WaveTTS: Tacotron-based TTS with Joint Time-Frequency Domain Loss

*Rui Liu[1,2] , Berrak Sisman[2,3], Feilong Bao[*1], Guanglai Gao[1], Haizhou Li[2]*

[1] Inner Mongolia University, China
[2] Dept. of Electrical and Computer Engineering, National University of Singapore, Singapore
[3] Information Systems Technology and Design, Singapore University of Technology and Design

{r.liu, berraksisman}@u.nus.edu, {csfeilong, csggl}@imu.edu.cn, haizhou.li@nus.edu.sg

Abstract

Tacotron-based text-to-speech (TTS) systems directly synthesize speech from text input. Such frameworks typically consist of a feature prediction network that maps character sequences to frequency-domain acoustic features, followed by a waveform reconstruction algorithm or a neural vocoder that generates the time-domain waveform from acoustic features. As the loss function is usually calculated only for frequency-domain acoustic features, that doesn't directly control the quality of the generated time-domain waveform. To address this problem, we propose a new training scheme for Tacotron-based TTS, referred to as WaveTTS, that has 2 loss functions: 1) time-domain loss, denoted as the waveform loss, that measures the distortion between the natural and generated waveform; and 2) frequency-domain loss, that measures the Mel-scale acoustic feature loss between the natural and generated acoustic features. WaveTTS ensures both the quality of the acoustic features and the resulting speech waveform. To our best knowledge, this is the first implementation of Tacotron with joint time-frequency domain loss. Experimental results show that the proposed framework outperforms the baselines and achieves high-quality synthesized speech.

1. Introduction

In recent years, Text-to-Speech (TTS) technology has advanced by leaps and bounds, from concatenative approach [1–3] to statistical parametric approach [4–7], and to deep learning [8]. They produce high quality and natural speech that rival human vocal production [4, 9]. TTS is also widely used in human-machine communications, such as robotics, call centres, games, entertainments, and healthcare applications.

With the advent of deep learning, neural approaches to TTS become mainstream, such as Tacotron [8], Tacotron2 [10] and its varieties [11–16]. The end-to-end TTS model is based on encoder-decoder framework, which has been widely adopted for sequence generation tasks, such as speech recognition [17–20], image translation [21] and neural machine translation [22, 23]. Tacotron-based TTS typically consists of two modules: 1) feature prediction, and 2) waveform generation. The main task of feature prediction network is to obtain frequency-domain acoustic features, while the waveform generation module is to convert frequency-domain acoustic features into time-domain waveform.

A typical Tacotron implementation adopts Griffin-Lim algorithm [24, 25] for phase reconstruction, that only uses a loss function derived from amplitude spectrogram in frequency domain. Such a loss function doesn't take the resulting

---

Speech Samples: https://ttslr.github.io/WaveTTS/.

waveform into consideration in the optimization process. As a result, there exists a mismatch between the Tacotron optimization and the expected waveform. We note that such mismatch also exists in many other speech processing tasks, such as speech separation [26], where we observe that, by incorporating time-domain loss function [27], one can improve the output speech quality. More recently, deep learning approach to speech enhancement methods with time-domain raw waveform outputs [28, 29] have also been investigated. However, we note that time-domain loss function has not been well explored in speech synthesis, which will be the focus of this paper.

Tacotron2 [10] has been proposed to achieve high quality synthesized voice. It addresses the waveform optimization problem by using WaveNet-based neural vocoder [30–34]. We note that WaveNet avoids the artifacts and deterioration caused by deterministic vocoders. It generates time-domain waveform samples conditioned on the predicted mel-spectrum features. Although Tacotron2 allows end-to-end learning of TTS directly from character sequences and speech waveforms, its feature prediction network is trained independently of the WaveNet vocoder. At run-time, the feature prediction network and WaveNet vocoder are artificially joined together. As a result, the framework suffers from the mismatch between frequency-domain acoustic features and time-domain waveform. It is reported that the samples generated from WaveNet occasionally become unstable, especially when less accurately predicted acoustic features are used as the local condition parameters [35, 36]. To overcome such mismatch, we propose to use joint time-frequency domain loss for TTS that effectively improves the synthesized voice quality.

