Modeling of Rakugo Speech and Its Various Speaking Styles: Toward Speech Synthesis That Entertains Audiences

SHUHEI KATO¹,², (Student Member, IEEE), YUSUKE YASUDA¹,², (Student Member, IEEE), XIN WANG², (Member, IEEE), ERICA COOPER², (Member, IEEE), SHINJI TAKAKI³, (Member, IEEE) AND JUNICHI YAMAGISHI²,⁴, (Senior Member, IEEE)

¹The Graduate University for Advanced Sciences (SOKENDAI), Hayama, Kanagawa, Japan
²National Institute of Informatics, Chiyoda, Tokyo, Japan
³Nagoya Institute of Technology, Nagoya, Aichi, Japan
⁴The University of Edinburgh, Edinburgh, UK

ABSTRACT We have been working on building rakugo speech synthesis as a challenging example of speech synthesis that entertains audiences. Rakugo is a traditional Japanese form of verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling and is popular even today. In rakugo, a performer plays multiple characters, and conversations or dialogues of the characters make the story progress. We modeled rakugo speech using the state-of-the-art Tacotron 2 and an enhanced version of it with self-attention to better consider long-term dependency. We also used global style tokens and manually labeled context features to enrich speaking styles. Through a listening test, we found that the speech synthesis models could not yet reach the professional level, but interestingly, some of the synthetic speech entertained the listeners as well as analysis-by-synthesis speech. Although there is room for improvement, this is an important stepping stone toward realization of entertaining speech synthesis at the professional level.

INDEX TERMS context, entertainment, global style tokens, rakugo, self-attention, speech synthesis, Tacotron, text-to-speech

I. INTRODUCTION

CAN machines read texts aloud like humans? The answer is yes, albeit under limited conditions. The mean opinion scores (MOSs) of some speech synthesis (text-to-speech; TTS) systems have already been the same as those of natural speech [1], [2]. These systems are trained with well-articulated read speech. Attempts to model speech with various speaking styles have also been actively investigated in deep-learning-based speech synthesis study [3]–[9].

However, can machines verbally entertain like humans? The answer is probably no. Verbal entertainment, including rakugo, which is a traditional Japanese form of verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling, entertains audiences through the medium of speech. In other words, speech in verbal entertainment does not just transfer information such as content and speaker emotions, personality, intention, etc. to listeners, but also stirring listeners’ emotions. Most of us would agree that verbal-entertainment performances by machines are quite unnatural or monotonic even if the content is appropriate. Some speech-synthesis-based rakugo performances, mostly including many manual interventions, have been submitted to online video platforms [10]–[12]. You might enjoy such performances depending on your view or subjectivity, but we think these performances have far poorer quality than those by professional rakugo performers. We believe such a gap between machines and humans should be filled to evolve human-machine relationships. To the best of our knowledge, this study, including our previous ones [13], [14], is the first machine-learning-oriented fundamental research on modeling rakugo speech for speech synthesis.

Some readers may wonder how different rakugo speech
is compared with that of audiobooks, which is an active research topic in the speech synthesis field. The main difference is that almost all parts of a rakugo story consist of conversations and dialogues of characters that are played by a performer from memory, and the conversations and dialogues make the story progress. There are few narrative sentences in the conversational parts. Therefore, we believe that improving rakugo speech synthesis would also help improve the quality of conversational speech synthesis. It should be noted that rakugo speech is more casually-pronounced than that of audiobooks because rakugo is performed from memory. In addition, expression of rakugo speech is far more diverse than that of audiobooks because conversations and dialogues of characters describe the entire story. Moreover, the Japanese language used in traditional rakugo stories is somewhat old-fashioned, and each character speaks a different dialect, sociolect, and idiolect according to his/her gender, age, social rank, and individuality. These casualness and diversity make it difficult to properly train a model. Since rakugo is a form of entertainment, we argue that whether audiences are being entertained is important. In other words, rakugo speech synthesis has to inherently entertain audiences and stir their emotions, whereas that of audiobooks has to at least transfer its content to listeners.

Before modeling rakugo speech, we first recorded and built a large rakugo speech database for our experiments because most commercial rakugo recordings are live recordings that contain noise and reverberation; thus, no rakugo speech databases suited to speech synthesis was available.

