FPETS: Fully Parallel End-to-End Text-to-Speech System

Dabiao Ma,*1 Zhiba Su,*1 Wenxuan Wang,*2 Yuhao Lu†1
1Turing Robot Co.,Ltd. Beijing, China
{madabiao, suzhiba, luyuhao}@uzoo.cn
2The Chinese University of Hong Kong, Shenzhen. Guangdong, China
wenxuanwang1@link.cuhk.edu.cn

Abstract
End-to-end Text-to-speech (TTS) system can greatly improve the quality of synthesised speech. But it usually suffers form high time latency due to its auto-regressive structure. And the synthesised speech may also suffer from some error modes, e.g. repeated words, mispronunciations, and skipped words.
In this paper, we propose a novel non-autoregressive, fully parallel end-to-end TTS system (FPETS). It utilizes a new alignment model and the recently proposed U-shape convolutional structure, UFANS. Different from RNN, UFANS can capture long term information in a fully parallel manner. Trainable position encoding and two-step training strategy are used for learning better alignments. Experimental results show FPETS utilizes the power of parallel computation and reaches a significant speed up of inference compared with state-of-the-art end-to-end TTS systems. More specifically, FPETS is 600X faster than Tacotron2, 50X faster than DCTTS and 10X faster than Deep Voice3. And FPETS can generates audios with equal or better quality and fewer errors comparing with other system. As far as we know, FPETS is the first end-to-end TTS system which is fully parallel.

Introduction
TTS systems aim to generate human-like speeches from texts. End-to-end TTS system is a type of system that can be trained on (text,audio) pairs without phoneme duration annotation(Wang et al. 2017). It usually contains 2 components, an acoustic model and a vocoder. Acoustic model predicts acoustic intermediate features from texts. And vocoder, e.g. Griffin-Lim (Griffin et al. 1984), WORLD (MORISE, YOKOMORI, and OZAWA 2016), WaveNet (van den Oord et al. 2016b), synthesizes speeches with generated acoustic features.

The advantages of end-to-end TTS system are threefold: 1) reducing manual annotation cost and being able to utilize raw data, 2) preventing the error propagation between different components, 3) reducing the need of feature engineering. However, without the annotation of duration information, end-to-end TTS systems have to learn the alignment between text and audio frame.

Most competitive end-to-end TTS systems have an encoder-decoder structure with attention mechanism, which is significantly helpful for alignment learning. Tacotron (Wang et al. 2017) uses an autoregressive attention (Bahdanau, Cho, and Bengio 2014) structure to predict alignment, and uses CNNs and GRU (Cho et al. 2014) as encoder and decoder, respectively. Tacotron2(Shen et al. 2018), which is a combination of the modified Tacotron system and WaveNet, also use an autoregressive attention. However, the autoregressive structure greatly limits the inference speed in the context of parallel computation. Deep voice 3 (Ping et al. 2018) replaces RNNs with CNNs to speed up training and inference. DCTTS (Tachibana, Uenoyama, and Aihara 2017) greatly speeds up the training of attention module by introducing guided attention. But Deep Voice 3 and DCTTS still have autoregressive structure. And those models also suffer from serious error modes e.g. repeated words, mispronunciations, or skipped words (Ping et al. 2018).

Low time latency is required in real world application. Autoregressive structures, however, greatly limit the inference speed in the context of parallel computation. (Ping et al. 2018) claims that it is hard to learn alignment without a autoregressive structure. So the question is how to design a non-autoregressive structure that can perfectly determine alignment?

In this paper, we propose a novel fully parallel end-to-end TTS system (FPETS). Given input phonemes, our model can predict all acoustic frames simultaneously rather than autoregressively. Specifically, we follow the commonly used encoder-decoder structure with attention mechanism for alignment. But we replace autoregressive structures with a recent proposed U-shaped convolutional structure (UFANS)(Ma et al. 2018), which can be fully parallel and has stronger representation ability. Our fully parallel alignment structure inference alignment relationship between all phonemes and audio frames at once. Our novel trainable position encoding method can utilize position information better and two-step training strategy improves the alignment
Experimental results show FPETS utilizes the power of parallel computation and reaches a significant speed up of inference compared with state-of-the-art end-to-end TTS systems. More specifically, FPETS is 600X faster than Tacotron2, 50X faster than DCTTS and 10X faster than Deep Voice3. And FPETS can generates audios with equal or better quality and fewer errors comparing with other system. As far as we know, FPETS is the first end-to-end TTS system which is fully parallel.

