Non-Autoregressive Neural Text-to-Speech

Kainan Peng[* 1] **Wei Ping**[* 1] **Zhao Song**[* 1] **Kexin Zhao**[* 1]

Abstract

In this work, we propose ParaNet, a non-autoregressive *seq2seq* model that converts text to spectrogram. It is fully convolutional and brings 46.7 times speed-up over the lightweight Deep Voice 3 at synthesis, while obtaining reasonably good speech quality. ParaNet also produces stable alignment between text and speech on the challenging test sentences by iteratively improving the attention in a layer-by-layer manner. Furthermore, we build the parallel text-to-speech system and test various parallel neural vocoders, which can synthesize speech from text through a single feed-forward pass. We also explore a novel VAE-based approach to train the *inverse autoregressive flow* (IAF) based parallel vocoder from scratch, which avoids the need for distillation from a separately trained WaveNet as previous work.

1. Introduction

Text-to-speech (TTS), also called speech synthesis, has long been a vital tool in a variety of applications, such as human-computer interactions, virtual assistant, and content creation. Traditional TTS systems are based on multi-stage hand-engineered pipelines (Taylor, 2009). In recent years, deep neural networks based autoregressive models have attained state-of-the-art results, including high-fidelity audio synthesis (van den Oord et al., 2016), and much simpler *seq2seq* pipelines (Sotelo et al., 2017; Wang et al., 2017; Ping et al., 2018b). In particular, one of the most popular neural TTS pipeline (a.k.a. "end-to-end") consists of two components (Ping et al., 2018b; Shen et al., 2018): (*i*) an autoregressive *seq2seq* model that generates mel spectrogram from text, and (*ii*) an autoregressive neural vocoder (e.g., WaveNet) that synthesizes raw waveform from mel spectro-

gram. This pipeline requires much less expert knowledge and only needs pairs of audio and transcript as training data.

However, the autoregressive nature of these models makes them quite slow at synthesis, because they operate sequentially at a high temporal resolution of waveform samples and spectrogram. Most recently, several models are proposed for parallel waveform generation (e.g., van den Oord et al., 2018; Ping et al., 2018a; Prenger et al., 2019; Kumar et al., 2019; Bińkowski et al., 2020; Ping et al., 2020). In the end-to-end pipeline, the models (e.g., ClariNet, WaveFlow) still rely on autoregressive component to predict spectrogram features (e.g., 100 frames per second). In the linguistic feature-based pipeline, the models (e.g., Parallel WaveNet, GAN-TTS) are conditioned on aligned linguistic features from phoneme duration model and F0 from frequency model, which are recurrent or autoregressive models. Both of these TTS pipelines can be slow at synthesis on modern hardware optimized for parallel execution.

In this work, we present a fully parallel neural TTS system by proposing a non-autoregressive text-to-spectrogram model. Our major contributions are as follows:

1. We propose ParaNet, a non-autoregressive attention-based architecture for text-to-speech, which is fully convolutional and converts text to mel spectrogram. It runs 254.6 times faster than real-time at synthesis on a 1080 Ti GPU, and brings 46.7 times speed-up over its autoregressive counterpart (Ping et al., 2018b), while obtaining reasonably good speech quality using neural vocoders.

2. ParaNet distills the attention from the autoregressive text-to-spectrogram model, and iteratively refines the alignment between text and spectrogram in a layer-by-layer manner. It can produce more stable attentions than autoregressive Deep Voice 3 (Ping et al., 2018b) on the challenging test sentences, because it does not have the discrepancy between the teacher-forced training and autoregressive inference.

3. We build the fully parallel neural TTS system by combining ParaNet with parallel neural vocoder, thus it can generate speech from text through a single feed-forward pass. We investigate several parallel vocoders, including the distilled IAF vocoder (Ping et al., 2018a) and Wave-Glow (Prenger et al., 2019). To explore the possibility of training IAF vocoder *without distillation*, we also

---

*Equal contribution . [1]Baidu Research, 1195 Bordeaux Dr, Sunnyvale, CA. Speech samples can be found in: `https://parallel-neural-tts-demo.github.io/`. Correspondence to: Wei Ping <weiping.thu@gmail.com>.

propose an alternative approach, WaveVAE, which can be trained from scratch within the variational autoencoder (VAE) framework (Kingma & Welling, 2014).

We organize the rest of paper as follows. Section 2 discusses related work. We introduce the non-autoregressive ParaNet architecture in Section 3. We discuss parallel neural vocoders in Section 4, and report experimental settings and results in Section 5. We conclude the paper in Section 6.

2. Related work

Neural speech synthesis has obtained the state-of-the-art results and gained a lot of attention. Several neural TTS systems were proposed, including WaveNet (van den Oord et al., 2016), Deep Voice (Arık et al., 2017a), Deep Voice 2 (Arık et al., 2017b), Deep Voice 3 (Ping et al., 2018b), Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2018), Char2Wav (Sotelo et al., 2017), VoiceLoop (Taigman et al., 2018), WaveRNN (Kalchbrenner et al., 2018), ClariNet (Ping et al., 2018a), and Transformer TTS (Li et al., 2019). In particular, Deep Voice 3, Tacotron and Char2Wav employ *seq2seq* framework with the attention mechanism (Bahdanau et al., 2015), yielding much simpler pipeline compared to traditional multi-stage pipeline. Their excellent extensibility leads to promising results for several challenging tasks, such as voice cloning (Arik et al., 2018; Nachmani et al., 2018; Jia et al., 2018; Chen et al., 2019). All of these state-of-the-art systems are based on autoregressive models.

RNN-based autoregressive models, such as Tacotron and WaveRNN (Kalchbrenner et al., 2018), lack parallelism at both training and synthesis. CNN-based autoregressive models, such as Deep Voice 3 and WaveNet, enable parallel processing at training, but they still operate sequentially at synthesis since each output element must be generated before it can be passed in as input at the next time-step. Recently, there are some non-autoregressive models proposed for neural machine translation. Gu et al. (2018) trains a feed-forward neural network conditioned on fertility values, which are obtained from an external alignment system. Kaiser et al. (2018) proposes a latent variable model for fast decoding, while it remains autoregressiveness between latent variables. Lee et al. (2018) iteratively refines the output sequence through a denoising autoencoder framework. Arguably, non-autoregressive model plays a more important role in text-to-speech, where the output speech spectrogram usually consists of hundreds of time-steps for a short text input with a few words. Our work is one of the first non-autoregressive *seq2seq* model for TTS and provides as much as 46.7 times speed-up at synthesis over its autoregressive counterpart (Ping et al., 2018b). There is a concurrent work (Ren et al., 2019), which is based on the autoregressive transformer TTS (Li et al., 2019) and

can generate mel spectrogram in parallel. Our ParaNet is fully convolutional and lightweight. In contrast to Fast-Speech, it has half of model parameters, requires smaller batch size (16 vs. 64) for training and provides faster speed at synthesis (see Table 2 for detailed comparison).

Flow-based generative models (Rezende & Mohamed, 2015; Kingma et al., 2016; Dinh et al., 2017; Kingma & Dhariwal, 2018) transform a simple initial distribution into a more complex one by applying a series of invertible transformations. In previous work, flow-based models have obtained state-of-the-art results for parallel waveform synthesis (van den Oord et al., 2018; Ping et al., 2018a; Prenger et al., 2019; Kim et al., 2019; Yamamoto et al., 2019; Ping et al., 2020).

