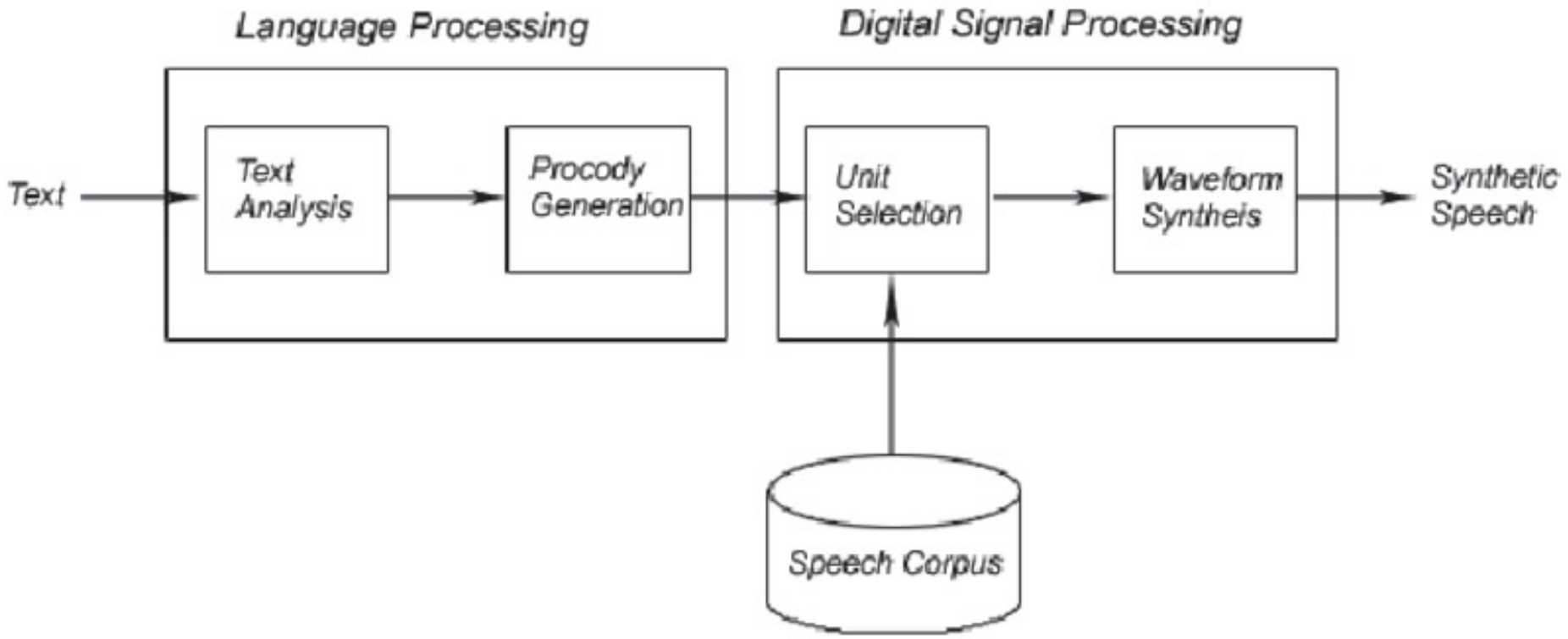
Moving the frontiers of Text-to-Speech with NLP

