Initial investigation of an encoder-decoder end-to-end TTS framework using marginalization of monotonic hard latent alignments

Yusuke Yasuda[1,3]*, Xin Wang*[1]*, Junichi Yamagishi*[1,2,3]

[1]National Institute of Informatics, Japan [2]The University of Edinburgh, Edinburgh, UK
[3]SOKENDAI (The Graduate University for Advanced Studies), Japan

yasuda@nii.ac.jp, wangxin@nii.ac.jp, jyamagis@nii.ac.jp

Abstract

End-to-end text-to-speech (TTS) synthesis is a method that directly converts input text to output acoustic features using a single network. A recent advance of end-to-end TTS is due to a key technique called attention mechanisms, and all successful methods proposed so far have been based on soft attention mechanisms. However, although network structures are becoming increasingly complex, end-to-end TTS systems with soft attention mechanisms may still fail to learn and to predict accurate alignment between the input and output. This may be because the soft attention mechanisms are too flexible. Therefore, we propose an approach that has more explicit but natural constraints suitable for speech signals to make alignment learning and prediction of end-to-end TTS systems more robust. The proposed system, with the constrained alignment scheme borrowed from segment-to-segment neural transduction (SSNT), directly calculates the joint probability of acoustic features and alignment given an input text. The alignment is designed to be hard and monotonically increase by considering the speech nature, and it is treated as a latent variable and marginalized during training. During prediction, both the alignment and acoustic features can be generated from the probabilistic distributions. The advantages of our approach are that we can simplify many modules for the soft attention and that we can train the end-to-end TTS model using a single likelihood function. As far as we know, our approach is the first end-to-end TTS without a soft attention mechanism.

Index Terms: text-to-speech synthesis, end-to-end, neural network

1. Introduction

End-to-end text-to-speech (TTS) synthesis method directly converts an input letter or phoneme sequence to an output acoustic feature sequence. All methods proposed so far have been based on an encoder-decoder sequence-to-sequence model with a soft attention mechanism [1, 2, 3, 4, 5, 6].

The aforementioned frameworks are very promising, and end-to-end TTS systems such as Tacotron 2 [5] can produce synthetic speech with a quality comparable to that of human speech. The use of neural waveform models is one of the reasons, but we believe that improved model architectures are another important reasons for the improved performance of the end-to-end TTS systems. For example, a transition from Tacotron [2] to Tacotron 2 [5] has extended the attention mechanism from additive attention [7] to location sensitive attention [8], post-net to improve predicted acoustic features, an additional stop flag prediction network, non-deterministic output generation using dropout during inference, and a significant increase in model parameter size. However, even if the attention network is well-trained, it may produce unacceptable errors such as skipping input words, repeating the same phrases, and prolonging the same sounds.

Therefore, efforts have been made to improve or constrain soft attention mechanisms for end-to-end TTS systems. Most of them aim to enforce a monotonic alignment structure in order to reduce alignment errors. Such attempts include location-sensitive attention [8], monotonic attention [9], and forward attention [10]. In addition, techniques other than the attention mechanism itself can enforce monotonic alignment, such as window masking [4, 10] and penalty loss for off-diagonal attention distribution [11]. Thus, although the quality of synthetic speech generated from end-to-end TTS systems is very high, the model architecture and its objective function are very complex, and many engineering tricks are used.

Can we construct an encoder-decoder end-to-end TTS system without using such complicated networks and engineering tricks? We hypothesize that the main cause is the soft attention mechanisms. Therefore, we propose a new end-to-end TTS system that can be optimized based on a likelihood function only. Furthermore, the proposed TTS system uses more explicit but natural constraints for speech signals instead of soft attention mechanisms and should make alignment learning more robust and efficient. The constrained alignment is conceptually borrowed from segment-to-segment neural transduction (SSNT) [12, 13] and is extended to continuous outputs [1]. Because the SSNT calculates the joint probability of an acoustic feature sequence and an alignment given the input text, we can compute the likelihood function by marginalizing all possible alignments. The alignments are hard and monotonic increase by definition.