Using this database, we modeled rakugo speech with two systems, Tacotron 2 [1] and an enhanced version of it with self-attention [15] (SA-Tacotron [16]). Tacotron 2 is the state-of-the-art speech synthesis system that can produce (read-aloud) speech as natural as human’s speech, as mentioned above. It has been reported that Tacotron-based systems can also model audiobook speech well. We introduced self-attention for further improvement. It has been reported that self-attention helps consider long-term dependency, and the combination of Tacotron and self-attention outperforms the original Tacotron [17] in pitch-accent languages such as Japanese [16]. We therefore used SA-Tacotron in this study. We also combined global style tokens (GSTs) [4] and/or manually labeled context features with Tacotron 2 and SA-Tacotron to enrich the speaking styles of synthetic rakugo speech.

This paper describes the continuation of research began in our previous study [14]. The differences from that study [14] and this one are as follows: in this study, 1) Tacotron-based systems were changed to have strictly the same architecture as the original Tacotron 2 [1] and implementation was refined, 2) SA-Tacotron was newly introduced, 3) systematic comparisons were conducted on Tacotron 2 and SA-Tacotron with and without the combination of GSTs and/or manually labeled context features, and 4) a more detailed listening test was conducted, especially asking listeners to answer how well they think they were entertained.

The rest of this paper is organized as follows. In Section II, we give an overview of rakugo. In Section III, we give the details of our rakugo speech database. In Section IV, we explain Tacotron 2 and introduce SA-Tacotron. In Section V, we present the experimental conditions, results, and discussion of a listening test and error analysis. We conclude the paper and mention future work in Section VI.

II. RAKUGO

A. OVERVIEW

Rakugo is a traditional Japanese form of one-person verbal entertainment similar to a combination of one-person stand-up comedy and comic storytelling. It has over 300 years of history, and is a popular form of entertainment even today in Japan. A theater that mainly produces rakugo is called a *yose*. In Tokyo, there are four major yoses, and rakugo is performed in each one every day of the year, even on January 1 (Figure 1). There are many other minor yoses. Rakugo is also performed at small to large halls, restaurants, coffeehouses, bookstores, shrines, temples, etc. almost every day. Thousands of CDs, DVDs, and streaming audio/videos of rakugo performances by present or former professional rakugo performers are available. Some TV and radio programs are broadcasted every week in Japan [22–28]. Amateur performances are also active. Some amateur rakugo performance societies at universities have produced professional performers.

Rakugo is generally divided into *Edo* (Tokyo) rakugo, which we focus on in this paper, and *Kamigata* rakugo, which has been developed in Osaka and Kyoto. A professional
FIGURE 2. Shumputei Shotaro [29], who is a professional rakugo performer, performing rakugo on a stage [30].

A rakugo performer is called a hanashika. In Edogawa rakugo, a hanashika is ranked at either of three levels, termed zenza (minor performer, assistant of stages, and housekeeper at their master/mistress’s house), futatsume (second-rank performer), and shin-uchi (first-rank performer). Only shin-uchis can take disciples. Usually, it takes about 3 to 5 years to be promoted from zenza to futatsume, and about 10 years to be promoted from futatsume to shin-uchi. About 600 performers are active as professionals in Edogawa rakugo as of 2019.

B. PERFORMANCE
During a performance, a rakugo performer sits down on a zabuton (cushion) and performs improvisationally or from memory alone on a stage (Figure 2). He/she plays multiple characters, and their conversations and dialogues make the story progress. In the main part of a story, to be mentioned later, almost all of the parts consist of conversations and dialogues between the characters played by the performer. In Edogawa rakugo, performers use only a sensu (folding fan) and a tenugui (hand towel) as props.

Rakugo performers tell stories in their performances. A rakugo story is composed of five parts: maeoki (greeting), makura (introduction), the main part, ochi (punch line), and musubi (conclusion) [31]. Maeoki is optional, so it may not appear during a performance. Some exceptional stories have musubi in place of ochi. Musubi is also used when performers terminate stories because of time limitations. Makura is often improvised, but during this, performers basically do not have conversations with the audience, unlike stand-up comedy. Ochi, the punch line, is the most important part of rakugo (the word “rakugo” is derived from “a story with ochi”).

Rakugo stories are generally divided into standards, which were established by about the 1920s, and modern stories, which were created after the 1930s. In this paper, we focus on standards. It should be noted that the Japanese language used in standards is slightly old-fashioned, and each character speaks a different dialect, sociolect, and idiolect of Japanese according to his/her gender, age, social rank, and individuality.