Model Architecture

Most competitive end-to-end TTS systems have an encoder-decoder structure with attention mechanism (Wang et al. 2017) (Ping et al. 2018). Following this overall architecture, our model consists of three parts, shown in Fig.1. The encoder converts phonemes into hidden states that are sent to decoder; the alignment module determines the alignment width of each phoneme, from which the number of frames that attend on that phoneme can be induced; The decoder receives alignment information and converts the encoder hidden states into acoustic features.

### Encoder

The encoder encodes phonemes into hidden states. It consists of 1 embedding layer, 1 dense layer, 3 convolutional layers, and a final dense layer. Some of TTS systems (Li et al. 2018) use self attention network as encoder. But we find that it does not make significant difference, both in loss value and MOS.

Alignment module determines the mapping from phonemes to acoustic features. We discard autoregressive structure, which is widely used in other alignment modules (Ping et al. 2018) (Wang et al. 2017) (Shen et al. 2018), for time latency issue. Our novel alignment module consists of 1 embedding layer, 1 UFANS (Ma et al. 2018) structure, trainable position encoding and several matrix multiplications, as depicted in Fig.1.

### Fully parallel UFANS structure

UFANS is a modified version of U-Net for TTS task aiming to speed up inference. The structure is shown in Fig.2. In alignment structure, UFANS is used to predict alignment width, which is similar to phoneme duration. Those pooling and up-sampling operation along the spatial dimension make the receptive field increases exponentially and high way connection enables the combination of different scales features.

For each phoneme $i$, we define the ‘Alignment Width’ $r_i$ which represents its relationship with frame numbers. Suppose the number of phonemes in an utterance is $N$, and UFANS outputs a sequence of $N$ scalars: $[r_0, r_1, ..., r_{N-1}]$.

Then we relate the alignment width $r_i$ to the acoustic frame index $j$. The intuition is that the acoustic frame with index $j = \sum_{k=0}^{i-1} r_k + \frac{1}{2} r_i$ should be the one that attends most on $i$-th phoneme. And we need a structure that satisfies the intuition.

### Position Encoding Function

Position encoding (Gehring et al. 2017) is a method to embed a sequence of absolute positions into a sequence of vectors. Sine and cosine position encoding has two very important properties that make it suitable for position encoding. In brief, function $g(x) = \sum_f \cos(\frac{x - s_f}{s_f})$ has a heavy tail that enables one acoustic frame to receive phoneme information very far away; The gradient function $|\dot{g}(s)| = |\sum_f \sin(\frac{x - s_f}{s_f})|$ is insensitive to the term $x - s$. We give a more detailed illustration in Appendix.

### Trainable Position Encoding

Some end-to-end TTS system, like deep voice 3 and Tacotron2, use sine and cosine functions of different frequencies and add those position encoding vectors to input embedding. But they both take position encoding as a supplement to help the training of attention module and the position encoding vectors remain constant. We propose a trainable position encoding, which is better than absolute position encoding in getting position information.

We define the absolute alignment position $s_i$ of $i$-th phoneme as:

$$s_i = \sum_{k=0}^{i-1} r_k + \frac{1}{2} r_i, i = 0, ..., T_p - 1, r_{-1} = 0 \quad (1)$$
Now choose $L$ float numbers log uniformly from range $[1.0, 10000.0]$ and get a sequence of frequencies $[f_0, \ldots, f_{L-1}]$. For $i$-th phoneme, the position encoding vector $vp_i$ of this phoneme is defined as:

$$vp_i = [vp_{i,sin}, vp_{i,cos}],$$

$$[vp_{i,sin}]_k = \sin\left(\frac{s_i}{f_k}\right),$$

$$[vp_{i,cos}]_k = \cos\left(\frac{s_i}{f_k}\right), \quad k = 0, \ldots, L - 1$$

