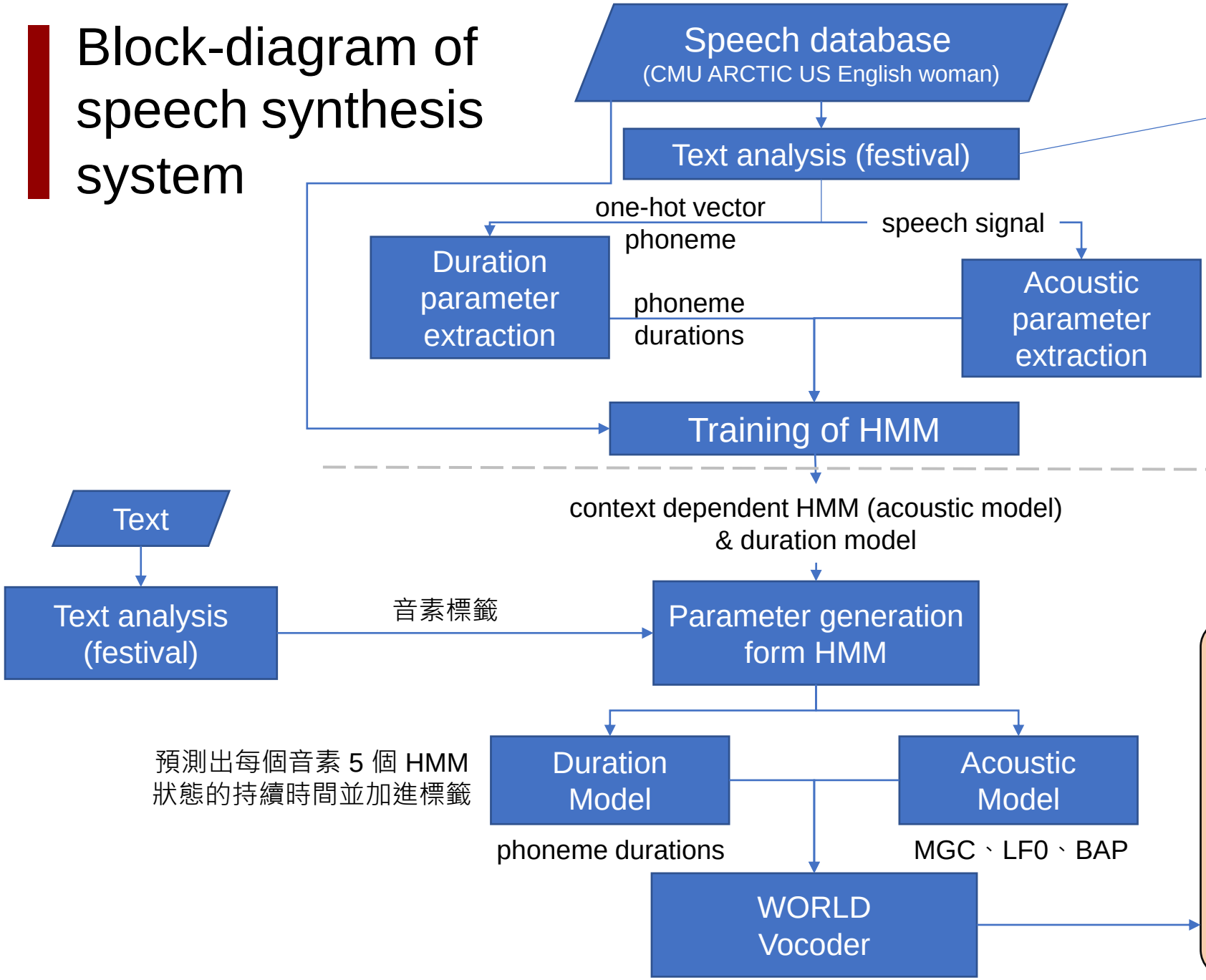
Traditional TTS design (block diagram with clear I/O & processing of each block)

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Block-diagram of speech synthesis system



1. 生成HTS(HMM-based speech synthesis system)格式標籤
2. 生成每個音素 5 個 HMM 狀態的標籤
3. 將標籤對齊 phoneme
4. 使用問題集生成問題 (ex：聲母韻母？詞性？重音？等等)，同時設定神經網路維度

- MGC (Mel-generalized Cepstrum),
- BAP (band aperiodicity),
- VUV (voiced or unvoiced),
- LF0 (log f0)

Output:

1. I came for information more out of curiosity than anything else.
2. Anyway, no one saw her like that.

Outline

Status report

- TTS (text to speech) 分類中，傳統技術為statistical parametric synthesis，該系統可以分為**文本分析**、**神經網路模型**參數預測 (聲學模型、持續時間模型)、**聲碼器**分析/合成 (聲碼器)三個階段，其中又分為訓練以及合成。

