國立交通大學
National Chiao Tung University

出國報告（出國類別： B 類：出國短期研究）

題目：整合韻律訊息之自動語音辨認研究

服務機關：電機工程學系
姓名職稱：江振宇 博士後研究員
前往國家：美國亞特蘭大喬治亞理工學院
出國期間：2012/01/02~2012/03/31
報告日期：2012/03/31

一、摘要

江振宇博士受「美國喬治亞理工學院-電機與電腦工程學院的信號與影像處理中心」李錦輝(Prof. Chin-Hui Lee)教授邀請，於民國 101 年 1 月 2 日~民國 101 年 6 月 1 日這段期間，至該中心進行訪問研究，將進行「整合韻律訊息之自動語音辨認研究」，於李錦輝老師提出的 detection-based ASR 架構下，整合江振宇博士提出的韻律模型，以期突破目前以 HMM 爲基礎之語音辨認技術。目前已完成建立基礎 5K 詞彙量之連續語音辨認器基礎系統實驗，亦詳細分析其辨識錯誤訊息，已完成韻律訊息改善辨認率之評估，並開始建立針對英文語音辨認所使用的韻律模型，與 2012 年本月 3 月 19 日起與另一位訪問學者 (Prof. Sabato Marco Siniscalchi, Department of Telematics, Kore University of Enna, Enna, Italy) 合作完成 A Bottom-Up Stepwise Prosodic Knowledge Integration Approach to Large Vocabulary Continuous Speech Recognition Using Weighted Finite State Machines 之工作，結果預計投稿明年之 ICASSP2013 國際會議。

二、目次

一、摘要 2

二、目次 3

三、本文 4

(一) 目的 2

(二) 過程 6

(二) 心得及建議 13

三、本文

（一）目的

利用韻律訊息於輔助語音辨認，在這幾年的語音處理研究中越來越受重視，其中最主要探討的議題有 1.加入韻律訊息的語音辨認架構設計(Framework of automatic speech recognition assisted with prosodic information)、2.韻律模型的設計及建立(design and construction of prosodic model)、3.韻律相關之聲學模型(prosody-dependent acoustic model)、4.韻律訊息相關之語音模型(prosody-dependent language model)、5.韻律模型對不同語者的調適(speaker adaptation for prosodic model)。我們基於先前對於「非監督式中文韻律標記及模式的研究」，近期已將此研究的韻律模型應用於輔助中文大詞彙連續語音辨認，並得到不錯的結果，大幅提升了詞、字以及音節辨認率，另外，在 prosody-dependent acoustic model 方面，近期亦利用其中的韻律斷點資訊建立 break-dependent 的聲學模型，在音節辨認的實驗裡可以得到些許的音節辨認率提升，亦可在 syllable lattice 上得到較高的正確音節涵蓋率(coverage rate)，另外的好處是可以在 syllable lattice 上提供 break 資訊，以利後級進行語言及韻律資訊整合，以得到更好的詞辨認率。

另一方面，目前語音辨認系統仍是以「聲學模型」和「語言模型」爲基礎的架構，其中聲學模型是以「隱藏式馬可夫模型」(Hidden Markov Model, HMM)爲主，而語言模型是以 n-gram model 爲最主要的實現方法。然而，此架構雖然能達到不錯的表現，但仍有許多基本的問題尚未解決，如：

1.以 HMM 爲基礎之聲學模型是以音或音素做爲模擬的目標，沒有直接利用發音位置及方法的強建語言屬性做爲模擬對向，另外由於 frame-based 處理的限制、segment duration 無法輕易模擬等原因，造成目前 state-of-art 的 HMM 模型能力已發展到極限。

2.聲學模型、語言模型在現行的語音辨認架構下，各自以常數之權重貢獻其分數，許多正確辨認出之語音單元，不是因爲聲學模型的鑑別貢獻其能力，而

是因爲語言模型的過度補償誤打誤撞得到的結果。這些聲學模型或語言模型，通常對於訓練語料或手上可得到的實驗語料是有其辨識能力，然而，當應用於實際應用的語音辨認環境，因模型未能真正學到信號或語言特性，表現即受影響。

