TalkNet: Non-Autoregressive Depth-Wise Separable Convolutional Model for Speech Synthesis

Stanislav Beliaev, Boris Ginsburg

NVIDIA, Santa Clara, USA
{stanislavv,bginsburg}@nvidia.com

Abstract

We propose TalkNet, a non-autoregressive convolutional neural model for speech synthesis with explicit pitch and duration prediction. The model consists of three feed-forward convolutional networks. The first network predicts grapheme durations. An input text is then expanded by repeating each symbol according to the predicted duration. The second network predicts pitch value for every mel frame. All networks are based on 1D depth-wise separable convolutional architecture. The explicit duration prediction eliminates word skipping and repeating. The quality of the generated speech nearly matches the best auto-regressive models – TalkNet trained on the LJSpeech dataset got a MOS of 4.08. The model has only 13.2M parameters, almost 2x less than the present state-of-the-art text-to-speech models. The non-autoregressive architecture allows for fast training and inference. The small model size and fast inference make TalkNet an attractive candidate for embedded speech synthesis.

1. Introduction

Neural Network (NN) based models for text-to-speech (TTS) have outperformed both concatenative and statistical parametric speech synthesis in terms of speech quality. Neural TTS systems typically have two stages. In the first stage, a model generates mel-spectrograms from text. In the second stage, a NN-based vocoder synthesizes speech from mel-spectrograms. Most NN-based TTS models have an encoder-attention-decoder architecture [1], which has been observed to have some common problems [2]:

1. A tendency to repeat or skip words due to attention failures. To handle this issue, one can use additional mechanisms to encourage monotonic attention [3, 4].
2. Slow inference relative to parametric models.
3. No easy way to control of word duration or voice speed.

We propose a new neural TTS model to address these three challenges. The model consists of three convolutional networks. The first network predicts grapheme durations. The explicit duration predictor replaces attention-based alignment to prevent word skipping and repeating. Next we expand an input text based on the grapheme predicted duration. The second network predicts pitch along expanded sequence. The third network generates mel-spectrograms from an expanded text and predicted pitch. Finally, we use an external vocoder - WaveGlow [5] or HiFi-GAN vocoder [6] - to synthesize audio from mel-spectrograms (see Figure 1).

To train the grapheme duration predictor, we use the ground truth alignment between input characters and speech. To obtain this alignment, we proposed a new method which uses an external automatic speech recognition (ASR) Connectionist Temporal Classification (CTC) based model. Instead of using the greedy output of ASR model like in DeepVoice [7], we extract the alignment by calculating the most probable path in the lattice using the Viterbi algorithm.

The quality of speech generated by TalkNet combined with a HiFi-GAN vocoder nearly matches the best auto-regressive models: TalkNet trained on the LJSpeech got a MOS of 4.08. The TalkNet architecture is based on 1D depth-wise separable convolutions, similar to QuartzNet [8]. As a result, TalkNet has only 13.2M parameters - almost 2X less than models with similar quality, e.g. the Transformer-based FastSpeech [2, 9], or FastPitch [10]. The non-autoregressive architecture enables fast parallel training and inference comparing to auto-regressive models like Tacotron 2 [3]. The TalkNet inference is 132 times faster than real-time with batch size 1, and 2531 times faster for batch size 32.

The high quality of generated speech combined with low compute and memory requirements make the TalkNet an attractive candidate for embedded speech synthesis.
2. Related work

A traditional parametric TTS pipeline has the following stages: grapheme-to-phoneme conversion, a phoneme duration predictor, an acoustic frame-level feature generator, and a vocoder [11]. Neural networks for TTS have been explored in 1990 [12, 13, 14, 15] but the speech quality was below the quality of traditional methods at that time. Deep learning re-introduced NNs to TTS in 2015: Zen et al [16, 17] proposed a hybrid NN-parametric TTS model, where deep NNs are used to predict the phoneme duration and to generate frame-level acoustic features. Merlin TTS system [18] supports various NN architectures, including a standard feed-forward NNs, and recurrent neural networks (RNNs) to predict acoustic features, which are then passed to a traditional vocoder to produce the speech.

DeepVoice models [7, 19] adopted the traditional TTS pipeline, but they replace all components with NNs. To train the phoneme duration predictor, an auxiliary CTC-based model for phonetic segmentation was used to annotate data with phoneme boundaries. Tacotron 2 [20, 3] one of the first end-to-end neural TTS model which are based on encoder-attention-decoder architecture. The encoder is composed of three convolutional layers plus a single bidirectional LSTM. The decoder is a RNN with location-sensitive monotonic attention.