The length of rakugo stories varies from story to story. Even if performers play the same story, the length can be varied from stage to stage due to time limitations or other situations. In a yose, one hanashika usually performs for about 15 minutes (only the last performer performs for about 30 minutes). In other stages or recordings, they may perform longer.

Rakugo stories are taught through oral instruction from a master/mistress to a disciple except in the case that the story was newly created. Performers may edit stories to increase the quality or match their own characteristics. They sometimes insert jokes not only in the makura but also in the main parts of the stories according to the circumstances during their performances.

The following is an example of a very short rakugo paragraph.

Tomi: Whoa! Oh no! Oh no! Oh no! Oh no!
Friend: Wait Tomi. What are you doing?
Tomi: Oh, I’m chasing after a thief.
Friend: Seriously? Aren’t you the fastest man in this town? He is unlucky.
Friend: Which direction did he escape?
Tomi: He’s catching up with me.

III. DATABASE
A. OVERVIEW
We built a rakugo speech database for our experiments because there was no rakugo speech database suited to speech synthesis. Most commercial rakugo recordings are live recordings that include noise and reverberation; therefore, we recorded the rakugo speech ourselves.

The recordings were conducted from July to September 2017. The performer was Yanagiya Sanza [32], a professional rakugo performer with over 20 years of experience and was promoted to shin-uchi in 2006. Only he was in the recording booth, and he did not face or receive any reaction from an audience (Figure 3). He performed 25 Edogawa rakugo standards, lasting from 6 to 47 minutes length (total 13.2 hours including pauses between sentences). We did not re-record any of the performance due to mispronunciation or restatements except in cases where the performer asked us to do so.

The first author carefully transcribed the pronunciation of the recorded speech. We did not define any special symbols for mispronunciation, fillers, or laughs. We used a comma only at a pause in a sentence, a period at the end of a sentence, and a question mark at the end of a sentence that ends with a quotation mark.
FIGURE 3. Yanagiya Sanza performing rakugo alone in recording booth.

TABLE 1. Symbols used in transcription of rakugo database.

| Phonemes | Vowels | a, e, i, o, u |
|----------|--------|--------------|
| Consonants | b, by, ch, d, dy, f, fy, g, gy, h, hy, j, k, kw, ky, m, my, n, N, ng, ny, p, py, r, ry, s, sh, t, ts, ty, v, w, y, z |
| Other | cl (stop consonant) |
| Pauses | pau (comma), sil (start of a sentence and period), qsil (question mark) |

with rising intonation. The ratio of question sentences, those having a question mark at the end, to the other sentences is about 3:7. We separated sentences according to the following criteria.

- A place we can separate sentences grammatically followed by a pause.
- A place where a turn-taking occurs.
- A place right after a rising intonation.

All the symbols used in the transcription are listed in Table 1. We did not use any accent symbols, although Japanese is a pitch-accent language, because the results of automatic morphological analysis and accent estimation are not usable due to the difference between the slightly old-fashioned Japanese dialects, which is spoken by characters in the stories, and the modern standard Japanese. Of course, manual labeling of accents is impractical because it is time-consuming.

B. CONTEXT LABELS

We also appended context labels to each sentence (Table 2). All the labels, excluding part, were defined by the first author because no well-known categories of them exists in rakugo.

We believe the role of the character is important because almost all speech in rakugo, especially in the main part, is composed of conversations or dialogues. The individuality of the character is a special category for fool characters, usually called Yotaro, who often appear in rakugo stories. We believe the condition of characters is also important because characters speak in various styles. All the styles were defined by the first author via carefully listening to speech and reading context. The relationship of the talking companions was defined because in conversations or dialogues (of two characters) in rakugo, one must be considered the superior and the other as the inferior. The n_companion (number of the talking companions) was defined because characters may talk to themselves or speak to one person or multiple persons. The distance to the talking companions was defined because characters may speak to someone near or far from them. In the context of a particular part of the story, we considered maeoki (greetings) and musubi (conclusion) as makura (introduction) and ochi (punch line), respectively.

