Speech Synthesis: Past, Present and Future

Yunlin Chen
2019.09.27

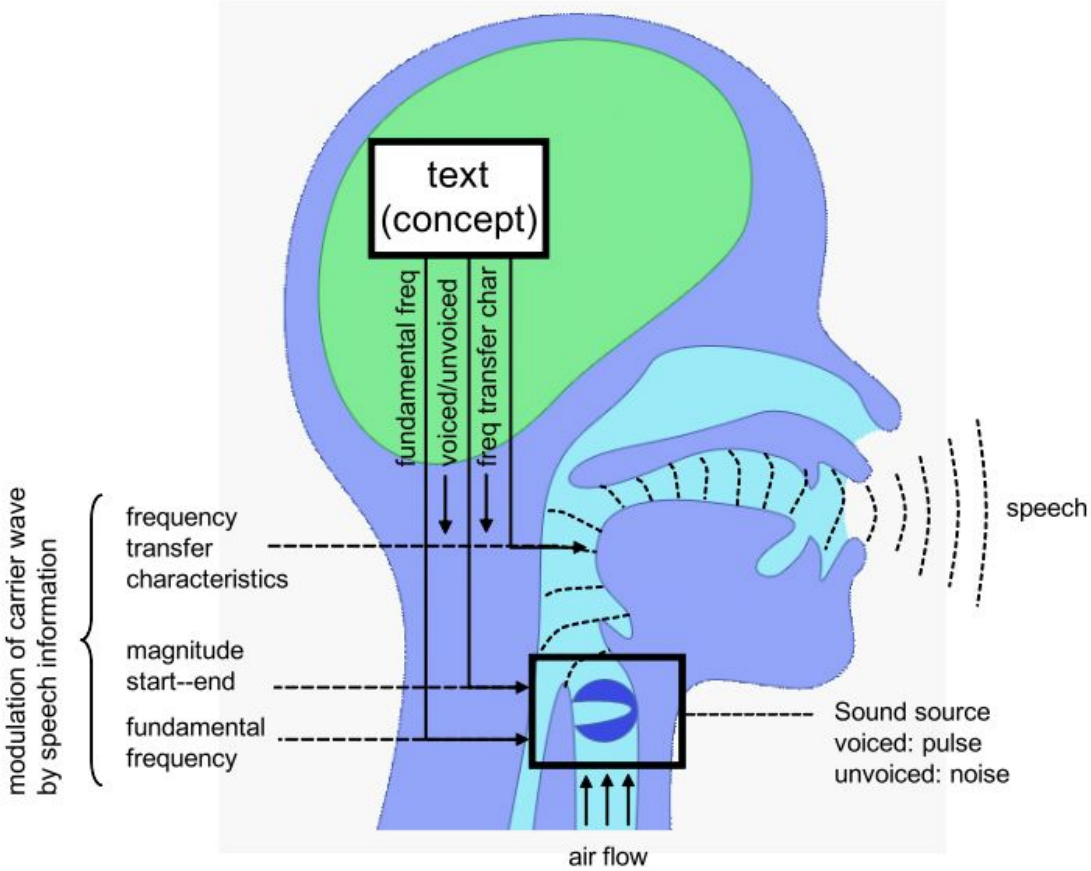
Reference

Statistical Approach to Speech Synthesis: Past, Present and Future

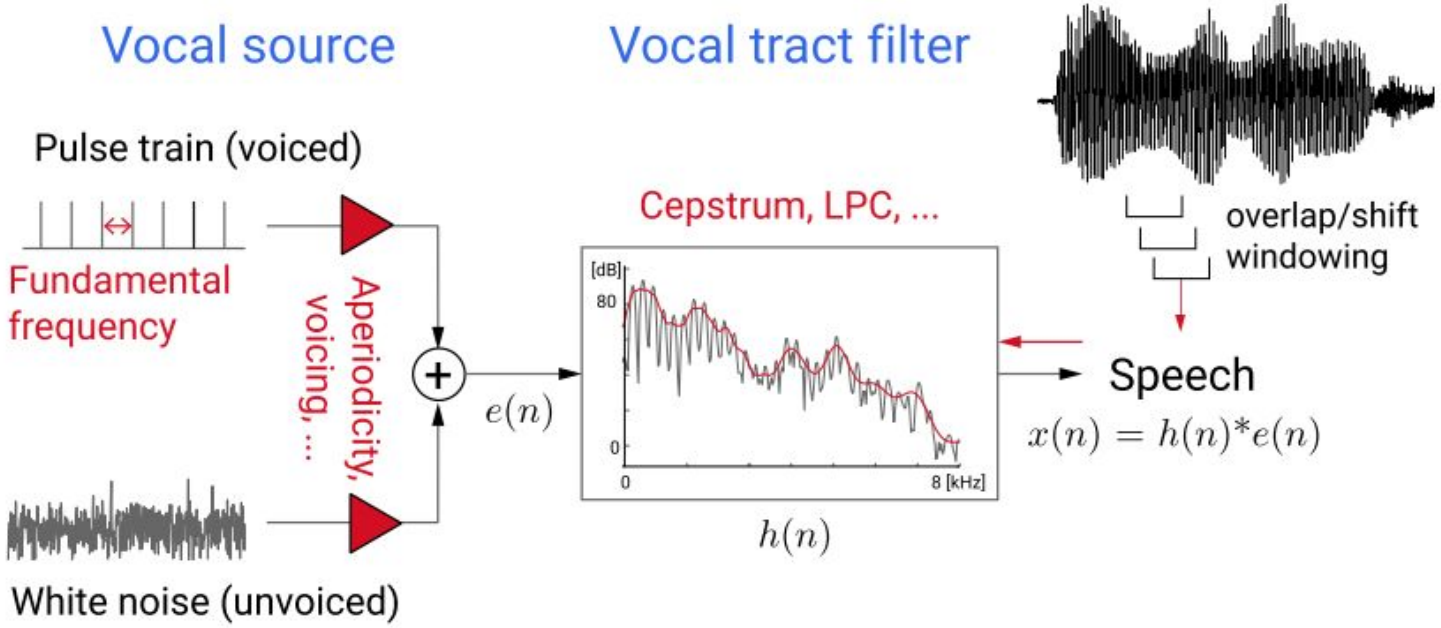
Keiichi Tokuda

The basic problem of statistical speech synthesis is quite simple: we have a speech database for training, i.e., a set of speech waveforms and corresponding texts; given a text not included in the training data, what is the speech waveform corresponding to the text? The whole text-to-speech generation process is decomposed into feasible subproblems: usually, text analysis, acoustic modeling, and waveform generation, combined as a statistical generative model. Each submodule can be modeled by a statistical machine learning technique: first, hidden Markov models were applied to acoustic modeling module and then various types of deep neural networks (DNN) have been applied to not only acoustic modeling module but also other modules. I will give an overview of such statistical approaches to speech synthesis, looking back on the evolution in the last couple of decades. Recent DNN-based approaches drastically improved the speech quality, causing a paradigm shift from concatenative speech synthesis approach to generative model-based statistical approach. However, for realizing human-like talking machines, the goal is not only to generate natural-sounding speech but also to flexibly control variations in speech, such as speaker identities, speaking styles, emotional expressions, etc. This talk will also discuss such future challenges and the direction in speech synthesis Research.

Speech Production Process



Speech Production Process



- **F0基频**对应激励部分的周期脉冲序列，如果我 们将声学信号分 为周期性信号与非周期信号的 话；
- **SP频谱**包络对应声道谐振部分时不变系统的冲激响 应
- **AP非周期**序列对应混合激励部分的非周期脉冲序列

TTS(Text-to-Speech)

- 语音合成技术:将文字转化为语音输出，即让计算机说话
- 日本学者fujisaki按照人在说话过程中的各种知识，将语音分为三个阶段
  a) 按照规则从文字到语音的合成(Text-To-Speech)
  b) 按照规则从概念到语音的合成(Concept-To-Speech)
  c) 按照规则从意向到语音的成(Intention-To-Speech)

意向 >> 语义表示 >> 概念 >> 语言编码 >> 文本

怎么判断机器合成的好坏

ABTest：

- 仔细听两段音频，从音质，流畅度，情感，韵律等 综合方面进行判别，选择相对较好的一个。

MOS：

- 一般是1-5分，大家依照各个分数区 间标准，进行打分，最后求平均（真实录音4.5 - 4.7）

业界的MOS：

HMM	NN	Unit Selection	Tacotron	Taco2+ Wavenet	**MeetVoice(ours)**
3.4	3.6	4.09	3.82	4.36	**4.1 - 4.2**

语音合成发展历史

- 1939，贝尔实验室第一个电子合成器VODER(共振峰)
- 1960年，瑞典科学家GFant提出了LPC
- 20世纪80年代，PSOLA技术的发展，推动拼接合成
- 90年代以来，可训练的语音合成(HTS)
- 2013，heiga提出了基于DNN的参数语音合成
- 近年来，端到端语音合成以及neural vocoder

Speech Synthesis - Probabilistic formulation

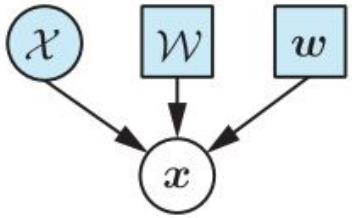
Probabilistic formulation of TTS

Random variables

$\mathcal{X}$	Speech waveforms (data)	Observed
$\mathcal{W}$	Transcriptions (data)	Observed
w	Given text	Observed
x	Synthesized speech	Unobserved

Synthesis

- Estimate posterior predictive distribution
  $\rightarrow p(x \mid w, \mathcal{X}, \mathcal{W})$
- Sample $\bar{x}$ from the posterior distribution



Probabilistic formulation of TTS

$$\hat{\mathcal{O}} = \arg\max_{\mathcal{O}} p(\mathcal{X} \mid \mathcal{O})$$ Extract *acoustic features*

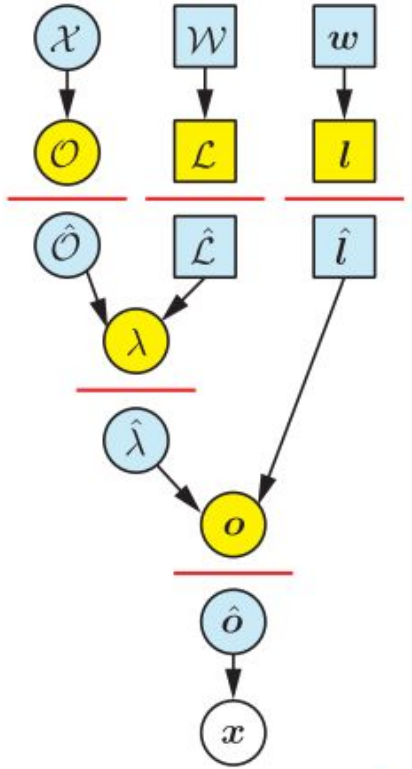
$$\hat{\mathcal{L}} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid \mathcal{W})$$ Extract *linguistic features*

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\mathcal{O}} \mid \hat{\mathcal{L}}, \lambda) p(\lambda)$$ Learn *mapping*

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\boldsymbol{l} \mid \boldsymbol{w})$$ Predict *linguistic features*

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ Predict *acoustic features*

$$\bar{\boldsymbol{x}} \sim f_{\boldsymbol{x}}(\hat{\boldsymbol{o}}) = p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$$ Synthesize waveform



Probabilistic formulation of TTS

$$\hat{\mathcal{O}} = \arg\max_{\mathcal{O}} p(\mathcal{X} \mid \mathcal{O})$$ Vocoder analysis

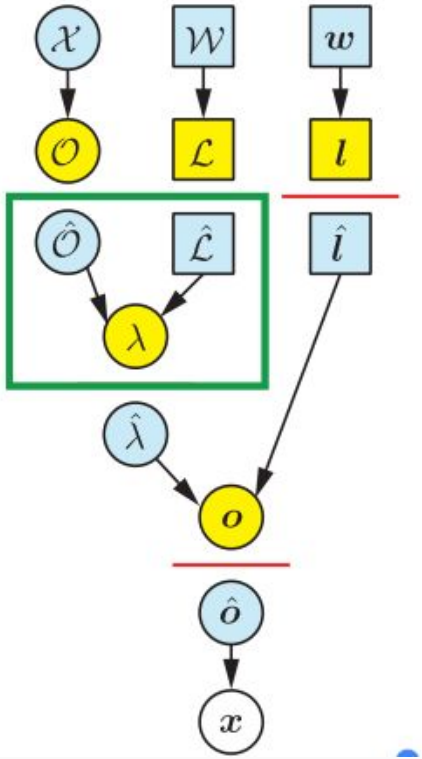
$$\hat{\mathcal{L}} = \arg\max_{\mathcal{L}} p(\mathcal{L} \mid \mathcal{W})$$ Text analysis

$$\hat{\lambda} = \arg\max_{\lambda} p(\hat{\mathcal{O}} \mid \hat{\mathcal{L}}, \lambda) p(\lambda)$$ ***Train HMMs***

$$\hat{\boldsymbol{l}} = \arg\max_{\boldsymbol{l}} p(\boldsymbol{l} \mid \boldsymbol{w})$$ Text analysis

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} p(\boldsymbol{o} \mid \hat{\boldsymbol{l}}, \hat{\lambda})$$ ***Parameter generation***

$$\bar{\boldsymbol{x}} \sim f_{\boldsymbol{x}}(\hat{\boldsymbol{o}}) = p(\boldsymbol{x} \mid \hat{\boldsymbol{o}})$$ Vocoder synthesis