(2)

Concatenating $vp_i, i = 0, \ldots, T_p - 1$ together, we get a matrix $P$ that represents position information of all the phonemes, denoted as ‘Key’, see Fig.1:

$$P = [vp_0^T, \ldots, vp_{T_p-1}^T]$$

(3)

And similarly, for the $j$-th frame of the acoustic feature, the position encoding vector $va_j$ is defined as:

$$va_j = [va_{j,sin}, va_{j,cos}],$$

$$[va_{j,sin}]_k = \sin\left(\frac{j}{f_k}\right),$$

$$[va_{j,cos}]_k = \cos\left(\frac{j}{f_k}\right), \quad k = 0, \ldots, L - 1$$

(4)

Concatenating all the vectors, we get the matrix $F$ that represents position information of all the acoustic frames, denoted as ‘Query’, see Fig.1:

$$F = [va_0^T, \ldots, va_{T_a-1}^T]$$

(5)

And now define the attention matrix $A$ as:

$$A = FP^T, \quad A_{ji} = vp_{i}^T va_{j}^T,$$

$$i = 0, \ldots, T_p - 1, \quad j = 0, \ldots, T_a - 1$$

(6)

That is, the attention of $j$-th frame on $i$-th phoneme is proportional to the inner product of their encoding vectors. This inner product can be rewritten as:

$$vp_{i}^T va_{j} = \sum_j (\cos\left(\frac{s_i}{f}\right)\cos\left(\frac{j}{f}\right) + \sin\left(\frac{s_i}{f}\right)\sin\left(\frac{j}{f}\right))$$

$$= \sum_j \cos\left(\frac{s_i - j}{f}\right)$$

(7)

It is clear when $j = s_i$, the $j$-th frame is the one that attends most on $i$-th phoneme. The normalized attention matrix $\hat{A}$ is:

$$\hat{A}, \hat{A}_{ji} = \frac{A_{ji}}{\sum_i A_{ji}}$$

(8)

Now $\hat{A}_{ji}$ represents how much $j$-th frame attends on $i$-th phoneme.

Then we use argmax to build new attention matrix $\overline{A}$:

$$\overline{A}_{ji} = \begin{cases} 1 & \text{if } i = \text{argmax}_{k \in [0, \ldots, T_p-1]} A_{jk} \\ 0 & \text{otherwise} \end{cases}$$

(9)

Now define the number of frames that attend more on $i$-th phoneme than any other phoneme to be its attention width $w_i$. From the definition of attention width, $\overline{A}$ is actually a matrix representing attention width $w_i$. The alignment width $r_i$ and $w_i$ are different but related.

For two adjacent absolute alignment positions $s_i$ and $s_{i+1}$, consider the two functions:

$$g_1(x) = \sum_j \cos\left(\frac{x - s_i}{f}\right), \quad g_2(x) = \sum_j \cos\left(\frac{x - s_{i+1}}{f}\right)$$

The values of the two functions only depend on the relative position of $x$ to $s_i$ and $s_{i+1}$. It is known function $g_1$ decreases when $x$ moves away from $s_i$ (locally, but it is sufficient here). So we have:
Training Strategy

We use Acoustic Loss, denoted as $\text{LOSS}_{\text{acou}}$, to evaluate the quality of generated acoustic features, which is $L_2$ norm between predicted acoustic features and ground truth features.

In order to train a better alignment model, we propose a two-stage training strategy. Our model focus more on alignment learning in stage 1. In stage 2 we fix the alignment module and train the whole system.

Stage 1: Alignment Learning In order to enhance the quality of alignment learning, we use convolutional decoder and design an alignment loss.

Convolutional Decoder: UFANS has stronger representation ability than vanilla CNN. But the learning of alignment will be greatly disturbed if using UFANS as decoder. The experimental evidences and analysis are shown in the next section.

So we replace UFANS decoder with a convolutional decoder. The convolutional decoder consists of several convolution layers with gated activation (van den Oord et al. 2016a), several Dropout (Srivastava et al. 2014) operations and one dense layer.