3. 目前的 HMM 聲學模型和 n-gram 語言模型皆是以 data driven 的方法訓練出來的統計式模型，優點在於目前可以輕易套用現成的 integrated search algorithms 便可以建立系統，但語音辨認不只是聲學模型和語言模型，其他重要語言知識資訊如：韻律、發音屬性特徵，並未整合至語音辨認的架構，雖然理論上可利用於語音辨認的語言學知識很多，但實際可用於語言辨認的技術，少之又少，原因是目前的研究風潮傾向相信統計方法，因爲統計方法有發展完整的理論，且簡單實作，因此忽略語音科學方面的知識進展。

由於以上的原因，喬治亞理工的李錦輝教授，於 2005 年左右，提出了一個稱爲「Detection-Based ASR」爲架構的語音辨認系統，此架構追本朔源地以 Human Speech Recognition 爲出發點設計，分爲 knowledge detection、integration、knowledge source，knowledge detection 目的爲偵測語音辨認所需要的訊息，這些訊息包含發音方法(Manner of articulation)、發音位置(place of articulation)、韻律訊息(prosodic information, i.e. stress, accent, prosodic boundary etc.)、spike train 等等。integration 的目的在整合 knowledge detection 得到的資訊，利用一些 pattern recognition 方法引入 knowledge source 進行整合，得到語音辨認的目標單元，如 phone、syllable 或 word。knowledge source 是指特定語言(language dependent)或所有語言(language independent)的特性或模型。在此架構下，語音處理中不同領域的專家，便可以專注於設計其擅長之 knowledge detector，並提供其熟悉之 domain knowledge，以便進行 knowledge integration，以期達到更佳之語音辨認效果。

基於上述之背景， 本計畫目的在於結合 1.李錦輝教授之 detection-based ASR 架構、2.交大語音處理實驗室所建立的韻律模型，擬將韻律訊息整合至

detection-based ASR 中，提供頻譜參數、語言參數以外的 knowledge source，期望韻律訊息能提供一個有效連結語音單位(或 segment unit)和高階語言參數(word)的橋樑，於 integration 時提供 suprasegmental 的資訊，使語音辨認系統能因爲韻律訊息的加入，達到較佳之語音辨認率。

（二）過程

本人於 2012 年的 1 月 2 日抵達喬治亞理工電機與電腦工程學院的信號與影像處理中心，當天便完成該學校訪問學者之報到手續，由於抵達該校的時間適逢美國的新年假期，因此起初的兩周 (2012 年 1 月 2 日~ 2012 年 1 月 19 日) 李錦輝教授並不在學校。本人於該中心臨時安排的工作座位(cubicle)後，於 2012 年 1 月 10 日開始進入該中心位於亞特蘭大第五街 75 號的 Centergy One 研究大樓，開始進行研究，正式的辦公室於於 1 月 23 日開始給本人使用。另外，由於此次訪問研究是由 2012 年 1 月 2 日至 2012 年 6 月 1 日，所以在剛到喬治亞理工的剛開始兩個禮拜，便開始積極找尋住宿地點，在 1 月 17 日左右確定住宿的地點，到此一切生活上住、行的問題皆已解決。

由 2012 年 1 月 19 日開始與李錦輝老師正式 meeting，固定每周三下午 5:00 到 6:00 進行密集的討論，以下分別以時間順序對於這段時間討論及初步結果進行簡述：