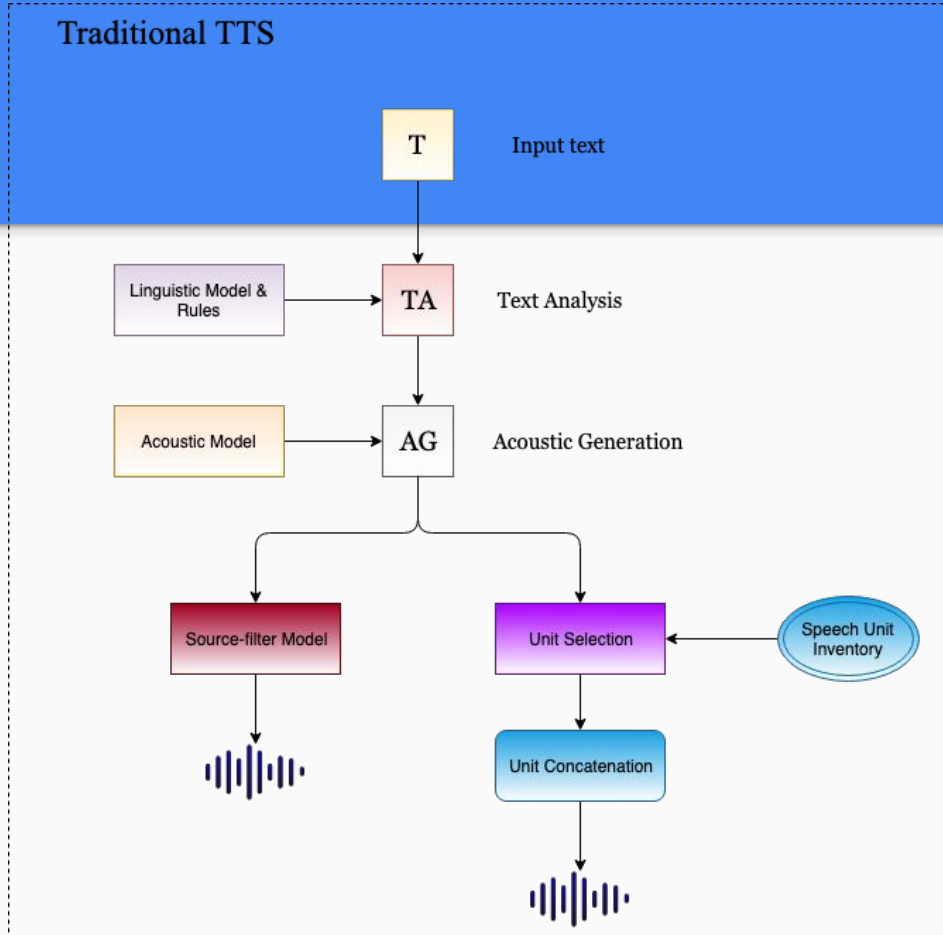


Speech Synthesis - Past

语音合成系统框架

- 文本处理以及分析(TA)- 前端 (front-end)
  - 文本: **90**后为中华人民共和国成立**70**周年准备了大礼
  - TN: **九零**后为中华人民共和国成立**七十**周年准备了大礼
  - 分词, 注音
  - 韵律预测
    九零后#1为中华人民#1共和国#2成立**七十**周年#3准备了大礼#4

- 声学模型参数以及语音生成 - 后端 (back-end)
  - 声学模型预测声学参数
  - 根据声学参数预测声音
    - Vocoder: 声学参数 -> speech
    - 拼接系统: 从现有的speech inventory找出合适的单元进行拼接



韵律预测

- 目的，给定待预测的文本及其分词、词性等信息，预测文本的韵律停顿等级：
  - 韵律词：一般为三个音节以下的语法词或词组，内部不出现节奏边界
  - 韵律短语：由一个或几个韵律词组成，具有相对稳定的短语语调模式和短语重音配置模式
  - 语调短语：在语法上相当于较短句子或较长的短语，韵律短语之间有音高重置
- 一个较小的韵律成分包含在一个更大韵律成分中
- 平均长度满足：韵律词 < 韵律短语 < 语调短语

传统算法

- 条件随机场(CRF)，基于无向图的概率图模型，具有表达长距离依赖性和交叠性特征的能力，能较好地解决标注偏置的问题，所有特征全局归一化，能够求得全局最优解

- 特征:词面，词性，词长(单个CRF)
  3个CRF：韵律词 、韵律短语、语调短语分开预测(上一级标注)

CRF的问题

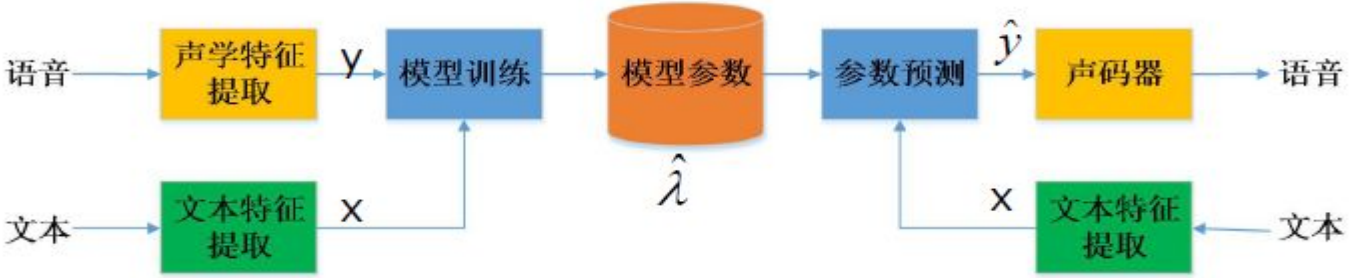
- 预测特征
  - 特征的选择和优化是决定CRF性能的关键因素，前后词窗长大小以及特征组合需要实验验证。
  - 基于词面特征，泛化能力一般，在标注数据不充分的情况模型有偏。
  - 受CRF建模能力和训练数据量的限制，一般使用前后3个词的信息，很难描述更远距离的词对当前词停顿的影响。
- 分别训练三个模型预测韵律词，韵律短语和语调短语，既有可能造成预测误差累计，又无法充分利用三种停顿之间的相关关系。当然训练一个模型，更不能考虑这些信息。

基于深度学习的韵律预测

- 使用多层双向LSTM-RNN神经网络代替CRF
  - 内嵌长距离轨迹建模的能力，可自动学习前后词信息对当前词停顿的影响，建模距离可以横跨整个句子
  - 一次性预测三种韵律停顿，充分考虑三种停顿之间的相互关系
  - 输入特征中加入词向量特征
    - 词向量是词在低维空间的分布式表示，其所具有的词间相似度信息可以大幅度提升韵律预测模型的泛化能力

Speech Synthesis - Past HMM

声学模型



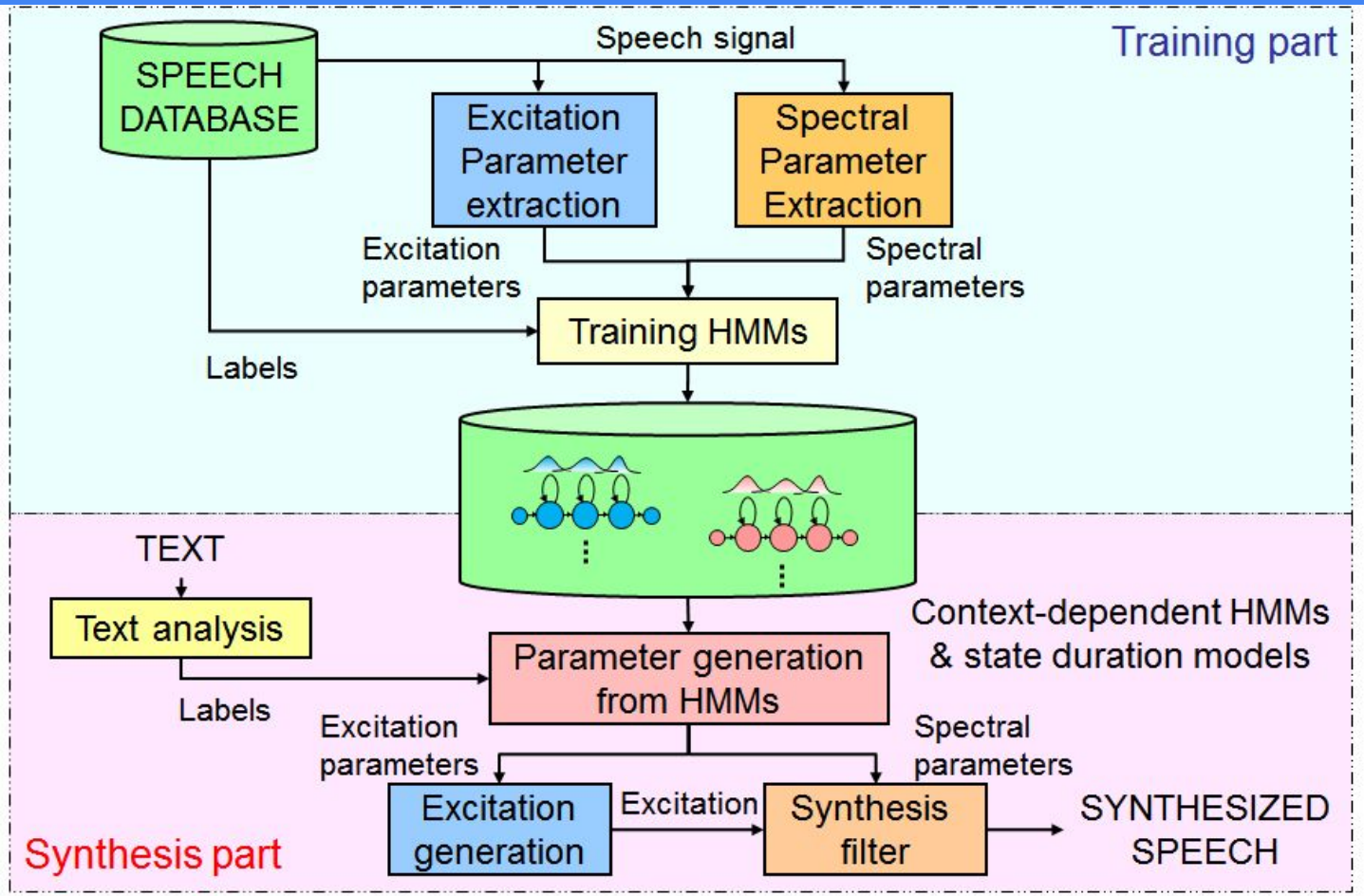
- 训练
  - 分别提取文本特征**x**和声学特征**y**；
  - 训练声学模型：

$$\hat{\lambda} = \arg\max p(y \mid x, \lambda)$$

- 预测
  - 从待合成文本提取文本特征**x**
  - 生成声学特征：

$$\hat{y} = \arg\max p(y \mid x, \hat{\lambda})$$

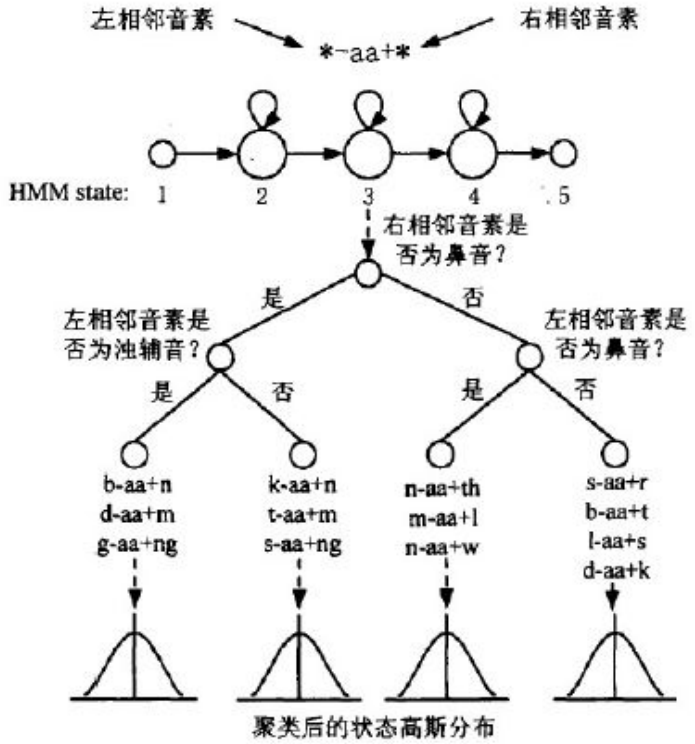
Acoustic - HMM



HMM - 特征

- 输入：文本(语境）特征
  - 音子名称，包括前后两个窗口quin-phone
  - 声调，当前音素的声调(也可以加入前后两个音节声调)
  - 层级相关，高层级单元中低层级单元的位置和数目
    语句>语调短语>韵律短语>韵律词>音节>音子>状态
  - 其他信息，例如当前音节的停顿、词性以及词长等信息
- 输出：声学特征
  - 谱特征， 41维的LSP特征，包括静态和一阶、二阶差分特征
  - 基频特征，log域基频值以及一阶、二阶差分特征
  - 时长，状态的帧数

HMM建模:基于决策树的模型聚类



- 作用
  - 建立文本特征到声学特征的映射关系。
- 目的
  - 训练中的数据稀疏问题；
  - 合成中的**unseen**语境；
- 缺点
  - 浅层建模，特征空间的线性划分；
  - 需要设计问题集；

HMM建模:HTS参数生成算法

- 决策树将文本特征x映射到了模型参数λ:
- 给定HMM模型参数λ，生成的参数序列o为:

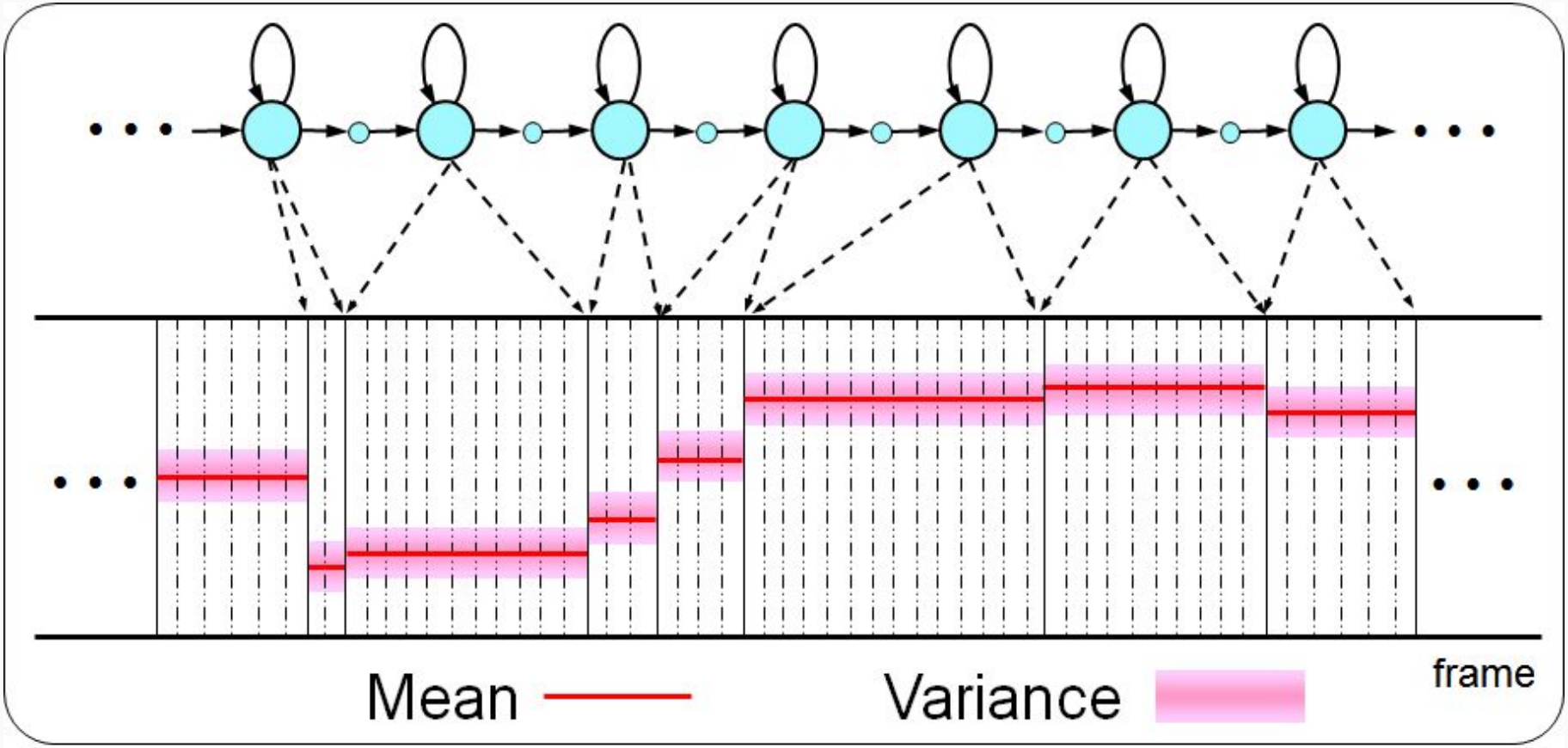
$$P(o|\lambda)=\sum_{q}P(o|q,\lambda)P(q|\lambda)\approx\max_{q}P(o|q,\lambda)P(q|\lambda)$$

$$\begin{cases}\hat{q}=\underset{q}{\operatorname{argmax}}\,P(q|w,\lambda)\\ \hat{o}=\underset{o}{\operatorname{argmax}}\,P(o|\hat{q},\lambda)\end{cases}$$

q：状态序列，由时长模型预测决定，

o：预测的声学参数序列

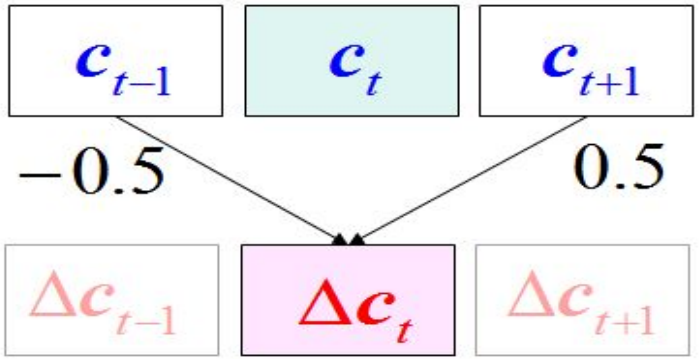
HMM建模:HTS参数生成算法

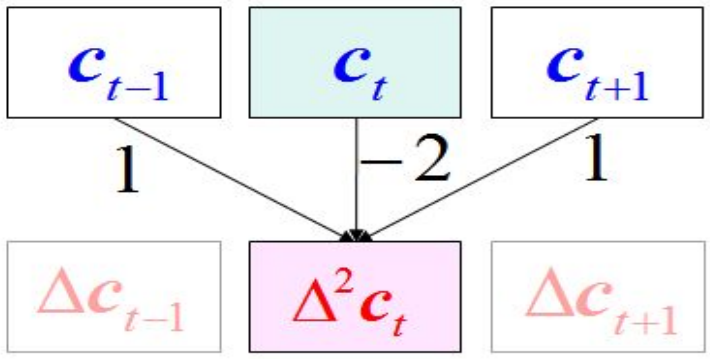


传统建模:HTS动态参数

$$\Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1})$$

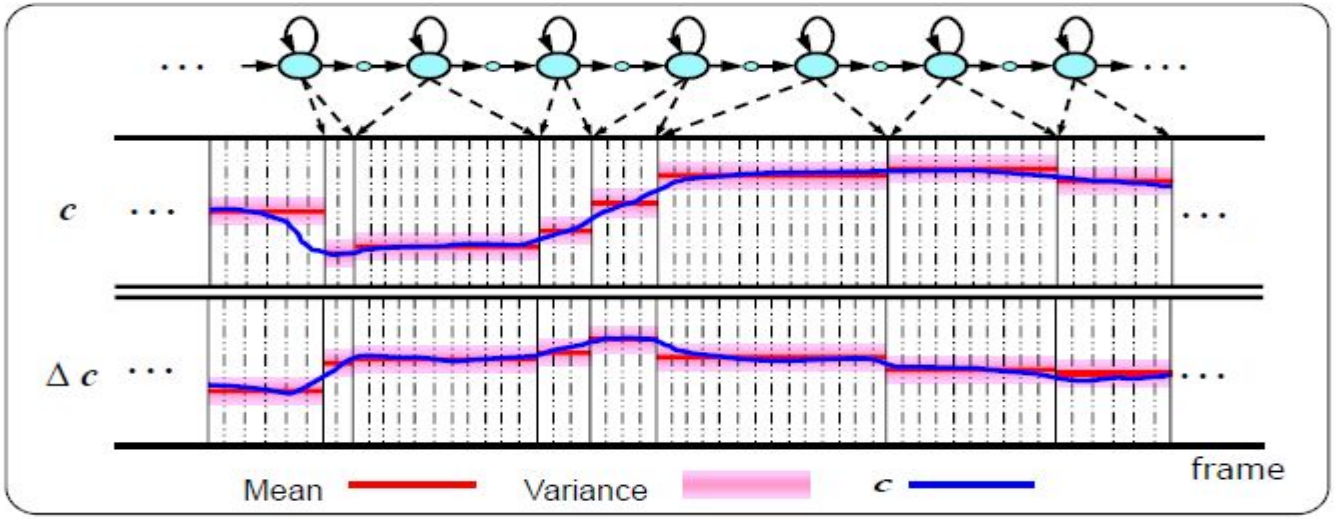
$$\Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1}$$





传统建模:HTS动态参数

$$\frac{\partial \log P(\mathbf{Wc} \mid \hat{\mathbf{q}}, \lambda)}{\partial \mathbf{c}} = \mathbf{0} \Longrightarrow \boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{q}}^{-1} \boldsymbol{W} \boldsymbol{c} = \boldsymbol{W}^{\mathrm{T}} \boldsymbol{\Sigma}_{\hat{q}}^{-1} \boldsymbol{\mu}_{\hat{q}} \ ,$$



Speech Synthesis - Past DNN

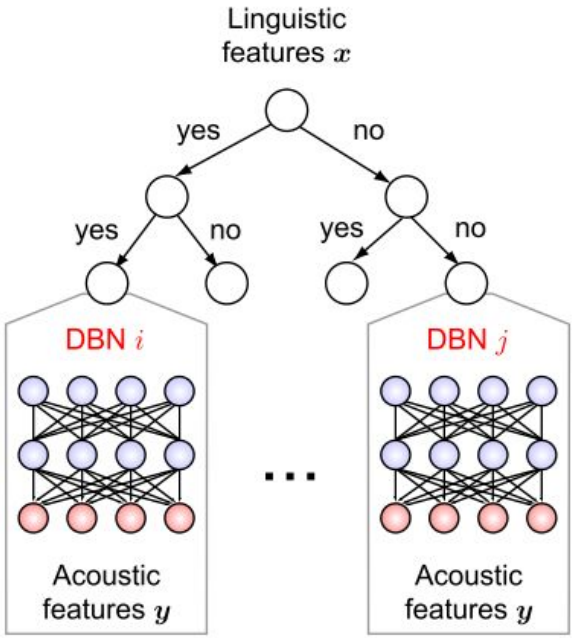
传统建模:HTS存在的问题

- HTS合成的过平滑问题导致合成效果平淡，音质不好
  - HMM基于建模，假设各状态之间独立，状态参数分布式统计平均的结果；
  - 决策树作为浅层线性模型无法建模复杂的映射关系，对输入特征和数据空间的线性分割会降低模型泛化能力；
  - 特征的短时相关特性
- 使用深度神经网络(LSTM-RNN)替代决策树直接对文本特征和声学特征映射关系进行建模
  - 基于帧建模替代HMM状态建模，可直接输出到声码器
  - 多层神经网络替代浅层的决策树

Deep learning-based approaches

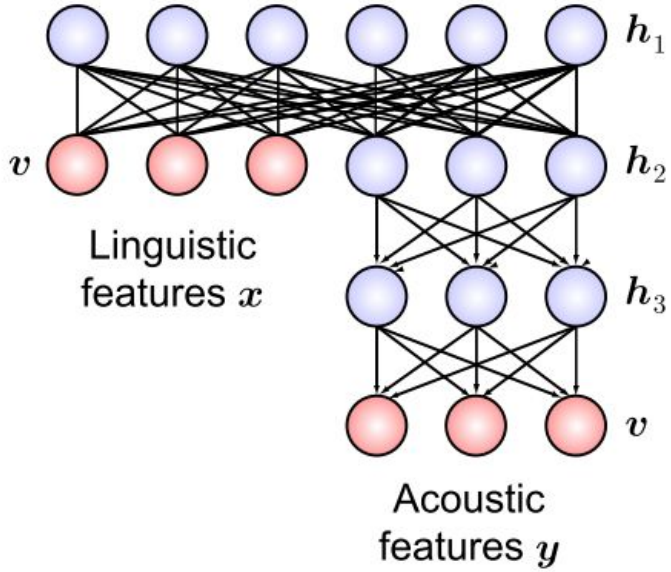
- Recent applications of deep learning to speech synthesis
  - HMM-DBN (USTC/MSR)
  - DBN (CUHK)
  - DNN (Google)
  - DNN-GP (IBM)
  - BLSTM (Microsoft)

HMM-DBN



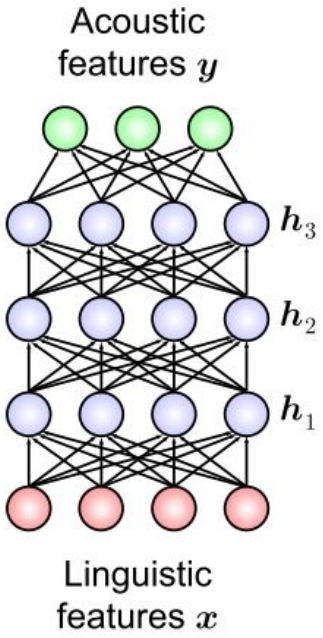
- Decision tree-clustered HMM with DBN state-output distributions
- DBNs replaces GMMs

DBN



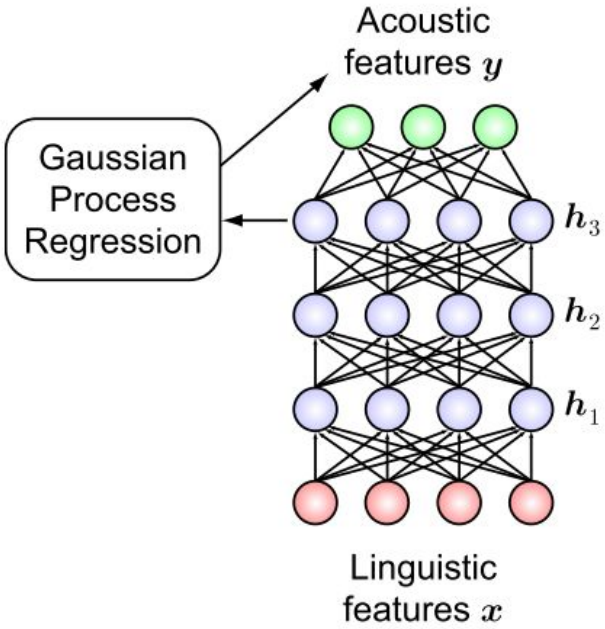
- DBN represents joint distribution of linguistic & acoustic features
- DBN replaces decision trees and GMMs

DNN



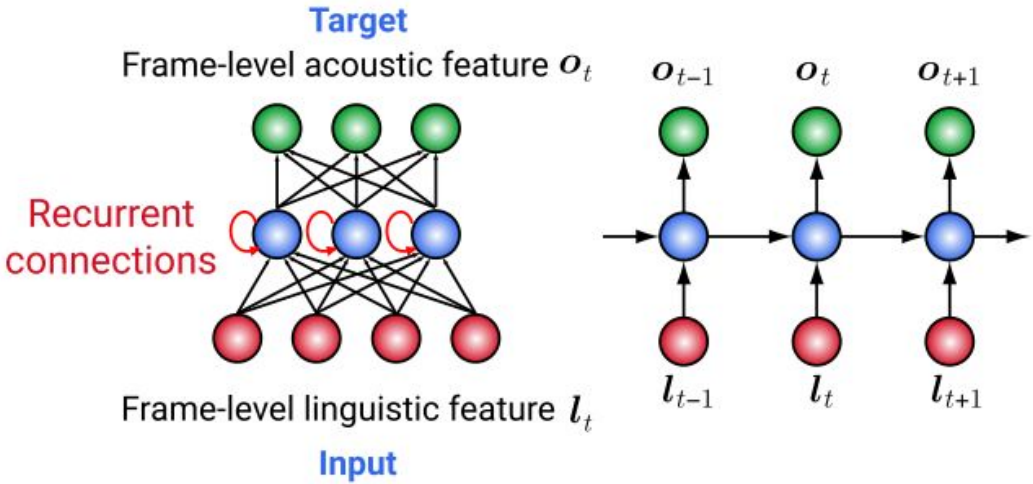
- DNN represents conditional distribution of acoustic features given linguistic features
- DNN replaces decision trees and GMMs

