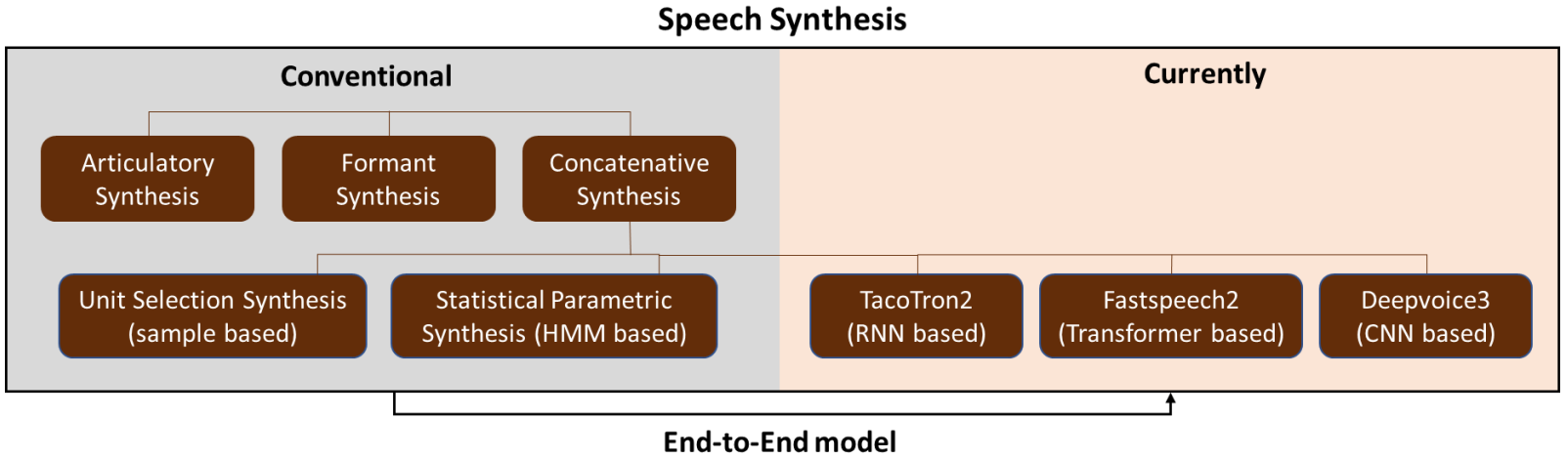
TTS survey

Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

Fig. 1: Taxonomy of Speech Synthesis Methods



Action item

- TTS survey: survey currently TTS & find a way to optimize

Status report

- 在 speech synthesis 分類中，傳統與現今技術我認為可以用 E2E model 作為分界點，而目前的聲學 (acoustic) 演算法主要分為三類，RNN based、Transformer based、CNN based。
- 目前的 TTS 追求的改進方案可以分類為幾類：追求生成速度、追求在低資源下生成可理解並自然的語音、追求模型生成的強健性 (robustness)、追求語音的表現性，具有韻律、情感等、追求 TTS model 的泛用性，而我的目標則是第四種，希望可以得到語音的音色、風格、韻律、語速等特徵。
- 現今 TTS model 可以拆解成文本分析、聲學模型、聲碼器三個部分，在 p.3 中顯示了E2E-TTS model 的演進，其中我所使用的架構為第三個(acoustic model + Vocoder)，因此我認為可以針對以下方式進行改進
  1. 找不僅是 acoustic model 且追求語音表現性的 TTS 更換目前所使用的 Transformer-TTS，因為在 [1] 提到 Transformer 為較初級的 TTS。
  2. 將 [1] 使用的 Parallel WaveGAN 換成 AR model 或是找更先進的 non-AR model，像是 HiFi-GAN。
  3. 找其他帶有風格的TTS遇到什麼問題，如何解決的，並應用到目前所使用的 Transformer 上，像是改進文本表示的詞嵌入或是文本訓練。

[1] W.-C. Huang, T. Hayashi, S. Watanabe, T. Toda, “The sequence-to-sequence baseline for the voice conversion challenge 2020: cascading ASR and TTS,” *arXiv preprint arXiv:2010.02434*, 2020.

E2E-TTS model

- 在 Stage 0 為傳統的 Statistical Parametric Synthesis(SPSS) 方法，透過文本分析將文字轉換成語言特徵，並透過聲學模型生成聲學特徵，最後透過聲碼器合成波型。

- 而後隨著神經網路的發展，不斷的簡化訓練模型，而我目前使用的 baseline 架構則屬於 stage 3 (acoustic model + Vocoder)，acoustic model 使用 Transformer，Vocoder 則使用 Parallel WaveGAN (PWG)，屬於non-AR 模型。

- 最後希望發展成完全端到端的 TTS 模型，E2E model 有著以下的優勢
  1. 降低開發與部屬成本，像是減少先備知識的需求，或是降低人工標記的成本。
  2. 越少模型串接，意味著越少的機會造成錯誤的傳遞。
  3. 傳統模型需要將語言與聲學特徵對齊，現今模型可以透過 attention 機制或是預測隱式來學習對齊。
  4. 語言特徵備簡化成字符或是音素。

