One TTS Alignment to Rule Them All

Rohan Badlani, Adrian Lancucki, Kevin J. Shih, Rafael Valle, Wei Ping, Bryan Catanzaro

{rbadlani, alancucki, kshih, rvalle, wping, bcatanzaro}@nvidia.com

NVIDIA | ICASSP 2022

Problem Statement and Motivation

Goal: Learn Speech-Text alignments online while training TTS models removing external dependencies

Obtaining accurate speech-text alignments is **hard but necessary for training TTS models** which is often obtained **using forced aligners OR training TTS models to obtain alignments.**

Problems with obtaining speech-text alignments:

1. Forced aligners generally have artifacts associated with them and are tied to the alphabet set.

2. Different languages have different alphabets: Its inefficient to learn a forced aligners for each language, alphabet pair.

Mathematical Formulation

To learn the alignments, we optimize the following objective that maximizes the probability of text given mel-spectrograms using the forward-sum algorithm used in Hidden Markov Models (HMMs). We **accelerate** the learning with a **static 2D beta binomial prior** to promote diagonal alignments.

$$\boxed{\Phi \in \mathbb{R}^{C_{txt} \times N}} \qquad \boxed{X \in \mathbb{R}^{C_{mel} \times T}}$$

Encoded Text — Encoded Mels

$$P(S(\Phi) \mid X; \theta) = \sum_{\mathbf{s} \in S(\Phi)} \prod_{t=1}^{T} P(s_t \mid x_t; \theta)$$

Where 's' a specific alignment between mel-spectrograms and text, S(Φ) is the set of all possible valid monotonic alignments; P(s_t|x_t) is the likelihood of a specific text token st=φ_i aligned for mel frame x_t at timestep t.

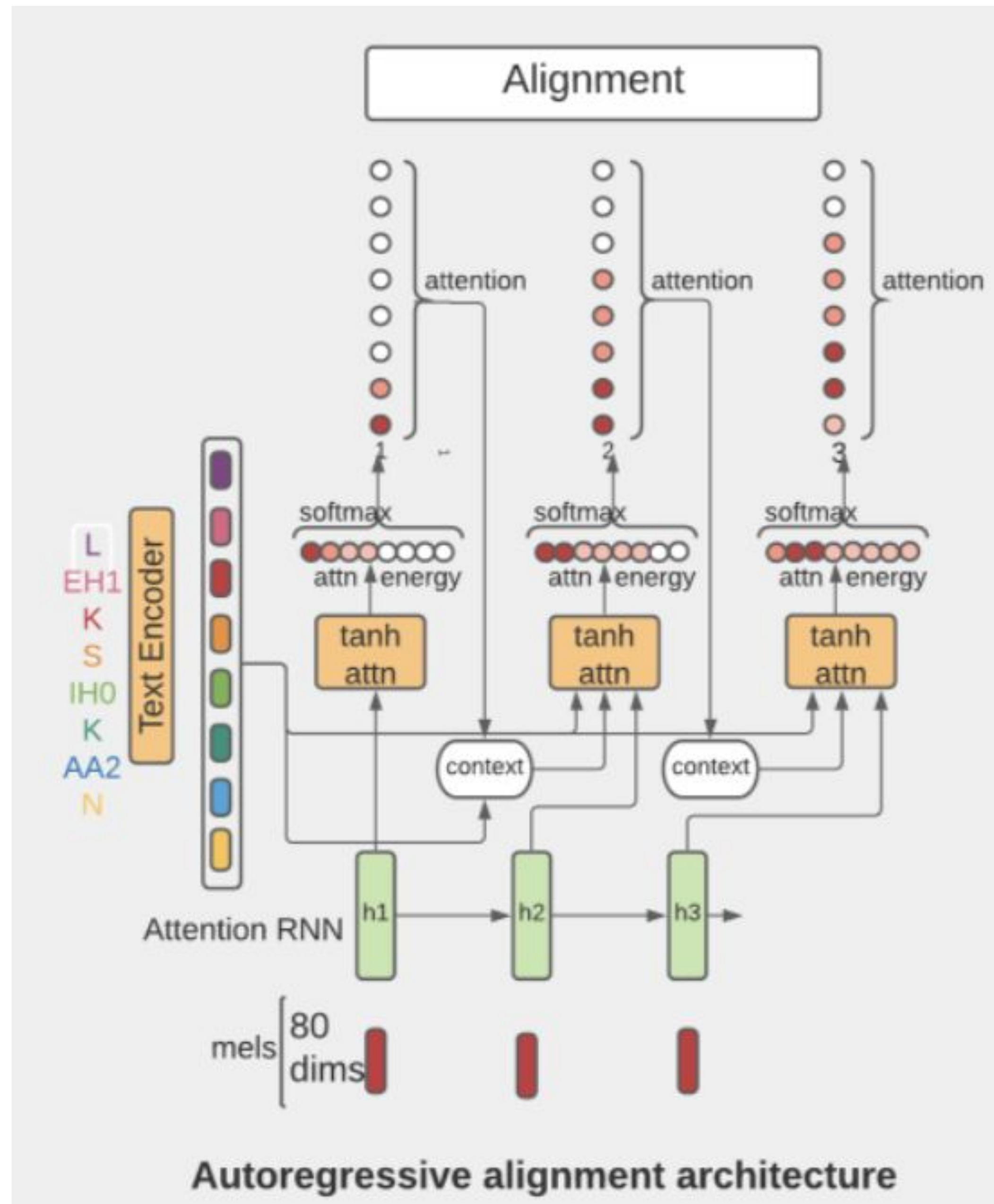
We maximize the above forward sum objective and call LforwardSum as the loss the minimizes the negative log likelihood given by above eq. For autoregressive models, this is the only loss. Since non-autoregressive models take durations as input during test time, we binarize the alignments (Viterbi algorithm) and minimze KL between soft and hard alignments