DNN-GP



- Uses last hidden layer output as input for Gaussian Process (GP) regression
- Replaces last layer of DNN by GP regression

LSTM



$$h_t = g\left(W_{hl} l_t + W_{hh} h_{t-1} + b_h\right)$$

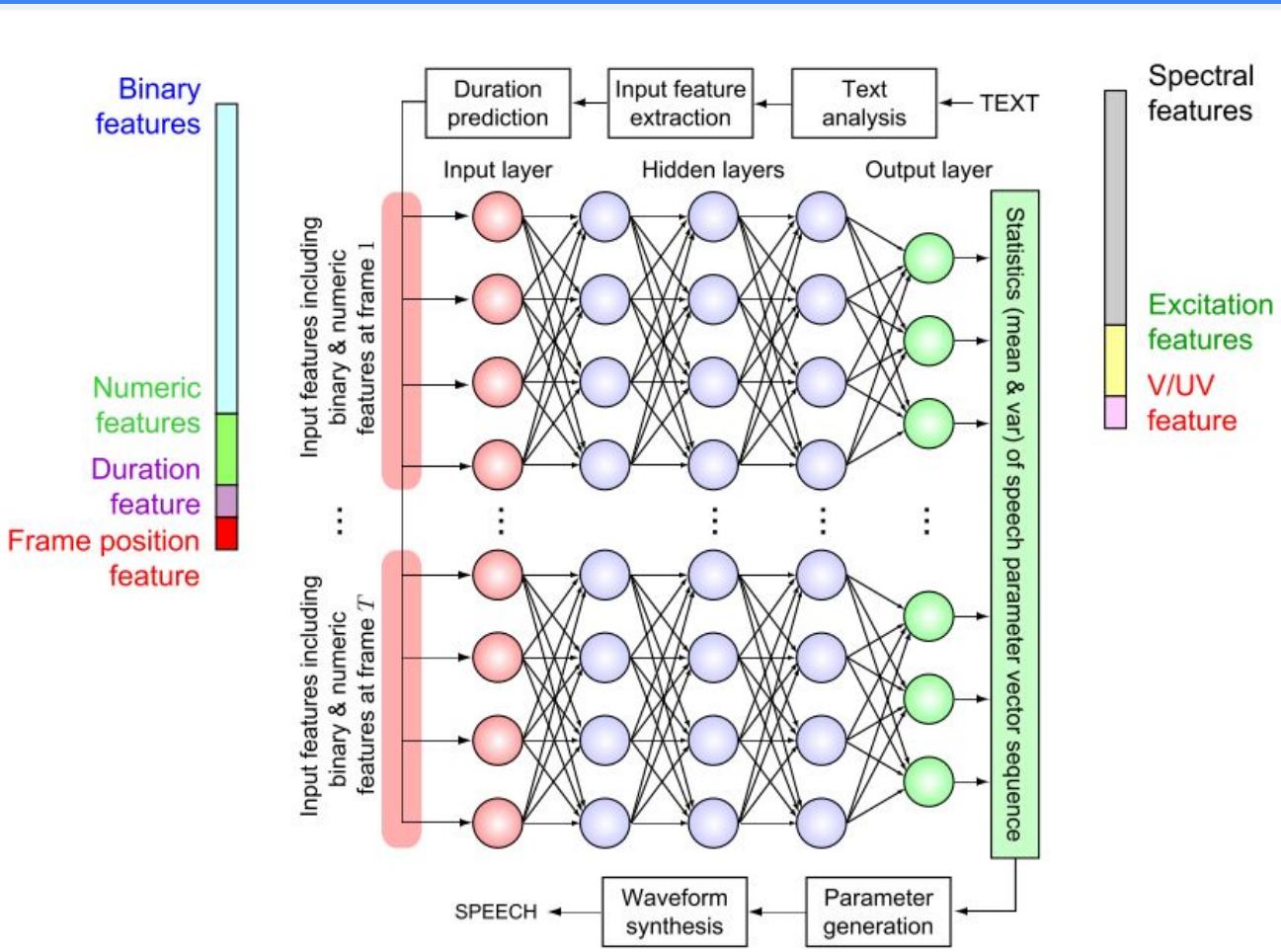
$$\hat{o}_t = W_{oh} h_t + b_o$$

$$\hat{\lambda} = \arg\min_{\lambda} \sum_t \|o_t - \hat{o}_t\|_2$$

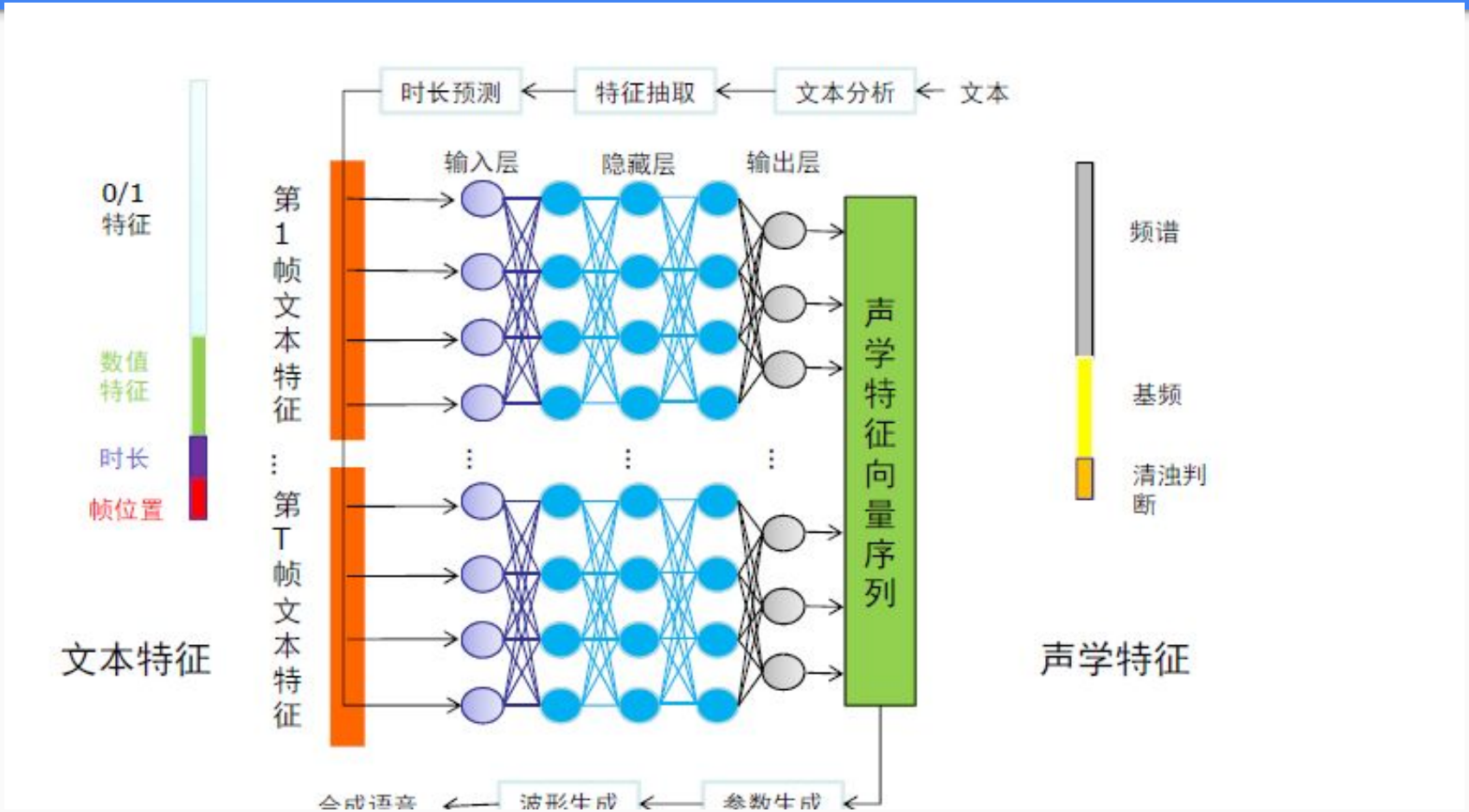
$$\lambda = \{W_{hl}, W_{hh}, W_{oh}, b_h, b_o\}$$

FFNN: $\hat{o}_t \approx \mathbb{E}\left[o_t \mid l_t\right]$ RNN: $\hat{o}_t \approx \mathbb{E}\left[o_t \mid l_1, \ldots, l_t\right]$

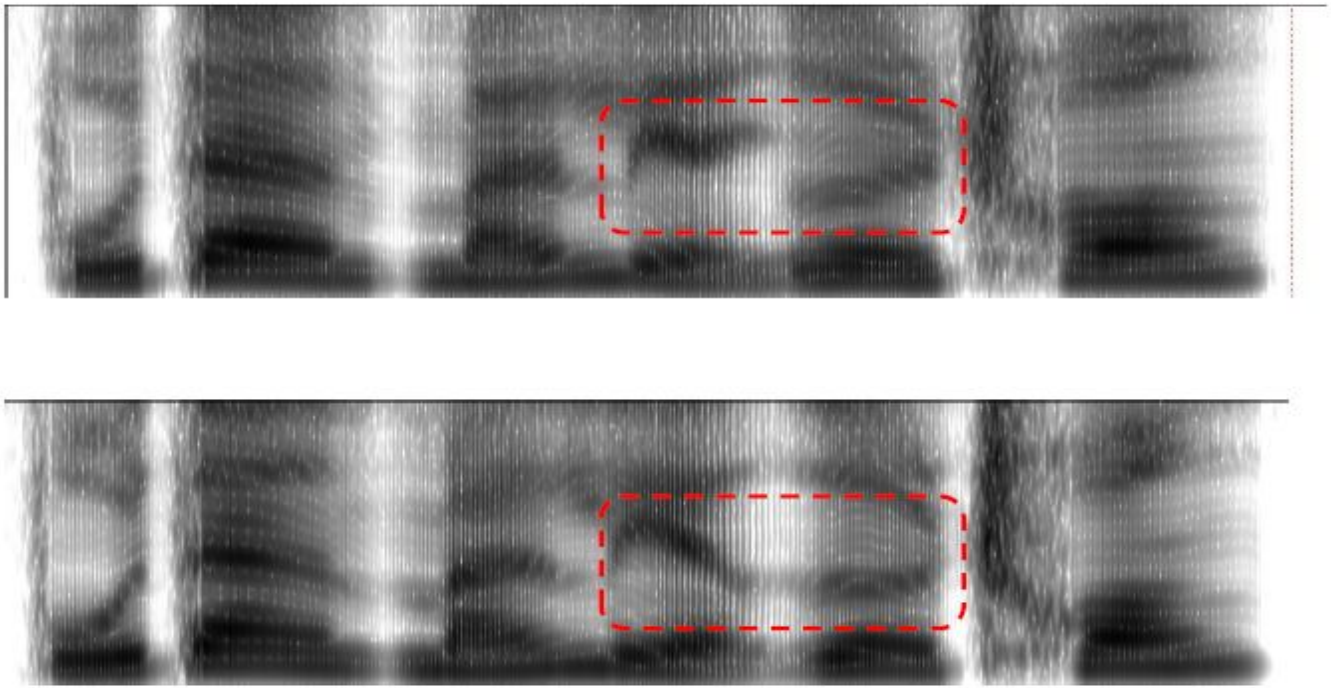
NN-Framework



NN-Framework



共振峰不连续的问题



HTS共振峰不连续的例子，LSTM会得到完善

Parametric - Summary

- **HMM**
  - Discontinuity due to step-wise statistics
  - Difficult to integrate feature extraction
  - Fragmented representation
- **Feedforward NN**
  - Easier to integrate feature extraction
  - Distributed representation
  - Discontinuity due to frame-by-frame independent mapping
- **(LSTM) RNN**
  - Smooth $\rightarrow$ Low latency

Speech Synthesis - Past

Concatenation Synthesis

Concatenative Speech Synthesis

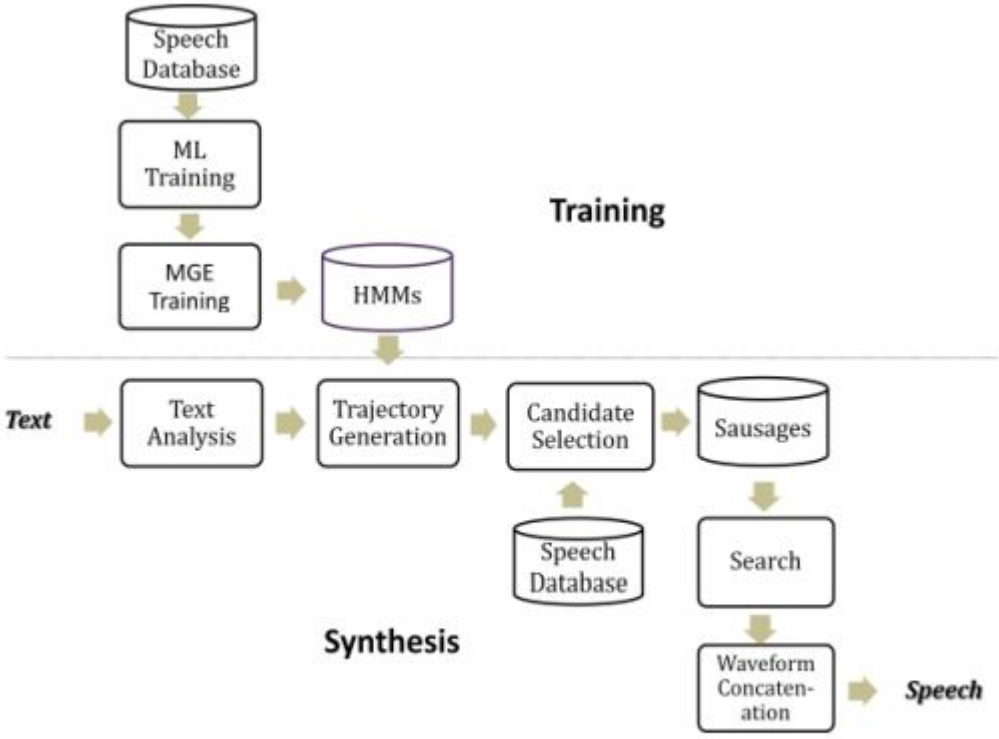


Figure 1: *Schematic diagram of HMM trajectory tiling based speech synthesis.*

Speech Synthesis - Present

Speech Synthesis - Present

End-to-End Model

E2E - Tacotron(2017.3)