Alignment Loss: We define an Alignment Loss, denoted as $\text{LOSS}_{\text{align}}$, based on the fact that the summation of alignment width should be equal or close to the frame length of acoustic features. We relax this restriction by using a threshold $\gamma$:

$$\text{LOSS}_{\text{align}} = \begin{cases} \gamma, & \text{if } \sum_{k=0}^{T_p-1} r_k - T_a < \gamma \\ \sum_{k=0}^{T_p-1} r_k - T_a, & \text{otherwise} \end{cases}$$

The final loss $\text{LOSS}$ is a weighted sum of $\text{LOSS}_{\text{acou}}$ and $\text{LOSS}_{\text{align}}$:

$$\text{LOSS} = \text{LOSS}_{\text{acou}} + \sigma \text{LOSS}_{\text{align}}$$

We choose 0.02 as alignment loss weight based on grid search from 0.005 to 0.3.

Stage 2: Overall Training In stage 2, we fix the well-trained alignment module and use UFANS as decoder to train the overall end-to-end system. Only acoustic Loss is used as objective function in this stage.

Experiments and Results

Dataset

LJ speech (Ito 2017) is a public speech dataset consisting of 13100 pairs of text and 22050 HZ audio clips. The clips vary from 1 to 10 seconds and the total length is about 24 hours. Phoneme-based textual features are given. Two kinds of acoustic features are extracted. One is based on WORLD vocoder that uses mel-frequency cepstral coefficients (MFCCs). The other is linear-scale log magnitude spectrograms and mel-band spectrograms that can be feed into Griffin-Lim algorithm or a trained WaveNet vocoder.

The WORLD vocoder uses 60 dimensional mel-frequency cepstral coefficients, 2 dimensional band aperiodicity, 1 dimensional logarithmic fundamental frequency,
Table 1: Hyper-Parameter

<table>
<thead>
<tr>
<th>Structure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder/DNN Layers</td>
<td>1</td>
</tr>
<tr>
<td>Encoder/CNN Layers</td>
<td>3</td>
</tr>
<tr>
<td>Encoder/CNN Kernel</td>
<td>3</td>
</tr>
<tr>
<td>Encoder/CNN Filter Size</td>
<td>1024</td>
</tr>
<tr>
<td>Encoder/Final DNN Layers</td>
<td>1</td>
</tr>
<tr>
<td>Alignment/UFANS layers</td>
<td>4</td>
</tr>
<tr>
<td>Alignment/UFANS hidden</td>
<td>512</td>
</tr>
<tr>
<td>Alignment/UFANS Kernel</td>
<td>3</td>
</tr>
<tr>
<td>Alignment/UFANS Filter Size</td>
<td>1024</td>
</tr>
<tr>
<td>CNN Decoder/CNN Layers</td>
<td>3</td>
</tr>
<tr>
<td>CNN Decoder/CNN Kernel</td>
<td>3</td>
</tr>
<tr>
<td>CNN Decoder/CNN Filter Size</td>
<td>1024</td>
</tr>
<tr>
<td>UFANS Decoder/UFANS layers</td>
<td>6</td>
</tr>
<tr>
<td>UFANS Decoder/UFANS hidden</td>
<td>512</td>
</tr>
<tr>
<td>UFANS Decoder/UFANS Kernel</td>
<td>3</td>
</tr>
<tr>
<td>UFANS Decoder/UFANS Filter Size</td>
<td>1024</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.15</td>
</tr>
</tbody>
</table>

Implementation Details

Hyperparameters of our model are shown in Table 1. Tacotron2, DCTTS and Deep Voice3 are used as baseline. The model configurations are shown in Appendix. Adam is used as optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e-4$. Each model is trained 300k steps.

All the experiments are done on 4 GTX 1080Ti GPUs, with batch size of 32 sentences on each GPU.

Main Results

We aim to design a TTS system that can synthesize speech quickly, high quality and with fewer errors. So we compare our FPETS with baseline on inference speed, MOS and error modes.