In this paper, we propose to add a time-domain loss function to the Griffin-Lim/ISTFT output of Tacotron-based TTS model at the training time. In other words, we use both frequency-domain loss and time-domain loss for the training of feature prediction model. We hypothesize that the feature prediction network will compensate the possible artifacts that Griffin-Lim process may introduce under the supervision of the time-domain loss. We use Griffin-Lim iteration followed by ISTFT to transform frequency-domain feature to time-domain waveform and use scale-invariant signal-to-distortion (SI-SDR) [37, 38] to measure the quality of the time-domain waveform. Our proposed idea shares a similar motivation with [39] in terms of the use of waveform loss. However, it differs from [39] in many ways, for example, we study Tacotron-based TTS, while [39] mostly deals with Wasserstein GAN-based TTS.

The main contributions of this paper include: 1) we study the use of time-domain loss for speech synthesis; 2) we improve Tacotron-based TTS framework by proposing a new training

scheme based on joint time-frequency domain loss; and 3) we propose to use SI-SDR metric to measure the distortion of time-domain waveform. The novel training scheme optimizes the frequency-domain acoustic features in a way that it leads to better time-domain waveform. To our best knowledge, this is the first implementation of joint training scheme on frequency and time domain for Tacotron-based TTS framework.

This paper is organized as follows: In Section 2, we present the Tacotron-based baseline TTS system. In Section 3, we present the novel idea of joint time-frequency domain loss, and formulate the training and run-time processes. We report the experimental results in Section 4. Section 5 concludes the study.

2. Baseline: Tacotron-based TTS

In this paper, we use a Tacotron-based framework [10] as a reference baseline. We illustrate the overall architecture of the *reference baseline* in Figure 1, that includes feature prediction model which contains encoder, attention-based decoder and Griffin-Lim algorithm for waveform reconstruction. The encoder (blue box in Figure 1) consists of two components, a CNN-based module that has 3 convolution layers, and a LSTM-based module that has a bidirectional LSTM layer [40–44]. The decoder (pale yellow box in Figure 1) consists of four components: a 2-layer pre-net, 2 LSTM layers, a linear projection layer and a 5-convolution-layer post-net. The decoder is a standard autoregressive recurrent neural network that generates the mel-spectrum features and stop tokens frame by frame.

During training, we optimize the feature prediction model to minimized the frequency-domain loss between the generated mel-spectrum features ($\hat{y}$) and the target mel-spectrum features (y).

3. WaveTTS

In this section, we study the use of a newly proposed time-domain loss function for Tacotron-based TTS. By applying a new training strategy that takes into account both time and frequency domain loss functions, we effectively reduce the mismatch between the frequency-domain features and the time-domain waveform, and improve the output speech quality. In addition, Griffin-Lim algorithm and SI-SDR metric are utilized to realize the calculation process of proposed loss term. The proposed framework is called as *WaveTTS* hereafter.

3.1. Time-domain and Frequency-domain Loss Functions

In WaveTTS, we define two objective functions during training: 1) frequency-domain loss, denoted as $Loss_F$, that is calculated with the mel-spectrum features in a similar way described in [10]; and 2) the proposed time-domain loss, denoted as $Loss_T$, that is obtained at waveform level at the output of Griffin-Lim iteration, that estimates time-domain signal from the mel-spectrum features. The two objective functions are illustrated in Figure 2.

The entire process can be formulated as follows. The encoder takes the character sequence $x = (x_1, x_2, ..., x_T)$ as input and converts the one-hot vector to continuous features representation h:

$$h_t = \text{Encoder}(h_{t-1}, x_t) \tag{1}$$

The decoder outputs a mel-spectrum feature $\hat{y}_t$ at each step t:

