FLOWTRON

Autoregressive Flow-based Network for Text-to-Mel-spectrogram Synthesis

Rafael Valle, Kevin J. Shih, Ryan Prenger, Bryan Catanzaro (Applied Deep Learning Research ADLR @ NVIDIA)

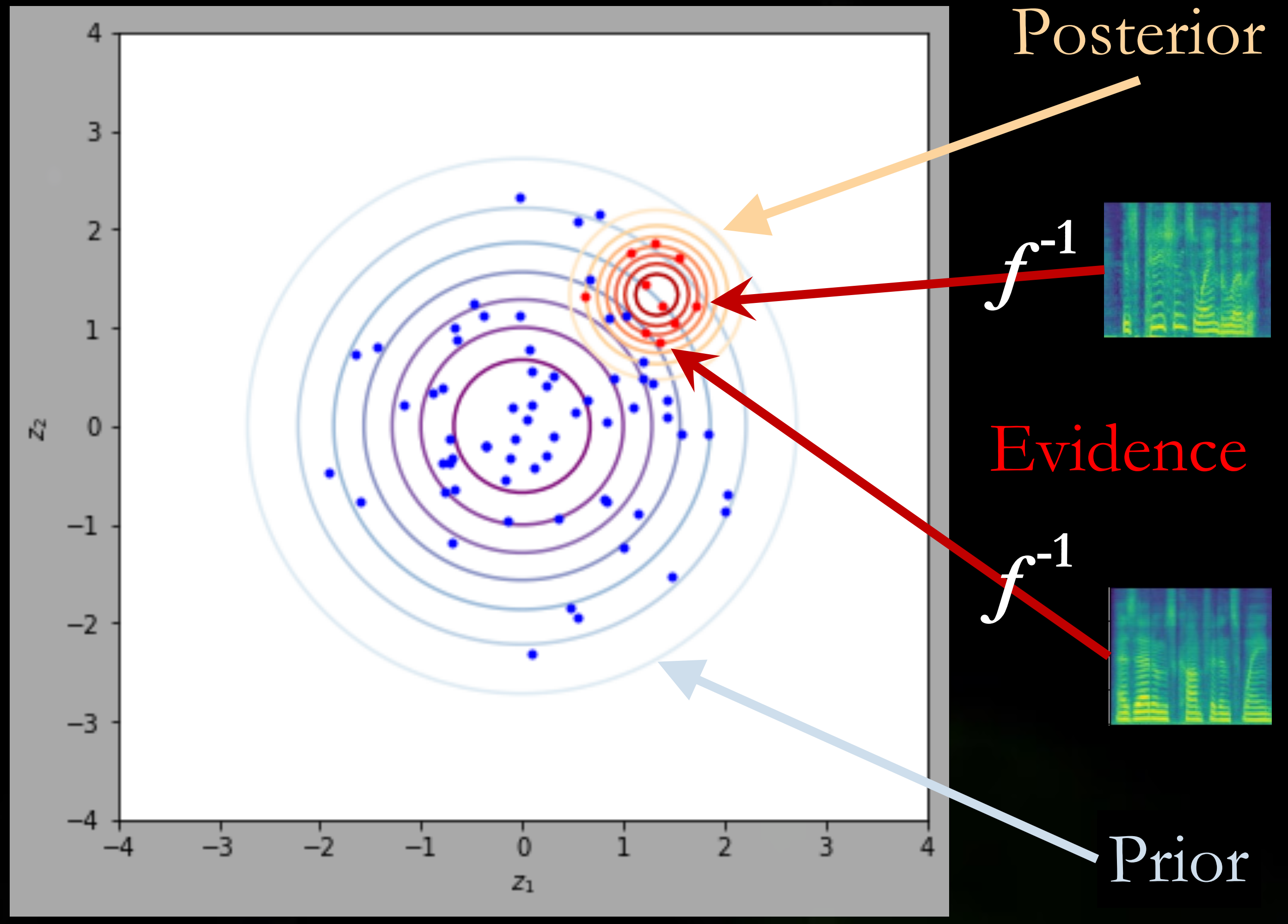
“Taming the non-textual information in speech is difficult because the non-textual is unlabeled.”

