PHONEMIC-LEVEL DURATION CONTROL USING ATTENTION ALIGNMENT FOR NATURAL SPEECH SYNTHESIS

Jungbae Park[1,2], *Kijong Han*[2], *Yuneui Jeong*[1], *Sang Wan Lee*[1,2,3*]

[1]Humelo Inc.
[2]Korea Advanced Institute of Science and Technology (KAIST)
[3]KAIST Institute for Artificial Intelligence
jb@humelo.com, han0ah@kaist.ac.kr, yjeong@humelo.com, sangwan@kaist.ac.kr

ABSTRACT

Recent attention-based end-to-end speech synthesis from text systems have achieved human-level performance. However, many approaches cause a sequence-to-sequence model to generate only averaged results of the input text, making it difficult to control the duration of utterance. In this study, we present a novel mechanism for phonemic-level duration control (PDC) in a nearly end-to-end manner in order to solve this problem. We used a teacher attention alignment generated by an annotation speech analyzer program. Our method is inspired by the idea that the duration of a phoneme is highly related to its phonemic features. These phonemic features are saved on the attention alignment by adding duration embedding to it. This enables the model to learn and control the phonemic and rhythmic features of speech. We also show that providing alignment information as a teacher loss term improves training speed and notably, makes the model better at controlling the speed of dramatic change in phonemic-level duration with subjective demonstration. As a result, we show that our PDC speech synthesis with alignment loss outperforms other baseline methods without losing the ability to control the duration of phonemes in extremely adjusted environments with faster convergence.

Index Terms— alignment loss, attention, deep learning, phonemic-level duration control, speech synthesis

1. INTRODUCTION

Sequence-to-sequence (seq2seq) generating deep neural network models [1], which are currently used as the base of many end-to-end speech synthesis systems, have been introduced to generate speech in an end-to-end manner in the context of machine translation [2–4]. However, the performance of seq2seq models is negatively affected when the input sequence is too long or the data have a variant length of inputs. The attention mechanism, allowing the model to focus on what it needs to learn, has been shown to resolve this dependency problem on the input dimension [5, 6].

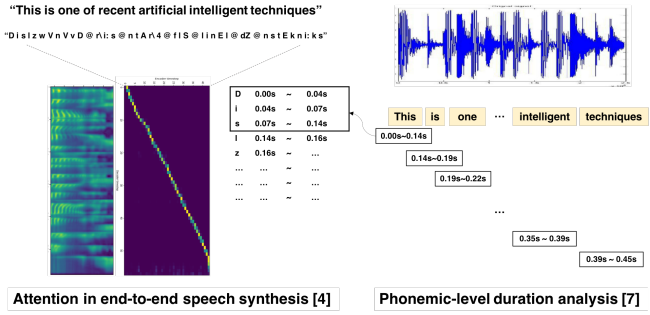

Fig. 1. Relation of attention alignment in end-to-end speech synthesis with phonemic-level duration extraction of [7]

As seq2seq models have developed, speech synthesis systems have incorporated each component of the complex pipeline of speech synthesis [4, 8–11]. This has successfully reduced the need for human annotation with conditioning of high-level features like speakers or high-level style features like prosody or style [12–14].

However, most existing attention-based end-to-end speech synthesis systems do not represent more than text. The phonemic or word-level duration control of speech while maintaining characteristics and tones of speech is essential to designing a customizable speech synthesis system. Unfortunately, the attention-based end-to-end speech synthesis requires a large number of training steps.

Nevertheless, the attention mechanism in seq2seq helps us to know which state the encoder is focusing on in a specific decoder time interval [15]. Ideally, well-trained speech recognition and speech synthesis systems can have a causal nature, so that the probability map of the attention, shortly the attention alignment, is mostly continuously aligned [16, 17] (See Fig. 1.a). This means the trained attention itself can glean information about duration from each encoder. Actually, some studies have shown that phonetic segmentation prediction could infer duration features per phoneme [7], thereby improving speech synthesis [18, 19].

