Towards Speech Entrainment: Considering ASR Information in Speaking Rate Variation of TTS Waveform Generation

Mayuko Okamoto
Nara Institute of Science and Technology
Japan
okamoto.mayuko.oi1@is.naist.jp

Sakriani Sakti
Nara Institute of Science and Technology
RIKEN, Advanced Intelligence Project AIP
Japan
ssakti@is.naist.jp

Satoshi Nakamura
Nara Institute of Science and Technology
RIKEN, Advanced Intelligence Project AIP
Japan
s-nakamura@is.naist.jp

***Abstract*—State-of-the-art text-to-speech (TTS) systems successfully produce speech with a high degree of intelligibility. But TTS systems still often generate monotonous synthesized speech, unlike natural utterances. Several existing studies have addressed the issue of modeling speaking style variations in TTSs. Unfortunately, scant research has discussed the dialog and entrainment context. In this paper, we address TTS waveform generation toward speech entrainment in human-machine communication and focus on the synchronization of speaking rates that may vary within an utterance, i.e., slowing down to emphasize specific words and distinguish elements to highlight. We assume a dialog system exists and concentrate on its speech processing part. To perform such a task, we develop (1) a multi-task automatic speech recognition (ASR) that listens to the conversation partner and recognizes the content and the speaking rate and (2) a generative adversarial network (GAN)-based TTS that produces the synthesized speech of the response while entraining with the partner's speaking rate. The evaluation is performed on a dialog corpus. Our results reveal that it is possible to entrain the input speech by synchronizing the speaking rate.**

***Index Terms*—text-to-speech synthesis, automatic speech recognition, speech entrainment, speaking-rate variation, generative adversarial network**

I. Introduction

In natural human communication, a phenomenon called speech entrainment may occur in which speakers A and B unconsciously synchronize their speech utterances and their conversational styles start to resemble each other [1]. Fig. 1 shows an example of a human-machine dialog in which speech entrainment or synchronization occurs on the emphasized words. Both parties utter the emphasized words (in bold and underlined) in a similar style (i.e., slowing down their speaking rate). Manson et al. [2] found that when entrainment of the speech rate happens, in which the dyads' speech rates converged from a conversation's beginning to its conversation, the success rate of negotiation and cooperation is likely to increase. Furthermore, the speaking rate may significantly influence how the listener perceives the speech. We may speak much more quickly during an emergency and may slow down our speaking rate to emphasize what we are saying. Therefore, it is critical to develop a speech synthesis system that can produce natural spoken dialogues by considering the other party and phrase the message with an appropriate speaking rate based on the situation.

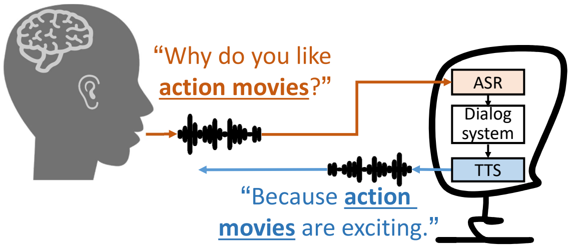
"Why do you like action movies?"
ASR
Dialog system
TTS
"Because action movies are exciting."

Fig. 1. Human-machine dialog in which speech entrainment or synchronization on emphasized words occurs (bold and underlined).

The resurgence of deep learning has revitalized the use of its frameworks for directly modeling text-to-speech synthesis (TTS) from the text-to-speech features. Several models have been proposed, such as an end-to-end deep neural network [3], [4], WaveNet [5], WaveRNN [6], and generative adversarial networks (GANs) [7]. They successfully produced speech with a high degree of intelligibility. The naturalness of generated speech has also significantly improved. But the speaking styles and speech expressions produced by the current TTS systems are typically averaged over training material that was mainly collected in reading-style speech; the utterances lack the variety and liveliness found in natural speech.