Animesh Prasad

Part I

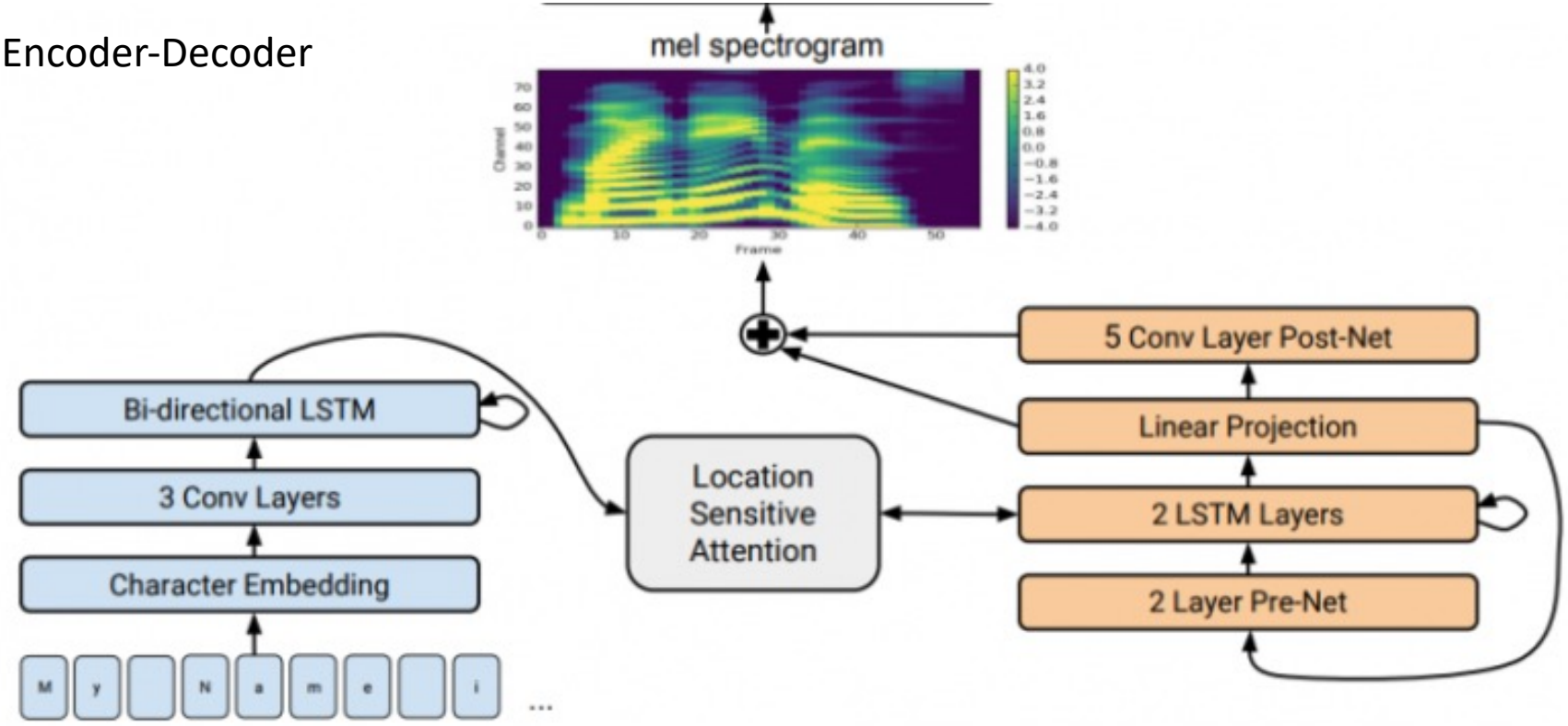
A brief introduction to TTS

Even before dawn of time ….