Attention-based models (e.g. Tacotron, Transformer-TTS, ParaNet etc) occasionally miss or repeat words [21]. To prevent word skipping and repeating, FastSpeech [2] proposes a feed-forward Transformer-based model with explicit length regulator. FastSpeech expands the sequence of phonemes according to a predicted duration in order to match the length of a mel-spectrogram sequence. The target phoneme duration is extracted from the attention alignment in an external pre-trained Tacotron 2 model. However, FastSpeech had several disadvantages: 1) the target mel-spectrograms distilled from teacher model suffer from information loss, 2) the phoneme duration extracted from the attention in teacher model is not accurate enough. FastPitch [10] is a logical extension of the FastSpeech model with explicit pitch modeling. Pitch is extracted and averaged over each grapheme duration, which helps make prediction process easier at the inference stage. FastPitch also relies on a pretrained attention TTS model to extract durations and it archives a notable quality increase in terms of Mean Opinion Score (MOS).

3. System architecture

TalkNet splits the text-to-spectrogram generation into three separate modules (see Figure 1). The first module, the duration predictor, aligns input graphemes in time with respect to the audio features. The second module, the pitch predictor, reconstructs pitch along expanded input tokens. The third module generates mel-spectrograms from time-aligned input sequence and pitch pattern. We use feed-forward CNNs for all modules, so both training and inference are non-autoregressive.

3.1. Grapheme duration predictor

This model predicts the length of the mel-spectrogram part corresponding to each grapheme in the input including punctuation. First, the grapheme duration predictor inserts a blank symbol ~ between every two input tokens. Next, it predicts the duration for each input token including blank. Finally, the sequence of input graphemes is expanded by repeating each token according to the predicted duration (see Figure 2).

![Grapheme duration prediction](image)

Figure 2: Grapheme duration prediction.

The grapheme duration predictor is a 1D time channel separable convolutional NN based on the QuartzNet architecture [8]. The model has five residual blocks with five sub-blocks per block. A sub-block consists of a 1D time-channel separable convolution, a \(1 \times 1\) pointwise convolution, batch normalization, ReLU, and dropout (see Figure 3). There are two additional layers: the grapheme embedding layer, and a \(1 \times 1\) convolution before the loss (see Table 1).

![Basic QuartzNet block](image)

Figure 3: Basic QuartzNet block. Both the grapheme duration predictor and the mel-spectrogram generator are 1D time-channel convolutional networks based on QuartzNet [8].

<table>
<thead>
<tr>
<th>Block</th>
<th># Sub Blocks</th>
<th># Output Channels</th>
<th>Kernel Size</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embed</td>
<td>1</td>
<td>64</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>Conv1</td>
<td>3</td>
<td>256</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>Conv2</td>
<td>5</td>
<td>256</td>
<td>{5, 7, 9, 11, 13}</td>
<td>0.1</td>
</tr>
<tr>
<td>Conv3</td>
<td>1</td>
<td>512</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>Conv1</td>
<td>1</td>
<td>32</td>
<td>1</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Parameters, M 2.3

Table 1: Both grapheme duration and pitch predictors has identical architectures based on QuartzNet-5x5.
We train a duration predictor using $L_2$ loss with logarithmic targets, similar to [2]. We used a log scale for large durations since the grapheme duration distribution has a long tail (see Figure 4). We also tried cross-entropy loss with each class corresponding to the character duration, but we found that speech generated with $L_2$ loss received a slightly higher MOS in our evaluation studies.

![Figure 4: The duration distribution for graphemes (left) and for blanks (right) based on CTC output for the LJSpeech dataset.](image)

### 3.2. Pitch predictor

The pitch prediction network is QuartzNet-5x5, similar to the duration predictor (see Table 1). The model has two heads: one head is binary cross-entropy (BCE) for non-voiced prediction, another – mean squared error (MSE) for non-zero values. Final loss is computed as sum of two head’s losses with equal weights coefficients. We scale MSE loss with global mean and standard deviation statistics to make it comparable to BCE (Figure 5). To train the pitch predictor we used as input the sequence of text tokens expanded with ground truth token durations. Following [24], we use Gaussian embedding layer for tokens with low duration and ground truth duration from a CTC ASR model, similar to [7]. We extracted the ground truth alignment from a CTC ASR model, similar to [7]. The CTC model assigns a probability to each of graphemes and its duration roughly corresponds to the length of the transition from one token to another.

We train ASR model with text with punctuation, converted to phonemes using g2p seq2seq model [28]. We use QuartzNet-5x5 [8] with a minor modification: we set the stride in the first convolutional layer to 1 to make the length of the CTC input equal to the mel-spectrogram. Model was trained on LibriTTS [29], and it achieves a phoneme error rate (PER) of 8.5% on LibriTTS test-clean.