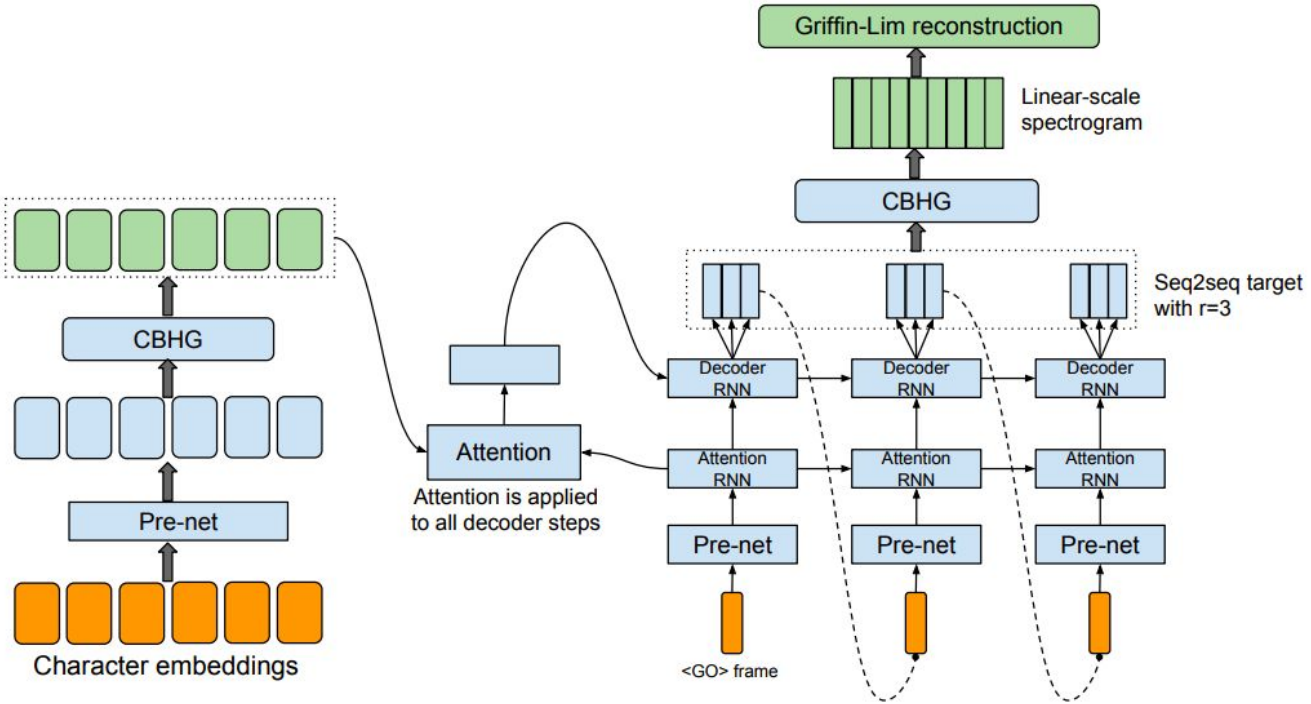


Figure 1: *Model architecture. The model takes characters as input and outputs the corresponding raw spectrogram, which is then fed to the Griffin-Lim reconstruction algorithm to synthesize speech.*

E2E - Uncovering Latent Style Factors for Expressive Speech Synthesis(2017.11)

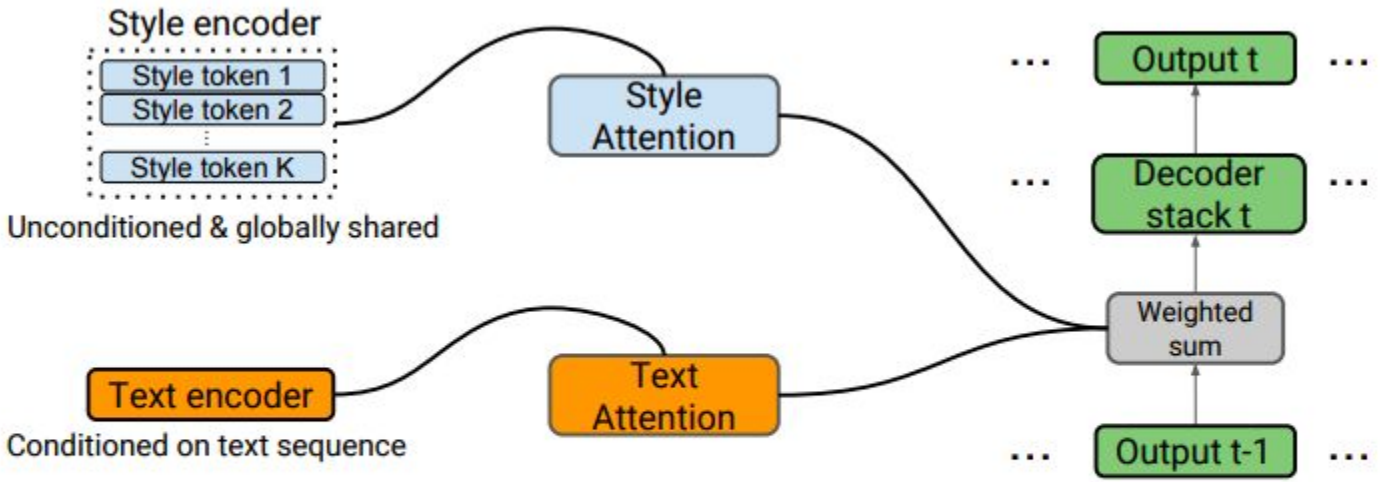
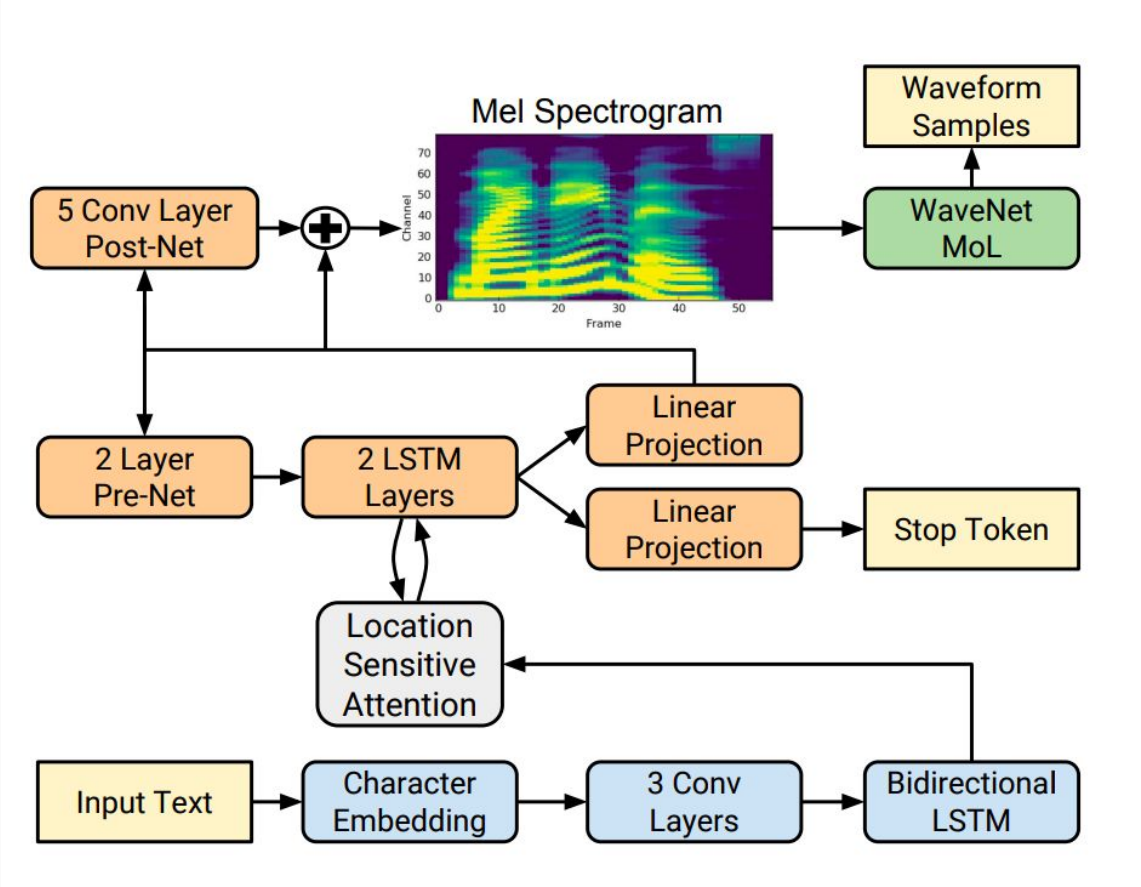
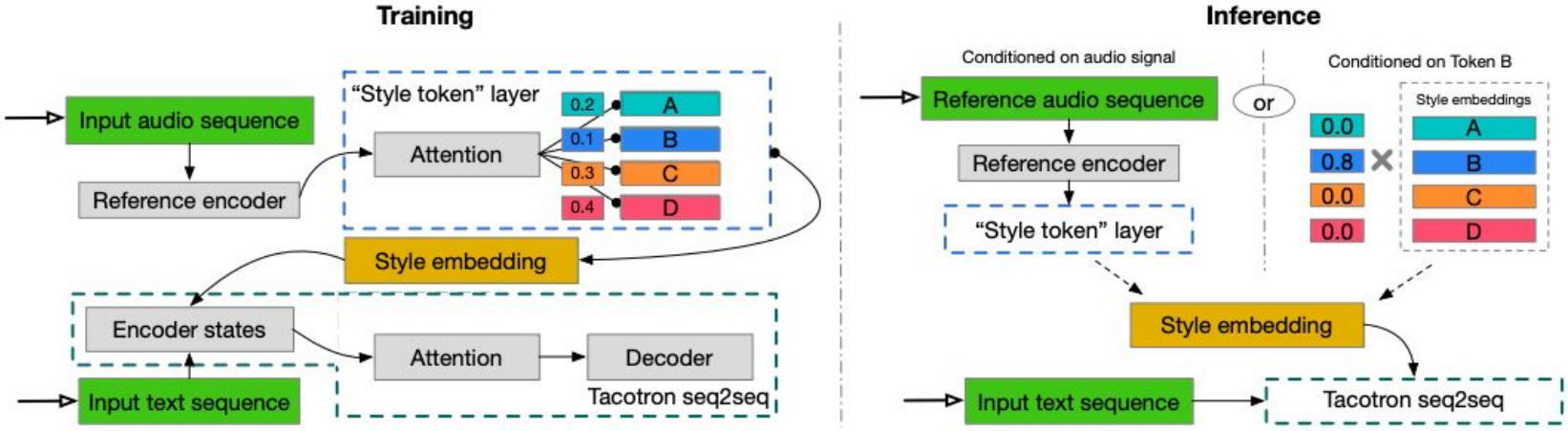


Figure 1: Model architecture cartoon based on Tacotron [2]. To learn style tokens in Tacotron, we add an additional style encoder and the corresponding attention pathway. See [2] for Tacotron architecture details.

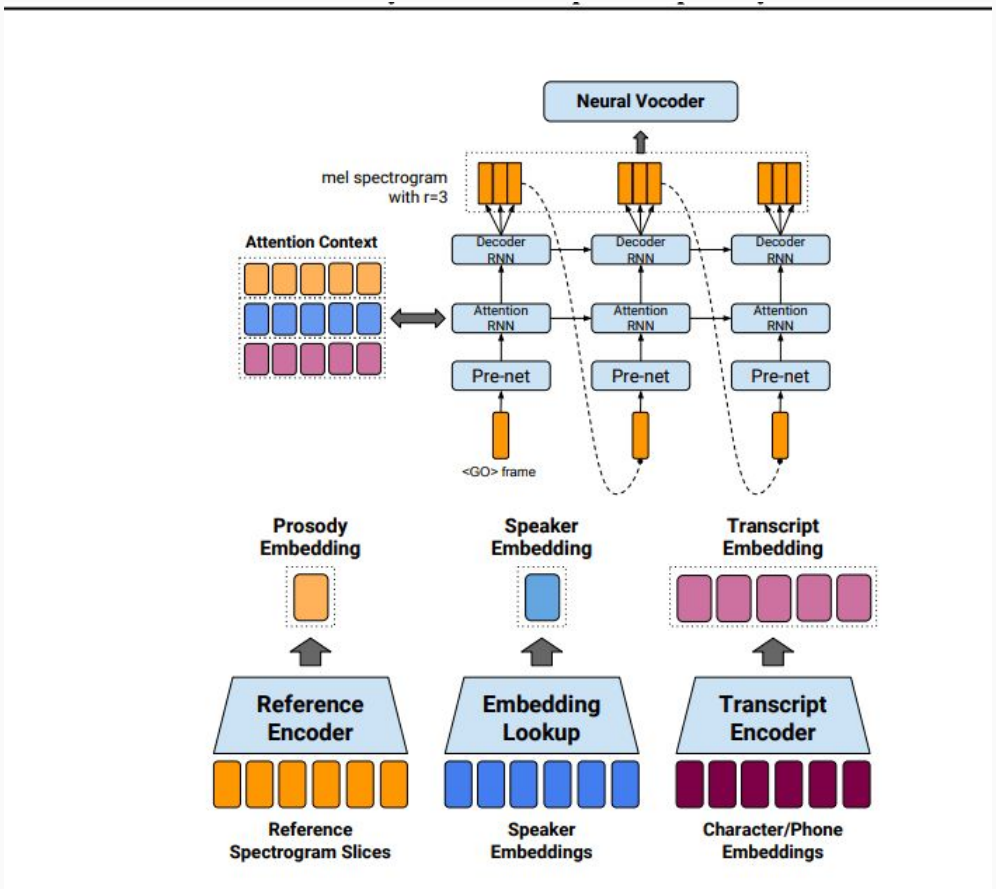
E2E - Tacotron2(2017.12)