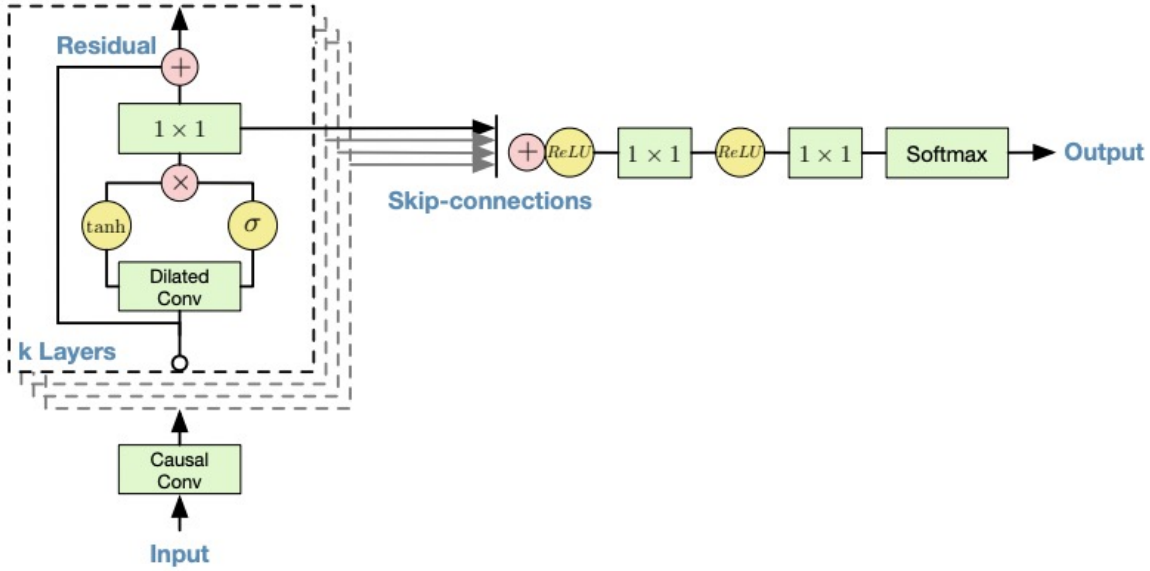


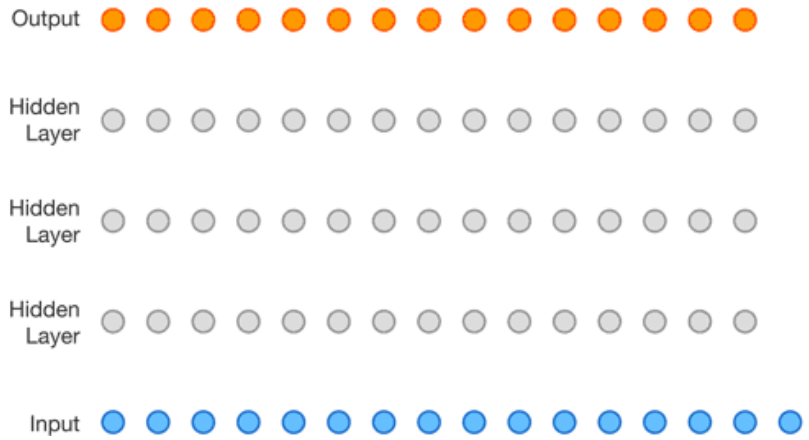
And then there was light

Basically Encoder-Decoder



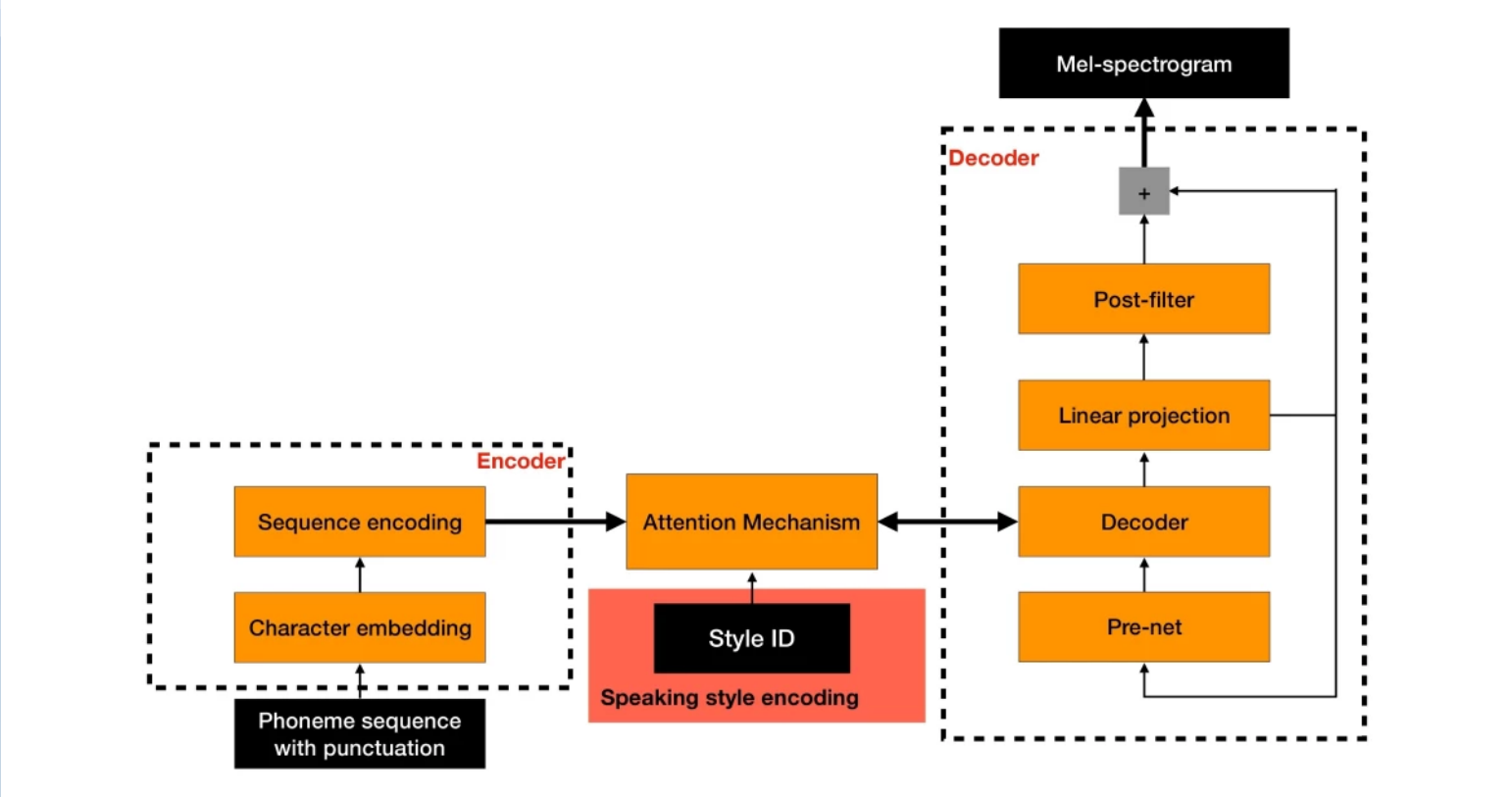
And there was more light





