IMPROVING MANDARIN END-TO-END SPEECH SYNTHESIS BY SELF-ATTENTION AND LEARNABLE GAUSSIAN BIAS

Fengyu Yang¹, Shan Yang¹, Pengcheng Zhu², Pengju Yan², Lei Xie¹*

¹Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xian, China
²Tongdun AI Lab

ABSTRACT

Compared to conventional speech synthesis, end-to-end speech synthesis has achieved much better naturalness with more simplified system building pipeline. End-to-end framework can generate natural speech directly from characters for English. But for other languages like Chinese, recent studies have indicated that extra engineering features are still needed for model robustness and naturalness, e.g. word boundaries and prosody boundaries, which makes the front-end pipeline as complicated as the traditional approach. To maintain the naturalness of generated speech and discard language-specific expertise as much as possible, in Mandarin TTS, we introduce a novel self-attention based encoder with learnable Gaussian bias in Tacotron. We evaluate different systems with and without complex prosody information and results show that the proposed approach has the ability to generate stable and natural speech with minimum language-dependent front-end modules.

Index Terms— Tacotron, end-to-end, speech synthesis, self-attention, Gaussian bias

1. INTRODUCTION

Traditional speech synthesis conducts text to speech mapping through a complex language-dependent front-end pipeline, which generally contains several hand-engineered modules, at least including part-of-speech tagging, pronunciation prediction and prosody labeling [1][2]. The prediction errors of these modules will inevitably accumulate to the later synthesis model [3]. To overcome this issue and simplify the speech generation process, an attention-based end-to-end framework named Tacotron [4] has recently been proposed to predict speech from raw text. Such model can directly synthesize speech from characters, which does not need the above front-end engineering modules. Considering the advantages of Tacotron, more follow-up end-to-end models are proposed to improve the quality or prosody of the generated speech [5][6][7].

Despite the good performance and simplified pipeline of end-to-end speech synthesis, as the input of the system, the text representation varies from language to language. For English, it has been proved that we can directly use simple character-level or phoneme-level representation to generate natural speech at decent level [4][5][8]. But for Mandarin Chinese, some recent work has found that simple phoneme sequences cannot produce suitable prosody with existing end-to-end systems [9]. Due to the character diversity, distinct tonal and boundary characteristics of Mandarin, traditional speech synthesis methods often inject a lot of prosodic cues like prosodic words (PW), prosodic phrases (PPH) and international phrases (IPH), to handle the proposed issue. These cues have also proved to be helpful for the end-to-end model [9]. But it makes the pipeline more complex again.

In this paper, we aim to achieve natural prosody on an end-to-end speech synthesis system for Mandarin Chinese with minimal extra information. Since self-attention shows its fine ability of global dependency modeling from simple phoneme sequences [5][10][11], we use it as encoder in Tacotron to capture global prosodic information. As for modeling local prosodic information, we introduce a learnable Gaussian bias to self-attention in our system because Gaussian distribution naturally focuses more on local relations of current positions [12][13][14].

In summary, to simplify the building pipeline and towards an end-to-end system, we propose a new structure which uses Self-Attention with learnable Gaussian bias as encoder in Tacotron. Experiments show that the proposed SAG-Tacotron, only using phoneme sequences as input in Chinese, can produce comparable natural and stable speech as the system using prosodic features. Readers can access some audio samples from our demo page ¹.

The rest of the paper is organized as follows. Section 2 introduces our baseline Tacotron structure. Section 3 describes the proposed structure named SAG-Tacotron focused on self-attention with learnable Gaussian bias. Section 4 introduces

* Lei Xie is the corresponding author.

This research work is partially supported by the National Key Research and Development Program of China (No.2017YFB1002102).

¹https://522659913.github.io/AS191245/
the experiments and results. We conclude this paper in Section 5.

2. BASELINE TACOTRON

In this section, we describe our slightly modified baseline Tacotron, as shown in Fig.1. We first convert inputs into character embeddings. Then a CBHG (1-D convolution bank + highway network + bidirectional GRU) module adapted from [16] transforms the pre-net outputs into the final encoder representation subsequently used by the location-sensitive attention module. The decoder is an autoregressive recurrent neural network, where a stack of GRUs with vertical residual connections produces the attention query at each decoder time step. We concatenate the context vector and the attention RNN cell output to form the input to the decoder RNNs. In parallel to the decoder RNN cell output, a 3-layer CNN is applied to the input text sequence, followed by a batch normalization and ReLU activation. Residual connection is applied between the above layers.