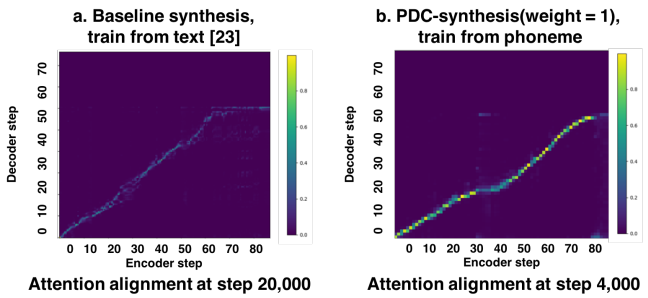

Fig. 2. Comparison of attention alignments. The color bar corresponds with the attention probability of which encoding cell would be focused on specific decoding cell

Motivated by these recent studies, this paper presents a novel model for a phonemic-level duration controllable (PDC) speech synthesis (shortly, PDC-synthesis). The proposed model attempts to resolve the issues described above with a duration embedding by a phonemic-level duration extractor, SPPAS [7] and by introducing the attention alignment loss term to guide attention-driven learning (See Fig. 1.b).

2. RELATED WORKS

Several attention-based seq2seq models have been suggested [4, 17, 20] in order to implement end-to-end speech synthesis systems. These models can successfully generate human-level speech by conditioning WaveNet [21, 22] on mel-spectrogram predictions [23] or by directly inferencing audio from text, distilling a Gaussian inverse autoregressive flow from an autoregressive WaveNet [24]. There are, however, some weaknesses in these models such as a lack of expressiveness of voice and necessity of huge number of training steps, as discussed in the introduction. According to a related paper [25], attention-based end-to-end speech synthesis has a limitation in that about 10 000 training steps are required to learn the attention alignment. Even worse is that models with guided-attention need to iterate over 20 000 training steps. On the other hand, our suggested model significantly reduces the training cost with more accurate attention alignment after only about 4000 steps, giving predicted duration with our attention alignment loss (Fig. 2).

Recently, embedding an unsupervised global style token has made speech synthesis models controllable in many styles [14]. However, its information embedding in a global manner hinders the ability to control duration in the word and phoneme level. Moreover, the nature of defining style tokens in an unsupervised manner may compromise the quality of interpretability [26]. The proposed approach in this study resolves these issues; it not only improves the ability to control duration in the phonemic level but also exhibits a relatively free range of control regardless of data distribution.

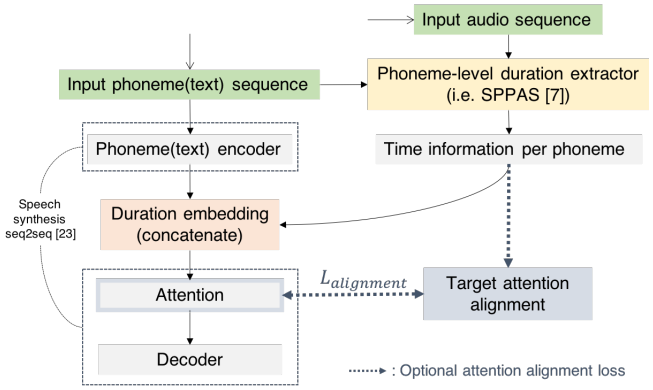

Fig. 3. Model diagram of phonemic-level duration control (PDC) speech synthesis. Attention alignment loss is denoted as $L_{alignment}$

3. APPROACHES AND MODEL ARCHITECTURES

Our model is constructed based on Tacotron 2 [23]. While the original work used $\{text, audio\}$ pairs for the training session, this study uses $\{phoneme, audio\}$ pairs. This is because 1) using phoneme instead of text allows for clear alignment of the attention (See Fig. 2) and 2) it makes a prediction of that duration extractor for each phoneme more accurate. In addition, although the original study used WaveNet vocoder [21] to synthesize more natural speech, we use Griffin-lim vocoder [27] as [4] to compare our model with the main seq2seq speech synthesis model in naturalness and controllability of duration in a phonemic level.

The architecture of our approach is shown in Fig. 3. The objective function of speech synthesis models, including [17, 23], is described as follows.

$$f(T) = M,$$

where T represents a text, and M is the mel-spectrogram of speech. The neural architectures using this objective are trained to generate mel-spectrogram of the given sentences. In addition to the original objective, another term was added to our model:

$$f(T, D) = M,$$

where D is the phoneme-sequence of the input text, paired with the duration of each phoneme. We then modified the baseline model [23] to incorporate the duration information. Details are described in the following subsections.