- 在傳統技術中無法直接使用音源檔進行訓練，需要先操作一連串的文本分析階段，這部分稱之為前端，可由 Festival (Speech Synthesis System) 達成，此階段目的為製作神經網路訓練用標籤，製作過程包含降頻、生成標籤、對齊、決定神經網路維度...等，標籤內容則包含了發音狀態、音素、為聲母韻母、詞性...等。

- **TTS詳細的input/output於次頁展示**，以下為訓練以及合成的流程
  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）- 神經網路(持續時間、聲學模型) - 聲碼器** -（Wav）

- 傳統技術與現今技術主要的分界點為開始使用End-to-End model，目的為簡化傳統複雜的過程，現今技術大大簡化了傳統文本分析階段所做的事情，以Transformer做說明，使用的架構為Acoustic Model + Vocoder，輸入為character或是phoneme，傳統技術中生成標籤、對齊...等事情都轉交給神經網路做處理。

(附錄) Details of blocks

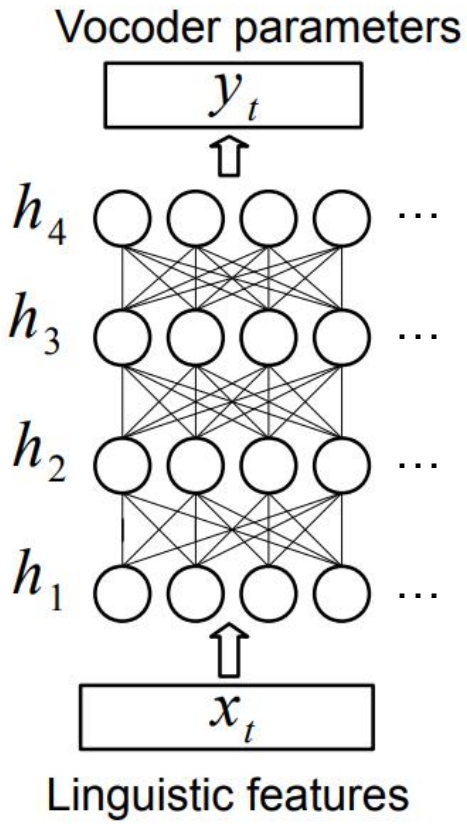
Forced alignment: 標準化(minmax、mean-variance)、對齊、提取特徵
- Label: 一個音素由幾個發聲狀態所組成 (設定為 5)

Question file: 416維

Duration model (DNN): 4*512
- input: 每個單元對應的長度、每個單元的向量
- output: 預測出每個音素 5 個狀態的持續時間

Acoustic model (DNN): 4*512
- output: mgc: 60; dmgc: 180; bap: 1; dbap: 3; lf0: 1; dlf0: 3

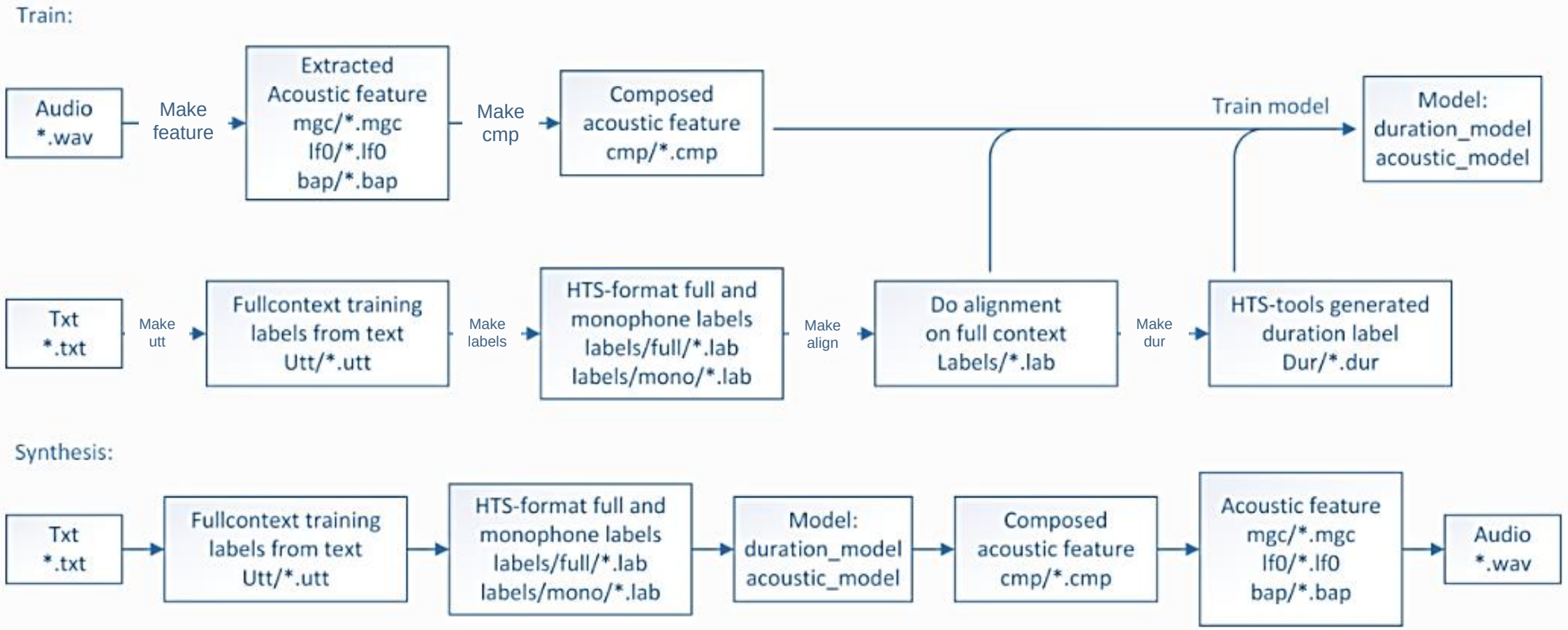
WORLD: 16kHz
- input
  - mgc: 60維
  - bap: 1維
  - lf0: 1維
- output: wav