The proposed framework and likelihood function are similar to those of a hidden Markov model (HMM) because all the possible alignments can be marginalized using a forward-backward algorithm over a trellis consisting of input labels and the output spectrum. However, unlike the conventional HMM or its mixture density network-based HMM [15], its transition and output probabilities are computed using encoder-decoder autoregressive neural networks like Tacotron. In other words, an input sequence is nonlinearly transformed into an encoded vector, and both the output and transition probabilities are computed in a non-linear autoregressive manner in the SSNT-based TTS.

Our system is still under development and the quality of synthetic speech is not perfect yet. However, we present our first evaluation on the performance of the current system in this paper. We focus on our system's description and its evaluation

[1]SSNT is a similar model to RNN transducer [14], a well-known method in speech recognition. A notable difference is that SSNT separates transition probability from output probability, which is a preferred feature for TTS which is a task to predict continuous output.

for a standard reading speech corpus in this paper. Its application to a verbal performance corpus is investigated in [16], and it shows effectiveness of our system in that corpus.

The rest of the paper is structured as follows. In Section 2, we overview end-to-end TTS systems with soft attention. In Section 3, we describe the new encoder-decoder TTS using marginalization of monotonic hard alignments and likelihood-based learning. In Section 4, we show experimental results using the system under development, and we summarize our findings in Section 5.

2. Overview of end-to-end TTS with soft attention

Table 1 summarizes major end-to-end TTS methods. All the existing systems consists of an encoder and decoder with attention as basic components. Most systems have a pre-net [2] at the entrance of the decoder [2, 4, 5, 6]. Although some of the earliest studies [1, 3] do not have the pre-net in the decoder, all the others do after [2] introduced it.

The choice of target acoustic features is crucial for the end-to-end approach [17]. Some of the earliest work chose vocoder parameters as a target [1, 3]. Vocoder parameters are challenging because they have a long sequence length caused by fine analysis conditions for reliable feature extraction. [2] used a mel spectrogram with coarse-grained analysis conditions as target features to reduce the gap of the length between input text and target acoustic feature sequences. All the other studies followed the same condition [2, 4, 5, 6].

The attention mechanism is the core of those approaches. Many attention mechanisms have been used for TTS. Additive attention [7] and dot-product attention [18] are content based families [8] that consider input content to align input to output features. GMM attention [9] is a location based attention [8] that consider input location only. Location sensitive attention [8] extends the additive attention by considering the previous alignment, so it has both properties of content-based and location-based attention. The systems using vocoder parameters [1, 3] seem to choose GMM attention. GMM attention has monotonic progress properties for modes of attention distributions for input location, so it is suitable for predicting long sequences like vocoder parameters. The systems based on CNN or self-attention to enable parallel training [4, 6] seem to use do-product attention which can be combined with positional encoding [19] for constructing sequential relations and an initial monotonic alignment in parallel. The systems based on RNN seem to use additive attention and its extension [2, 5, 10, 17].

In addition to the decoder, some systems have a post-net, an additional network that predicts acoustic features. A post-net was originally introduced to convert acoustic features to different acoustic features that were suitable for an adopted waveform synthesis method, for example, from mel spectrograms to linear spectrograms [2] or mel spectrograms to vocoder parameters [4]. In recent studies the role of the post-net was to improve the acoustic features predicted by the decoder to improve quality further [5, 6]. The post-net introduces an additional loss term in the objective function.

Recent methods predicts binary stop flags to determine when to stop prediction [5, 4, 6]. As opposed to predicting the fixed length output, the stop flag enables avoiding unnecessary computation. The stop flag introduces an additional loss term in the objective function.

Speech is nondeterministic. An exactly identical utterance can not be reproduced in speech. To enable the randomness, [5] enables dropout in pre-net during inference.