![Figure 5: Pitch predictor loss is a sum of two heads: one head is binary cross-entropy (BCE) for non-voiced prediction, another head is mean squared error (MSE) for non-zero values.](image)

### 3.3. Mel-spectrogram generator

The mel-spectrogram generator is QuartzNet-9x5 with a mean squared error (MSE) loss, see Table 2. Mel-spectrogram generator is based on QuartzNet-9x5.

![Table 2: Mel-spectrogram generator is based on QuartzNet-9x5.](image)

<table>
<thead>
<tr>
<th>Block</th>
<th># Sub Blocks</th>
<th># Output Channels</th>
<th>Kernel Size</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>Embed</td>
<td>1</td>
<td>256</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>Conv1</td>
<td>3</td>
<td>256</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>$B_i \ldots B_n$</td>
<td>5</td>
<td>256</td>
<td>${5,7,9,13,15,17}$</td>
<td>0.1</td>
</tr>
<tr>
<td>$B_f \ldots B_n$</td>
<td>5</td>
<td>512</td>
<td>${21,23,25}$</td>
<td>0.1</td>
</tr>
<tr>
<td>Conv2</td>
<td>1</td>
<td>1024</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>Conv3</td>
<td>1</td>
<td>80</td>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>Parameters, M</td>
<td></td>
<td></td>
<td></td>
<td>8.5</td>
</tr>
</tbody>
</table>

for validation, and 500 samples for testing. The text was lower-cased while leaving all punctuation intact. We convert ground truth audio to mel-spectrograms using a Short-Time Fourier Transform (STFT) with 50 ms Hann window and 12.5 ms frame hop. Ground truth pitch was extracted using pysptk library [26] with values aligned with mel-spectrogram frames.

### 4. Training

The model was trained on LJSpeech dataset [25] which was split into three sets: 12,500 samples for training, 100 samples for validation, and 500 samples for testing. The text was lower-cased while leaving all punctuation intact. We convert ground truth audio to mel-spectrograms using a Short-Time Fourier Transform (STFT) with 50 ms Hann window and 12.5 ms frame hop. Ground truth pitch was extracted using pysptk library [26] with values aligned with mel-spectrogram frames.

### 4.2. Ground truth duration extraction

The non-autoregressive model rely on audio/text alignment based on the prediction of duration for each input token. To obtain the ground truth durations one can use an external aligner, for example Montreal Forced Aligner [27]. We extracted the ground truth alignment from a CTC ASR model, similar to [7]. The CTC model assigns a probability to each of graphemes from the phonetic alphabet plus an auxiliary blank symbol ∼. The blank symbol acts as an intermediate state between two neighbouring graphemes, and its duration roughly corresponds to the length of the transition from one token to another.

We train ASR model with text with punctuation, converted to phonemes using g2p seq2seq model [28]. We use QuartzNet-5x5 [8] with a minor modification: we set the stride in the first convolutional layer to 1 to make the length of the CTC input equal to the mel-spectrogram. Model was trained on LibriTTS [29], and it achieves a phoneme error rate (PER) of 8.5% on LibriTTS test-clean.

Instead of just using the CTC output which can have errors, we apply the Viterbi alignment algorithm between ground truth text tokens and NN output with log probabilities. The algorithm finds most probable alignment by estimating maximum joint probability for prefix of ground truth input tokens using dynamic programming. The algorithm has two passes: forward pass – to calculate maximum sum of output log probabilities over time, and backward pass – to traverse best path in reverse mode and count number of steps we stays at each grapheme. The overall algorithm similar to CTC alignment, but instead of summing over all correct paths, we use only the most probable path.

### 4.3. Grapheme duration and pitch predictors training

The NN for both grapheme duration and pitch was trained using the Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, a weight decay of $10^{-6}$ and gradient norm clipping of 1.0. We used a cosine decay learning rate policy starting from $10^{-3}$ and decreasing to $10^{-5}$ with a 2% warmup. We used a batch of 256 for one 16GB GPU and scaled learning rate for multi-GPU setups. Training duration predictor for 200 epochs takes $\approx 1.3$ hours on one V100 GPU, and 11 minutes on a DGX1 server with 8 GPUs in mixed precision [30]. Training pitch predictor takes...
about 2 hours on one V100 GPU, and 25 minutes on DGX1. The duration predictor accuracy is around 81%, but it covers approximately 95% of classes within an absolute distance of 1.

4.4. Mel-spectrogram generator training

We train the mel-spectrogram generator with the same parameters as above. We used a batch size of 64 for one 16GB GPU, and scale the learning rate for a multi-GPU setup. Training for 200 epochs takes ≈ 8 hours for one V100 GPU and less than 2 hours for DGX1.

