

Linguistic Analysis for English/Mandarin Speech Synthesis System

洪翌翔　Yi-Hsiang Hung
國立屏東大學資訊科學系
Department of Computer Science
National Pingtung University
gbaian10@gmail.com

黃奕欽　Yi-Chin Huang
國立屏東大學資訊科學系
Department of Computer Science
National Pingtung University
ychuangnptu@mail.nptu.edu.tw

鄧廣豐 Guang-Feng Deng
資訊工業策進會
Institute for Information Industry, Taipei, Taiwan
raymalddeng@iii.org.tw

摘要

本論文將藉由文脈分析的處理，實作出一套中英夾雜的語音系統。在語音模型的建模上，採取統計式模型中的隱藏式馬可夫模型(Hidden Markov Model)做為基礎針對中文以及英文進行處理。在系統的實作中，首先在合成語音前先將文字做前語言處理切割成中文和英文的部分，接著將中文與英文分別已預先訓練好的的中文/英文之語音模型分別進行合成，最終將各自合成的部份進行語音段的串接。其中，由於中文以及英文為不同的語言，為了維持整段話的連貫性，若整個句子以中文句當作主體，並且將此中英夾雜句中的英文字的部份，透過其詞性分析(POS Analysis)找出其詞性後，將此英文字置換成與其詞性相同的中文字(Substitute Word, 縮寫為 SW)，使其與原英文字的詞性相同，在中文主體句中，則透過置換過後的中文句來進行文脈分析，挑選合適的中文語音模型，並用來為合成整段中文句子，並且將合成好的英文部分替換回該句中完成中英文夾雜的句子。透過實驗分析顯示，透過文脈的分析，能夠幫助合成的句子的語流較為順暢，因而提升中英夾雜句的何成語音更為自然。

關鍵字 ： 中英夾雜句、隱藏式馬可夫模型、文脈分析、語音串接、語音合成

Abstract

In this study, we analysis the effect of the linguistic information for the English/Mandarin speech synthesis system. In order to construct the acoustic models for both languages, we

adopted the Hidden Markov Model. For the system implementation, we firstly detected the language segments for each language of the input bilingual sentence, and then independently generate the feature sequences for each language. However, for generating fluent synthesized speech, the linguistic information should be taken into account. Here, if the bilingual sentence is mainly written in Mandarin with a few English words, we firstly analyze the Part-Of-Speech information for the English words. Then, we adopted some substitute words (SW) to translate the English parts into Mandarin which have the same POS tags as their corresponding English words. Finally, The entire sentence consists of only one language and could be analyzed linguistically and keep its context information. Finally, the synthesized speech should be more fluent since the contextual linguistic information is used for choosing the suitable acoustic model sequence. In order to construct the original bilingual speech utterance, the English segment is substituted back to the synthesized speech. Experimental results showed that adding the contextual linguistic information is indeed helpful for generating fluent speech for the bilingual sentences.

Keywords: English/Mandarin bilingual sentence, Hidden Markov Model, Linguistic analysis, Speech concatenation, Speech synthesis

一、緒論

(一)、研究動機

隨著世界朝向國際化的發展，不同語言之間的交流越來越盛行，不管是在學界、業界，都免不了會接觸到不同語言間的各式各樣問題，而某些特定的專有名詞翻成當地語言時，時常會有無法充實表達其原意的困境，亦或者可能發生類似繁中與簡中的翻譯完全不同甚至到了相反的意義(如:'行'與'列')。此時為了避免問題，我們常常會在語言中直接使用該詞的原本發音來溝通，這在人與人之間或許並沒有什麼太大的問題，但實際上因為中文跟英文不管在發音還是整個語言以即文字的結構上都有著非常大的不同，這使的如果要合成一句多語言的語音合成容易發生語調不順暢的問題，為此我們必須找一個方法來解決此問題。本論文提出一種依照前後文分析文脈關係以及根據其詞性來確認中文的發音方式，使得合成中英夾雜語句時能保持中文的整體脈絡，進而提升合成語音的流暢性與自然度。