3.1. Phonemic-Level Duration Embedding

We extracted phonemes and duration for each phoneme utilizing SPPAS [14]. We refer to each phoneme as p_i in this study. SPPAS also gives start and end times of each phoneme in milliseconds. We refer it as $start(p_i)$ and $end(p_i)$, respectively. This is illustrated in Fig. 1.b. In our model, the vector

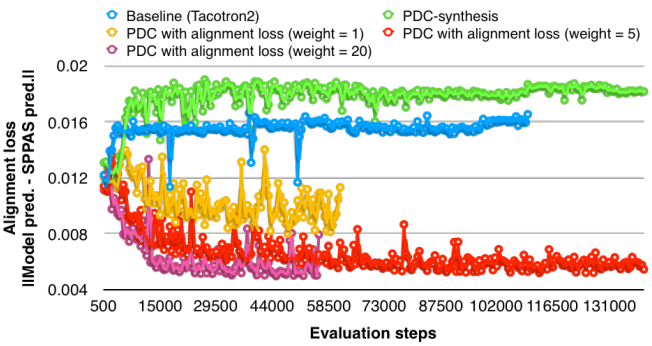

Fig. 4. Comparison of alignment loss of each model

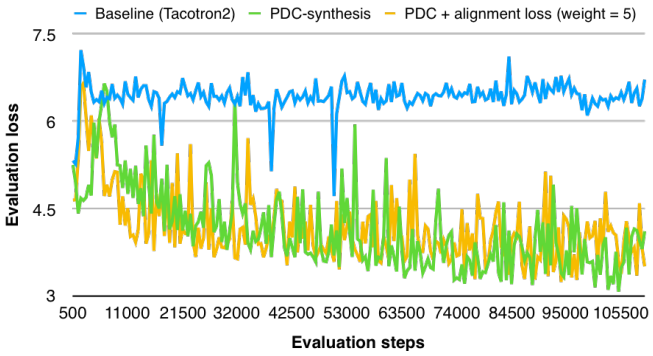

Fig. 5. Evaluation loss results

x_i, the input for the i-th recurrent unit (i.e., the i-th cell of a long-short term memory (LSTM)) in encoder, is defined as follows.

$$x_i = \{conv(v_{p_i}), start(p_i), end(p_i), dur(p_i)\},$$

where v_{p_i} represents the embeddings for i-th phoneme, and $conv(v_{p_i})$ represents the convolution layer, which generates a 509-dimensional vector. This is the same design as [9] except that the input unit changed from the character to phoneme.

The output of the cells of the encoder LSTM is fed into the attention, which gives alignment information about which portion of input the decoder should concentrate on when it generates a specific time range of mel-spectrogram. Thus, the duration information should be inputted into the encoder so that the attention can utilize it. For this purpose, we added three concatenated floating-point scalar elements to the input vector, $start(p_i)$, $end(p_i)$, and $dur(p_i)$, which were extracted from the phonemic-level duration extractor. Each represents start time, end time and duration for the phoneme p_i, respectively.

3.2. Attention Alignment Loss Term to Learn Duration

We added the following attention alignment loss (hereafter referred to as the alignment loss) term to the loss function to ensure that the model aligned duration of phonemes properly during the training session:

$$\alpha * \frac{1}{|\ A\ |} \mid A_{pred} - A_{target} \mid_2^2$$

where α is a weight term of the attention alignment and $A = (a_{ij})$ is the alignment matrix. The subscript $pred$ represents the alignment that is trained in the attention of the model, and the subscript $target$ represents the true alignment, which is constructed from the duration information of the input extracted from the phonemic-level duration extractor. $a_{ij} \in [0.0, 1.0]$ means how much the decoder should focus on the i-th phoneme when it generates the j=th frame of the mel-spectrogram.

The true alignment is constructed as follows. Firstly, all elements in a vector are filled with zeros. For every i-th phoneme, we iterate over the frames j. In this step, we set $a_{ij} = 1.0$ if the time range of the j-th frame is within the time range of the time range of i-th phoneme. SPPAS also gives the error range for each duration. Thus, if the time range is in the middle of the error range, we can set the value proportional to the overlapping ratio of the time and error range. In this study, we tested our model with various weight values, $\alpha \in \{1, 5, 20\}$ to explore the effects of the alignment loss.