IV. TACOTRON 2 / SA-TACOTRON BASED RAKUGO SPEECH SYNTHESIS

A. TACOTRON 2

Tacotron 2 [1] is the state-of-the-art speech synthesis system, which produces read-aloud speech as natural as human’s speech. Some Tacotron-based systems can model expressive speech including audiobook’s speech well [4], [5], [7]. We therefore argue that modeling rakugo speech using Tacotron 2 is reasonable.

The architecture of the Tacotron-2-based rakugo TTS model is shown in Figure 4. This model is slightly modified from the original one to learn alignments more robustly and faster.

The main network is composed of an encoder, a decoder, and an attention network, which maps each time step of the encoder and that of the decoder. The encoder converts an input phoneme sequence into a hidden feature representation, which will be taken to predict an output mel spectrogram. Each input phoneme is embedded into a 512-dimensional vector. The vector can then be concatenated to a style embedding, which is described in IV-C, a context embedding, which is derived from the input context labels listed in Table 2, or concatenated vector of them. The (concatenated) vector is passed into 3 convolutional neural networks (CNNs) each containing 512 filters with a $5 \times 1$ shape (same padding [33]), followed by batch normalization [34] and rectifier linear units (ReLU) activations. The output of the final CNN is then passed into a single bi-directional [35] long short-term memory (LSTM) [36] that has 512 units (256 units per direction) to generate the final encoded features.

The output sequence of the encoder is used by an attention network that compresses the full encoded sequence as a fixed-length context vector for each decoder time step. We used the forward attention with transition agent [37] instead of the location sensitive attention [38], which is used in the original Tacotron 2, to learn the alignment between the
TABLE 2. Context labels (*hanashika* refers to speech not by any characters).

| Group               | Name                  | Details                                                                 |
|---------------------|-----------------------|-------------------------------------------------------------------------|
| ATTR (attribute)    | role of character     | gender: *hanashika*, male, female; age: *hanashika*, child, young, middle-aged, old; social rank: *hanashika*, samurai, artisan, merchant, other townsperson, countryperson, with other dialog, modern, other  |
|                     | individuality of character | *hanashika*, fool                                                           |
| COND (condition)    | condition of character | neutral, admiring, admonishing, affected, angry, begging, buttering up, cheerful, complaining, confident, confused, convinced, crying, depressed, drinking, drunk, eating, encouraging, excited, fearing, feeling sketchy, feeling sick, feeling sleep, feeling sorry, feeling suspicious, find it easier than expected, freezing, frustrated, ghostly, happy, hesitating, interested, justifying, *kakegoe* (shout/call), loud voice, laughing, leaning on, lecturing, looking down, panicked, pet directed speech, playing dumb, putting up with, rebellious, refusing, sad, seducing, shocked, shouting, small voice, soothing, straining, surprised, swaggering, teasing, telling off, tired, trying to remember, underestimating, unpleasant |
| SIT (situation)     | relationship to talking companions | *hanashika*, narrative, soliloquy, superior, inferior |
|                     | n_companion: number of talking companions | *hanashika*, narrative, soliloquy, one, two or more |
|                     | distance to talking companions | *hanashika*, narrative, near, middle, far |
| STR (structure)     | part of story         | makura, main part, ochi                                                  |

**FIGURE 4.** Overall network structure of Tacotron-2-based rakugo TTS. Network structure of reference encoder and style token layer is shown in the Figure 6.

code and decoder time steps more robustly and faster. The forward attention algorithm has a left-to-right initial alignment, which is useful because all the alignments of encoder-decoder TTS should be left-to-right because the output speech has to be produced from the beginning to the end of the input text. For more details, please refer to [37].

In the decoder, the predicted mel spectrum in the previous time step is first passed into a pre-net, a feed-forward network (FFN) that has 2 fully-connected layers of 256 ReLU units. The pre-net output and the context vector from the attention network at the previous time step are concatenated and passed into 2 unidirectional LSTMs each containing 1,024 units. Using the output sequence of the LSTMs and encoder output, the context vector at the current time step is calculated. Then the concatenation of the output of the LSTMs and the context vector is passed into an FFN that has a fully-connected layers of 80 linear units to generate a mel spectrum. To predict a residual for improving the reconstruction, the predicted spectrum is passed into a post-net, 5 CNNs each containing 512 filters with a $5 \times 1$ shape (same padding) followed by batch normalization. Then tanh activations are applied except for the final layer. The summation of the former predicted mel spectrum and the output of the post-net is the final target mel spectrum.