After the context vector and decoder RNN cell output are concatenated, the decoder receives the input at each decoder time step. We concatenate the context vector and the attention query at each decoder time step. The position encoding (PE) is used to reconstruct waveforms through Griffin-Lim [17].

3. PROPOSED SAG-TACOTRON

3.1. Motivation

For Mandarin, we aim to use minimal extra text analysis modules to maintain the naturalness of generated speech, and achieve comparable performance with those systems with complex prosodic information. As self-attention has strong ability in global dependency modeling, we replace the baseline CBHG encoder with the self-attention based one to improve the performance of prosodic phrasing. This is inspired by Transformer [11], where self-attention plays a vital role in modeling global dependency. At the same time, to overcome the problem of self-attention’s dispersing the distributions of attention [12][18], we introduce a learnable Gaussian bias to enhance localness modeling. The architecture of this system named SAG-Tacotron can be seen in Fig. 2.

3.2. Self-attention based encoder

In encoder pre-net, a 3-layer CNN is applied to the input text embeddings over input sequence, followed by a batch normalization and ReLU activation. Since the self-attention block does not contain any sequential information, we also inject positional information like that in Transformer:

$$PE_{pos,2i} = \sin (pos/10000^{2i/d})$$

$$PE_{pos,2i+1} = \cos (pos/10000^{2i/d})$$

where pos is the current position, d is the feature dimension and i is the current dimension. The position encoding (PE) is then fed to the self-attention block, which consists of a self-attention layer followed by a fully connected layer with tanh activation. Residual connection is applied between the above layers.

For each $head_i$ in multi-head self-attention [11], Given a sequence $x$ of n elements, we want to get a latent representation $head_i$ with the same length n:

$$Head_i = \sum_{j=1}^{n} ATT(Q, K) V$$

where Q, K and V represent queries, keys and values, respectively, which are three separate weight matrices from the outputs of the encoder pre-net module. Additionally ATT($\cdot$) is the normed weight from the following score function:

$$ATT(Q, K) = \text{softmax}(energy)$$

$$energy = \frac{QK^T}{\sqrt{d}}$$

where $\sqrt{d}$ is the scaling factor with d being the dimensionality of layer states. Finally all the h outputs are concatenated and projected to obtain the final attention values:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, ..., head_h)W^O$$

where $W^O$ is the parameter matrix of the final linear layers to conduct projection.

This block is used as encoder to replace CBHG module in our baseline system.
3.3. Learnable Gaussian bias

In sequence-to-sequence models, closer position is extremely important for Mandarin. In this case, we want to enhance encoder’s local contributions of neighboring states in self-attention. Take Fig. 3 as an example. We first learn a Gaussian bias being centered around the phoneme ”e5”, with a window size being 3 (in practice, it is a learned float number in our model). Then we regularize the distribution of attention with the learned Gaussian bias to produce the final distribution. Obviously the final results pay more attention to the local context around the phoneme ”e5”.

![Fig. 3. Gaussian Bias (D=3)](image)

In details, we place a Gaussian bias $G$ to mask the logit similarity $\textit{energy}$ in Eq. (7) to enhance its localness modeling ability and speech naturalness. Gaussian distribution naturally focuses more on the closer positions.

$$\text{ATT}(Q, K) = \text{softmax}(\text{energy} + G)$$

(7)

$G \in \mathbb{R}^{N \times N}$ belonging to $(-1; 0]$ measures the relation between current query $x_i$ and position $j$:

$$G_{ij} = -\frac{(j - P_i)^2}{2\sigma_i^2}$$

(8)

where $P_i$ is the central position of $x_i$ from the given input sequences $x = (x_1, x_2, \ldots, x_n)$ of $n$ elements, and $\sigma_i$ is the standard deviation. How to choose suitable $P_i$ and $\sigma_i$ is the key problem of the learnable Gaussian bias based self-attention. Since the prediction of each central position depends on its corresponding query vector, we can predict the central position $P_i$ from $x_i$:

$$P_i = N \cdot \text{sigmoid}(\sigma_p^T \text{tanh}(W_p x_i)),$$

(9)

where $N$ is the sequence length to make $P_i$ lie in $(0; N)$, sigmoid activation is applied because its output is in $(0; 1)$ and $W_p$ is the model parameter matrices. Then the final predicted position is $[P_i]$.

