PAMA-TTS: PROGRESSION-AWARE MONOTONIC ATTENTION FOR STABLE SEQ2SEQ TTS WITH ACCURATE PHONEME DURATION CONTROL

Yunchao He *Jian Luan* *Yujun Wang*

Xiaomi Corporation, Beijing, China

ABSTRACT

Sequence expansion between encoder and decoder is a critical challenge in sequence-to-sequence TTS. Attention-based methods achieve great naturalness but suffer from unstable issues like missing and repeating phonemes, not to mention accurate duration control. Duration-informed methods, on the contrary, seem to easily adjust phoneme duration but show obvious degradation in speech naturalness. This paper proposes PAMA-TTS to address the problem. It takes the advantage of both flexible attention and explicit duration models. Based on the monotonic attention mechanism, PAMA-TTS also leverages token duration and relative position of a frame, especially countdown information, i.e. in how many future frames the present phoneme will end. They help the attention to move forward along the token sequence in a soft but reliable control. Experimental results prove that PAMA-TTS achieves the highest naturalness, while has on-par or even better duration controllability than the duration-informed model.

Index Terms— Alignment guidance, duration control, attention mechanism, seq2seq TTS, speech synthesis

1. INTRODUCTION

Text-To-Speech is a typical sequence-to-sequence modeling task. In general, its input is a grapheme or phoneme sequence while the output is a much longer sequence of acoustic parameters at the frame level. In recent popular encoder-decoder architectures, the attention mechanism demonstrated strong capability in mapping two sequences with different lengths [1] and achieved high naturalness in TTS tasks [2, 3, 4]. However, for unseen texts, it may also bring errors like missing and repeating phonemes, unexpected long silence, and even failure to produce speech completely [5, 6, 7]. Many efforts have been made to enhance the attention robustness by constraining the attention to meet locality, monotonicity, and completeness, such as Forward attention [5], Stepwise monotonic attention [6], and Location-Relative attentions [7]. However, none of them constrained how many frames one token should occupy. Without it, phonemes in an unseen text may still be articulated extremely short or too long in synthesized speech.

Other than attention-based methods, many studies utilize a separate duration model to implement the sequence upsampling. Fastspeech [8], Fastspeech2 [9], and DurIAN [10] duplicate encoder outputs according to the phoneme duration. Non-Attentive Tacotron [11] implements upsampling with Gaussian weights. The ground-truth duration is obtained from external forced-alignment tools [9, 10, 11, 12, 13] or by internal joint training [14, 15]. Regardless of how the alignment is obtained and how the duplicated tokens are smoothed, duration-informed methods always show naturalness degradation due to hard duration control.

Differentiable duration models [16, 17] are also designed. They need no phoneme alignment guidance but to optimize duration model parameters by minimizing the final spectrogram reconstruction loss directly. For the duration loss, only the total duration of phonemes in a sequence is taken into account. It improved the naturalness of duration-informed methods. However, in such networks, the output of the duration model may not physically stand for phoneme duration. Particularly, when the predicted duration of one word is adjusted when inference, the durations of other words in the synthesized speech are often affected unexpectedly.

This paper proposes a Proceeding-Aware Monotonic Attention (PAMA[1]) for sequence-to-sequence TTS to realize accurate phoneme duration control without naturalness degradation. The neural network is based on Tacotron2 but the Location Sensitive Attention (LSA) is replaced by stepwise monotonic attention [6]. Besides, a soft guidance attention matrix is generated from ground-truth alignment to benefit both the efficiency of attention training and the correctness of learned alignment. At the same time, an auxiliary duration model is trained with the same alignment label. From the duration model, latent duration representation and backward position embedding are offered to attention memory and query respectively. The main contributions of this paper include:

- Design an innovative guidance attention matrix for alignment constraint. The guidance is soft at phoneme boundaries since there are no solid ground-truth breaks;
- Introduce latent duration representation into encoder output as attention memory. With this information, alignment loss converges faster and more stably;
- Introduce backward frame position within phoneme

[1] Audio examples: https://pama-tts.github.io/

into prenet output as an attention query. In this way, the generation of current spectrum conditions on not only the preceding spectrogram but also how many future frames the present phoneme should end within. The former ensures the spectrum smoothness while the latter helps the phoneme duration control. Their impacts are balanced by the network dynamically.