Variational autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) has been applied for representation learning of natural speech for years. It models either the generative process of raw waveform (Chung et al., 2015; van den Oord et al., 2017), or spectrograms (Hsu et al., 2019). In previous work, autoregressive or recurrent neural networks are employed as the decoder of VAE (Chung et al., 2015; van den Oord et al., 2017), but they can be quite slow at synthesis. In this work, we employ a feed-forward IAF as the decoder, which enables parallel waveform synthesis.

3. Text-to-spectrogram model

Our parallel TTS system has two components: 1) a feed-forward text-to-spectrogram model, and 2) a parallel waveform synthesizer conditioned on mel spectrogram. In this section, we first present an autoregressive model derived from Deep Voice 3 (DV3) (Ping et al., 2018b). We then introduce ParaNet, a non-autoregressive text-to-spectrogram model (see Figure 1).

3.1. Autoregressive architecture

Our autoregressive model is based on DV3, a convolutional text-to-spectrogram architecture, which consists of three components:

- **Encoder**: A convolutional encoder, which takes text inputs and encodes them into internal hidden representation.

- **Decoder**: A *causal* convolutional decoder, which decodes the encoder representation with an *attention* mechanism to log-mel spectragrams in an *autoregressive* manner with an ℓ_1 loss. It starts with a 1×1 convolution to preprocess the input log-mel spectrograms.

- **Converter**: A *non-causal* convolutional post processing network, which processes the hidden representation from the decoder using both past and future context information and predicts the log-linear spectrograms with an ℓ_1 loss. It enables bidirectional processing.

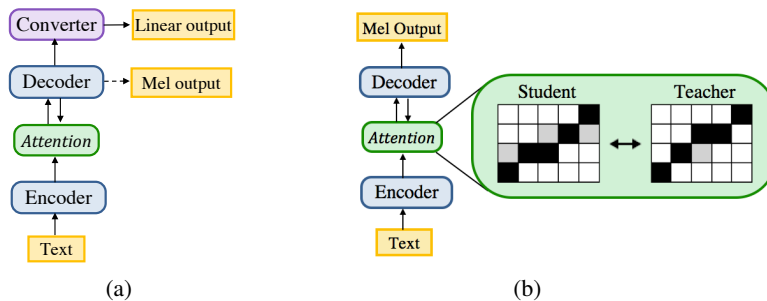All these components use the same 1-D *convolution block*

Figure 1. (a) Autoregressive *seq2seq* model. The dashed line depicts the autoregressive decoding of mel spectrogram at inference. (b) Non-autoregressive ParaNet model, which distills the attention from a pretrained autoregressive model.
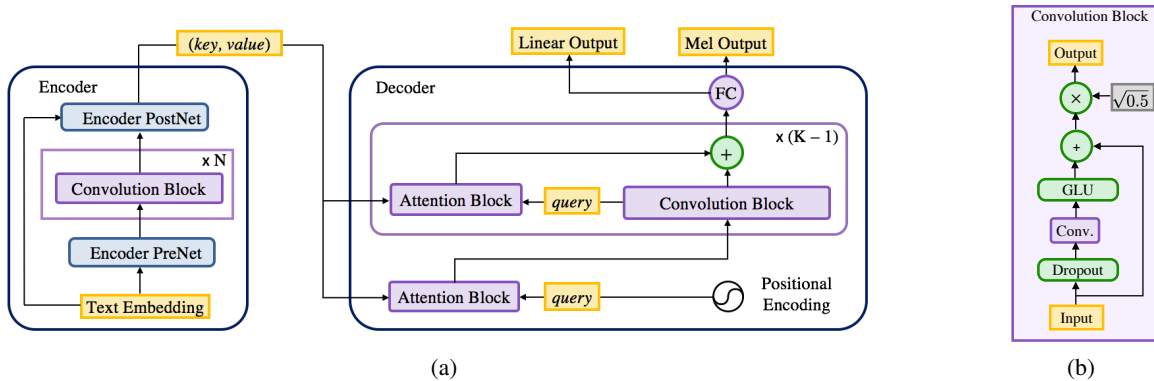

Figure 2. (a) Architecture of ParaNet. Its encoder provides *key* and *value* as the textual representation. The first attention block in decoder gets positional encoding as the *query* and is followed by non-causal convolution blocks and attention blocks. (b) Convolution block appears in both encoder and decoder. It consists of a 1-D convolution with a gated linear unit (GLU) and a residual connection.

with a gated linear unit as in DV3 (see Figure 2 (b) for more details). The major difference between our model and DV3 is the decoder architecture. The decoder of DV3 has multiple attention-based layers, where each layer consists of a causal convolution block followed by an attention block. To simplify the *attention distillation* described in Section 3.3.1, our autoregressive decoder has only one attention block at its first layer. We find that reducing the number of attention blocks does not hurt the generated speech quality in general.

3.2. Non-autoregressive architecture

The proposed ParaNet (see Figure 2) uses the same encoder architecture as the autoregressive model. The decoder of ParaNet, conditioned solely on the hidden representation from the encoder, predicts the entire sequence of log-mel spectrograms in a feed-forward manner. As a result, both its training and synthesis can be done in parallel. Specially, we make the following major architecture modifications from the autoregressive text-to-spectrogram model to the non-autoregressive model:

1. **Non-autoregressive decoder**: Without the autoregressive generative constraint, the decoder can use *non-causal* convolution blocks to take advantage of future context information and to improve model per-

formance. In addition to log-mel spectrograms, it also predicts log-linear spectrograms with an ℓ_1 loss for slightly better performance. We also remove the 1×1 convolution at the beginning, because the decoder does not take log-mel spectrograms as input.

2. **No converter**: Non-autoregressive model removes the *non-causal* converter since it already employs a *non-causal* decoder. Note that, the major motivation of introducing non-causal *converter* in DV3 is to refine the decoder predictions based on bidirectional context information provided by non-causal convolutions.

3.3. Parallel attention mechanism

It is challenging for the feed-forward model to learn the accurate alignment between the input text and output spectrogram. In particular, we need the full parallelism within the attention mechanism. For example, the location-sensitive attention (Chorowski et al., 2015; Shen et al., 2018) improves attention stability, but it performs sequentially at both training and synthesis, because it uses the cumulative attention weights from previous decoder time steps as an additional feature for the next time step. Previous non-autoregressive decoders rely on an external alignment system (Gu et al., 2018), or an autoregressive latent variable model (Kaiser
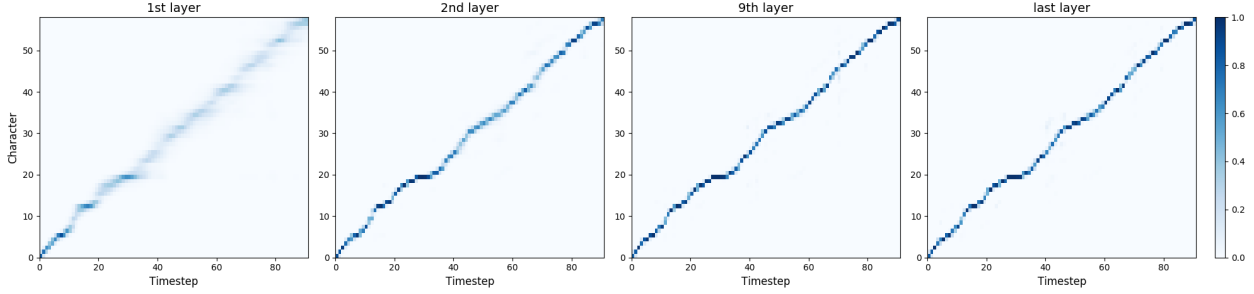
Figure 3. Our ParaNet iteratively refines the attention alignment in a layer-by-layer way. One can see the 1st layer attention is mostly dominated by the positional encoding prior. It becomes more and more confident about the alignment in the subsequent layers.

et al., 2018).

