International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN: 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)
doi : https://doi.org/10.32628/IJSRSET229643

Autoregressive Speech-To-Text Alignment is a Critical Component of Neural Text-To-Speech (TTS) Models

Barath M, Bhuvan M
Senior High School Graduate, Kendriya Vidyalaya, Bangalore, Karnataka, India

Article Info
Volume 9, Issue 6
Page Number : 310-316

Publication Issue :
November-December-2022

Article History
Accepted : 05 Dec 2022
Published: 15 Dec 2022

ABSTRACT

Autoregressive speech-to-text alignment is a critical component of neural text-to-speech (TTS) models. Commonly, autoregressive TTS models rely on an attention mechanism to train these alignments online--but they are often brittle and fail to generalize in long utterances or out-of-domain text, leading to missing or repeating words. Non-autoregressive endto end TTS models usually rely on durations extracted from external sources. Our work exploits the alignment mechanism proposed in RAD -, which can be applied to various neural TTS architectures. In our experiments, the proposed alignment learning framework improves all tested TTS architectures—both autoregressive (Flowtron and Tacotron 2) and non-autoregressive (FastPitch, FastSpeech 2, RAD-TTS). Specifically, it improves alignment convergence speed of existing attention-based mechanisms; simplifies the training pipeline; and makes models more robust to errors on long utterances. Most importantly, it also improved the perceived speech synthesis quality when subject to expert human evaluation.

Keywords: Neural Text-To-Speech, RAD-TTS, TTS models, Artificial intelligence (AI), RecSLAM

I. INTRODUCTION

Neural text-to-speech (TTS) models often produce natural sounding speech for in-domain text, though they can have pronounced issues with comprehension and repetition when processing out-of-domain texts. A typical neural TTS model consists of a series of stages: mapping spoken words to their corresponding representations, generating audio files based on this representation, then aligning those sounds with what you've already heard. Older autoregressive TTS algorithms relied on an automatic system called content which matched the visual stimuli that are being encoded. However newer approaches use both content and location sensitivity called attention. Additionto being dependent on alignments from outside sources, these models can sufferfrom inefficient training methods, require well-crafted scheduling to maintain stable development, and may be difficultto expand across multiple languages if there are no pre-existing tools available or the output doesn't match what we're looking for. Ideally, the alignment should be trained end-to-end as part of the TTS program so that it would simplify an already long process. It would also help if it could make progress quickly enough so that other elements in the process won't slow down too much waiting for it. Perhaps most importantly, its final quality level needs to be as good or even better than what you'd get with externally sourced alignments alone.This report leverages the alignment framework discussed in which simplifies alignment learning in several TTS models. We show that it is able to easily convert all TTS models into a simpler end-to-end pipeline with much better convergence rates, and improved robustness against long utterances.We improve upon prior work on alignments in autoregressive speech synthesis by including a constraint that directly maximizes the probability of text given speech mel-spectrograms; we demonstrate how this approach can also be used online to learn alignments in parallel TTS models, thus eliminating the need for external aligners or alignments obtained from an already trained TTS model.Plus, we examined what an initial static alignment would do to help guide attentional learning.We show that our framework improves the performance of auto-regressive and parallel models when it comes to convergence rates in speech text alignments, their closeness to hand-annotated durations - but most importantly, they also tend to sound less robotic than other competing methods. In conclusion, this experiment showed that TTS models trained with our guidance on alignment-learning had reduced repetition and missed words during playback; improved stability on long sequences synthesis; and overall improved quality according to human evaluation.

II. Features

1) Speech AI. Large Language Models.

Artificial intelligence (AI) has transformed synthesized speech from monotone robocalls and decades-old GPS navigation systems to the polished tone of virtual assistants in smartphones and smart speakers.It has never been so easy for organizations to use customized state-of-the-art speech AI technology for their specific industries and domains.Speech AI is being used to power virtual assistants, scale call centers, humanize digital avatars, enhance AR experiences, and provide a frictionless medical experience for patients by automating clinical note-taking.

2) Accelerated Computing for Enterprise IT. Colocation. Networking

Modern applications are transforming every business, from data analytics for better business forecasting, to AI for autonomous vehicles, to advanced visualization for medical diagnosis. NVIDIA Accelerated Computing platforms provide the infrastructure to power these applications, no matter where they are run. An accelerated system is the next phase in the evolution of computers. Just like how all Smartphone's today have processors for graphics and AI, so too will every server and workstation have compute accelerators to power today's modern applications, including AI, visualization, and autonomous machines. Many of these systems will also have data processing units, which accelerate the network, storage and security services that are central to cloud native and cloud computing frameworks.