2. RELATED WORKS

Although PAMA-TTS calculates attention alignment vector recursively in the same way as stepwise monotonic attention in [6], both attention query and memory of them are different. For query, PAMA-TTS adds backward position information for token proceeding awareness. For memory, PAMA-TTS adds latent duration representation for efficient and stable training convergence.

VAENAR-TTS [18] introduces a latent variable Z to help soft attention alignment, in which Z implicitly stands for phoneme duration. However, there are no phoneme level duration labels to guide Z explicitly. Besides, VAENAR-TTS leverages both annealing reduction factor and causality mask to help attention-based alignment learning other than applies monotonic constraint.

Moreover, the attention alignment loss in PAMA-TTS is quite similar to PAG in [19]. However, PAMA-TTS generates guidance matrices in a softer way for better flexibility, since the results of a forced alignment tool may have slight distortion, especially on found data.

3. ARCHITECTURE

The architecture of PAMA-TTS is shown in Fig. 1. Tacotron2 [3] with stepwise monotonic attention [6] is employed as the backbone. Modified modules are highlighted and will be illustrated below one by one.

3.1. Text Encoder & Phoneme Classifier

The text encoder takes a sequence of token IDs as inputs and outputs the latent representation of them, which consist of regular phonemes, tones, prosodic boundaries, and silence. The tokens are placed in a carefully designed order to build up input sequences as demonstrated in Fig. 2.

Since tones and most prosodic boundaries (except intonation phrase boundary #3, which can be regarded as silence or short pause as well) do not correspond to any acoustic frames in speech, a filter is applied to skip the hidden states of them as shown in Fig. 2. A similar strategy is used in DurIAN [10], but they remove only prosodic boundaries. Moreover, the trimmed encoder output is fed into a phoneme classifier to ensure the token location information remains. Both above designs aim at making the subsequent alignment learned by the attention mechanism more meaningful.

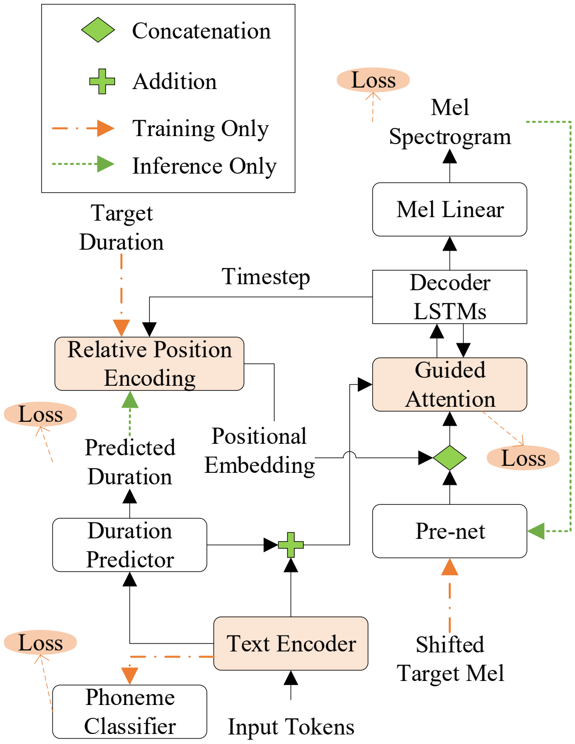

Fig. 1. Architecture diagram of PAMA-TTS. The yellow and green dotted lines are turned on only for the training and inference stage respectively.

The encoder structure is the same as that of Tacotron2, i.e. three convolutional layers followed by a BLSTM layer. For the phoneme classifier, a single feed-forward layer with softmax cross-entropy loss is employed.

3.2. Guided Attention Matrix

Guided attention is used to help the attention module learn a correct mapping between phoneme sequence and acoustic frames efficiently. Previous work [19] used time-aligned phoneme sequences obtained by forced alignment to generate hard guidance matrices. Considering the existence of alignment errors, this paper improves the guidance matrix to have fuzzy weights at phoneme boundaries as shown in Fig. 3.