In this work, we present several simple & effective techniques, which could obtain accurate and stable attention alignment. In particular, our non-autoregressive decoder can iteratively refine the attention alignment between text and mel spectrogram in a layer-by-layer manner as illustrated in Figure 3. Specially, the decoder adopts a dot-product attention mechanism and consists of K attention blocks (see Figure 2 (a)), where each attention block uses the per-time-step *query* vectors from convolution block and per-time-step *key* vectors from encoder to compute the attention weights (Ping et al., 2018b). The attention block computes *context* vectors as the weighted average of the *value* vectors from the encoder. The non-autoregressive decoder starts with an attention block, in which the *query* vectors are solely positional encoding (see Section 3.3.2 for details). The first attention block then provides the input for the convolution block at the next attention-based layer.

3.3.1. ATTENTION DISTILLATION

We use the attention alignments from a pretrained autoregressive model to guide the training of non-autoregressive model. Specifically, we minimize the cross entropy between the attention distributions from the non-autoregressive ParaNet and a pretrained autoregressive teacher. We denote the attention weights from the non-autoregressive ParaNet as $W_{i,j}^{(k)}$, where i and j index the time-step of encoder and decoder respectively, and k refers to the k-th attention block within the decoder. Note that, the attention weights $\{W_{i,j}^{(k)}\}_{i=1}^M$ form a valid distribution. We compute the *attention loss* as the average cross entropy between the ParaNet and teacher's attention distributions:

$$l_{\text{atten}} = -\frac{1}{KN} \sum_{k=1}^K \sum_{j=1}^N \sum_{i=1}^M W_{i,j}^t \log W_{i,j}^{(k)}, \qquad (1)$$

where $W_{i,j}^t$ are the attention weights from the autoregressive teacher, M and N are the lengths of encoder and decoder, respectively. Our final loss function is a linear combination of l_{atten} and ℓ_1 losses from spectrogram predictions. We set

the coefficient of l_{atten} as 4, and other coefficients as 1 in all experiments.

3.3.2. POSITIONAL ENCODING

We use a similar positional encoding as in DV3 at every attention block (Ping et al., 2018b). The positional encoding is added to both *key* and *query* vectors in the attention block, which forms an inductive bias for monotonic attention. Note that, the non-autoregressive model solely relies on its attention mechanism to decode mel spectrograms from the encoded textual features, without any autoregressive input. This makes the positional encoding even more crucial in guiding the attention to follow a monotonic progression over time at the beginning of training. The positional encodings $h_p(i, k) = \sin\left(\omega_s i / 10000^{k/d}\right)$ (for even i), and $\cos\left(\omega_s i / 10000^{k/d}\right)$ (for odd i), where i is the time-step index, k is the channel index, d is the total number of channels in the positional encoding, and ω_s is the *position rate* which indicates the average slope of the line in the attention distribution and roughly corresponds to the speed of speech. We set ω_s in the following ways:

- For the autoregressive teacher, ω_s is set to one for the positional encoding of *query*. For the *key*, it is set to the averaged ratio of the time-steps of spectrograms to the time-steps of textual features, which is around 6.3 across our training dataset. Taking into account that a reduction factor of 4 is used to simplify the learning of attention mechanism (Wang et al., 2017), ω_s is simply set as $^{6.3}/4$ for the *key* at both training and synthesis.

- For ParaNet, ω_s is also set to one for the *query*, while ω_s for the *key* is calculated differently. At training, ω_s is set to the ratio of the lengths of spectrograms and text for each individual training instance, which is also divided by a reduction factor of 4. At synthesis, we need to specify the length of output spectrogram and the corresponding ω_s, which actually controls the speech rate of the generated audios (see Section II on demo website). In all of our experiments, we simply set ω_s to be $^{6.3}/4$ as in autoregressive model, and the length of output spectrogram as $^{6.3}/4$ times the length of input text.

Such a setup yields an initial attention in the form of a diagonal line and guides the non-autoregressive decoder to refine its attention layer by layer (see Figure 3).

3.3.3. ATTENTION MASKING

Inspired by the attention masking in Deep Voice 3, we propose an attention masking scheme for the non-autoregressive ParaNet at synthesis:

- For each *query* from decoder, instead of computing the softmax over the entire set of encoder *key* vectors, we compute the softmax only over a fixed window centered around the *target position* and going forward and backward several time-steps (e.g., 3). The *target position* is calculated as $\lfloor i_{\text{query}} \times 4/6.3 \rfloor$, where i_{query} is the time-step index of the *query* vector, and $\lfloor \rfloor$ is the rounding operator.

We observe that this strategy reduces serious attention errors such as repeating or skipping words, and also yields clearer pronunciations, thanks to its more condensed attention distribution. Note that, this attention masking is shared across all attention blocks once it is generated, and does not prevent the parallel synthesis of the non-autoregressive model.

4. Parallel waveform model

As an indispensable component in our parallel neural TTS system, the parallel waveform model converts the mel spectrogram predicted from ParaNet into the raw waveform. In this section, we discuss several existing parallel waveform models, and explore a new alternative in the system.

4.1. Flow-based waveform models

Inverse autoregressive flow (IAF) (Kingma et al., 2016) is a special type of normalizing flow where each invertible transformation is based on an autoregressive neural network. IAF performs synthesis in parallel and can easily reuse the expressive autoregressive architecture, such as WaveNet (van den Oord et al., 2016), which leads to the state-of-the-art results for speech synthesis (van den Oord et al., 2018; Ping et al., 2018a). However, the likelihood evaluation in IAF is autoregressive and slow, thus previous training methods rely on probability density distillation from a pretrained autoregressive WaveNet. This two-stage distillation process complicates the training pipeline and may introduce pathological optimization (Huang et al., 2019).

RealNVP (Dinh et al., 2017) and Glow (Kingma & Dhariwal, 2018) are different types of normalizing flows, where both synthesis and likelihood evaluation can be performed in parallel by enforcing bipartite architecture constraints. Most recently, both of them were applied as parallel neu-

ral vocoders and can be trained from scratch (Prenger et al., 2019; Kim et al., 2019). However, these models are less expressive than their autoregressive and IAF counterparts. One can find a detailed analysis in WaveFlow paper (Ping et al., 2020). In general, these bipartite flows require larger number of layers and hidden units, which lead to huge number of parameters. For example, a WaveGlow vocoder (Prenger et al., 2019) has 87.88M parameters, whereas IAF vocoder has much smaller footprint with only 2.17M parameters (Ping et al., 2018a), making it more preferred in production deployment.

4.2. WaveVAE

Given the advantage of IAF vocoder, it is interesting to investigate whether it can be trained *without* the density distillation. One related work trains IAF within an autoencoder (Huang et al., 2019). Our method uses the VAE framework, thus it is termed as WaveVAE. In contrast to van den Oord et al. (2018) and Ping et al. (2018a), Wave-VAE can be trained from scratch by jointly optimizing the encoder $q_\phi(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{c})$ and decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{c})$, where $\boldsymbol{z}$ is latent variables and $\boldsymbol{c}$ is the mel spectrogram conditioner. We omit $\boldsymbol{c}$ for concise notation hereafter.