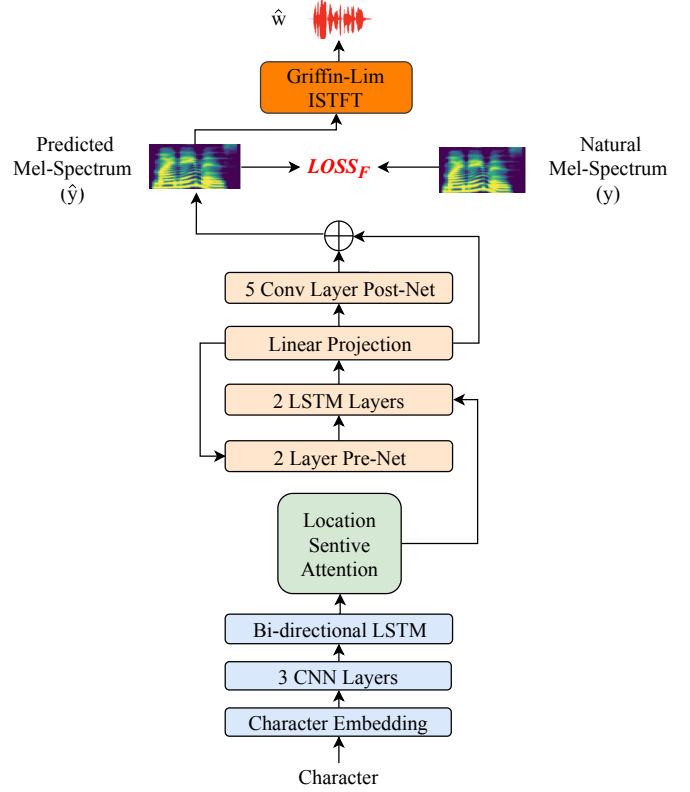$$\hat{y}_t = \text{Decoder}(y_{t-1}, \sigma(h_t)) \tag{2}$$

Figure 1: Block diagram of Tacotron-based reference baseline that has three modules, encoder, attention-based decoder, and Griffin-Lim reconstruction algorithm.

where $\sigma()$ represents a function to calculate the context vector by using location-sensitive attention mechanism [45].

We first calculate $Loss_F$ in a similar way as that in [10]. $Loss_F$ ensures that the generated mel-spectrum is close the natural mel-spectrum. $Loss_F$ is given as follows,

$$Loss_F(\hat{y}, y) = \sum_{t=1}^{T'} L_2(\hat{y}_t, y_t) \tag{3}$$

where T' is the total number of the sequences in training data. L_2 loss function is used to minimize the error which is the sum of the all the squared differences between the true value and the predicted value.

We then propose the use of a time-domain loss $Loss_T$, that is applied to the Griffin-Lim output. This is to reduce the mismatch between the optimized frequency-domain output and the actual time-domain waveform [39]. The implementation details of $Loss_T$ will be explained in Section 3.2.

Overall, the proposed WaveTTS framework has two loss functions. $Loss_F$ minimizes the loss between converted and original mel-spectrum, while $Loss_T$ minimizes this loss at waveform level. We add a weighting coefficient λ to balance the two losses. The training criterion of the whole model is defined as:

$$Loss = Loss_F + \lambda \cdot Loss_T \tag{4}$$

Algorithm 1 also shows the complete training process of our proposed WaveTTS. WaveTTS model predicts the mel-spectrum features $\hat{y}$ from the given input character sequence $x = (x_1, x_2, ..., x_T)$, and then converts the estimated and target mel-spectrum to the time-domain signal $\hat{w}, w$ using Griffin-Lim based ISTFT algorithm (blue content in Algorithm 1). Finally,

the joint loss function given in Equation 4 is used to optimize the WaveTTS model.

3.2. Implementation of time-domain Loss

3.2.1. Time-domain loss

We adopt Griffin-Lim algorithm [24], followed by ISTFT to generate the time-domain waveform. Griffin-Lim algorithm has been widely used in speech synthesis [10, 46] for its simplicity, that can be formulated as follows:

$$\hat{A} = \varepsilon(\hat{y}) \quad (5)$$

$$\hat{\mathbf{X}} = \text{Griffin–Lim}(\hat{A}) \quad (6)$$

where $\hat{y}$ represents the predicted mel-spectrum sequences, and $\hat{A}$ represents their amplitude. ε is the function that calculates the amplitude of the given input mel-spectrum sequences, which is followed by the Griffin-Lim algorithm, that estimates a complex valued spectrum, while minimizing the change to the input amplitude $\hat{A}$. ISTFT transforms the estimated complex valued spectrum to time-domain signals. The details of the Griffin-Lim algorithm are given in Algorithm 2. P_S is the metric projection onto a set S. Here, C is the set of consistent spectrums, and A is the set of spectrums whose amplitude is the same as the given one.

It's worth mentioning that the Griffin-Lim algorithm usually requires many iterations, as shown in Algorithm 2, at run-time to obtain a high-quality audio signal. It is an optimization process independent of Tacotron training. We would like the Tacotron feature prediction network to generate acoustic features, that not only are close to those of natural speech in frequency-domain, but also allow Griffin-Lim to produce speech that is close to natural speech in time-domain.