(二)、相關研究

近年來語音合成系統廣泛被使用的主要有兩種，一是基於大型語料庫樣本做串接，如:單元選擇(Unit selection approach) [1]，另一種則是基於統計方法。如:隱藏式馬可夫模型(HMM-based approach) [2]。單元選擇合成雖然有著極佳的合成音質，但卻需要非常大量的語料庫做支持，所以在製作語料成本上有著極大的代價。而隱藏式馬可夫模型對語料庫的需求則不像前者需求這麼大。

語音相關的應用也非常多，例如在不同語速下的應用 [3]、合成歌唱合成系統 [4]、情緒轉換 [5]、多語言語音合成 [6]、基於深度神經網路在多語言間的應用 [7]等，其中關於 [7]DNN 部份所合成的多語言系統為使用單一合成器，並且在合成出來的語音上有著不錯的結果，但其缺點是必須選擇語系相似的語言，且通常要找到精通多語言的相同語者是困難的，對此問題也有 [8]這些使用類似音素來進行不同語言間資料的補足的方法，且中英合成中也有使用混合兩語言決策樹的方法。

(三)、隱藏式馬可夫模型系統概述

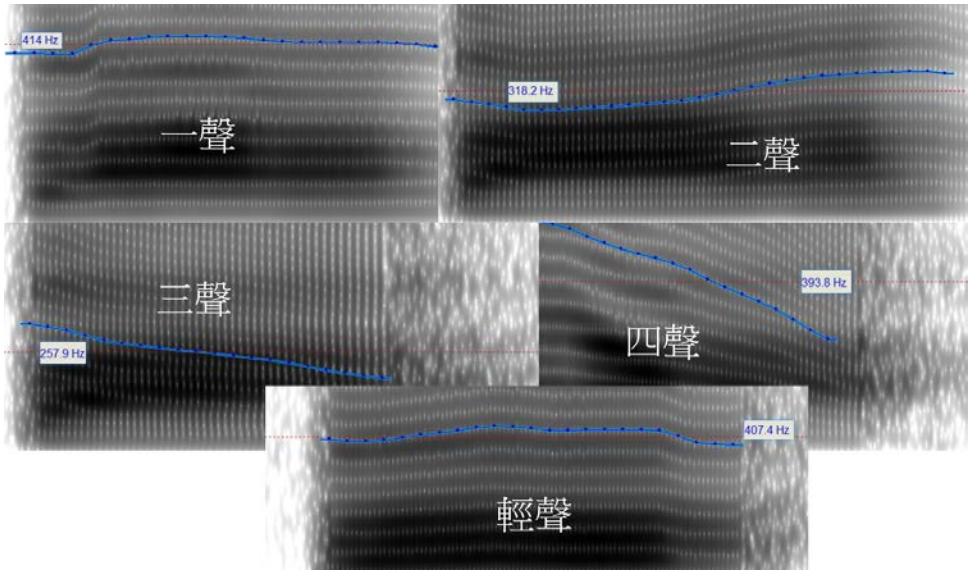
近年來隱藏式馬可夫模型在語音合成的領域中已經有著舉足輕重的地位，因為其所合出的語音流暢度已經不亞於傳統串接合成的結果，且其所訓練出來的模型所需要的儲存空間也相對較小，有較高的攜帶性。在本文中所使用的基於隱藏式馬可夫模型的語音合成系統(HMM-based Speech Synthesis System, HTS)，是由 HTS 工作團隊 [9]所研發，該技術由 Hidden Markov Model Toolkit(HTK) [10]研發修改而來，HTS 團隊提供了一個便於開發的研究平台，並有效幫助 HMM 的訓練。

本研究目標為建立一套基於 HMM 的語音模型，主要可分為訓練與合成兩個階段。訓練階段時由聲音語料藉由 SPTK [11]估算頻譜及音韻參數，聲音語料所相對應的文字經由文字分析器(Text analysis)產生相應的文脈訊息，在經由問題集(Question set)分類樹產生與文脈訊息相關的 HMM 模型。合成階段將欲合成的語音文字當作輸入丟入文字分析器中產生相應的文脈訊息，再經由問題集分類樹找出該段文字對應的 HMM 模型序列，再經由 HMM 模型產生對應的頻譜及音韻參數，最後合成出所需的語音訊號當作輸出。