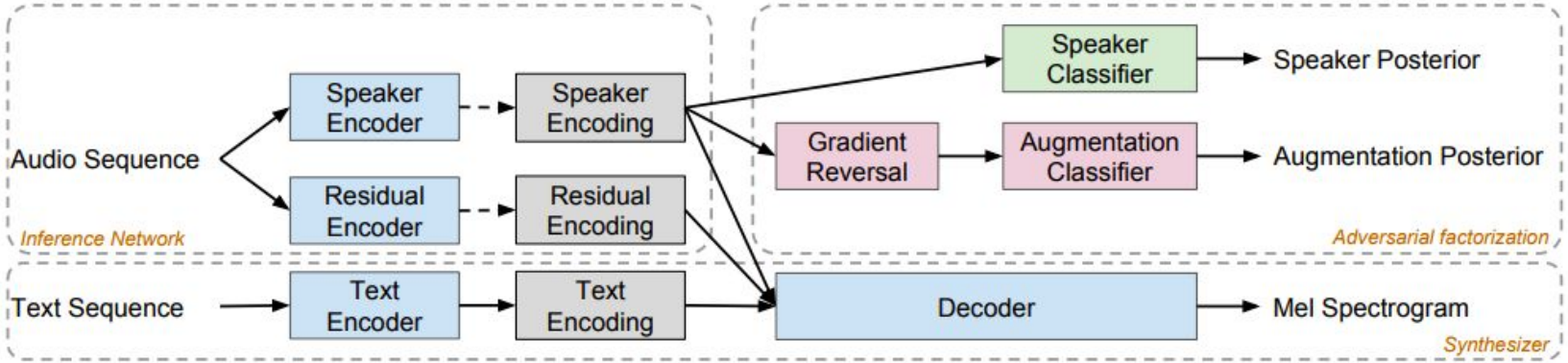
E2E - Style Token(2017.12)



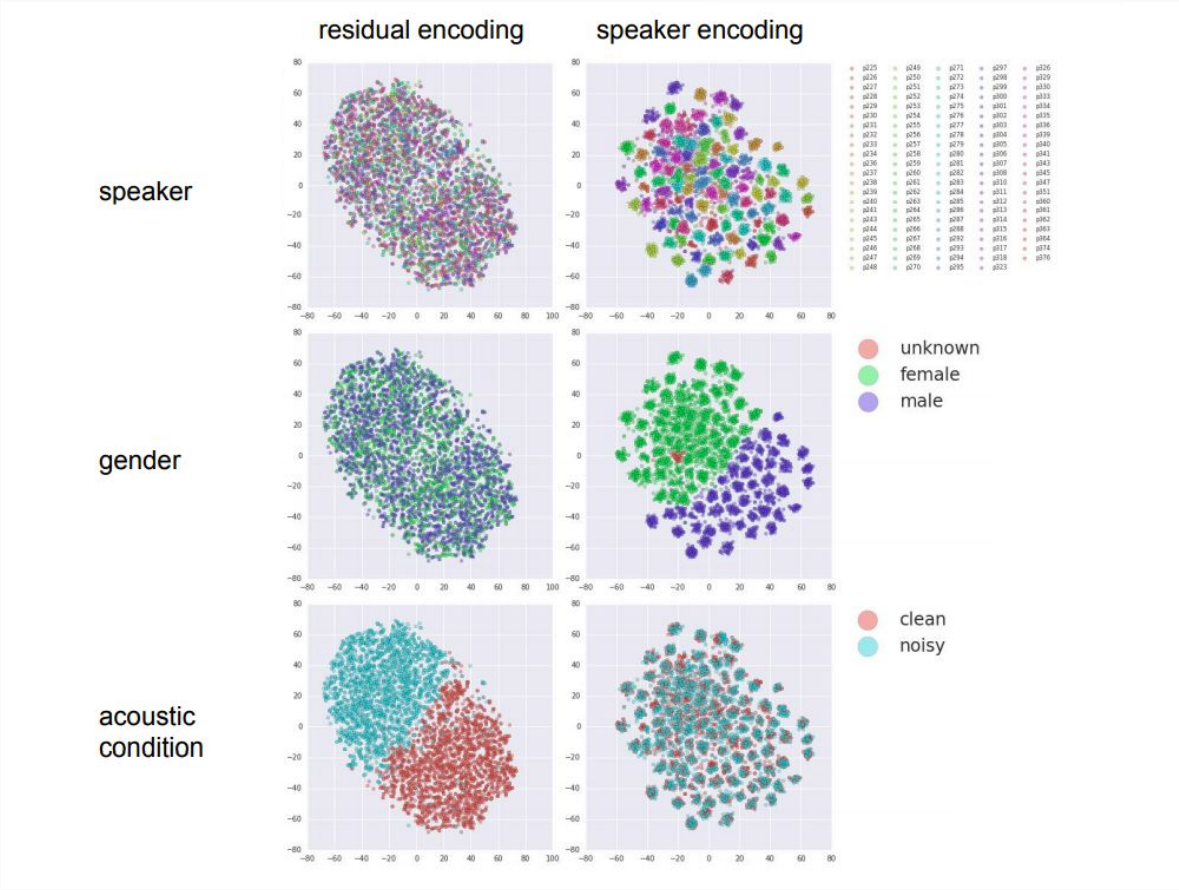
E2E - Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron(2018.03)



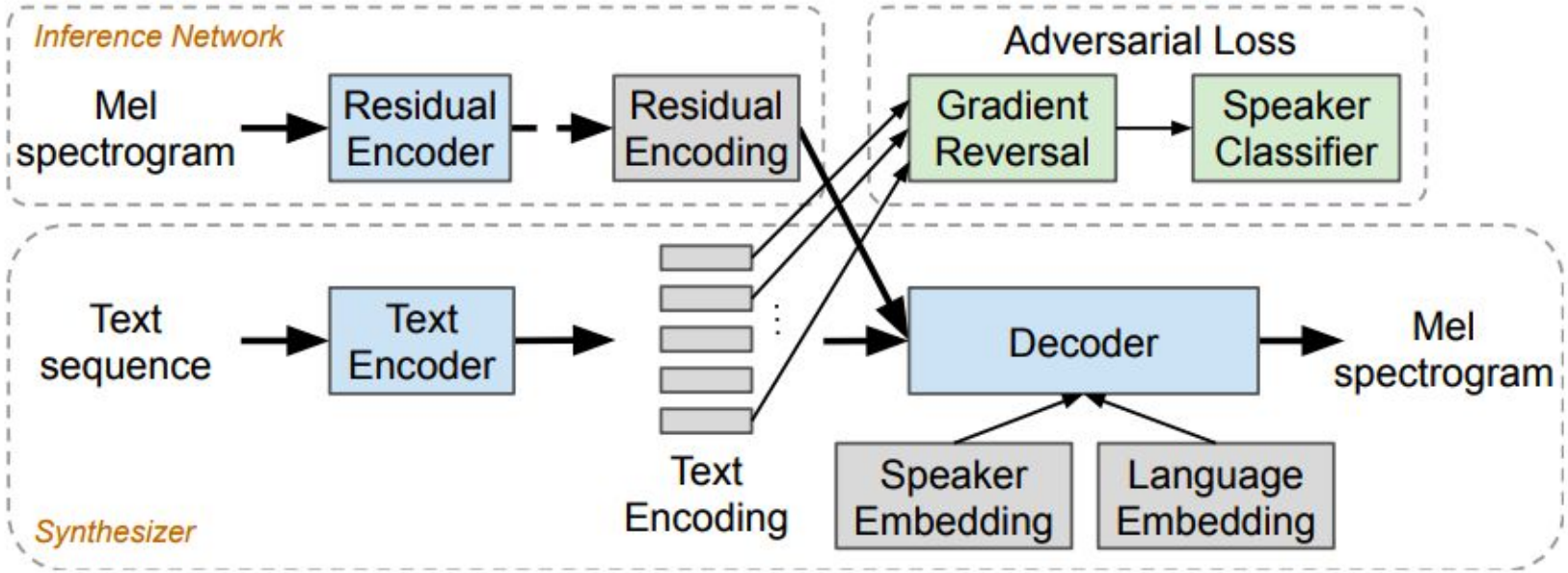
E2E - Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization (2018.11)



E2E - Disentangling Correlated Speaker and Noise for Speech Synthesis via Data Augmentation and Adversarial Factorization (2018.11)



E2E - Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning (2019.07)



End-to-End Model Summary

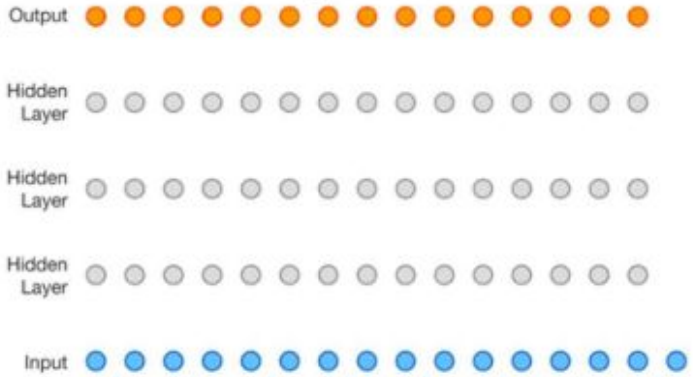
https://google.github.io/tacotron/

Speech Synthesis - Present

Neural Vocoder

Wavenet(2016.09)

- A generative model directly on the raw audio waveform (AR) $p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$
- Stacked dilated casual convolutional layers
  * Model longer sequence with limited parameters



Wavenet(2016.09)

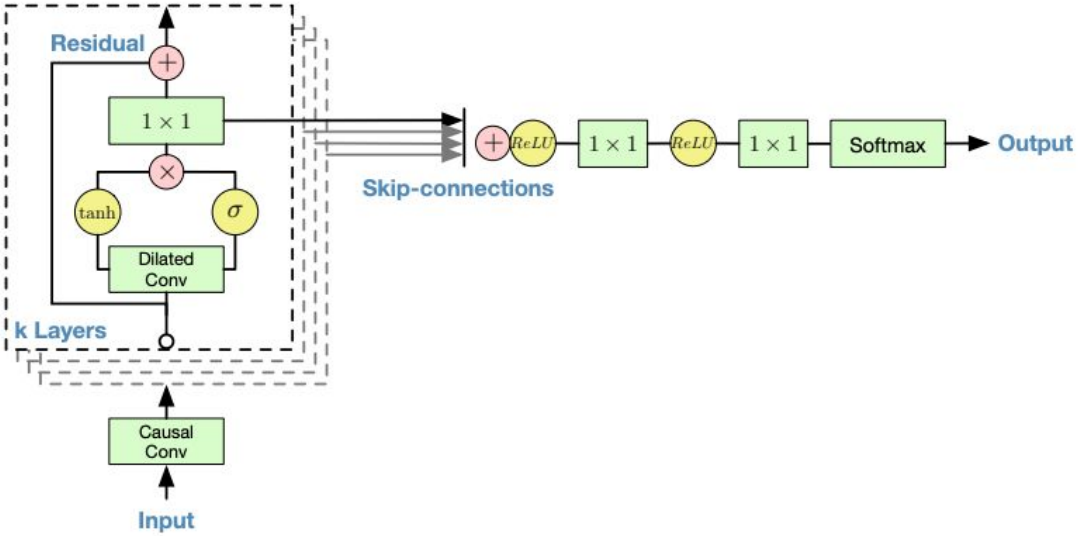
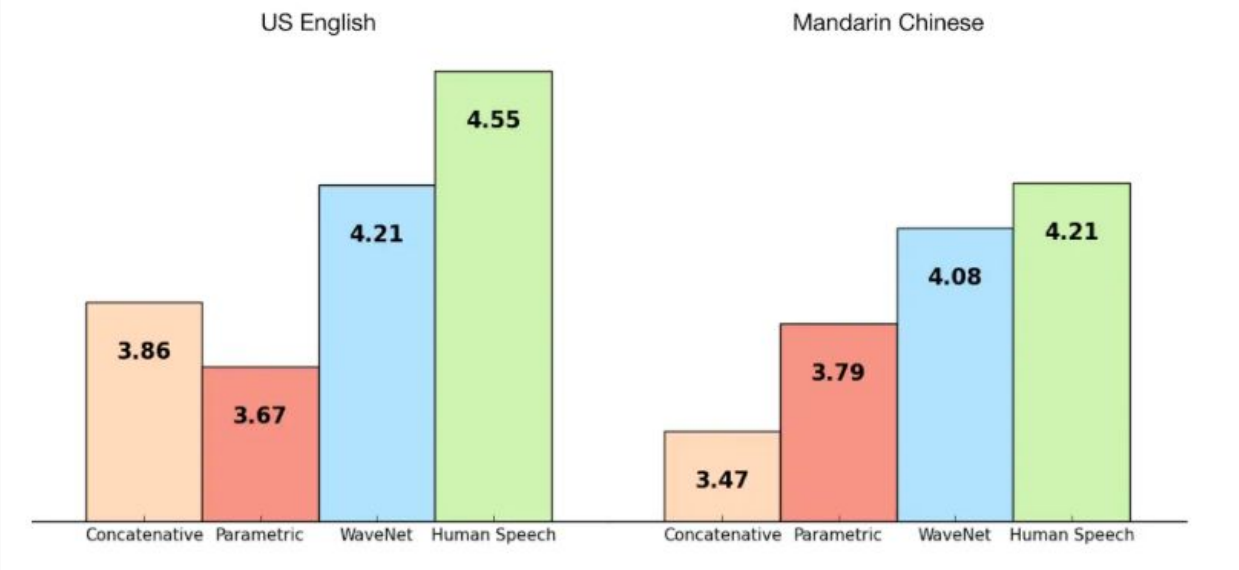


Figure 4: Overview of the residual block and the entire architecture.