1. Baseline: Baseline system described in Sec. 2 using simple inputs. Fig. 4 shows an example of the simple inputs.

2. Baseline-prosody: Baseline system described in Sec. 2 using complex inputs. Fig. 5 shows complex inputs with different levels of prosodic boundaries, which contain phones, tones, character segments, PW tags, PPH tags and IPH tags. The meanings of prosodic tags can be seen in Table 1.

3. SAE-Tacotron: Self-attention as encoder without Gaussian bias described in Sec. 3 using simple inputs as in Fig. 4.

4. SAG-Tacotron: Self-attention as encoder with Gaussian bias described in Sec. 3 using simple inputs as in Fig. 4.

4. EXPERIMENTS

4.1. Basic setups

In the experiments, we use a publicly-available Chinese corpus 2, which contains about 12 hours of speech from a female speaker, recorded in a professional studio. The training and testing sets are composed of 9900 and 100 utterances respectively. For the text side, we convert the sentences into phone sequences. The corpus also provides manually-labelled prosodic boundaries in three levels, i.e., prosodic words (PW), prosodic phrase (PPH) and intonational phrase (IPH), which are used as input features in the baseline systems. In practice, a model has to be trained to predict the three levels of boundaries and prediction errors are unavoidable. In the training and test, we use the ground truth labels to get the performance upper bound.

4.2. System comparison

To investigate the performance of the proposed method with or without the prosodic information, we conduct prosody analysis, Mel cepstral distortion (MCD), error rates and mean opinion score (MOS) to evaluate the performance of different systems. Different systems built are described as follows.

1. Baseline: Baseline system described in Sec. 2 using simple inputs. Fig. 4 shows an example of the simple inputs.

2. Baseline-prosody: Baseline system described in Sec. 2 using complex inputs. Fig. 5 shows complex inputs with different levels of prosodic boundaries, which contain phones, tones, character segments, PW tags, PPH tags and IPH tags. The meanings of prosodic tags can be seen in Table 1.

3. SAE-Tacotron: Self-attention as encoder without Gaussian bias described in Sec. 3 using simple inputs as in Fig. 4.

4. SAG-Tacotron: Self-attention as encoder with Gaussian bias described in Sec. 3 using simple inputs as in Fig. 4.

5. Transformer: Transformer in [5], using simple inputs as in Fig. 4.

2http://www.data-baker.com/hc_znv_1.html
Table 1. Prosodic tags

<table>
<thead>
<tr>
<th>Tag</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>^</td>
<td>beginning silence</td>
</tr>
<tr>
<td>$</td>
<td>ending silence</td>
</tr>
<tr>
<td>#0</td>
<td>character segment</td>
</tr>
<tr>
<td>#1</td>
<td>prosodic word (PW)</td>
</tr>
<tr>
<td>#2</td>
<td>prosodic phrase (PPH)</td>
</tr>
<tr>
<td>#3</td>
<td>intonational phrase (IPH)</td>
</tr>
</tbody>
</table>

4.3. Model details

For the encoder part in SAE-Tacotron, it contains 3 feed-forward layers with ReLu activation as pre-net and dropout regularization as described in Tacotron [4]. The pre-net is followed by 6 self-attention encoder blocks, where 8-head multi-head attention is applied to learn the latent representation for each self-attention layer. And the position-wise feed-forward sub-module in each self-attention block consists of two linear transformations with 2048 and 512 hidden units. We conduct ReLU activation in the output of the first linear transformation, of which the hidden size for query, key, and value is 512. Then, we apply dropout in the self-attention weights, sub-layer outputs and the ReLU activation with probability 0.1.

The experimental settings for Transformer totally follow [5]. In encoder and decoder part, the hyper-parameters of self-attention and feed-forward modules are the same with the above corresponding encoder module in SAE-Tacotron. For the attention between encoder and decoder, we also use 8-head multi-head attention with the attention space dimension of 128.

4.4. Results

4.4.1. Robustness test

From previous studies, in the autoregressive model, wrong attention alignment may appear at the encoder-decoder attention mechanism, leading to repeating and skipping problems. To evaluate the robustness of different systems, generating errors on the test set are listed in Table 2. It can be seen clearly that Transformer TTS is not robust to these common cases in Mandarin as getting 12% word error rate. The baseline Tacotron systems may exist skipping problem occasionally whether they use simple or complex inputs or not. while SAE-Tacotron and SAG-Tacotron system can solve this problem and are more stable no matter we introduce self-attention with or without Gaussian bias to baseline Tacotron. SAG-Tacotron and SAE-Tacotron can effectively eliminate word repeating and skipping problems to improve intelligibility. As Transformer has bad robustness with current configuration, we do not consider it in the following detailed investigation.