Several studies addressed this issue by modeling speaking style variations. Yoshimura et al. modeled the spectrum, pitch, and duration in a hidden Markov model (HMM)-based speech synthesis to generate speech that resembles various speaker's voices [8]. Yamagishi et al. addressed emotional expressivity and speaking style variability in an HMM-based speech synthesis [9]. But these techniques were based on the HMM framework [10]. For deep learning, Skerry-Ryan et al. [11] augmented an end-to-end Tacotron with explicit prosody

978-1-7281-9896-5/20/$31.00 ©2020 IEEE

controls for expressive speech synthesis. Unfortunately, only a few studies have addressed the speaking rate issue.

Recently, Wang et al. proposed "global style tokens" (jointly trained within Tacotron) that can control synthesis speech by varying the speed and speaking style [12]. However, since its information embedding is stored globally, controlling the duration at the word and phoneme levels is challenging. Although Park et al. introduced a mechanism for the phonemic-level duration within sequence-to-sequence frameworks [13], their system required phoneme input instead of text, which may be too complicated for actual users. Their proposed methods' effectiveness were also only evaluated through simulated data with unnatural speaking-rate variations. Their approach also failed to consider the effect on the other party. In fact, it is very difficult to find existing works that discuss the speaking style variation of TTS in the context of dialog and entrainment. Levitan et al. [14] is one of the few works that implemented acoustic-prosodic entrainment in TTS for a spoken dialog system. However, since the speech volume and speaking rate were assumed to be constant throughout all of the utterances, the speech feature controller is done at the utterance level and performed independently from the ASR and TTS processes.

In reality, humans vary their speaking rates and tend to slow down to emphasize words to distinguish elements of focus in an utterance. Therefore, in this paper, we address TTS waveform generation toward speech entrainment in human-machine communication and focus on the synchronization of speaking rates that may vary within an utterance. We simultaneously process both the content of the speech utterances and the additional information of speaking rates using multi-task ASR and multi-sources TTS.

II. Overview of Proposed Framework

Figure 2 shows our overall proposed framework. We are mainly interested in the speech processing part, assuming a dialog system exists. Therefore, our focus is on developing (1) a multi-task automatic speech recognition (ASR) that listens to its conversation partner and recognizes the content and the speaking rate, and (2) a generative adversarial network (GAN)-based TTS that produces the synthesized speech of the response while being entrained with the partner's speaking rate. Each of these components is described in the following sections.

A. Multi-task ASR system for speech recognition and speaking-rate identification

Instead of only recognizing what has been said, our proposed ASR (Fig. 2 left side) performs multi-task learning for recognizing the text transcription and speaking rate. It is based on a sequence-to-sequence framework [15], [16] that consists of an encoder, a decoder, and attention modules. It directly models the conditional probability of $P([\hat{\mathbf{Y}}, \hat{\mathbf{Z}}]|\mathbf{X})$, where $\mathbf{X}$ is a sequence of the framed speech features with length S and $[\hat{\mathbf{Y}}, \hat{\mathbf{Z}}]$ is a sequence of text transcription $\hat{\mathbf{Y}}$ and speaking rate $\hat{\mathbf{Z}}$ with length T. We gave speaking-rate information for each text output by the discretized symbol of three types speaking rates: "N" for standard read speech (normal), "S" for slow, and "F" for fast speech. For speaking-rate information in the phoneme level, the ASR was trained to output phoneme sequences.

In this network, the encoder transforms input speech sequence $\mathbf{X}$ to hidden representative information $\mathbf{h}^e = [h^e_1, ..., h^e_S]$, and the decoder predicts target sequence probability $[\hat{\mathbf{Y}}, \hat{\mathbf{Z}}]$, given the previous output, current context information c_t, and current decoder hidden state h^d_t. Context information c_t is produced by attention modules [17] at time *t* based on the encoder and decoder hidden states.

B. GAN-TTS for speech generation with variable speaking rates considering ASR input

Our proposed TTS is based on a GAN-TTS framework [7] (Fig. 2, right side). However, instead of receiving only the text transcription as in a standard GAN-TTS approach, our proposed TTS has multi-sources input, which is a sequence of text transcription $\mathbf{Y}$ and speaking rate $\mathbf{Z}$ with length T. Then the task becomes producing speech acoustic features $\hat{\mathbf{X}}$ with the defined speaking rate.

