TTS survey

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

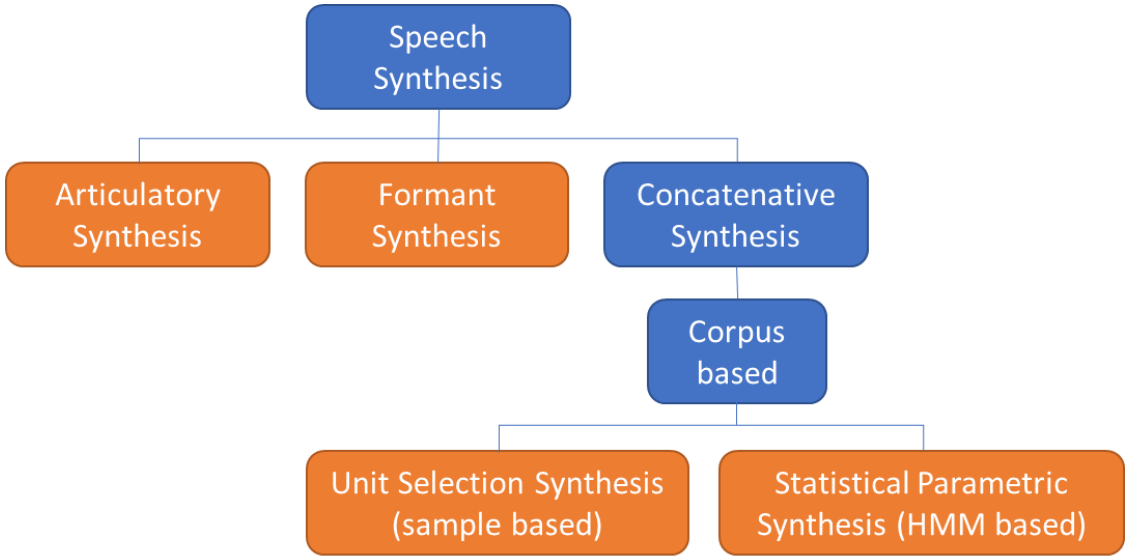

Fig. 1: Taxonomy of Speech Synthesis Methods [1]

Action item

- TTS survey: 上週找到傳統 TTS 共有四種方式，
  並將其技術或 building block 說明清楚

Status report

- 傳統 TTS 可以分為以下三種，其中拼接合成又分為兩種
  1. **發音合成 (Articulatory Synthesis)**：為模擬人類發聲器官的行為來產生語音
  2. **共振峰合成 (Formant Synthesis)**：根據 source-filter model 的規則生成語音
  3. 拼接合成 (Concatenative Synthesis)
     - **單元選擇合成 (Unit Selection Synthesis)**：從資料庫中選擇適合的單元進行拼接
     - **統計參數合成 (Statistical Parametric Synthesis)**：生成語音所需要的聲學參數，然後透過數學方法恢復語音，其中包含了文本分析、參數預測 (聲學模型)、聲碼器分析/合成 (聲碼器)三部分

[1] A. K. Rubeena and J. S. Chitode, “Concatenative speech synthesis: a review,” in *Proc. IJCA*, Vol. 136, no.3, Feb., 2016.

Articulatory Synthesis

介紹

- 基於人類聲道模型和該部位發音過程合成語音的計算技術，可以通過幾種方式控制聲道的形狀，在 VocalTractLab (VTL) 找到有開源的發音合成工具，其中為人類聲道的 3D 模型，模型由嘴唇、下頜、軟顎、舌體、舌尖...等部位組成，通常有 7~11 個參數來描述不同的發音動作，其中聲道形狀可以藉由拖曳模型中的控制點進行改變，也可以由參數做設定。
- 這種合成器的控制參數是：聲門下壓力、聲帶張力和不同發音 器官的相對位置，然後再現對應於聲道形狀的發音模型，並透過數學模擬計算通過聲道的氣流量。
- 該技術面臨的問題是獲得準確的三維聲道表示和使用有限的參數集對系統進行建模，因為對於複雜的人類發音器官缺乏了解[2]，因此沒辦法合成高品質語音。

VTL Tool

- 在 VTL 工具中可以點擊工具欄的 vocal tract shape 叫出元音、輔音設定列表 (Fig. 2左下角)，可以透過多樣函數定義聲道形狀。
- 發音器的形狀和/或位置由多個聲道參數定義，這些參數可以通過拖動聲道圖片中的黃色控制點進行交互更改。