(1) 2012 年 1 月 2 日~2012 年 2 月 1 日

- 與李錦輝老師討論 HMM-based Text-to-Speech System 相關技術內容，評估採用 speaker independent acoustic model 進行語者聲音調式的可能性，本人並開始整理 TIMIT 語料庫之相關語音及語言參數，準備開始訓練 speaker independent 的 HMM-based speech synthesizer。
- 廣泛蒐集並研讀李錦輝老師有關 Detection-based ASR 的相關論文，討論以下兩種 detection-based ASR 整合 prosody knowledge 的方法，第一種方法是設計韻律事件偵測器，由語音信號中的 F0 及 energy 資訊偵測韻律事件，如 Fig.1 所示，擬於 detection-based ASR 中，於 knowledge detector 部分加入 detectors of prosodic

events (DPEs)，這些 prosodic events 包括：stress、accent、phrase boundary 等等，DPEs 將利用語音信號中抽出的 F0 資訊及頻譜資訊偵測 prosodic events (PEs)，提供 PEs 的機率或分數提供後級的 knowledge integrator 使用，而 knowledge integrator 將利用 PEs、發音資訊(articulation)、詞典(lexicon)以及語言模型(language model)等訊息，使用相關的模型、利用 pattern recognition 技術將欲辨認之語言單元(linguistic unit)辨認出來。

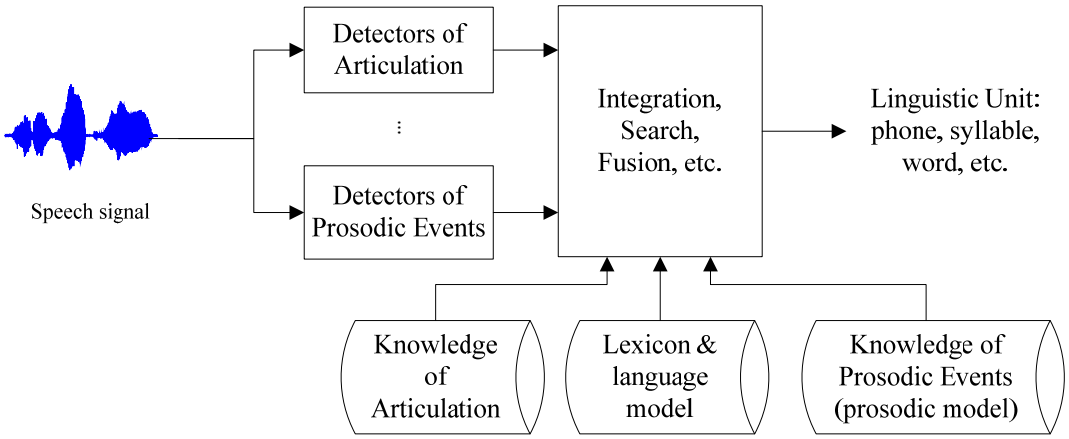


Fig. 1: A scheme diagram of the prosody-based knowledge integration for automatic speech recognition

第二種方法，是採用 Bottom-Up Stepwise Prosodic-Knowledge-Integration Approach，如 Fig. 2 所示，Attribute Detectors 由語音信號中偵測出每個 frame 屬於哪種發音屬性，再藉由 Evidence Merger 將這些發音屬性的偵測結果轉成 language dependent 的 phoneme 事後機率，接下來用 phoneme access module 利用 finite state machine 將 frame phoneme 事後機率 sequence 展開成 phoneme network，此 phoneme network 的 search space 會相當的大，因此將使用 phone-level 的 prosodic model 對這個 phoneme network 進行重新計分和刪除不可能的 node，在此同時，也可利用 phoneme network 上的語音切割資訊，以 phoneme 時長資訊對於語者的語速進行估計，以便於加入 prosodic model 時對於 prosodic model 進行語速調適或是對於韻律特徵參數進行正規化。除了對於語者的語速進行估