Since many contextual factors (e.g., phoneme identity, word stress, etc.) might affect the speech's prosodic characteristic, we first generate a full-context label from a given text, which is the most common approach in standard HMM-based and GAN-based TTS frameworks. This is done by a front-end text processing block that extracts the linguistic features from a given input text.

As described in Fig. 2, a front-end text processing block extracts the linguistic features from a given input text. Since many contextual factors (e.g., phoneme identity, word stress, etc.) might affect speech's prosodic characteristic, generating a full-context label from a given text is the most common way in standard GAN-TTS, which is also well-known in a HMM-based TTS framework. Fig. 3 shows an example of a full-context label that is comprised of the following factors:

- Phoneme level:
  - $p_1, ..., p_5$: {second preceding, preceding, current, succeeding, second succeeding} phoneme;
  - p_6, p_7: position of current phoneme in the current word (forward and backward);
- Syllable level:
  - $a_1, ..., a_3$: {type of syllable stress, number of phonemes} in the preceding syllable.
  - $b_1, ..., b_6$: {type of syllable stress, number of phonemes, position in word and phrase, number of syllables before and after} in the current syllable.
  - $c_1, ..., c_3$: {type of syllable stress, number of phonemes} in the succeeding syllable.
- Word level:
  - d_1, d_2: {part-of-speech, number of syllables} in the preceding word.
  - $e_1, ..., e_8$: {part-of-speech, number of syllables, position in phrase, number of content words before and after} in the current word.

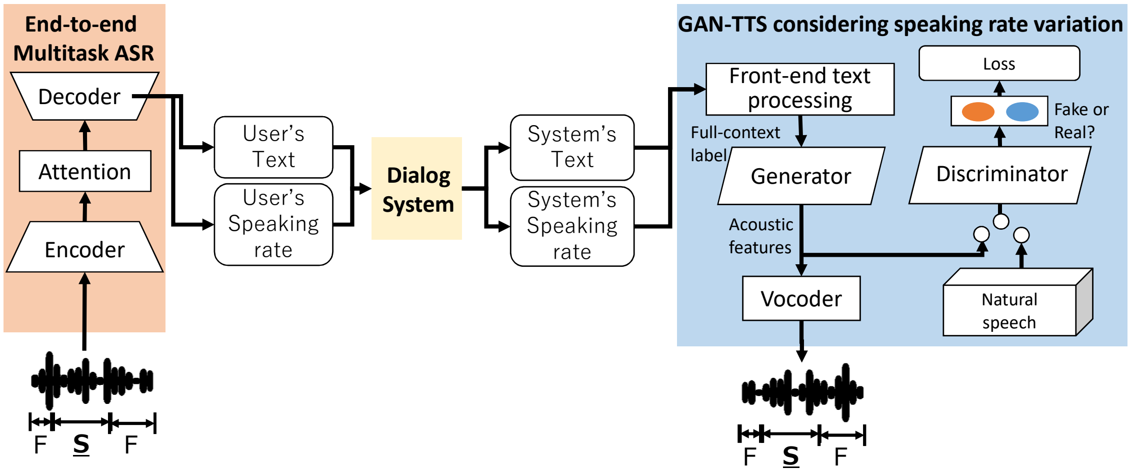

Fig. 2. Overview of the proposed framework.

$p_1 * p_2 \text{-}\ p_3 + p_4 = p_5 @ p_6 \text{-} p_7$
/A:$a_1\_a_2\_a_3$
/B:$b_1\text{-}b_2\text{-}b_3@b_4\text{-}b_5\&b_6\text{-}b_7\#b_8\text{-}b_9\$b_{10}\text{-}b_{11}!b_{12}\text{-}b_{13};b_{14}\text{-}b_{15}|b_{16}$
/C:$c_1+c_2+c_3$
/D:$d_1\_d_2$
/E:$e_1+e_2@e_3\ +e_4\&e_5+e_6\ \#e_7+e_8$
/F: $f_1\text{-}f_2$
/G:$g_1\text{-}g_2$
/H:$h_1=h_2\ @h_3=h_4|h_5$ /I:$i_1=i_2$
/J:$j_1+j_2\text{-}j_3$