Let's denote the predicted and the original mel-spectrum as $(\hat{y}, y)$. We apply Griffin-Lim and ISTFT to generate $(\hat{w}, w)$, with $\hat{y} \rightarrow \hat{w}$ and $y \rightarrow w$. We keep the same number of Griffin-Lim iterations to ensure that Griffin-Lim behaves the same between the two transformation pairs, $\hat{y} \rightarrow \hat{w}$ and $y \rightarrow w$. We measure the distortion between $\hat{w}$ and w with a time-domain loss, that forces the speech waveform generated from the predicted network to be as close as possible to that generated from the mel-spectrum of natural speech.

3.2.2. Scale-Invariant Signal-to-Distortion (SI-SDR)

In speech synthesis, we optimize the feature prediction network to minimize the discrepancy between the synthesized waveform and the target natural speech that is supervised by a loss function. We propose a time-domain loss function, $Loss_T$, that is based on scale-invariant signal-to-distortion (SI-SDR). SI-SDR has been introduced as a time-domain objective measure in source separation [47–49] to compare two time-domain speech signals. We adopt SI-SDR to measure the discrepancy between the generated waveform and the target natural speech. To our best knowledge, this is the first implementation of SI-SDR for time-domain loss calculation to improve TTS quality.

We note that SI-SDR is evaluated only during training, and not required at run-time inference. During training, the predicted time-domain waveform $\hat{w}$ and the target speech w have identical duration. Similarly, the predicted mel-spectrum $\hat{y}$ and target mel-spectrum y also share the same frame length, that facilitates the SI-SDR calculation. As a greater SI-SDR indicates better quality, to turn it into a loss function, we take the negative value of SI-SDR as the loss function,
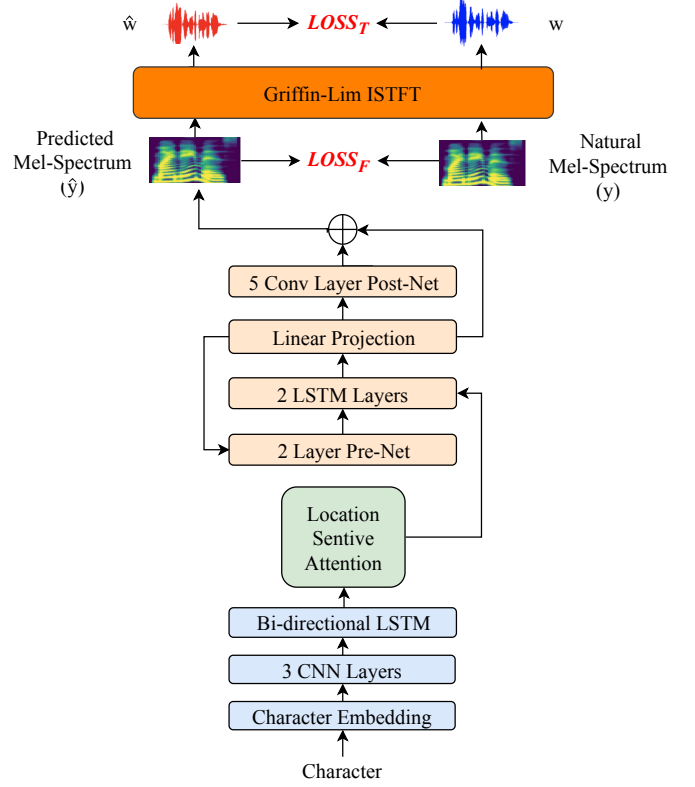

Figure 2: Block diagram of proposed WaveTTS model using both frequency-domain loss ($Loss_F$) and time-domain loss ($Loss_T$).

$$Loss_T(\hat{w}, w) = -10 \log_{10} \frac{\|\alpha w\|^2}{\|\alpha w - \hat{w}\|^2} \quad (7)$$

where

$$\alpha = \frac{\hat{w}^T w}{\|w\|^2} = \underset{\alpha}{\text{argmin}} \|\alpha w - \hat{w}\|^2 \quad (8)$$

SI-SDR is expressed in decibel (dB) and defined in the range of $[-\infty, \infty]$, so is $Loss_T$.