4.2.1. ENCODER

The encoder of WaveVAE $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is parameterized by a Gaussian autoregressive WaveNet (Ping et al., 2018a) that maps the ground truth audio $\boldsymbol{x}$ into the same length latent representation $\boldsymbol{z}$. Specifically, the Gaussian WaveNet models x_t given the previous samples $x_{<t}$ as $x_t \sim \mathcal{N}\big(\mu(x_{<t}; \phi), \sigma(x_{<t}; \phi)\big)$, where the mean $\mu(x_{<t}; \phi)$ and scale $\sigma(x_{<t}; \phi)$ are predicted by WaveNet, respectively. The encoder posterior is constructed as,

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \prod_t q_\phi(z_t \mid x_{\leq t}),$$

$$\text{where } q_\phi(z_t \mid x_{\leq t}) = \mathcal{N}\big(\frac{x_t - \mu(x_{<t}; \phi)}{\sigma(x_{<t}; \phi)}, \varepsilon\big).$$

Note that, the mean $\mu(x_{<t}; \phi)$ and scale $\sigma(x_{<t})$ are applied for "whitening" the posterior distribution. We introduce a trainable scalar $\varepsilon > 0$ to decouple the global variation, which will make optimization process easier. Given the observed $\boldsymbol{x}$, the $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ admits parallel sampling of latents $\boldsymbol{z}$. One can build the connection between the encoder of WaveVAE and the teacher model of ClariNet, as both of them use a Gaussian WaveNet to guide the training of IAF for parallel wave generation.

4.2.2. DECODER

Our decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is parameterized by the one-step-ahead predictions from an IAF (Ping et al., 2018a). We let $\boldsymbol{z}^{(0)} = \boldsymbol{z}$ and apply a stack of IAF transformations

from $z^{(0)} \to \dots z^{(i)} \to \dots z^{(n)}$, and each transformation $z^{(i)} = f(z^{(i-1)}; \theta)$ is defined as,

$$z^{(i)} = z^{(i-1)} \cdot \sigma^{(i)} + \mu^{(i)}, \qquad (2)$$

where $\mu_t^{(i)} = \mu(z_{<t}^{(i-1)}; \theta)$ and $\sigma_t^{(i)} = \sigma(z_{<t}^{(i-1)}; \theta)$ are shifting and scaling variables modeled by a Gaussian WaveNet. One can show that, given $z^{(0)} \sim \mathcal{N}(\mu^{(0)}, \sigma^{(0)})$ from the Gaussian prior or encoder, the per-step $p(z_t^{(n)} \mid z_{<t}^{(0)})$ also follows Gaussian with scale and mean as,

$$\sigma^{\text{tot}} = \prod_{i=0}^{n} \sigma^{(i)}, \quad \mu^{\text{tot}} = \sum_{i=0}^{n} \mu^{(i)} \prod_{j>i}^{n} \sigma^{(j)}. \qquad (3)$$

Lastly, we set $x = \epsilon \cdot \sigma^{\text{tot}} + \mu^{\text{tot}}$, where $\epsilon \sim \mathcal{N}(0, I)$. Thus, $p_\theta(x \mid z) = \mathcal{N}(\mu^{\text{tot}}, \sigma^{\text{tot}})$. For the generative process, we use the standard Gaussian prior $p(z) = \mathcal{N}(0, I)$.

4.2.3. Training Objective

We maximize the evidence lower bound (ELBO) for observed x in VAE,

$$\max_{\phi, \theta} \ \mathbb{E}_{q_\phi(z|x)} \big[\log p_\theta(x|z) \big] - \text{KL}\big(q_\phi(z|x) \,\|\, p(z)\big), \quad (4)$$

where the KL divergence can be calculated in closed-form as both $q_\phi(z|x)$ and $p(z)$ are Gaussians,

$$\begin{aligned} &\text{KL}\big(q_\phi(z|x) \,\|\, p(z)\big) \\ &= \sum_t \log \frac{1}{\varepsilon} + \frac{1}{2}\Big(\varepsilon^2 - 1 + \big(\frac{x_t - \mu(x_{<t})}{\sigma(x_{<t})}\big)^2\Big). \end{aligned}$$

The reconstruction term in Eq. (4) is intractable to compute exactly. We do stochastic optimization by drawing a sample z from the encoder $q_\phi(z|x)$ through the reparameterization trick, and evaluating the likelihood $\log p_\theta(x|z)$. To avoid the "posterior collapse", in which the posterior distribution $q_\phi(z|x)$ quickly collapses to the white noise prior $p(z)$ at the early stage of training, we apply the annealing strategy for KL divergence, where its weight is gradually increased from 0 to 1, via a sigmoid function (Bowman et al., 2016). Through it, the encoder can encode sufficient information into the latent representations at the early training, and then gradually regularize the latent representation by increasing the weight of the KL divergence.

STFT loss: Similar to Ping et al. (2018a), we also add a short-term Fourier transform (STFT) loss to improve the quality of synthesized speech. We define the *STFT loss* as the summation of ℓ_2 loss on the magnitudes of STFT and ℓ_1 loss on the log-magnitudes of STFT between the output audio and ground truth audio (Ping et al., 2018a; Arık et al., 2019; Wang et al., 2019). For STFT, we use a 12.5ms frame-shift, 50ms Hanning window length, and we set the FFT size to 2048. We consider two STFT losses

in our objective: (i) the STFT loss between ground truth audio and reconstructed audio using encoder $q_\phi(z|x)$; (ii) the STFT loss between ground truth audio and synthesized audio using the prior $p(z)$, with the purpose of reducing the gap between reconstruction and synthesis. Our final loss is a linear combination of VAE objective in Eq. (4) and the STFT losses. The corresponding coefficients are simply set to be one in all of our experiments.

5. Experiment

In this section, we present several experiments to evaluate the proposed ParaNet and WaveVAE.

5.1. Settings

Data: In our experiment, we use an internal English speech dataset containing about 20 hours of speech data from a female speaker with a sampling rate of 48 kHz. We down-sample the audios to 24 kHz.

Text-to-spectrogram models: For both ParaNet and Deep Voice 3 (DV3), we use the mixed representation of characters and phonemes (Ping et al., 2018b). The default hyperparameters of ParaNet and DV3 are provided in Table 1. Both ParaNet and DV3 are trained for 500K steps using Adam optimizer (Kingma & Ba, 2015). We find that larger kernel width and deeper layers generally help improve the performance of ParaNet. In terms of the number of parameters, our ParaNet (17.61 M params) is $2.57\times$ larger than the Deep Voice 3 (6.85M params) and $1.71\times$ smaller than the FastSpeech (30.1M params) (Ren et al., 2019). We use an open source reimplementation of FastSpeech [1] by adapting the hyperparameters for handling the 24kHz dataset.

Neural vocoders: In this work, we compare various neural vocoders paired with text-to-spectrogram models, including WaveNet (van den Oord et al., 2016), ClariNet (Ping et al., 2018a), WaveVAE, and WaveGlow (Prenger et al., 2019). We train all neural vocoders on 8 Nvidia 1080Ti GPUs using randomly chosen 0.5s audio clips.