3) Design and Simulation. Overview. Metaverse.

The term 'simulation' has come to refer to a wide variety of forms of learning and activity. Definition is problematic, not least because the field is fast-moving and conceptions are being altered at fundamental levels by new technology and practice. A good example of this is the usual distinction that is made between symbolic and experiential simulations. Symbolic simulations 'depict the characteristics of a particular population, system or process through symbols; and the user performs experiments with variables that are a part of the program's population

4) Robotics and Edge Computing.

With the wide penetration of smart robots in multifarious fields, Simultaneous Localization and Mapping (SLAM) technique in robotics has attracted growing attention in the community. Yet collaborating SLAM over multiple robots still remains challenging due to performance contradiction between the intensive graphics computation of SLAM and the limited computing capability of robots tailored to heterogeneous edge resource conditions. Extensive evaluations show RecSLAM can achieve up to 39% processing latency reduction over the state-of-the-art. Besides, a proof-of-concept prototype is developed and deployed in real scenes to demonstrate its effectiveness.

5) HPC and AI. Simulation and Modeling. Scientific Visualization.

Modeling and Simulation (M&S) offer adequate abstractions to manage the complexity of analyzing big data in scientific and engineering domains. Unfortunately, big data problems are often not easily amenable to efficient and effective use of High Performance Computing (HPC) facilities and technologies. Furthermore, M&S communities typically lack the detailed expertise required to exploit the full potential of HPC solutions while HPC specialists may not be fully aware of specific modeling and simulation requirements and applications.

III. How it works

This program takes encoded text input Φ ∈ R Ctxt × N and aligns it to mel-spectrograms X ∈ R Cmel ×T where T is number of mel frames and N is the text length. The alignment is done by splitting up the input into sections of T consecutive frames in both spectrogram and input and finding the correspondence of each section to its corresponding section in the other matrix.

3.1 Alignment learning objective

To learn the alignment between mel-spectrograms X and text Φ, we use the alignment learning objective proposed in RADTTS. The idea is to find a maximum likelihood estimate of the probability that a given frame in X is generated from Φ. we constrain the alignment between text and speech to be monotonic, in order to avoid missing or repeating tokens.

$$P (S (\Phi) | X; \theta) = X s \in S(\Phi) YT t=1 P(st | xt; \theta)$$

The above formulation of the alignment learning objective does not depend on how the likelihood P(st = φi xt) is obtained, which makes it possible to explore both bottom-up and top-down approaches to the problem of alignment. Hence, it can be applied to both autoregressive and parallel models.

3.2 Autoregressive TTS Models

Autoregressive TTS models typically use a sequential formulation of attention to learn online alignments. TTS models such as Tacotron and Flowtron use a content based attention mechanism that relies only on decoder inputs and the current attention hidden state to compute an attention map between encoder and decoder steps. Autoregressive TTS models such as Tacotron and Flowtron use a content based attention mechanism that relies only on decoder inputs and the current attention hidden state to compute an attention map between encoder and decoder steps. In these autoregressive models, the alignment of input and output has a direct effect on the likelihood of a misstep in the alignment. Alignment learning is tightly coupled with the decoder and can be learned with the mel-spectrogram reconstruction objective.

However, it has been observed that the likelihood of a misstep in the alignment increases with the length of the utterance. This results in catastrophic failure on long sequences and out-of domain text.

IV. How it is used

The autoregressive setup for Flowtron uses the standard stateful content based attention mechanism and a hybrid attention mechanism that uses both content and location based features for Tacotron2. The standard stateful content based attention mechanism is not a novel approach to solving this problem but it can be effective in some contexts, whereas the hybrid attention mechanism is more robust and versatile because it can leverage both content and location based features.

Use of a Tacotron2 encoder to obtain the sequence of encoded text representations (φ enc i) N i=1 and an attention RNN to produce a sequence of states ht.