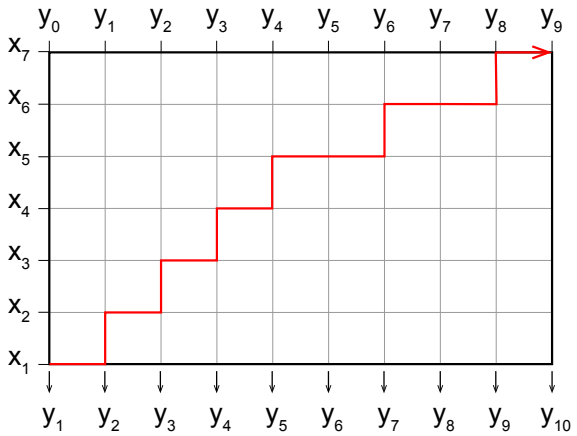

Figure 1: *Trellis structure of our model. A path that connects $\boldsymbol{x}$ and $\boldsymbol{y}$ represents alignment.*

Although the aforementioned techniques have contributed to the advancement of the end-to-end TTS, some issues remain unsolved by these techniques. All content-based attentions and their extensions do not guarantee forward progress of alignment from the previous position. They include forward attention [10], which extends the location sensitive attention by incorporating monotonic transition formulation. The GMM attention [9] has monotonic progress properties for the mode position of each Gaussian component. However, it sometimes gives broad, mode-free, or multi-modal distributions that result in muffled speech, possibly because it does not consider the content of input. Monotonic attention [20] is the only attention mechanism that guarantees monotonic progress by switching to hard attention during inference from soft attention during training, but it is not successful for text-to-speech synthesis [4].

Learning stop flags seems to be trivial because it is just a binary classification task and because only one boundary has a flag turn from not to stopping to stopping. However, it tends to overfit because its loss has an order of magnitude lower than the acoustic feature loss term. To alleviate this problem, [6] introduced higher weighting at the boundary where a flag reaches a stop state in the binary loss term. The stop flag is an extra complexity to implement such a trivial feature.

Enabling randomness for generated speech by dropout is not widely utilized because the dropout is normally disabled during prediction.

3. Proposed SSNT-based TTS

3.1. Model definition and learning of SSNT-based TTS

Let us denote the input letter or phone sequence as $\boldsymbol{x}_{1:I} = \{x_1, \cdots x_I\}$, where x_i is a i-th letter or phone of the input sequence. We then use $\boldsymbol{y}_{1:J} = \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_J\}$ and $\boldsymbol{y}_j \in \mathbb{R}^D$ to denote an output acoustic feature sequence and acoustic features at time j, respectively. Our approach is to model the output acoustic feature sequence $\boldsymbol{y}_{1:J}$ given a letter or phone sequence $\boldsymbol{x}_{1:I}$ by marginalizing all possible alignments over a trellis con-

Table 1: *Summary of major end-to-end TTS methods. All existing methods use soft attention mechanism for implicit alignment learning.*

System	Network	Alignment	Decoder output	Post-net output	Waveform synthesis
Char2Wav [1]	RNN	GMM	Vocoder	-	SampleRNN
Tacotron [2]	RNN	Additive	Mel	Linear	Griffin-Lim
VoiceLoop [3]	Memory buffer	GMM	Vocoder	-	WORLD
Deep Voice 3 [4]	CNN	Dot-product	Mel	Linear/Vocoder	Griffin-Lim/WORLD/WaveNet
Tacotron 2 [5]	RNN	Location-sensitive	Mel	Mel	WaveNet
Transformer [6]	Self-attention	Dot-product	Mel	Mel	WaveNet
SSNT (proposed)	**RNN**	**Hard**	**Mel**	**-**	**WaveNet**

sisting of the input and output sequences:

$$p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}) = \sum_{\forall \mathbf{z}} p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I}), \tag{1}$$

where $\mathbf{z} = \{z_1, \cdots z_J\} = \{1, 1, , \cdots I-1, I\}$ represents one of the possible paths of the trellis. Figure 1 shows trellis structure of our model. We then use the concept of SSNT [12, 13]. More specifically, we factorize the joint probability of Eq. (1) into an alignment transition probability and emission probability for acoustic features with the 1st-order Markov assumption of $\mathbf{z}$:

$$p(\boldsymbol{y}_{1:J}, \mathbf{z} \mid \boldsymbol{x}_{1:I}) \approx \prod_{j=1}^{J} p(z_j \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I}) p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I}) \tag{2}$$

and we use neural networks to compute the two probabilities.