We train two 20-layer WaveNets with residual channel 256 conditioned on the predicted mel spectrogram from ParaNet and DV3, respectively. We apply two layers of convolution block to process the predicted mel spectrogram, and use two layers of transposed 2-D convolution (in time and frequency) interleaved with leaky ReLU ($\alpha = 0.4$) to upsample the outputs from frame-level to sample-level. We use the Adam optimizer (Kingma & Ba, 2015) with a batch size of 8 and a learning rate of 0.001 at the beginning, which is annealed by half every 200K steps. We train the models for 1M steps.

We use the same IAF architecture as ClariNet (Ping et al., 2018a). It consists of four stacked Gaussian IAF blocks,

---

[1] `https://github.com/xcmyz/FastSpeech`

Table 1. Hyperparameters of autoregressive text-to-spectrogram model and non-autoregressive ParaNet in the experiment.

Hyperparameter	Autoregressive Model	Non-autoregressive Model
FFT Size	2048	2048
FFT Window Size / Shift	1200 / 300	1200 / 300
Audio Sample Rate	24000	24000
Reduction Factor r	4	4
Mel Bands	80	80
Character Embedding Dim.	256	256
Encoder Layers / Conv. Width / Channels	7 / 5 / 64	7 / 9 / 64
Decoder PreNet Affine Size	128, 256	N/A
Decoder Layers / Conv. Width	4 / 5	17 / 7
Attention Hidden Size	128	128
Position Weight / Initial Rate	1.0 / 6.3	1.0 / 6.3
PostNet Layers / Conv. Width / Channels	5 / 5 / 256	N/A
Dropout Keep Probability	0.95	1.0
ADAM Learning Rate	0.001	0.001
Batch Size	16	16
Max Gradient Norm	100	100
Gradient Clipping Max. Value	5.0	5.0
Total Number of Parameters	6.85M	17.61M

Table 2. The model footprint, synthesis time for 1 second speech (on 1080Ti with FP32), and the 5-scale Mean Opinion Score (MOS) ratings with 95% confidence intervals for comparison.

Neural TTS system	# parameters	synthesis time (ms)	MOS score
DV3 + WaveNet	6.85 + 9.08 = 15.93 M	181.8 + 5×10^5	4.09 ± 0.26
ParaNet + WaveNet	17.61 + 9.08 = 26.69 M	3.9 + 5×10^5	4.01 ± 0.24
DV3 + ClariNet	6.85 + 2.17 = 9.02 M	181.8 + 64.9	3.88 ± 0.25
ParaNet + ClariNet	17.61 + 2.17 = **19.78** M	**3.9 + 64.9**	3.62 ± 0.23
DV3 + WaveVAE	6.85 + 2.17 = 9.02 M	181.8 + 64.9	3.70 ± 0.29
ParaNet + WaveVAE	17.61 + 2.17 = 19.78 M	3.9 + 64.9	3.31 ± 0.32
DV3 + WaveGlow	6.85 + 87.88 = 94.73 M	181.8 + 117.6	3.92 ± 0.24
ParaNet + WaveGlow	17.61 + 87.88 = 105.49 M	3.9 + 117.6	3.27 ± 0.28
FastSpeech (re-impl) + WaveGlow	31.77 + 87.88 = **119.65** M	**6.2 + 117.6**	*3.56 ± 0.26*

which are parameterized by [10, 10, 10, 30]-layer WaveNets respectively, with the 64 residual & skip channels and filter size 3 in dilated convolutions. The IAF is conditioned on log-mel spectrograms with two layers of transposed 2-D convolution as in ClariNet. We use the same teacher-student setup for ClariNet as in Ping et al. (2018a) and we train a 20-layer Gaussian autoregressive WaveNet as the teacher model. For the encoder in WaveVAE, we also use a 20-layers Gaussian WaveNet conditioned on log-mel spectrograms. For the decoder, we use the same architecture as the distilled IAF. Both the encoder and decoder of WaveVAE share the same conditioner network. Both of the distilled IAF and WaveVAE are trained on ground-truth mel spectrogram. We use Adam optimizer with 1000K steps for distilled IAF. For WaveVAE, we train it for 400K because it converges much faster. The learning rate is set to 0.001 at the beginning and annealed by half every 200K steps for both models.

We use the open source implementation of WaveGlow with default hyperparameters (residual channel 256) [2], except change the sampling rate from 22.05kHz to 24kHz, FFT window length from 1024 to 1200, and FFT window shift from 256 to 300 for handling the 24kHz dataset. The model is trained for 2M steps.

5.2. Results

Speech quality: We use the crowdMOS toolkit (Ribeiro et al., 2011) for subjective Mean Opinion Score (MOS) evaluation. We report the MOS results in Table 2. The ParaNet can provide comparable quality of speech as the autoregressive DV3 using WaveNet vocoder (MOS: 4.09 vs. 4.01). When we use the ClariNet vocoder, ParaNet can still provide reasonably good speech quality (MOS:

[2] https://github.com/NVIDIA/waveglow

Table 3. Attention error counts for text-to-spectrogram models on the 100-sentence test set. One or more mispronunciations, skips, and repeats count as a single mistake per utterance. The non-autoregressive ParaNet (17-layer decoder) with attention mask obtains the fewest attention errors in total. For ablation study, we include the results for two additional ParaNet models. They have 6 and 12 decoder layers and are denoted as ParaNet-6 and ParaNet-12, respectively.

Model	Attention mask	Repeat	Mispronounce	Skip	Total
Deep Voice 3		12	10	15	37
ParaNet	No	**1**	**4**	**7**	**12**
ParaNet-12		5	7	5	17
ParaNet-6		4	11	11	26
Deep Voice 3		1	4	3	8
ParaNet	Yes	**2**	**4**	**0**	**6**
ParaNet-12		4	6	2	12
ParaNet-6		3	10	3	16

3.62) as a fully feed-forward TTS system. WaveVAE obtains worse results than distilled IAF vocoder, but it can be trained from scratch and simplifies the training pipeline. When conditioned on predicted mel spectrogram, WaveGlow tends to produce constant frequency artifacts. To remedy this, we applied the denoising function with strength 0.1, as recommended in the repository of WaveGlow. It is effective when the predicted mel spectrograms are from DV3, but not effective when the predicted mel spectrograms are from ParaNet. As a result, the MOS score degrades seriously. We add the comparison with FastSpeech after the paper submission. Because it is costly to relaunch the MOS evaluations of all the models, we perform a separate MOS evaluation for FastSpeech. Note that, the group of human raters can be different on Mechanical Turk, and the subjective scores may not be directly comparable. One can find the synthesized speech samples in: `https://parallel-neural-tts-demo.github.io/`.

Synthesis speed: We test synthesis speed of all models on NVIDIA GeForce GTX 1080 Ti with 32-bit floating point (FP32) arithmetic. We compare the ParaNet with the autoregressive DV3 in terms of inference latency. We construct a custom 15-sentence test set (see Appendix A) and run inference for 50 runs on each of the 15 sentences (batch size is set to 1). The average audio duration of the utterances is 6.11 seconds. The average inference latencies over 50 runs and 15 sentences are 0.024 and 1.12 seconds for ParaNet and DV3, respectively. Hence, our ParaNet runs 254.6 times faster than real-time and brings about 46.7 times speed-up over its small-footprint autoregressive counterpart at synthesis. It also runs 1.58 times faster than FastSpeech. We summarize synthesis speed of TTS systems in Table 2. One can observe that the latency bottleneck is the autoregressive text-to-spectrogram model, when the system uses parallel neural vocoder. The ClariNet and WaveVAE vocoders have much smaller footprint and faster synthesis speed than WaveGlow.