4.1 Parallel Models of TSS

Recently, researchers have developed a system in which the alignment learning module is decoupled from the Mel decoder as a standalone aligner. The benefits of this system are that it is more efficient and has less latency than the traditional approach to TTS models, allowing for better understanding of what is being said. Furthermore, as the duration is factored out from the decoder, there is no need for an initial frame alignment between speech and text, which saves processing time for all three modules and reduces latency for this particular step in the pipeline. We compute the soft alignment distribution based on the learned pairwise affinity between all text tokens and mel frames, which is normalized with softmax across the text domain

Di,j = distL2(φenci , x encj),

Asoft = softmax(−D, dim = 0)

4.2 Architecture of TSS

A parallel model is a sequence modeling technique that does not require any information about the input sequence to be specified beforehand and is instead inferred from the input data as it goes along. Parallel models can either be generative or discriminative and are considered an example of a statistical model. The alignment of the input sequence to the output sequence can be specified beforehand by determining the number of output samples for every input phoneme, equivalent to a binary alignment map. However, attention models produce soft alignment maps, constituting a train-test domain gap. As such, the Viterbi algorithm is a powerful tool for finding the most likely monotonic path through the soft alignment map in order to convert soft alignments (A soft) to hard alignments (A hard).

4.3 Acceleration of alignment

Faster convergence of alignments means faster training for the full TTS model, as the decoder needs a stable alignment representation to build upon. Training relies on a stable alignment representation to build upon, so the slower it takes to find an initial alignment, the longer it takes to train and find a final alignment that is useful in generating speech. To speed up this process, we use a static 2D prior with uniform scaling near the corners and more aggressive scaling near the center of Mel-spectrograms during training. . Although our formulation with the 2D static prior is slightly different than Tachibana et al [18], but we believe both should yield similar results

V. Experiment

We compare the effectiveness of the alignment learning framework by comparing its performance in terms of convergence speed, distance from human annotated ground truth durations, and speech quality. For autoregressive models like Flowtron and Tacotron 2, we compare with the baseline alignment methods therein. For FastPitch, we compare with an alignment method that relies on an external TTS model (Tacotron2) to obtain token durations. For the parallel models: FastSpeech 2 and RAD-TTS, we compare against an alignment method that obtains durations from the MFA aligner. We use LJ dataset for all our experiments.

5.1 Convergence Rate

To compare the convergence rate of different alignment methods, we use the mean mel-cepstral distance (MCD). MCD compares the distance between synthesized and ground truth mel-spectrograms aligned temporarily with dynamic time warping (DTW). MCD is calculated by taking the mean of all distances between synthesized and ground truth mel-spectrograms when aligning with DTW. The difference in convergence rates can be observed by comparing the MCD values obtained by different methods.

Parallel models such as RAD-TTS, FastPitch, and FastSpeech2 with the alignment framework converge at the same rate as their baseline models using a forced aligner. Furthermore, its been noted that even without forcing an alignment algorithm to be applied, these parallel models converged at the same rate as their baseline models. The model that benefits the most from using the alignment framework is Flowtron. It has two autoregressive flows running in opposing directions, each with their own learned alignment. Notably, the second autoregressive flow is performed on top of the autoregressive outputs of the previous flow. This means that if the alignment in the first flow fails, so will the second. The second flow can only be added after the first has converged in order to train properly. Prior attempts to train both flows simultaneously have resulted in poor minima where neither flow has learned to align with each other. By using just the attention prior, we are now able to train at least two flows simultaneously, with further improvements with adding the unsupervised alignment learning Lalign objective. This significantly reduces training time and improves convergence of Flowtron.

5.2 Alignment Sharpness

The alignment objective consistently makes the attention distribution sharper with more connected alignment paths. This suggests that models with Lalign produce more confident and continuous alignments, and by extension, continuous speech without repeating or missing words.

5.3 Durational Analysis

To measure the effectiveness of an unsupervised alignment loss, we examine the difference between average durations from our model based alignments and human-annotated durations. For autoregressive models, we extract every part after adjusting for human-annotations and find which one has highest attention weights among those two sequences (current + next). This will give us a binary line segmented alignment which allows us to find how long each phoneme is said in both ground truth text as well as our own text. As shown by the Figure, it seems that by using our method, the time taken to reach convergence rates are shorter than when basing on just the baseline or even worse - no supervision at all! Thus it can be inferred that this unsupervised approach provides better results since it doesn't depend on external data besides what could already been seen before hand.