According to statistics on large data, most alignment errors of phonemes are within 3 frames. Therefore, the weights at boundaries in the guidance matrix are linearly transitioned from 0 to 1 in six frames with a step size of 0.2. Then, a mean square error is computed as alignment loss as

$$L_{align} = \frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{N}(W_{ij} - \alpha_{ij})^2 \tag{1}$$

where T, N denote the number of spectrogram frames and filtered tokens, $W, \alpha \in \mathbb{R}^{N \times T}$ are the guidance matrix and attention weight matrix, respectively.

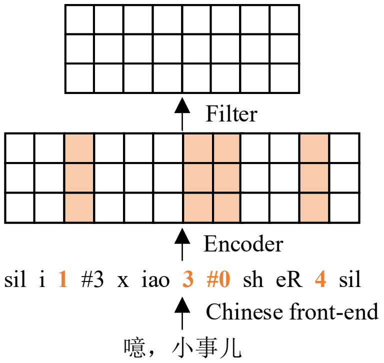

Fig. 2. An illustration of how to skip some hidden states (highlighted in yellow) of encoder output. The symbols #0, #1, #2, and #3 denote the boundaries of syllables, prosodic words, intermediate phrases, and intonational phrases respectively. The numbers 1-5 denote tones of the previous syllable.

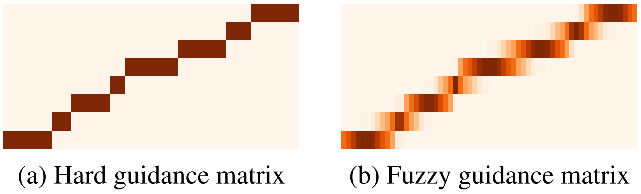
(a) Hard guidance matrix (b) Fuzzy guidance matrix

Fig. 3. An example of hard and fuzzy alignment guidance. Weights in (b) change smoothly at phoneme boundaries.

3.3. Progression-Aware Monotonic Attention

The proposed PAMA is based on stepwise monotonic attention. To make the monotonic attention aware of the mapping progression between phonemes and spectrogram, two additional pieces of information is leveraged: one is a latent duration code for attention memory, and the other is a relative position embedding for attention query.

The latent duration code is from the last hidden layer of a duration predictor and transformed by a linear layer. For each phoneme, its duration code is added with its encoder output to generate key and value for the attention mechanism. In this way, the attention's key and value vectors carry duration information more explicitly.

The relative position embedding is a concatenation of two vectors from learnable look-up tables. One is for the forward position within a phoneme, which implies the distance to the beginning of the token. The other is for the backward position, which denotes the distance to the end of the token. Both of the two distances are ceilinged with a constant C. For each acoustic frame, its relative positional embedding is concatenated with the output of prenet to generate an attention query.

Generally speaking, prenet output only carries information of the preceding spectrogram. The injection of relative position embedding, especially bringing the knowledge that how many future frames the current phoneme should end within, helps the attention be more premeditated.

At the training stage, the forward and backward positions are both derived from forced alignment labels. At the inference stage, the forward position is calculated according to the attention weights of preceding steps and the backward position is estimated from the predicted duration. To convert the forward /backward distance into a learnable vector, an embedding lookup layer is used in which two lookup tables are learned for forward and backward distances separately.

3.4. Training Loss

The overall loss is a weighted sum of four parts as

$$L = L_{mel} + \alpha_1 L_{pc} + \alpha_2 L_{dur} + \alpha_3 L_{align} \tag{2}$$

where L_{mel}, L_{pc}, L_{dur}, and L_{align} denote MSE loss of Mel-spectrogram reconstruction, Cross-Entropy (CE) loss for the phoneme classifier, L1 loss for duration predictor, and MSE loss for guided attention, respectively. Their weights are set as $\alpha_1 = 0.005, \alpha_2 = 0.025, \alpha_3 = 0.25$ empirically.

Here, the stop token predictor [3] is not used. Instead, the decoder is assumed to stop when attention has stayed at the last token for the predicted duration time.

4. EXPERIMENTS

4.1. Training Setup

We evaluated the proposed model on an internal corpus, which was from a non-professional female speaker, containing about 10 hours of speech data (about 12,000 utterances). The audios were collected in native mandarin Chinese and resampled into 16 kHz, 16-bit mono wave format.

