Learning to Auto-correct for High-quality Spectrograms

A New NAR-TTS Architecture Named ReActSpeech

Zhiyang Zhou & Shihui Liu

Beijing Bombax XiaoIce Technology Co., Ltd

Contact Information:

TTS Research

Beijing Bombax XiaoIce Technology Co., Ltd

67 North 4th Ring West Road, Haidian, Beijing, China

Email: zhouzhiyang@xiaobing.ai

Github: github.com/atomicoo

ICASSP 2023

4 - 10 JUNE, RHODES ISLAND, GREECE

Abstract

Non-autoregressive text-to-speech (TTS) has achieved impressive inference speedup but at the cost of inferior voice quality. The fundamental reason lies in the gap between the complexity of data distributions and the capability of modeling methods. Previous works utilize either simplifying data distributions or enhancing modeling methods to alleviate the problem. In this work, we propose a new architecture ReActSpeech to explicitly learn to "auto-correct" for high-quality spectrograms. Specifically, ReActSpeech utilizes a redistribution module to improve (correct) extracted alignments automatically, and an iterative decoder called revisor to refine (correct) spectrograms iteratively. Extensive experiments conducted on several benchmarks show that ReActSpeech can greatly alleviate the above problem and achieve a nice tradeoff between training time, inference speed, and output quality.

Introduction

Autoregressive text-to-speech (AR-TTS) models use autoregressive decoding where the decoder generates a spectrogram step by step, and the generation of the latter frame depends on previously generated ones. Instead of sequential decoding, non-autoregressive text-to-speech (NAR-TTS) models generate the whole target spectrogram simultaneously. To enable parallel decoding, NAR-TTS imposes a conditional independence assumption among frames in a spectrogram, yielding significantly faster inference speed than AR-TTS. However, since intrinsic dependencies within spectrogram are omitted, NAR-TTS suffers from severe **inconsistency problems**, leading to inferior voice quality, which is manifested in the **fracture and over-smoothed** of spectrograms, especially when capturing highly dependent and multimodal distributions of spectrograms.

Dependent and multimodal distributions increase the uncertainty for model training and cause severe inconsistency problem in NAR models, which is observed in many generation tasks. There are some common ways to tackle dependent and multimodal distributions: (1) using loss functions or modeling methods that can well fit the distributions; and (2) introducing some input variables or simplifying the target data distributions.

In this work, a new non-autoregressive architecture **ReActSpeech** is proposed to explicitly learn to "**auto-correct**" for high-quality spectrograms, which can greatly alleviate the above problem. Specifically, it contains two parts: one is a redistribution module, which aims to redistribute alignments for improving (correct) alignments extracted by auto forced alignment tool, and the other is an iterative decoder called revisor, which is employed to refine (correct) spectrograms iteratively.

Main Objectives

- A redistribution module is designed to improve the alignments accuracy automatically.
- An iterative decoder and progressive training strategy is designed to improve the capacity of refinement.
- A multiple frequency bands discriminator with GAN training pipeline is employed to further tackle the over-smoothness problem.

Methodology

Preliminary Study

We calculate the marginal distributions ($P(y(t,f)|x = ph)$) and the joint distributions along frequency and time dimensions ($P(y(t,f1),y(t,f2)|x=ph)$ and $P(y(t,f),y(t+1,f)|x=ph)$), where $y(t,f)$ denotes to the data point in the t-th frame and the f-th frequency bin in the ground-truth spectrogram. As shown in Figure 1 & 2.

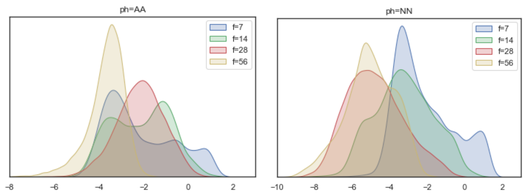

Figure 1: The marginal distributions $P(y(t,f)|x=ph)$ for different phonemes. We choose 2 phonemes ($ph = AA, NN$) and 4 frequency bins ($f = 7, 14, 28, 56$).

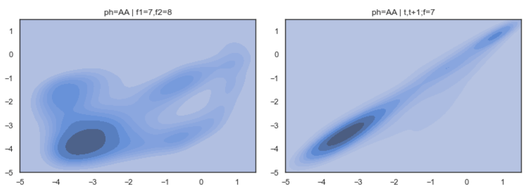

Figure 2: The joint distributions along frequency and time dimensions. (a) f-dim: $P(y(t,f1),y(t,f2)|x=ph)$. (b) t-dim: $P(y(t,f),y(t+1,f)|x=ph)$.