Inference Speed

The inference speed evaluates time latency of synthesizing a one-second speech, which includes data transfer from main memory to GPU global memory, GPU calculations and data transfer back to main memory. As is shown in Table 2, our FPETS model is able to greatly take advantage of parallel computations and is significantly faster than other systems.

MOS

Harvard Sentences List 1 and List 2 are used to evaluate the mean opinion score (MOS) of a system. The synthesized audios are evaluated on Amazon Mechanical Turk using crowdMOS method (Protasio Ribeiro et al. 2011). The score ranges from 1 (Bad) to 5 (Excellent). As is shown in Table 3, our FPETS is no worse than other end-to-end system. The MOS of WaveNet-based audios are lower than expected since background noise exists in these audios.

Robustness Analysis

Attention-based neural TTS systems may run into several error modes that can reduce synthesis quality. For example, repetition means repeated pronunciation of one or more phonemes, mispronunciation means wrong pronunciation of one or more phonemes and skip word means one or more phonemes are skipped.

In order to track the occurrence of attention errors, 100 sentences are randomly selected from Los Angeles Times, Washington Post and some fairy tales. As is shown in Table 4, Our FPETS system is more robust than other systems.

Alignment Learning Analysis

Alignment learning is essential for end-to-end TTS system which greatly affects the quality of generated audios. So we further discuss the factors that can affect the alignment quality.

100 audios are randomly selected from training data, denoted as origin audios. Their utterances are fed to our system to generate audios, denoted as re-synthesized audios. The method to evaluate the alignment quality is objectively computing the difference of the phoneme duration between origin audios and their corresponding re-synthesized audios. The phoneme durations are obtained by hand. Figure 4 is the labeled phonemes of audio 'LJ048-0033'. Here only results with mel-band spectrograms using Griffin-Lim algorithm are shown. For MFCCs, results are similar.

We compare alignment quality between different alignment model configurations. Table 6 shows the overall results on 100 audios. Table 5 is a case study which shows how phoneme-level duration is affected by different model.

Position Encoding Function and Alignment Quality

We replace the Sine and Cosine position encoding function with Gaussian function. As Table 6 shows, the experimental results show that the model can not learn correct alignment...
Table 4: Robustness Comparison

<table>
<thead>
<tr>
<th></th>
<th>Repeats</th>
<th>Mispronunciation</th>
<th>Skip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tacotron2</td>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>DCTTS</td>
<td>2</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Deep Voice 3</td>
<td>1</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>FPETS</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 4: The upper is real audio of ‘LJ048-0033’, the lower is the re-synthesized audio from alignment learning model.

Figure 5: Attention plot of text: This is the destination for all things related to development at stack overflow.

Related Works

FastSpeech (Ren et al. 2019), which is proposed in the same period, can also generate acoustic features in a parallel way. Specifically, it extracts attention alignments from an auto-regressive encoder-decoder based teacher model for phoneme duration prediction. It is used by a length regulator to expand the source phoneme sequence to match the length of the target mel-spectrogram sequence for parallel mel-spectrogram generation. Using the phoneme duration extracted from a teacher model is a creative work to solve the problem that model can’t inference in a parallel way. However, it’s speed is still not fast enough to satisfy industrial application, especially it can’t speed up when batch size is increased.

Our FPETS has lower time latency and faster than FastSpeech. On average FPETS generates 10ms per sentence under GTX 1080ti GPU and FastSpeech is 170ms per sentence under Tesla V100 GPU, which is known faster than GTX 1080ti. And FPETS can also automatically specify the phoneme duration by trainable position encoding.

Conclusion

In this paper, a new non-autoregressive, fully parallel end-to-end TTS system, FPETS, is proposed. Given input phonemes, FPETS can predict all acoustic frames simultaneously rather than autoregressively. Specifically FPETS utilize a recent proposed U-shaped convolutional structure, which can be fully parallel and has stronger representation ability. The fully parallel alignment structure infers the alignment relationship between all phonemes and audio frames at once. The novel trainable position encoding method can utilize position information better and two-step training strategy improves the alignment quality.