Attention error analysis: In autoregressive models, there is a noticeable discrepancy between the teacher-forced training and autoregressive inference, which can yield accumulated errors along the generated sequence at synthesis (Bengio et al., 2015). In neural TTS, this discrepancy leads to miserable attention errors at autoregressive inference, including (*i*) repeated words, (*ii*) mispronunciations, and (*iii*) skipped words (see Ping et al. (2018b) for detailed examples), which is a critical problem for online deployment of attention-based neural TTS systems. We perform an attention error analysis for our non-autoregressive ParaNet on a 100-sentence test set (see Appendix B), which includes particularly-challenging cases from deployed TTS systems (e.g. dates, acronyms, URLs, repeated words, proper nouns, and foreign words). In Table 3, we find that the non-autoregressive ParaNet has much fewer attention errors than its autoregressive counterpart at synthesis (12 vs. 37) without attention mask. Although our ParaNet distills the (teacher-forced) attentions from an autoregressive model, it only takes textual inputs at both training and synthesis and does not have the similar discrepancy as in autoregressive model. In previous work, attention masking was applied to enforce the monotonic attentions and reduce attention errors, and was demonstrated to be effective in Deep Voice 3 (Ping et al., 2018b). We find that our non-autoregressive ParaNet still can have fewer attention errors than autoregressive DV3 (6 vs. 8), when both of them use the attention masking.

5.3. Ablation study

We perform ablation studies to verify the effectiveness of several techniques used in ParaNet, including attention distillation, positional encoding, and stacking decoder layers to refine the attention alignment in a layer-by-layer manner. We evaluate the performance of a non-autoregressive ParaNet model trained without attention distillation and find that it fails to learn meaningful attention alignment. The synthesized audios are unintelligible and mostly pure noise. Similarly, we train another non-autoregressive ParaNet

model without adding positional encoding in the attention block. The resulting model only learns very blurry attention alignment and cannot synthesize intelligible speech. Finally, we train two non-autoregressive ParaNet models with 6 and 12 decoder layers, respectively, and compare them with the default non-autoregressive ParaNet model which has 17 decoder layers. We conduct the same attention error analysis on the 100-sentence test set and the results are shown in Table 3. We find that increasing the number of decoder layers for non-autoregressive ParaNet can reduce the total number of attention errors, in both cases with and without applying attention mask at synthesis.

6. Conclusion

In this work, we build a feed-forward neural TTS system by proposing a non-autoregressive text-to-spectrogram model. The proposed ParaNet obtains reasonably good speech quality and brings 46.7 times speed-up over its autoregressive counterpart at synthesis. We also compare various neural vocoders within the TTS system. Our results suggest that the parallel vocoder is generally less robust than WaveNet vocoder, when the front-end acoustic model is non-autoregressive. As a result, it is interesting to investigate small-footprint and robust parallel neural vocoder (e.g., WaveFlow) in future study.

References

Arık, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Raiman, J., Sengupta, S., and Shoeybi, M. Deep Voice: Real-time neural text-to-speech. In *ICML*, 2017a.

Arık, S. O., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., and Zhou, Y. Deep Voice 2: Multi-speaker neural text-to-speech. In *NIPS*, 2017b.

Arik, S. O., Chen, J., Peng, K., Ping, W., and Zhou, Y. Neural voice cloning with a few samples. In *NIPS*, 2018.

Arık, S. Ö., Jun, H., and Diamos, G. Fast spectrogram inversion using multi-head convolutional neural networks. *IEEE Signal Processing Letters*, 2019.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015.

Bińkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., Cobo, L. C., and Simonyan, K. High fidelity speech synthesis with adversarial networks. In *ICLR*, 2020.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, 2016.

Chen, Y., Assael, Y., Shillingford, B., Budden, D., Reed, S., Zen, H., Wang, Q., Cobo, L. C., Trask, A., Laurie, B., et al. Sample efficient adaptive text-to-speech. In *ICLR*, 2019.

Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pp. 577–585, 2015.

Chung, J., Kastner, K., Dinh, L., Goel, K., Courville, A. C., and Bengio, Y. A recurrent latent variable model for sequential data. In *NIPS*, 2015.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real NVP. In *ICLR*, 2017.

Gu, J., Bradbury, J., Xiong, C., Li, V. O., and Socher, R. Non-autoregressive neural machine translation. In *ICLR*, 2018.

Hsu, W.-N., Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Wang, Y., Cao, Y., Jia, Y., Chen, Z., Shen, J., et al. Hierarchical generative modeling for controllable speech synthesis. In *ICLR*, 2019.

Huang, C.-W., Ahmed, F., Kumar, K., Lacoste, A., and Courville, A. Probability distillation: A caveat and alternatives. In *UAI*, 2019.

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., and Moreno, I. L. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *NIPS*, 2018.

Kaiser, Ł., Roy, A., Vaswani, A., Pamar, N., Bengio, S., Uszkoreit, J., and Shazeer, N. Fast decoding in sequence models using discrete latent variables. In *ICML*, 2018.

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A. v. d., Dieleman, S., and Kavukcuoglu, K. Efficient neural audio synthesis. In *ICML*, 2018.

Kim, S., Lee, S.-g., Song, J., and Yoon, S. FloWaveNet: A generative flow for raw audio. In *ICML*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *NIPS*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *ICLR*, 2014.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. Improving variational inference with inverse autoregressive flow. In *NIPS*, 2016.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., and Courville, A. C. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Advances in Neural Information Processing Systems*, pp. 14910–14921, 2019.

Lee, J., Mansimov, E., and Cho, K. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *EMNLP*, 2018.

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M., and Zhou, M. Neural speech synthesis with transformer network. AAAI, 2019.

Nachmani, E., Polyak, A., Taigman, Y., and Wolf, L. Fitting new speakers based on a short untranscribed sample. In *ICML*, 2018.

Ping, W., Peng, K., and Chen, J. ClariNet: Parallel wave generation in end-to-end text-to-speech. *arXiv preprint arXiv:1807.07281*, 2018a.

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. Deep Voice 3: Scaling text-to-speech with convolutional sequence learning. In *ICLR*, 2018b.

Ping, W., Peng, K., Zhao, K., and Song, Z. WaveFlow: A compact flow-based model for raw audio. In *ICML*, 2020.

Prenger, R., Valle, R., and Catanzaro, B. WaveGlow: A flow-based generative network for speech synthesis. In *ICASSP*, 2019.

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. Fastspeech: Fast, robust and controllable text to speech. In *Advances in Neural Information Processing Systems*, pp. 3165–3174, 2019.

Rezende, D. J. and Mohamed, S. Variational inference with normalizing flows. In *ICML*, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

Ribeiro, F., Florêncio, D., Zhang, C., and Seltzer, M. Crowd-MOS: An approach for crowdsourcing mean opinion score studies. In *ICASSP*, 2011.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R., et al. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *ICASSP*, 2018.

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., and Bengio, Y. Char2wav: End-to-end speech synthesis. *ICLR workshop*, 2017.

Taigman, Y., Wolf, L., Polyak, A., and Nachmani, E. VoiceLoop: Voice fitting and synthesis via a phonological loop. In *ICLR*, 2018.