Fig. 3. Example of full-context label

  - f_1, f_2: {part-of-speech, number of syllables} in the succeeding word.
- Phrase level:
  - g_1, g_2: number of {syllables and words} in the preceding phrase.
  - $h_1, ..., h_5$: {number of syllables and words, utterance position, TOBI endtone} of the current phrase.
  - i_1, i_2: number of {syllables and words} in the succeeding phrase.
- Utterance level:
  - $j_1, ..., j_3$: number of {syllables, words, and phrases} in the utterance;

To achieve a GAN-TTS that controls the speaking rate variations at the phoneme level, we incorporate the information of the speaking-rate variations ("N," "S," and "F") within the phoneme symbols of the full-context label by directly attaching it to the phoneme label in the full-context label by specifically modifying the $p_1, ..., p_5$ label into $p_1 + N/S/F, ..., p_5 + N/S/F$. For example, the pentaphone label of "$hello \rightarrow (hh, eh, l, ow)$" is

$$pau * hh - eh + l = ow.$$

With slow-speaking-rate information, it becomes

$$pau * hhS - ehS + lS = owS.$$

The "pau" for the silent part ignores the speaking-rate information, since there is no change in the generated acoustic features regardless of the speed information.

After that, we trained the GAN-TTS based on the text transcription and the incorporation of the speaking-rate information. The framework consists of two types of neural networks: generator G and discriminator D. Its training procedure is employed by an adversarial process in which the two models (generator G and discriminator D) compete. In other words, the generator learns to create the speech output that causes the discriminator to misrecognize the generated result as natural speech with a speaking-rate variation, and the discriminator learns to accurately distinguish between natural and synthetic speech produced by generator G. Further details of GAN-TTS technology are available [7].

III. Experimental Setup

We used several types of data within this study. The first one is the CMU ARCTIC database that was constructed at the Carnegie Mellon University. It consists of phonetically balanced sentences (1132 speech utterances) uttered by a US English speaker at a normal reading speed [18]. To construct data with speaking-rate variations, we artificially modified the speaking rate using SoundExchange (Sox)[1] software by slowing down the original rate by 75% and speeding it up by 125%, resulting in "slow" and "fast" sample data. We chose these parameters where the resulting slow and fast speech still sounded natural. Next we recorded the natural speech data uttered by one female and one male who spoke as naturally as possible and simultaneously produced speech with three different speaking rates, as in the data samples. We got 6,792 utterances (two speakers and three speaking rates) for analysis as well as model learning and evaluation. We denote this data as CMU ARCTIC SPK-RATE. Further details of the data construction and analysis are available in [19]. In this study, to train and evaluate our GAN-TTS we used 3396 utterances

[1] Sox – http://sox.sourceforge.net/

from the female speaker, including 3003 utterances for the training set, 378 utterances for the development set, and 15 utterances for the test set.

Since the CMU ARCTIC SPK-RATE data remain too limited for training our multi-task ASR, we made a larger dataset based on a basic travel expression corpus (BTEC) [20]. We prepared 51,500 sentences of the English BTEC text and utilized our GAN-TTS to generate speech waveforms with three speaking rates: slow, normal, and fast. Then with 3396 natural speech utterances CMU ARCTIC SPK-RATE, we trained the proposed multi-task ASR, which used 157,896 speech utterances (training set: 148350 utterances, development set: 8031 utterances, and test set: 1515 utterances).

Last, to investigate the effectiveness of the proposed waveform generation toward speech entrainment in human-machine communication, we also utilized the Coached Conversational Preference Elicitation dataset that consists of 502 English dialogs with 12,000 annotated utterances between a user and an assistant who discussed movies in natural language (CCPE-M dataset) [21]. It was collected using Wizard-of-Oz methodology between two paid crowd-workers: an assistant and a user. We selected the dialog conversations between assistant and user that shared the same nouns and tagged them as "slow" and the other words as "fast" to simulate the speaking rates that change within a single utterance. After that, we utilized the previously trained GAN-TTS to generate a speech waveform that slowed down to emphasize the nouns as the elements of focus in an utterance.