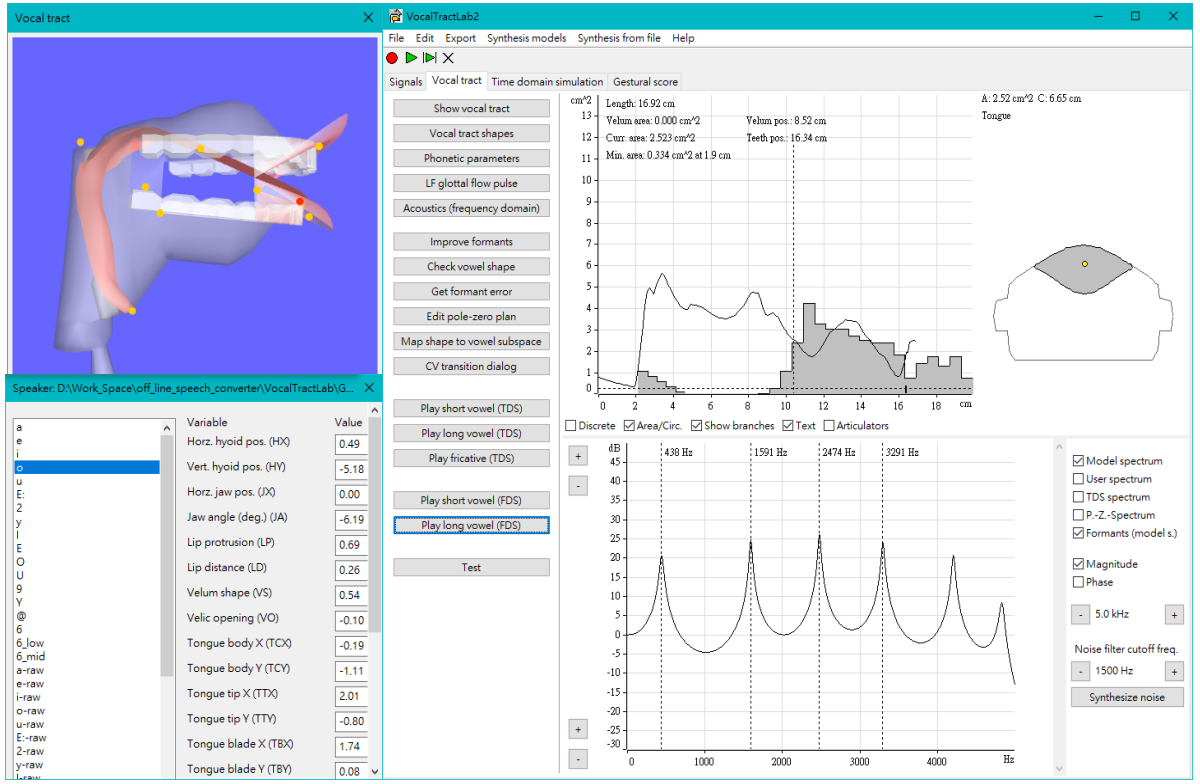

Fig. 2: VocalTractLab (VTL): 一個發音語音合成工具

Original

Synthesis

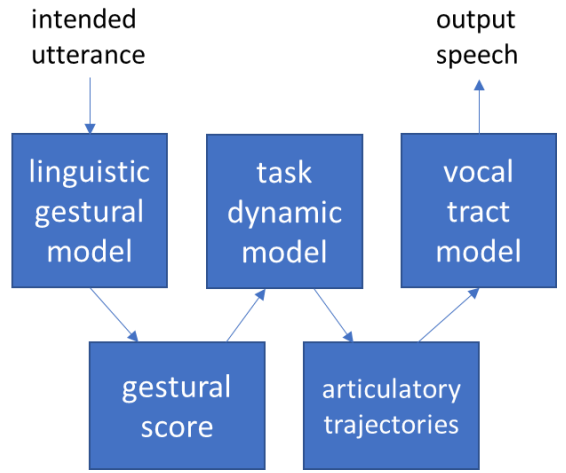

Fig. 3: 系統流程

[2] A. Balyan, S. S. Agrawal and A. Dev, “Speech synthesis: a review,” in *Proc. IJERT*, Vol. 2, June, 2013.