(四)、章節概述

本論文主要敘述中英夾雜語音合成系統為了確保連貫性，而使用 SW 替換英文字來合成整句中文字以保持整體中文脈絡，本論文主要分成五節:第一章 ： 緒論，主要說明研究動機、相關研究與討論、以及系統概述。第二章 ： 中英文語音模型，定義分別介紹中

文、英文語音模型的定義。第三章 ： 中英夾雜語音合成系統實作，詳述中英切割的方法、中文音素模型的建立、前後文相關之問題集決策樹。第四章 ： 實驗結果及討論，說明實驗目的、語料庫來源以及內容、分析及討論結果。第五章 ： 結論，總結整篇論文的結論。



圖一、中文五聲變化

二、中英文語音模型定義

(一)、中文模型定義

中文約有 420 個不含聲調的基本單元，加入五聲變化後，則有超過 1200 個含聲調的單元。若直接對這約 1200 個模型進行訓練或許可以達到不錯的結果，但是這需要非常非常大量的語料庫來訓練才有可能，若資料量不足可能導致某些音訓練不足難以發出正確的音調，甚至可能發生某些音連一個資料都沒有的情況。

為了解決模型過多導致訓練不足的問題我們必須盡可能壓低模型的數量，本論文中，我們以“Segmental Tonal Phone Model” (STPM) [12]做為定義我們中文模型的方法，對此我們必須將模型考慮至聲母、韻母以及五聲來進行設計。然而五聲的變化差別在音高上，所以要區別五聲就必須得觀察五聲音高的的變化，圖一分別為五聲的頻譜圖以及其音高(圖中聲音分別為:巴、拔、把、爸、吧)。如圖所示，音頻的範圍大概可分為高音頻範圍(H)、中音頻範圍(M)、低音頻範圍(L)，五聲的變化趨勢分別為:

一聲:H→H 、 二聲:L→H 、 三聲:L→L 、 四聲:H→L 、 輕聲:M→M

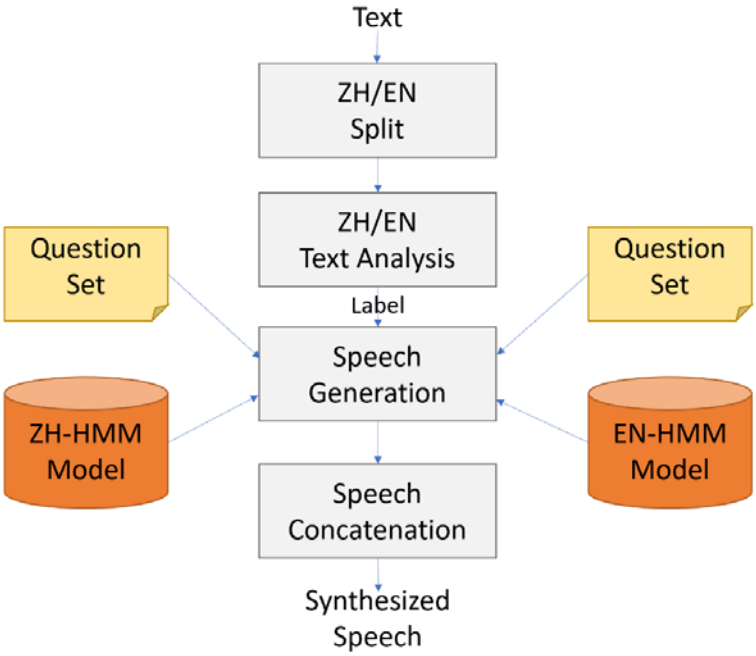
為了能表達五聲的變化趨勢，我們將中文的基本單元分割成三個音素模型:

C+V1+V2

其中 C 為第一個音素模型用來表示聲母的音，V1+V2 則同時用來表示韻母以及五聲變化趨勢的前(V1)後(V2)，以此方法最後我們可能如下圖 107 個音素模型(含一個 pause 模型)。相對於原本將所有中文約 1200 多個音的模型數量，此方法大大減少了模型的數量

表一、音素模型範例

	Syllable	C	V1	V2
ㄏㄨㄟˋ	huei4	hu	eiH	erL
ㄕˊ	shr2	shr	shrL	shrH
ㄧㄡˇ	yiou3	yi	ouL	ouL0



圖二、中英夾雜合成系統流程圖

(二)、英文模型定義

英文模型方面則直接基於 ARPAbet 標示並以國際音標(IPA)為準來做為我們的音素模型，其中包含 13 個元音、3 個雙元音、27 個輔音，共 43 個音。元音又因發音位置可細分至前、中、後元音，輔音也可依發音位置分成塞音、擦音、半元音、鼻音、塞擦音等。其中值得注意的是音標中的[l]、[m]、[n]雖然在音標內是同個樣子，但實際上根據是否在母音前，看似相同的音標會有不同的發音，這在 ARPAbet 標示法中需以不同的記號

做標記。

三、中英夾雜語音合成系統實作

本系統在合成文字前須先將原始文字做中英文字切割的前處理，在得到中英文後，先將整個句子中的英文以一個可識別的 SW 做替換並將整段文字丟入中文合成器直接合成出整段中文句子以保留整段中文的文脈資訊。英文單字則丟入英文合成器做合成，最後在將合好的英文單字取代前面的 SW 做語音串接，整個架構的流程圖如圖二所示。

(一)、中英夾雜文字切割

使用者輸入文字後(圖二. ZH/EN Split 的步驟)，本系統主要以 Unicode 來識別中英文字的區別。在 Unicode 中中文的範圍為 19968~40869，其餘部分包含半形空白當作英文部分(非中文)。

當輸入一段文字後，逐一由前至後針對每一個文字做中英文字判別，並使用一個變數來儲存當前文字屬於中文還是英文的狀態做紀錄，同時有兩個陣列分別來儲存由上一次狀態轉換前至現在的中/英文字收集器，當遇到狀態轉換時候就代表中英文字做了切換，此時就把所有文字收集器的中/英文丟入結果集內，當所有文字都判別完畢後，在把文字收集器內的結果放入結果集就將原始句子切割完成。