Our multi-task ASR system is an attention-based encoder-decoder model [15], [16] that consists of three stacked bidirectional long short-term memory (BiLSTM) encoders, a single layer LSTM, and multi-layer perceptron (MLP)-based attention [22] components. Log-scaled Mel-spectrograms were fed into a fully connected layer and transformed by a LeakyReLU ($l = 1e-2$) [23] activation function. This model doesn't need any language model or word dictionary. For the TTS system, we followed the PyTorch implementation of GAN-TTS [7].

IV. Experiment Results

A. ASR Performance

First, we investigated how the additional task of speaking-rate identification affected the speech recognition performance. We used the test set described in the previous section. The phoneme error rate (PER) of the standard ASR was 11.68%, and the PER of our proposed multi-task ASR was 12.04%. Although the proposed ASR is a multi-tasking ASR, the phoneme sequence can be estimated with almost the same accuracy as a standard single-task ASR. The results reveal that the additional task did not significantly affect the performance of the speech recognition part, which shows a positive indication.

The error rate of the speaking-rate identification was 27.15%, which exceeds the phoneme sequence recognition result. The error matrix is shown in Table I. Note that the errors in it were calculated by extracting only the sentences with the same number of correct speaking-rate labels and recognition speaking-rate labels. In other words, we excluded the sentences in which the number of phonemes increased or decreased compared to the correct sentence due to phoneme recognition errors. <spc> is a delimiter tag between words. Based on Table I, there was no case where the slow-speaking rate was incorrectly recognized as the fast speaking rate or vice versa. However, many utterances in the normal speaking rate were mistakenly recognized as fast- or slow-speaking rates because while recording the normal speech, the speaker still sometimes uttered the sentence slightly slower or faster. Labels based on the dynamic speaking style of the speakers might be necessary.

TABLE I
Confusion matrix of speaking-rate recognition performance from proposed multi-task ASR.

res / ref	Fast	Normal	Slow	<spc>
Fast	**99.584** %	0.098 %	0 %	0.318 %
Normal	28.790 %	**50.484** %	19.436 %	1.290 %
Slow	0 %	0.024 %	**99.903** %	0.073 %
<spc>	0.350 %	0.489 %	0.033 %	**99.127** %

B. TTS Performance

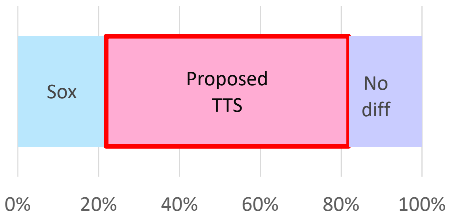

Fig. 4. Results of subjective evaluation on naturalness: Sox baseline versus proposed method.

Next we evaluated our proposed GAN-TTS model. Given an input sentence, we first generated full-context labels using part of the tools from the HMM/DNN-based Speech Synthesis System (HTS) [24], [25]. After that, we included speaker variation information into the labels based on the proposed methods described in Section II-B. For comparison, we applied Sox, which changed the speed on the synthesized output produced with the "normal" data as the baseline.

We used a preference (AB) test to evaluate the performance and subjectively assessed the speech's naturalness. 11 subjects (7 males, 4 females) participated. From speech utterances in the test set, we showed them paired-by-paired with a random order, and asked them to answer which voice sounded more natural: A, B, or no difference (denoted as "No diff"). The results in Fig. 4 indicate that the synthesized speech utterances from the proposed method are more natural than those from the baseline.