4. Experiments

We report the TTS experiments on LJSpeech database [1], which consists of 13,100 short clips with a total of nearly 24 hours of speech from one single speaker reading about 7 non-fiction books. We develop four systems for a comparative study:

- *Tacotron-GL*: Tacotron-based baseline model [10], that has only frequency-domain loss function. Griffin-Lim algorithm is used to generate the waveform at run-time.

- *Tacotron-WN*: Tacotron-based baseline model [10], that has only frequency-domain loss function. Pre-trained WaveNet vocoder is used to generate the waveform at run-time.

- *WaveTTS-GL*: proposed WaveTTS model is trained with joint time-frequency domain loss. Griffin-Lim algorithm is used during training and run-time phases.

- *WaveTTS-WN*: proposed WaveTTS model is trained with joint time-frequency domain loss. Griffin-Lim algorithm is used during training and the pre-trained WaveNet vocoder is used to synthesize speech at run-time.

[1] https://keithito.com/LJ-Speech-Dataset/

Algorithm 1: Training WaveTTS model using time-frequency domain loss.

Input:
    Training set: $D = \{x, y, w\}$
    x: character sequence
    y: frequency-domain acoustic feature sequence
Output:
    Θ: TTS model
Hyper-parameter:
    N: epoch number
    n: batch size
    η: learning rate
    λ: weight for time-domain loss
Begin
1: Initialize TTS model Θ
2: **for** epoch = 1,2,...N **do**
3:    **for** batch = 1,2,...n **do**
4:       output $\hat{y}$:
          $\hat{y} = \Theta(x)$
5:       generate $\hat{w}$ and w using the Griffin-Lim
        algorithm:
          $\hat{A} = \varepsilon(\hat{y})$
          $A = \varepsilon(y)$
          $\hat{\mathbf{X}} = \text{Griffin–Lim}(\hat{A})$
          $\mathbf{X} = \text{Griffin–Lim}(A)$
          $\hat{w} = \text{ISTFT}(\hat{\mathbf{X}})$
          $w = \text{ISTFT}(\mathbf{X})$
6:       update Θ with $Loss_F$ and $Loss_T$:
          $\Theta \leftarrow \bigtriangledown_\Theta (Loss_F(y, \hat{y}) + \lambda \cdot Loss_T(w, \hat{w}))$
7:    **end for**
8: **end for**
9: return Θ
End

Algorithm 2: Griffin-Lim algorithm.

Input:
    $\angle c_0$: the initial phase
    A: amplitude
Output:
    $\mathbf{X}$: complex valued spectrum
Begin
1: Initialize $\mathbf{X}_0 = A \cdot e^{\cdot i \angle c_0}$
2: **for** iteration = 1,2,...n **do**
    $\mathbf{X}_n = P_\mathcal{C}(P_\mathcal{A}(\mathbf{X}_{n-1}))$
3: **end for**
4: return $\mathbf{X}_n$
End

We also compare these systems with the ground truth speech, denoted as *GT*. The comparison of the systems is also summarized in Table 1.

4.1. Experimental Setup

The 80-channel mel-spectrum is extracted with 12.5ms frame shift and 50ms frame length. It is normalized to zero-mean and unit-variance as the reference target. The decoder predicts only one non-overlapping output frame at each decoding step. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 10^{-3} exponentially decaying to 10^{-5} starting with 50k iterations. We also apply L_2 regularization with weight 10^{-6}. Hyper-parameter λ in Equation 4 is empirically set as 10^{-3}. All models are trained with a batch size of 32. The final models are trained with 100k steps for all systems. At run-time, Tacotron-GL and WaveTTS-GL use Griffin-Lim algorithm with 64 iterations, while Tacotron-WN and WaveTTS-WN use pretrained WaveNet vocoder.

4.2. Subjective Evaluation

We conduct listening experiments to evaluate the quality of the synthesized speech. We first evaluate the sound quality of the synthesized speech in terms of mean opinion score (MOS) among GT, Tacotron-GL, Tacotron-WN and the proposed WaveTTS-GL and WaveTTS-WN, that is reported in Figure 3.