In parallel with the spectrogram prediction, the concatenation of the output of the decoder LSTMs and the attention vector is projected to a scalar and activated by a sigmoid function to predict the probability of the completion of the
output sequence. This probability is called “stop token”.

Training is conducted through minimizing the summation of the mean squared errors (MSEs) of the spectrograms both before and after the post-net, the binary cross entropy loss of the stop token, and the L2 regularization loss. During training, dropout [39] is applied to each layer of the FFN in the pre-net and the CNNs in the post-net with probability 0.5 for regularization. Zoneout [40] is also applied to each LSTM with probability 0.1 for further regularization.

B. SA-TACOTRON

For further improvement, we enhanced the Tacotron 2 above with self-attention [15] (SA-Tacotron). Self-attention can capture long-term dependency well, and it was reported that the enhanced version of Tacotron with self-attention exceeds the original Tacotron [17] in pitch-accent languages such as Japanese [16]. We therefore enhanced the Tacotron 2 above with self-attention. The network structure is shown in Figure 5.

The network structure is similar to that presented in [16], except that the encoder and decoder used in our study are based on Tacotron 2, while those in [16] are based on Tacotron.

In the encoder, a self-attention block is inserted after the bi-directional LSTM. A self-attention block consists of a self-attention, followed by a fully-connected layer with tanh activation and residual connections. This block is expected to capture the long-term dependency inside the input phoneme sequence. The number of heads in the multi-head attention [41] in the self-attention is 2, and the dimension is 32. During training, dropout is applied to the multi-head attention with probability 0.05 for regularization. The encoder generates two output sequences, one is from the bi-directional LSTM, and the other from the self-attention block.

The output sequences of the encoder are input into two attention networks. The output sequence from the bi-directional LSTM is used by a forward attention with transition agent, the same architecture as the Tacotron 2 described in [IV-A]. On the other hand, the output sequence from the self-attention block is used by an additive attention [42]. The context vectors from the two attentions are concatenated and used in the decoder.

The structure of the decoder is the same as that of the decoder of the Tacotron 2 described in [IV-A] except that a self-attention block is inserted after the LSTMs. This block is expected to capture the long-term dependency inside the output sequence. The number of heads in the multi-head attention in the self-attention is 2, and the dimension is 1,568 (1024 + 512 + 32).

C. GLOBAL STYLE TOKENS WITH TACOTRON 2 AND SA-TACOTRON

We also use GSTs [4] to enrich the speaking style of synthesized speech and make characters distinguishable from each other. The GST framework was proposed as a prosody transfer approach. In this framework, we assume that TTS systems have access to a reference audio file from which we can borrow prosody and the speaking style, which are transferred to synthetic speech produced by the TTS system. Its role is to extract the prosody and speaking style that cannot be explained by the text input.

The architecture we used is basically the same as the original one, except some parameters (Figure 5). An input or reference audio sequence, 80-dimensional mel spectrogram, is passed into a reference encoder. The reference encoder is composed of 6-layer 2D convolution layers with batch normalization and a 128-unit gated recurrent unit (GRU) [43]. Each convolution layer is made up of $3 \times 3$ filters with $2 \times 2$ stride (same padding) and ReLU activation. Batch normalization is applied to each layer. The number of filters in the layers are 128, 128, 256, 256, 512, and 512. The output of the final layer is passed into the GRU. The final state of the GRU is then passed into a network called style token layer.

The style token layer is composed of 10 randomly initialized 512-dimensional embeddings called style tokens and a multi-head attention. The output of the reference encoder and tanh-activated tokens are then mapped by the multi-head attention. Any number of heads of the attention can be used if the dimension of style tokens can be divided by this number. If the number of head is $h$, the dimension of tokens is $512/h$. We use 8 heads in this study based on the results of preliminary experiments. The attention calculates the weights over tokens, and the weighted summation of tokens is treated as a style embedding, which will be concatenated to the phoneme embeddings output from the encoder of the Tacotron-2-based or SA-Tacotron-based model. The style embedding vector is constant within a sentence.

V. EXPERIMENTS

A. PURPOSE OF EXPERIMENTS

We modeled rakugo speech with two different types of models, Tacotron-2-based [IV-A] and SA-Tacotron-based [IV-B] models. We conducted a listening test to compare their performances and assess how they are accepted by the public.