4. SIMULATION RESULTS

4.1. Dataset and Experiment Setups

We used LJSpeech 1.1 dataset [28] for the demonstration. In this experiment, to show the control range of our model, we synthesized speech data in four different speed categories, labeled as {*very fast*, *fast*, *slow*, *very slow*}. The first eight phonemes for every twenty-phoneme chunk in a phoneme-sequence were controlled in such a way that it encodes a particular duration based on the category ({30, 70, 110, 150} ms). The duration of the other twelve phonemes was fixed at 70 ms. We set *very fast* and *very slow* categories to be unusual cases in the LJSpeech dataset, while moderately controlled categories were in the usual cases of the dataset. We tested two suggested models: a PDC-synthesis model that had a duration embedding with no alignment loss and a PDC-synthesis model with the alignment loss [1].

4.2. Alignment Prediction of Attention in Evaluation

As described in Section 3.2., the alignment loss is defined as a norm of the difference between the target alignment generated by a phonemic-level extractor and the current attention alignment. This alignment loss term can be a measure of whether the attention of the model is well-trained. (See

[1] Audio samples are available in https://jungbaepark.github.io/20190517-pdcsynthesis

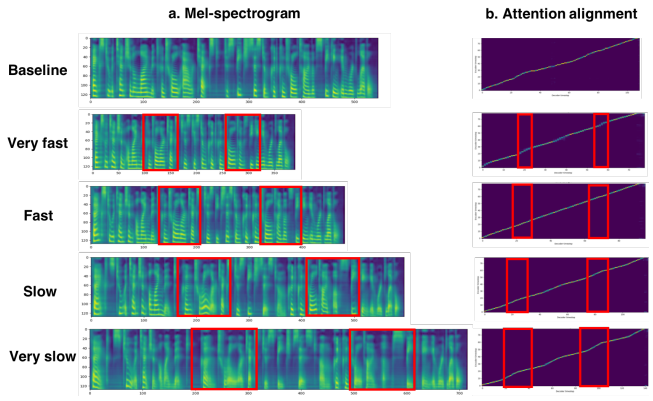

Fig. 6. Examples of phonemic control. Red rectangles indicate controlled areas for phones to be faster or slower.

Usual in data?	Category	Baseline [23]+G.L.	PDC	PDC+1*A.L.	PDC+5*A.L.	PDC+20*A.L.
Unusual	Very fast	-	**2.69±0.70**	2.26±0.83	1.67±0.55	2.02±0.56
Usual	Fast	2.59±0.80	**3.64±0.87**	2.92±0.73	3.21±0.66	2.21±0.66
	Slow		2.88±0.73	2.50±0.61	**3.02±0.79**	2.07±0.78
Unusual	Very slow	-	2.02±0.65	1.92±0.59	**2.31±0.68**	1.78± 0.58

Table 1. MOS for naturalness, 95% confidence interval (CI)

Usual in data?	Category	PDC	PDC+1*A.L.	PDC+5*A.L.	PDC+20*A.L.
Unusual	Very fast	3.10±0.97	2.10±1.19	1.91±1.19	**3.14±1.17**
Usual	Fast	**3.86±0.89**	3.57±1.22	**3.86±0.99**	3.57±1.05
	Slow	3.38±1.13	3.19±1.26	**3.71±1.03**	3.24±1.23
Unusual	Very slow	2.71±1.24	2.33±1.13	**3.19±1.44**	2.71±1.35

Table 2. MOS for duration control, 95% CI

Fig. 4) Because SPPAS itself has an error in the prediction procedure, PDC-synthesis with no alignment loss term has more alignment loss compared to the baseline. With higher α, the weight term of attention alignment, the alignment loss becomes much smaller as the model is trained. This means the attention alignment may be similar to the duration predicted by SPPAS library.

4.3. Inference and Evaluation

For evaluation loss (See Fig. 5), with comparison to the baseline, both simple PDC-synthesis and PDC-synthesis with the alignment loss converged faster than the baseline. This indicates the duration embedding from the duration extractor helps training and evaluation, even though the duration predicted by SPPAS was not exact (as illustrated in Fig. 4). PDC-synthesis, especially, with the alignment loss term expedites attention learning as shown in Fig. 2 (attention alignment is completed in less than 4000 steps).

Generated mel-spectrograms and predicted attention alignment samples by PDC-synthesis are shown in Fig. 6. The controlled area that makes speech have a specific duration is marked as red line quadrangles. Notably, the duration control, even in unusual cases such as the {*very fast*, *very slow*} models, successfully infers attention alignment.

4.4. Subject Evaluation

For evaluation, we measured the mean opinion score (MOS) that quantifies naturalness. Several subjects were asked to rate the naturalness and the completeness of the duration control of the stimuli in a five-point Likert scale score.

