Conventional TTS: statistical parametric synthesis

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

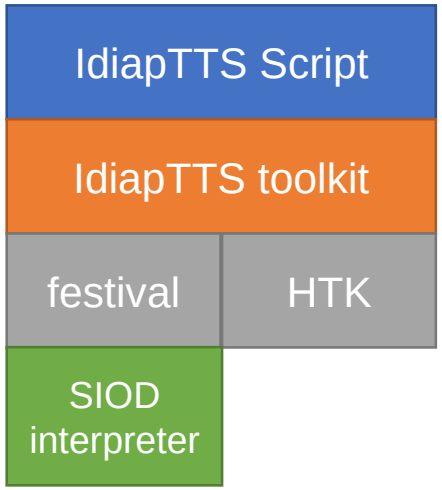
Outline

Action item

- Conventional TTS: statistical parametric synthesis implement

Status report

- **TTS 的 input/output**
  - TTS 系統總共分為 4 個階段，分別是文本分析、時間持續模型、聲學模型、聲碼器
  - 以下分別為訓練與合成階段 **(詳細 input/output 於次頁展示)**
    1. 訓練
       - (One-hot vector phoneme) - 持續時間模型 - (phoneme durations)
       - (Wav) - 聲學模型 - (聲學特徵)
    2. 合成
       - (Text) - 前端 (文本分析) - 持續時間模型 - 聲學模型 - 聲碼器 - (Wav)
- **TTS implement**
  - 目前使用 IdiapTTS 專案，在前端文本處理的部分推測為 festival 安裝失敗。
  - 先前在 Ubuntu20.04 使用 gcc4.8 編譯 festival 失敗而使用虛擬機降版至 Ubuntu18.04，使用 gcc4.8 編譯成功，但執行腳本時卻出現 SIOD error。
  - 解決方案
    1. 嘗試將 Ubuntu 降版至16.04，有看到文件說 gcc4.4 為編譯完最穩定的版本。
    2. 使用 Merlin toolkit，利用提供的範例跳過前端處理的部分，直接進行語音合成。



圖一：IdiapTTS 架構組成

HTK：The Hidden Markov Model Toolkit，用於建構或是操作隱式馬可夫模型的工具包。應用於語音識別、語音合成、字符辨識等。
Festival：Speech Synthesis System，基於 SIOD 命令解釋器進行控制。

The building block of statistical parametric synthesis

訓練：

one-hot vector phoneme → Duration Model → phoneme durations

Wav → Acoustic Model → MGC：由 MFCC 降維
LF0：對 f0 取 log
BAP： band aperiodicity
VUV：voiced or unvoiced

合成：

Text → frontend → 生成音素標籤 → Duration Model → 預測出每個音素 5 個 HMM 狀態的持續時間並加進標籤 → Question generate → 聲、韻母
韻律
詞性
.
.
等等
特徵劃分

→ Acoustic Model → MGC、LF0、BAP、VUV → WORLD → Wav

(附錄) Statistical Parametric Synthesis implement steps

執行步驟

- 安裝 concatenative speech synthesis TTS 工具（IdiapTTS）、環境以及各套件
- 以 IdiapTTS 工具為主進行實驗腳本，執行腳本前須下載預訓練模型與下載 LJSpeech 資料庫
- 使用 source cmd.sh 連結腳本與 IdiapTTS 工具（問題1）
- 實驗步驟可分為三大項

1. **特徵生成**
   - 創建資料庫
   - 創建強制對齊的 HTK 標籤
   - 從 HTK 標籤生成問題集（可決定神經網路維度）
   - 從對齊的 HTK 特徵中提取音素持續時間
   - 使用 World/PyWorld 從音檔中提取聲學特徵
2. **訓練**持續時間與聲學**模型**
   - 持續時間模型預測每個音素的 5 個 HMM 狀態
   - 聲學模型預測（MGC、LF0、VUV、BAP）聲學特徵
3. **合成**
   - 使用 WORLD 聲碼器生成音檔

HTK：The Hidden Markov Model Toolkit，用於建構或是操作隱式馬可夫模型的工具包。應用於語音識別、語音合成、字符辨識等。

(附錄) 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。

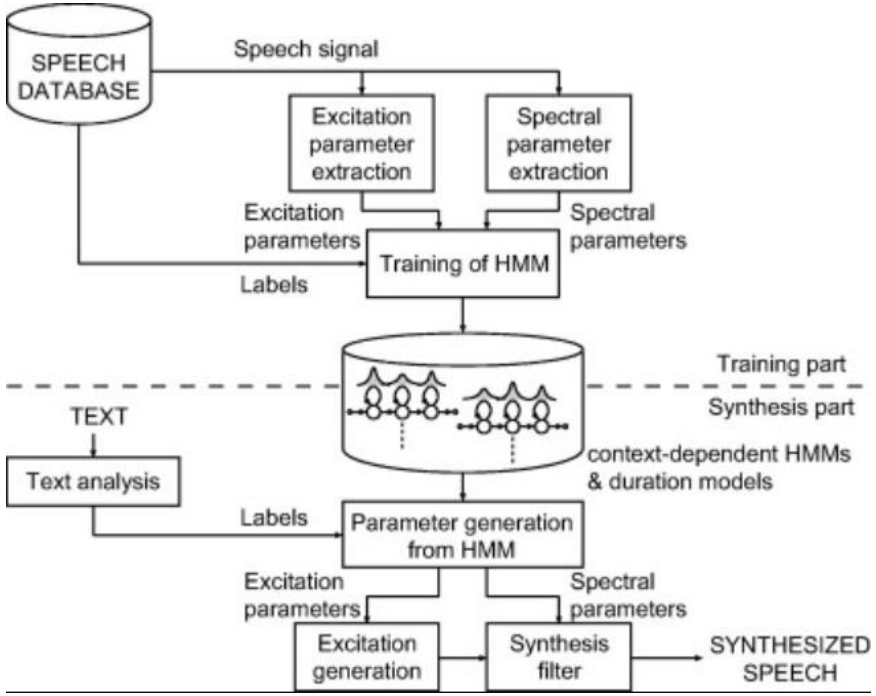


Fig. 8: Block-diagram of HMM-based speech synthesis system (HTS) [3]

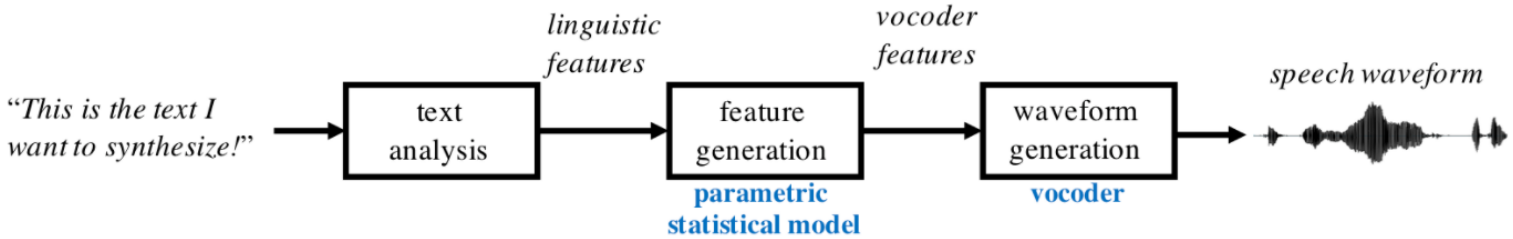


Fig. 9: A schematic view of an SPSS system

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