Since rakugo is a form of entertainment and the conversations or dialogues of the characters make the story progress, we are interested in not only the naturalness of synthesized speech, but also how accurately listeners distinguish characters, how well listeners understand the content of the speech, and how well the speech entertains listeners.

We also analyzed alignment errors of synthesized speech. Almost all the encoder-decoder sequence-to-sequence TTS including the Tacotron-based and SA-Tacotron ones above use a soft attention mechanism to map each time step of the encoder and that of the decoder. In speech synthesis, alignments should be left-to-right. However, the soft attention mechanism does not have such restriction, so it may cause alignment errors. The stop token used in the Tacotron-based and SA-Tacotron-based models may fail to predict the termination of a sentence.
In addition, we analyzed pitch-accent errors. As mentioned in III-A, our database does not have accent labels due to practical limitations, even though Japanese is a pitch-accent language; therefore, pitch-accent errors can occur more easily.

B. MODELS AND SAMPLES USED IN EXPERIMENTS

We used 16 of all 25 stories in the database because the annotation of the database is a work in progress. They are about 4.3 hours long except for pauses between sentences, and contain 7,341 sentences. We used 6,437 sentences for training, 715 for validation, and 189 for testing. The training and validation sets did not include very short (< 0.5 s) or very long (≥ 20 s) speech to reduce alignment errors during training.

We trained several models for the experiments.

- **Tacotron** is a Tacotron-2-based model, and no style embeddings or context features are input.
- **Tacotron-GST-8** is a Tacotron-2-based model with GSTs with an 8-head multi-head attention. It should be noted that the reference audio of the test set is natural speech itself.
- **Tacotron-ATTR** is a Tacotron-2-based model with manually labeled context features belonging to ATTR (role and individuality) only. The dimension of context embedding is 4.
- **Tacotron-context** is a Tacotron-2-based model with all the manually labeled context features. The dimension

2This is a choice by design since this makes comparisons of *-GST-8 with other systems (*-ATTR and *-context) that use manually labeled context features fairer.
of context embedding is 68.

- **Tacotron-GST-8-ATTR** and **Tacotron-GST-8-context** is a Tacotron-2-based model with a combination of GSTs with an 8-head multi-head attention and manually labeled context features belonging to ATTR, all contexts, respectively.
- **SA-Tacotron, SA-Tacotron-GST-8, SA-Tacotron-ATTR, SA-Tacotron-context, SA-Tacotron-GST-8-ATTR, and SA-Tacotron-GST-8-context** are the same models as Tacotron, Tacotron-GST-8, Tacotron-ATTR, Tacotron-context, Tacotron-GST-8-ATTR, and Tacotron-GST-8-context, respectively, but they are based on SA-Tacotron instead of Tacotron 2.

We trained each model for about 2,000 epochs (mini-batch size: 128, 120,000 steps) with an initial learning rate of 0.001 and exponential decay of learning rate. The optimization method was Adam. The number of mel filters for input spectrograms was 80. The spectrograms were converted from 48 kHz/16 bit waveforms with 50 ms-long frame, 12.5 ms frame shift, Hann window, and 4,096-long fast Fourier transform. Values of the spectrograms were transformed into 0 mean and 1 standard deviation at each dimension over all the data.

Predicted mel spectrograms were converted into waveforms by using a WaveNet vocoder trained with natural mel spectrograms and waveforms of all the training, validation, and test sets. The sampling rate of the output waveform was 16 kHz.

**C. LISTENING TEST**

1) **Experimental conditions**

We selected a set of sentences comprising a short story as materials for the listening test. A total of 189 sentences comprising 13 short stories were prepared, and sentence-level audio files were concatenated as one audio file per story. Because the speech samples were predicted sentence by sentence, and pauses between sentences were not predicted, the pauses between sentences used in the test were the same as those of real audio recordings. These should be predicted using models, but that is out of the scope of this paper. Listeners evaluated speech NOT sentence by sentence but in a whole story. Analysis-by-synthesis (AbS; copy synthesis) speech was also used for the test. The concatenated audio files were normalized to $-26\,\text{dB}_{\text{vol}}$ by sv56 [47].