Formant Synthesis

- 共振峰合成是將聲道為一系列數字諧振器建模，是一種加法合成形式合成語音，想法是生成週期性、非週期性訊號或噪聲源，並將他們送到諧振器電路或濾波器生成。
- 在2019年Christian d'Heureuse 開發了 KlattSyn - Klatt Formant Synthesizer。
- Klatt 合成器可以給定多樣參數 (source 的特徵、頻率、頻寬、時間...等) 合成語音，但合成器只能合成單個音。
- 合成器可以使用串聯或是並聯方式連接諧振器來實現。
- 元音、濁音、半元音和送氣音採用連續音軌產生，擦音和塞音的爆發採用平行音軌產生，合成器由一組大約 60 個參數（輔音和變量）控制。

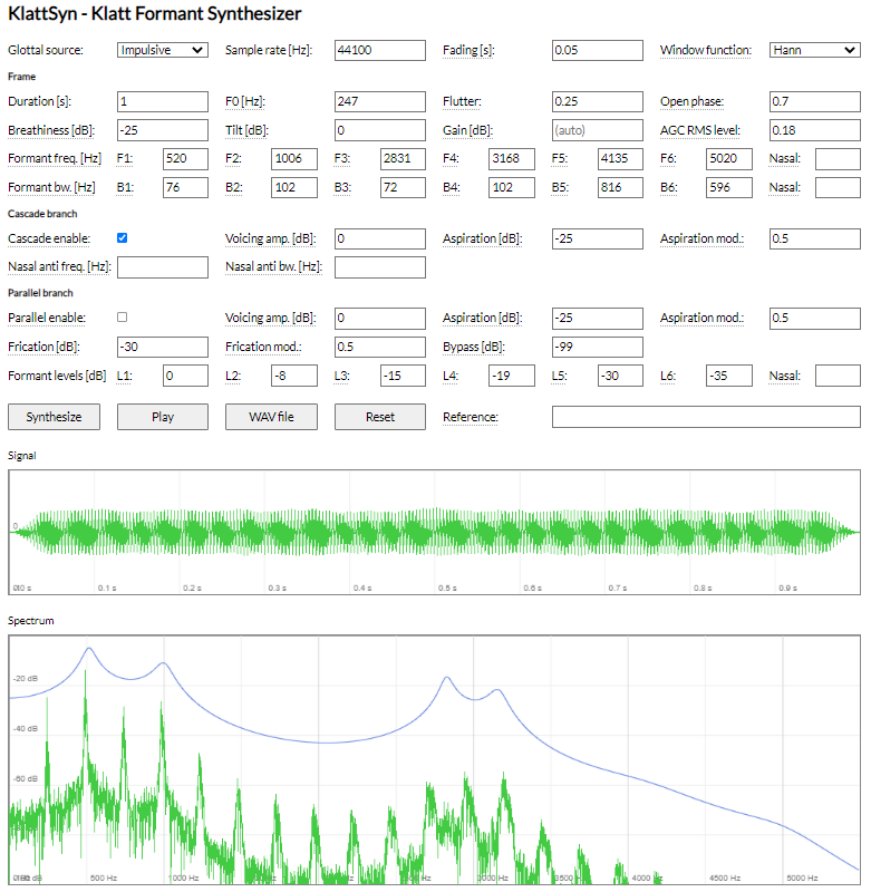

Fig. 4: Klatt 合成器

- **系統流程**
  - 元音：方波 -> 並行帶通濾波器 -> 混頻器 -> 放大/失真 -> 輸出
  - 輔音：噪聲 -> ADSR -> 輸出
    - Attack：觸發聲音時，Attack的量值可以決定多久聲音會到達最高峰。
    - Decay：聲音到達最高峰後會除漸削弱，但此階段並不會完全消逝。
    - Sustain：Decay的終點即是Sustain所設定的數值。
    - Release：聲音由Sustain衰減到聽不見的時間。