The listeners rate of the quality is on a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. The MOS values reported in Figure 3 are calculated by taking the arithmetic average of all scores assigned the subjects who have passed the validation question test. We keep the linguistic content the same among different models so as to exclude other interference factors. 15 subjects participate in these experiments, and each one of them listens to 120 synthesized speech samples. We have three observations through the experiments:

1. The importance of joint time-frequency domain loss: We compare Tacotron-GL and WaveTTS-GL to observe the effect of joint time-frequency domain loss. We believe that this is a fair comparison as both frameworks use Griffin-Lim algorithm for waveform generation during training and/or run-time. As can be seen in Figure 3, WaveTTS-GL outperforms Tacotron-GL by achieving 3.30 MOS value, while Tacotron-GL achieves only 3.18.

2. The performance of WaveTTS with a neural vocoder at run-time: We compare Tacotron-WN and WaveTTS-WN to investigate how well the predicted mel-spectrum features perform with WaveNet vocoder. We observe that even though WaveTTS is trained with Griffin-Lim algorithm, it performs better than Tacotron when WaveNet vocoder is available at run-time. This shows how well our proposed WaveTTS performs with other neural vocoders.

3. Griffin-Lim vs WaveNet vocoder at run-time: We compare WaveTTS-GL and WaveTTS-WN in terms of voice quality. We note that both frameworks are trained under the same conditions. However, WaveTTS-WN uses WaveNet vocoder for waveform generation at run-time. As expected, WaveTTS-WN outperforms WaveTTS-GL.

We also conduct A/B preference tests to assess speech quality of proposed frameworks. In A/B preference tests, the listeners are asked to compare the quality and naturalness of the synthesized speech samples from different systems, and select the better one. 15 listeners were invited to participate in all the tests. 80 samples were randomly selected from 200 converted samples from each system. Figure 4 shows the speech

		Training Phase		Run-time Inference
		Loss Function	**Waveform Generation**	**Waveform Generation**
Baseline	**Tacotron-GL**	$Loss_F$	NA	Griffin-Lim [24]
	Tacotron-WN	$Loss_F$	NA	WaveNet vocoder [34]
Proposed	**WaveTTS-GL**	$Loss_F + \lambda \cdot Loss_T$	Griffin-Lim [24]	Griffin-Lim [24]
	WaveTTS-WN	$Loss_F + \lambda \cdot Loss_T$	Griffin-Lim [24]	WaveNet vocoder [34]

Table 1: Comparison of the frameworks in terms of loss function and waveform generation during training and run-time.
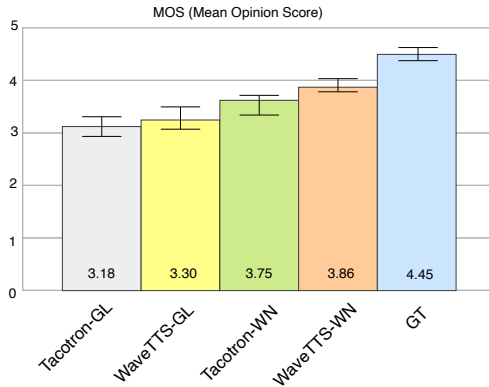
Figure 3: Comparison of mean opinion scores (MOS) between Tacotron-GL, WaveTTS-GL, Tacotron-WN, WaveTTS and ground truth (GT), with 95% confidence interval.
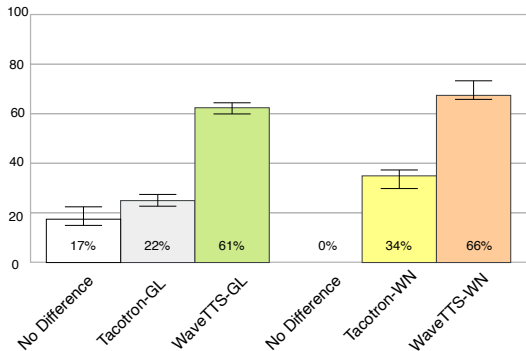

Figure 4: The preference test between (1) Tacotron-GL and WaveTTS-GL, and (2) Tacotron-WN and WaveTTS-WN, with 95% confidence interval.

quality test results, which suggests that our proposed WaveTTS framework outperforms the baseline system for both Griffin-Lim and WaveNet vocoder settings at run-time.

We further conduct another A/B preference test to examine the effect of the number of Griffin-Lim iterations on the WaveTTS performance. To calculate the time-domain loss, WaveTTS needs to generate the synthesized waveform during training. For rapid turn-around, we only apply 1 and 2 Griffin-Lim iterations for phase reconstruction, and investigate the effect in terms of voice quality. Figure 5 shows A/B preference test results on both WaveTTS-GL and WaveTTS-WN. We observe that the single iteration of Griffin-Lim algorithm presents a better performance than 2 iterations.