We conducted MOS tests. In each evaluation round, listeners listened to the speech of all the 13 stories each synthesized using one of the models listed in Table 4 for AbS speech. For each listener, the story-system combinations and their permutation were randomly selected. One of the speech was presented on each screen, and listeners answered four MOS-based questions:

1) How accurately do you think you can distinguish each character?
2) How accurately do you think you can understand the content?
3) How well do you think you are entertained?
4) How well do you think you can understand the content?

We used a five-point MOS scale. A listener was allowed to answer one evaluation round only because listeners would remember the content of the stories. A total of 183 listeners participated in 183 evaluation rounds.

2) **Results**

The results are shown in Figure 7. For statistical analysis, we conducted Brunner-Munzel test [48] with Bonferroni correction among the scores for all the model combinations. For Q1–Q3, we can see AbS speech was superior to all the Tacotron-based and SA-Tacotron-based models. Regarding Q4, AbS speech was also superior to many of the Tacotron-based and the SA-Tacotron-based models, but there are no significant differences between some of the (SA-)Tacotron models and AbS speech.

The score range for Q4 was 0.7–0.8 lower than those for Q1–Q3.

3) **Discussions**

As mentioned in [4], Tacotron 2 can produce speech as natural as human in the case of well-articulated read speech. However, in the case of rakugo speech, all the TTS models including Tacotron 2 could not achieve the same MOSs regarding naturalness as that for AbS speech. Regarding distinguishability of characters and understanding content, there were also significant differences between the MOS for each model and that for AbS. In other words, speech synthesis currently cannot reach the professional level of rakugo performance. These results match to our prediction.

The most important point of our experiment was whether the listeners were entertained by synthetic rakugo speech, and we obtained an interesting result: there were no significant differences among the MOSs for some of the TTS models and that for AbS speech. This means that speech synthesis can have a role beyond just information transfer.

For further analysis, we calculated correlation relationships among Q4, the question for evaluating entertaining, and the other questions. The results are listed in Table 6. The correlation coefficient between Q4 and Q2 (distinguishability of characters) and that between Q4 and Q3 (understanding content) were higher than that between Q4 and Q1 (naturalness). This suggests that we should improve the distinguishability of characters and understandability of content, not naturalness, of synthesized speech to further entertain listeners.

**D. ERROR ANALYSIS**

1) **Details of analysis**

To understand the results in detail, we calculated alignment and pitch-accent error rates for the test set.
To calculate alignment error rates, the first author carefully listened to all the synthesized speech for the test sentences produced with each model and checked whether any alignment errors occur per sentence. Prolonging phonemes, skipping phonemes, repeating phrases, and late terminations were defined as alignment errors.

The first author also carefully listened to all the synthesized speech for the test sentences produced with each model and counted the number of pitch-accent errors. We regarded a pitch-accent phrase pronounced in a different accent from that of the recorded speech as an accent error even if the synthesized accent is acceptable as natural Japanese. The total number of pitch-accent phrases in the test set was 1,089.

2) Results

Alignment and pitch-accent error rates for the test set are listed in Tables 4 and 5 respectively.
TABLE 4. Alignment error rates (%) for the test set.

| System                  | Total | Prolonging phonemes | Skipping phonemes | Repeating phrases | Late termination |
|-------------------------|-------|---------------------|-------------------|-------------------|-----------------|
| Tacotron                | 8.5   | 1.6                 | 4.8               | **0.5**           | 1.6             |
| Tacotron-GST-8          | 11.1  | 3.2                 | 6.4               | 1.6               | 1.1             |
| Tacotron-ATTR           | 11.6  | 1.1                 | 9.0               | 2.7               | 0.5             |
| Tacotron-context        | 7.4   | 2.1                 | 4.2               | 1.1               | **0.0**         |
| Tacotron-GST-8-ATTR     | 11.6  | 2.1                 | 8.5               | 1.6               | 1.1             |
| Tacotron-GST-8-context  | 11.6  | 3.2                 | 5.8               | 3.2               | 1.1             |
| SA-Tacotron             | 10.1  | 3.2                 | 5.8               | 2.1               | **0.0**         |
| SA-Tacotron-GST-8       | 8.5   | 2.1                 | 4.8               | 2.1               | **0.0**         |
| SA-Tacotron-GST-8-ATTR  | 10.1  | 1.6                 | 7.4               | 1.6               | 1.1             |
| SA-Tacotron-context     | 9.0   | 2.1                 | 5.8               | 1.1               | 0.5             |
| SA-Tacotron-GST-8-ATTR  | **6.9** | **0.0**           | **2.7**           | 1.0               | 3.2             |
| SA-Tacotron-GST-8-context | 11.1 | 2.7                 | 5.3               | 4.2               | 0.5             |

