1. Abstract

Motivations
- Modern end-to-end TTS models (e.g., Tacotron) are data-hungry, they require a sizeable set of high-quality data in the format of <text, audio> pairs for training
- Expensive to collect
- Hard to be applied to low-resource languages

Goal & Contributions
- To improve data efficiency ▶ TTS model requires less paired data for training to produce speech of reasonable quality
- In this work, a semi-supervised training framework is proposed:
  - Leverage textual and acoustic knowledge contained in large, publicly available text and speech corpora ▶ cheap, unpaired, (potentially) noisy
  - Separately pre-train the Tacotron encoder and decoder followed by fine-tuning
  - Tacotron is able to produce intelligible speech with < 30 mins of paired training data ▶ Lower the bar of end-to-end TTS training

3. Experiments & Results

Setup
- Internal single-speaker US English dataset for training (fine-tuning)
- For encoder conditioning:
  - Neural network language model as the word embedding module
  - Pre-trained on English Google News 200B corpus
- For decoder pre-training:
  - VCTK speech portion
  - 44 hours of speech from 109 speakers, most have British accents

Griffin-Lim as the “vocoder” to convert the spectrograms to waveforms
- For faster experimental cycles
- Generating high-fidelity audio is not the main focus of this work

Notations:
- T-Base: baseline Tacotron model
- T-Enc: Tacotron model that only incorporates encoder conditioning
- T-Dec: Tacotron model that only incorporates decoder pre-training
- T-Enc-Dec: Tacotron model that incorporates the above two

Data requirements of the baseline Tacotron
- Goal: To find out maximum amount of data that could almost never successfully train a baseline Tacotron to produce intelligible speech
- Decreased the amount of training data from 40 hours to just 12 minutes
- 10 ~ 40 hours: almost equally good synthesis
- 3 ~ 10 hours: minor degradation, but still sounds good
- 12 ~ 24 minutes (already cover all phonemes): gibberish
- We will demonstrate the effectiveness of the proposed semi-supervised framework using only 24 minutes of paired data

2. Proposed Approach

Encoder Conditioning
- Goal: To exploit textual knowledge contained in large text corpora
- Method: Train word vectors on large text corpora and use them as auxiliary inputs to a TTS model
  1) Embed each word in the input text into a word vector
  2) Add the word vector sequence on either encoder input (the character/phoneme embedding sequence) or encoder top (the encoder output sequence to be consumed by the decoder)
  3) Combining methods: concatenation or attention

Illustration:

Objective Test
- Mel cepstral distortion (MCD): measures the distance between synthesis and ground truth in the mel cepstrum space
- Evaluation set contains 631 unseen sentences (30 mins in total)
- MCD results correlate well with our subjective perception
- Results (the smaller the better, best marked in bold):

<table>
<thead>
<tr>
<th>T-Base</th>
<th>T-Enc</th>
<th>T-Enc-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCD</td>
<td>0.06</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>12.89</td>
<td>12.46</td>
</tr>
<tr>
<td></td>
<td>13.03</td>
<td>12.71</td>
</tr>
<tr>
<td></td>
<td>12.09</td>
<td>12.27</td>
</tr>
</tbody>
</table>

Findings:
- T-Enc, T-Dec, and T-Enc-Dec all achieve much lower MCD than T-Base
- Best T-Enc config. is concatenating the word vectors at the encoder top
- In terms of conditioning location: encoder top ▶ encoder input
- In terms of conditioning method: concatenation ▶ attention
- T-Dec outperforms all the other semi-supervised variants, but not by much: in our SxS test, the raters also did not strongly prefer one over the other two

Subjective Test
- Side-by-side comparisons
- 7-point rating scale
- 1000 unseen phrases of different lengths
- Rater preference results:

<table>
<thead>
<tr>
<th>T-Base</th>
<th>Enc-Dec</th>
<th>T-Enc-Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preference (%) &amp; p-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65.1</td>
<td>71.6</td>
</tr>
<tr>
<td></td>
<td>31.6</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td>16.1</td>
<td>18.2</td>
</tr>
</tbody>
</table>

Results on other amounts of data
- Train (fine-tune) models on varying amounts of paired data
- Figure x- and y-axis are the amount of paired data and the resultant MCD, respectively

Results of subjective and objective tests are consistent:
- Raters strongly preferred T-Enc and T-Dec over T-Base by more than 60%
- All semi-supervised variants (T-Enc, T-Dec, and T-Enc-Dec) were similarly preferable by the raters
- Both demonstrate the effectiveness of our framework

or
https://joo.gl/3ITEvm