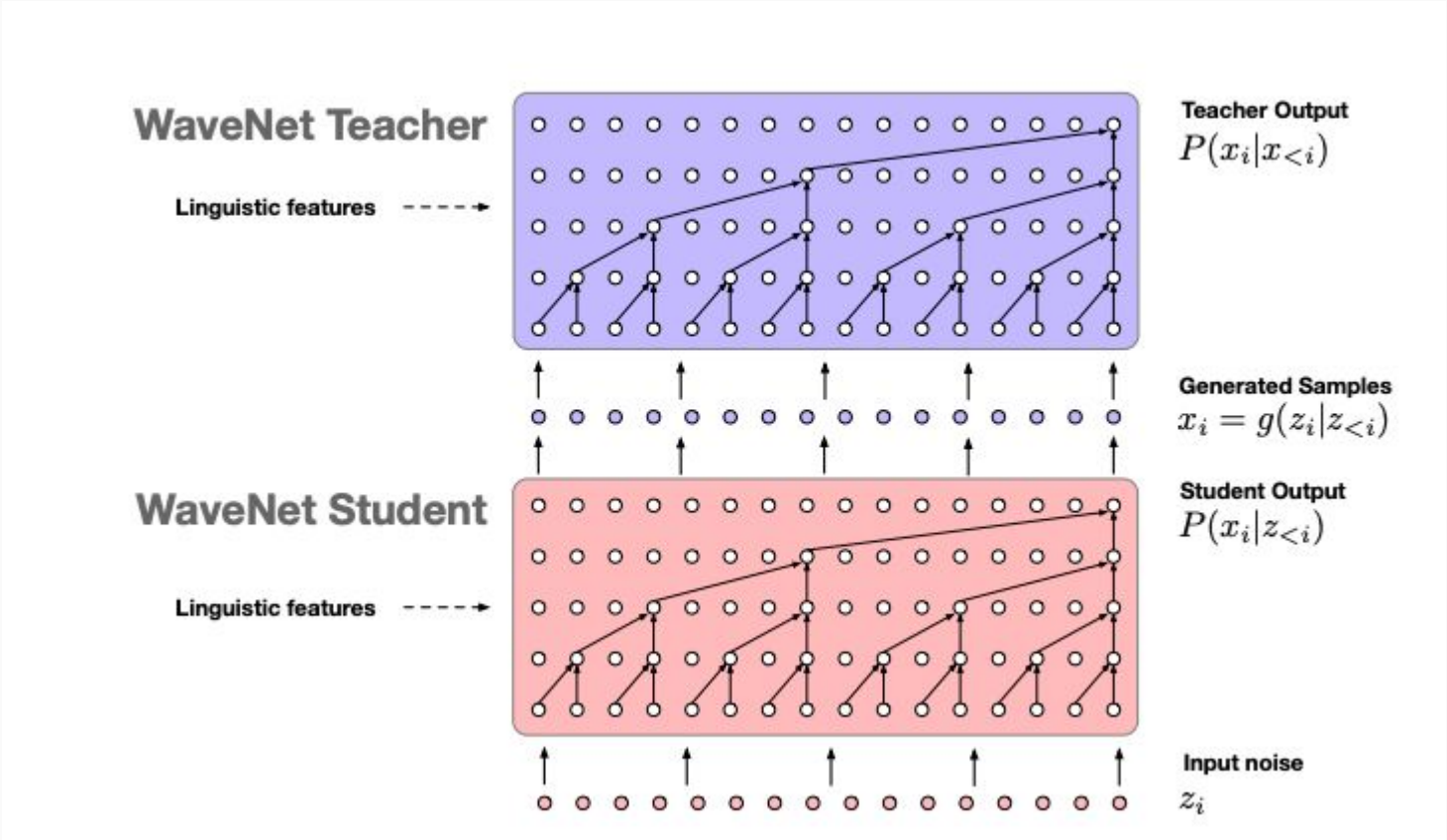
Wavenet(2016.09)



Wavenet(2016.09)

Inference too SLOW !!!!!

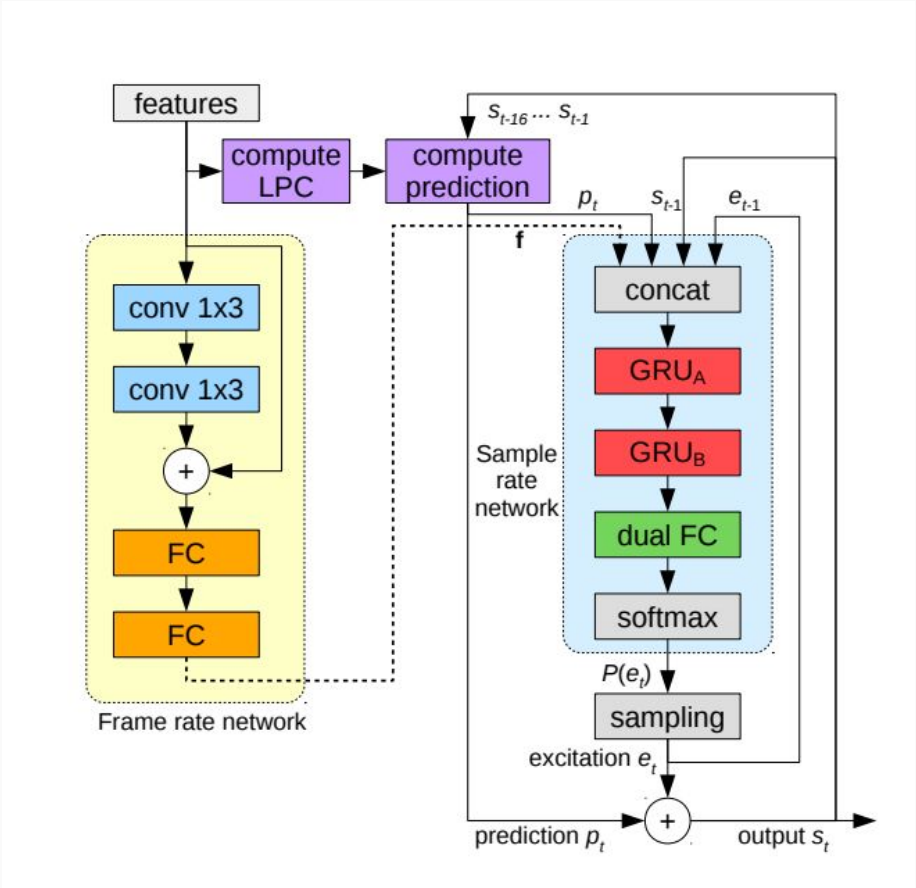
Parallel Wavenet(2017.11)



Parallel Wavenet(2017.11)

1000x faster than Wavenet

LPCNet(2018.11)



Present - others

- Autoregressive
  - WaveNet, SampleRNN, WaveRNN, ...
- Normalizing flow
  - WaveGlow, Parallel WaveNet, ClariNet, FloWaveNet, WaveFlow ...
- Combining with source filter model
  - LPCNet, ExcitNet, GlotNet, LP-WaveNet, …
- Introducing signal processing technique
  - SubbandWaveNet, FFTNet, ...

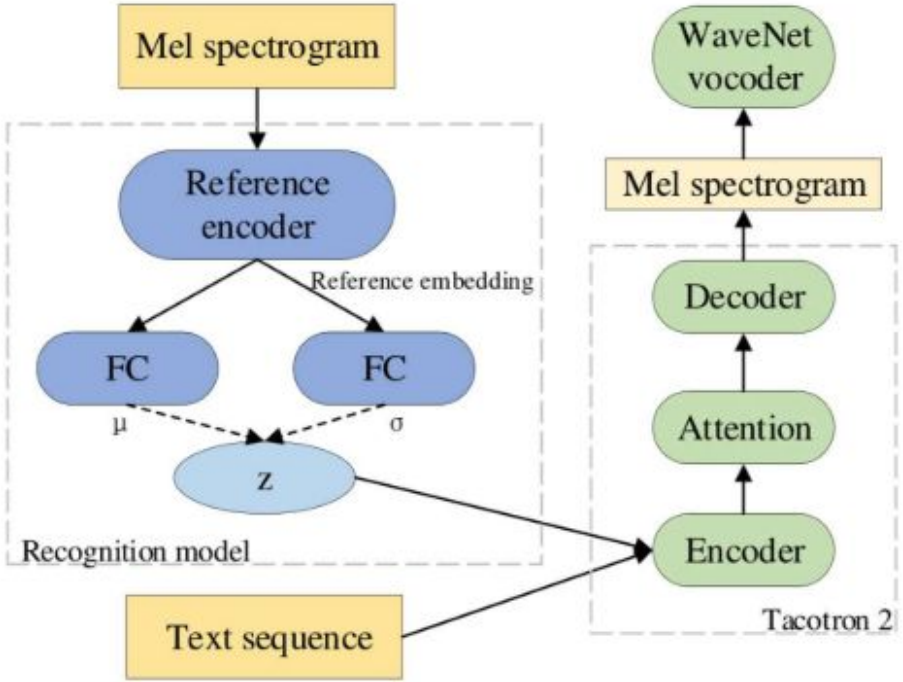
Speech Synthesis - Future

Universal model and transfer learning across Styles, Speakers and Languages

Transfer learning: across styles

Transfer Learning: across styles

- Variational Autoencoder captures the style embedding of each utterance to enable the across style transfer learning.



Transfer learning: across speaker

Transfer learning: across speakers

- Greatly reduce the data requirements to build high quality TTS model.
- 3 minutes recordings result in high similarity.

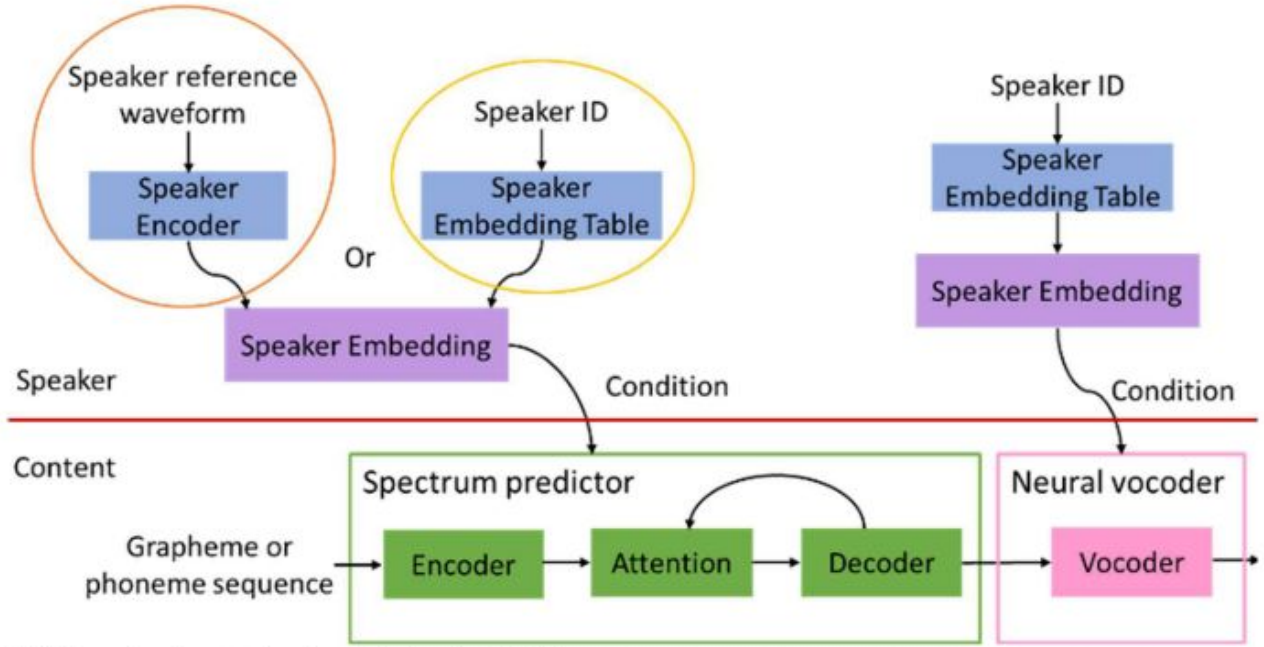


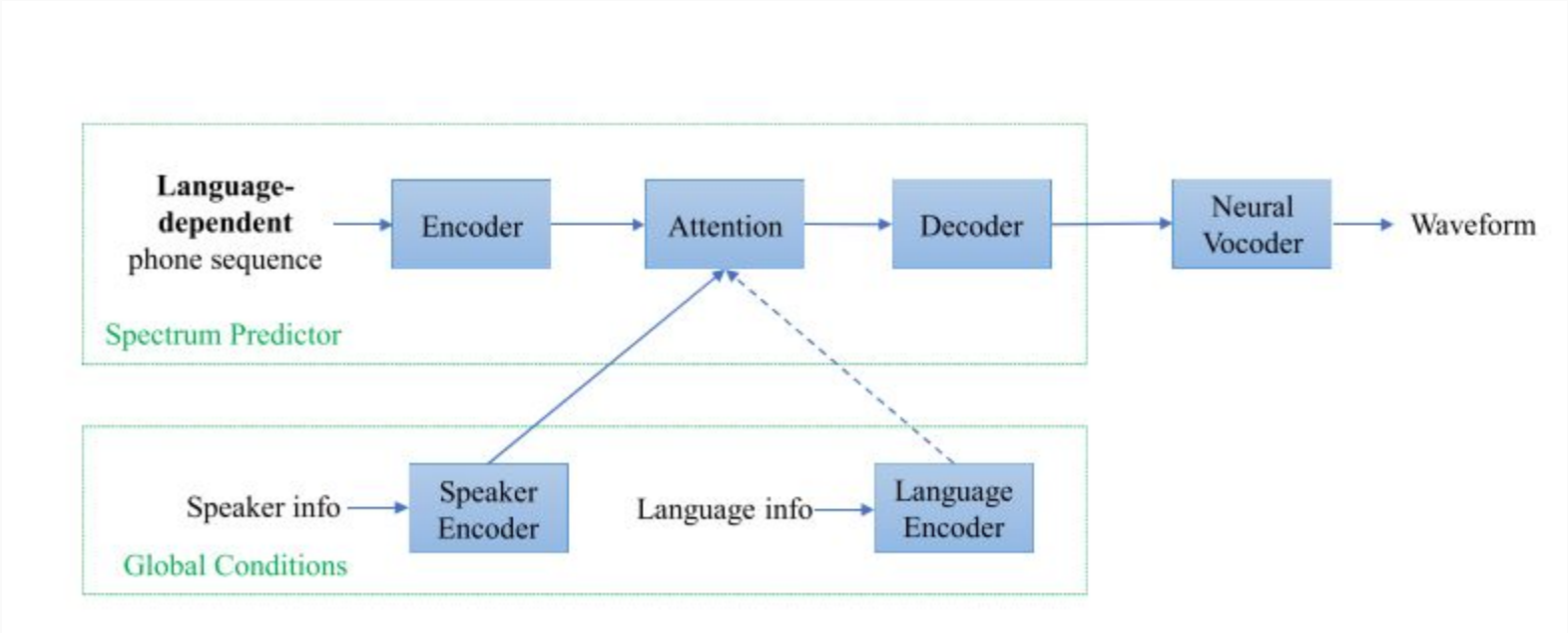
Table 1. MOS for naturalness on in-domain test set with 95% confidence intervals

Recording	Neural TTS	FF-LSTM
4.57±0.1	4.16±0.07	3.65±0.1

Table 2. MOS for speaker similarity with 95% confidence intervals.