$$\boxed{\mathcal{L}_{align} = \mathcal{L}_{ForwardSum}.}$$

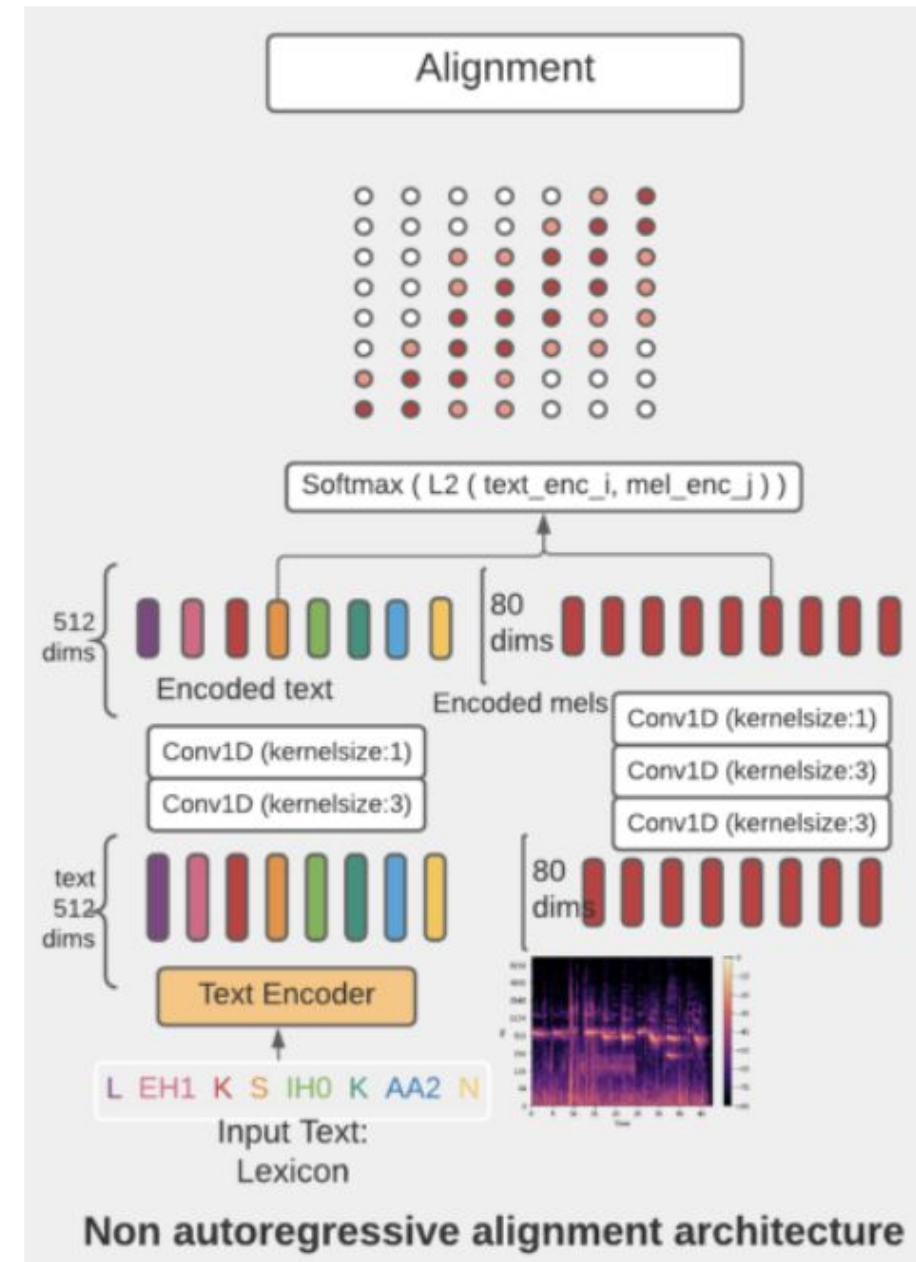
$$\boxed{\begin{aligned}\mathcal{L}_{bin} &= A_{hard} \odot \log A_{soft}, \\ \mathcal{L}_{align} &= \mathcal{L}_{ForwardSum} + \mathcal{L}_{bin}.\end{aligned}}$$