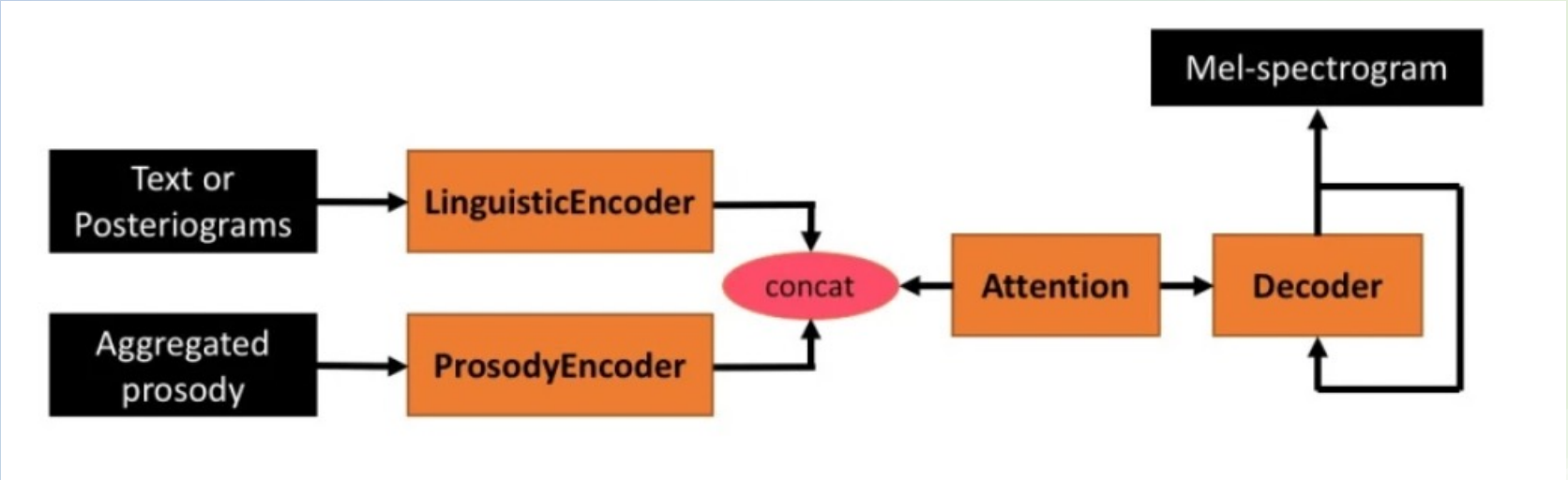
But now there was a whole new world ….

Multi-Speaker
Multi-Style
Multi-language



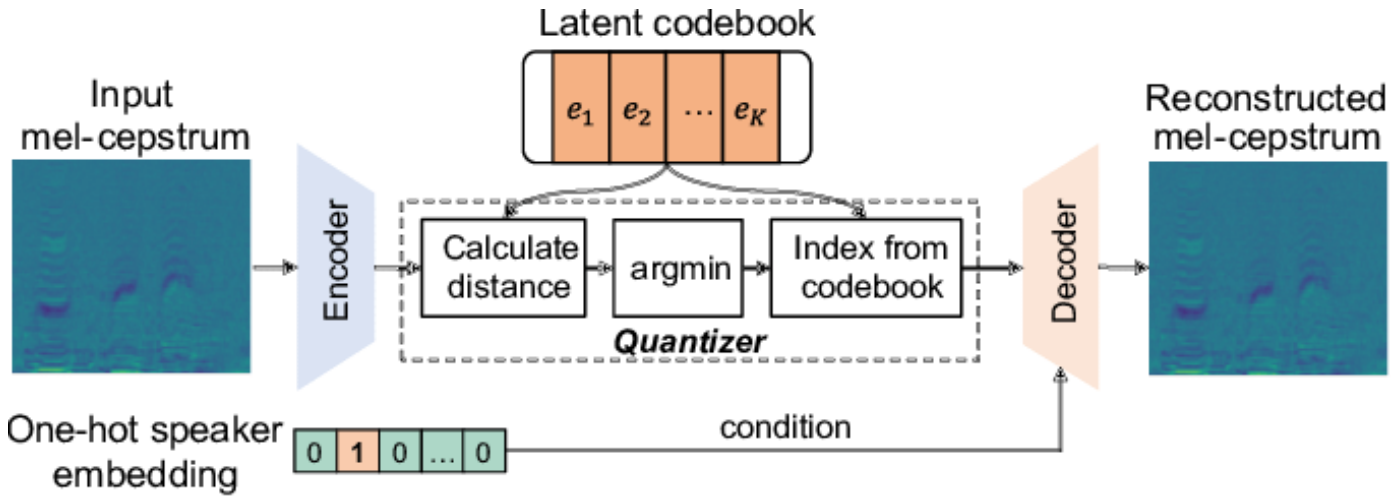
But now there was a whole new world ….