5. Conclusion

In this paper, we propose a new Tacotron implementation, called WaveTTS. The traditional framework calculates only frequency-domain loss to update the network parameters, that
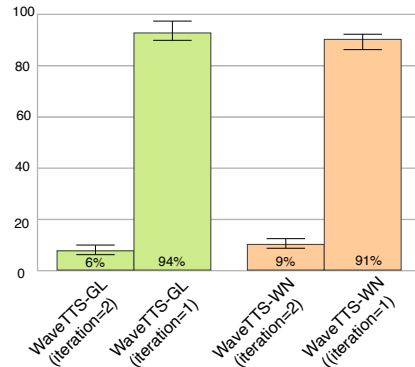

Figure 5: The preference test between (1) WaveTTS-GL (2 iterations) and WaveTTS-GL (1 iteration), and (2) WaveTTS-WN (2 iterations) and WaveTTS-WN (1 iteration), with 95% confidence interval.

doesn't directly control the quality of the generated time-domain waveform. The proposed WaveTTS is unique in a sense that it calculates both time-domain and frequency-domain loss, and optimizes the model for generating high-quality synthesized voice. We propose to use scale-invariant signal-to-distortion (SI-SDR) as the loss function. Even though the proposed model is trained with Griffin-Lim algorithm for time-domain loss calculation, it performs remarkable well with both Griffin-Lim and WaveNet vocoder at run-time. Experimental results show that the proposed framework outperforms the baselines and achieves high-quality synthesized speech. To our best knowledge, this is the first implementation of Tacotron-based TTS model with joint time-frequency domain loss.

As a future work, we will investigate the training phase of joint time-frequency domain loss with a neural vocoder for high-quality TTS.

6. Acknowledgements

7. References

[1] Andrew J Hunt and Alan W Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE, 1996, vol. 1, pp. 373–376.

[2] Heiga Zen, Andrew Senior, and Mike Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 ieee international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7962–7966.

[3] Thomas Merritt, Robert AJ Clark, Zhizheng Wu, Junichi Yamagishi, and Simon King, "Deep neural network-guided unit selection synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5145–5149.

[4] Heiga Zen, Keiichi Tokuda, and Alan W Black, "Statistical parametric speech synthesis," *speech communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[5] Keiichi Tokuda, Yoshihiko Nankaku, Tomoki Toda, Heiga Zen, Junichi Yamagishi, and Keiichiro Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[6] Heiga Zen and Haşim Sak, "Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4470–4474.

[7] Zhizheng Wu and Simon King, "Investigating gated recurrent networks for speech synthesis," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5140–5144.

[8] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: A fully end-to-end text-to-speech synthesis model," in *INTERSPEECH*, 2017, pp. 4006–4010.

[9] Paul Taylor, *Text-to-speech synthesis*, Cambridge university press, 2009.

[10] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[11] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *Proceedings of the 35th International Conference on Machine Learning. PMLR*, vol. 80, pp. 4693–4702, 2018.

[12] Wei-Ning Hsu, Yu Zhang, Ron Weiss, Heiga Zen, Yonghui Wu, Yuan Cao, and Yuxuan Wang, "Hierarchical generative modeling for controllable speech synthesis," in *International Conference on Learning Representations (ICLR)*, 2019.

[13] Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, David Kao, and Tom Bagby, "Semi-supervised generative modeling for controllable speech synthesis," in *International Conference on Learning Representations (ICLR)*, 2020.

[14] Rui Liu, Feilong Bao, Guanglai Gao, and Yonghe Wang, "Mongolian text-to-speech system based on deep neural network," in *National Conference on Man-Machine Speech Communication*. Springer, 2017, pp. 99–108.

[15] Rui Liu, Berrak Sisman, Jingdong Li, Feilong Bao, Guanglai Gao, and Haizhou Li, "Teacher-student training for robust tacotron-based tts," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[16] Jingdong Li, Hui Zhang, Rui Liu, Xueliang Zhang, and Feilong Bao, "End-to-end mongolian text-to-speech system," in *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2018, pp. 483–487.

[17] Alex Graves and Navdeep Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[18] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio, "End-to-end attention-based large vocabulary speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4945–4949.