發音合成結果

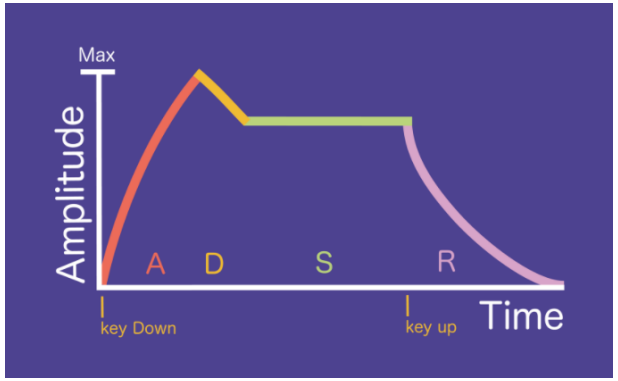

Fig. 5: ADSR

Unit Selection Synthesis

- 合成語音的單位可以是音素、雙音素、音節、單詞或句子，並儲存在資料庫中，持續的時間沒有限制，時間大概從10毫秒到1秒，使用大型數據庫合成，目標就是要合成自然的聲音。
- 主要合成品質的因素為連接點聲學特性的連續性，像是基頻、幅度、語速以及數據庫中具有適當韻律的語音單元的可用性，另外還有取得度量和權重較好的算法，也是影響音質的關鍵。
- 基本的單元選擇前提是我們可以通過從自然語音數據庫中選擇適當的子詞單元來合成新的自然發音的話語。
- 對於合成技術有兩種方式
    1. 選擇模型概念為目標成本，從資料庫中做選擇時候選單元與所選單元的匹配程度，以及連接的成本，定義了兩個選單元的結合程度。
    2. 使用聚類做法，預先計算目標成本，相同類型的單元被聚集到一個決策樹中，該決策樹詢問有關在合成時可用的特徵（例如，語音和韻律上下文）的問題。
- 找了一個 GitHub 專案，資料庫中使用 44 個英文音素做拼接，以下為合成結果。

合成結果：Hi, my name is Joe.

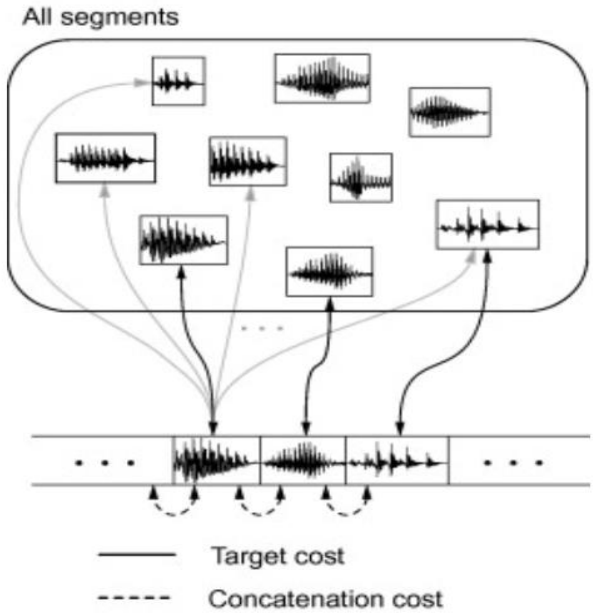

Fig. 6: 一般單元選擇方案 [3]

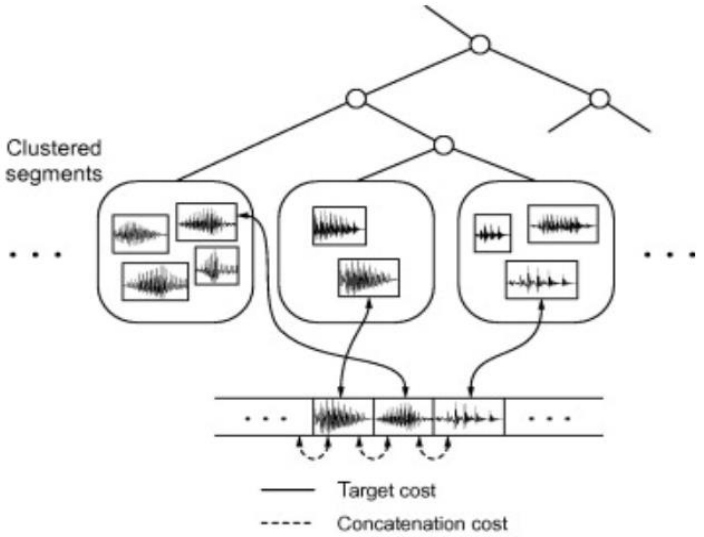

Fig. 7: 基於聚類的單元選擇方案 [3]

[3] H. Zen, K. Tokuda and A. W. Black, “Statistical parametric speech synthesis,” in *Proc. Speech Communication*, Vol. 51, 2009.