Encoding
Prosody

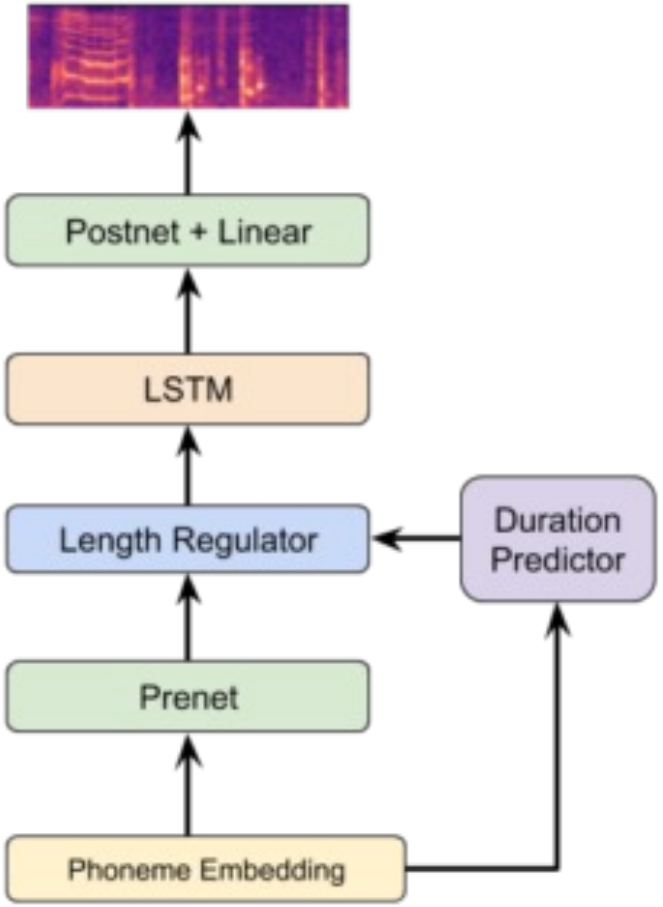


Training: Original
Inference: Mean, Transfer

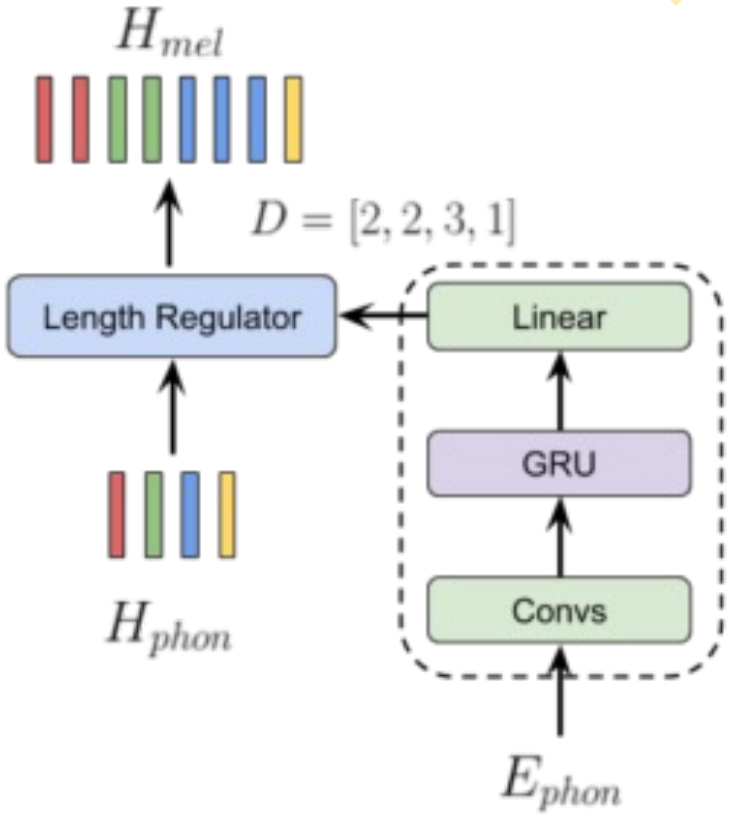
- Predict latent during inference