[19] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[20] Zhehuai Chen, Mahaveer Jain, Yongqiang Wang, Michael L Seltzer, and Christian Fuegen, "End-to-end contextual speech recognition using class language models and a token passing decoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6186–6190.

[21] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim, "Attribute manipulation generative adversarial networks for fashion images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10541–10550.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2014.

[23] Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al., "Google's multilingual neural machine translation system: Enabling zero-shot translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.

[24] Daniel Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[25] Yoshiki Masuyama, Kohei Yatabe, Yuma Koizumi, Yasuhiro Oikawa, and Noboru Harada, "Deep griffin–lim iteration," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 61–65.

[26] DeLiang Wang and Jitong Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[27] Yuxuan Wang and DeLiang Wang, "A deep neural network for time-domain signal reconstruction," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4390–4394.

[28] Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 006–012.

[29] Chang-Le Liu, Szu-Wei Fu, You-Jin Lee, Yu Tsao, Jen-Wei Huang, and Hsin-Min Wang, "Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks," *arXiv preprint arXiv:1909.11909*, 2019.

[30] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," in *2016 ISCA Speech Synthesis Workshop*, 2016, pp. 125–133.

[31] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 27, no. 6, pp. 1085–1097, 2019.

[32] Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura, "Adaptive wavenet vocoder for residual compensation in gan-based voice conversion," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 282–289.

[33] Berrak Sisman, Mingyang Zhang, and Haizhou Li, "A voice conversion framework with tandem feature sparse representation and speaker-adapted wavenet vocoder," in *Proc. Interspeech 2018*, 2018, pp. 1978–1982.

[34] Tomoki Hayashi, Akira Tamamori, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "An investigation of multi-speaker training for wavenet vocoder," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 712–718.

[35] Berrak Sisman, *Machine Learning for Limited Data Voice Conversion*, Ph.D. thesis, 2019.

[36] Patrick Lumban Tobing, Yi-Chiao Wu, Tomoki Hayashi, Kazuhiro Kobayashi, and Tomoki Toda, "Voice conversion with cyclic recurrent neural network and fine-tuned wavenet vocoder," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6815–6819.

[37] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[38] Morten Kolbæk, Zheng-Hua Tan, Søren Holdt Jensen, and Jesper Jensen, "On loss functions for supervised monaural time-domain speech enhancement," *arXiv preprint arXiv:1909.01019*, 2019.

[39] Yi Zhao, Shinji Takaki, Hieu-Thi Luong, Junichi Yamagishi, Daisuke Saito, and Nobuaki Minematsu, "Wasserstein gan and waveform loss-based acoustic model training for multi-speaker text-to-speech synthesis systems using a wavenet vocoder," *IEEE Access*, vol. 6, pp. 60478–60488, 2018.

[40] Rui Liu, Jingdong Li, Feilong Bao, and Guanglai Gao, "The imu speech synthesis entry for blizzard challenge 2019," *Proceedings of Blizzard Challenge*, vol. 2019, 2019.

[41] Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang, and Yonghe Wang, "Improving mongolian phrase break prediction by using syllable and morphological embeddings with bilstm model.," in *Interspeech*, 2018, pp. 57–61.

[42] Rui Liu, Feilong Bao, Guanglai Gao, Hui Zhang, and Yonghe Wang, "A lstm approach with sub-word embeddings for mongolian phrase break prediction," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2448–2455.

[43] Rui Liu, FeiLong Bao, Guanglai Gao, Hui Zhang, and Yonghe Wang, "Phonologically aware bilstm model for mongolian phrase break prediction with attention mechanism," in *Pacific Rim International Conference on Artificial Intelligence*. Springer, 2018, pp. 217–231.

[44] Rui Liu, Feilong Bao, and Guanglai Gao, "Building mongolian tts front-end with encoder-decoder model by using bridge method and multi-view features," in *International Conference on Neural Information Processing*. Springer, 2019, pp. 642–651.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[46] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura, "Vqvae unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019," *Proc. Interspeech 2019*, pp. 1118–1122, 2019.

[47] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 696–700.

[48] Fahimeh Bahmaninezhad, Jian Wu, Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, Meng Yu, and Dong Yu, "A comprehensive study of speech separation: spectrogram vs waveform separation," *Proc. Interspeech 2019*, 2019.

[49] Shrikant Venkataramani, Ryley Higa, and Paris Smaragdis, "Performance based cost functions for end-to-end speech separation," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 350–355.