To constrain the alignment probability as left-to-right with a self transition, an alignment transition variable with two possible values, $a_{i,j} = \{\texttt{Emit}, \texttt{Shift}\}$ is further introduced for the alignment probability of Eq. (2):

$$p(z_j = i \mid z_{j-1}, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I}) = \begin{cases} 0 & z_{j-1} > i \\ p(a_{i,j} = \texttt{Emit}) & z_{j-1} = i \\ p(a_{i-1,j} = \texttt{Shift}) p(a_{i,j} = \texttt{Emit}) & z_{j-1} = i-1 \\ 0 & z_{j-1} < i-1 \end{cases} \tag{3}$$

Note that the `Emit` transition keeps the input position, while `Shift` transition proceeds and reads one more input. Please also note that $p(a_{i,j} = \texttt{Emit}) = 1 - p(a_{i,j} = \texttt{Shift})$ and that a neural network predicts only one of them.

The emission probability was a discrete distribution because SSNT was originally proposed for NLP tasks such as abstractive sentence summarization, morphological inflection generation, and machine translation and because its outputs are words, the emission probability was a discrete distribution. In our case, the output is continuous, and hence we have to define our own emission probability distribution function. In this paper, we simply use a multivariate Gaussian distribution as the emission probability of the acoustic features:

$$p(\boldsymbol{y}_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I}) = \mathcal{N}(\boldsymbol{y}_j; \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \tag{4}$$

Note that $\boldsymbol{\mu}$ is predicted by an encoder-decoder network with autoregressive feedback similar to Tacotron.

Our model can be trained by minimizing the negative log likelihood:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= -\log p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta}) \\ &= -\log \alpha(I, J) \end{aligned} \tag{5}$$

Here $\alpha(I, J)$ is a forward variable of the forward-backward algorithm at the final input position I and the final time step J. The final forward variable can be calculated recursively:

$$\begin{aligned} &\text{For } j = 1, \\ &\quad \alpha(i, 1) = p(z_1 = 1 \mid \boldsymbol{x}_{1:I}) p(y_1 \mid z_1, \boldsymbol{x}_{1:I}) \\ &\text{For } j > 1, \\ &\quad \alpha(i, j) = p(y_j \mid \boldsymbol{y}_{1:j-1}, z_j, \boldsymbol{x}_{1:I}) \cdot \\ &\quad \big\{ \alpha(i-1, j-1) p(z_j = i \mid z_{j-1} = i-1, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I}) \\ &\quad + \alpha(i, j-1) p(z_j = i \mid z_{j-1} = i, \boldsymbol{y}_{1:j-1}, \boldsymbol{x}_{1:I}) \big\} \end{aligned} \tag{6}$$

The gradients of the negative log likelihood with respect to $\boldsymbol{\theta}$ can be computed using the standard back-propagation algorithm. For more efficient gradient computation, the gradient can be computed with the following form:

$$\frac{\partial \log p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{1}{p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\partial p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})}{\partial \alpha(i, j)} \frac{\partial \alpha(i, j)}{\partial \boldsymbol{\theta}} \tag{7}$$

Then the back-propagation can be combined via backward operation of the forward-backward algorithms by introducing a backward variable $\beta(i, j)$ and utilizing the relation $\frac{\partial p(\boldsymbol{y}_{1:J} \mid \boldsymbol{x}_{1:I}; \boldsymbol{\theta})}{\partial \alpha(i, j)} = \beta(i, j)$ [12, 21].