External Duration makes a comeback…

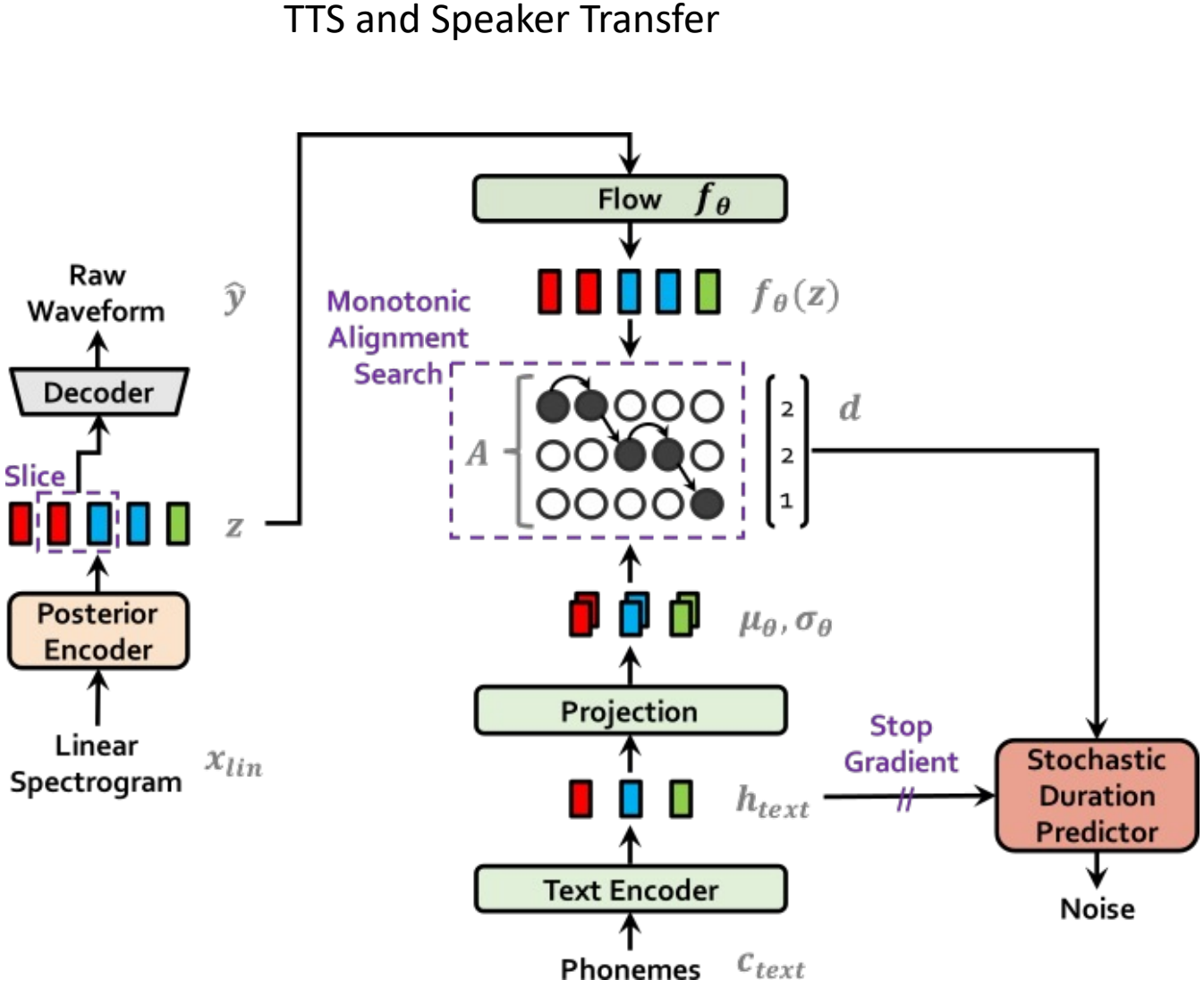


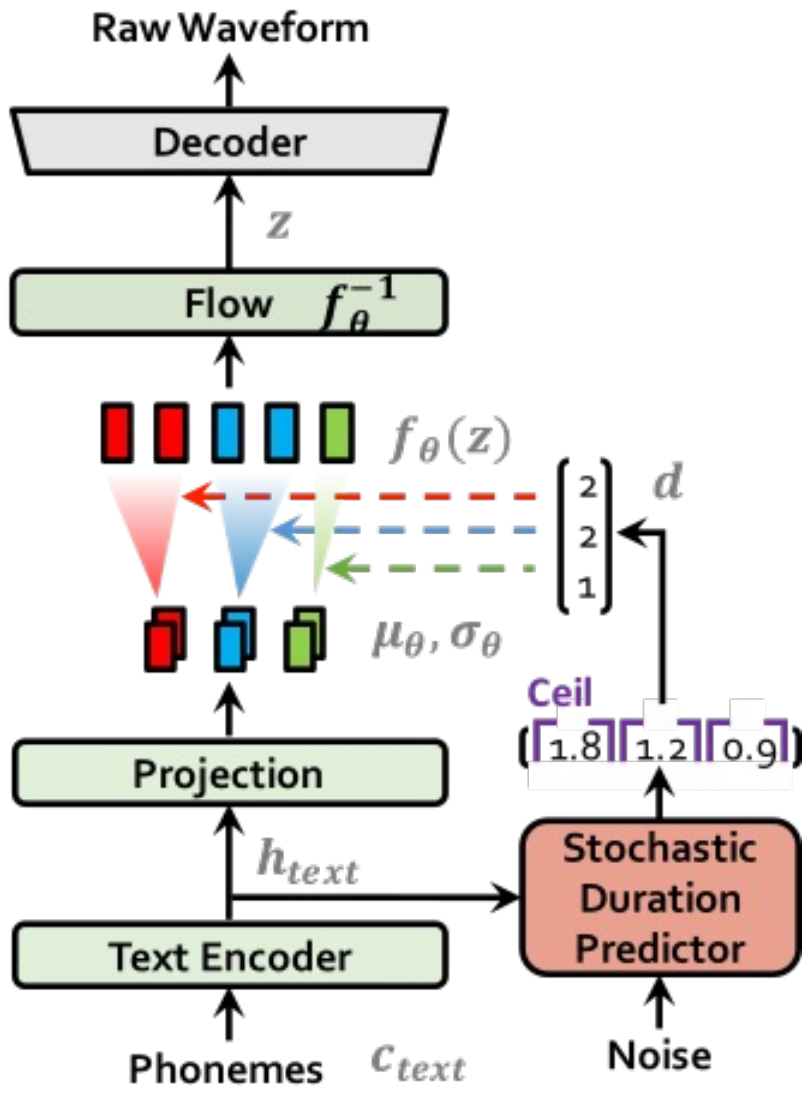
Forward Tacotron



Length Regulator with Duration Predictor

TTS and Speaker Transfer