“How can we formulate this problem such that we model the non-textual information without using labels ?”

Formulation

Flowtron is a generative model that learns an invertible mapping f^{-1} from speech data (x) to a latent space (z) that can be used to modulate non-textual aspects of speech synthesis.

With this formulation, during inference we can sample from the prior or collect evidence with speech characteristics of interest and sample from the posterior.



Invertible Transformations

$$(\log \boldsymbol{s}_t, \boldsymbol{b}_t) = NN(\boldsymbol{z}_{1:t-1}, \boldsymbol{text})$$

$$\boldsymbol{f}(\boldsymbol{z}_t) = (\boldsymbol{z}_t - \boldsymbol{b}_t) \div \boldsymbol{s}_t$$

$$\boldsymbol{f}^{-1}(\boldsymbol{z}_t) = \boldsymbol{s}_t \odot \boldsymbol{z}_t + \boldsymbol{b}_t$$

Latent Distributions

$$\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I})$$

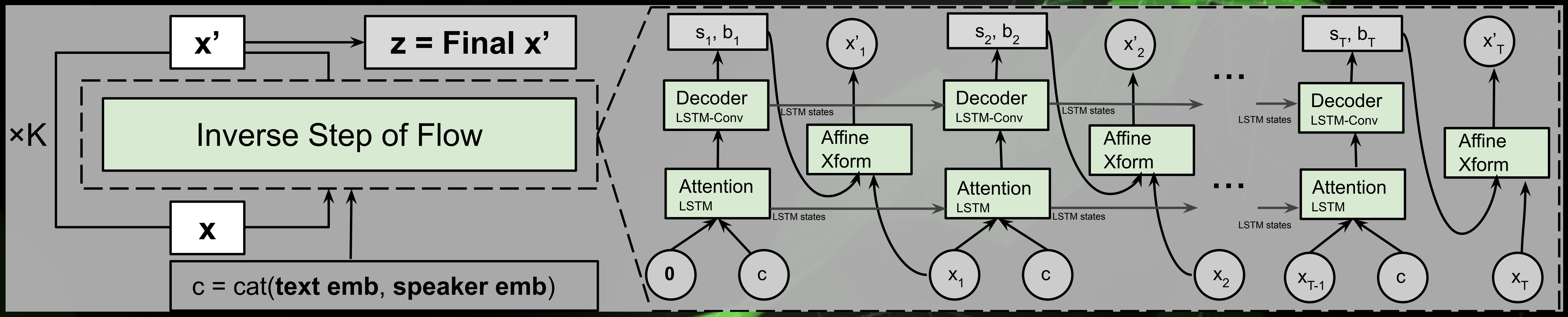
$$\boldsymbol{z} \sim \sum_k \hat{\phi}_k \mathcal{N}(\boldsymbol{z}; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$$

Exact Log-Likelihood Evaluation

$$\log p_\theta(\boldsymbol{x}) = \log p_\theta(\boldsymbol{z}) + \sum_{i=1}^{k} \log|\det(\boldsymbol{J}(\boldsymbol{f}_i^{-1}(\boldsymbol{x})))|$$

Architecture

Flowtron revamps the architecture from Tacotron 2 by removing the Prenet and Postnet layers that were previously thought to be essential to learn attention and to produce sharp harmonics and well resolved formants. (Shen et al., 2017)