計及正規化以外，也必須對語者的基頻軌跡進行正規化。

接下來將 phoneme network 和 lexicon 進行 compose 將 phoneme network 展開成 word network，這部分也將會把 search space 大量的增加，因此也必須利用 Intra-word prosodic model 將不可能的 lexicon entries 刪除或是重新計分，期望能降低 word network 的大小，並且將 word network 中不符合韻律特性的詞，以 intra-word prosodic model，將其分數給予逞罰，降低後級語言模型 (LM) 造成混淆的可能。最後一級是進行 LM 及 prosodic model 重新計分，此階段便可以使用跨數個詞的超音段 prosodic information，將原本以 LM information 無法辨識正確的結果進行重新計分，以期增加詞辨識率。

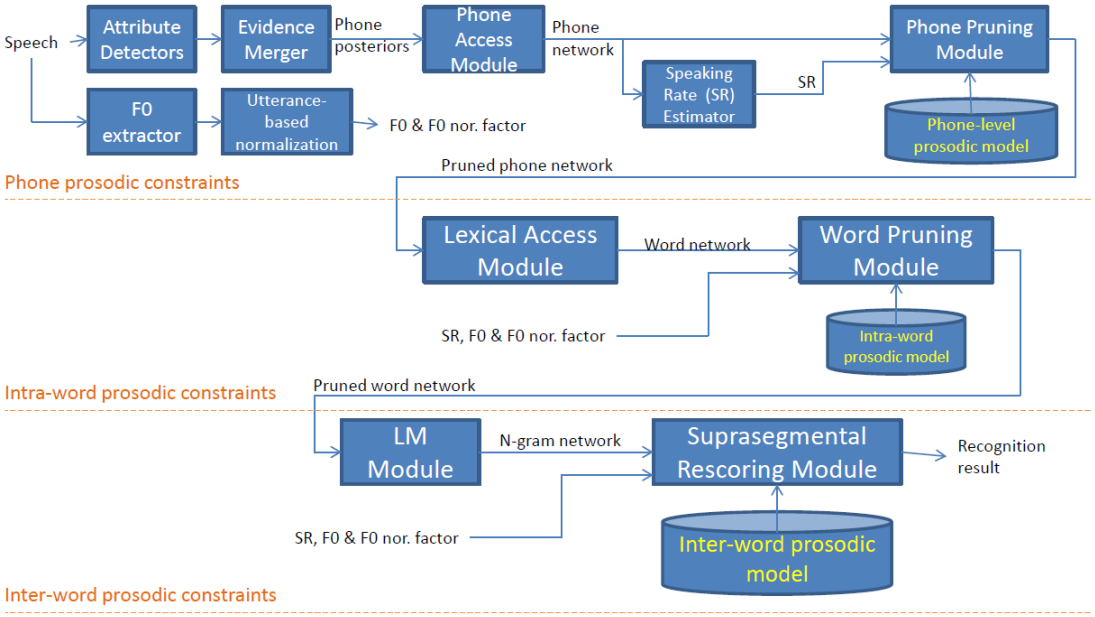


Fig. 2: A bottom-up stepwise prosodic knowledge integration approach

經過詳盡的討論後，決定以第二種「A bottom-up stepwise prosodic knowledge integration approach」方法進行「整合韻律訊息之自動語音辨認研究」。

(2) 2012 年 2 月 2 日~2012 年 3 月 1 日

這段期間，本人開始建立英文語音辨認的基礎系統，藉由分析其辨識錯誤的地方，加強 prosodic model 的設計。此基礎系統使用 WSJ0 SI-84 進行 cross word

tri-phone AM的訓練，語言模型使用NIST baseline bigram/trigram-backoff language model，測是語料爲 WSJ0 Nov92，共有 8 位語者，330 個 utterance，詞彙量爲 5K，表一爲其基礎辨識率之辨識結果。

表一：WSJ0 Nov92 測式語料辨識結果

LM	Correctness	Accuracy	Hit	Deletion	Substitution	Insertion	Total
Bigram	93.82	92.19	5022	32	299	87	5353
Trigram	95.89	94.60	5133	22	198	69	5353