As shown in Table 1, both simple PDC-synthesis (denoted as PDC) and PDC-synthesis with alignment loss (denoted as A.L.) ($\alpha = 5$) showed better performance in the usual cases of data than the baseline, Tacotron2, with Griffin-lim (denoted as G.L.) vocoder. Simple PDC-synthesis showed great performance in unusually *very fast* cases, and the best cases of PDC-synthesis with alignment showed better performance in abnormally *very slow* cases.

On the other hand, as shown in Table 2, PDC-synthesis with high weighted alignment loss showed better performance in controlling duration for speech synthesis.

5. CONCLUSION AND DISCUSSION

We have proposed a PDC-synthesis, phonemic-level duration controllable speech synthesis that features duration embedding from a phonemic-level duration extractor. We also have defined a novel attention alignment loss that compares the attention alignment to the target information on predicted duration from phonemic-level duration. Notably, we have shown that the alignment loss can be used to check whether an attention-based s speech synthesis model is well-trained.

We also have demonstrated that the PDC-synthesis can help to reduce training costs significantly by improving the convergence speed. In addition, both simple PDC-synthesis and PDC-synthesis with the alignment loss generate more natural speech when the duration control range is in the data spectrum. Even in unusual cases involving out-of-range data, a simple PDC-synthesis performed inference in unusually *very fast* cases. On the other hand, the best cases of PDC-synthesis with alignment showed better performance in abnormally *very slow* cases. Nevertheless, to get more natural speech, a distilling process from better vocoder such as WaveNet [21, 22] may be added. Because the well-trained WaveNet can reduce noises from incomplete predicted mel-spectrogram during synthesis, there is an open question that combining PDC-synthesis with WaveNet would be natural even in unusual cases of data distribution. Lastly, for the completeness of duration control, we have shown that PDC-synthesis with the alignment loss term helps the model to have better completeness in duration control during inference, both in usual and unusual cases of data.

6. REFERENCES

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *In Advances in neural information processing systems*, pp. 3104–3112, 2014.

[2] A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[3] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *arXiv preprint arXiv:1603.01354*, 2016.

[4] W. Yuxuan et al., "Tacotron: Towards end-to-end speech synthesis," in *arXiv preprint arXiv*, 2017, p. 1703.10135.

[5] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[6] M. T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *CoRR*, vol. abs/1508.04025, 2015.

[7] B. Bigi, "Sppas: a tool for the phonetic segmentations of speech," in *The eighth international conference on Language Resources and Evaluation*, 2012, pp. 1748–1755.

[8] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[9] S. Xu, W. Wang, and B. Xu, "First step towards end-to-end parametric tts synthesis: Generating spectral parameters with neural attention," in *In Proceedings Interspeech*, 2016, pp. 2243–2247.

[10] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 84–96, 2018.

[11] S. O. Arik et al., "Deep voice: Real-time neural text-to-speech," *CoRR*, vol. abs/1702.07825, 2017.

[12] S. O. Arik et al., "Deep voice 2: Multi-speaker neural text-to-speech," *CoRR*, vol. abs/1705.08947, 2017.

[13] R. J. Skerry-Ryan et al., "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *CoRR*, vol. abs/1803.09047, 2018.

[14] Y. Wang et al., "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *CoRR*, vol. abs/1803.09017, 2018.

[15] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, vol. 37, pp. 2048–2057.

[16] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964.

[17] J. Sotelo et al., "Char2wav: End-to-end speech synthesis," 2017.

[18] S. R. Vignesh, S. A. Shanmugam, and H. A. Murthy, "Significance of pseudo-syllables in building better acoustic models for indian english tts," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5620–5624.

[19] A. Baby et al., "Deep learning techniques in tandem with signal processing cues for phonetic segmentation for text to speech synthesis in indian languages," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, 2017, pp. 3817–3821.

[20] W. Ping et al., "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. 6th International Conference on Learning Representations.*, 2018.

[21] A. Oord et al., "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[22] A. Oord et al., "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, vol. 80, pp. 3918–3926.

[23] J. Shen et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *arXiv preprint arXiv*, 2017, p. 1712.05884.

[24] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," in *arXiv preprint arXiv*, 2018, p. 1807.0728.

[25] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4784–4788.

[26] D. R. Liu et al., "Improving unsupervised style transfer in end-to-end speech synthesis with end-to-end speech recognition," .

[27] D. Griffin and J. 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.

[28] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.