Advantages over other models

No hand-engineered loss functions

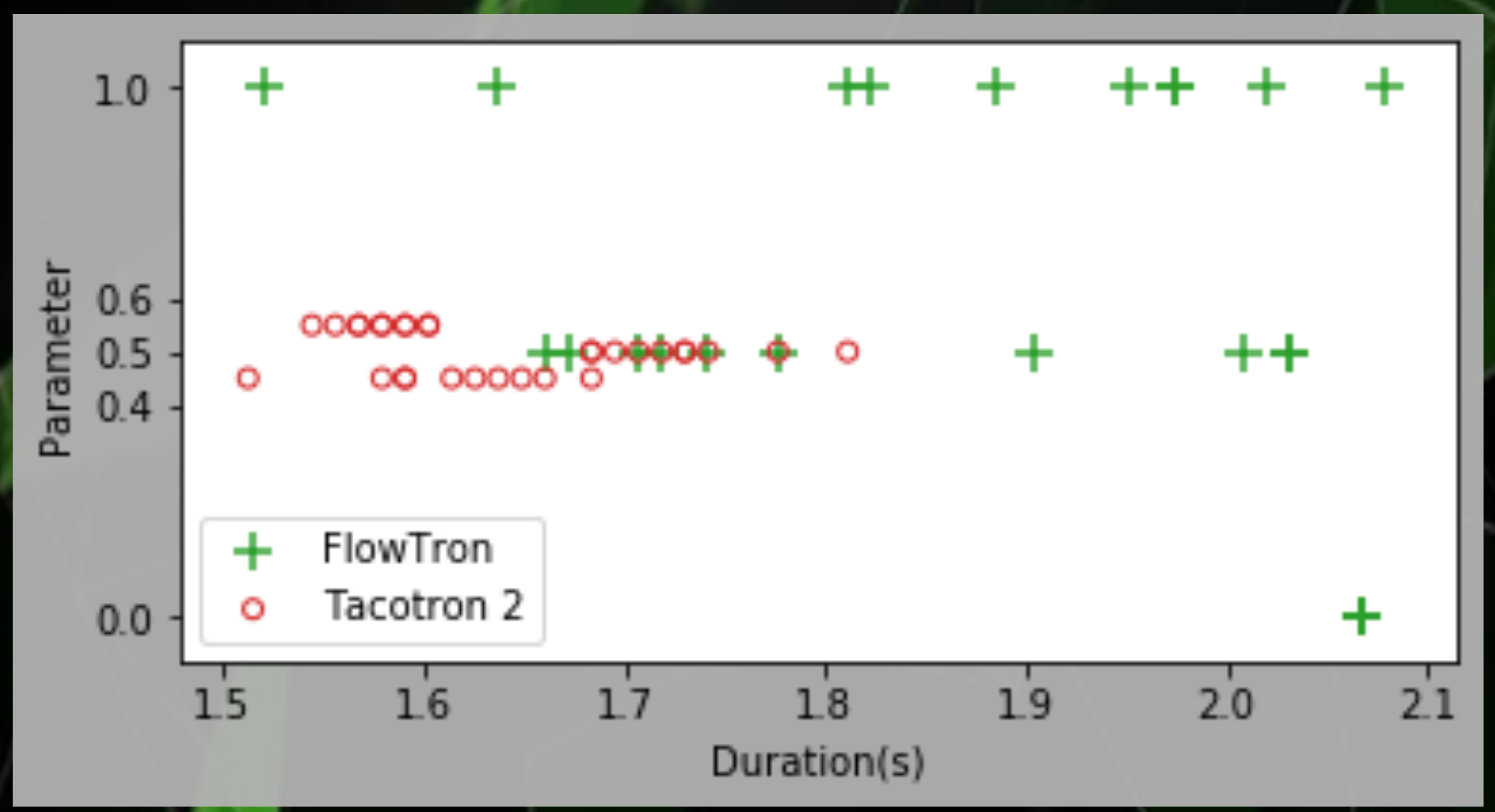
Optimized by maximizing the exact likelihood of the training data

Control over the amount of variety in synthesized speech

Interpolation between prior and posterior over time due to variable length embeddings

Posterior sampling for style transfer of hard to label speech characteristics

Distribution of sentence durations



Mean Opinion Scores (MOS)

Source	Flows	MOS
Real	-	4.27 ± 0.13
Flowtron	2	3.66 ± 0.16
Tacotron 2	-	3.52 ± 0.17