Part II

Unlocking NLP For TTS

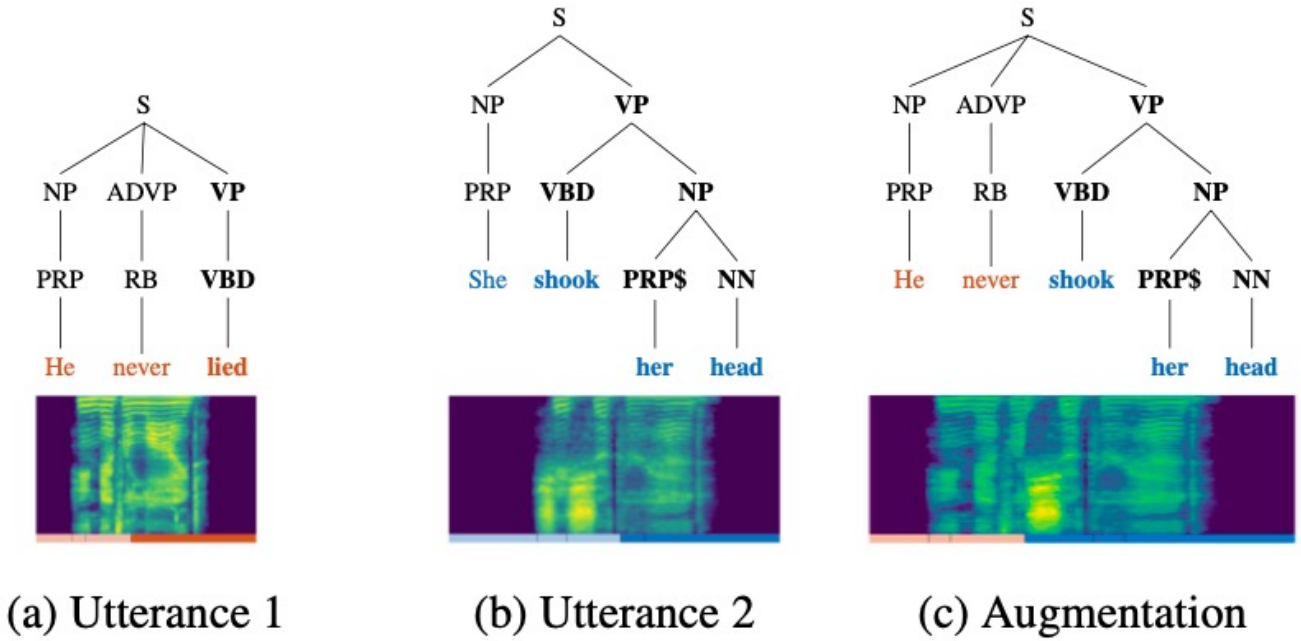
Distribution augmentation for low resource expressive TTS

ICASSP 2022

Alexa AI

Mateusz Lajszczak, Animesh Prasad et al.

