Thanks all the reviewers for the comments and suggestions!

To Reviewer #1

- **About 2-head attention** Compared with the attention alignments in machine translation, the attention in TTS is monotonous and much simpler than machine translation. Therefore, we choose fewer attention heads. Our preliminary experiments also show that more heads do not bring much difference to Transformer TTS. We will report the numbers in the new version.
- **Generated v.s. Groundtruth mel-spectrogram** Knowledge distillation is widely used in non-autoregressive machine translation [1] and speech synthesis [2] for transfer knowledge from the autoregressive teacher model to the non-autoregressive student model. The intuitive explanation is that the teacher model can generate data with smoother distribution and less noise, which is easy to be fitted by the student model [1][2]. The ablation study in Table 4 demonstrates the effectiveness of knowledge distillation.
- **Unmatched hyperparams** Our FastSpeech model differs from the original Transformer TTS model in that we use Conv1D instead of the original dense network after multi-head attention. So in our experiment, FastSpeech model is initialized from the teacher model with the same configuration, but not the original Transformer TTS model as shown in Appendix A. We will explicitly point out this in the new version of the paper.
- **Missing hyperparams in Appendix** Thanks for your reminder. We will add the missing hyperparams of pre-net and post-net in the new version of paper. For Transformer TTS, the hidden dimension and the number of layers of CNN are 512, 5 for post-net, and are 512, 3 for pre-net. For FastSpeech, the hidden dimension of the decoder's final linear layer is 80.
- **Batch size for inference** The batch size is 1 for inference evaluation, in order to simulate the scenario of online production. Many previous works (e.g., [1][2][3]) use this batch size to evaluate the inference latency.
- **About the inconsistent latency numbers** There is a typo in Table 2. The unit of latency should be "second". We will fix it in the new version of the paper.
- **Robustness test for Tacotron 2** We evaluate the robustness of Tacotron 2 on the 50 hard sentences. The repeating, skipping and error sentences are 4, 11 and 12 respectively, and the error rate is 24%. We will add the results in the new version of the paper.
- **The reference for CMOS evaluation** We will add the reference for CMOS in the new version of the paper.

[1] Gu, Jiatao, et al. "Non-autoregressive neural machine translation." ICLR 2018.

[2] Oord, Aaron van den, et al. "Parallel wavenet: Fast high-fidelity speech synthesis." ICML 2018.

[3] Prenger, Ryan, et al. "Waveglow: A flow-based generative network for speech synthesis." ICASSP 2019.

To Reviewer #2

- **About pronunciation dictionary** Our grapheme-to-phoneme tool works like this: it first looks up the pronunciation dictionary, and if the dictionary does not contain the word, it predicts the phoneme using a grapheme-to-phoneme model.
- **Enforce monotonicity** Thanks for your suggestion. We observe from the experiments that there are roughly two kinds of attention in Transformer TTS: diagonal and non-diagonal. If we enforce all the attentions to be diagonal, the performance of Transformer TTS will drop, because the non-diagonal attentions are also very helpful to the model. So we select the diagonal attentions rather than enforce all attentions to be monotonic.
- **Add the citation** We will add this citation in the new version of the paper.

To Reviewer #3

- **About the hard alignments** We've tried to add Gaussian smoothing to soften the inputs of the FFT block in the mel side. However, it doesn't improve the performance. We think the Conv1D and self-attention can mix the information from neighbors and play a role in softening the hard copied hidden states.
- **About diagonal alignments** We visualize the attention of each head in Transformer TTS and find that nearly half of the attentions are always diagonal for each data pair. We also find that if an attention has high focus rate, it is always diagonal. We find that some errors are closely related to the poor attention: jumping attention causes skipping and non-monotonic attention causes repeating. We think that diagonal attentions can ensure correct alignments in transformer TTS. We could not find any cases that have no diagonal attention at all.
- **About the title** Thanks for your advice! We will change the title accordingly.