A proprietary front-end engine was used to convert input texts into token sequences, which contain phonemes, tones, prosodic boundaries, and silence marks. Besides, a Kaldi-based forced alignment tool [20] was used to obtain phoneme duration labels from recordings.

Two variants of Tacotron2 are used as baselines. One replaces the attention mechanism in Tacotron2 with a duration informed length regulator (called TLR), and the other employs stepwise monotonic attention (called TSW). The post-net module is removed due to limited effectiveness. The reduction factor is set to 1 for a better quality of speech.

The same pre-trained LPCNet [21] is used as a vocoder to generate audio signals from the predicted Mel-spectrogram.

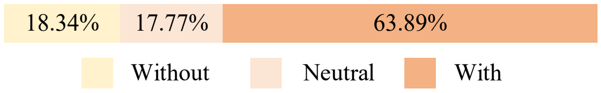

Fig. 4. Preference test between PAMA-TTS with and without relative position embedding, which is at a p <0.01 level.

Table 1. The MOS with 95% confidence intervals for the proposed method (PAMA), ground-truth samples (GT), and two baselines (TLR and TSW). The ground truth is obtained via analysis-synthesis.

Models	MOS
GT	4.54 ± 0.12
TLR	4.22 ± 0.14
TSW	4.38 ± 0.18
PAMA	**4.41** ± 0.14

Table 2. The mean absolute errors (ms) of phoneme duration between the predicted by duration model and the segmented from the synthetic speech by forced alignment. The duration factor is used to scale the predicted duration to control the speech rate of synthetic speech.

Model	Duration Factor		
	0.75	1.0	1.5
TLR	9.72	7.53	14.67
TSW	-	65.92	-
PAMA	**8.48**	**6.68**	**13.54**

4.2. Evaluation Setup

Two objective evaluations were conducted using 1,000 sentences. Firstly, the duration consistency was measured to show the duration controllability of models, which was calculated as the mean absolute errors (MAE) between the phoneme duration predicted by the duration predictor and that from a forced aligner. For TSW, phoneme duration was estimated from the attention results as $d_i = \sum_{t=1}^{T}[argmax_n \alpha_{n,t} = i]$, where d_i was the duration of the ith phoneme, and $\alpha \in \mathbb{R}^{N \times T}$ was the final attention matrix. Secondly, the phoneme error rate (PER) given by an automatic speech recognition (ASR) model was adopted as the metric to measure the robustness of different models. The ASR model was based on a TDNN-LSTM structure and trained on nearly 100,000 hours of recordings collected from various Xiaomi mobile phones.

Subjective evaluations were conducted using 30 sentences. They were not included in the training data. The naturalness of the synthetic speech was evaluated through the mean opinion score (MOS) test and AB preference test. 16 native listeners participated in the test, and the speech samples were shuffled in each test.

4.3. Results & Discussion

As shown in Table 1, the proposed model (PAMA) gets the highest mean opinion score. TLR shows slightly mechanical rhythm while TSW has clarity issues in some cases.

Table 3. The phoneme error rates (PER) of different models on 1,000 test sentences using a reduction factor (DF) of 1.0, 0.75 and 1.5. For TSW, speed modulation is almost infeasible.

DF	Error	TLR	TSW	**PAMA**
1.0	Sub	1.92	1.68	1.69
	Del	0.27	1.13	0.23
	Ins	0.20	0.21	0.20
	PER	2.39	3.02	**2.12**
0.75	Sub	4.23	-	3.16
	Del	1.84	-	0.93
	Ins	0.37	-	0.42
	PER	6.44	-	**4.51**
1.5	Sub	1.60	-	1.43
	Del	0.16	-	0.15
	Ins	0.27	-	0.28
	PER	2.03	-	**1.86**

Results of the AB preference test shown in Fig. 4 confirm the importance of procession-awareness for attention. If the relative position embedding is not leveraged, the naturalness of synthetic speech has remarkable degradation.