C. TTS considering ASR outputs within dialog context

Last, we investigate the waveform generation in the context of dialog and entrainment and focus on the synchronization

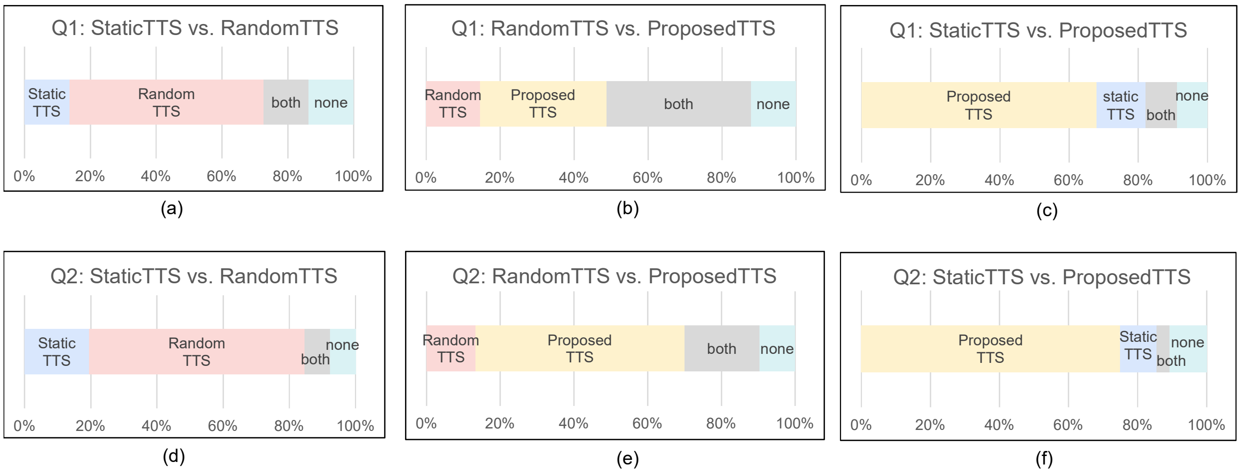

Fig. 5. Results of subjective evaluation on speaking-rate synchronization: StaticTTS and RandomTTS baseline versus ProposedTTS.

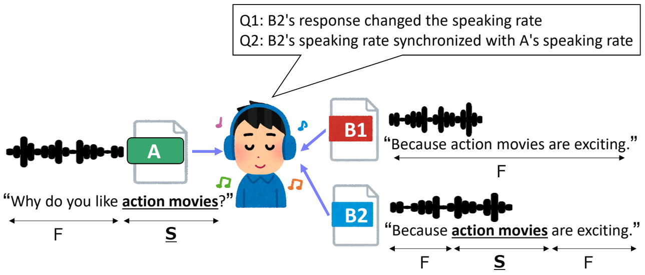
Q1: B2's response changed the speaking rate
Q2: B2's speaking rate synchronized with A's speaking rate
A
"Why do you like action movies?"
F
S
B1
"Because action movies are exciting."
F
B2
"Because action movies are exciting."
F
S
F

Fig. 6. ABX preference test procedure: subjects listened to speech utterances from speaker A, and then two responses from speaker B (B1 and B2). We asked them to answer which voice B has speaking rate variation and entrain to voice A: B1, B2, both of them, or none of them.

of speaking rates that may vary within an utterance, i.e., slowing down to emphasize specific words and distinguish elements of focus in the utterance. Since our interest lies in the speech processing part, we assumed the dialog system has no errors. We used the CCPE-M dataset for the evaluation and denoted the human as the assistant and passed the conversation speech utterances to the ASR and the machine as the user and passed the corresponding text responses to the GAN-TTS and generated the speech waveforms.

For comparison, we also developed the following: (1) a standard GAN-TTS that generated speech waveform without considering a speaking-rate factor and denoted it as StaticTTS; (2) a GAN-TTS that generated speech waveform while randomly varying the speaking rate and denoted it as RandomTTS. The evaluation was done with another preference (ABX) test to subjectively assess whether speech entrainment can be perceived. 15 subjects (11 males, 4 females) participated in the experiments. Fig. 6 illustrates ABX preference test procedure. Assuming there is a dialog conversation between speaker A and B, subjects listened to speech utterances from speaker A, and then two responses from speaker B (B1 and B2), and we asked them two questions:

- Q1: Which response of Speaker B changed the speaking rate (B1, B2, both of them, or none of them)?
  This question confirmed whether the subjects could perceive speaking-rate variation within a single utterance.
- Q2: Which response of Speaker B reflected the speech entrainment of the speaking rate of speaker A (B1, B2, both of them, or none of them)?
  This question confirmed whether subjects could perceive the synchronization of speaking rate between utterances from speakers A and B.