We analyze the accuracy of alignments provided to train the duration predictor by a new metric called relative boundary difference rate (RBDR), which is formulated as:

$$RBDR = \frac{min(\hat{E}, E_{gt}) - max(\hat{B}, B_{gt})}{max(\hat{E}, E_{gt}) - min(\hat{B}, B_{gt})}$$

, where B_{gt} and E_{gt} denote to the ground-truth begin and end boundaries of phoneme, respectively. As shown in Figure 3.

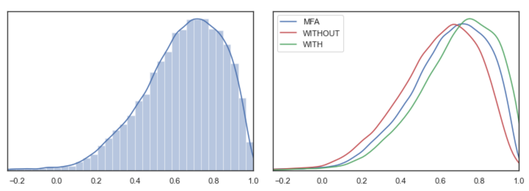

Figure 3: Alignments accuracy. (a) Alignments accuracy of MFA. (b) Alignments accuracy comparison (redistribution module).

Overall

The architecture of ReActSpeech is shown in Figure 4.

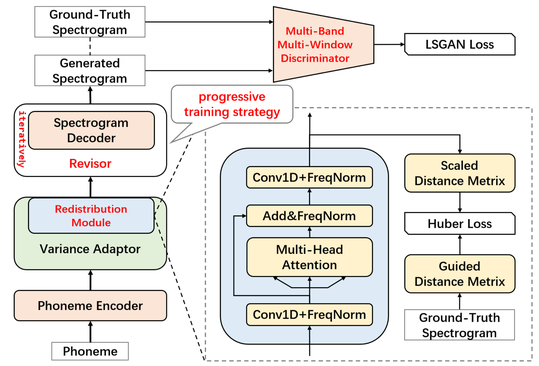

Figure 4: Overall Architecture.

Adaptor

The adaptor consists of several variance predictors, a length regulator (like FastSpeech 2), and a redistribution module. The redistribution module utilizes a variant of FFT Block to extract the cross-position information and redistribute the alignments.

Furthermore, we introduce a scaled distance matrix calculated from ground-truth spectrogram to guide the jointly training of alignments with our model. The main motivation is that we analyze the semantic distribution of hidden speech representations should be similar to the distribution of corresponding spectrogram. The scaled distance matrix describes the scaled L2 distance between every two frames of spectrogram or hidden speech representations.

Revisor

To tackle inconsistency problems, an iterative decoder called revisor is proposed to improve NAR-TTS by repeatedly refining previously generated spectrogram. Instead of enforcing NAR-TTS to generate accurate spectrogram by one-pass decoding, the revisor is expected to revise inferior spectrogram through several refinements.

For keeping the consistency with iterative decoding, a progressive iterative training strategy is utilized to further improve the training procedure. Specifically, we randomly fed a percentage of ground-truth frames of target spectrogram into revisor while training and gradually reduce the percentage to 0%.

Discriminator

We notice that the low, medium and high frequency bands of spectrograms have different natures. So we propose a multiple frequency bands multiple random window discriminator.

Results

We conduct subjective and objective evaluations to comprehensively measure the performance of each model.

Model	MOS
GT	4.536 ± 0.04
GT(Mel+HifiGAN)	4.464 ± 0.04
Tacotron 2	3.904 ± 0.07
FastSpeech 2 [1]	3.890 ± 0.07
ReActSpeech (Ours)	$\mathbf{4.237 \pm 0.06}$

Method	CMOS
ReActSpeech	0.000
w/o redistribution	-0.212
w/o iterative decoding	-0.135
w/o adversarial	-0.188

Table 1: MOS & CMOS. (a) MOS (95% CI). (b) CMOS comparison (Ablation).

Model	Training Time (h)	Inference Speed (RTF)	Inference Speedup
Tacotron 2	69.78	9.34×10^{-1}	/
FastSpeech 2 [1]	23.58	2.43×10^{-2}	38.4×
ReActSpeech (Ours)	**9.21**	$\mathbf{1.59 \times 10^{-2}}$	**58.7×**

Table 2: Training time and inference latency comparison. To make a fair comparison, we train all models (python implementation) on 1 NVIDIA V100 GPU, with batch size of 32, and inference on the same device with batch size of 1.

Conclusions

- **ReActSpeech** with a redistribution module and an iterative decoder has been proposed, which can learn to auto-correct for high-quality spectrograms.
- **ReActSpeech** outperformed the baseline largely on both the voice quality and the speech prosody. Furthermore, it achieves a nice tradeoff between fast training, fast inference, and high quality.

Additional Information

- Samples: https://wavelandspeech.github.io/reactspeech/.
- Appendix: https://wavelandspeech.github.io/reactspeech/pdfs/supplementary.pdf.
- Organization: https://www.xiaoice.com/.

References

[1] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*, 2021.