5.4 Pair wise Opinion Scores

To measure the effectiveness of an unsupervised alignment loss, we examine the difference between average durations from our model based alignments and human-annotated durations. For autoregressive models, we extract every part after adjusting for human-annotations and find which one has highest attention weights among those two sequences (current + next). This will give us a binary line segmented alignment which allows us to find how long each phoneme is said in both ground truth text as well as our own text. As shown by Figure 1, it seems that by using our method, the time taken to reach convergence rates are shorter than when basing on just the baseline or even worse - no supervision at all! Thus it can be inferred that this unsupervised approach provides better results since it doesn't depend on external data besides what could already been seen before hand.

5.5 Robustness to Errors on Long Utterances

We measure the character error rate between synthesized and input texts when we evaluate the robustness of the alignments on long utterances. Using

14,045 full sentences from Libri TTS dataset, we generate sequences of speech with a model trained on LJ Speech and then use Jasper to recognize them. As you can see in Figure below (which plots CER), autoregressive models with Lalign are less prone to errors than other methods such as parallel models. Parallel models - like RAD-TTS - don't experience alignment issues because they predict how much time is left for speech generation based on average sentence lengths in a corpus; however, this does not apply to autoregressive models which can only deal with one sentence at a time.

VI. Pros

Parallel TTS models provide a lot of flexibility in choosing the architecture to formulate the distribution. The process, which includes both autoregressive and parallel architectures, combines guidance in the form of forward-sum, Viterbi, and diagonal priors with attention-based online alignment training. This ensures stability and fast convergence while eliminating the need for costly forced aligners.

VII. Cons

One of the difficulties of text-to-speech algorithms is that the alignment learning module and the Mel decoder are coupled together.

Attention models produce soft alignment maps, constituting a train-test domain gap. Models with a focus on attention generate maps of soft alignment, which creates a difference between the training and testing domains.

VIII. Conclusion

We present a new process for TTS alignment to improve voice recognition rates. The process, which includes both autoregressive and parallel architectures, combines guidance in the form of forward-sum, Viterbi, and diagonal priors with attention-based online alignment training. This ensures stability and fast convergence while eliminating the need for costly forced aligners. To ensure we reach optimal alignment performance our approach has been tested on synthesizing texts of arbitrary lengths that were chosen among those typically encountered by such systems.

IX. Acknowledgement

This project is done by student Bhuvan Shridhar and Barath.M , 12th Grade students under the guidance of Associate Professor Pavan Kumar, Department of Computer Science, Kendriya Vidyalaya ASC Centre, Bangalore under Central Board of Secondary Education, Delhi, India.

X. REFERENCES

[1]. MikikoBazeley is a Senior ML Operations and Platform Engineer at Mailchimp. She has extensive experience as an engineer, data scientist, and data analyst for startups and high-growth companies leveraging machine learning and data for consumer and enterprise facing products. She actively contributes content around best practices for developing ML products as well as speaking and mentoring non-traditional candidates in building careers in data science.

[2]. Nvidia, https://www.nvidia.com/en-in/data-center/solutions/accelerated-computing/

[3]. Simulations, learning and the metaverse: changing cultures in legal education Paul Maharg (Glasgow Graduate School of Law) Martin Owen, (Futurelab)

[4]. Edge Robotics: Edge-Computing-Accelerated Multi-Robot Simultaneous Localization and Mapping, Liekang Zeng, Xu Chen, Ke Luo, Zhi Zhou, Shuai Yu

[5]. Why High-Performance Modelling and Simulation for Big Data Applications Matters,

Clemens Grelck, Ewa Niewiadomska-Szynkiewicz, Marco Aldinucci, Andrea Bracciali & Elisabeth Larsson

[6]. Image on iStocks, Licence details, Creator: MF3d, Credit: Getty Images

[7]. Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," CoRR, vol. abs/1703.10135, 2017. [Online]. Available: http://arxiv.org/abs/1703.10135

[8]. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," CoRR, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[9]. R. Valle, K. Shih, R. Prenger, and B. Catanzaro, "Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis," 2020.

[10]. Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," arXiv preprint arXiv:2006.04558, 2020.

[11]. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. Fox, and ´ R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019, pp. 3171–3180.

Cite this article as :

Barath M, Bhuvan M, "Autoregressive Speech-To-Text Alignment is a Critical Component of Neural Text-To-Speech (TTS) Models", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 9 Issue 6, pp. 310-316, November-December 2022. Available at doi : https://doi.org/10.32628/IJSRSET229643
Journal URL : https://ijsrset.com/IJSRSET229643