The results are shown in Fig. 5 where (a),(b), and (c) are the results for $Q1$, and (d), (e), and (f) are the results for $Q2$. We expected the subjects to perceive the speaking-rate variation on the speech waveform generated by the RandomTTS and the ProposedTTS without perceiving the speech waveform generated by the StaticTTS. However, when we compared the StaticTTS and RandomTTS for Q2 (Fig. 5(d)), the subjects still perceived speech entrainment with RandomTTS. This indicates that varying speaking rates within a single utterance with the same style as speaker A might still be useful even if the emphasized words were not synchronized. But from Fig. 5(e), the results indicate that the synthesized speech utterances from the proposed method still outperformed the RandomTTS baseline in reflecting speech entrainment on the speaking rate of speaker A.

V. Conclusion

We examined speech synthesis that controls speaking rates based on the speaking rate of another party for a speech dialogue system that can communicate more naturally. First, we proposed a multi-tasking ASR for utterance and speech rate recognition using a sequence-to-sequence model and showed that texts can be recognized with almost the same accuracy as an ordinary ASR that did not recognize the speaking rate. Next we proposed a GAN-TTS that can control the speaking rate in phoneme units. A subjective evaluation of the naturalness of speech showed that the proposed method generated more natural speech than artificially manipulating the waveform of the synthesized speech. Finally, we proposed a GAN-TTS for speech generation that changes the speaking rate at the phoneme level based on the ASR output. Our results revealed that the proposed method outperformed the baseline in reflecting the speech entrainment on the partner's speaking rate. In the future, we will apply our proposed ASR and TTS in a complete spoken dialog system and investigate the effect of speech entrainment within human-machine communication.

Acknowledgment

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101 and JP17K00237.

References

[1] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2012, pp. 11–19.

[2] J. H. Manson, G. A. Bryant, M. M. Gervais, and M. A. Kline, "Convergence of speech rate in conversation predicts cooperation," *Evolution and Human Behavior*, vol. 34, no. 6, pp. 419–426, 2013.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013, pp. 7962–7966.

[4] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[5] A. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[6] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[7] 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, p. 84–96, 2018.

[8] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proceedings of Eurospeech*, 1999, p. 2347–2350.

[9] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis," *IEICE Transactions on Information and Systems*, vol. E88D, pp. 502–509, 03 2005.

[10] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation from HMM using dynamic features," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995, p. 660–663.

[11] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[12] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," *arXiv preprint arXiv:1803.09017*, 2018.

[13] J. Park, K. Han, Y. Jeong, and S. W. Lee, "Phonemic-level duration control using attention alignment for natural speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 5896–5900.

[14] R. Levitan, S. Benus, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar," in *Proceedings of INTERSPEECH*, 2016, pp. 1166–1170.

[15] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Proceedings of the Advances in neural information processing systems (NIPS)*, 2015, p. 577–585.

[16] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," pp. 4960–4964, 2016.

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

[18] J. Kominek and A. Black, "The CMU arctic speech databases," in *Proceedings of SSW*, 2004.

[19] "Phoneme-level speaking rate variation on waveform generation using gan-tts."

[20] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto, "Creating corpora for speech-to-speech translation," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH)*, 2003, pp. 381–384.

[21] F. Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi, "Coached conversational preference elicitation: A case study in understanding movie preferences," in *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*.

[22] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[23] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," *arXiv preprint arXiv:1505.00853*, 2015.

[24] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2000, p. 1315–1318.

[25] "The HMM-based speech synthesis system (HTS)," http://hts.ics.nitech.ac.jp.