此結果已達到要求之基礎系統效能，並於相關文獻知結果相符，以此結果本人開始進行辨識錯誤之分析，歸納出以下幾種錯誤：

(1)詞頻率高的「冠詞」、「指代詞」或容易發生互相取代行錯誤 (substitution)，如 the、that、a，這些詞通常發音變異較多，韻律上表現出短時長、低音高、低能量。

(2)「動詞時態」、「單複數」容易產生 substitution，如 price-priced、unit-units

(3)母音相同或類似的單音節詞容易產生 substitution，如 too-to、cards-car、fight-cite、lost-last 等等。

(4)單音節的 function word 容易互相產生 substitution，如 a-of、of-at、but-a、and-in、a-it 等等，推測其原因是因爲英文中的 function word 在韻律上表現出短時長、低音高、低能量，並且發音變異較多，因此辨認困難。

(5)一個詞辨認成多個詞。此種錯誤通常是發生在詞頻低的多音節詞，LM 的分數過於強調，以致正確的詞被辨認錯誤。

(6)多個詞辨認成一個詞。此種錯誤發生於發音及韻律上容易 linking 的 word sequence，如 it is-as、has a-hasn't、is in-isn't 、expect it-expected 等。

(7)詞串辨認成另一個詞串。此種錯誤常是因爲兩種詞串之發音類似或一樣，且 LM 比較 favor 某個詞串的機率，造成詞邊界的判定錯誤，如：a move to base it-on the two basic、to end-two and、very little we can-really can't、it may

curb-and maker of。

(8)刪除性錯誤 (deletion)的情形很少，皆發生在 function words

(9)插入性錯誤 (insertion)的情形較 deletion 多一些，亦皆發生於 function words

歸納以上的錯誤分析，prosodic model可能對於以上第(5)、(6)、(7)有所助益，Figs. 3-8 爲針對韻律訊息可能輔助語音辨認之範例，其中Fig. 3 將被李錦輝老師於今年 ICASSP 2012 (2012/3/28) 擔任 plenary talks 的 keynote speaker 時(http://www.icassp2012.com/CLee.asp) 簡報時使用，用來解釋目前 prosody-knowledge integration的實用潛力。