3.2. Inference of SSNT-based TTS

During prediction, the alignment variable z_j is incremented by sampling from the Bernoulli distribution with a parameter obtained by normalizing the two nonzero cases of Eq. (3), namely, $z_j = z_{j-1} + k$, where $k \sim \text{Bernoulli}(p)$ and

$$p = \frac{p(a_{i-1,j} = \texttt{Shift}) p(a_{i,j} = \texttt{Emit})}{p(a_{i,j} = \texttt{Emit}) + p(a_{i-1,j} = \texttt{Shift}) p(a_{i,j} = \texttt{Emit})}.$$

Note that this is different from a transition matrix in HMM synthesis which models duration with exponential distribution. Our system models transition of alignment as a latent variable so our system do not define distribution of duration like HMM synthesis.

In the original SSNT, its prediction stops when the EOS token is the output. In our case, its prediction stops when the alignment variable reaches the final input position[2] causing all of the input sequence to be used.

Acoustic features may be sampled from the Gaussian distribution using the conditional input at the sampled alignment

[2] [13] uses the same criteria to stop prediction to consume full input.

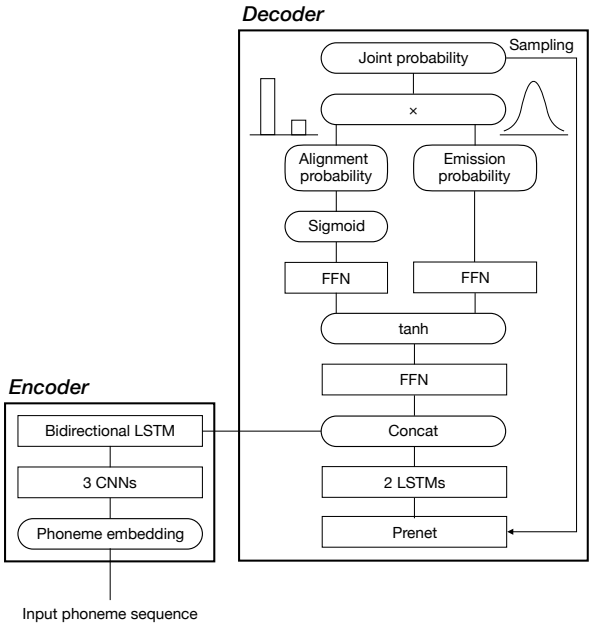

Figure 2: *Detailed network structure of the SSNT-based TTS system.*

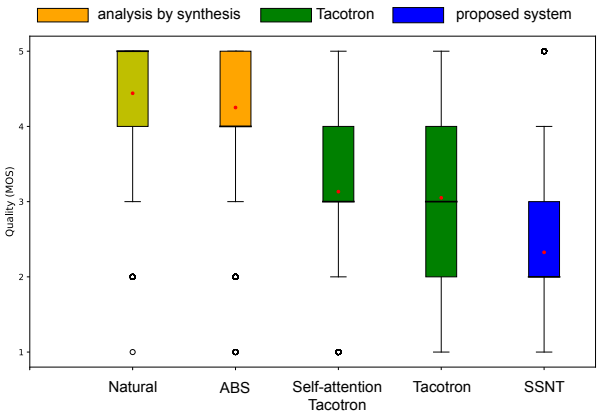

Figure 3: *MOS scores of subjective evaluation.*

position. However, in this paper, we simply use the mean of the prediction probability distribution as the generated acoustic features. We will investigate the random sampling strategy after we properly estimate a full covariance of Eq. (4).

3.3. Network structure of SSNT-based TTS

Figure 2 shows the detailed network structure of the system for SSNT-based speech synthesis system. The network consists of an encoder and decoder. The encoder processes either a letter or phone sequence. We then use the CNN-based encoder [5], which consists of a stack of convolutional layers, and a bidirectional LSTM layer [22].