Recording	Neural TTS	FF-LSTM
4.74±0.1	4.64±0.07	4.09±0.1

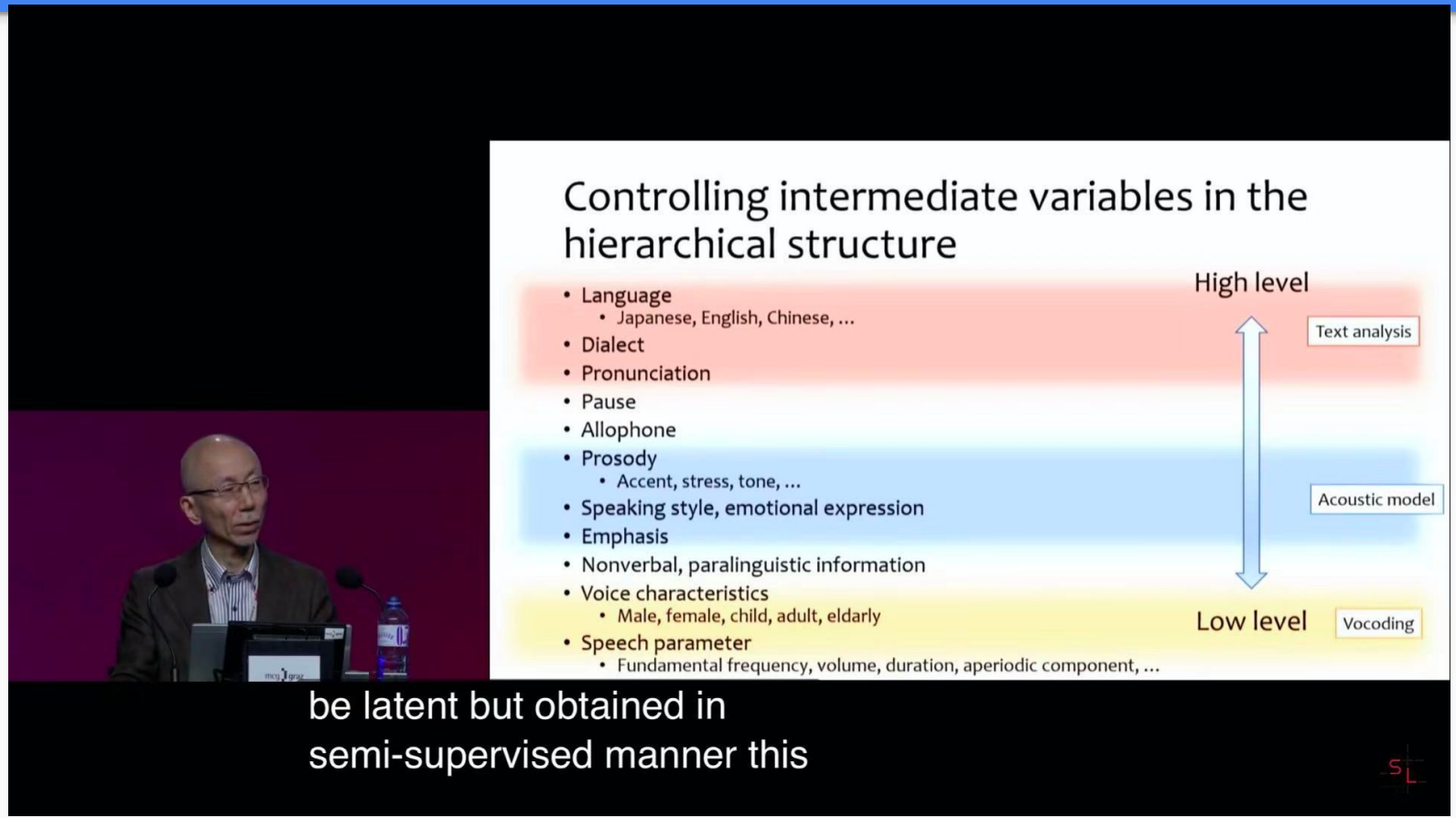
Transfer learning: across language



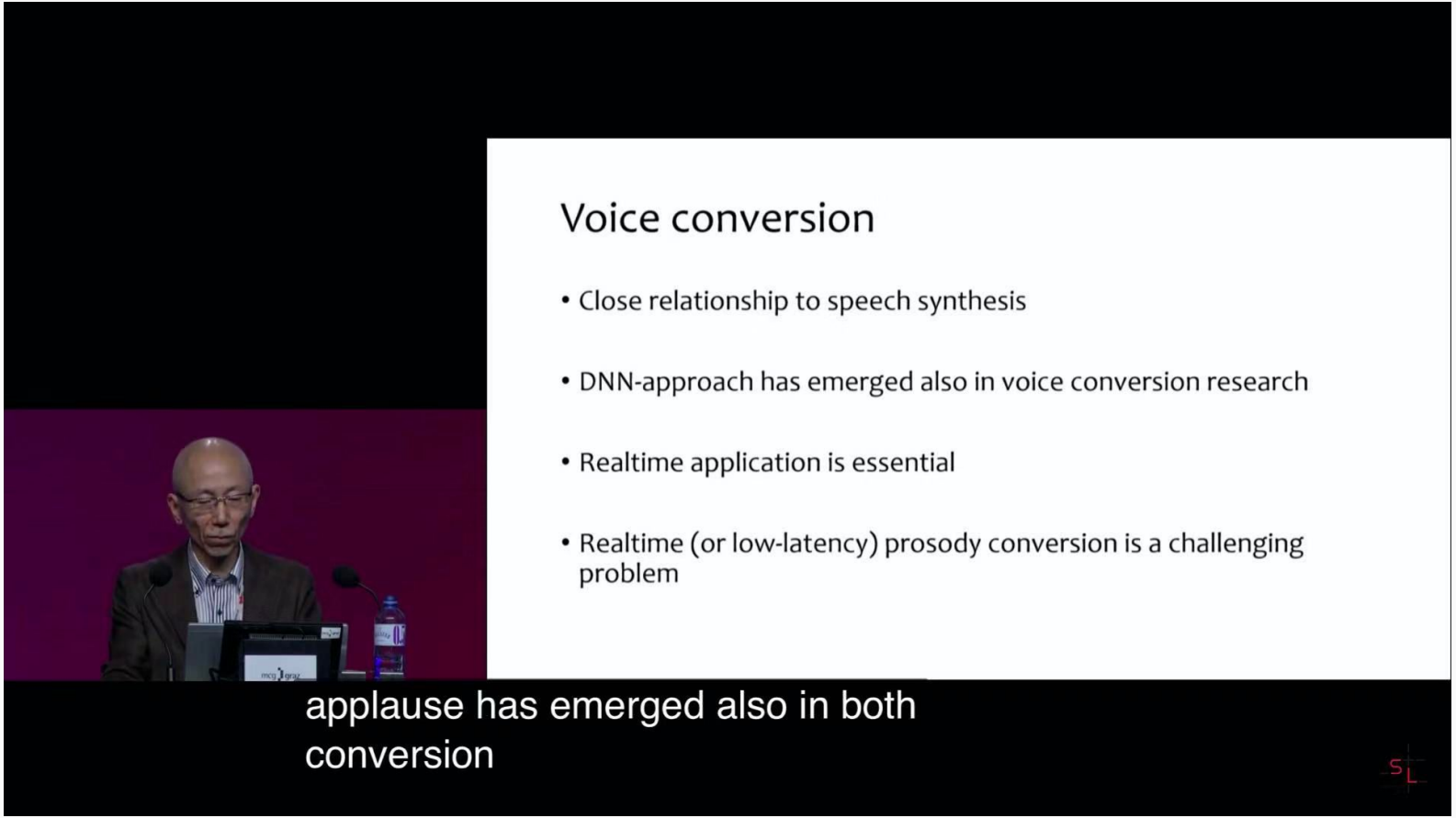
Future - more

- Emotion TTS
- Singing synthesis
- General speech for TTS
- TTS with less T or without T
- Voice Conversion
- AI 虚拟说话人
- ...

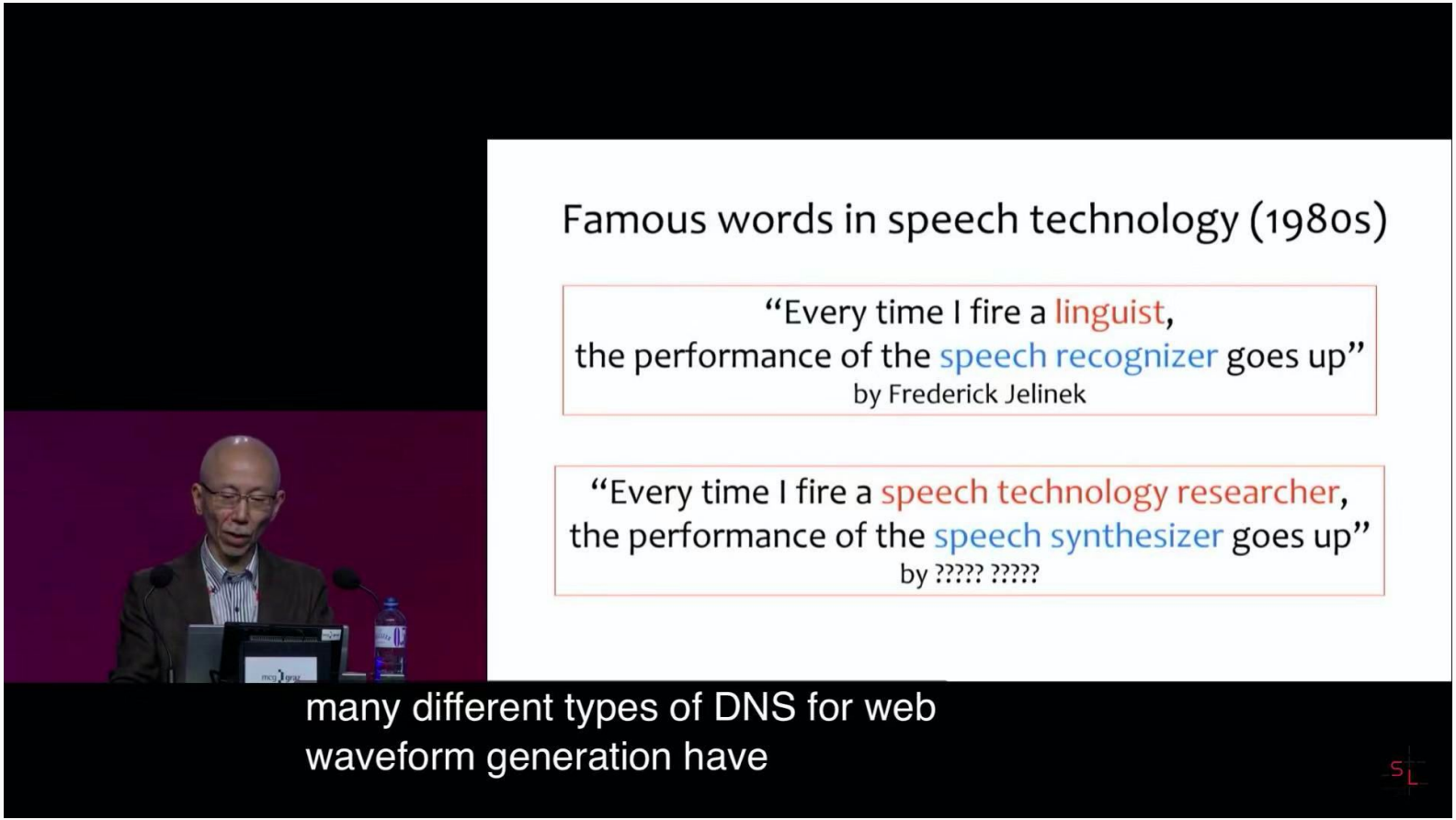
Future - more


Controlling intermediate variables in the hierarchical structure
High level
• Language
• Japanese, English, Chinese, ...
• Dialect
• Pronunciation
• Pause
• Allophone
• Prosody
• Accent, stress, tone, ...
• Speaking style, emotional expression
• Emphasis
• Nonverbal, paralinguistic information
• Voice characteristics
• Male, female, child, adult, eldarly
• Speech parameter
• Fundamental frequency, volume, duration, aperiodic component, ...
Text analysis
Acoustic model
Low level
Vocoding
be latent but obtained in
semi-supervised manner this

Future - more


Voice conversion
• Close relationship to speech synthesis
• DNN-approach has emerged also in voice conversion research
• Realtime application is essential
• Realtime (or low-latency) prosody conversion is a challenging problem
applause has emerged also in both conversion

Future - more


Famous words in speech technology (1980s)
"Every time I fire a linguist,
the performance of the speech recognizer goes up"
by Frederick Jelinek
"Every time I fire a speech technology researcher,
the performance of the speech synthesizer goes up"
by ????? ?????
many different types of DNS for web
waveform generation have

Thanks!