To check the duration controllability of different systems, MAE and PER are calculated for three duration factors (DF). We find stepwise monotonic attention is very weak at speech rate control. When attention score bias is shifted within a small range [-3, 3], the speech rate has a very slight change. However, if a greater shifting is applied, serious word skipping /repeating issues occur frequently. Therefore, only TLR and PAMA are evaluated for duration modification. Table 2 compares the capability of duration control. It shows PAMA has on-par or even fewer duration errors than TLR, and an overwhelming advantage over TSW. Table 3 compares the robustness with an ASR tool, in which PAMA has much fewer deletion errors than TSW and even better than TLR on overall performance.

5. CONCLUSION

This paper introduced progression-aware monotonic attention for robust sequence-to-sequence speech synthesis. The proposed model (PAMA-TTS) demonstrates that injecting the duration and relative position information into attention can achieve a better balance between the robustness and naturalness of synthetic speech. Besides, it enables accurate control of phoneme duration. Subjective and objective evaluation results show that PAMA-TTS outperforms the attention-based model on robustness and duration controllability while outperforms the duration-informed model on naturalness. Progression-aware monotonic attention is proved to be feasible for token length control and may be extended to other similar applications easily.

6. REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, p. 6000–6010.

[2] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Samy Bengio Zhifeng Chen, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[3] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[4] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI*, 2019, vol. 33, pp. 6706–6713.

[5] Jing-Xuan Zhang, Zhen-Hua Ling, and Li-Rong Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *Proc. ICASSP*, 2018, pp. 4789–4793.

[6] Mutian He, Yan Deng, and Lei He, "Robust sequence-to-sequence acoustic modeling with stepwise monotonic attention for neural tts," in *Proc. Interspeech*, 2019, p. 1293–1297.

[7] Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby, "Location-relative attention mechanisms for robust long-form speech synthesis," in *Proc. ICASSP*, 2020, pp. 6189–6193.

[8] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019, p. 3165–3174.

[9] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-toend text to speech," in *arXiv:2006.04558*, 2020.

[10] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi Lei, Dan Su, and Dong Yu, "Durian: Duration informed attention network for multimodal synthesis," in *arXiv:1909.01700*, 2019.

[11] Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu, "Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling," in *arXiv:2010.04301*, 2020.

[12] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J. Weiss, and Yonghui Wu, "Parallel tacotron: Non-autoregressive and controllable tts," in *Proc. ICASSP*, 2021, p. 5694–5698.

[13] Takuma Okamoto, Tomoki Toda, Yoshinori Shiga, and Hisashi Kawai, "Tacotron-based acoustic model using phoneme alignment for practical neural text-to-speech systems," in *Proc. ASRU*, 2019, p. 7254–7258.

[14] Zhen Zeng, Jianzong Wang, Ning Cheng, Tian Xia, and Jing Xiao, "Aligntts: Efficient feed-forward text-to-speech system without explicit alignment," in *Proc. ICASSP*, 2020, p. 6714–6718.

[15] Dan Lim, Won Jang, Gyeonghwan O, Heayoung Park, Bongwan Kim, and Jaesam Yoon, "Jdi-t: Jointly trained duration informed transformer for text-to-speech without explicit alignment," in *arXiv:2005.07799*, 2020.

[16] Jeff Donahue, Sander Dieleman, Mikołaj Binkowski, Erich Elsen, and Karen Simonyan, "End-to-end adversarial text-to-speech," in *arXiv:2006.03575*, 2020.

[17] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, RJ Skerry-Ryan, and Yonghui Wu, "Parallel tacotron 2: A non-autoregressive neural tts model with differentiable duration modeling," in *arXiv:2103.14574*, 2021.

[18] Hui Lu, Zhiyong Wu, Xixin Wu, Xu Li, Shiyin Kang, Xunying Liu, and Helen Meng, "Vaenar-tts: Variational auto-encoder based non-autoregressive text-to-speech synthesis," in *Proc. Interspeech*, 2021, pp. 3775–3779.

[19] Xiaolian Zhu, Yuchao Zhang, Shan Yang, Liumeng Xue, and Lei Xie, "Pre-alignment guided attention for improving training efficiency and model stability in end-to-end speech synthesis," *IEEE Access*, vol. 7, pp. 65955–65964, 2019.

[20] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[21] Jean-Marc Valin and Jan Skoglund, "Lpcnet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*. IEEE, 2019, pp. 5891–5895.