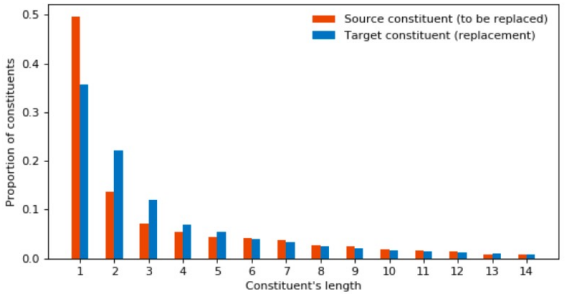
Subtree Substitution

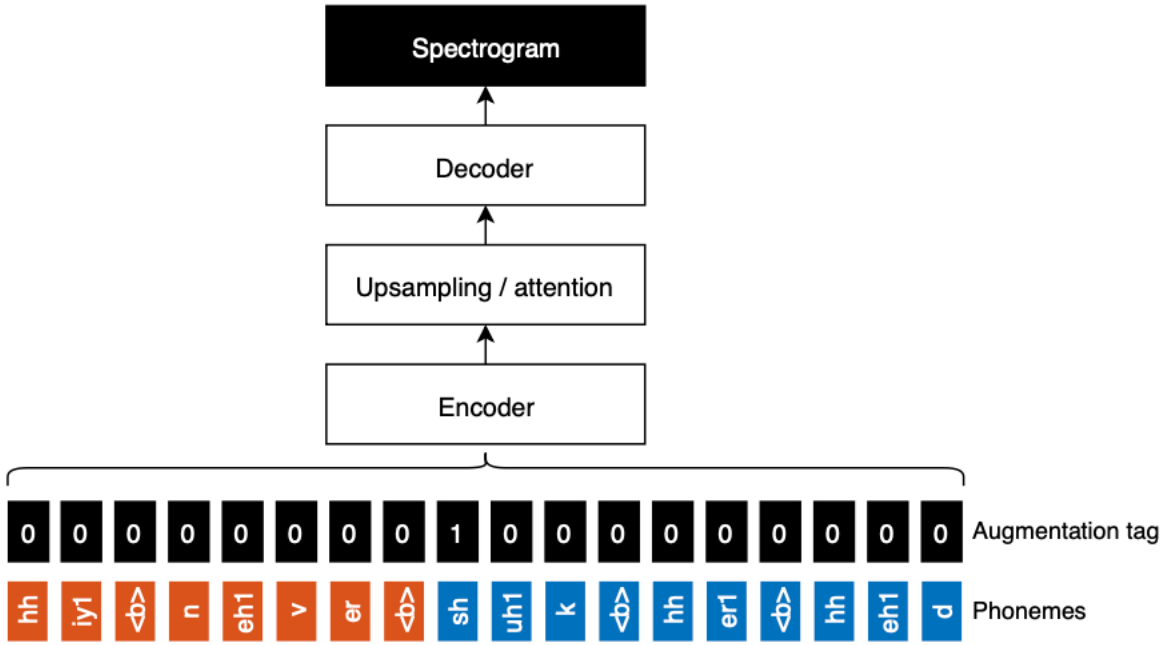


(a) Utterance 1 (b) Utterance 2 (c) Augmentation

- Standard TTS augmentation: Get many speakers, train a canonical model
  - Finetune later for target speaker
  - Or embed new speaker

- Single Speaker ?





Backwards Encoder:
P(right|flag)

Forward Encoder:
P(left|flag)

During training:
P(right,flag,left)

During inference :
P(left, right)

Results (Stability)

Dataset	WER		PER	
	Base	Ours	Base	Ours
D_1^{10h}	0.74	**0.26**	0.59	**0.07**
D_1^{5h}	1.00	**0.19**	1.00	**0.05**

Table 1: Word and phoneme error rates (bolded is better) of attention-based models for baseline (Base) and proposed models (Ours), relative to the WER and PER of 5h baseline model.

Results (Naturalness; Attention -A)

Dataset	% Preference			Test loss	
	Base	Ours	None	Base	Ours
D_1^{10h}	38	**41.1**	20.9	0.034	0.028
D_1^{5h}	33.5	**46**	20.5	0.039	0.031

Table 2: Preference ratings (bolded is preferred) between baselines (Base) and proposed models (Ours) with architecture (A).

Results (Naturalness; External Duration -B)

Dataset	% Preference			Test loss	
	Base	Ours	None	Base	Ours
D_1^{10h}	35.6	**40.3**	24.1	0.035	0.034
D_1^{5h}	37	**41.4**	21.6	0.037	0.036
D_1^{2h}	34.7	**43.2**	22.1	0.040	0.039
D_2^{m}	34.5	**38.1**	27.4	0.047	0.045
D_2^{f}	35.3	**38.3**	26.4	0.048	0.047

Table 3: Preference ratings (bolded is preferred) between baselines (Base) and proposed models (Ours) with architecture (B).

Results (Effect of Conditioning)

Dataset	% Preference			
	Base	W/o conditioning	With conditioning	None
D_1^{10h}	39.8	38.3	-	21.9
	-	37.9	**41.7**	20.4

Table 4: Preference ratings (bolded is preferred) for: 1) baseline (Base) vs proposed model without augmentation conditioning (W/o conditioning); 2) proposed model without augmentation conditioning (W/o conditioning) vs with conditioning (With conditioning).