Taylor, P. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

van den Oord, A., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *NIPS*, 2017.

van den Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G. v. d., Lockhart, E., Cobo, L. C., Stimberg, F., et al. Parallel WaveNet: Fast high-fidelity speech synthesis. In *ICML*, 2018.

Wang, X., Takaki, S., and Yamagishi, J. Neural source-filter-based waveform model for statistical parametric speech synthesis. In *ICASSP*, 2019.

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., and Saurous, R. A. Tacotron: Towards end-to-end speech synthesis. In *Interspeech*, 2017.

Yamamoto, R., Song, E., and Kim, J.-M. Probability density distillation with generative adversarial networks for high-quality parallel waveform generation. *arXiv preprint arXiv:1904.04472*, 2019.

Appendix

A. 15-Sentence Test Set

The 15 sentences used to quantify the inference speed up in Table 2 are listed below (note that % corresponds to pause):

1. WHEN THE SUNLIGHT STRIKES RAINDROPS IN THE AIR%THEY ACT AS A PRISM AND FORM A RAINBOW%.
2. THESE TAKE THE SHAPE OF A LONG ROUND ARCH%WITH ITS PATH HIGH ABOVE%AND ITS TWO ENDS APPARENTLY BEYOND THE HORIZON%.
3. WHEN A MAN LOOKS FOR SOMETHING BEYOND HIS REACH%HIS FRIENDS SAY HE IS LOOKING FOR THE POT OF GOLD AT THE END OF THE RAINBOW%.
4. IF THE RED OF THE SECOND BOW FALLS UPON THE GREEN OF THE FIRST%THE RESULT IS TO GIVE A BOW WITH AN ABNORMALLY WIDE YELLOW BAND%.
5. THE ACTUAL PRIMARY RAINBOW OBSERVED IS SAID TO BE THE EFFECT OF SUPER IMPOSITION OF A NUMBER OF BOWS%.
6. THE DIFFERENCE IN THE RAINBOW DEPENDS CONSIDERABLY UPON THE SIZE OF THE DROPS%.
7. IN THIS PERSPECTIVE%WE HAVE REVIEWED SOME OF THE MANY WAYS IN WHICH NEUROSCIENCE HAS MADE FUNDAMENTAL CONTRIBUTIONS%.
8. IN ENHANCING AGENT CAPABILITIES%IT WILL BE IMPORTANT TO CONSIDER OTHER SALIENT PROPERTIES OF THIS PROCESS IN HUMANS%.
9. IN A WAY THAT COULD SUPPORT DISCOVERY OF SUBGOALS AND HIERARCHICAL PLANNING%.
10. DISTILLING INTELLIGENCE INTO AN ALGORITHMIC CONSTRUCT AND COMPARING IT TO THE HUMAN BRAIN MIGHT YIELD INSIGHTS%.
11. THE VAULT THAT WAS SEARCHED HAD IN FACT BEEN EMPTIED EARLIER THAT SAME DAY%.
12. ANT LIVES NEXT TO GRASSHOPPER%ANT SAYS%I LIKE TO WORK EVERY DAY%.
13. YOUR MEANS OF TRANSPORT FULFIL ECONOMIC REQUIREMENTS IN YOUR CHOSEN COUNTRY%.
14. SLEEP STILL FOGGED MY MIND AND ATTEMPTED TO FIGHT BACK THE PANIC%.
15. SUDDENLY%I SAW TWO FAST AND FURIOUS FEET DRIBBLING THE BALL TOWARDS MY GOAL%.

B. 100-Sentence Test Set

The 100 sentences used to quantify the results in Table 3 are listed below (note that % corresponds to pause):

1. A B C%.
2. X Y Z%.
3. HURRY%.
4. WAREHOUSE%.
5. REFERENDUM%.
6. IS IT FREE%?
7. JUSTIFIABLE%.
8. ENVIRONMENT%.
9. A DEBT RUNS%.
10. GRAVITATIONAL%.
11. CARDBOARD FILM%.
12. PERSON THINKING%.
13. PREPARED KILLER%.
14. AIRCRAFT TORTURE%.
15. ALLERGIC TROUSER%.
16. STRATEGIC CONDUCT%.
17. WORRYING LITERATURE%.
18. CHRISTMAS IS COMING%.
19. A PET DILEMMA THINKS%.
20. HOW WAS THE MATH TEST%?

21. GOOD TO THE LAST DROP%.
22. AN M B A AGENT LISTENS%.
23. A COMPROMISE DISAPPEARS%.
24. AN AXIS OF X Y OR Z FREEZES%.
25. SHE DID HER BEST TO HELP HIM%.
26. A BACKBONE CONTESTS THE CHAOS%.
27. TWO A GREATER THAN TWO N NINE%.
28. DON'T STEP ON THE BROKEN GLASS%.
29. A DAMNED FLIPS INTO THE PATIENT%.
30. A TRADE PURGES WITHIN THE B B C%.
31. I'D RATHER BE A BIRD THAN A FISH%.
32. I HEAR THAT NANCY IS VERY PRETTY%.
33. I WANT MORE DETAILED INFORMATION%.
34. PLEASE WAIT OUTSIDE OF THE HOUSE%.
35. N A S A EXPOSURE TUNES THE WAFFLE%.
36. A MIST DICTATES WITHIN THE MONSTER%.
37. A SKETCH ROPES THE MIDDLE CEREMONY%.
38. EVERY FAREWELL EXPLODES THE CAREER%.
39. SHE FOLDED HER HANDKERCHIEF NEATLY%.
40. AGAINST THE STEAM CHOOSES THE STUDIO%.
41. ROCK MUSIC APPROACHES AT HIGH VELOCITY%.
42. NINE ADAM BAYE STUDY ON THE TWO PIECES%.
43. AN UNFRIENDLY DECAY CONVEYS THE OUTCOME%.
44. ABSTRACTION IS OFTEN ONE FLOOR ABOVE YOU%.
45. A PLAYED LADY RANKS ANY PUBLICIZED PREVIEW%.
46. HE TOLD US A VERY EXCITING ADVENTURE STORY%.
47. ON AUGUST TWENTY EIGTH%MARY PLAYS THE PIANO%.
48. INTO A CONTROLLER BEAMS A CONCRETE TERRORIST%.
49. I OFTEN SEE THE TIME ELEVEN ELEVEN ON CLOCKS%.
50. IT WAS GETTING DARK%AND WE WEREN'T THERE YET%.
51. AGAINST EVERY RHYME STARVES A CHORAL APPARATUS%.
52. EVERYONE WAS BUSY%SO I WENT TO THE MOVIE ALONE%.
53. I CHECKED TO MAKE SURE THAT HE WAS STILL ALIVE%.
54. A DOMINANT VEGETARIAN SHIES AWAY FROM THE G O P%.
55. JOE MADE THE SUGAR COOKIES%SUSAN DECORATED THEM%.
56. I WANT TO BUY A ONESIE%BUT KNOW IT WON'T SUIT ME%.
57. A FORMER OVERRIDE OF Q W E R T Y OUTSIDE THE POPE%.
58. F B I SAYS THAT C I A SAYS%I'LL STAY AWAY FROM IT%.
59. ANY CLIMBING DISH LISTENS TO A CUMBERSOME FORMULA%.
60. SHE WROTE HIM A LONG LETTER%BUT HE DIDN'T READ IT%.
61. DEAR%BEAUTY IS IN THE HEAT NOT PHYSICAL%I LOVE YOU%.
62. AN APPEAL ON JANUARY FIFTH DUPLICATES A SHARP QUEEN%.
63. A FAREWELL SOLOS ON MARCH TWENTY THIRD SHAKES NORTH%.
64. HE RAN OUT OF MONEY%SO HE HAD TO STOP PLAYING POKER%.
65. FOR EXAMPLE%A NEWSPAPER HAS ONLY REGIONAL DISTRIBUTION T%.
66. I CURRENTLY HAVE FOUR WINDOWS OPEN UP%AND I DON'T KNOW WHY%.
67. NEXT TO MY INDIRECT VOCAL DECLINES EVERY UNBEARABLE ACADEMIC%.
68. OPPOSITE HER SOUNDING BAG IS A M C'S CONFIGURED THOROUGHFARE%.
69. FROM APRIL EIGHTH TO THE PRESENT%I ONLY SMOKE FOUR CIGARETTES%.
70. I WILL NEVER BE THIS YOUNG AGAIN%EVER%OH DAMN%I JUST GOT OLDER%.
71. A GENEROUS CONTINUUM OF AMAZON DOT COM IS THE CONFLICTING WORKER%.
72. SHE ADVISED HIM TO COME BACK AT ONCE%THE WIFE LECTURES THE BLAST%.
73. A SONG CAN MAKE OR RUIN A PERSON'S DAY IF THEY LET IT GET TO THEM%.
74. SHE DID NOT CHEAT ON THE TEST%FOR IT WAS NOT THE RIGHT THING TO DO%.