Architecture for learning Speech-Text Alignments

The model architecture diagrams for obtaining soft alignments between text and speech for autoregressive and parallel TTS models



Autoregressive alignment architecture



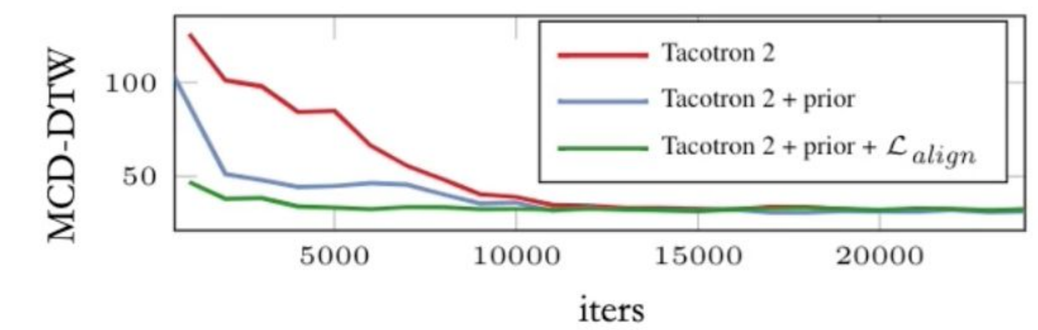
Non autoregressive alignment architecture

Alignment framework consistently improves over all baselines

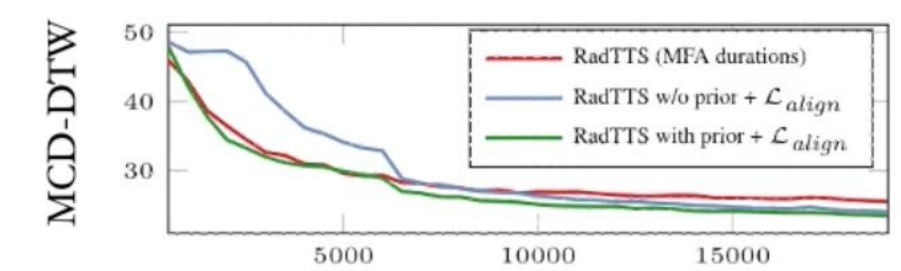
Table 1: *Pairwise preference scores judged by human raters, shown with 95% confidence intervals. Scores above 0.5 indicate models trained with $\mathcal{L}_{align}$ were preferred by majority of raters.*

Model	Alignment Framework vs Baseline
Tacotron 2	0.556 ± 0.068
Flowtron ($\sigma = .5$)	0.635 ± 0.065
RAD-TTS ($\sigma = .5$)	0.639 ± 0.066
FastPitch	0.565 ± 0.068
FastSpeech2	0.521 ± 0.067

Faster Convergence



Convergence rate of autoregressive tacotron2 with and without alignment framework



Convergence rate of non autoregressive RADTTS with and without alignment framework

Better Pronunciation

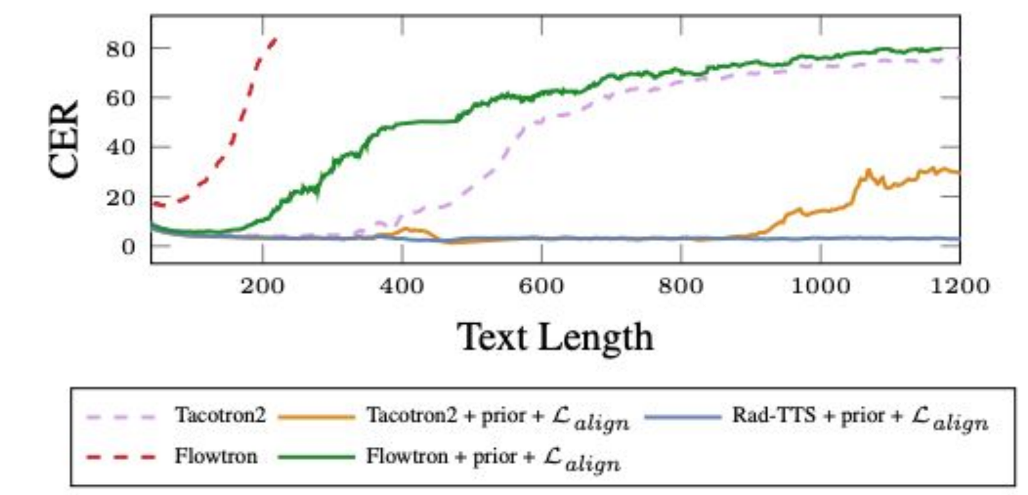


Figure 5: *Character error rate of different models at different text lengths. Models that use the alignment framework make fewer mistakes with increased utterance length.*

Takeaways and Conclusions

- Eliminates the dependency on external aligners by learning speech-text aligners online. This simplifies the training pipeline of TTS models.
- The same alignment learning framework can support multiple languages and alphabets
- Improves pronunciation of several TTS models and leads to faster convergence of TTS models.

Demo, Samples and source Code available at: https://nv-adlr.github.io/one-tts-alignment