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: It is inefficient to learn a forced aligner 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.

\[
\Phi \in \mathbb{R}^{C_{xt} \times N} \\
X \in \mathbb{R}^{C_{mel} \times T}
\]

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

Where \(s\) is a specific alignment between mel-spectrograms and text, \(S(\Phi)\) is the set of all possible valid monotonic alignments; \(P(s, x)\) is the likelihood of a specific text token \(s_{t=1,1}\) aligned for mel frame \(x_{1,1}\) at timestep \(t\).

We maximize the above forward sum objective and call ForwardSum as the loss that 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 minimize KL between soft and hard alignments

\[
L_{\text{align}} = L_{\text{ForwardSum}}
\]

\[
L_{\text{base}} = L_{\text{ForwardSum}} + L_{\text{base}}
\]

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 \(L_{\text{align}}\) were preferred by majority of raters.

<table>
<thead>
<tr>
<th>Model</th>
<th>Alignment Framework vs Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tacotron 2</td>
<td>0.556 ± 0.068</td>
</tr>
<tr>
<td>Flowtron ((\sigma = .5))</td>
<td>0.635 ± 0.065</td>
</tr>
<tr>
<td>RAD-TTS ((\sigma = .5))</td>
<td>0.639 ± 0.066</td>
</tr>
<tr>
<td>FastPitch</td>
<td>0.563 ± 0.068</td>
</tr>
<tr>
<td>FastSpeech2</td>
<td>0.521 ± 0.067</td>
</tr>
</tbody>
</table>

Better Pronunciation

- 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.

Takeaways and Conclusions

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