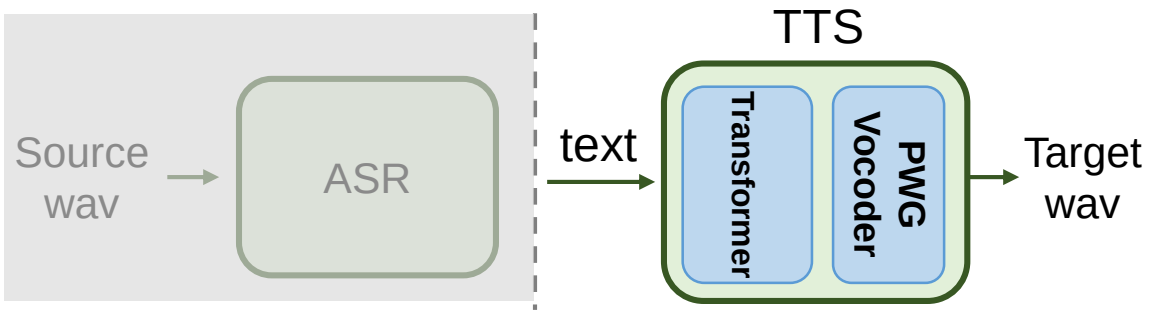
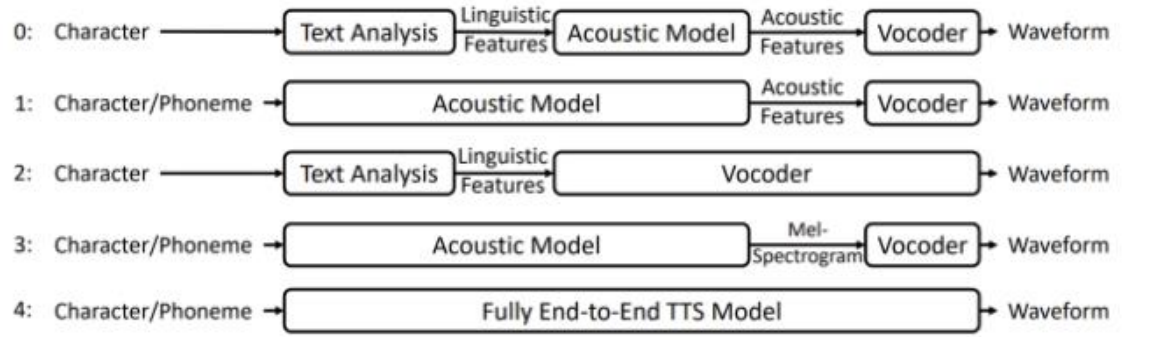