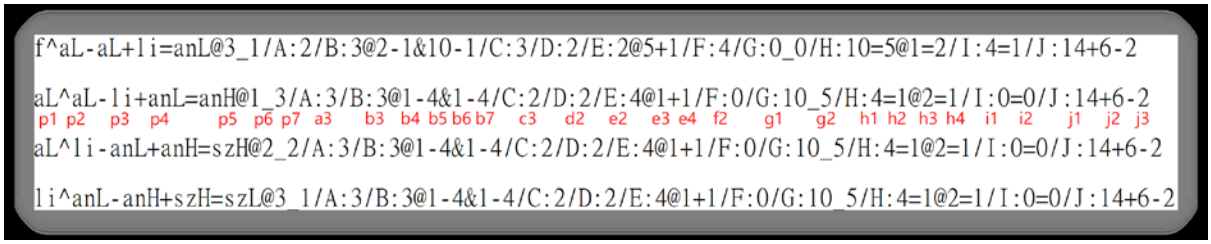
(二)、語音串接

1、文脈分析

為了使中英文連接處有著更自然的發音與韻律，我們必須保有整句中文字中的前後關係並記錄下來，因此在合成語音前必須先對文字做前處理(圖二. ZH/EN Text Analysis)，為此我們必須定義以下中文文脈資訊，並加以記錄成 Label 檔。細節如下所列: **音素(Extended Final):**在中文中即表示一個聲母/韻母/介音，如:ㄅ、ㄆ、ㄧ、ㄨ、ㄢ、ㄥ ；**音節(Syllable):**在中文中即表示一個字，如:我、剛、回、家；**音韻詞(Prosodic Word)** 多個音節組合成的一個詞，如: 螞蟻、無法、我們；**音韻短語(Prosodic Phrase):** 多個音韻詞組成的一小段話，通常兩個短語間會有停頓，所以短語很容易發生在連接詞上；**句子(Utterance):** 一個以上的短語組成一個句子，即我們要合成的目標語句。以上文脈資訊會依照其前前一個單元、前一個單元、當前單元、下一個單元、下下一個單元以及當前單元層級由前至後和由後至前數來的位置都納入考慮，並將以上的韻律階層可由分類回歸決策樹(Classification and Regression Tree, CART) [13]得到。

藉由以上文脈資訊記錄，我們可由 Label 檔中看出整句話的斷詞是如何斷的，在圖

三中，原句子為”家裡網路出了問題無法連伺服器”當中的’連’字，由 e2 可知當前音韻詞由四個組成，由 b4 可知他是由前數來第一個音節，即代表’連’在斷詞中被斷成”連伺服器”；英文方面則是在文脈分析上與記錄音節及重音(stress)方面[14]，與中文有著類似的記錄方法。



圖三、Label 檔範例

2、問題集

根據中文模型定義以及上述所記錄的文脈資訊，便可開始設計問題集之決策樹以讓模型達到最佳狀態，對此我們將對音素模型考慮以下五大類問題 [15]

(1) 音素相關(Phoneme related): 若為韻母 : 其音高範圍(H/M/L)；若為聲母 : 單一或由兩個音素組成(即是否含介音)；聲母發音類別 : 塞音、塞擦音、鼻音、擦音、邊音、唇音、舌尖音、舌根音、舌面音、翹舌音、齒舌音；韻母發音類別 : 單韻、複韻、聲隨韻、捲舌韻；音素在音節中的位置 : 由前/後數來位置。

(2) 音節相關(Syllable related): 音節中音素的數量 : 考慮前/當前/下一個音節；在音韻詞、音韻片語中的位置 : 由由前/後數來位置。

(3) 音韻詞相關(Prosodic Word related): 音韻詞中音節的數量 : 考慮前/當前/下一個音韻詞；音韻詞在音韻片語中的位置 : 由前/後數來位置。

(4) 音韻短語相關(Prosodic Phrase related): 音韻片語中音節、音韻詞的數量 : 考慮前/當前/下一個音韻片語；音韻片語在整段話的位置 : 由前/後數來位置。

(5) 句子相關(Utterance related): 句子中音節、音韻詞、音韻短語的數量。

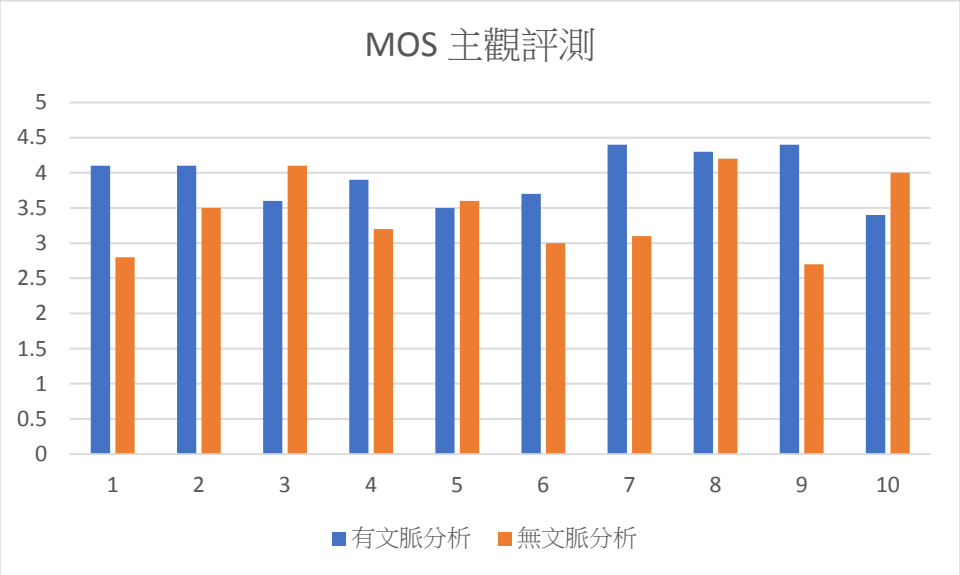
當其在訓練階段透過問題集分類並訓練完成模型後，在合成語音接段時就可再次從此問題集(圖 3. Question Set)中從所需合成文字的 Label 中找到其對應的模型並合成該語音，最後在將中英文部分合併，由英文字取代 SW 並得到最中的語音輸出結果。