圖一：前饋神經網路(DNN)

(附錄) Merlin 流程圖

整體流程圖和時長模型 & 聲學模型



(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。

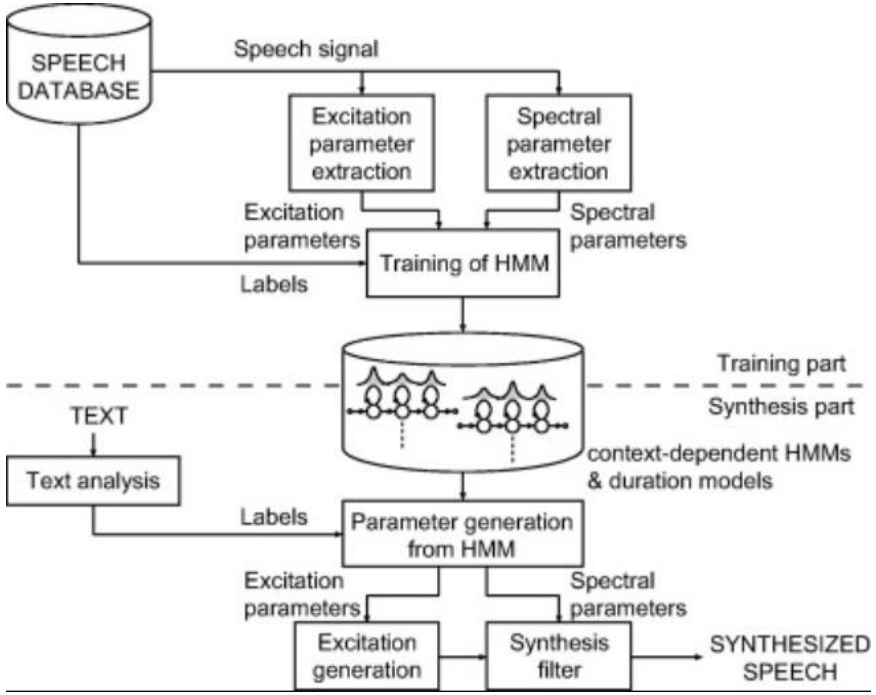


Fig. 8: Block-diagram of HMM-based speech synthesis system (HTS) [3]

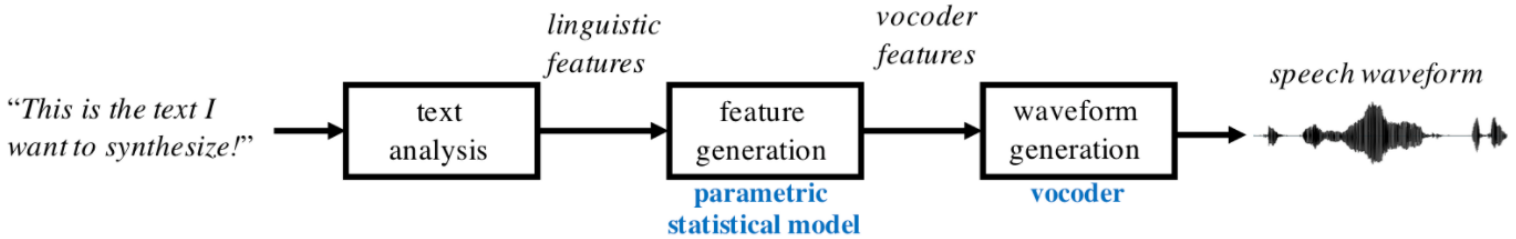


Fig. 9: A schematic view of an SPSS system

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