The decoder processes the acoustic feature sequence in an autoregressive manner. The predicted acoustic features from the previous time step are first processed by the pre-net [2], which consists of fully connected layers with ReLU activation [23] and dropout regularization [24]. The pre-net's output is passed through a stack of LSTM layers [25]. The LSTM stack's output is concatenated with the encoder's output and then transformed by a fully connected layer with tanh activation. The output of tanh nonlinearity is passed to two networks. One is a fully connected layer with sigmoid activation to compute the alignment transition probability $p(a_{i,j})$ of Eq. (3). The other is a linear layer to compute the emission probability of Eq. (4).

4. Experiments

4.1. Experimental conditions

We used the same conditions as in our previous experiment [17]. A Japanese speech corpus from the ATR Ximera dataset [26] was used. This corpus contains a total of 28,959 utterances from a professional female speaker and is around 46.9 hours in duration. We used manually annotated phonemes labels [27]. The phoneme labels had 58 classes, including silences, pauses, and short pauses. All sentences start and end with a silence symbol. Although Japanese is a pitch-accented language, we did not use accentual type labels in this paper. To train our proposed systems, we trimmed the beginning and ending silence from the utterances, after which the duration of the corpus was 33.5 hours. Fixed length silences were prepended and appended to target mel spectrogram. These data preparation made any phoneme symbols inappropriate to be skipped. The frame size and shift used for the mel spectrogram were 50 ms and 12.5 ms, respectively. We used 27,999 utterances for training, 480 for validation, and 480 for testing.

Phoneme embedding vectors have 256 dimensions. For the encoder, we used the same conditions as [5]. For the decoder, we used two pre-net layers with 256 and 128 dimensions, two LSTM layers with 256 dimensions each, and two fully connected layers for context projection with 256 dimensions each. We applied zoneout regularization [28] to all LSTM layers with probability 0.1 as in [5]. We set the reduction factor [2] to be 2. The Adam optimizer was used for training [29]. The validation loss was steady so we stopped training at 510 epochs by checking quality of predicted spectrogram in validation set. Finally, we used μ-law WaveNet for the waveform generation [30].

4.2. Subjective evaluations

We conducted a listening test to measure the quality of synthetic speech. We chose Japanese Tacotron and self-attention Tacotron without accentual type labels [17] as baselines in this experiment. All the synthetic speech waveforms were generated using the identical WaveNet model, which was trained using natural mel spectrograms with a 12.5 ms frame shift and 16 kHz sampling frequency.

We recruited 104 native speakers of Japanese as listeners by crowdsourcing. The listeners evaluated 40 audio samples that contained eight randomly selected sentences generated by each of five systems in a random order in a single test set. The systems included natural speech and analysis-by-synthesis in addition to our system and the two aforementioned baseline systems. One listener could evaluated at most ten test sets. One audio sample was evaluated five times, and we got a total of 19,200 data points.

Figure 3 shows the results of the subjective evaluation. Among the baseline systems, self-attention Tacotron got 3.13$\pm$0.03, and Japanese Tacotron got 3.05$\pm$0.03. The scores of the baseline systems are consistent with the previous work [17]. The scores were relatively low because we did not

provide any pitch accentual type labels, even though Japanese is a pitch-accented language. Our system got a MOS score of 2.33 ± 0.03. Unfortunately, it was rated lower than the baseline systems.

To understand the reasons, we did a simple investigation of generated audio files and discovered out our model overestimated the phoneme duration, especially for pauses within a sentence. Pauses had much longer duration than other phonemes, but, its acoustic features had less useful information; hence, deciding when `Shift` transition should be made would be difficult. Figure 4 shows an example sample that has alignment error of overestimation of pause duration. In fact, the MOS score of sentences that did not include pauses was 2.75 ± 0.06 and 3.24 ± 0.05 for SSNT and Tacotron, respectively, and the score difference was smaller. We also discovered that our method needed longer training time due to the marginalization process. The performance might be improved after using sufficiently long training periods.