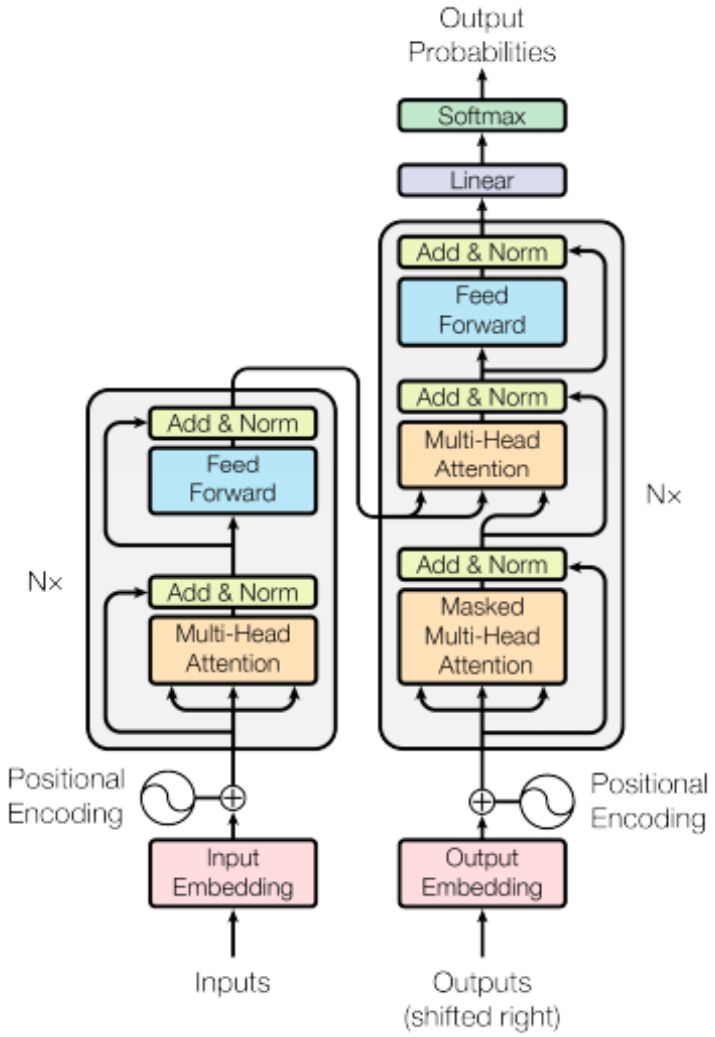
Fig. 2: Baseline ASR + TTS structure



Stage	Models
0	SPSS [416, 356, 415, 425, 357]
1	ARST [375]
2	WaveNet [254], DeepVoice 1/2 [8, 87], Par. WaveNet [255], WaveRNN [150], HiFi-GAN [23]
3	DeepVoice 3 [270], Tacotron 2 [303], FastSpeech 1/2 [290, 292], WaveGlow [279], FloWaveNet [163]
4	Char2Wav [315], ClariNet [269], FastSpeech 2s [292], EATS [69], Wave-Tacotron [385], VITS [160]

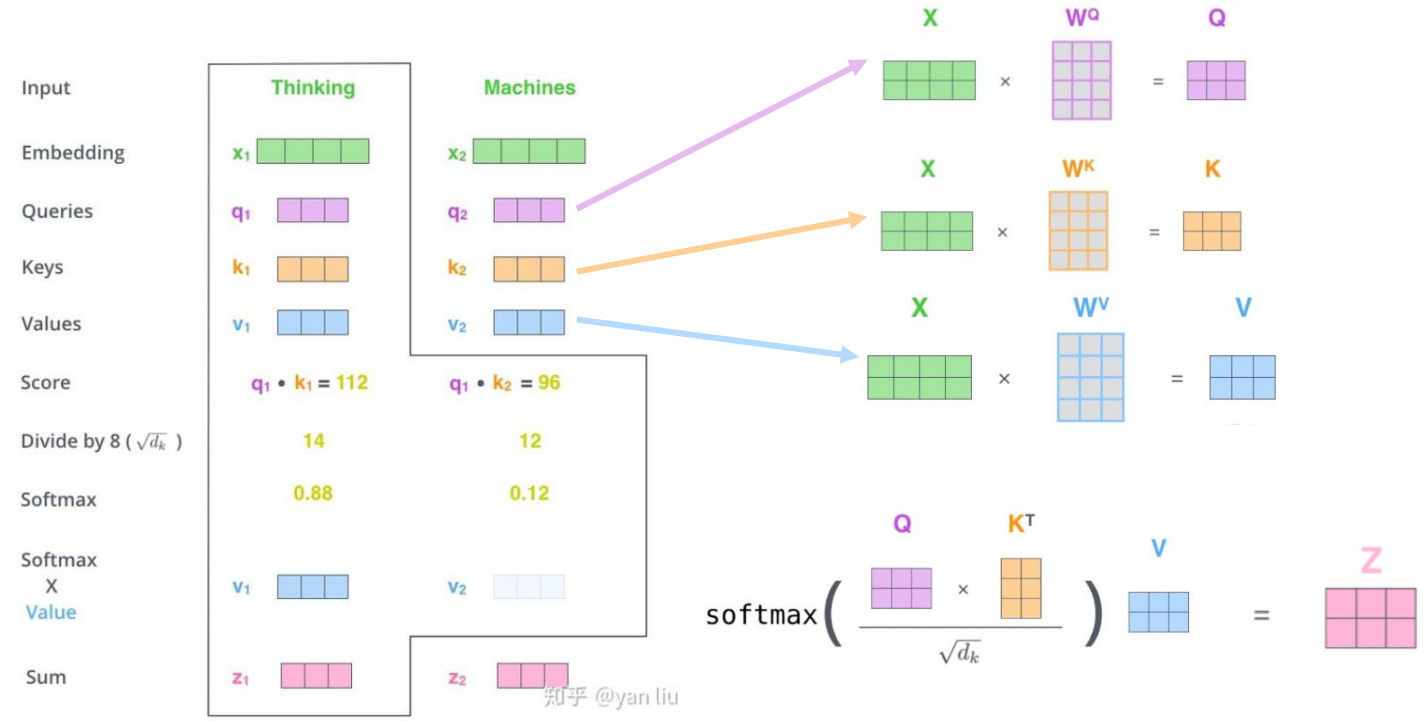
Fig. 3: The progressively end-to-end process for TTS models

Transformer



Transformer定義為：第一個完全依賴 self-attention 去計算 input、output 之間關係並不使用序列對齊的 RNN 或 CNN 的 transduction model。

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



Query，Key，Value的概念取自於資訊檢索系統，舉個簡單的搜索的例子來說。
當你在網購平臺搜索某件商品（年輕女士冬季穿的紅色薄款羽絨服）時
你在**搜尋引擎**上輸入的內容便是 **Query**
然後搜尋引擎根據Query為你匹配 **Key**（例如商品的種類，顏色，描述等）
然後根據Query和Key的相似度得到匹配的內容（**Value**）。