EAT: Enhanced ASR-TTS Framework for Self-supervised ASR

Murali Karthick Baskar 1 Shinji Watanabe 2 Ramon Astudillo 3 Lukaš Burget 1 Jan Černocký 1

Abstract

Self-supervised training has drawn major benefits for unsupervised speech recognition tasks. In this paper, we propose an enhanced ASR-TTS (EAT) framework using cycle-consistency algorithm by introducing two main features: 1) The ASR→TTS pipeline is included with a language model reward to penalise the ASR hypotheses before forwarding to TTS. 2) In TTS→ASR, a hyper-parameter is introduced to scale the attention context from synthesized speech before sending it to ASR to handle out-of-domain data. We explore the effectiveness of the EAT framework under different data conditions and latest architectures. The results show that EAT reduces the performance gap between supervised and unsupervised to 17.1% and 20.1% on Librispeech and BABEL corpus respectively.

1. Introduction

The primary motivation behind this work are as follows:

To improve the cycle-consistency training method for better seq2seq speech recognition performance

To handle out-of-domain data effectively

Reduce over-fitting for certain class labels while training with supervised data

To train an effective speech synthesis model using semi-supervised cycle-consistency training approach

This paper initially explores latest architectures to fit in ASR and TTS for cycle-consistency training. The performance of the cycle-consistency training is improved to match the stat-of-the-art results in Librispeech. A common character set is defined to handle languages limited to having ASCII characters.

1.1. Constraints in using TTS

Figure ?? shows the log-Mel filterbank spectrum of ground truth speech, ASR→TTS and TTS→ASR. The spectrum generated by TTS in ASR→TTS pipeline is closer to ground truth and the primary reason behind this is that the ground truth speech is available to perform teacher-forcing (using previous ground truth label to predict next target label). This sophistication is not available for TTS in TTS→ASR pipeline and thus the spectrum deviates from the ground truth. Even the speech and silent predictions are wrongly predicted in the TTS→ASR pipeline. This scenario is observed in our experiments when there is a data mismatch between for pre-trained and retrained TTS.

1.1.1. Multilingual Training

This is a transfer-learning approach where the model learned from some source language is used as pre-initialization to train a target model. The pre-trained model is trained using 10 BABEL languages as described in (?). Varying amounts of Pashto data is used to re-train the model and the results are reported in the following table 1. The results show that only retraining decoder with text only data can provide benefit under data mismatched conditions. Further investi-
gation is required to understand the effect of TTS output in TTS→ASR pipeline.

<table>
<thead>
<tr>
<th>Parallel Data (#hrs)</th>
<th>%WER with unpaired data</th>
<th>Baseline</th>
<th>Speech</th>
<th>Text</th>
<th>Speech+text</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>87.5</td>
<td>77.9</td>
<td>71.1</td>
<td>67.6</td>
<td></td>
</tr>
<tr>
<td>+RNNLM</td>
<td>71.4</td>
<td>67.6</td>
<td>61.2</td>
<td>59.8</td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>61.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+RNNLM</td>
<td>52.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>