四、實驗結果及討論

(一)、實驗目的與語料庫介紹

實驗將請受試者分別對每個句子來評論有分析並考慮前後文且記錄文脈訊息來一次合成整個句子後再替換掉 SW 所合出的語音，以及與沒有分析前後文並分開合成每段文字後直接合併的語音，比較哪一個句子較為順暢並評分，但由於實驗中英夾雜語句時容易使受試者混淆，所以本實驗僅以純中文並無將 SW 換回英文的方法來讓受試者接受本次測試。在實驗中所使用的語料庫分為中英兩部分。中文語料庫使用資策會所錄製的女性語者，主要內容為新聞語料，總共有 5102 個句子，92388 個字，英文語料庫方面使用 CMU ARCTIC 語料庫[16]。共有 1132 個句子，其中包含 10045 個單字(2974 個不重複的單字)，39153 個音素。

本次實驗目的在於確認有分析前後文文脈是否會影響句子的流暢性，評分方式採平均主觀值分數(Mean Opinion Score, MOS)，由 10 位受試者對 10 組隨機順序的不同的語句做評分，受試者成員包含同班同學、指導老師、社交平台上的朋友等，語音的自然度越高則給越高分(最高 5 分)，越不自然則給越低分(最低 1 分)，最後得到的 10 組句子的平均分數長條圖如圖四。其中，經由計算平均的結果後發現，有文脈分析較無文脈分析平均來得高(3.94 vs. 3.42)