5. Results

5.1. Audio quality

We conduct the MOS (mean opinion score) evaluation for generated speech using Amazon Mechanical Turk. We compared five types of samples: 1) ground truth audio, 2) ground truth mel-spectrogram converted with HiFi-GAN vocoder, 3) Tacotron 2 + HiFi-GAN, and 4) FastPitch + HiFi-GAN 5) TalkNet + HiFi-GAN. We used NVIDIA’s implementation for Tacotron 2 + HiFi-GAN, and 4) FastPitch + HiFi-GAN 5) TalkNet + HiFi-GAN. We tested 100 audio samples with 10 people per sample. The scores ranged from 1.0 to 5.0 with a step of 0.5. TalkNet speech quality comes quite close to Tacotron 2 and FastPitch (see Table 3).

Table 3: MOS scores with 95% confidence interval

<table>
<thead>
<tr>
<th>Model</th>
<th>MOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth speech</td>
<td>4.32 ± 0.04</td>
</tr>
<tr>
<td>Ground truth mel + HiFi-GAN</td>
<td>4.20 ± 0.04</td>
</tr>
<tr>
<td>Tacotron 2 + HiFi-GAN</td>
<td>4.20 ± 0.04</td>
</tr>
<tr>
<td>FastPitch + HiFi-GAN</td>
<td>4.13 ± 0.05</td>
</tr>
<tr>
<td>TalkNet + HiFi-GAN</td>
<td>4.08 ± 0.05</td>
</tr>
</tbody>
</table>

TalkNet is very robust with respect to missing or repeated words compared to auto-regressive TTS models such as Tacotron 2 or Transformer TTS. We evaluated the robustness of TalkNet on 50 hard sentences from the FastSpeech paper [2] and found that TalkNet practically eliminates missed or repeated words.

5.2. Inference

In the inference mode, we first insert blank symbols into the tokenized input text between every two graphemes. The obtained sequence is passed through the grapheme duration predictor. The input sequence is expanded with each token repeated according to the predicted duration. The second network predicts pitch: one head predicts silence parts, another head predicts values for voiced parts corrected with global statistics. The third network generates the mel-spectrogram from the expanded grapheme sequence and predicted pitch values.

We compare TalkNet inference with Tacotron 2 and FastPitch [10]. To measure the inference speed, we generate mel-spectrograms with a batch size equal to 1 and 32 for 100 samples from the LJSpeech test split. The average mel-spectrogram length is 563.93 frames. TalkNet inference is significantly faster than Tacotron 2. And it scales well with increasing the batch size (see Table 4). Since TalkNet does not use an attention mechanism, the inference speed practically doesn’t depend on the input length.

Table 4: Inference real-time factor (RTF) for mel-spectrogram generation (without vocoder). The RTF was measured with batch size 1 and 32 using a V100 GPU and averaged over 100 test samples from LJSpeech using PyTorch inference with AMP FP16 mode.

<table>
<thead>
<tr>
<th>Model</th>
<th>RTF, Batch size 1</th>
<th>RTF, Batch size 32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tacotron 2 [3]</td>
<td>13</td>
<td>233</td>
</tr>
<tr>
<td>FastPitch [10]</td>
<td>573</td>
<td>3502</td>
</tr>
<tr>
<td>TalkNet</td>
<td>132</td>
<td>2531</td>
</tr>
</tbody>
</table>

6. Conclusions

In this paper, we present TalkNet, a non-autoregressive fully convolutional neural speech synthesis system. The model is composed of three convolutional networks: a grapheme duration predictor, a pitch predictor and a mel-spectrogram generator. The model does not require another text-to-speech model as a teacher. The explicit duration predictor practically eliminates skipped or repeated words. TalkNet achieves a comparable level of speech quality to Tacotron 2 and FastPitch. The model is very compact. It has only 13.2M parameters, almost 2x less than similar neural TTS models: Tacotron-2 has 28.2M, and FastSpeech has 30.1M parameters. Training TalkNet takes only around 2 hours on a server with 8 V100 GPUs. The parallel mel-spectrogram generation makes the inference significantly faster and optimal for scaling with increasing the batch size. This makes TalkNet a great candidate for using in full-fledged TTS system and with streaming mode.

The model, the training recipes, and audio samples will be released as part of the NeMo toolkit [31].

7. Acknowledgments

The authors thank Jocelyn Huang, Rafael Valle, Felix Kreuk, Adrian Lancucki, Vitaly Lavrukhin, Jason Li, Christopher Parisien, and Joao Felipe Santos for the helpful feedback and review.

8. References