Statistical Parametric Synthesis

- 統計參數語音合成(SPSS) 主要比較對象為 unit Selection Synthesis。
- 生成語音所需要的聲學參數，然後透過數學方法恢復語音，其中包含了文本分析、參數預測 (聲學模型)、聲碼器分析/合成 (聲碼器)三部分。
- SPSS 系統可以看作是 ASR 的鏡像系統：ASR 系統嘗試使用機器學習模型將語音從聲學特徵轉換為一串單詞，而 SPSS 系統嘗試使用機器學習模型將一串單詞轉換為聲學特徵或直接轉換為聲波波形。
- ASR 和 SPSS 系統通常都使用大量語音數據及其轉錄進行訓練，從而產生一組描述語音數據統計特徵的參數，因此稱為“統計參數”語音合成。
- 首先從語音數據庫中提取語音的參數表示，包括頻譜和激勵參數(mfcc, lsf, f0..等)，然後使用一組生成模型（例如，HMM）對其進行建模。最大似然 (ML) 標准通常用於估計模型參數，最後從語音的參數表示中重建語音波形。

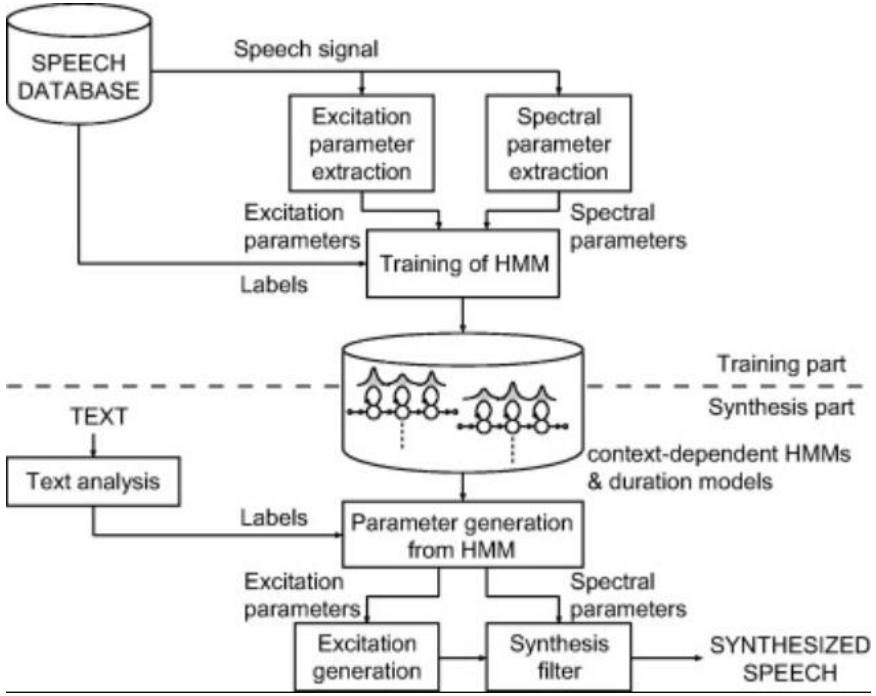

Fig. 8: Block-diagram of HMM-based speech synthesis system (HTS) [3]

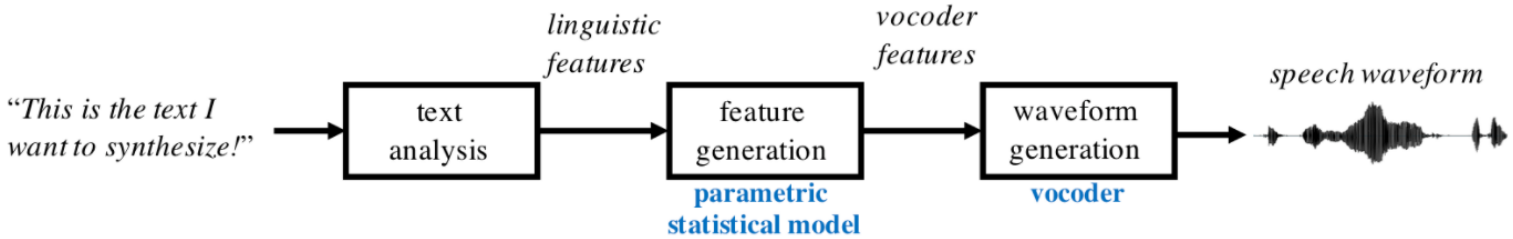

Fig. 9: A schematic view of an SPSS system

[3] H. Zen, K. Tokuda and A. W. Black, “Statistical parametric speech synthesis,” in *Proc. Speech Communication*, Vol. 51, 2009.