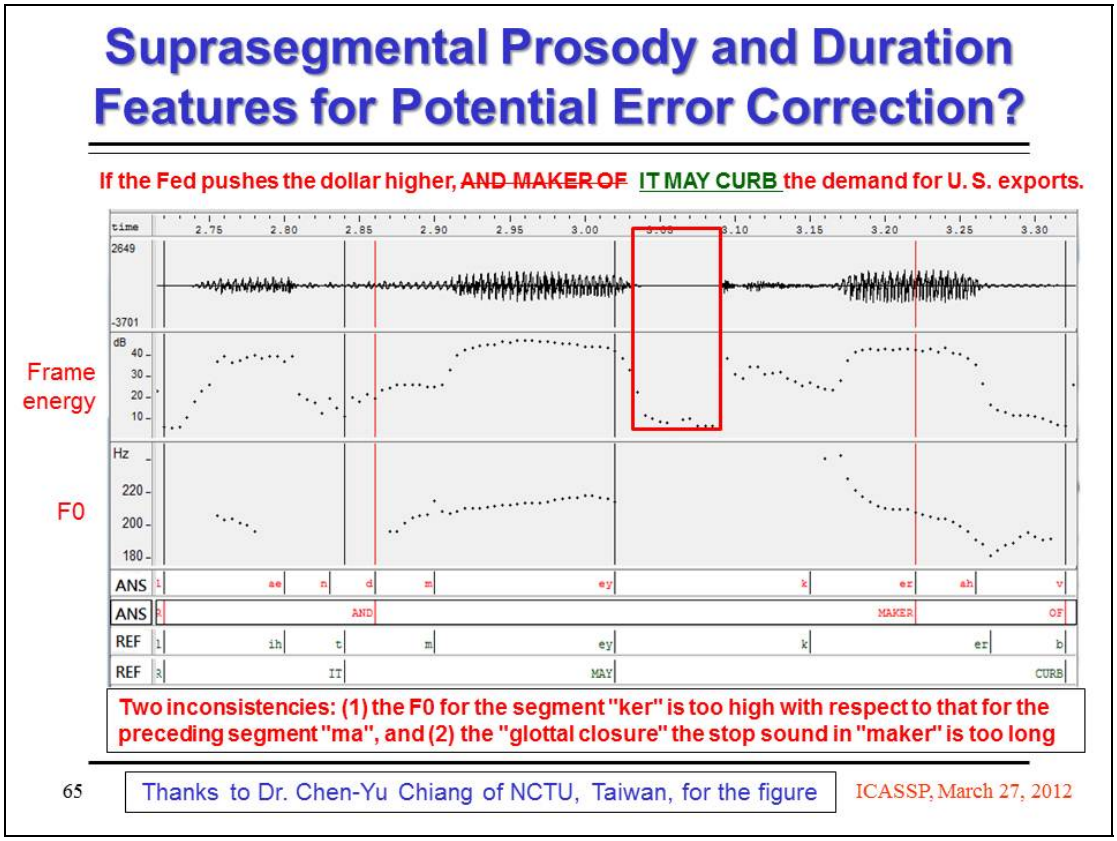
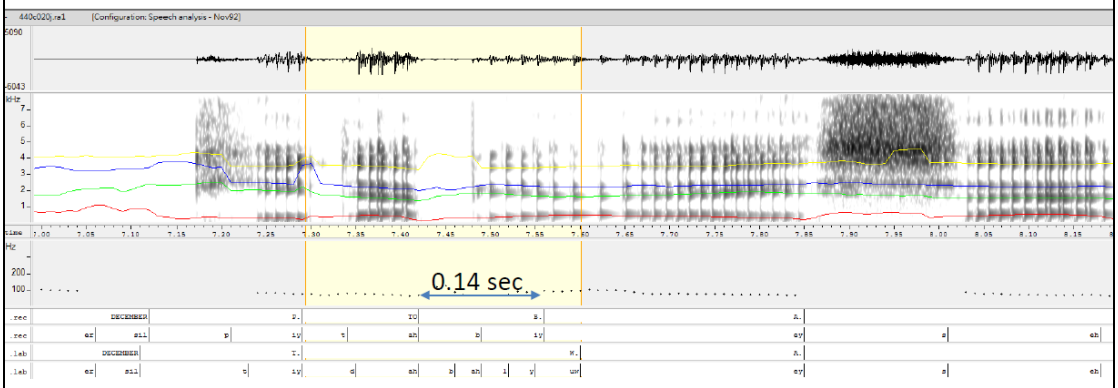


Fig.3

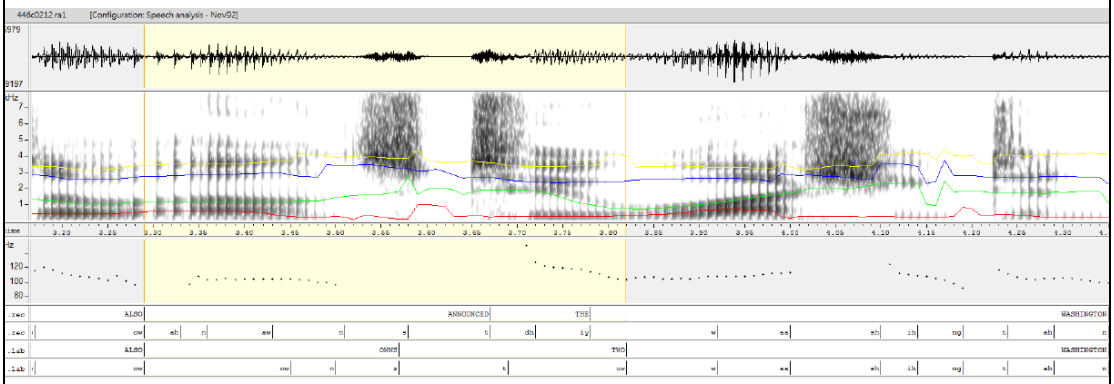
- The independent committee will recommend that holders accept the offer at a meeting expected to be held in December, T. W. A. said. (440c020j reference)
- The independent committee will recommend that holders accept the offer at a meeting expected to be held in December, T. to B. A. said. (440c020j answer)



