基於深度學習之中文文字轉台語語音合成系統初步探討

A Preliminary Study on Deep Learning-based Chinese Text to Taiwanese Speech Synthesis System

許文漢 Wen-Han Hsu, 曾證融 Cheng-Jung Tseng, 廖元甫 Yuan-Fu Liao
國立臺北科技大學電子工程系

Department of Electronic Engineering, National Taipei University of Technology

jeff3136169@gmail.com, t107368030@ntut.edu.tw, yfliao@ntut.edu.tw

王文俊 Wern-Jun Wang, 潘振銘 Chen-Ming Pan
中華電信實驗室
Chunghwa Telecom Laboratories
wernjun@cht.com.tw, chenming@cht.com.tw

摘要

台語在台灣歷史悠久，使用的族群眾多，有著很重要的存在價值。語音合成在追求跟人類一樣的聲音以及語調的同時，語言的多樣性也是一個需要深入探討的領域。本論文針對目前較少有的台語語音合成系統來作探討，利用翻譯模型 Chinese to Taiwanese(C2T)將輸入的中文文字轉成台羅拼音數字調(TLPA)，再將拼音輸入 Tacotron2 模型(Text to Spectrogram)後輸出頻譜，最後由 WaveGlow 模型(Spectrogram to Waveform)來實現語音合成。同時有架設網頁可供使用者一同來測試成效。

本文 C2T 機器翻譯的實驗方面採取三種模式，包括(1)輸入中文字詞，先進行斷詞，再輸出每個中文詞的台語台羅（Tâi-lô）拼音。(2)輸入中文字元串，直接輸出台羅拼音串。(3)輸入中文字元串，輸出台語的台羅拼音串與台語詞的斷詞關係。若不考慮聲調，方法(1)的 syllable error rate(SER)為 15.66%。而方法(2)的 SER 更可達 6.53%。這表示我們所用的 sequence-to-sequence 模型確實可以正確地將輸入的中文字元串，直接輸出台羅拼音串。

在台語語音合成品質實驗方面，我們找了 20 位聽者，各聽取 15 句不同內容的合成音檔後，以平均主觀意見進行評分(mean opinion score，MOS，完全不像人講話的聲音為 1 分，完全像真人講話聲音為 5 分)。總計收集到 300 個評分，最後得到我們系統的 MOS

得分為 4.30 分。這表示我們所用的 Tacrtron2 與 WaveGlow 模型確實可以正確將台羅拼音串轉成台語語音。此外此系統的語音合成速度為一秒可合成約 3.5 秒之音檔，的確可以達到即時語音合成的要求。

關鍵詞：機器翻譯、臺灣閩南語羅馬字拼音、台語語音合成

Abstract

This paper focuses on the development and implementation of a Chinese Text-to-Taiwanese speech synthesis system. The proposed system combines three deep neural network-based modules including (1) a sequence-to-sequence-based Chinese characters to Taiwan Minnanyu Luomazi Pinyin (shortened to as Tâi-lô) machine translation (called C2T from now on), (2) a Tacotron2-based Tâi-lô pinyin to spectrogram and (3) a WaveGlow-based spectrogram to speech waveform synthesis subsystems.

Among them, the C2T module was trained using a Chinese-Taiwanese parallel corpus (iCorpus) and 9 dictionaries released by Academia Sinica and collected from internet, respectively. The Tacotron2 and Waveglow was tuned using a Taiwanese speech synthesis corpus (a female speaker, about 10 hours speech) recorded by Chunghwa Telecom Laboratories. At the same time, a demonstration Chinese Text-to-Taiwanese speech synthesis web page has also been implemented.

From the experimental results, it was found that (1) the best syllable error rate (SER) of 6.53% was achieved by the C2T module, (2) and the average MOS score of the whole speech synthesis system evaluated by 20 listeners gains 4.30. These results confirm that the effectiveness of integration of C2T, Tacrtron2 and WaveGlow models. In addition, the real-time factor of the whole system achieved 1/3.5.

Keywords: Machine Translation, Taiwanese Speech Synthesis, Tacotron2, Waveglow