4.3. Discussions

We designed the alignment structure to be hard and monotonic, which enabled us to avoid some alignment errors that are commonly observed in soft-attention-based approaches. Such alignment errors include muffling, skipping, and repeating. Muffling error is caused by an attention distribution without a sharp mode, skipping error is caused by discontinuous attention, and repeating error is caused by a repeat of backward jumps of the mode of attention. These errors were not observed in our method because they could not happen by definition.

However, we still observed different types of alignment errors in our samples, such as prolongations of vowel duration. We also observed that even though the alignment looked acceptable, that is, a monotonic increase occurred, wrong phonemes were sometimes generated due to a poorly trained model. When the alignment was not properly specified, its emission probability could be learned from different acoustic segments. This is a disadvantage of SSNT's hard alignments. However, we also found an advantage with the hard alignment. From our informal listening of generated speech, compared with soft-attention-based methods, speech generated from our method tended to have relatively distinct pronunciation.

5. Conclusions

We proposed a new method for end-to-end TTS without soft attention. In contrast with soft attention based methods, our method has a simpler architecture and ensures monotonic alignment structure by design. Our method represents an alignment variable as a latent variable, and the model can be trained by maximizing the total probability that can be derived by marginalizing the latent alignments. During inference, the alignment variable can be randomly sampled from the learned distribution, and the inference simply stops when the alignment reaches the final input position.

Thanks to the design of hard and monotonic alignment, our proposed method could avoid some alignment errors that were commonly observed in soft attention based approaches. Our method also replaced many engineering features in soft attention based approaches with a probabilistic approach, which included advanced attention mechanisms to enforce a monotonic alignment, stop flag prediction network, and nondeterministic inference by dropout.

Although our analysis revealed that some generated speech

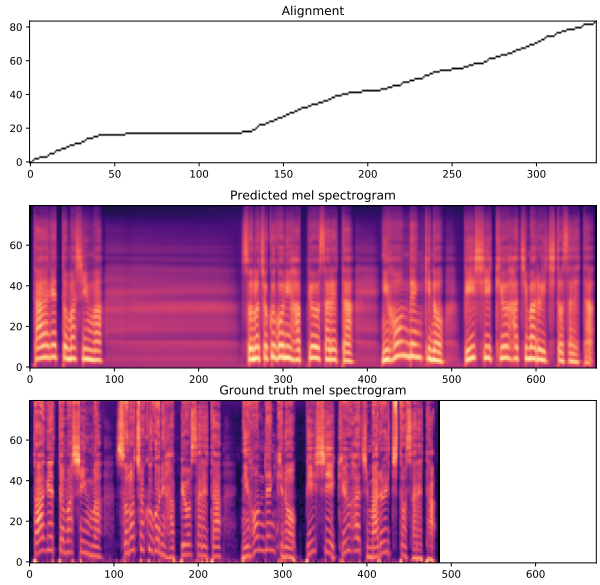

Figure 4: *A sample that shows overestimation of pause duration.*

contained another type of alignment errors, and although a subjective evaluation showed that the quality of synthetic speech from our system was not yet competitive with soft attention based methods on the reading speech corpus yet, our other research showed the effectiveness of the proposed method on a verbal performance corpus that is much more challenging data than reading speech [16].

Our future work includes performance optimization for faster training of SSNT-based TTS. We expect fast training will help to reduce the remaining alignment errors induced by insufficient training time. In addition, we plan to use a more complex probability distribution function for the emission probability of the acoustic features. In this study, we chose the isotropic Gaussian distribution for the emission probability. This was not an ideal choice because the target mel spectrogram had a clear correlation across frequency dimensions. We expect that full covariance modeling will enable random sampling of acoustic features and that a sufficiently complex probability distribution will lead to higher quality generated speech.

6. Acknowledgements

This work was partially supported by a JST CREST Grant (JPMJCR18A6, VoicePersonae project), Japan, and by MEXT KAKENHI Grants (16H06302, 17H04687, 18H04120, 18H04112, and 18KT0051), Japan. The numerical calculations were carried out on the TSUBAME 3.0 supercomputer at the Tokyo Institute of Technology.