B. is too short. If it was a spelling “B.”, the duration should be longer!

Fig. 4

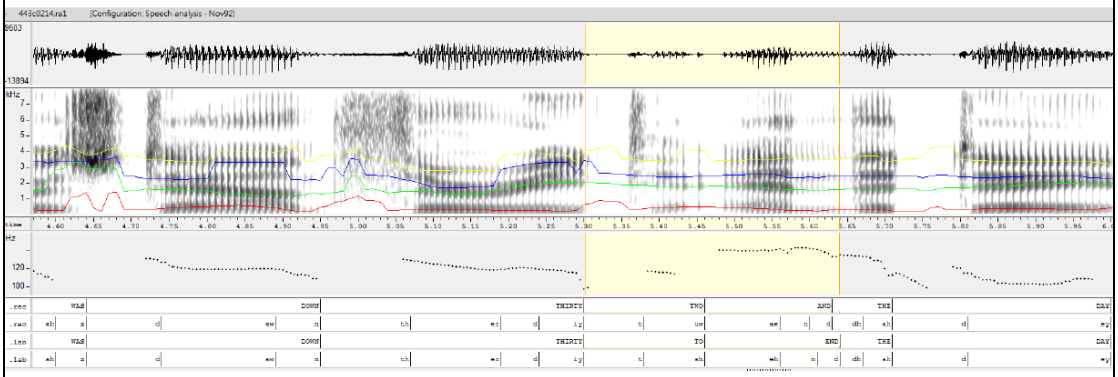
- Closely held Times Publishing also owns two Washington-based publications -- Congressional Quarterly, which covers Capitol Hill, and Governing, which covers state and local governments. (446c0212 reference)
- Closely held Times Publishing also announced the Washington-based publications -- Congressional Quarterly, which covers Capitol Hill, and Governing, which covers state and local governments. (446c0212 answer)



The article “the” should not have higher F0 than the content word “announced”

Fig. 5

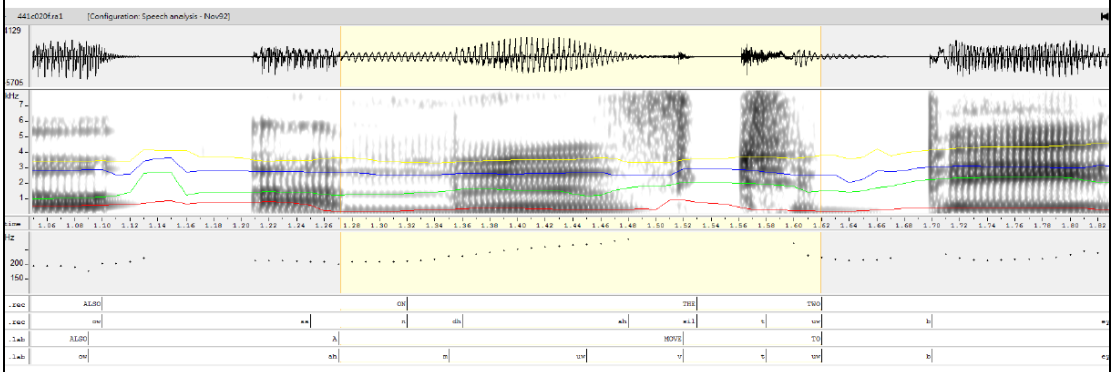
- Nissan lost thirty to one thousand, five hundred twenty and Toyota was down thirty to end the day at two thousand, six hundred twenty. (443c0214, reference)
- Nissan lost thirty to one thousand, five hundred twenty and Toyota was down thirty two and the day at two thousand, six hundred twenty. (443c0214, answer)



The F0 of “and” is too high to “two”.