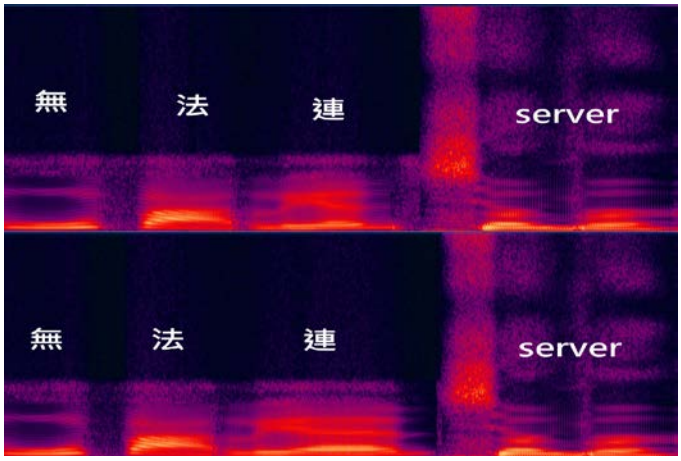
圖五、MOS 主觀評估

(二)分析與討論

從實驗平均和長條圖來看，多數句子都顯示有分析文脈後的結果較好。以下將用一個正式的中英夾雜句子作為範例來觀察頻譜圖和文脈中如何知道為什麼不分析前後文會導

致句子不自然。

在此使用圖四中文脈的句子”家裡網路出了問題無法連伺服器”，但實際上其實我們想合的是 Server 而不是伺服器，所以當使用者輸入”家裡網路出了問題無法連 Server”時候，我們的文字前處理器會先判斷 Server 是個名詞並換成一個作為名詞用的 SW 後再合成整句中文，此時文脈中的’連’被斷成”連+某個名詞 SW”，其被當為動詞後面有個受詞。但如果我們沒有使用文脈分析，則合成句子時候會先合成”家裡網路出了問題無法連”，再另外合成一個 Server 後直接做串接，此時因為’連’為最後一個字，後面並無受詞也沒有其他句子使其成連接詞，最後斷詞就將他斷成”無法連”，後面原本該有的名詞被落單成一個單字了，使的整句頓時失去了連貫，其兩者頻譜圖比較如圖六，圖中上半不為有分析文脈，下半部則無分析，可看出無分析文脈時，連與 sever 上出現了斷層。最後使的整句話念起來不通順，由此可見文脈分析對於一句話的重要性。



圖六、文脈分析頻譜圖比較

五、結論

本論文為中英夾雜語音合成系統，在一句中英混雜的語句以 SW 替換英文字並藉此合出整句考慮前後文關係的中文以保證句子的流暢性，最後在以合成的英文換掉替換用的 SW 來還原原本的中英夾雜句子。實作方法以隱藏式馬卡夫模型並在音素模型以及句子的文脈資訊使用問題集之決策樹來進行模型優化。

在實驗數據上顯示，本論文所提出的將句子整句合成在將字替換回原本的字，確實比每段字分開合成後在合併來的好，因為其保有了整句話的流暢性而非像是各個單字拼接而成。

參考文獻 [References]

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, in *Acoustics, Speech and Signal Processing (ICASSP)*, vol.1, pp. 373-376, 1996.

[2] Yoshimura, Takayoshi, et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis." *Sixth European Conference on Speech Communication and Technology*. 1999.

[3] 江振宇, 黃啟全, 王逸如, 余秀敏, & 陳信宏. (2012). 可變速中文文字轉語音系統. *中文計算語言學期刊*, *17*(1), 27-41.

[4] Ju-Yun Cheng, Yi-Chin Huang, and Chung-Hsien Wu. "合成單元與問題集之定義於隱藏式馬可夫模型中文歌聲合成系統之建立 (Synthesis Unit and Question Set Definition for Mandarin HMM-based Singing Voice Synthesis)." *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*. 2013.

[5] 吳尚鴻. "基於隱藏式馬可夫模型之中文語音合成與吼叫情緒轉換." 清華大學電機工程學系所學位論文 (2010): 1-66.

[6] Chia-Ping Chen, Yi-Chin Huang, Chung-Hsien Wu, & Kuan-De Lee. (2014). Polyglot speech synthesis based on cross-lingual frame selection using auditory and articulatory features. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1558-1570.

[7] Reddy, M. Kiran, and K. Sreenivasa Rao. "DNN-based Bilingual (Telugu-Hindi) Polyglot Speech Synthesis." *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018.

[8] Qian, Y., Cao, H., & Soong, F. K. (2008, December). HMM-based mixed-language (Mandarin-English) speech synthesis. In *2008 6th International Symposium on Chinese Spoken Language Processing* (pp. 1-4). IEEE.

[9] Zen, H., Nose, T., Yamagishi, J., Sako, S. and Tokuda, K., The HMM-based Speech Synthesis System (HTS) Version 2.0, 2007. http://hts.sp.nitech.ac.jp/

[10] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X.Y., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., The Hidden Markov Model Toolkit (HTK) Version 3.4, 2006. http://htk.eng.cam.ac.uk/

[11] SPTK Working Group, “Reference Manual for Speech Signal Processing Toolkit Ver 3.3,”. http://sp-tk.sourceforge.net/

[12] T. Lin, and L.-J. Wang, “Phonetic Tutorials”, Beijing University Press, pp. 103-121, 1992.

[13] Hsia, C. C., Wu, C. H., & Wu, J. Y. (2010). Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(8), 1994-2003.

[14] Zen, H. (2006). An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*, *133*.

[15] 謝雲飛, “語音學大綱”, 臺灣學生書局, 民國 63 年.

[16] Kominek, John, Alan W. Black, and Ver Ver. "CMU ARCTIC databases for speech synthesis." (2003).