TABLE 5. Pitch-accent error rates (%) for the test set.

| System                  | Pitch-accent error rates (%) |
|-------------------------|------------------------------|
| Tacotron                | 13.9                         |
| Tacotron-GST-8          | 14.1                         |
| Tacotron-ATTR           | 12.3                         |
| Tacotron-context        | 11.7                         |
| Tacotron-GST-8-ATTR     | 14.3                         |
| Tacotron-GST-8-context  | 14.6                         |
| SA-Tacotron             | 14.9                         |
| SA-Tacotron-GST-8       | 15.4                         |
| SA-Tacotron-ATTR        | 16.3                         |
| SA-Tacotron-context     | **8.8**                     |
| SA-Tacotron-GST-8-ATTR  | 9.6                          |
| SA-Tacotron-GST-8-context | 15.6             |

We can see that both error rates differed by system. However, there was no obvious relationships between alignment or pitch-accent error rates and the MOSs of the listening test.

3) Discussions

The reason why there were no obvious relationships between the error rates and the MOSs is probably that the evaluation was done paragraph by paragraph and not sentence by sentence. The listeners would make an effort to evaluate the whole paragraph and not focus on detailed errors. It is unknown how alignment or pitch-accent errors affect the degree of entertainment, fewer errors will be better. Self-attention, GSTs, and manually labeled context features, all of which are technical improvements we attempted in this paper, reduced errors as a whole. However, just a combination of them (SA-Tacotron-GST-8-context) was not the best model. Further technical development and finding the best combination are future tasks.

VI. CONCLUSION

Toward a speech synthesis entertaining audiences, we modeled rakugo speech using the state-of-the-art Tacotron 2 and an enhanced version of it with self-attention (SA-Tacotron) to better consider long-term dependency and compared their outputs. We first built a rakugo database because no suitable databases or commercial recordings existed. We then trained TTS models with Tacotron 2 and the SA-Tacotron. We also used global style tokens (GSTs) and manually labeled context features to enrich speaking styles. Through a listening test, we found that the TTS models could not produce professional-quality speech, but interestingly, some of the synthetic speech entertained the listeners as well as the analysis-by-synthesis speech. From the listening test results, no models was obviously superior to the other models. However, self-attention, GSTs, and manually labeled context features reduced alignment and pitch-accent errors as a whole.

We succeeded to synthesize somewhat natural and comic rakugo speech that entertains audiences. We believe this is an important step toward realization of genuine entertaining TTS. Our future plans are as follows: 1) modeling pauses between sentences to achieve fully-synthesized rakugo speech, 2) estimating speaking styles from manually labeled context features and/or input texts, and 3) synthesizing longer stories and evaluating them.

REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Sauvour, Y. Agiomyrgiannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Calgary, AB, Canada, Apr 15–20 2018, pp. 4779–4783.
[2] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, “Neural speech synthesis with transformer network,” in Proc. AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu, HI, USA, Jan 27 – Feb 1 2019.
[3] Y. Wang, R. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Sauvour, “Uncovering latent style factors for expressive speech synthesis,” in Proc. Conference on Neural Information Processing Systems (NIPS) Machine Learning for Audio Signal Processing Workshop, Long Beach, CA, USA, Dec 4–9 2017, pp. 4006–4010.
[4] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Sauvour, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in Proc. International Conference on Machine Learning (ICML), Stockholm, Stockholm, Sweden, Jul 10–15 2018, pp. 5180–5189.
[5] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Sauvour, “Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,” arXiv:1803.09047 [cs.CL], Mar 24 2018.
[6] T. Wood, “Varying speaking styles with neural text-to-speech,” Alexa Blogs, Nov 19 2018. [Online]. Available: https://developer.amazon.com/zh/blogs/alexa/post/7ab9665a-0536-4be2-aaad-18281ec59af8/varying-speaking-styles-with-neural-text-to-speech
[7] W.-N. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, “Hierarchical generative modeling for controllable speech synthesis,” in Proc. International Conference on Learning Representations (ICLR), New Orleans, LA, USA, May 6–9 2019.