Fig. 6

- Also, a move to base it abroad will have tax advantages. (441c020f, reference)
- Also, a on the two basic abroad will have tax advantages. (441c020f, answer)



The article “the” should not be too long (0.2 sec) and the F0 of “two” is lower than that of “the”.

Fig. 7

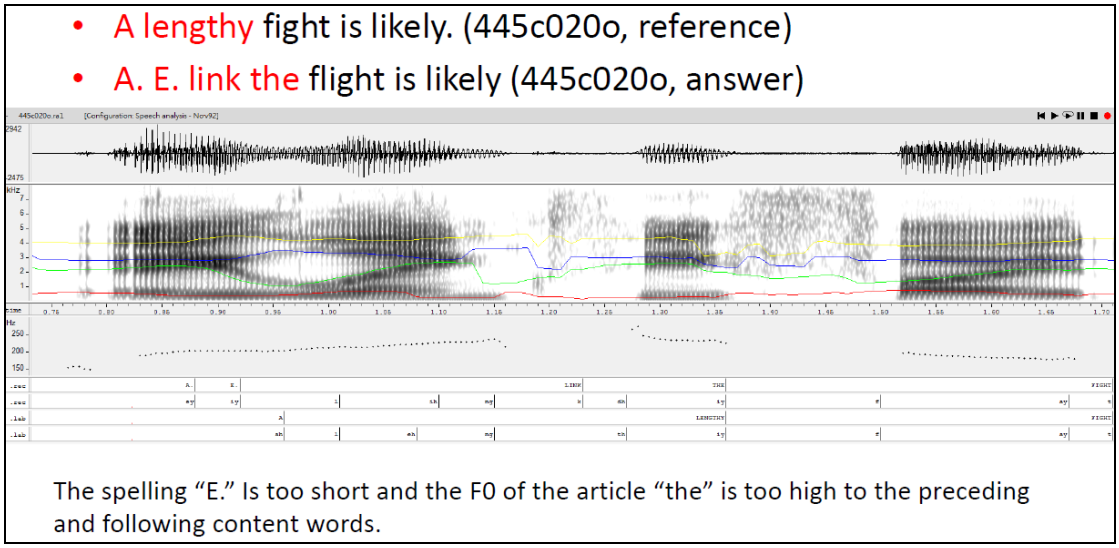


Fig. 8

(3) 2012 年 3 月 1 日~迄今

進行 prosodic modeling，對於 phoneme network 以 phone-level prosodic model 進行刪除及重新計分，預計於 3 月 19 日開始與 Prof. Sabato Marco Siniscalchi, Department of Telematics, Kore University of Enna, Enna, Italy 合作。

（三）心得及建議

此次十分感謝校方能支持這次爲其 5 個月的訪問研究，讓本人有機會能與李錦輝老師密切的研究討論。在討論的過程中，重新檢視過去累積的研究經驗與研究方法，受益良多。另外，在研究的過程中，發現到美國在語音處理的技術能夠累積，很多原因是因爲美國重視研究的建設基礎，如語音資料庫的蒐集、整理以及標準化，皆是由多所學校及美國政府 (如國防部) 合作而成，並非少數人力完成，所有的基礎系統皆有公開且標準化的模型，讓所有進行研究的學者有所依規，讓所有比較公平，因此學術界的經驗可以系統性的累積。反觀台灣於基礎工作上就比較不受重視，一開始可能可以快速的完成特定的研究，然而久而久之，學術上的累積便因爲缺乏穩固的基礎而進展變慢，這的確是令人值得反思的地方。