Table 2. Robustness evaluation of different systems

<table>
<thead>
<tr>
<th>Method</th>
<th>Repeats</th>
<th>Skips</th>
<th>Error Sentences</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>Baseline-prosody</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1%</td>
</tr>
<tr>
<td>SAE-Tacotron</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>SAG-Tacotron</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Transformer</td>
<td>2</td>
<td>11</td>
<td>12</td>
<td>12%</td>
</tr>
</tbody>
</table>

4.4.2. Prosody Analyses

We plot spectrogram and F0 of a test sentence generated by baseline, baseline-prosody, SAE-Tacotron and SAG-Tacotron, as shown in Fig. 6. The corresponding transcript of the generated speech is shown Fig. 4. We marked two positions of high F0 to show the distinct pitch in this sentence. From the ground-truth waveform, we found that the distinct pitch appear near "de5" and "san1" which refers to 1:5 in demo website 3. Listening to the sample, we can easily capture the high pitch near "de5" and "san1" in SAE-Tacotron, SAG-Tacotron and ground-truth waveform. But in baseline and baseline-prosody, we may capture high pitch in different positions. And observing the value of F0 marked, SAG-Tacotron’s value is the closest to that of the ground-truth waveform. As for the whole trajectory pattern of F0, it is obvious that the SAG-Tacotron’s pattern is most similar to that of the ground-truth.

4.4.3. Objective test

We used MCD to measure the quality of the learned spectrum on the testing set. Lower MCD indicates better spectral quality. Dynamic time warping (DTW) is applied to perform the alignment since the frames of the predicted Mel spectrums may not match those from the ground truth recording. The objective results are shown in Table 3.

Table 3. MCD evaluation of different systems

<table>
<thead>
<tr>
<th>Method</th>
<th>MCD(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>6.004</td>
</tr>
<tr>
<td>Baseline-prosody</td>
<td>5.814</td>
</tr>
<tr>
<td>SAE-Tacotron</td>
<td>5.861</td>
</tr>
<tr>
<td>SAG-Tacotron</td>
<td>5.775</td>
</tr>
</tbody>
</table>

The results show that prosodic information is important for an end-to-end Mandarin system to achieve a lower MCD. With prosody boundary information, MCD is reduced by 0.19. When we do not use prosodic information and just change the encoder module with self-attention – the

3https://522659913.github.io/AS191245/
SAE-Tacotron system, we can acquire an MCD close to that of Baseline-prosody. Most importantly, the SAG-Tacotron system achieves the lowest MCD score, which means that self-attention and Gaussian bias both contribute to the mel-spectrum’s similarity to the natural speech.

4.4.4. Subjective test

We also conducted subjective tests to evaluate the quality of the synthesized speech. There are 20 Chinese listeners with normal hearing taking part in the evaluation, where 30 randomly selected utterances are provided in each testing session. The participants were asked to rate the overall impression on the naturalness of the testing samples. The MOS results are shown in Table 4.

We can easily see that the gains between the baseline and our proposed method SAG-Tacotron is 0.11 while the gap between our proposed method SAG-Tacotron and the system with rich prosodic inputs is 0.02.

From the MOS results, we can also found the importance of prosodic information in the baseline system, which directly affects the prosody of generated speech according to the listeners’ feedback. When replacing CBHG encoder with the self-attention module, we found the SAE-tacotron system outperforms the baseline with simple text representation. Besides, injecting learnable Gaussian bias to self-attention in SAE-Tacotron can further improve the model performance. The results indicate that even with simple text inputs, the SAG-Tacotron can reach a comparable level with baseline-prosody in the subjective test. In this case, we can avoid the use of a large number of prosodic data to train a prosodic model in end-to-end speech synthesis.

5. CONCLUSION

In this paper, we propose a novel self-attention based encoder with learnable Gaussian bias to generate stable and natural speech with minimum language-dependent modules. Self-attention is used to acquire global dependency and learnable Gaussian bias is used to acquire local dependency. Experimental results on Mandarin speech synthesis indicate that the proposed SAG-Tacotron has strong ability to simplify the building pipeline and towards end-to-end with natural and stable generated speech.

6. REFERENCES