FPETS can utilize the power of parallel computation and reach a significant speed up of inference compared with state-of-the-art end-to-end TTS systems. More specifically, FPETS is 600X faster than Tacotron2, 50X faster than DCTTS and 10X faster than Deep Voice3. And FPETS can generates audios with equal or better quality and fewer errors comparing with other system. As far as we know, FPETS is the first end-to-end TTS system which is fully parallel.
Table 5: A case study about phoneme-level comparison of alignment quality. Real duration and the predicted duration by our alignment method, using Gaussian as position encoding function, using fixed position encoding, using UFANS as decoder are shown.

<table>
<thead>
<tr>
<th></th>
<th>P</th>
<th>R</th>
<th>AY</th>
<th>ER</th>
<th>T</th>
<th>UW</th>
<th>N</th>
<th>OW</th>
<th>V</th>
<th>EH</th>
<th>M</th>
<th>B</th>
<th>ER</th>
<th>T</th>
<th>W</th>
<th>EH</th>
</tr>
</thead>
<tbody>
<tr>
<td>real</td>
<td>5.35</td>
<td>7.28</td>
<td>15.48</td>
<td>13.43</td>
<td>4.96</td>
<td>3.44</td>
<td>3.36</td>
<td>5.44</td>
<td>4.72</td>
<td>7.20</td>
<td>4.56</td>
<td>1.92</td>
<td>7.12</td>
<td>5.36</td>
<td>3.36</td>
<td>3.84</td>
</tr>
<tr>
<td>resynth</td>
<td>3.55</td>
<td>7.97</td>
<td>13.28</td>
<td>11.37</td>
<td>4.88</td>
<td>4.00</td>
<td>6.19</td>
<td>5.27</td>
<td>5.46</td>
<td>6.39</td>
<td>3.56</td>
<td>2.08</td>
<td>6.13</td>
<td>5.69</td>
<td>4.34</td>
<td>3.03</td>
</tr>
<tr>
<td>resynth-Gauss</td>
<td>6.31</td>
<td>6.03</td>
<td>5.78</td>
<td>6.11</td>
<td>6.59</td>
<td>6.73</td>
<td>6.74</td>
<td>6.76</td>
<td>6.75</td>
<td>6.75</td>
<td>6.77</td>
<td>6.80</td>
<td>6.84</td>
<td>6.82</td>
<td>6.79</td>
<td>6.78</td>
</tr>
<tr>
<td>resynth-fixenc</td>
<td>7.41</td>
<td>7.35</td>
<td>11.40</td>
<td>10.46</td>
<td>4.04</td>
<td>4.60</td>
<td>2.95</td>
<td>6.41</td>
<td>4.30</td>
<td>7.86</td>
<td>5.26</td>
<td>2.45</td>
<td>9.21</td>
<td>6.77</td>
<td>3.90</td>
<td>2.85</td>
</tr>
<tr>
<td>resynth-UFANS</td>
<td>4.08</td>
<td>8.09</td>
<td>9.41</td>
<td>8.45</td>
<td>6.90</td>
<td>5.70</td>
<td>5.21</td>
<td>5.71</td>
<td>5.98</td>
<td>5.29</td>
<td>4.87</td>
<td>5.20</td>
<td>5.43</td>
<td>5.14</td>
<td>5.10</td>
<td>5.37</td>
</tr>
</tbody>
</table>

Table 6: Comparison of alignment quality between different configurations. Sine-Cosine or Gaussian encoding function, trainable or fixed position encoding and CNN or UFANS decoder are evaluated based on their average difference between their duration prediction and real duration length on 100 audios.

<table>
<thead>
<tr>
<th>Encoding func</th>
<th>Trainable?</th>
<th>Decoder</th>
<th>Average-diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>Trainable</td>
<td>CNN</td>
<td>2.58</td>
</tr>
<tr>
<td>Sin-Cos</td>
<td>Fixed</td>
<td>CNN</td>
<td>1.96</td>
</tr>
<tr>
<td>Sin-Cos</td>
<td>Trainable</td>
<td>UFANS</td>
<td>1.80</td>
</tr>
<tr>
<td>Sin-Cos</td>
<td>Trainable</td>
<td>CNN</td>
<td><strong>0.85</strong></td>
</tr>
</tbody>
</table>

References