7. References

[1] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proc. ICLR (Workshop Track)*, 2017.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[3] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "VoiceLoop: Voice fitting and synthesis via a phonological loop," in *Proc. ICLR*, 2018.

[4] W. Ping, K. Peng, A. Gibiansky, S. Ö. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep Voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, 2018.

[5] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461368

[6] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895

[7] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. ICLR*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1409.0473

[8] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proc. NIPS*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 577–585.

[9] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013. [Online]. Available: http://arxiv.org/abs/1308.0850

[10] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *Proc. ICASSP*. IEEE, 2018, pp. 4789–4793.

[11] 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, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 4784–4788. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461829

[12] L. Yu, J. Buys, and P. Blunsom, "Online segment to segment neural transduction," in *Proc. EMNLP*, J. Su, X. Carreras, and K. Duh, Eds. The Association for Computational Linguistics, 2016, pp. 1307–1316.

[13] L. Yu, P. Blunsom, C. Dyer, E. Grefenstette, and T. Kociský, "The neural noisy channel," in *Proc. ICLR*, 2017.

[14] A. Graves, "Sequence transduction with recurrent neural networks," *CoRR*, vol. abs/1211.3711, 2012. [Online]. Available: http://arxiv.org/abs/1211.3711

[15] K. Tokuda, K. Hashimoto, K. Oura, and Y. Nankaku, "Temporal modeling in neural network based statistical parametric speech synthesis," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 106–111. [Online]. Available: http://dx.doi.org/10.21437/SSW.2016-18

[16] S. Kato, Y. Yasuda, X. Wang, E. Cooper, S. Takaki, and J. Yamagishi, "Rakugo speech synthesis using segment-to-segment neural transduction and style tokens — toward speech synthesis for entertaining audiences," submitted to SSW10, 2019.

[17] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, May 2019, pp. 6905–6909.

[18] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. EMNLP*, L. Màrquez, C. Callison-Burch, J. Su, D. Pighin, and Y. Marton, Eds., 2015, pp. 1412–1421. [Online]. Available: http://aclweb.org/anthology/D/D15/D15-1166.pdf

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010. [Online]. Available: http://papers.nips.cc/paper/7181-attention-is-all-you-need

[20] C. Raffel, M. Luong, P. J. Liu, R. J. Weiss, and D. Eck, "Online and linear-time attention by enforcing monotonic alignments," in *Proc. ICML*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 2017, pp. 2837–2846.

[21] J. Eisner, "Inside-outside and forward-backward algorithms are just backprop (tutorial paper)," in *Proceedings of the Workshop on Structured Prediction for NLP@EMNLP 2016, Austin, TX, USA, November 5, 2016*, K. Chang, M. Chang, A. M. Rush, and V. Srikumar, Eds. Association for Computational Linguistics, 2016, pp. 1–17. [Online]. Available: https://doi.org/10.18653/v1/W16-5901

[22] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997. [Online]. Available: https://doi.org/10.1109/78.650093

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, J. Fürnkranz and T. Joachims, Eds., 2010, pp. 807–814.

[24] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: https://doi.org/10.1162/neco.1997.9.8.1735

[26] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, and K. Tokuda, "XIMERA: A concatenative speech synthesis system with large scale corpora," *IEICE Transactions on Information and System (Japanese Edition)*, pp. 2688–2698, 2006.

[27] H.-T. Luong, X. Wang, J. Yamagishi, and N. Nishizawa, "Investigating accuracy of pitch-accent annotations in neural-network-based speech synthesis and denoising effects," in *Proc. Interspeech*, 2018, pp. 37–41.

[28] D. Krueger, T. Maharaj, J. Kramár, M. Pezeshki, N. Ballas, N. R. Ke, A. Goyal, Y. Bengio, A. Courville, and C. Pal, "Zoneout: Regularizing RNNs by randomly preserving hidden activations," in *Proc. ICLR*, 2017.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2014.

[30] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016. [Online]. Available: http://arxiv.org/abs/1609.03499