75. HE SAID HE WAS NOT THERE YESTERDAY%HOWEVER%MANY PEOPLE SAW HIM THERE%.

76. SHOULD WE START CLASS NOW%OR SHOULD WE WAIT FOR EVERYONE TO GET HERE%?

77. IF PURPLE PEOPLE EATERS ARE REAL%WHERE DO THEY FIND PURPLE PEOPLE TO EAT%?

78. ON NOVEMBER EIGHTEENTH EIGHTEEN TWENTY ONE%A GLITTERING GEM IS NOT ENOUGH%.

79. A ROCKET FROM SPACE X INTERACTS WITH THE INDIVIDUAL BENEATH THE SOFT FLAW%.

80. MALLS ARE GREAT PLACES TO SHOP%I CAN FIND EVERYTHING I NEED UNDER ONE ROOF%.

81. I THINK I WILL BUY THE RED CAR%OR I WILL LEASE THE BLUE ONE%THE FAITH NESTS%.

82. ITALY IS MY FAVORITE COUNTRY%IN FACT%I PLAN TO SPEND TWO WEEKS THERE NEXT YEAR%.

83. I WOULD HAVE GOTTEN W W W DOT GOOGLE DOT COM%BUT MY ATTENDANCE WASN'T GOOD ENOUGH%.

84. NINETEEN TWENTY IS WHEN WE ARE UNIQUE TOGETHER UNTIL WE REALISE%WE ARE ALL THE SAME%.

85. MY MUM TRIES TO BE COOL BY SAYING H T T P COLON SLASH SLASH W W W B A I D U DOT COM%.

86. HE TURNED IN THE RESEARCH PAPER ON FRIDAY%OTHERWISE%HE EMAILED A S D F AT YAHOO DOT ORG%.

87. SHE WORKS TWO JOBS TO MAKE ENDS MEET%AT LEAST%THAT WAS HER REASON FOR NOT HAVING TIME TO JOIN US%.

88. A REMARKABLE WELL PROMOTES THE ALPHABET INTO THE ADJUSTED LUCK%THE DRESS DODGES ACROSS MY ASSAULT%.

89. A B C D E F G H I J K L M N O P Q R S T U V W X Y Z ONE TWO THREE FOUR FIVE SIX SEVEN EIGHT NINE TEN%.

90. ACROSS THE WASTE PERSISTS THE WRONG PACIFIER%THE WASHED PASSENGER PARADES UNDER THE INCORRECT COMPUTER%.

91. IF THE EASTER BUNNY AND THE TOOTH FAIRY HAD BABIES WOULD THEY TAKE YOUR TEETH AND LEAVE CHOCOLATE FOR YOU%?

92. SOMETIMES%ALL YOU NEED TO DO IS COMPLETELY MAKE AN ASS OF YOURSELF AND LAUGH IT OFF TO REALISE THAT LIFE ISN'T SO BAD AFTER ALL%.

93. SHE BORROWED THE BOOK FROM HIM MANY YEARS AGO AND HASN'T YET RETURNED IT%WHY WON'T THE DISTINGUISHING LOVE JUMP WITH THE JUVENILE%?

94. LAST FRIDAY IN THREE WEEK'S TIME I SAW A SPOTTED STRIPED BLUE WORM SHAKE HANDS WITH A LEGLESS LIZARD%THE LAKE IS A LONG WAY FROM HERE%.

95. I WAS VERY PROUD OF MY NICKNAME THROUGHOUT HIGH SCHOOL BUT TODAY%I COULDN'T BE ANY DIFFERENT TO WHAT MY NICKNAME WAS%THE METAL LUSTS%THE RANGING CAPTAIN CHARTERS THE LINK%.

96. I AM HAPPY TO TAKE YOUR DONATION%ANY AMOUNT WILL BE GREATLY APPRECIATED%THE WAVES WERE CRASHING ON THE SHORE%IT WAS A LOVELY SIGHT%THE PARADOX STICKS THIS BOWL ON TOP OF A SPONTANEOUS TEA%.

97. A PURPLE PIG AND A GREEN DONKEY FLEW A KITE IN THE MIDDLE OF THE NIGHT AND ENDED UP SUNBURNT%THE CONTAINED ERROR POSES AS A LOGICAL TARGET%THE DIVORCE ATTACKS NEAR A MISSING DOOM%THE OPERA FINES THE DAILY EXAMINER INTO A MURDERER%.

98. AS THE MOST FAMOUS SINGLER-SONGWRITER%JAY CHOU GAVE A PERFECT PERFORMANCE IN BEIJING ON MAY TWENTY FOURTH%TWENTY FIFTH%AND TWENTY SIXTH TWENTY THREE ALL THE FANS THOUGHT HIGHLY OF HIM AND TOOK PRIDE IN HIM ALL THE TICKETS WERE SOLD OUT%.

99. IF YOU LIKE TUNA AND TOMATO SAUCE%TRY COMBINING THE TWO%IT'S REALLY NOT AS BAD AS IT SOUNDS%THE BODY MAY PERHAPS COMPENSATES FOR THE LOSS OF A TRUE METAPHYSICS%THE CLOCK WITHIN THIS BLOG AND THE CLOCK ON MY LAPTOP ARE ONE HOUR DIFFERENT FROM EACH OTHER%.

100. SOMEONE I KNOW RECENTLY COMBINED MAPLE SYRUP AND BUTTERED POPCORN THINKING IT WOULD TASTE LIKE CARAMEL POPCORN%IT DIDN'T AND THEY DON'T RECOMMEND ANYONE ELSE DO IT EITHER%THE GENTLEMAN MARCHES AROUND THE PRINCIPAL%THE DIVORCE ATTACKS NEAR A MISSING DOOM%THE COLOR MISPRINTS A CIRCULAR WORRY ACROSS THE CONTROVERSY%.